Springer Proceedings in Mathematics & Statistics

Ke Chen *Editor*
Anton Ravindran *Managing Editor*

Forging Connections between Computational Mathematics and Computational Geometry

Papers from the 3rd International Conference on Computational Mathematics and Computational Geometry

Springer

Springer Proceedings in Mathematics & Statistics

Volume 124

More information about this series at http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Ke Chen *Editor*
Anton Ravindran *Managing Editor*

Forging Connections between Computational Mathematics and Computational Geometry

Papers from the 3rd International Conference on Computational Mathematics and Computational Geometry

Springer

GSTF
STEERING INNOVATION, SERVING SOCIETY

Editor
Ke Chen
Department of Mathematical Sciences
The University of Liverpool
Liverpool, UK

Managing Editor
Anton Ravindran
President, Global Science and Technology Forum
Singapore

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-16138-9 ISBN 978-3-319-16139-6 (eBook)
DOI 10.1007/978-3-319-16139-6

Library of Congress Control Number: 2015958098

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Foreword

This volume of conference proceedings contains a collection of research papers presented at the 3rd Annual International Conference on Computational Mathematics, Computational Geometry & Statistics (CMCGS 2014) organized by Global Science and Technology Forum, held in Singapore on 3–4 February 2014.

The CMCGS 2014 conference is an international event for the presentation, interaction, and dissemination of new advances relevant to computational mathematics, computational geometry, and statistics research. As member of the Board of Governors, GSTF, I would like to express my sincere thanks to all those who have contributed to the success of CMCGS 2014.

A special thanks to all our speakers, authors, and delegates for making CMCGS 2014 a successful platform for the industry, fostering growth, learning, networking, and inspiration. We sincerely hope you find the conference proceedings enriching and thought-provoking.

Preface

We are pleased to welcome you to the 3rd Annual International Conference on Computational Mathematics, Computational Geometry & Statistics (CMCGS 2014) organized by Global Science and Technology Forum, held in Singapore on 3–4 February 2014.

The CMCGS 2014 conference continuously aims to foster the growth of research in mathematics, geometry, statistics, and its benefits to the community at large. The research papers published in the proceedings are comprehensive in that it contains a wealth of information that is extremely useful to academics and professionals working in this and related fields.

It is my pleasure to announce the participation of leading academics and researchers in their respective areas of focus from various countries at this event. The Conference Proceedings and the presentations made at CMCGS 2014 are the end result of a tremendous amount of innovative work and a highly selective review process. We have received research papers from distinguished participating academics from various countries. There will be "BEST PAPER AWARDS" for authors and students, to recognize outstanding contributions and research publications.

We thank all authors for their participation and we are happy that they have chosen CMCGS 2014 as the platform to present their work. Credit also goes to the Program Committee members and review panel members for their contribution in reviewing and evaluating the submissions and for making CMCGS 2014 a success.

Anton Ravindran

Program Committee

PROGRAM CHAIR

Dr. Jörg Fliege
Professor
School of Mathematics
The University of Southampton, UK

Co-EDITORs-IN-CHIEF

Prof. Ke Chen
Director of Centre for Mathematical Imaging Techniques
Department of Mathematical Sciences
The University of Liverpool, UK

Prof. C. Raju
Professor, Quantitative Methods & Operations Management
Chairman—Post Graduate Programme
Indian Institute of Management Kozikode, India

PROGRAMCOMMITTEE MEMBERS

Prof. Marc Demange
Professor of Operations Research
ESSEC Business School, Paris

Prof. Luca Bonaventura
Research Assistant Professor of
Numerical Analysis
Laboratory for Modeling and
Scientific Computing MOX
Politecnico di Milano, Italy

Dr. Pamini Thangarajah
Associate Professor/Mathematics
Coordinator
Department of Mathematics, Physics
and Engineering
Mount Royal University, Calgary,
Alberta, Canada

Dr. Selin Damla Ahipasaoglu
Assistant Professor
Engineering Systems and Design
Singapore University of Technology
and Design

Prof. Jun Zou
Department of Mathematics
The Chinese University of Hong
Kong, Hong Kong

Prof. B. Bollobás
Honorary Professor
Department of Pure Mathematics &
Mathematical Statistics
University of Cambridge, UK

Prof. Hassan Ugail
Director, Centre for Visual
Computing
University of Bradford, UK

Dr. Ping Lin
Professor, Department of
Mathematics
University of Dundee, UK

Dr. Julius Kaplunov
Professor
Applied Mathematics
Brunel University, UK

Dr. R. Ponalagusamy
Professor
Department of Mathematics
National Institute of Technology
Tiruchirappalli, India

Dr. A. K. Singh
Professor
Department of Mathematics
Banaras Hindu University
Varanasi, India

Dr. Nandadulal Bairagi
Associate Professor & Coordinator
Centre for Mathematical
Biology and Ecology
Jadavpur University
Kolkata, India

Dr. Kallol Paul
Associate Professor
Department of Mathematics
Jadavpur University
India

Dr. Khanindra Chandra Chowdhury
Department of Mathematics
Gauhati University
India

Dr. D. Deivamoney Selvam
Professor, Department of Mathematics
National Institute of Technology
Tiruchirappalli, India

Contents

Part II Pure Mathematics

Part III Computational Geometry

Part IV Statistics

Part I
Computational Mathematics

An Augmented Lagrangian Approach with Enhanced Local Refinement to Simulating Yield Stress Fluid Flows Around a Spherical Gas Bubble

Jianying Zhang

Abstract We simulate the flow of a yield stress fluid around a gas bubble using an augmented Lagrangian approach with piecewise linear equal-order finite elements for both the velocity and the pressure approximations. An enhanced mesh adaptive strategy based on this element-pair choice is also proposed to render the yield surfaces with desired resolution. The corresponding numerical scheme is formulated for general Herschel–Bulkley fluids. Numerical results on Bingham fluid flows around a slowly rising spherical gas bubble are provided to showcase the improvement on the previously proposed algorithm in (Zhang, Int J Numer Methods Fluids 69:731–746, 2012).

Keywords Viscoplastic fluids • Yield surfaces • Bingham fluid flows • Augmented Lagrangian method • Gas bubble dynamics

Introduction

A viscoplastic fluid is a complex fluid with yield stress. It deforms only when the applied shear stress exceeds the yield stress. Many multicomponent industrial fluids are viscoplastic [1]. The existence of the yield stress allows a gas bubble of a certain size or shape to be trapped in a viscoplastic fluid when the buoyancy of the bubble is insufficient to break the yield stress or undergo deformations when the surface tension becomes less dominant compared to the yield stress effect of the surrounding fluid [2].

Theoretically, viscoplastic fluids are generalized Newtonian fluids governed by discontinuous constitutive laws which lead to implicitly defined and highly nonlinear viscous terms in the corresponding momentum equations. This makes the simulation of viscoplastic fluid flows rather difficult. Our main interest in this work

J. Zhang (✉)
Department of Mathematics, Western Washington University, Bellingham, WA, USA
e-mail: Jianying.Zhang@wwu.edu

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.04_1

is to capture the yield surfaces around a fixed-shape bubble. A detailed review of various numerical approaches to simulating viscoplastic fluid flows in the literature can be found in [3]. To keep the actual viscoplastic feature of the fluid of interest, we are in favor of the variational approach in the presented work.

Mathematical Formulation

The Constitutive Laws for the Liquid and the Gas Regions

We consider the flow of a yield stress fluid around a slowly rising spherical gas bubble in a long cylindrical tube. The center of the bubble is located on the symmetric axis of the tube. In the following, the subscript "l" associated with a physical quantity indicates that the quantity is in the liquid region and the subscript "g" indicates it in the gas region. Consequently, we denote the liquid and gas densities by ρ_l and ρ_g, the liquid and gas viscosities by μ_l and μ_g.

The yield stress fluids considered here are generalized Newtonian fluids, a class of non-Newtonian fluids. For such fluids, the rate of strain $\dot{\gamma}_{ij}$ and the deviatoric stress τ_{ij} are related through a constitutive equation of form $\tau_{ij} = \eta\left(\dot{\gamma}\right)\dot{\gamma}_{ij}$ with $\dot{\gamma} = \sqrt{\frac{\dot{\gamma}_{ij}\dot{\gamma}_{ij}}{2}}$, where $\eta = \eta\left(\dot{\gamma}\right)$ is termed the effective viscosity. Denote the second invariant of the deviatoric stress by $\dot{\tau} = \sqrt{\frac{\tau_{ij}\tau_{ij}}{2}}$. If $\tau\left(\dot{\gamma} \to 0^+\right) = \tau_Y > 0$, then the models are viscoplastic, with yield stress τ_Y.

A typical viscoplastic model is the Herschel–Bulkley model with the following scaled constitutive relation:

$$\tau_{ij} = \left(\gamma^{n-1} + \frac{B}{\gamma}\right)\gamma_{ij} \quad \text{if } \tau > B, \gamma = 0 \quad \text{if } \tau \leq B$$

with n being the power-law index.

This is an extension of the power-law model to a fluid with a yield stress $\widehat{\tau}_Y$. The dimensionless parameter $B = \frac{\widehat{\tau_Y}\widehat{R}}{\widehat{\mu}\widehat{V}}$, termed the Bingham number, denotes the ratio of yield stress to viscous stress. Here $\widehat{\mu}$ represents the kinematic viscosity, $\widehat{R}$ and $\widehat{V}$ are the reference spatial and velocity scales, respectively.

For the Herschel–Bulkley model, the effective viscosity is defined from $\tau = \eta\left(\dot{\gamma}\right)\dot{\gamma}$. Setting $n = 1$ and $B = 0$ returns the Newtonian model, $\eta = 1$. Setting $n = 1$, we recover the popular Bingham model. Note that for the Herschel–Bulkley model, if $B > 0$, then $\eta \to \infty$ as $\gamma \to 0$.

The gas inside the bubble is assumed to be Newtonian, which leads to the constitutive equation $\tau_{g,ij} = \mu_g\dot{\gamma}_{g,ij}$.

Equations of Motion and Boundary Conditions

We study the behavior of viscoplastic fluid around a slowly rising gas bubble, for which the corresponding equations of motion and boundary conditions are derived in [2]. Let $u(x_1, x_2, x_3) = (u_1(x_1, x_2, x_3), u_2(x_1, x_2, x_3), u_3(x_1, x_2, x_3))$ be the velocity of the two-phase fluid at the location $x = (x_1, x_2, x_3)$. The gravitational force is assumed to be the only body force acting on the two-phase system.

The equations of motion in the fluid region Ω_l are given by the generalized incompressible Navier–Stokes equations:

$$\rho_l \frac{Du_i}{Dt} = -\frac{\partial p}{\partial x_i} + \rho_l g_i + \frac{\partial \tau_{l,ij}}{\partial x_j}, \quad \nabla \cdot u = 0 \tag{1}$$

with $\frac{Du_i}{Dt} = \frac{\partial u_i}{\partial t} + u \cdot \nabla u_i$ being the material derivative and $g = (0, 0, -9.8)^T$ the acceleration of the gravitational field.

The gas region Ω_g is generally compressible and the Navier–Stokes equations associated with the mass conservation inside the bubble are given by

$$\rho_g \frac{Du_i}{Dt} = -\frac{\partial p}{\partial x_i} + \rho_g g_i + \frac{\partial \tau_{g,ij}}{\partial x_j},$$

$$\frac{\partial \rho_g}{\partial t} + \nabla \cdot \left(\rho_g u\right) = 0. \tag{2}$$

In (1) and (2), $\frac{\partial p}{\partial x_i}$, for $i = 1, 2, 3$, represent the three components of the gradient field of the pressure p. With Einstein's notation, the viscous terms $\frac{\partial \tau_{k.ij}}{\partial x_j}$, for $k = l, g$ and $i = 1, 2, 3$, stand for the divergences of the three row vectors in the deviatoric stress tensors $\tau_{l,ij}$ and $\tau_{g,ij}$ correspondingly. Due to the unknown yield surfaces, these viscous terms are implicitly defined in (1).

The boundary of the fluid region Ω_l, denoted by $\partial\Omega$, consists of two portions: One is the collection of the cylinder walls, denoted by $\partial\Omega_w$; the other is the bubble surface, denoted by $\partial\Omega_g$. That is, $\partial\Omega = \partial\Omega_l \cup \partial\Omega_g$. Boundary conditions on $\partial\Omega$ shall be specified as complements for (1) and (2).

The no-slip boundary condition is imposed on the cylinder walls, i.e., $u = 0$ on $\partial\Omega_w$. In addition, a set of jump conditions are imposed on the bubble surface. Let n_b be the outer unit normal on the bubble surface; the continuity of the velocity across the bubble surface amounts to

$$\left(u_l - u_g\right) \cdot n_b = 0 \quad \text{on} \quad \partial\Omega_g.$$

For $k = l, g$ and $i = 1, 2, 3$, let $\sigma_{k,ij}(u) = -p_k I + \tau_{k,ij}(u)$ be the stress tensor for the corresponding liquidor gas region, where p_l and p_g stand for the pressure in the liquid and the gas regions, respectively. Let t_1 and t_2 be two linearly independent unit tangent vectors on the bubble surface. The continuity of the tangential traction

$$t_i^T\left(\sigma_{l,ij}-\sigma_{g,ij}\right)\cdot n_b=0$$

and the jump of the normal traction

$$n_b^T\left(\sigma_{l,ij}-\sigma_{g,ij}\right)\cdot n_b=\xi\left(\frac{1}{R_1}+\frac{1}{R_2}\right)$$

are also imposed, where ξ is the surface tension, R_1 and R_2 are the radii of curvature in the principle directions, and $\frac{1}{R_1}+\frac{1}{R_2}$ is then twice the mean curvature of the bubble surface.

For a viscoplastic fluid around a slowly rising gas bubble, the nondimensionalized equations of motion are derived and simplified in [2] as

$$0=-\frac{\partial p}{\partial x_i}+g_i+\frac{\partial \tau_{l,ij}}{\partial x_j},\quad \nabla\cdot u=0 \tag{3}$$

and

$$0=-\frac{\partial p}{\partial x_i}+\varepsilon\rho_g g_i+\delta\Delta u_i,\quad \nabla\cdot u=0 \tag{4}$$

with the boundary conditions

$$t_i^T\tau_{l,ij}n_b=0,\quad -p_l+p_g+n_b^T\tau_{l,ij}n_b=\beta\left(\frac{1}{R_1}+\frac{1}{R_2}\right).$$

Here $g=(0,0,-1)^T$ is the unit acceleration of the gravitational field, B is the Bingham number, and β is the dimensionless surface tension. The dimensionless parameters $\varepsilon=\frac{\rho_g^*}{\rho_l}$ and $\delta=\frac{\mu_g R_0^{n-1}}{\mu_l U^{n-1}}$ assuming the bubble radius at the injection pressure is R_0, the steady-state bubble velocity is U, and the gas density at the injection pressure is ρ_g^*.

The Augmented Lagrangian Method

The variational reformulation and its application to Bingham fluid flows date back to the pioneer work of Duvaut and Lions [4], in which a desired flow motion is captured by solving an equivalent variational inequality whose minimizer set is proven to be the solution set of the momentum equations with the associated constitutive law. For a viscoplastic fluid, the discontinuity of the constitutive law brings in a nonlinear and non-differentiable yield stress term in the corresponding variational reformulation. The augmented Lagrangian method (ALM) [5, 6], as an effective numerical technique, resolves this difficulty by introducing an auxiliary variable to relax the undesired yield stress term and then adding an augmented constraint. Consequently, the original problem is decoupled into a series of element-wise optimization problems, each of which can be solved with standard optimization techniques. This is also the virtue of the ALM.

For a slowly rising gas bubble, its motion can be assumed axisymmetric. That is, the velocity field u is identical in each cross section along the radial direction. Hence, u can be written as $u\,(r,\theta,z) = (u_1\,(r,z)\,,0,u_3\,(r,z))$ in cylindrical coordinates (r,θ,z). It is then enough to simulate the original three-dimensional two-phase system, set up in Ω, in one of its cross sections, denoted by D. In the following presentation, we use D_l and D_g to denote the liquid and gas regions in the corresponding cross section. This assumption simplifies the three-dimensional vector field to two-dimensional. However, the rate of strain tensor $\dot{\gamma}_{ij}\,(u)$ is still a 3×3 symmetric matrix.

Define the admissible set

$$A = \left\{v \in \left(H_0^1(D)\right)^2 : \nabla \cdot v = 0 \ \text{ in } \ D\right\}.$$

It can be shown via integration by parts that the desired vector field u for the steady-state two-phase problem (3)–(4) with the boundary conditions is the one that satisfies the following constrained variational inequality:

$$a\,(u,v-u) + j(v) - j(u) \geq L\,(v-u)\,, \quad \forall v \in \mathrm{A} \tag{5}$$

where $j(v) = B\int\limits_{D_l} \dot{\gamma}(v)$,

$$a\,(u,v) = \frac{1}{2}\left(\int\limits_{D_l} \dot{\gamma}^{n-1}(u)\dot{\gamma}_{ij}\,(u)\dot{\gamma}_{ij}\,(v) + \delta\int\limits_{D_g} \dot{\gamma}^{n-1}(u)\dot{\gamma}_{ij}\,(u)\dot{\gamma}_{ij}\,(v)\right).$$

$$L(v) = -\int\limits_{D_l} v_3 - \varepsilon\rho_g\int\limits_{D_g} v_3 - \int\limits_{\partial D_g} \beta\left(\frac{1}{R_1} + \frac{1}{R_2}\right)v \cdot n_b$$

Note that $a\,(\cdot,\cdot)$, referred to as the viscous dissipation rate in some of the literature, is linear in its argument v for general Herschel–Bulkley fluids and bilinear in either of its argument for Bingham fluids, i.e., when $n = 1$. The force term $L\,(\cdot)$ is linear in its argument, whereas the yield stress dissipation rate $j\,(\cdot)$ is nonlinear and non-differentiable in its argument.

It can be shown [6] that the desired vector field u in (5) is the one that minimizes

$$J(v) = \frac{1}{n+1}a\,(v,v) + j(v) - L(v) \tag{6}$$

over the function space A. Our numerical algorithm is based on this variational equality. For Bingham fluids, the existence and uniqueness of the minimizer can be shown by directly applying Theorem 4.1 and Lemma 4.1 from Chapter 1 of [6].

The ALM is implemented following the Uzawa-type iterations:

Step 1: Solve an elliptic problem for the velocity. The finite element method is naturally preferred.

Step 2: Update the pressure based on the incompressible constraint.

Step 3: Solve element-wise optimization problems for the rate of strain tensor.

Step 4: Update the Lagrange multiplier corresponding to the augmented constraint.

Various ways of choosing the finite element spaces for the velocity and the pressure exist, such as $P_2 - P_1$, $P_1 - C_0$, based on the Babuska–Brezzi stability condition [7]. Differently, our work is based on that of Latch'e and Vola [8] where the piecewise linear equal-order $P_1 - P_1$ element spaces are chosen for the velocity and the pressure, and the piecewise constant approximation is made for the rate of strain tensor. This arrangement is cost-effective and matches the coherence constraint between velocity and the rate of strain perfectly. In addition, an implicit updating scheme [8] is applied in Step 2 in order to accelerate and stabilize the convergence of the pressure.

Numerical Implementation of ALM

Difficulty in solving the optimization problem (6) is caused by the nonlinear and non-differentiable yield stress term $j(\cdot)$. This can be resolved by applying the ALM [6].

Let W be the collection of symmetric 3×3 tensors with L^2 entries. The key idea of the ALM is to relax the nonlinear yield stress term $\dot{\gamma}_{ij}(u)$ to an auxiliary term $w_{ij} \in W$ so that the minimization problem (6) can be reformulated into a constrained minimization problem [3] solved following the Uzawa-type iterations.

Initialize p, w_{ij}, and s_{ij}.

Step 1: With fixed p, w_{ij}, and s_{ij}, solve the following for u:

$$\begin{aligned} -\int_D p\nabla \cdot v + \int_{D_l} v_3 + \varepsilon\rho_g \int_{D_g} v_3 + \int_D \dot{\gamma}_{ij}(v)s_{ij} \\ + R\int_D \left(\dot{\gamma}_{ij}(u) - w_{ij}\right)\dot{\gamma}_{ij}(v) = 0, \quad \forall v \in V. \end{aligned} \tag{7}$$

Step 2: With fixed u and s_{ij}, solve the following for w_{ij}:

$$\min_{w_{ij}\in W} \frac{1}{n+1}\left(\int\limits_{D_l} w^{n+1} + \delta \int\limits_{D_g} w^{n+1}\right) + B\int\limits_{D} w - \int\limits_{D} w_{ij}s_{ij} \\ -R\int\limits_{D} \dot{\gamma}_{ij}(u)w_{ij} + R\int\limits_{D} w^2. \tag{8}$$

Step 3: With fixed u, update p based on the globally stabilized constraint

$$\int\limits_{D} q\nabla\cdot u + c_b\int\limits_{D} \nabla p\cdot\nabla q = 0, \quad \forall q\in H_0^1(D) \tag{9}$$

where the second term on the left-hand side of the equation is a Brezzi–Pitkäranta stabilization term used for the pressure updating, with $c_b > 0$ being the stabilization parameter [8].

A larger value of c_b yields faster convergence in the pressure update, but the result could be too much off the true value. So in practical c_b should not be set too large in order to keep the real structure of the pressure.

Step 4: With fixed u and w_{ij}, update $s_{ij}\in W$ based on the yield stress constraint

$$\int\limits_{D} \left(\dot{\gamma}_{ij}(u) - w_{ij}\right)t_{ij} = 0, \quad \forall t_{ij}\in W. \tag{10}$$

The virtue of the above decomposition coordination process is to decouple the original problem (6) into several subproblems (7)–(10), each of which can be solved with less effort.

Yield Surface Detection with an Enhanced Mesh Adaptive Strategy

We propose the following refinement algorithm as an enhancement of the local refinement algorithm proposed in [3], based on the solution property determined by the piecewise linear equal-order element pair.

Initialize with a uniform unstructured triangulation of some pre-chosen size h.

Step 1: For each edge, compute its larger and smaller $\dot{\gamma}$ values, V_w^1 and V_w^2. The larger and smaller $\dot{\gamma}$ values for an edge is defined as the larger and smaller $\dot{\gamma}$ values in the adjacent triangles sharing this common edge.

Step 2: For each edge, compute the ratio $R_w = \frac{V_w^1 - V_w^2}{L}$ with L being the length of the edge. R_w is a reasonable indicator of the corresponding edge location. For a common edge inside either the fluid or the solid region, R_w should be moderately small. An edge with larger ratio R_w is supposed to be closer to a yield surface.

Step 3: For each element, make the longest edge refinement if the largest R_w value among its three edges is larger than some preset value C_r.

Numerical Results

Consider a gas bubble of unit radius in a cylindrical container filled with Bingham fluids of different Bingham numbers. The base radius and the height of the cylinder are set to 5 and 10 units, respectively. The coordinates of the system are set such that both the centers of the bubble and of the cylinder are located at the origin. The surface tension is set to 0 for simplification. It can be added though without crucial numerical efforts.

The analytical solution for Newtonian fluid flow around a gas bubble rising at a steady speed U has been derived in [9]. The validity of our numerical algorithm in the Newtonian case has been confirmed in [3] by comparing the corresponding U numerical and analytical results with $U = 1/3$.

In Bingham fluids, we still set $U = 1/3$ at the instant of computing. Unlike it is in the Newtonian case, the bubble actually tends to deform due to the non-Newtonian effects of the surrounding fluid flows. In [10], the transient deformations of a gas bubble in various Bingham fluids have been investigated. Our goal is to understand how the unyielded regions are affected by the Bingham number B for a fixed-shape spherical gas bubble. Of course this captures the flow behavior only at a certain moment.

Adopting the enhanced mesh adaptivity technique introduced above, yield surfaces around the gas bubble in various Bingham fluids can be captured with desirable resolution more effectively. The number of elements drops from 7,000–8,000 range as required in [3] to 4,500–5,500 here in order to achieve comparable resolution, which considerably lowers the numerical cost.

Yield surfaces in Bingham fluids with $B = 0.08, 0.12, 0.15, 0.2, 0.4, 0.55$ are detected and shown in Fig. 1. In addition, yield surfaces in Bingham fluids with $B = 0.542, 0.544, 0.546, 0.549$ are detected and shown in Fig. 2. A “+” sign is plotted at the center of each element in which the numerical indicator of the second invariant of the deviatoric stress $\dot{\gamma} < TOL$, some preset threshold. We set $TOL = 10^{-8}$ here. The density of the plots provides additional information on the unyielded regions. The highly concentrated plots generate the yield surfaces, the boundaries between the solid and the fluid regions, whereas the sparse plotted regions are considered to lie in the interior of the unyielded regions.

As shown in the figures, the yield surfaces first expand towards the gas bubble as B increases. During the evolution of the unyielded region, small inner unyielded regions next to the left and right of bubble surface arise as B increases beyond 0.15 and then expand towards the outer unyielded region with the increment of B. The inner and outer regions eventually merge as B keeps increasing. This growth property of the unyielded regions is consistent with the one observed in [10]. Yet the growth becomes less noticeable as B reaches 0.55, and the entire fluid region is

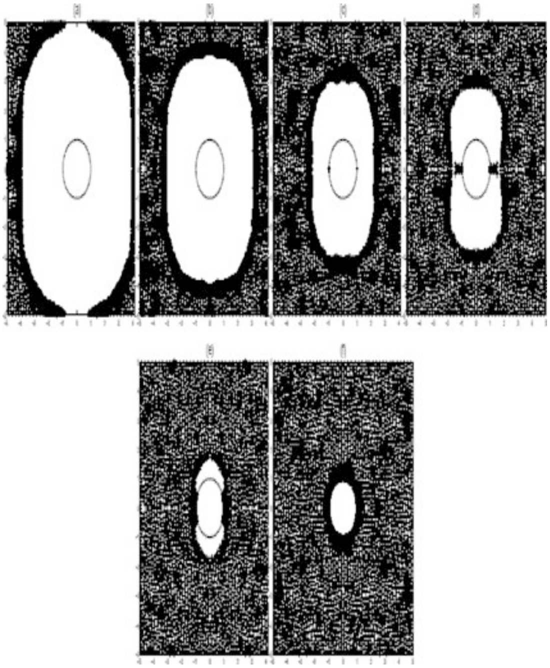


Fig. 1 Yield surfaces around a unit radius gas bubble in various Bingham fluid flows in a cylinder with radius 5 and height 10. The corresponding Bingham numbers are (**a**) B = 0.08, (**b**) B = 0.12, (**c**) B = 0.15, (**d**) B = 0.2, (**e**) B = 0.4, and (**f**) B = 0.55

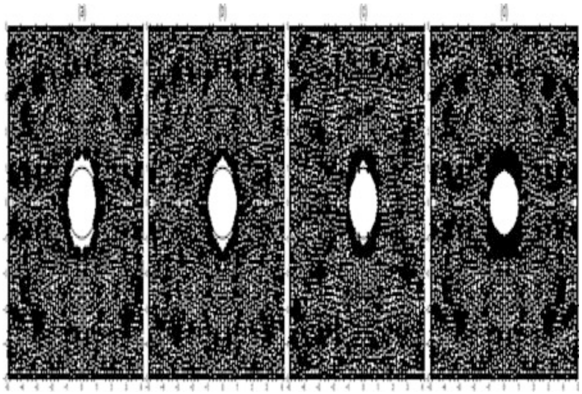


Fig. 2 Yield surfaces around a unit radius gas bubble in various Bingham fluid flows in a cylinder with radius 5 and height 10. The corresponding Bingham numbers are (**a**) B = 0.542, (**b**) B = 0.544, (**c**) B = 0.546, and (**d**) B = 0.549

observed to be unyielded based on our computation. The evolution of yield surfaces in Bingham fluids with B values right below this critical number can be clearly seen in Fig. 2.

Our result numerically verifies that $B = 0.55$ is the critical Bingham number at which the gas bubble gets completely trapped. This is consistent with the conclusion drawn in [3], considered as an improvement of the sufficient but unnecessary stopping criterion, $B > 1/\sqrt{3}$, analytically derived in [2].

Conclusions and Future Investigations

The key feature of a bubble propagating through a viscoplastic fluid is that the yield stress can prevent it from moving. The proposed numerical algorithm allows us to numerically determine the minimal Bingham number at which a spherical bubble gets completely trapped in a Bingham fluid. With the enhanced mesh adaptivity strategy, the yield surfaces can be detected with desired resolution and at a faster speed.

As one of the future investigations, it would be interesting to extend the presented work to a dynamical two-phase system, in which a gas bubble is moving in viscoplastic fluids and undergoing a series of deformations. This can be achieved by coupling the momentum equations with the level set equation [11, 12] that tracks the bubble interface.

In addition, the proposed mesh adaptivity strategy is obviously not optimal. More enhanced mesh adaptivity strategies are to be developed in order to further improve the resolution of the yield surfaces while keeping the numerical cost as low as possible.

Acknowledgments The author would like to thank Dr. I. Frigaard for valuable discussions and suggestions during her stay as a postdoctoral researcher at the University of British Columbia. The author is also grateful for Dr. D. Vola's generous help with some implementation issues on the related numerical algorithm.

References

1. Bird, R.B., Dai, G.C., Yarusso, B.J.: The rheology and flow of viscoplastic materials. Rev. Chem. Eng. **1**, 1–70 (1982)
2. Dubash, N., Frigaard, I.A.: Conditions for static bubbles in visco-plastic fluids. Phys. Fluids **16**(12), 4319–4330 (2004)
3. Zhang, J.: An augmented Lagrangian approach to simulating yield stress fluid flows around a spherical gas bubble. Int J Numer Methods Fluids **69**, 731–746 (2012)
4. Duvaut, G., Lions, J.L.: Inequalities in Mechanics and Physics. Springer, Berlin (1976)
5. Fortin, M., Glowinski, R.: Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems. North-Holland, Amsterdam (1983)
6. Glowinski, R.: Numerical Methods for Nonlinear Variational Problem. Springer, New York (1984)
7. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer, New York (1991)
8. Latch'e, J.-C., Vola, D.: Analysis of the Brezzi-Pitkäranta stabilized Galerkin scheme for creeping flows of Bingham fluids. SIAM J. Numer. Anal. **42**(3), 1208–1225 (2004)
9. Batchelor, K.: An Introduction to Fluid Dynamics. Cambridge University Press, Cambridge (1967)
10. Tsamopoulos, J., Dimakopoulos, Y., Chatzidai, N., Karapetsas, G., Pavlidis, M.: Steady bubble rise and deformation in Newtonian and viscoplastic fluids and conditions for bubble entrapment. J. Fluid Mech. **601**, 123–164 (2008)
11. Sussman, M., Smereka, P., Osher, S.: A level set approach for computing solutions to incompressible two-phase flow. J. Comput. Phys. **114**, 146–159 (1994)
12. Zapryanov, Z., Tabakova, S.: Dynamics of Bubbles, Drops and Rigid Particles. Fluid Mechanics and its Applications. Kluwer Academic, Boston (1999)

On Classical Solution in Finite Time of BGK-Poisson's Equations

Slim Ben Rejeb

Abstract BGK model is a collision operator for the evolution of gases which satisfies several fundamental properties. Different collision operators for gas evolutions have been introduced earlier, but none of them could satisfy all the basic physical properties: conservation, positivity, correct exchange coefficients, and entropy inequality. However, contrary to Boltzmann model which has a quadratic form, the BGK model presents a heavy nonlinearity which explains the complexity of this analysis.

The existence of solution to the periodic Boltzmann BGK model (Bhatnagar–Gross–Krook) coupled with Poisson's equation has been proved by the same author.

In this paper we are interested to the uniqueness of such solution.

Keywords Kinetic equations • BGK model • Maxwellian • Boltzmann's equations • Plasma's physics • Schauder's fixed point • Poisson's equation • Fluid equations

Introduction

We study the initial value problem of BGK model [2] coupled with Poisson's equation, which is a simple relaxation model introduced by Bhatnagar, Gross, and Krook to mimic Boltzmann flows, where $f(x, v, t)$ is the density of plasma particles at time t in the space of position x and velocity v and $\phi(x, t)$ is the electric field potential of the plasma.

The existence and uniqueness problems of the BGK model were proved by Perthame and Pulvirenti [4] but without coupling with Poisson's equation.

Ukay and Okabe [5] had proved the existence and uniqueness of (f, ϕ) for the Vlasov–Poisson equation (without collision term).

S.B. Rejeb (✉)
Department of Mathematics, Science Faculty, Jazan University, P.O. Box 2097,
Jazan, Kingdom of Saudi Arabia
e-mail: benrejeb_slim@yahoo.fr

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.14_2

Here we have a complete problem, BGK model coupled with Poisson's equation. We will start by giving the construction of the solution [1], and in the second time, we prove the uniqueness of such solution.

In periodic Case the dimensionless Boltzmann BGK model coupled with Poisson's equation in one space dimension is written as

$$\begin{cases} L_f f = M[f] - f, & (x,v) \in \Omega \times \mathbb{R},\ t \geq 0 \\ f(t=0) = f_0(x,v), & x \in \Omega,\ v \in \mathbb{R} \\ -\phi_{xx} = \int_{\mathbb{R}} f(x,v,t)dv, & \phi(0) = \phi(L) = 0. \end{cases} \tag{1}$$

$\Omega =]0, L[$.

where

$$L_f = \partial_t + v\partial_x + E(x,t)\partial_v \tag{2}$$

$$E(x,t) = -\frac{1}{2}\phi_x \tag{3}$$

$$M[f] = \frac{\rho}{(2\pi T)^{1/2}} \exp(-\frac{|u-v|^2}{2T}) \tag{4}$$

$M[f]$ is the Maxwellian associated to f, where

$$(\rho, \rho u, \rho(u^2+T)) = \int_{\mathbb{R}} (1, v, v^2) f(v)dv \tag{5}$$

Remark 1. The notation L_f for the deferential operator in (2) is chosen to see that it depends on f according to (3) and (1).

Existence of Solution

The electric field potential is given by

$$\phi(x,t) = \int_{\Omega} G(x-y)(\int_{\mathbb{R}^2} f(y,v,t)dv)dy \tag{6}$$

where G is the fundamental solution of Δ_x (see [5] or [1]).

Using ϕ, we can solve the initial value problem of the first- order partial differential equation:

$$\begin{cases} L_f f + f = M[f], & (x, v) \in \Omega \times \mathbb{R}, \quad t \geq 0 \\ f(t = 0) = f_0(x, v), & x \in \Omega, \quad v \in \mathbb{R} \end{cases} \tag{7}$$

We can easily solve Eq. (7) using the characteristics $(X(t), V(t))$ which will be denoted as (X_t, V_t):

$$\begin{cases} \dfrac{dX}{dt} = V(t), \quad X(s) = x, \\ \dfrac{dV}{dt} = E(X(t), t), \quad V(s) = v. \end{cases} \tag{8}$$

The solution of Eq. (7) is given implicitly as

$$\begin{aligned} f(x, v, t) = e^{-t} f_0(X(0, x, v, t), V(0, x, v, t)) \\ + \int_0^t e^{-(t-s)} M[f](X_s, V_s, s) ds \end{aligned}$$

In this way we have assigned a function f to a given function g which we will denote by $f = \Phi(g)$. So we shall specify a set S of functions g in such a way that the map Φ defined on S can be shown to have a fixed point, with the aid of Schauder's fixed-point theorem and that any fixed point of Φ in S is a classical solution of (1).

Class of Functions

For any set $\Xi \subset \mathbb{R}^2 \times \mathbb{R}^+$, we denote $B^{l+\sigma}(\Xi)$ the set of all continuous and bounded functions defined on Ξ having continuous and bounded l^{th} derivatives which are uniformly Holder continuous in Ξ with exponent σ, where l is a positive integer and $0 \leq \sigma \leq 1$.

Notations

For $f \in L^1(\mathbb{R}),\ q \geq 0,\ f \geq 0$, we denote
$N_q(f) = \sup_{v \in \mathbb{R}}(|v|^q f(v))\text{and}\mathbb{N}_q(f) = \sup_{v \in \mathbb{R}}((1 + |v|^q) f(v)$ and for $\tau \geq 0$, we introduce

$$\Omega_\tau = \Omega \times]0, \tau[\text{ and } Q_\tau = \Omega \times \mathbb{R} \times]0, \tau[.$$

Lemma 1. *For $f \geq 0$ and $f(v) \in L^1(\mathbb{R}, (1+v^2)dv)$, we have*

$$\begin{aligned} &i)\ \frac{\rho}{T^{1/2}} \leq N_0(f), \\ &ii)\ \rho(T+u^2)^{\frac{(q-1)}{2}} \leq C_q N_q(f),\ \ q > 3, \\ &iii)\ \sup_{v\in\mathbb{R}}\{|v|^q M[f]\} \leq C_q N_q(f),\ \ q > 3. \end{aligned} \tag{9}$$

where (ρ, u, T) are given by (5). See [4] for the proof.

Proposition 2. *Suppose that f is a solution of (1). Then*

$$\mathbb{N}_q(f) \leq C_q \exp(C_q t) \text{ for } q = 0 \text{ or } q \geq 3$$

Proof. $i)$ We shall first prove the case $q = 0$,

From (9) (i) we have $M[f] \leq CN_0(f)$ where C is a positive constant. From (1) we have

$$\begin{aligned} &\frac{d}{dt}(e^t f(X_t, V_t, t)) \leq Ce^t \sup_{x\in\Omega} N_0(f)(t), \\ &\Rightarrow e^t f(X_t, V_t, t)) \leq f_0(x, v) + C\int_0^t e^s \sup_{x\in\Omega} N_0(f)(s)ds, \\ &\Rightarrow e^t f(x, v, t)) \leq f_0(X(0, x, v, t), V(0, x, v, t)) \\ &\quad +C \textstyle\int_0^t e^s \sup_{x\in\Omega} N_0(f)(s)ds, \\ &\Rightarrow e^t \sup_{x\in\Omega} N_0(f)(t) \leq ||f_0||_\infty + C\int_0^t e^s \sup_{x\in\Omega} N_0(f)(s)ds. \end{aligned}$$

The Gronwall lemma ended the proof.

ii) case where $q > 3$

We denote $f_q = (1 + |v|^q)f$; writing the equation verified by f_q, we get

$$L_f f_q = (1 + |v|^q)M[f] - f_q + v|v|^{(q-2)}Ef \tag{10}$$

where E can be written as

$$E(x,t) = \int_\Omega K(x,y)(\int_\mathbb{R} g(y,v,t)dv)dy. \tag{11}$$

K is a bounded kernel and can be easily deduced from (3) and (6). We can easily see that

$$|E(x,y)| \leq \sup_{(x,y)\in\Omega} |K(x,y)|||f_0||_{L^1(\Omega\times\mathbb{R})} \tag{12}$$

For the values of $|v| \geq 1$, we have

$$L_f f_q + f_q \leq C_q \mathbb{N}_q(f) \tag{13}$$

The Gronwall lemma applied to the map
$t \longrightarrow e^t \sup\limits_{x \in \Omega} \mathbb{N}_q(f)(t)$ gives the proof.

The case $|v| < 1$ is easy to prove from (i).

Lemma 3. *We suppose that there exist $V_1 \in \mathbb{R}$ and $C_0 > 0$, such that:*

i) *f_0 is not depending on x,*
ii) *f_0 is increasing at $]-\infty, V_1[$ and decreasing at $]V_1, +\infty[$,*
iii) $\displaystyle\int_{|v-V_1|>2\tau} f_0(v)dv \geq C_0$.

Then

$$\rho(t) \geq C_0 e^{-t}.$$

The proof is detailed in [1].

Proposition 4. *For $f_0 \geq 0$ we suppose:*

i) *$f_0 \in B^1(\Omega \times \mathbb{R})$,*
ii) *there exist $A_0 > 0$ such that,*

$$\sup_{v \in \mathbb{R}} \{(1 + |v|^q) f_0(x, v)\} = A_0 < +\infty,$$

then $\forall t \in [0, \tau]$, $\exists A(t) < +\infty,\ B(t) \in \mathbb{R}_+^*$ verify

$$\begin{aligned} &i)\ \ 0 < B(t) \leq T(t) \leq A(t), \\ &ii)\ \ u(t) \leq A(t). \end{aligned} \tag{14}$$

Proof. From (9) (i), we get

$$T^{1/2} \geq C \frac{\rho}{N_0(f)} \geq C_1(t),$$

and from (9) (ii) we get (14) (i) and (ii).
Indeed

$$\rho(T + u^2)^{\frac{q-1}{2}} \leq C_q N_q(f),$$

Thus,

$$(T + u^2) \leq A(t).$$

Definition 5. We denote S as the class of functions satisfying

$$S = \{g \in B^{\delta}(Q_\tau); ||g||_{B^{\delta}(Q_\tau)} \leq A_1, \sup_v((1 + |v|^q)g) \\ \leq A_2, \forall (x,t) \in \Omega \times]0, \tau[\},$$

where A_1 and A_2 are positive constants.

For $g \in S$ we consider f a solution of

$$\begin{cases} L_g f = M[g] - f, & (x,v) \in \Omega \times \mathbb{R}, t \geq 0 \\ f(t=0) = f_0(x,v), & x \in \Omega, v \in \mathbb{R} \\ -\phi_{xx} = \int_{\mathbb{R}} g(x,v,t)dv, & \phi(0) = \phi(L) = 0 \end{cases} \tag{15}$$

and

$$E(x,t) = \int_{\Omega} K(x,y)(\int_{\mathbb{R}} g(y,v,t)dv)dy.$$

We denote

$$f = \Phi(g)$$

We have to prove that Φ is a continuous map from S to itself, which will prove the existence of a solution in S.

The solution of (15) is given by

$$f(x,v,t) = e^{-t} f_0(X(0,x,v,t), V(0,x,v,t)) \\ + \int_0^t e^{-(t-s)} M[g](X_s, V_s, s)ds \tag{16}$$

where we noted: From (9) (i), we get

$$M[g] \leq \frac{\rho(x,t)}{T^{1/2}} \leq C N_0(g)$$

by virtue of (16) and the condition imposed to f_0 in proposition (4), it's easy to see that $f \in S$ if $g \in S$.

We consider a sequence $g^n \in S$ and $g^\infty \in B^0(Q_\tau)$ verify $||g^n - g^\infty||_{B^0(Q_\tau)} \longrightarrow 0$ when $n \longrightarrow +\infty$.

Lemma 6. *S is a compact convex subset of $B^0(Q_\tau)$*

The proof is detailed in [1].

Theorem 7. *With the conditions of proposition (4) for f_0, the problem (1) has one solution (f, ϕ).*

Before the proof, we introduce these notations.

Notations

First, we denote

$$f^n = \Phi(g^n) \qquad \text{and} \qquad f^\infty = \Phi(g^\infty)$$

and for any function $\mathcal{F}$

$$\Delta\mathcal{F} = \mathcal{F}^n - \mathcal{F}^\infty.$$

Proof.

$$\begin{aligned} e^t \Delta f(x, v, t) = f_0(X_0^n, V_0^n) - f_0(X_0^\infty, V_0^\infty) + \\ \int_0^t e^s [M[g^n](X_s^n, V_s^n, s) - M[g^\infty](X_s^\infty, V_s^\infty, s)] ds \end{aligned} \tag{17}$$

On one hand, we have

$$f_0(X_0^n, V_0^n) - f_0(X_0^\infty, V_0^\infty) \leq ||f_0||_{B^1(\Omega\times\mathbb{R})} N_n^s(X, V)$$

On the other hand,

$$\begin{aligned} &M[g^n](X_s^n, V_s^n, s) - M[g^\infty](X_s^\infty, V_s^\infty, s) \\ &\quad = (M[g^n] - M[g^\infty])(X_s^n, V_s^n, s) + \\ &\quad M[g^\infty](X_s^n, V_s^n, s) - M[g^\infty](X_s^\infty, V_s^\infty, s) \end{aligned}$$

By virtue of (14) (i), it is easily seen that $M[g]$ has at least the same regularity as $g \in S$ then

$$\begin{aligned} &|M[g^\infty](X_s^n, V_s^n, s) - M[g^\infty](X_s^\infty, V_s^\infty, s)| \leq \\ &\quad ||M[g^\infty]||_{B^\delta(Q_\tau)} N_n^s(X, V)^\delta \end{aligned}$$

It remains then to estimate the term:

$$(M[g^n] - M[g^\infty])(X_s^n, V_s^n, s).$$

We pose for $\theta \in [0,1]$,

$$(\rho^n_\theta, u^n_\theta, T^n_\theta) = \theta(\rho^\infty, u^\infty, T^\infty) + (1-\theta)(\rho^n, u^n, T^n)$$

We denote M^n_θ the Maxwellian associated to $(\rho^n_\theta, u^n_\theta, T^n_\theta)$. We have

$$\begin{aligned} |M[g^n] - M[g^\infty]|(X^n_s, V^n_s, s) \leq |\Delta\rho(X^n_s, s)\frac{\partial M^n_\theta}{\partial \rho}| \\ +|\Delta u(X^n_s, s)\frac{\partial M^n_\theta}{\partial u}| + |\Delta T(X^n_s, s)\frac{\partial M^n_\theta}{\partial T}| \end{aligned} \tag{18}$$

the derivatives of M^n_θ verify:

$$\begin{aligned} &|\frac{\partial M^n_\theta}{\partial \rho}| \leq C(T^n_\theta)^{-1/2}, \\ &|\tfrac{\partial M^n_\theta}{\partial u}| \leq C\rho^n_\theta(T^n_\theta)^{-1}, \\ &|\tfrac{\partial M^n_\theta}{\partial T}| \leq C\rho^n_\theta(T^n_\theta)^{-3/2}. \end{aligned}$$

To conclude from (18), we shall need the estimates:

$$(i) : |\Delta\rho(X^n_s, s), \quad (ii) : |\Delta u(X^n_s, s)|, \quad (iii) : |\Delta T(X^n_s, s)|$$

which can be estimated by

$$\begin{aligned} &(i) : |\Delta\rho(X^n_s, s)| = |\int_{\mathbb{R}} (g^n - g^\infty)(X^n_s, w, s)dw| \\ &(ii) : |\Delta u(X^n_s, s)| \leq C|\Delta u(X^n_s, s)\rho^\infty(X^n_s, s)| \leq \\ &\qquad C|(\Delta(\rho u)u^\infty)(X^n_s, s)| + |u^n\Delta\rho(X^n_s, s)| \\ &(iii) : |\Delta T(X^n_s, s)| \leq C|\Delta T(X^n_s, s)\rho^\infty(X^n_s, s)| \leq \\ &\qquad C|\Delta(\rho T))(X^n_s, s)| + |T^n(\Delta\rho)(X^n_s, s)|. \end{aligned}$$

Hence $q > 3$, the dominated convergence theorem applied to (17) ended the proof.

We conclude that Φ is a continuous map in S; then it has a fixed point in S, which is a solution of BGK-Poisson's equation (1).

The Schauder's fixed-point theorem does not allow to show the uniqueness of the solution; this question will be solved in the next section.

Remark 8. The conditions (*ii*) and (*iii*) imposed to f_0 in lemma (3) can be generalized as

$(ii)'$ There is a finite sequence $(V_n)_{n\in\mathbb{N}}$ such that f_0 is increasing at $]-\infty, V_0[$ and decreasing at $]V_n, +\infty[$.

$(iii)'$ There exist $C_0 > 0$ such that $\displaystyle\int_{-\infty}^{V_0-2\tau} f_0(v)dv + \int_{V_n+2\tau}^{+\infty} f_0(v)dv \geq C_0$

Remark 9. The condition (ii) imposed to f_0 is not excessive; the distribution of particles has generally this shape, like Gaussian curves, for example.

Uniqueness of the Solution

Assume, in addition to the hypothesis in proposition (4), the following conditions on f_0:

$$\begin{aligned} &\partial_x f_0 \in L^1(\mathbb{R}) \cap C^1(\Omega \times \mathbb{R}) \\ &\sup_{v\in\mathbb{R}}(1 + |v|^q)\partial_x f_0 < A_0 \end{aligned} \tag{19}$$

The aim of this section is to prove the uniqueness of the solution (f, ϕ) to (1) constructed above in class of such functions that

$$\begin{aligned} &i) \quad f \in B^0(Q_\tau) \cap C^1(Q_\tau) \\ &ii) \quad \rho = \int_{\mathbb{R}} f dv \in B^\delta([0, \tau]; B^1(\Omega)) \cap B^\delta(\Omega_\tau) \end{aligned} \tag{20}$$

Let $(f_i, \phi_i), i = 1, 2$ be any two solutions of (1) satisfying (20).

Subtracting the two equations for $i = 1$ and $i = 2$, writing $f = f_1 - f_2$ and $\phi = \phi_1 - \phi_2$, we obtain

$$\begin{cases} L_{f_1} f + (E_1 - E_2)\partial_v f_2 = M[f_1] - M[f_2] - f, \\ f|_{t=0} = 0 \\ -\phi_{xx} = \int_{\mathbb{R}} f(x, v, t) dv \end{cases} \tag{21}$$

Integrating the first equation with respect to v (first fluid equation), we get

$$\begin{cases} \partial_t \rho + \partial_x(\rho u) = 0 \\ \rho(t = 0) = 0 \end{cases} \tag{22}$$

which can be written as

$$\begin{cases} \partial_t \rho + u\partial_x \rho = -(\partial_x u)\rho \\ \rho(t = 0) = 0 \end{cases} \tag{23}$$

$$\begin{cases} \dfrac{d}{dt}\rho(x_t, t) = -\partial_x u(x_t, t)\rho(x_t, t) \\ \rho(t = 0) = 0 \end{cases} \tag{24}$$

where x_t are the characteristic associated to (23).

Finally, we can write

$$\begin{cases} \rho(x, t) = -\displaystyle\int_0^t \partial_x u(\tilde{x}_s, s)\rho(\tilde{x}_s, s)ds \\ \rho(t = 0) = 0 \end{cases} \tag{25}$$

$\tilde{x}(t)$ are the characteristics adjusted according to their properties (see [1]):

$$\begin{cases} \sup\limits_{x\in\Omega} \rho(x, t) \leq \displaystyle\int_0^t \sup\limits_{x\in\Omega} |\partial_x u(x, s)| \sup\limits_{x\in\Omega} \rho(x, s)ds \\ \rho(t = 0) = 0 \end{cases} \tag{26}$$

Thus, with the aid of Gronwall lemma, we can prove that $\rho(x, t) = 0$ which imply that $f(x, v, t) = f_1(x, v, t) - f_2(x, v, t) = 0$ and gives the uniqueness of the solution.

Finally, from (12) we can see that the electric field potential $\phi(x, t) = \phi_1(x, t) - \phi_2(x, t) = 0$.

Theorem 1. *According to the conditions of (4) and (19), the solution of (1) constructed in theorem (7) is unique.*

Conclusion

We have proved the uniqueness of the solution of BGK model coupled with Poisson's equation.

This document completes the result of the existence, given in [1] by the same author where we constructed a solution and demonstrated the existence to such a problem.

References

1. Ben Rejeb, S.: On the existence of conditions of a classical solution of BGK-Poisson's equation in finite time. Int. J. Appl. Math. **39**, 4 (2009)
2. Bhatnagar, P.L., Gross, E.P., Krook, M.: A model for collision processes in gases. Phys. Rev. **94**, 511–525 (1954)

3. Saint-Raymond, L.: From the BGK Boltzmann model to the Euler equations of incompressible fluids. Bull. Sci. Math. **126**(6), 493–506 (2002)
4. Perthame, B., Pulviremti, M.; Weighted L^{∞} bounds and uniqueness for the Boltzmann B.G.K model. Arch. Rat. Mech. Anal. **125**(3), 289–295 (1993)
5. Ukai, S., Okabe, T.: On classical solutions in the large in time of the tow-dimensional Vlasov's equation. Osaka J. Math. **15**, 245–261 (1978)

A Note on Lanczos Algorithm for Computing PageRank

Kazuma Teramoto and Takashi Nodera

Abstract We now study the Lanczos algorithm for computing the PageRank vector. This algorithm is based on biorthogonalization, which transforms a nonsymmetric matrix into a tridiagonal matrix to compute PageRank. This generates better approximation of the largest eigenvalue at early stage of iterations. We propose a practical scheme of the Lanczos biorthogonalization algorithm with SVD scheme for computing PageRank. Numerical results show that the proposed algorithm converges faster than the existing Arnoldi method in the computation time.

Keywords PageRank • Lanczos method • Eigenvalue problem

Introduction

PageRank is the essential approach for ranking a Web page whereby a page's status id decided according to the link structure of the Web. This model has been used by Google as a part of its contemporary search engine equipment. Nowadays, the precise ranking procedures and computation schemes used by Google are no longer public evidence, but the PageRank model has taken on the life of its own and has received important consideration in the scientific and technology society in the last 10 years. PageRank is essentially the fixed distribution vector of the Markov chain whose transition matrix is a convex combination of a Web link graph and precise rank-one matrix. A major parameter in the model is a damping factor, a scalar that settles the weight given to a Web link graph in the model. The weighted PageRank

K. Teramoto (✉)
Graduate School of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama 223-8522, Japan
e-mail: trmtkzm011010@a5.keio.jp

T. Nodera
Department of Mathematics, Keio University, 3-14-1, Hiyoshi, Kohoku, Yokohama 223-8522, Japan
e-mail: nodera@math.keio.ac.jp

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.15_3

is the elements of the dominant eigenvector of the modified adjacency matrix as follows:

$$A = \alpha P + (1 - \alpha) E$$

where P is a column stochastic matrix, α is a damping factor, and E is a rank-one matrix. The specified derivation can be seen in Kamvar et al. [7].

Recently, the computation of eigenpair (eigenvalue and eigenvector) of nonsymmetric matrices is one of the most major tasks in many scientific and technology applications. A typical example, nowadays, is the computation of PageRank for the link structure of the Web. Due to the great size and sparsity of the matrix, factorization schemes are considered infeasible. Instead, iterative schemes are used, where the computation is dominated by matrix–vector products. Detailed descriptions of this problem are available, and the algorithms can be found in lots of references (i.e., [2, 5–9]), where P is a column stochastic matrix, α is a damping factor, and E is a rank one matrix. The specified derivation can be seen in Kamvar et al. [7].

In recently, the computation of eigenpair (eigenvalue and eigenvector) of nonsymmetric matrices are one of the most major tasks in many scientific and technology applications. Typical example, nowadays, is the computation of PageRank for the link structure of Web. Due to the great size and sparsity of the matrix, factorization schemes are considered infeasible. Instead, iterative schemes are used, where the computation is dominated by matrix–vector products. Detailed descriptions of this problem are available, and the algorithms can be found in lots of references (i.e., [2, 5–9]).

The power method was firstly considered for computing PageRank. For detailed properties of PageRank for using the power method, we refer the reader to Kamvar et al. [7]. However, the power method has its disadvantage. For some given matrices, the power method converges very slowly.

Although methods were suggested to accelerate its convergence, improvement was not important. As a selection, a procedure using orthogonalization such as the Arnoldi method was suggested [6].

In this paper, we propose and investigate a new algorithm for computing the PageRank vector, using a combination of the Lanczos biorthogonalization algorithm and SVD (singular value decomposition). This remainder of the paper is organized as follows. In the section "Arnoldi Method," we sketch the brief description of Arnoldi method of the PageRank vector. Then in the section "Lanczos Algorithm for Computing PageRank," we will propose a new Lanczos algorithm with SVD scheme. In the section "Numerical Experiments," the results of numerical experiments obtained by running the MATLAB codes are reported. At last, in the section "Conclusion," we draw some conclusions.

Arnoldi Method

In this section, we describe the brief introduction of the Arnoldi method [1] for computing the PageRank vector. The Arnoldi method, which is given Algorithm 1, builds an orthonormal basis for Krylov subspace given by

$$K_m\left(A, \boldsymbol{q_0}\right) = \left\{\boldsymbol{q_0}, A\boldsymbol{q_0}, \ldots, A^{m-1}\boldsymbol{q_0}\right\}$$

where the Krylov subspace is restricted to be of fixed dimension m and q_0 is an initial vector which satisfies $|\boldsymbol{q_0}|_2 = 1$. From Algorithm 1, the following relations hold:

$$AQ_m = Q_m H_m + h_{m+1}, m\boldsymbol{q_m}e_m^T$$

$$Q_m^T A Q_m = H_m$$

where $Q_m = [\boldsymbol{q_0}, \boldsymbol{q_1}, \ldots, \boldsymbol{q_m}] \in R_{n\times n}$ is a column–orthogonal matrix and $H_m = \left\{h_{i,j}\right\} \in R^{m\times m}$ is a Hessenberg matrix [6].

Since H_m is an orthogonal projection from A to K_m, we can use the eigenvalue of H_m as an approximate eigenvalue of A. If y is the eigenvector of H_m, then $Q_m\boldsymbol{y}$ is the approximate eigenvector of A, since it is known that the largest eigenvalue of a PageRank matrix is 1. The Arnoldi-type method was proposed by Golub and Greif [4] and we call it as Algorithm 2. In Algorithm 2, we are used to compute the singular value decomposition instead of computing the eigenvalue of H_m [6].

We can see that when m increases, the total computation cost of this method in every cycle is increasing, while the total iterations are decreasing. Unsuitably, it can be very difficult to know how to choose m a priori and if too small a value is chosen, the convergence may stall. Consequently, it is difficult to choose the optimal value of m to minimize the total computation cost (CPU time).

Algorithm 1 [Q_m, H_m]=Arnoldi(A, q_0, m)

```
Compute $q_1 = q_0/||q_0||_2$
for $j = 1, 2, ...,$ to $m$ do
    Compute $w_j = Aq_j$
    for $k = 1, 2...$ to $j$ do
        $h_{kj} = q_k^T w_j$
        $w_j = w_j - h_{kj} q_k$
    end for
    Compute $h_{j+1,j} = ||w_j||_2$
    if $h_{j+1,j} = 0$ then
        stop and exit
    else
        Set $q_{j+1} = w_j/h_{j+1,j}$
    end if
end for
```

Algorithm 2 Arnoldi-type method

```
Choose $\boldsymbol{q}_0$ with $\|\boldsymbol{q}_0\|_2 = 1$
for $l = 1, 2, \dots$, until convergence do
    Compute $[Q_m, H_{m+1,m}]$=Arnoldi$(A, \boldsymbol{q}_0, m)$
    Compute singular value decomposition $H_{m+1,m} - [I; 0] = U\Sigma V^T$
    Compute $\boldsymbol{q}_0 = Q_m \boldsymbol{v}_m$
    Compute $\boldsymbol{r} = \sigma_m Q_{m+1} \boldsymbol{u}_m$
    if $\|\boldsymbol{r}\|_1$ <TOL then
        stop and exit
    end if
end for
```

Lanczos Algorithm for Computing PageRank

The Lanczos biorthogonalization algorithm is an extension to nonsymmetric matrices of a symmetric Lanczos algorithm. One such extension, the Arnoldi method, has been seen in [10], but the nonsymmetric Lanczos biorthogonalization algorithm is quite different in the concept from Arnoldi method (see Saad [10, pp. 207–221]). In this section, we propose a new algorithm for computing PageRank by using the Lanczos algorithm. This algorithm is one of the Krylov subspace methods which transform symmetric matrices into tridiagonal matrices. To deal with nonsymmetric matrices, we will use the Lanczos biorthogonalization algorithm, an extension that allows the Lanczos algorithm to be applied to nonsymmetric matrices. This will build a pair of biorthogonal bases for the two Krylov subspaces as follows

$$K_m\left(A, \boldsymbol{v_1}\right) = \left\{\boldsymbol{v_1}, A\boldsymbol{v_1}, \dots, A^{m-1}\boldsymbol{v_1}\right\}$$

$$K_m\left(A^T, w_1\right) = \left\{\boldsymbol{w_1}, A^T\boldsymbol{w_1}, \dots, \left(A^T\right)^{m+1}\boldsymbol{w_1}\right\}$$

Also, from the Lanczos algorithm (Algorithm 3), the following three-term recurrence formulae are given as follows:

$$\begin{aligned}\widehat{\boldsymbol{v}}_{\mathbf{j+1}} &= \beta_{j+1}\boldsymbol{v}_{j+1} = A\boldsymbol{v}_j - \alpha_j\boldsymbol{v}_j - \beta_j\boldsymbol{v}_{j-1}\\ \widehat{\boldsymbol{w}}_{j+1} &= \delta_{j+1}\boldsymbol{w}_{j+1} = A^T\boldsymbol{w}_j - \alpha_j\boldsymbol{w}_j - \delta_j\boldsymbol{w}_{j-1}\end{aligned}$$

where $\alpha_j = \left(A\boldsymbol{v}_j, \boldsymbol{w}_j\right)$. Additionally, β_{j+1} and δ_{j+1} are a parameter for normalizing $\boldsymbol{v}_{j+1}$ and $\boldsymbol{w}_{j+1}$. They satisfy the following equalities:

$$\boldsymbol{w}_{j+1} = \frac{\widehat{\boldsymbol{w}}_{j+1}}{\beta_{j+1}}$$

$$\boldsymbol{v}_{j+1} = \frac{\widehat{\boldsymbol{v}}_{j+1}}{\delta_{j+1}}$$

Algorithm 3 Lanczos bi-orthogonalization algorithm

```
Choose two vectors $\boldsymbol{v}_1$, $\boldsymbol{w}_1$ such that$(\boldsymbol{v}_1, \boldsymbol{w}_1) = 1$
Set  $\beta_1 = \delta_1 \equiv 0, \boldsymbol{w}_0 = \boldsymbol{v}_0 \equiv 0$
for $j = 1, 2, ...,$ to $m$ do
    $\alpha_j = (A\boldsymbol{v}_j, \boldsymbol{w}_j)$
    $\widehat{\boldsymbol{v}}_{j+1} = A\boldsymbol{v}_j - \alpha_j \boldsymbol{v}_j - \beta_j \boldsymbol{v}_{j-1}$
    $\widehat{\boldsymbol{w}}_{j+1} = A^T \boldsymbol{w}_j - \alpha_j \boldsymbol{w}_j - \delta_j \boldsymbol{w}_{j-1}$
    $\delta_{j+1} = |(\widehat{\boldsymbol{v}}_{j+1}, \widehat{\boldsymbol{w}}_{j+1})|^{1/2}$
    if $\delta_{j+1} = 0$ then
        then stop
    end if
    $\beta_{j+1} = (\widehat{\boldsymbol{v}}_{j+1}, \widehat{\boldsymbol{w}}_{j+1}) / \delta_{j+1}$
    $\mathbf{w}_{j+1} = \widehat{\boldsymbol{w}}_{j+1} / \beta_{j+1}$
    $\mathbf{v}_{j+1} = \widehat{\boldsymbol{v}}_{j+1} / \delta_{j+1}$
end for
```

where $\boldsymbol{v}_{j+1}$ and $\boldsymbol{w}_{j+1}$ can be selected in any manner to ensure that

$$\left(\boldsymbol{v}_{\mathrm{j}+1}, \boldsymbol{w}_{j+1}\right) = 1$$

As a result, β_{j+1} and δ_{j+1} satisfy the following equality:

$$\delta_{\mathrm{j}+1}\beta_{j+1} = \left(\widehat{\boldsymbol{v}}_{j+1}, \widehat{\boldsymbol{w}}_{j+1}\right)$$

As long as this condition is satisfied, β_{j+1} and δ_{j+1} are free to select. Accordingly, we select these formulae as follows:

$$\delta_{j+1} = \left|\left(\widehat{\boldsymbol{v}}_{j+1}, \widehat{\boldsymbol{w}}_{j+1}\right)\right|^{\frac{1}{2}},$$

$$\beta_{j+1} = \frac{\widehat{\boldsymbol{v}}_{j+1}, \widehat{\boldsymbol{w}}_{j+1}}{\delta_{j+1}}$$

As a result of above, the following equation is also satisfied:

$$\beta_{j+1} = \pm\delta_{j+1}$$

By using these formulae, we obtain Algorithm 3. Now we consider the following matrices:

$$V_m = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m)$$
$$W_m = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m)$$

Since these matrices are biorthogonal matrices, they satisfy the equality:

$$W_m^T W_m = I.$$

Algorithm 4 Lanczos algorithm for computing PageRank

```
Choose two vectors $\boldsymbol{p}_1$, $\boldsymbol{q}_1$ such that $(\boldsymbol{p}_1, \boldsymbol{q}_1) = 1$
[T,P,Q]=Lanczos(A, $\boldsymbol{p}_1$, $\boldsymbol{q}_1$, $m$)
Compute SVD $T_m - I = U\Sigma V^T$
Compute eigenvector $P\boldsymbol{v}_m$
```

Table 1 Computation cost of the Lanczos algorithm

Line	Operation	Cost
4	Matrix vector product	$3mn^2$
5	Matrix vector product and vector scaling	$2mn^2 + 2mn$
6	Matrix vector product and vector scaling	$2mn^2 + 2mn$
7	Vector product and norm computation	$3mn$
8	Vector product and vector scaling	$3mn$
9	Vector scaling	mn
10	Vector scaling	mn
Total		$7mn^2 + 12mn$

We define the following tridiagonal matrix T_m to use in the Algorithm 3.

$$T_m = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \delta_2 & \alpha_2 & \beta_3 & & \\ & & \cdots & & \\ & & \delta_{m-1} & \alpha_{m-1} & \beta_m \\ & & & \delta_m & \alpha_m \end{pmatrix}.$$

From the three-term recurrence formulae, the following relations hold:

$$\begin{aligned} AV_m &= V_m T_m + \delta_{m+1} \boldsymbol{v}_{\boldsymbol{m}+\boldsymbol{1}} e_m^T, \\ A^T W_m &= W_m T_m^T + \beta_{m+1} \boldsymbol{w}_{\boldsymbol{m}+\boldsymbol{1}} e_m^T, \\ W_m^T A V_m &= T_m \end{aligned}$$

Next, using the results obtained above, we propose the Algorithm 4 for computing the PageRank vector.

We now present the computation cost of the Lanczos algorithm in Table 1, m is the dimension of tridiagonal matrix, and $\boldsymbol{v}_{\boldsymbol{m}}$ is the right singular vector corresponding to the minimal singular value. The benefit of the Lanczos algorithm is that the total computation cost is small. But one drawback of the Lanczos algorithm is that we don't know the optimal value of m. If the value of m increases, accuracy may be improved, but the total computation cost also increases. From these facts, it can be very difficult to know how to choose the optimal value of m a priori.

Table 2 Iterations and computation time for test matrix 1 (4,298 × 4,298)

	Arnoldi		Lanczos	
α	*IT*	Time (s)	*IT*	Time (s)
0.85	4	0.884	10	0.677
0.90	4	0.904	10	0.677
0.95	4	0.904	10	0.677
0.99	5	1.121	10	0.677

Numerical Experiments

In this section, numerical results will be presented that compare the two methods described in the previous sections on the test problems. All computing of numerical experiments were done on the PC with 3.6 GHz and an eight-gigabyte memory in using MATLAB R2012b. We will show these results to demonstrate the efficiency of the Lanczos algorithm with SVD scheme. Here, we present that the test matrices, Death Penalty, are obtained from the Web page [3]. First of all, we choose the stopping criterion of the Arnoldi-type method as follows.

$$|\sigma_m \; Q_{m+1}\boldsymbol{u_m}|_1 \leq \; 1.0 \times 10^{-6}.$$

The computation cost of the Arnoldi method for one iteration,

$$2mn^2 + 2m\,(m+1)\,n + 3mn$$

(see [6]), and the Lanczos method is $7mn^2 + 12mn$ from Table 1 in the former section “Lanczos Algorithm for Computing PageRank.” The Lanczos algorithm requires more computation cost than the Arnoldi method, but the total computation cost is less than Arnoldi method (Table 2). However, m increases, the total computational cost is increasing, and more computational cost may require than the Arnoldi method. Hence, it is important to choose the optimal value of m. In order to choose the number of m and check the accuracy, we inspect the relative error between the exact PageRank vector and computation PageRank vector as follows:

$$|m - \boldsymbol{c}|_2,$$

where m is the exact PageRank vector and $\boldsymbol{c}$ is the computation PageRank vector. Test matrix 1 is an ill-conditioned matrix, and test matrix 2 is a well-conditioned matrix (see [3]). From Fig. 1, we can see that norm of relative error is completely different by each of the matrix. Consequently, it is difficult to choose the optimal value of m to minimize the total computation cost (CPU time), systematically. We need aped up to compute the approximate value of the PageRank vector. From these points, we consider the value of $m = 10$ as a practical value in these problems.

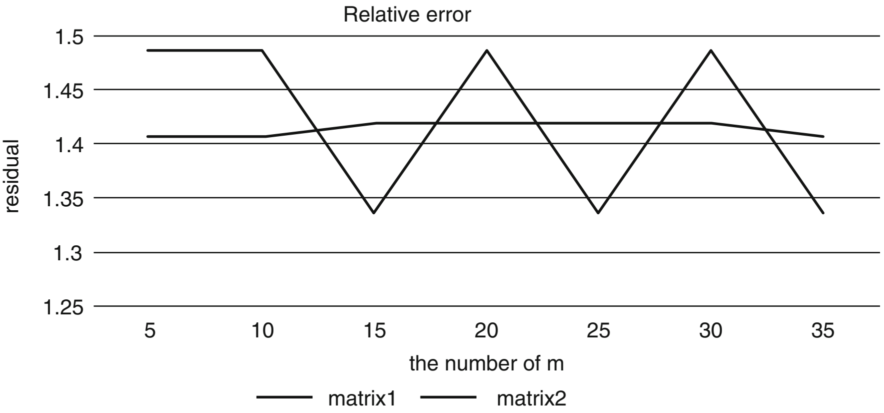


Fig. 1 Relative error vs. iterations for test matrix 1 and 2 ([3])

Conclusion

In this paper, we proposed the new algorithm to compute the PageRank vector, using a combined scheme of the Lanczos algorithm and SVD. Since the computation time is dependent on the dimension m of the tridiagonal matrix, numerical results showed that the computation time is constant. The proposed algorithm has the advantages which do not depend on α. If the accuracy of computation is not sufficient, we need to increase the number of α. Future work may include investigating how to adaptively estimate the parameter m and exploring the performance of the proposed algorithm as an acceleration technique for a variety of procedures of the PageRank computation. At last, there are lots of applications which are based on random walks. They are analogous in essence to the PageRank computation.

References

1. Arnoldi, W.E.: The principle of minimized iteration in the solution of the matrix eigenvalue problem. Q. Appl. Math. **9**, 17–29 (1951)
2. Bryan, K., Leise, T.: The $25,000,000,000 eigenvector: the linear algebra behind Google. SIAM Rev. **48**, 569–581 (2006)
3. Death Penalty, http://snap.stanford.edu/data/index.html
4. Golub, G.H., Greif, C.: An Arnoldi-type algorithm for computing PageRank. BIT Numer. Math. **46**, 756–771 (2006)
5. Gang, W., Yimin, W.: A power-Arnoldi algorithm for computing PageRank. Numer. Linear Algebra Appl. **14**, 521–546 (2007)
6. Yin, J.-F., Yin, G.-J., Ng, M.: On adaptively accelerated Arnoldi method for computing PageRank. Numer. Linear Algebra Appl. **19**, 73–85 (2012)
7. Kamvar, S.D., Haveliwala, T.H., Manning, C.D., Golub, G.H.: Exploiting the block structure of the web for computing PageRank. Technical Report, SCCM-03-02, Stanford University (2003)

8. Langville, A., Meyer, C.: Google's PageRank and Beyond: The Science of Search Engine Ranking. Princeton University Press, Princeton (2006)
9. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bring order to web, Stanford Digital Libraries, 1999, http://dbpubs.stanford.edu:8090/pub/1999-66
10. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. SIAM, Philadelphia (2003)

Superconvergence of Discontinuous Galerkin Method to Nonlinear Differential Equations

Helmi Temimi

Abstract In this paper, we investigate the superconvergence criteria of the discontinuous Galerkin (DG) method applied to one-dimensional nonlinear differential equations. We show numerically that the p-degree finite element (DG) solution is $O(\Delta x^{p+2})$ superconvergent at the roots of specific combined Jacobi polynomials. Moreover, we used these results to construct efficient and asymptotically exact a posteriori error estimates.

Keywords Discontinuous Galerkin method • Nonlinear boundary value problem • Superconvergence

Introduction

The (DG) method was left asleep for many decades till the last twenty years when it attracted the attention of several researchers and gained much popularity due to its wide application and flexibility. Celiker and Cockburn [4] showed that the p-degree (DG) solution and its derivative are, respectively, $O(\Delta x^{p+2})$ and $O(\Delta x^{p+1})$ superconvergent at the p-degree right Radau and at p-degree left Radau polynomials. Adjerid and Temimi [3] provided an original (DG) formulation applied to higher-order initial value problem without introducing an auxiliary variable in order to transform it into a first-order system. They showed that the *p-degree* DG solution is $O(\Delta t^{p+2})$ superconvergent at the roots of the $(p+1-m)$*-degree* Jacobi polynomial $P^{m,0}_{p+1-m}(\tau)$ and that the DG solution and its first $m-1$ derivatives are $O(\Delta t^{2p+2-m})$ superconvergent at the end of each step where m is the order of the ordinary differential equation (ODE).

H. Temimi (✉)
Department of Mathematics & Natural Sciences, Gulf University for Science & Technology,
P.O. Box 7207, Hawally 32093, Kuwait
e-mail: Temimi.H@gust.edu.kw

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.24_4

The paper is organized as follows: in section "The Discontinuous Galerkin Method", we present the DG formulation for the nonlinear differential equations. In section "Error Estimation for Nonlinear Boundary Value Problem", we provide an error estimation of DG method for nonlinear boundary value problem along with several numerical results. In section "Conclusion", we conclude with a few remarks.

The Discontinuous Galerkin Method

Let us consider the following nonlinear boundary value problem:

$$u'' + f(u) = g(x),\ a < x < b, \tag{1a}$$

subjected to the Dirichlet boundary conditions

$$u(a) = u_l, \qquad u(b) = u_r. \tag{1b}$$

In order to implement the discontinuous Galerkin (DG) method, we first create a partition, $x_k = k\,\Delta x,\ k = 0, 1, 2, \cdots, N+1,\ \Delta x = \frac{b-a}{N+1}$ with $I_k = (x_k, x_{k+1})$ and define the piecewise polynomial spaces

$$S^{n,p} = \{U\ :\ U|_{I_k} \in \mathcal{P}_p\}, \tag{2}$$

where $\mathcal{P}_p$ denotes the space of Legendre polynomials of degree p which will be adopted as basis functions.

We define the weak discontinuous Galerkin (DG) formulation for (1) by multiplying (1a) by a test function and then integrating over I_k. After integrating by parts, we obtain

$$\begin{gathered} u'v|_{x_k}^{x_{k+1}} - uv'|_{x_k}^{t_{x+1}} + \int_{x_k}^{x_{k+1}} uv''dx- \\ \int_{x_k}^{x_{k+1}} f(u)vdx = \int_{x_k}^{x_{k+1}} gvdx \end{gathered} \tag{3}$$

Let us replace u by $U_k(x) = U|_{[x_k,x_{k+1}]} \in \mathcal{P}_p$ and v by $V \in \mathcal{P}_p$ in (3), we obtain for $k = 0, 1, 2, \cdots, N$ and $\forall\ V \in \mathcal{P}_p$

$$\begin{gathered} \hat{U}_k'(x_{k+1})V(x_{k+1}^-) - \hat{U}_k'(x_k)V(x_k^+)- \\ \hat{U}_k(x_{k+1})V'(x_{k+1}^-) + \hat{U}_k(x_k)V'(x_k^+)+ \\ \int_{x_k}^{x_{k+1}} U_k V''dt - \int_{x_k}^{x_{k+1}} f(U_k)Vdx = \int_{x_k}^{x_{k+1}} gvdx. \end{gathered} \tag{4}$$

where $\hat{U}_k(x_k)$, $\hat{U}_k(x_{k+1})$, $\hat{U}'_k(x_k)$, and $\hat{U}'_k(x_{k+1})$ are called numerical fluxes. These terms arise from a double integration by parts, and an appropriate choice of these fluxes will define a stable DG method. Therefore, let us choose for $k = 1, 2, , \cdots, N-1$

$$\hat{U}_k(x_{k+1}) = U_k(x_{k+1}^-), \qquad \hat{U}_k(x_k) = U_{k-1}(x_k^-)$$

and

$$\hat{U}'_k(x_{k+1}) = U'_{k+1}(x_{k+1}^+), \qquad \hat{U}'_k(x_k) = U'_k(x_k^+)$$

and

$$\hat{U}_0(a) = u_l, \qquad \hat{U}_N(b) = u_r$$

and

$$\hat{U}'_N(b) = U'_N(b^-) - \frac{p}{\Delta x}(U_N(b^-) - u_r)$$

Therefore, the discrete formulation consists of determining $U_k(x) = U|_{[x_k, x_{k+1}]} \in \mathcal{P}_p$, such that $\forall\ V \in \mathcal{P}_p$

$$\begin{aligned} &U'_1(x_1^+)V(x_1^-) - U'_0(a^+)V(a^+) - U_0(x_1^-)V'(x_1^-) + \\ &\int_a^{x_1} U_0 V'' dx - \int_a^{x_1} f(U_0)V dx = -u_l V'(a^-) \end{aligned} \tag{5a}$$

and for $k = 1, 2, , \cdots, N-1$

$$\begin{aligned} &U'_{k+1}(x_{k+1}^+)V(x_{k+1}^-) - U'_k(x_k^+)V(x_k^+) - \\ &U_k(x_{k+1}^-)V'(x_{k+1}^-) + U_{k-1}(x_k^-)V'(x_k^+) + \\ &\int_{x_k}^{x_{k+1}} U_k V'' dx - \int_{x_k}^{x_{k+1}} (U_k)V dx = 0 \end{aligned} \tag{5b}$$

and

$$\begin{aligned} &U'_N(b^-)V(b^-) - U'_N(x_N^+)V(x_N^+) + U_{N-1}(x_N^-)V'(x_N^+) + \\ &\int_{x_N}^{b} U_N V'' dx - \int_{x_N}^{b} f(U_N)V dx = u_r V'(b^-) \end{aligned} \tag{5c}$$

Error Estimation for Nonlinear Boundary Value Problem

Error Estimation

Let us recall $P_k^{\alpha,\beta}(\tau)$ are Jacobi polynomials defined by the Rodrigues formula

$$P_k^{\alpha,\beta}(\tau) = \frac{(-1)^k}{2^k k!}(1-\tau)^{-\alpha}(1+\tau)^{-\beta}\frac{d^k}{d\tau^k}[(1-\tau)^{\alpha+k}(1+\tau)^{\beta+k}],$$
$$\alpha, \beta > -1,\ k = 0, 1, \cdots . \tag{6}$$

We note that Jacobi polynomials satisfy the orthogonality condition

$$\int_{-1}^{1}(1-\tau)^{\alpha}(1+\tau)^{\beta}P_k^{\alpha,\beta}(\tau)P_l^{\alpha,\beta}(\tau)d\tau = c_k\delta_{kl}, \tag{7}$$

where $c_k > 0$ and δ_{kl} is the Kronecker symbol equal to 1 if $k = l$ and 0, otherwise. We further note that $P_k^{0,0} = P_k$, the k^{th}-degree Legendre polynomial.

Lemma 1. *Let u and U be respectively solutions of (1) and (5); therefore,*

$$u - U = Q_{p+1}(\xi) + O(\Delta x^{p+1}) \tag{8}$$

where the leading term of the discretization error is given by

$$Q_{p+1}(\xi) = \alpha_{p+1}(\xi - 1)\left[P_p^{1,0}(\xi) + \left(\frac{p+1}{p}\right)^2 P_{p-1}^{1,0}(\xi)\right]. \tag{9}$$

□

and

Lemma 2. *Let $u \in C^{2p+2}$ and $U_k \in \mathcal{P}_p,\ p \geq 2$, be the solutions of (1) and (5); then the DG solution is superconvergent at the interior points x_j^8*

$$e(x_j^*) = O(\Delta x^{p+2}),\ j = 1, \cdots, p,$$

where x_j^* are the shifted roots of $\left[P_p^{1,0}(\xi) + \left(\frac{p+1}{p}\right)^2 P_{p-1}^{1,0}(\xi)\right]$ *to each element.*

□

Moreover, we construct a *posteriori* error estimates for 1.

We replace $u = e_k + U_k$ in (1a), multiply by a test function V and, integrate over $[x_k, x_{k+1}]$ to obtain

$$\int_{x_k}^{x_{k+1}} \left[(e_k + U_k)'' + f(e_k + U_k)\right] V dx = \int_{x_k}^{x_{k+1}} g(x) V dx, \tag{10}$$

where $U_k \in \mathcal{P}_p$ is the DG solution of (1) and e_k the discontinuous Galerkin error on $[x_k, x_{k+1}]$ defined by

$$e_k \approx E_k = \alpha_{p+1,k} \tilde{P}_{p+1}(x) \tag{11}$$

where $\tilde{P}_{p+1}(x)$ is obtained by mapping

$$P_{p+1}(\xi) = (\xi - 1)\left[P_p^{1,0}(\xi) + (\frac{p+1}{p})^2 P_{p-1}^{1,0}(\xi)\right]$$

to $[x_k, x_{k+1}]$.

Testing against $V = \tilde{P}_{p+1}$ and solving (10) for $\alpha_{p+1,k}$.

we measure the accuracy of a *posteriori* error estimates using the effectivity index defined by

$$\theta = \frac{\sqrt{\sum\limits_{k=0}^{N} ||E_k||_k^2}}{\sqrt{\sum\limits_{k=0}^{N} ||u - U_k||_k^2}}.$$

A *posteriori* error estimates are considered asymptotically exact, if the effectivity index converges to 1 under Δx or p refinements.

Computational Results

Let us consider the following nonlinear boundary value problem:

$$u'' + \ln(u) = \exp(x) + x, \ 0 < x < 1, \tag{12a}$$

subjected to

$$u(0) = 1, \qquad u(1) = \exp(1). \tag{12b}$$

The exact solution for 12 is $u(x) = \exp(x)$. We solve (12) using a uniform mesh with $\Delta x = 0.1$ and $p = 2, 3, 4, 5$, and we plot the error $u - U$ versus x in Figs. 1 and 2.

In Fig. 3, we plot $||e||$ versus N in a log-log graph in order to obtain the convergence rates for e. We conclude that $||e|| = \mathcal{O}(\Delta x^{p+1})$. Figure 4

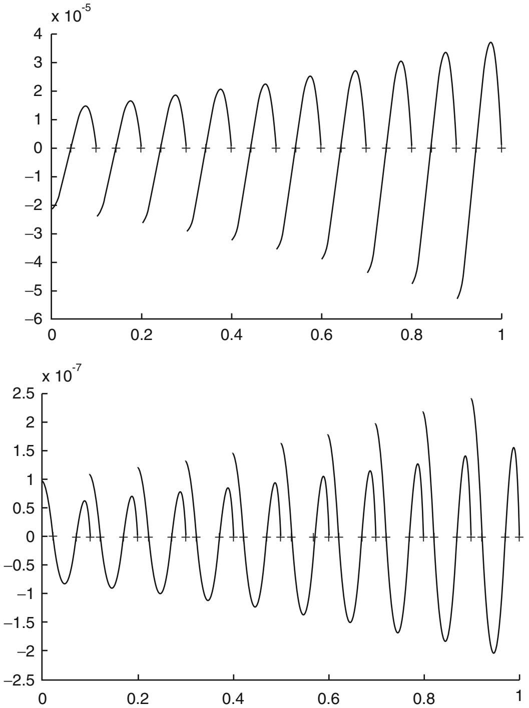


Fig. 1 The error curves of $u - U$ for problem (12) on uniform meshes ($N = 10$) for $p = 2$ (*top*) and $p = 3$ with roots of mapped polynomials $\left[P_p^{1,0}(\xi) + \left(\frac{p+1}{p}\right)^2 P_{p-1}^{1,0}(\xi)\right]$ are marked by *plus symbol*

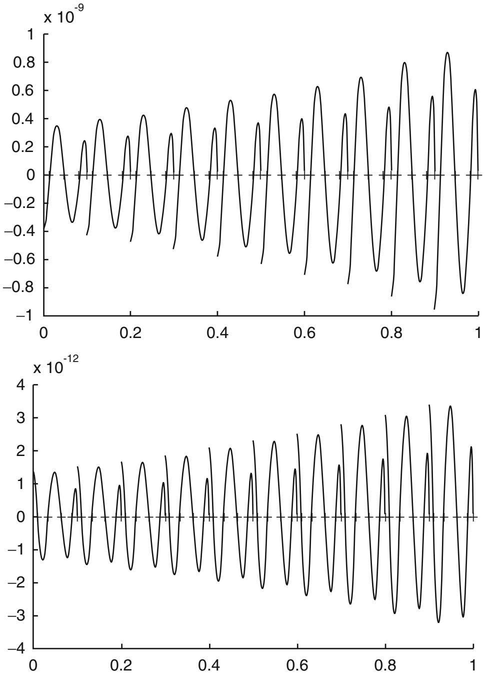


Fig. 2 The error curves of $u - U$ for problem (12) on uniform meshes ($N = 10$) for $p = 4$ (*top*) and $p = 5$ with roots of mapped polynomials $\left[P_p^{1,0}(\xi) + \left(\frac{p+1}{p}\right)^2 P_{p-1}^{1,0}(\xi)\right]$ are marked by *plus symbol*

Fig. 3 The $\mathcal{L}^2$-norm of the error $||e||$ using $p = 2, 3, 4$ versus N

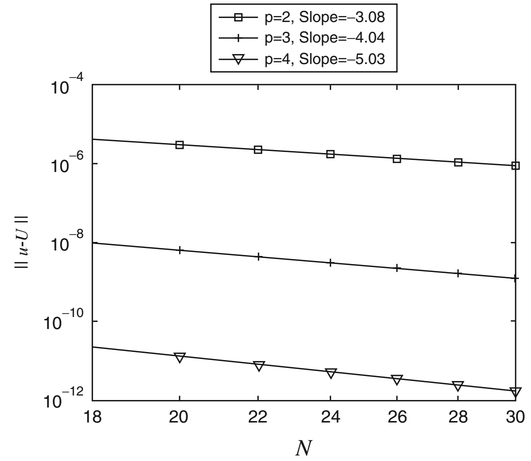


Fig. 4 The $\mathcal{L}^{\infty}$-norm of the error e at roots of mapped polynomials $\left[P_p^{1,0}(\xi) + \left(\frac{p+1}{p}\right)^2 P_{p-1}^{1,0}(\xi)\right]$ using $p = 2, 3, 4$ versus N

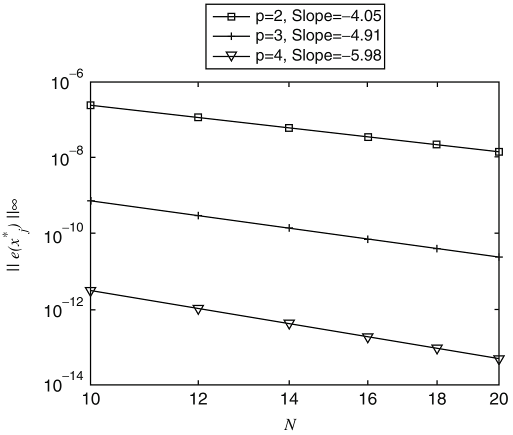


Table 1 The effectivity indices of the DG method for problem (12) versus N using $p = 2, 3, 4$

θ			
N	$p = 2$	$p = 3$	$p = 4$
20	9.8823e−001	1.0013	9.8758
25	9.9060e−001	1.0010	9.9056
30	9.9218e−001	1.0009	9.9376
35	9.9330e−001	1.0007	9.9649
40	9.9414e−001	1.0006	1.0018

shows that the error is superconvergent at roots x_j^* of the mapped polynomials $\left[P_p^{1,0}(\xi) + \left(\frac{p+1}{p}\right)^2 P_{p-1}^{1,0}(\xi)\right]$. Thus, $||e(x_j^*)||_{\infty} = \mathcal{O}(\Delta x^{p+2})$. Table 1 exhibits the convergence of the effectivity index to 1 under p and Δx refinement which reveals that a *posteriori* error estimates are asymptotically exact

Conclusion

In this paper, we present an error estimation of the discontinuous Galerkin method developed by Cheng and Shu [5] applied to nonlinear boundary value problems. We show that the leading term of the DG error is proportional to $P_{p+1}(\xi) = (\xi - 1)\left[P_p^{1,0}(\xi) + \left(\frac{p+1}{p}\right)^2 P_{p-1}^{1,0}(\xi)\right]$; therefore, the p-degree finite element (DG) solution is $O(\Delta x^{p+2})$ superconvergent at the shifted roots of $P_{p+1}(\xi)$. We use these superconvergence results to compute efficient and asymptotically exact a *posteriori* error estimates.

Acknowledgements We acknowledge the Gulf University for Science and Technology (GUST) for partially supporting this research in terms of conference grants.

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions. Dover, New York (1965)
2. Adjerid, S., Devine, K.D., Flaherty, J.E., Krivodonova, L.: A posteriori error estimation for discontinuous Galerkin solutions of hyperbolic problems. Comput. Methods Appl. Mech. Eng. **191**, 1097–1112 (2002)
3. Adjerid, S., Temimi, H.: A discontinuous Galerkin Method for higher-order ordinary differential equations. Comput. Methods Appl. Mech. Eng. **197**, 202–218 (2007)
4. Celiker, F., Cockburn, B.: Superconvergence of the numerical traces for discontinuous Galerkin and hybridized methods for convection-diffusion problems in one space dimension. Math. Comput. **76**(257), 67–96 (2007)
5. Cheng, Y., Shu, C.-W.: A discontinuous Galerkin finite element method for time dependent partial differential equations with higher order derivatives. Math. Comput. **77**, 699–730 (2008)
6. Cockburn, B., Shu, C.-W.: TVB Runge-Kutta local projection discontinuous Galerkin methods for scalar conservation laws II: general framework. Math. Comput. **52**, 411–435 (1989)
7. Cockburn, B., Shu, C.-W.: The local discontinuous Galerkin finite element method for convection-diffusion systems. SIAM J. Numer. Anal. **35**, 2240–2463 (1998)
8. Le Saint, P., Raviart, P.: On a finite element method for solving the neutron transport equations. In: de Boor, C. (ed.) Mathematical Aspects of Finite Elements in Partial Differential Equations, pp. 89–145. Academic, New York (1974)
9. Szego, G.: Orthogonal Polynomials. American Mathematical Society, Rhode Island (1975)

A Least Squares Approach for Exponential Rate of Convergence of Eigenfunctions of Second-Order Elliptic Eigenvalue Problems

Lokendra K. Balyan

Abstract In this paper, we show the convergence estimates for eigenvalues and eigenvectors of second-order elliptic eigenvalue problems using spectral element method. A least squares approach is used to prove that the eigenvalues and eigenvectors converge exponentially in P, degree of polynomials, when the boundary of the domains is to be assumed sufficiently smooth and the coefficients of the differential operator are analytic.

Keywords Approximation of eigenvalues and eigenvectors • Compact operator • Nonconforming • Spectral element method • Elliptic operators • Smooth boundary

Introduction

The study of eigenvalues/eigenfunctions is very important in science and engineering, particularly in physics, civil engineering, and aeronautical engineering. In this paper the author has proved the rate of convergence of eigenvalues and eigenvectors of elliptic operator where the coefficients of the operator and the boundary of the domain are sufficiently smooth. In [7, 11] have proved the similar convergence result for elliptic system on non-smooth domain. The author et al. have discussed the least squares nonconforming spectral element method for elliptic eigenvalue problems in [4]. We approximate eigenvalues and corresponding subspaces of generalized eigenfunctions are in terms of compact operators [8].

We shall consider a compact operator $\mathcal{T} : V \to V, V$ is a Banach space, and a family of compact operators $\mathcal{T}^P : V \to V$, such that $\mathcal{T}^P \to \mathcal{T}$ in H^1 norm, as $P \to \infty$. We obtained the estimates which show how the eigenvalues and eigenvectors of $\mathcal{T}$ are approximated by those of $\mathcal{T}^P$ [9]. In [5] Bramble and Osborn developed spectral approximation results for a class of compact operators on a Hilbert space and applied them to obtain convergence estimates for the eigenvalues

L.K. Balyan (✉)
IIIDM-Jabalpur, Jabalpur, MP 482005, India
e-mail: lokendra.balyan@gmail.com

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.32_5

and generalized eigenvectors of non-self-adjoint elliptic boundary value problems. Later in [9], Osborn presented the spectral approximation results for compact operators on a Banach space. He formulated the results in terms of the norm on the underlying Banach spaces.

We use least squares approach to discretize the partial differential equations. In literature, there exist several methods for differential equations, but each method has its own conditions under which their performance is measurable. However, the least squares method works on uniform policy, and it has a unified formulation for the numerical solution of all kinds of differential equations. Least squares approach for spectral element methods needs less degrees of freedom to obtain the prescribed level of accuracy; however, the amount of work that needs to be done for each degree of freedom is higher. Variation formulation approach to find out the solution of eigenvalues problems must impose the boundary condition on test and trail functions. To avoid this, least squares methods for elliptic eigenvalue problems were first proposed by Bramble and Schatz on smooth domains [6]. They constructed a finite dimensional subspace of $H^2(\Omega)$ for which the method employs C^1 elements and the functions need not satisfy the boundary conditions. However, the requirement of first-order continuity on the finite element space causes the numerical solution of the problem to be computationally demanding, and it is difficult to parallelize the method.

We minimize the sum of the residuals in the partial differential equations and a fractional Sobolev norm of the residuals in the boundary conditions and enforce continuity by adding a term which measures the jump in the function and its derivatives at inter-element boundaries in fractional Sobolev norms to the functional being minimized.

The outline of the paper is as follows. In section "Discretization and Prior Estimations", we define the elliptic eigenvalue problem and discuss how to discretize the domain. We close this section by stating the stability theorem for elliptic eigenvalue problems on smooth domain. In section "Main Result", the convergence estimates for eigenvalues and eigenvectors are presented.

Discretization and Prior Estimations

In this section we state our problem over the domain whose boundaries are smooth. We discretize the domain and define the nonconforming spectral element functions over the domain.

Consider the second-order elliptic eigenvalue problem on domain $\Omega \subseteq \mathcal{R}^2$. We seek the eigenvalue λ and eigenfunction $u(x)$ satisfying

$$\begin{aligned} \mathcal{L}u &= \lambda u(x) \text{ in } \Omega, \\ Bu &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{1}$$

where the operator $\mathcal{L}u$ is defined

$$-\sum_{i,j=1}^{2}\frac{\partial}{\partial x_i}\left(a_{i,j}(x)\frac{\partial u}{\partial x_j}\right)+\sum_{i=1}^{2}b_i(x)\frac{\partial u}{\partial x_i}+c\,(x)u,$$

and the coefficients of elliptic operators $a_{i,j}(x) = a_{j,i}(x)$, $b_i(x)$ and $c(x)$ are analytic functions of x. Further, the boundary operator $Bu = u$ denotes the Dirichlet boundary condition and $Bu = \left(\frac{\partial u}{\partial N}\right)_A$ denotes the Neumann boundary condition, where $\left(\frac{\partial u}{\partial N}\right)_A$ is the conormal derivative of u. Let A denote the 2×2 matrix whose entries are given by $A_{i,j}(x) = a_{i,j}(x)$, for $i, j = 1, 2$ [3]. Then $\left(\frac{\partial u}{\partial N}\right)_A$ is defined as

$$\left(\frac{\partial u}{\partial N}\right)_A = \sum_{i,j=1}^{2} a_{i,j}n_j\frac{\partial u}{\partial x_i}, \tag{2}$$

where $n(x) = (n_1, n_2)$ is the exterior unit normal to $\partial\Omega$ at x. Here, $\partial\Omega$ denotes the boundary of the domain.

We divide Ω, Fig. 1, into a fixed number of quadrilateral elements $\Omega_1, \Omega_2, \ldots, \Omega_r$. Let us define the set of nonconforming spectral element functions $\{u_1, u_2, \ldots, u_r\}$ as follows. The spectral element function $\hat{u}_i$ is defined on S as

$$\hat{u}_i(\xi, \eta) = \sum_{k=0}^{P}\sum_{l=0}^{P} a_{k,l}\xi^k\eta^l. \tag{3}$$

Define a smooth mapping M_i from the master squares $S = (-1, 1)^2$ to Ω_i by

$$x_j = (X_i)_j(\xi, \eta) \qquad \text{for } j = 1, 2.$$

Then u is given by

$$u_i(x_1, x_2) = \hat{u}_i(M_i^{-1}(x_1, x_2)).$$

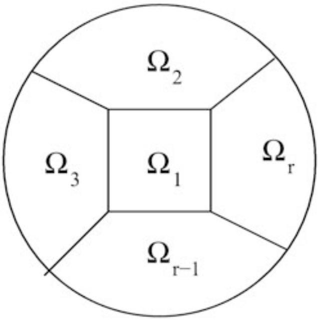


Fig. 1 The domain Ω is divided into elements $\Omega_1, \Omega_2, \ldots \Omega_r$

We now describe the terms which shall be needed in sequel. We define

$$(u_m)_{x_1} = (\hat{u}_m)_\xi \xi_{x_1} + (\hat{u}_m)_\eta \eta_{x_1},$$
$$(u_m)_{x_2} = (\hat{u}_m)_\xi \xi_{x_2} + (\hat{u}_m)_\eta \eta_{x_2}.$$

Let γ_s be a side common to the elements Ω_m and Ω_n. We may assume that γ_s is the image of $\eta = -1$ under the mapping M_m which maps S to Ω_m and also the image of $\eta = 1$ under the mapping M_n which maps S to Ω_n. We now define the jumps at the inter-element Boundaries:

$$\|[u]\|^2_{0,\gamma_s} = \|\hat{u}_m(\xi,-1) - \hat{u}_n(\xi,1)\|^2_{0,I}, \tag{4}$$
$$\|[(u_{x_1})]\|^2_{1/2,\gamma_s} = \|(u_m)_{x_1}(\xi,-1) - (u_n)_{x_1}(\xi,1)\|^2_{1/2,I},$$
$$\|[(u_{x_2})]\|^2_{1/2,\gamma_s} = \|(u_m)_{x_2}(\xi,-1) - (u_n)_{x_2}(\xi,1)\|^2_{1/2,I}.$$

Here I = (-1,1).

Now we define

$$\int_{\Omega_i} \mid \mathcal{L}u_i \mid^2 dx_1 dx_2 = \int_S \mid \mathcal{L}u_i \mid^2 J_i \, d\xi d\eta. \tag{5}$$

Here $J_i(\xi,\eta)$ is the Jacobian of the mapping M_i from S to Ω_i. Let $\hat{\mathcal{L}}_i \hat{u}_i = \sqrt{J_i}\,\mathcal{L}u_i$ and $\hat{F}_i(\xi,\eta) = f(M_i(\xi,\eta))\sqrt{J_i(\xi,\eta)}$ [4].

Thus,

$$\int_{\Omega_i} \mid \mathcal{L}u_i \mid^2 dx_1 dx_2 = \int_S \mid \hat{\mathcal{L}}_i \hat{u}_i \mid^2 \, d\xi d\eta. \tag{6}$$

Define the quadratic form

$$\mathcal{V}^P(\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_r) = \textstyle\sum_{i=1}^r \left\| \hat{\mathcal{L}}_i \hat{u}_i \right\|^2_{0,S} \tag{7}$$
$$+ \textstyle\sum_{\gamma_j \in \Omega} \left(\|[u]\|^2_{0,\gamma_j} + \|[u_{x_1}]\|^2_{\frac{1}{2},\gamma_j} + \|[u_{x_2}]\|^2_{\frac{1}{2},\gamma_j} \right)$$
$$+ \textstyle\sum_{\gamma_j \in \partial\Omega} \|Bu\|^2_{\lambda,\gamma_j} \,.$$

Here $\lambda = 3/2$ if $Bu = u$ and $\lambda = 1/2$ if $Bu = \left(\frac{\partial u}{\partial N}\right)_A$. Let γ_s be a side of the element Ω_m corresponding to $\eta = 1$ such that $\gamma_s \subseteq \partial\Omega$, and let $Bu = u$ denote the Dirichlet boundary conditions [4]. Then

$$\|Bu\|^2_{\frac{3}{2},\gamma_s} = \|u_m(\xi,1)\|^2_{0,I} + \|\frac{\partial u_m}{\partial T}(\xi,1)\|^2_{\frac{1}{2},I}.$$

Here $\frac{\partial u_m}{\partial T}$ denotes the tangential derivative along the curve γ_s. Similarly if $Bu = \left(\frac{\partial u}{\partial N}\right)_A$, the Neumann boundary condition, we can define $\|Bu\|^2_{\frac{1}{2},\gamma_s}$.

Then the following stability theorem holds which is a special case of Theorem 2.1 for smooth domains in [7].

Theorem 1. *There is a constant* $C > 0$ *such that*

$$\sum_{i=1}^{r} \| \hat{u}_i \|_{2,S}^2 \leq C\,(lnP)^2\, \mathcal{V}^P(\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_r). \tag{8}$$

Main Result

Let $\mathcal{T} : V \to V$ be a compact operator, V is a Banach space, and a family of compact operators $\mathcal{T}^P : V \to V$. In this paper, we present the convergence estimates for eigenvalues and eigenvectors in terms of compact operator by using the results of [9].

Let $f \in H^s(\Omega)$ with $s > 0$. Then by the shift theorem for elliptic boundary value problems $u \in H^{s+2}(\Omega)$ [1]. Henceforth, V will denote the space $H^1(\Omega)$. Let $\mathcal{T}$ denote an operator defined by

$$\mathcal{T}(f) = u,$$

where $f \in H^1(\Omega)$. Then $\mathcal{T}$ is a compact operator from $V \to V$. It follows that (λ, u) is an eigenpair of (1) if and only if $(\mu = \lambda^{-1}, u)$ is an eigenpair of the operator T [3, 9].

Theorem 1. *Let the coefficients of the elliptic operator and the boundary of the domain* Ω *be analytic. Let* λ *and* $\lambda_j(P)$ *be an eigenvalue of* $\mathcal{T}$ *and* $\mathcal{T}^P$*, respectively, such that* $\lambda_j(P)$ *converges to* λ*. Let* $u_j(P)$ *be an unit eigenvector of* $\mathcal{T}^P$ *corresponding to* $\lambda_j(P)$*. Then there exists an eigenvector* u *of* $\mathcal{T}$ *such that*

$$| \lambda - \lambda_j(P) | \leq \|(\mathcal{T} - \mathcal{T}^P) f\| \leq c\, e^{-bP}, \quad \text{and}$$
$$\|u - u_j(P)\| \leq \|(\mathcal{T} - \mathcal{T}^P) f\| \leq c\, e^{-bP}.$$

Here c *and* b *denote constants.*

Proof. Let $u \in H^s(\Omega)$. Define $\hat{U}_i(\xi, \eta) = u(M_i(\xi, \eta))$ for $(\xi, \eta) \in S$ and $\hat{v}_i(\xi, \eta) = \Pi^P \hat{U}_i$ be the orthogonal projection of $\hat{U}_i$ into the space of polynomials of degree P in each variable separately with respect to the inner product in $H^2(S)$. Then the following estimates hold from Theorem 4.46 of [10]:

$$\|\hat{U}_i - \hat{v}_i\|_{2,S}^2 \leq C_s P^{-2s+8} \|\hat{U}_i\|_{s,S}^2, \tag{9}$$

where $s \geq 5$ and $P \geq s - 3$, where $C_s = c\, e^{2s}$.

Let $\mathcal{L}_i$ denote the differential operator with analytic coefficients such that

$$\int_{\Omega_i}\int (\mathcal{L}v(x,y))^2 dxdy = \int_S\int (\mathcal{L}_i\hat{v}(\xi,\eta))^2 d\xi d\eta,$$

where

$$\mathcal{L}_i\hat{v}(\xi,\eta) = \mathcal{A}_i\hat{v}_{\xi\xi} + 2\mathcal{B}_i\hat{v}_{\xi\eta} + \mathcal{C}_i\hat{v}_{\eta\eta} + \mathcal{D}_i\hat{v}_{\xi} + \mathcal{E}_i\hat{v}_{\eta} + \mathcal{F}_i\hat{v}.$$

Since u and M_i are analytic, we can show that there exist constants C and d such that

$$\mid D_{\xi}^{\alpha_1} D_{\eta}^{\alpha_2}\hat{U}_i \mid\leq C\, d^s s!$$

for all $(\xi,\eta)\in S$ and $|\alpha|\leq s$ [2, 11].

Now, we estimate the residuals in the partial differential equation

$$\left\|\mathcal{L}_i\hat{v}_i - \hat{F}_i\right\|_{0,S}^2 \leq \left\|\mathcal{L}_i\hat{v}_i - \mathcal{L}_i\hat{U}_i\right\|_{0,S}^2 \leq C\left\|\hat{v}_i - \hat{U}_i\right\|_{2,S}^2. \tag{10}$$

Hence,

$$\left\|\mathcal{L}_i\hat{v}_i - \hat{F}_i\right\|_{0,S}^2 \leq C_s P^{-2s+8}(C\, d^s s!)^2. \tag{11}$$

Next, we show how to estimate the jumps defined in functional

$$\left(\|[v]\|_{0,\gamma_j}^2 + \|[v_{x_1}]\|_{\frac{1}{2},\gamma_j}^2 + \|[v_{x_2}]\|_{\frac{1}{2},\gamma_j}^2\right)$$

for any $\gamma_j\subset\Omega$. Here γ_j is a side common to Ω_m and Ω_n for some m and n. Then

$$\begin{aligned}\|[v]\|_{0,\gamma_j}^2 &= \int_{-1}^{1}(\hat{v}_m(\xi,-1)-\hat{v}_n(\xi,1))^2\,d\xi\\ &\leq 2\left(\int_{-1}^{1}\left(\hat{v}_m(\xi,-1)-\hat{U}_m(\xi,-1)\right)^2 d\xi\right.\\ &\quad\left.+\int_{-1}^{1}\left(\hat{v}_n(\xi,1)-\hat{U}_n(\xi,1)\right)^2 d\xi\right)\\ &\leq C_s P^{-2s+8}(C\, d^s s!)^2. \end{aligned}\tag{12}$$

Now,

$$\begin{aligned}\|[v_{x_1}]\|_{\frac{1}{2},\gamma_j}^2 &= \left\|\left((\hat{v}_m)_\xi\,\xi_{x_1} + (\hat{v}_m)_\eta\,\eta_{x_1}\right)(\xi,-1)\right.\\ &\quad\left.-\left((\hat{v}_n)_\xi\,\xi_{x_1} + (\hat{v}_n)_\eta\,\eta_{x_1}\right)(\xi,1)\right\|_{\frac{1}{2},I}^2\end{aligned}$$

$$
\begin{aligned}
\leq\ 4\Bigg(&\left\|\left(\left((\hat{v}_m)_\xi-(\hat{U}_m)_\xi\right)\xi_{x_1}\right)(\xi,-1)\right\|^2_{\frac{1}{2},I}\\
&+\left\|\left(\left((\hat{v}_m)_\eta-(\hat{U}_m)_\eta\right)\eta_{x_1}\right)(\xi,-1)\right\|^2_{\frac{1}{2},I}\\
&+\left\|\left(\left((\hat{v}_n)_\xi-(\hat{U}_n)_\xi\right)\xi_{x_1}\right)(\xi,1)\right\|^2_{\frac{1}{2},I}\\
&+\left\|\left(\left((\hat{v}_n)_\eta-(\hat{U}_n)_\eta\right)\eta_{x_1}\right)(\xi,1)\right\|^2_{\frac{1}{2},I}\Bigg).
\end{aligned}
$$

We have

$$\|ab\|_{1/2,I}\ \leq\ C\,\|a\|_{1,\infty,I}\ \|b\|_{1/2,I}\,.$$

Putting all these estimates together, we can conclude that

$$\|[v_{x_1}]\|^2_{\frac{1}{2},\gamma_j}\ \leq\ C_s P^{-2s+8}(C\ d^s s!)^2. \tag{13}$$

Similarly we can show that

$$\|[v_{x_2}]\|^2_{\frac{1}{2},\gamma_j}\ \leq\ C_s P^{-2s+8}(C\ d^s s!)^2, \tag{14}$$

and the remaining terms can be estimated in the same manner.

Hence, we can conclude that

$$\mathcal{R}^P(\hat{v}_1,\ldots,\hat{v}_r)\ \leq\ K\left(C_s P^{-2s+8}(C\ d^s s!)^2\right). \tag{15}$$

where the functional $\mathcal{R}^P(\hat{v}_1,\hat{v}_2,\ldots,\hat{v}_r)$ is defined as per (7).

$$
\begin{aligned}
\mathcal{R}^P(\hat{v}_1,\hat{v}_2,\ldots,\hat{v}_r) &= \textstyle\sum_{i=1}^r\|\hat{\mathcal{L}}_i\hat{v}_i-\hat{F}_i\|^2_{0,S}\\
&+\textstyle\sum_{\gamma_j\in\Omega}\left(\|[v]\|^2_{0,\gamma_j}+\|[v_{x_1}]\|^2_{\frac{1}{2},\gamma_j}+\|[v_{x_2}]\|^2_{\frac{1}{2},\gamma_j}\right)\\
&+\textstyle\sum_{\gamma_j\in\partial\Omega}\|Bv\|^2_{\lambda,\gamma_j}\,.
\end{aligned}
$$

We now use Stirling's formula

$$n!\sim\sqrt{2\pi n}\,e^{-n}n^n.$$

Let $s=\Upsilon P$, where Υ is a constant. Then

$$
\begin{aligned}
\mathcal{R}^P(\hat{v}_1,\ldots,\hat{v}_r)\ &\leq K(2\pi\Upsilon P)C\ e^{2s}P^8e^{-2s}\left(\frac{ds}{P}\right)^{2s}\\
&\leq C\,\Upsilon P^9\,(d\,\Upsilon)^{2\Upsilon P}\,.
\end{aligned} \tag{16}
$$

We choose Υ so that $d\,\Upsilon<1$.

Then using Theorem 3.1 of [11], there exist constants $c, b > 0$ such that the estimate

$$\mathcal{R}^P(\hat{v}_1, \ldots, \hat{v}_r) \leq C\, e^{-bP} \tag{17}$$

holds.

Let the set of spectral element functions $\{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_r\}$ be our approximate solution which minimizes the functional $\mathcal{R}^P(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_r)$.

Thus,

$$\mathcal{R}^P(\hat{w}_1, \ldots, \hat{w}_r) \leq C\, e^{-bP}. \tag{18}$$

Therefore, we can conclude that

$$\mathcal{R}^P(\hat{v}_1 - \hat{w}_1, \ldots, \hat{v}_r - \hat{w}_r) \leq C\, e^{-bP}, \tag{19}$$

Hence, by using the stability theorem, there exist constants k and c such that

$$\sum_{i=1}^{r} \|\hat{v}_i - \hat{w}_i\|_{2,S}^2 \leq k\, e^{-cP}$$

holds.

Now, we make corrections so that the corrected solution is conforming and belongs to $H^1(\Omega)$.

Thus, the error estimate

$$\|v_c - w_c\|_{H^1(\Omega)} \leq k\, e^{-cP} \tag{20}$$

holds for P large enough. Here, k and c denote generic constants.

Clearly

$$\|u - v_c\|_{H^1(\Omega)} \leq k\, e^{-cP}. \tag{21}$$

Hence,

$$\|u - w_c\|_{H^1(\Omega)} \leq k\, e^{-cP}. \tag{22}$$

Here k and c denote generic constants. Define the operator

$$\mathcal{T}^P(f) = w_c$$

Then $\mathcal{T}^P$ is an operator from $V \to V$, where $V = H^1(\Omega)$. Then $\mathcal{T}^P$ is a compact operator since its image is finite.

Therefore, Eq. (22) can be written as

$$\|(\mathcal{T} - \mathcal{T}^P) f\|_{H^1(\Omega)} \leq k\, e^{-cP}. \tag{23}$$

By using the Theorems (1, 2) of [9], we obtain the results.

Conclusion

We have shown the convergence estimates for the eigenvalues and eigenvectors using spectral element methods when the boundary of the domain is smooth and the coefficients of differential operator are analytic. We estimate the errors for eigenvalues and eigenvectors using least squares methods. We used fully nonconforming approach and considered the jumps between two elements. It has been shown that the eigenvalues and eigenvectors of second-order elliptic operator exponentially converge.

References

1. Adams, R.A.: Sobolev Spaces. Academic, New York (1975)
2. Babuska, I., Guo, B.Q.: The h-p version of the finite element method on domains with curved boundaries. SIAM J. Numer. Anal. **25**(4), 837–861 (1988)
3. Babuska, I., Osborn, J.: Eigenvalue problems. In: Ciarlet, P.G., Lions, J.L. (eds.) Finite element methods (part II). Handbook of numerical analysis, vol. 2. (1991)
4. Balyan, L.K., Dutt, P., Rathore, R.K.S.: Least squares h-p spectral element method for elliptic eigenvalue problems. Appl. Math. Comput. **218**(19), 9596–9613 (2012)
5. Bramble, J.H., Osborn, J.E.: Rate of convergence estimates for non-selfadjoint eigenvalue approximation. Math. Comput. **27**, 525–549 (1973)
6. Bramble, J.H., Schatz, A.H.: Rayleigh-Ritz-Galerkin methods for Dirichlet's problem using subspaces without boundary conditions. Commun. Pure Appl. Math. **23**, 653–675 (1970)
7. Dutt, P., Kishore, N., Upadhyay, C.S.: Nonconforming h-p spectral element methods for elliptic problems. Proc. Indian Acad. Sci. (Math. Sci.), Springer **117**(1), 109–145 (2007)
8. Kato, T.: Perturbation theory for linear operators, 2nd edn. Springer, Berlin/New York (1980)
9. Osborn, J.E.: Spectral approximation for compact operators. Math. Comput. **29**, 712–725 (1975)
10. Schwab, Ch.: p and h-p Finite Element Methods. Clarendon Press, Oxford (1998)
11. Tomar, S.K.: h-p spectral element methods for elliptic problems on non-smooth domains using parallel computers, Ph.D. Thesis (IIT Kanpur, INDIA), 2001; Reprint available as Tec. Rep. no. 1631, Department of Applied Mathematics, University of Twente, The Netherlands. http://www.math.utwente.nl/publications

Multivariable Polynomials for the Construction of Binary Sensing Matrices

R. Ramu Naidu, Phanindra Jampana, and Sastry S. Challa

Abstract In compressed sensing, the matrices that satisfy restricted isometry property (RIP) play an important role. But to date, very few results for designing such matrices are available. Of interest in several applications is a matrix whose elements are 0's and 1's (in short, $0, 1$-matrix), excluding column normalization factors. Recently, DeVore (J Complexity 23:918–925, 2007) has constructed deterministic $0, 1$-matrices that obey sparse recovery properties such as RIP. The present work extends the ideas embedded in DeVore (J Complexity 23:918–925, 2007) and shows that the $0, 1$-matrices of different sizes can be constructed using multivariable homogeneous polynomials.

Keywords Compressed sensing • Restricted isometry property • Deterministic construction • Multivariable homogeneous polynomials

Introduction

Compressed sensing (CS) aims at recovering high-dimensional sparse vectors from considerably fewer linear measurements. The problem of sparse recovery through l_0 norm minimization (i.e., minimization of number of nonzero components in the solution to be obtained) is not tractable. Chen and Donoho [7], Donoho et al. [11], Candes [3], Candes et al. [4], Candes and Tao [6] and Cohen et al. [8] have made several pioneering contributions and have reposed the problem as a simple linear programming problem (LPP). They have then established the conditions that ensure the stated equivalence between the original l_0 problem and its reposed version.

R.R. Naidu (✉) • S.S. Challa
Department of Mathematics, Indian Institute of Technology Hyderabad,
Yeddumailaram, A.P. 502 205, India
e-mail: ma11p003@iith.ac.in; csastry@iith.ac.in

P. Jampana
Department of Chemical Engineering, Indian Institute of Technology Hyderabad,
Yeddumailaram, A.P. 502 205, India
e-mail: pjampana@iith.ac.in

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.36_6

It is known that RIP is one sufficient condition to ensure the equivalence. Random matrices such as Gaussian or Bernoulli as their entries satisfy RIP with high probability.

DeVore [9] has constructed deterministic RIP matrices with 0's and 1's as elements (excluding column normalization factors) using general univariate polynomials on finite fields. In the present work, we extend the ideas developed in [9] and construct $0, 1$-matrices that satisfy RIP with different row and column sizes using homogeneous multivariable polynomials.

The paper is organized into several sections. In sections "Sparse Recovery from Linear Measurements" and "On the Equivalence Between P_0 and P_1 Problems", we present basic CS theory and conditions that ensure the equivalence between l_0-norm problem and its LPP version problems. Motivated by the deterministic construction methodology proposed in [9], we present in section "Deterministic CS Matrices Through Multivariable Homogeneous Polynomials" our extension, which is based on homogeneous two-variable polynomials over finite fields. In section "On Constructing Circulant RIP Matrices", we extend the ideas to deal with circulant matrices. While in section "Generalization from Two Variables to **n** Variables", we state the results for n-variable homogeneous polynomials. In the last section, we present our concluding remarks.

Sparse Recovery from Linear Measurements

As stated already, CS refers to the problem of reconstruction of an unknown vector $u \in \mathbb{R}^m$ from the linear measurements $y = (\langle u, \psi_1 \rangle, \ldots, \langle u, \psi_n \rangle) \in \mathbb{R}^n$ with $\langle u, \psi_j \rangle$ being the inner product between u and ψ_j. The basic objective in CS is to design a recovery procedure based on the sparsity assumption on u when the number of measurements n is much small compared to m. Sparse representations seem to have merit for various applications in areas such as image/signal processing and numerical computation.

A vector $u \in \mathbb{R}^m$ is k-spare if it has at most k nonzero coordinates. The problem of obtaining the sparse vector from its linear measurements may be posed as

$$P_0 : \min_v \|v\|_0 \text{ subject to } \quad \Psi v = y. \tag{1}$$

Here, $\|v\|_0 = |\{i \mid v_i \neq 0\}|$.

Donoho and others [7, 11] have provided the conditions under which the solution to P_0 is the same as that of the following LPP:

$$P_1 : \min_v \|v\|_1 \text{ subject to } \quad \Psi v = y. \tag{2}$$

Here, $\|v\|_1$ denotes the l_1-norm of the vector $v \in \mathbb{R}^m$. Denote the solution to P_1 by $f_\Psi(y)$ and solution to P_0 by $u^0_\Psi(y) \in \mathbb{R}^m$.

On the Equivalence Between P_0 and P_1 Problems

Definition 1. The mutual coherence $\mu(\Psi)$ of a given matrix Ψ is the largest absolute normalized inner product between different columns of Ψ. Denoting the k-th column in Ψ by ψ_k, the **mutual coherence** is given by

$$\mu(\Psi) = \max_{1 \leq i,j \leq m,\, i \neq j} \frac{|\, \psi_i^T \psi_j \,|}{\|\psi_i\|_2 \|\psi_j\|_2}. \tag{3}$$

It is known [11] that for μ-coherent matrices Ψ, one has

$$u_\Psi^0(y) = f_\Psi(y) = u, \tag{4}$$

provided u is k-sparse with $k < \frac{1}{2}\ (1 + \frac{1}{\mu})$. Donoho [10] has given sufficient conditions on the matrix Ψ for (4) to hold.

Candes and Tao [5] have introduced the following isometry condition on matrices Ψ and have established its important role in CS.

Definition 2. An $n \times m$ matrix Ψ is said to satisfy the **Restricted Isometry Property (RIP)** of order k with constant δ_k if for all k-sparse vectors $x \in \mathbb{R}^m$, we have

$$(1 - \delta_k)\, \|x\|_{l_2}^2 \leq \|\Psi x\|_{l_2}^2 \leq (1 + \delta_k)\, \|x\|_{l_2}^2\,. \tag{5}$$

Candes and Tao [5] have shown that whenever Ψ satisfies RIP of order $3k$ with $\delta_{3k} < 1$, the CS reconstruction error satisfies the following estimate:

$$\|u - f_\Psi(\Psi u)\|_{l_2^m} \leq C k^{\frac{-1}{2}} \sigma_k(u)_{l_1^m}, \tag{6}$$

where $\sigma_k(u)_{l_1^m}$ denotes the l_1 error of the best k-term approximation and the constant C depends only on δ_{3k}. This means that the bigger the value of k for which we can verify the RIP, then the better guarantee we have on the performance of Ψ.

One of the important problems in CS theory deals with constructing CS matrices that satisfy the RIP for the largest possible range of k. It is known that the widest range possible is $k \leq C \frac{n}{\log(\frac{m}{n})}$ [1, 9, 13, 14]. However, the only known matrices that satisfy the RIP for this range are based on random constructions.

Baraniuk et al. [1] have verified the RIP for random matrices with some probability using the following concentration of measure inequality:

$$\Pr(\left| \|\Psi(\omega)u\|_{l_2}^2 - \|u\|_{l_2}^2 \right| \geq \epsilon\, \|u\|_{l_2}^2) \leq 2e^{-nc_0(\epsilon)},\ 0 < \epsilon < 1. \tag{7}$$

where the probability is taken over all $n \times m$ matrices $\Psi(\omega)$ and $c_0(\epsilon)$ is a constant dependent only on ϵ such that for all $\epsilon \in (0, 1)$, $c_0(\epsilon) > 0$.

There are, however, no deterministic constructions for k being equal to $\frac{n}{\log(\frac{m}{n})}$. To the best of our knowledge, designing the good deterministic constructions of RIP matrices is still an open problem. DeVore [9], Nelson and Temlyakov [16] have constructed deterministic RIP matrices. DeVore has constructed 0, 1-RIP matrix using univariable polynomials. In the present work, we attempt to extend DeVore's work and construct 0, 1-matrices based on multivariable homogeneous polynomials. The advantage of using such polynomials is that matrices of different sizes can be constructed.

Deterministic CS Matrices Through Multivariable Homogeneous Polynomials

In this section, we present our deterministic construction procedure that is based on multivariable homogeneous polynomials. To begin with, we consider homogeneous polynomials in two variables, and later on we extend our methodology using homogeneous polynomials in n variables. As in [9], we shall consider only the case where $\mathbb{F}$ has a prime order and hence is a field of integers modulo p ($\mathbb{Z}_p$). The results we prove can be established for other finite fields also.

Given any integer $2 < r \leq p$, define $\mathcal{P}_r$ to be the set of all homogeneous polynomials in two variables of degree r over $\mathbb{Z}_p$. Let $Q(x, y) \in \mathcal{P}_r$ be represented as $Q(x, y) = \sum_{i+j=r} a_{ij} x^i y^j$, where the coefficients $a_{ij} \in \mathbb{Z}_p$. Clearly the cardinality of $\mathcal{P}_r$ is $m = p^{r+1} - 1$ (-1 for removing the zero polynomial). Any $Q(x, y) \in \mathbb{P}_r$, is a mapping from $\mathbb{Z}_p \times \mathbb{Z}_p$ to $\mathbb{Z}_p$. The graph of $Q(x, y)$ is $\mathcal{G}(Q) = \{(x, y, Q(x, y)) \mid (x, y) \in \mathbb{Z}_p \times \mathbb{Z}_p\} \subseteq \mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p$, with $|\mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p| = p^3 = n$.

We order the elements of $\mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p$ lexicographically as $(0, 0, 0), \ldots (0, p-1, p-1), (1, 0, 0), \ldots (p-1, p-1p-1)$. For any $Q(x, y) \in \mathcal{P}_r$, denote V_Q the vector indexed on $\mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p$ which takes value one at any ordered pair from the graph of $Q(x, y)$ and takes the value zero otherwise. There are exactly p^2 ones in V_Q. Define the matrix Φ with columns $V_Q, Q(x, y) \in \mathcal{P}_r$, with these columns ordered lexicographically with respect to the coefficients of the polynomials. Then the size of the matrix is $n \times m$, that is, $p^3 \times (p^{r+1} - 1)$. The following theorem [15] provides an upper bound on the number of zeros of a homogeneous polynomial $Q \in \mathcal{P}_r$ in $\mathbb{F}_q^2$.

Theorem 1. *Let $f \in \mathbb{F}_q[x_1, \ldots x_j]$ be homogeneous polynomial with $deg(f) = r \geq 1$. Then the equation $f(x_1, \ldots x_j) = 0$ has at most $r(q^{j-1} - 1) + 1$ solutions in $\mathbb{F}_q^j$, where $\mathbb{F}_q$ is a finite field of characteristic p and $q = p^i, i \in \mathbb{Z}^+$.*

The following proposition relates the RIP constant δ_k and μ.

Proposition 2. *Suppose that $\phi_1, \ldots, \phi_m$ are the unit norm columns of the matrix Φ and have coherence μ. Then Φ satisfies RIP of order k with constant $\delta_k = (k-1)\mu$.*

Using the above Theorem 1 and Proposition 2 [2, 12], we prove that the matrix Φ so constructed satisfies RIP, as detailed below:

Theorem 3. *The matrix* $\Phi_0 = \frac{1}{p}\Phi$ *satisfies the RIP with* $\delta_k = \frac{k-1(r(p-1)+1)}{p^2}$ *for any* $k < \frac{p^2}{r(p-1)+1} + 1$.

Proof. Let V_Q, V_R be two different columns from Φ. For any $Q, R \in \mathcal{P}_r$ with $Q \neq R$, there are at most $r(p-1)+1$ values of $\mathbb{Z}_p \times \mathbb{Z}_p$ such that $Q(x, y) = R(x, y)$.

Therefore, the inner product between any two different columns of Φ is at most equal to $r(p-1)+1$. It follows that coherence of the matrix Φ is $\mu(\Phi)$, which is at most equal to $\frac{r(p-1)+1}{p^2}$. From the above proposition 4.2, Φ satisfies the RIP with $\delta_k = (k-1)\frac{r(p-1)+1}{p^2}$. For $\delta_k < 1$, $k < \frac{p^2}{r(p-1)+1} + 1$. Hence, Φ satisfies RIP of order k. □

Remark 1. Since $n = p^3$, $m = p^{r+1} - 1$, and $k - 1 < \frac{p^2}{r(p-1)+1}$, we get $p = n^{\frac{1}{3}}, r = \frac{\log(m+1)}{\log p} - 1$. Consequently,

$$k(n,m) \asymp \frac{n^{\frac{2}{3}} \log(n)^{\frac{1}{3}}}{n^{\frac{1}{3}} \log\left(\frac{m+1}{n^{\frac{1}{3}}}\right) + \log\left(\frac{n^{\frac{2}{3}}}{m+1}\right)}. \tag{8}$$

From this we can easily get that $n^{\frac{1}{3}} \ll k(n,m)$.

Remark 2. The size of the matrix obtained in [9] is $p^2 \times p^{(r+1)}$ and of our matrix is $p^3 \times p^{(r+1)} - 1$. Therefore, it may be concluded that our methodology allows us to construct a different class of 0, 1-matrices. So far in this construction, the field that is considered is $\mathbb{Z}_p$. If we consider any finite field $\mathbb{F}_q, q = p^i, i \in \mathbb{Z}^+$ in place of $\mathbb{Z}_p$, then sizes of Devore's and our matrices become $p^{2n} \times p^{n(r+1)}$, $p^{3n} \times p^{n(r+1)} - 1$, respectively.

On Constructing Circulant RIP Matrices

Motivated by the results of [9], we also extend our construction to deal with circulant matrices $\Phi = (\phi_{ij})$. The circulant matrix is completely determined by its first l columns. Subsequent blocks of l columns are obtained by cyclic shifts of the rows of the first block. For choosing the first l columns of Φ, we define an equivalence relation on $\mathcal{P}_r$. For $P, Q \in \mathcal{P}_r$, $P \backsim Q$ iff there exists $0 \neq \lambda \in \mathbb{Z}_p$ such that $P(x, y) = \lambda Q(x, y), \forall (x, y) \in \mathbb{Z}_p \times \mathbb{Z}_p$. Given $P \in \mathcal{P}_r$, let $[P] = \left\{\lambda^{-1} P | 0 \neq \lambda \in \mathbb{Z}_p\right\}$ and then $|\,[P]\,| = p-1, |\mathcal{P}_r| = p^{r+1}-1$. Therefore, there are $\frac{p^{r+1}-1}{p-1}$ distinct equivalence classes in $\mathcal{P}_r$. Let Γ_r be a set which consists of one representation from each of the equivalence classes. Then the cardinality of Γ_r is $\frac{p^{r+1}-1}{p-1}$. The polynomials from Γ_r represent the first l columns. Hence, $l = |\Gamma_r|$. Since $n = p^3$ and these l columns can be shifted in cyclic way p^3 times, the matrix can be written as

$$\phi_{ij} = \phi_{((i-k) \mod p^3)(j \mod l)}$$

where $kl \leq j < (k+1)l$ for $0 \leq i < p^3$ and $0 \leq j < m := (\frac{p^{r+1}-1}{p-1})p^3 = p^{r+3} + p^{r+2} + \ldots + p^3$.

Now define the circulant matrix Φ_0 of size $n \times m$ whose first l columns are $V_Q, Q(x,y) \in \Gamma_r$, written in lexicographic order with respect to the coefficients of the polynomials. We can find ones in shifted columns in the following way:

Consider the i^{th}-block, $0 \leq i \leq n-1$. As $n = p^3$, we can write $i = a+bp+cp^2$, where $a, b, c \in \{0, \ldots, p-1\}$. Each column in this block will be the cyclic shift of a column V_Q in the first block. The entry in the (x, y, z) position of V_Q will now occupy the position (x', y', z') in the i^{th}-block, where $z' = z + i = z + a \ (mod \ p), y' = y + b \ or \ y + b + 1 \ (mod \ p), x' = x + c \ or \ x + c + 1 (mod \ p)$. Since the ones in V_Q occur precisely in the position $(x, y, Q(x, y))$, the new ones in the corresponding column of block i will occur at (x', y', z'), where $z' = Q(x, y) + a \ (mod \ p), y' = y + b \ or \ y + b + 1 \ (mod \ p), x' = x + c \ or \ x + c + 1 (mod \ p)$.

The following theorem [15] provides an upper bound on the number of zeros of a multivariable polynomial $Q(x, y)$ in $\mathbb{F}_q^2$.

Theorem 4. *Let $f \in \mathbb{F}_q[x_1, \ldots x_n]$ be a polynomial with $deg(f) = r \geq 1$. Then the equation $f(x_1, \ldots x_n) = 0$ has at most rq^{n-1} solutions in $\mathbb{F}_q^n$, where the $\mathbb{F}_q$ is the finite field of characteristic p, $q = p^n, n \in \mathbb{Z}^+$.*

The following lemma bounds the inner product of any two columns of Φ_0.

Lemma 5. *The inner product between any two different columns of the matrix Φ_0 is $|V \cdot W| \leq 2^4 rp$.*

Proof. Let V and W be any two columns of Φ_0. Then there exist $Q, R \in \Gamma_r$ such that V and W are the cyclic shift of vectors V_Q, V_R. As we have observed above, there are integers a_0, b_0, c_0 (depending only on V) such that any one in column V occurs at a position (x', y', z') iff $z' = Q(x, y) + a_0 \ (mod \ p), \ y' = y + b_0 + \epsilon_0 \ (mod \ p), x' = x + c_0 + \epsilon_1 (mod \ p)$ with $(x, y) \in \mathbb{Z}_p \times \mathbb{Z}_p$, $\epsilon_0, \epsilon_1 \in \{0, 1\}$. Similarly a one occurs in column W at position (x'', y'', z'') iff $z'' = R(\overline{x}, \overline{y}) + a_1 \ (mod \ p), \ y'' = \overline{y} + b_1 + \epsilon_0' \ (mod \ p), \ x'' = \overline{x} + c_1 + \epsilon_1' (mod \ p)$ with $(\overline{x}, \overline{y}) \in \mathbb{Z}_p \times \mathbb{Z}_p$ and $\epsilon_0', \epsilon_1' \in \{0, 1\}$. The inner product $V \cdot W$ counts the number of row positions for which there is a one at common places of these two columns. In other words, that is, $(x', y', z') = (x'', y'', z'')$, which is the same as

$$\begin{aligned} x + c_0 + \epsilon_1 &= \overline{x} + c_1 + \epsilon_1', \\ y + b_0 + \epsilon_0 &= \overline{y} + b_1 + \epsilon_0' \\ Q(x, y) + a_0 &= R(\overline{x}, \overline{y}) + a_1 (mod\ \ p). \end{aligned} \tag{9}$$

Case 1: If $Q \neq R$, we fix one of the 2^4 possibilities for $\epsilon_0, \epsilon_0', \epsilon_1, \epsilon_1'$. (9) implies that $\overline{x} = x + d,\ \overline{y} = y + e\ and\ R(x + d, y + e) = Q(x, y) + a, (mod\ \ p)$ where $a = a_0 - a_1, d = c_0 + \epsilon_1 - c_1 - \epsilon_1', e = b_0 + \epsilon_0 - b_1 - \epsilon_0'$. Since $Q \neq R$ and Q, R are homogeneous, we have $R(. + (d, e)) \neq Q(.) + a, \forall (x, y) \in \mathbb{Z}_p \times \mathbb{Z}_p$. Therefore, $R(. + (d, e)) - Q(.) - a$ is a nonzero, nonhomogeneous polynomial of degree r. In this case, the only possible $(x, y)'$s which satisfy (9) are zeros of $R(. + (d, e)) - Q(.) - a$. As $R(. + (d, e)) - Q(.) - a$ is nonhomogeneous two-variable polynomial, from theorem 4, it has at most rp roots. Since there are 2^4 possibilities for $\epsilon_0, \epsilon_0', \epsilon_1, \epsilon_1', |V \cdot W| \leq 2^4 rp$.

Case 2: If $Q = R$, then (9) implies that $\overline{x} = x + d, \overline{y} = y + e\ and\ Q(x + d, y + e) = Q(x, y) + a$. Suppose $Q(. + (d, e)) \equiv Q(.) + a$, which implies that $Q(. + (d, e)) - a \equiv Q(.)$. Since $Q(. + (d, e)) - a$ is a nonhomogeneous polynomial, Q is nonhomogeneous, which is a contradiction. Hence, $Q(. + (d, e)) - Q(.) - a$ is a nonzero, nonhomogeneous polynomial of degree at most r. Consequently, it has at most rp roots, which implies that $|V \cdot W| \leq 2^4 rp$. □

Therefore, the coherence of the circulant matrix is $\mu(\Phi_0) \leq \frac{2^4 rp}{p^2}$.

Theorem 6. *The circulant matrix* $\Phi_1 = \frac{1}{p}\Phi_0$ *has the RIP with* $\delta_k = 2^4(k - 1)\frac{r}{p}$ *whenever* $k - 1 < \frac{p}{2^4 r}$.

Proof. Proof is the same as that of Theorem 3. □

Generalization from Two Variables to n Variables

In the above construction, we have used homogeneous polynomials in two variables over $\mathbb{Z}_p$. This idea can, however, be extended to deal with homogeneous polynomials in n variables over $\mathbb{Z}_p$. Given any integer $2 < r < p$, define $\mathcal{P}_r$ to be a set of all homogeneous polynomials in n variables of degree r over $\mathbb{Z}_p$. Let $f_n(r)$ be a cardinality of a set of all integer solutions such that $i_1 + i_2 + \ldots + i_n = r$. Then using the following relation

$$\sum_{k=1}^{n} k(k + 1)(k + 2)\ldots(k + m) = \frac{n(n + 1)(n + 2)\ldots(n + m + 1)}{m + 2}, \tag{10}$$

and the following recursive formula

$$\begin{aligned} f_2(r) &= r+1 \\ f_3(r) &= \sum_{i=0}^{r} f_2(r-i) \\ f_4(r) &= \sum_{i=0}^{r} f_3(r-i) \\ &\vdots \\ f_n(r) &= \sum_{i=0}^{r} f_{n-1}(r-i), \end{aligned} \tag{11}$$

one may easily show that $f_n(r) = \binom{r+(n-1)}{n-1}$. Therefore, the cardinality of $\mathcal{P}_r = p^{f_n(r)} - 1 = p^{\binom{r+(n-1)}{n-1}} - 1$. If we construct the matrix Φ_0 using these polynomials just as in the two-variable case, the size of the matrix becomes $p^{n+1} \times p^{\binom{r+(n-1)}{n-1}} - 1$.

Remark 3. In the general case, it is difficult to obtain asymptotic estimates due to the presence of $\binom{r+(n-1)}{n-1}$ terms in the value for column size. The following theorem concludes that the matrix Φ_0 so defined is RIP compliant.

Theorem 7. *The matrix* $\Phi = \frac{1}{\sqrt{p^n}}\Phi_0$ *satisfies the RIP with* $\delta_k = \frac{k-1(r(p^{n-1}-1)+1)}{p^n}$ *for any* $k < \frac{p^n}{r(p^{n-1}-1)+1} + 1$.

Proof. Proof is the same as that of Theorem 3. □

Using the afore-discussed ideas, if we construct the circulant matrices Φ with respect to homogeneous r^{th} degree polynomials in n variables, then they satisfy RIP, as summarized below.

Theorem 8. *The circulant matrix* $\Phi_0 = \frac{1}{\sqrt{p^n}}\Phi$ *has the RIP with* $\delta_k = 2^{2n}(k-1)\frac{r}{p}$ *whenever* $k - 1 < \frac{p}{2^{2n}r}$.

Proof. Proof is the same as that of Theorem 3. □

Concluding Remarks

Inspired by the deterministic construction procedure proposed by DeVore [9], we have constructed a different class of CS matrices by using homogeneous multivariable polynomials. It should be emphasized here that the comparison in terms of recovery properties between the matrices obtained in this work and the ones in [9] does not make sense as both classes of matrices have different sizes. Despite it, we have noticed that the bound on the sparsity in our construction is wider as compared to that of Devore's matrix. However, the construction methodology becomes more relevant if it is capable of being used for constructing RIP compliant matrix of any order. In addition, for use in real-life applications, the columns of the matrix should contain an unequal number of 1's. Our future work shall attempt to address these issues.

Acknowledgements The first author gratefully acknowledges the support (Ref No. 20-6/2009(i)EU-IV) that he receives from UGC, Government of India. The last author is thankful to DST (SR/FTP/ETA-054/2009) for the partial support that he received.

References

1. Baraniuk, R., Davenpor, M., De Vore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constr. Approx. **28**(3), 253–263 (2008)
2. Bourgain, J., Dilworth, S., Ford, K., Konyagin, S., Kutzarova, D.: Explicit constructions of RIP matrices and related problems. Duke Math. J. **159**, 145–185 (2011)
3. Candes, E.: The restricted isometry property and its implications for compressed sensing. Comptes Rendus Mathematique **346**, 589–592 (2008)
4. Candes, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**, 1207–1223 (2006)
5. Candes, E., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**, 42–4215 (2005)
6. Candes, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
7. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Rev. **43**, 129–159 (2001)
8. Cohen, A., Dahmen, W., DeVore, R.: Compressed sensing and best k-term approximation. J. Am. Math. Soc. **22**(1), 211–231 (2009)
9. DeVore, R.A.: Deterministic constructions of compressed sensing matrices. J. Complexity **23**, 918–925 (2007)
10. Donoho, D.: Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006)
11. Donoho, D.L., Elad, M., Temlyakov, V.N.: Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inf. Theory **52**, 6–18 (2006)
12. Elad, M.: Sparse and Redundant Representations; from Theory to Applications in Signal and Image Processing. Springer, Berlin (2010)
13. Garnaev, A., Gluskin, E.: The widths of a Euclidean ball. Dokl. Akad. Nauk USSR **277**, 1048–1052 (1984); English transl. in Soviet Math. Dokl. **30**, 200–204 (1984)
14. Kashin, B.S.: Widths of certain finite-dimensional sets and classes of smooth functions. Izv. Akad. Nauk SSSR, Ser. Mat. **41**, 334–351 (1977); English transl. in Math. USSR IZV. **11**, 317–333 (1978)
15. Lidl, R., Niederreiter H.: Finite Fields, 2nd edn., pp. 275–277. Cambridge University Press, New York (1997)
16. Nelson, J.L., Temlyakov, V.N.: On the size of incoherent systems. J. Approx. Theory **163**(9), 1238–1245 (2011)

An Ant Colony Algorithm to Solve the Container Storage Problem

Ndèye Fatma Ndiaye, Adnan Yassine, and Ibrahima Diarrassouba

Abstract In this chapter we treat the container storage problem in port terminal. We study the storage of inbound containers in a port wherein straddle carriers are used as means of transport instead of internal trucks. In this work, unlike to the one that we did in Moussi et al. (LNCS 7197:301–310, 2012), reshuffles are not completely prohibited but are minimized. We consider additional constraints operational and propose a linear mathematical model. For numerical resolution we design an ant colony-based algorithm, named CSP-ANT. Several performed simulations prove the effectiveness of our algorithm.

Keywords Storage container • Ant colony algorithm • CPLEX

Introduction

In a seaport, the management of all containers is ensured by the container terminal. Its performance is an important criterion of the port productivity. This justifies many research work done concerning it. Generally there are three types of containers: inbound, outbound, and transshipment containers. Inbound containers are unloaded from ships by quay cranes (QC), then transported and placed to their storage locations by straddle carriers (SC), and at the end claimed by external trucks (ET). Outbound containers are brought by ET, also stored in the container yard, and so

N.F. Ndiaye (✉)
Laboratory of Applied Mathematics of Le Havre, University of Le Havre,
25 rue Philippe Lebon, B.P. 540, Le Havre Cedex 76058, France
e-mail: farlou@live.fr

A. Yassine
Superior Institute of Logistics Studies (ISEL), Quai Fissard, B.P. 1137,
Le Havre Cedex 76063, France
e-mail: adnan.yassine@univ-lehavre.fr

I. Diarrassouba
University Institute of Technology, Place Robert Schuman, Le Havre Cedex 76600, France
e-mail: diarrasi@univ-lehavre.fr

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.37_7

loaded onto ships later. Transshipment containers are unloaded from ships, then temporarily stacked in the storage areas, before being loaded onto other vessels. In this chapter we focus on the storage of inbound containers. Most of the work regarding this topic aims to minimize the number of reshuffles. In fact, these are unproductive movements requiring to move some containers in order to reach another. In [1], three different strategies of storage are proposed by Sauri and Martin. They have designed a mathematical model based on probabilistic distribution function, which minimize reshuffles. In [2], a segregation strategy is tackled by Kim and Bae. They do not mix containers which are unloaded from different vessels. In [3], a genetic algorithm is implemented by Jinxin and Qixin, to allocate storage spaces for inbound containers. They minimize the number of congestions, the waiting time of trucks, and the unloading time of containers. In [4], a modern container terminal is considered by Yu et Qi. To minimize reshuffles, they firstly resolve the block space allocation problem for newly arriving inbound containers. Then after the retrieval of some containers, they deal with the reorganization problem. Previously in [5], we have proposed an effective hybrid algorithm to store inbound containers in a terminal wherein reshuffles are not allowed. But with the continuous increase of the containership sizes, there may be situations where the available space is not sufficient to store all containers without causing any reshuffles. Therefore, we tackle in this chapter the case of storage where reshuffles are minimized instead of prohibited. We also determine the exact location assigned to every container, unlike the other papers which specify in the best case the assigned stack to each container. So, we propose a linear mathematical model and an ant colony-based algorithm to solve this problem. The remainder of this chapter is organized as follows: in the second section we explain the details of the studied problem, the mathematical model is highlighted in the third section, the fourth part is dedicated to our ant colony-based algorithm, in the fifth section we present our numerical results, and in the sixth and last section we give a conclusion.

Context

We deal with the storage problem of inbound containers in a modern container terminal. When a containership comes to port, the inbound containers are unloaded by the quay cranes and then placed on quays. Thereafter, they are collected individually by the straddle carriers which store them to the container yard. The collection order is the same as the unloading order. This avoids congestions on quays and decreases the waiting time of straddle carriers. We determine the exact storage location of each container and assume the six following main hypotheses.

1. Containers stored in a stack have the same sizes.
2. In a stack, containers are arranged following the order that they are unloaded from ships.
3. Each stack has a capacity equal to the difference between the maximum height authorized and the number of containers which are already within it.

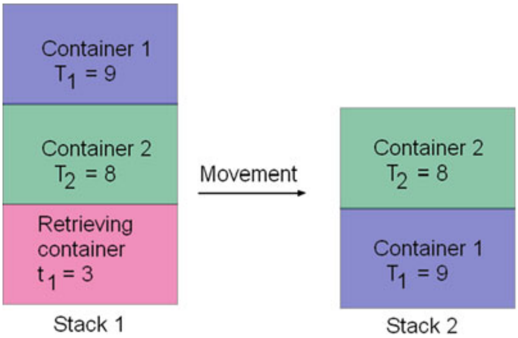


Fig. 1 A reshuffle

4. We take into account the departure times of containers which are already stacked before the actual storage period.
5. After every reshuffle, the moved containers are placed in an empty stack.
6. If a stack is empty then containers will be stored within it following the decreasing order of their departure times, else two cases can occur.

(6a) If containers assigned to this stack have departure times inferior or equal to those of containers which are already within, then they will be stored following the descending order of their departure times.

(6b) Otherwise, the newly incoming containers will be stored following the ascending order of their departure times. Thereby, upon retrieving the containers which are at the bottom, the remaining containers will be arranged following the descending order of their departure times into another stack, as shown in Fig. 1.

Mathematical Modeling

Notations

We use the following notations:

Indices
p: stack
i: empty slot
k: container

Data
N: number of containers
N_p: number of stacks
c_p: number of empty slots in the stack p
r_p: type of container which can be placed in the stack p

t_p: departure time of the container which was on the top of the stack p at the beginning of the new storage period
R_k: type of the container k
T_k: departure time of the container k
O_k: unloading order of the container k from ships
d_p^k: traveled distance to transport the container k from quay to stack p
M: a great integer

Decision variables

$$x_{p,i}^k = \begin{cases} 1 \text{ If container } k \text{ is assigned to the empty slot } i \text{ in the stack } p \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad \text{Otherwise} \end{cases}$$

$$y_p^k = \begin{cases} 1 \text{ If container } k \text{ is assigned to stack } p \text{ and this may cause reshuffles} \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad \text{Otherwise} \end{cases}$$

The Proposed Model

We model the problem as follows.

$$\text{Minimize} \quad \sum_{k=1}^{N}\sum_{p=1}^{N_p}\sum_{i=1}^{c_p} x_{p,i}^k d_p^k + M\sum_{k=1}^{N}\sum_{p=1}^{N_p} y_p^k \tag{1}$$

$$\sum_{p=1}^{N_p}\sum_{i=1}^{c_p} x_{p,i}^k = 1, \ \forall k = 1,\ldots,N \tag{2}$$

$$\sum_{k=1}^{N} x_{p,i}^k \leq 1, \ \forall p = 1,\ldots,N_p, \ i = 1,\ldots,c_p \tag{3}$$

$$\sum_{p=1}^{N_p}\sum_{i=1}^{c_p} x_{p,i}^k R_k - r_p = 0, \ \forall k = 1,\ldots,N \tag{4}$$

$$\sum_{k=1}^{N} (N - O_k)\, x_{p,i}^{k} \geq \sum_{k=1}^{N} \Big(N - O_k\Big) x_{p,i+1}^{k} \qquad (5)$$
$$\forall p = 1, \ldots, N_p,\;\; i = 1, \ldots, c_{p-1}$$

$$y_p^k \leq \sum_{i=1}^{c_p} x_{p,i}^{k}, \;\; \forall p = 1, \ldots, N_p,\; k = 1, \ldots, N \qquad (6)$$

$$\big(T_k - t_p\big) \sum_{i=1}^{c_p} x_{p,i}^{k} \leq M y_p^k \qquad (7)$$
$$\forall p = 1, \ldots, N_p,\;\; k = 1, \ldots, N$$

$$\sum_{k=1}^{N} T_k x_{p,i}^{k} - \sum_{k=1}^{N} T_k x_{p,i+1}^{k} \geq -M \sum_{k=1}^{N} y_p^k \qquad (8)$$
$$\forall p = 1, \ldots, N_p,\;\; i = 1, \ldots, c_{p-1}$$

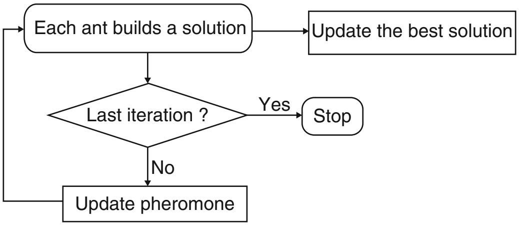
$$\sum_{k'=1}^{N} (M - T_{k'})\, x_{p,i}^{k'} - \sum_{k'=1}^{N} \Big(M - T_{k'}\Big) x_{p,i+1}^{k'} \geq M \left(y_p^k - 1\right) \qquad (9)$$
$$\forall p = 1, \ldots, N_p,\;\; i = 1, \ldots, c_{p-1},\;\; k = 1, \ldots, N$$

The objective function (1) minimizes simultaneously the number of reshuffles and the total distance traveled by SC between quays and the container yard. Constraints (2) require that each container is assigned to a single location. Constraints (3) ensure that several containers are not assigned to a same empty slot. Constraints (4) secure the compatibility between containers and stacks. Constraints (5) guarantee that in each stack, containers are arranged following the increasing order of their unloading from ships. Constraints (6) and (7) determine the number of induced reshuffles. If no rehandle is occasioned in a stack, constraints (8) impose that containers are stored there following the descending order of their departure times. Otherwise, if one of the containers assigned to a stack has a departure time superior to those of containers which are already within, then constraints (9) force that the newly stacked containers are arranged following the ascending order of their departure times.

Ant Colony Algorithm

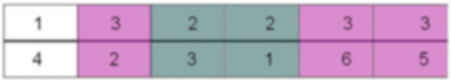
In 1996, Dorigo et al. [6] have created a new optimization method based on ant colonies, which is named the *ant system*. Natural ants have the ability to find the shortest path between their anthill and a place where there are food. At the beginning of the research, each ant follows the path which seems to be the shortest. Throughout the collection of food, each ant puts down a natural substance called pheromone along all paths that it has taken. Therefore, when an ant detects the presence of pheromone on a path, it concludes that this leads to food and follows it if it is shorter than the others. As the pheromone evaporates over time, at the end, it will remain only on the shortest path.

The ant colony algorithm mainly includes two steps: the construction of a solution by an ant and the update of pheromone. It can be summarized by the following flowchart.



Solution Representation

A solution is formulated as an array having two rows and *N* columns. Stacks are mentioned in the first row, and in the second there are containers. For example, let's assume that we have to store six containers in three stacks. A solution can be represented as follows:

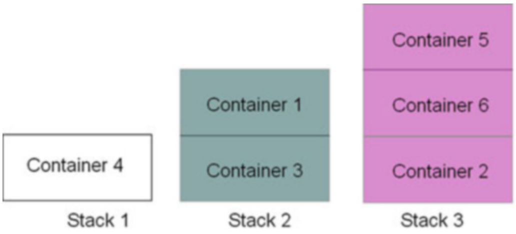


1	3	2	2	3	3
4	2	3	1	6	5

This means the following assignment:

- Container 4 to stack 1
- Containers 3 and 1 to stack 2
- Containers 2, 6, and 5 to stack 3

The appearance order is important to know the exact location of a container, if several containers are assigned to a same stack. In this case, the first container which appears in the solution is assigned to the lower location and so on. In fact containers of the example must be stacked in the following manner:



Construction Method of a Solution

In a solution, each column is a couple of container and stack which are compatible. Before the beginning of the solution search, we must firstly construct the set of couples. For this, we look for all pairs (*p,k*) of container and stack which satisfies two conditions:

- $c_p > 0$ (ensures that the stack *p* is not full)
- $R_k = r_p$ (verifies that the stack *p* and the container *k* have the same type)

To construct a solution, each ant chooses arbitrarily a starting duo. Then the set of options is updated in order to delete all pairs which can occasion violation of constraints if they are added to the solution. For each of the remaining couples a probability is calculated, and the ant chooses the one which has the highest probability. Whenever a duo is added to the solution, the set of options is updated, and the probability of each remaining option is determined in order to choose the next pair and so on until there is no available couple.

At the end of the construction of each solution, a check is performed to ensure that all containers are assigned. If a solution has a number of columns inferior to the number of containers, that means all containers are not assigned. So, this solution is eliminated and the ant starts the construction of another one.

Update the Set of Couples

In order to secure the feasibility of the solutions, the set of options is updated every time a pair (*k,p*) is chosen. For this, all couples having container *k* are deleted. Similarly, if the stack *p* is full, then all options containing it are eliminated. For every other pair, checks are done to prevent violations of the order constraints (5), (8), and (9). Thus each duo (k',p') such that $O_k > O_k{}'$ is obliterated, likewise if ($T_k \leq t_p$ and $T_{k'} > T_k$) or ($T_k > t_\mathrm{p}$ and $T_k{}' < T_k$).

Probability Calculation

Let S be the current set of options, α and β two positive real numbers, (p,k) an element of S, and $\rho_{(p,k)}$ the pheromone rate of (p,k).

The probability formula is

$$P_{(p,k)} = \frac{\left(\rho_{(p,k)}\right)^{\alpha} \times \left(\frac{1}{d_p^k}\right)^{\beta}}{\sum\limits_{(p,k)\in S} \left(\rho_{(p,k)}\right)^{\alpha} \times \left(\frac{1}{d_p^k}\right)^{\beta}}$$

where d_p^k is the traveled distance to transport the container k from the quay to stack the p and $\frac{1}{d_p^k}$ the visibility of (p,k).

Update Pheromone

At the beginning of the algorithm, each member of the initial set of couples has a pheromone rate equal to the maximum threshold ρ_{max}. But since in real life, pheromone is a substance which evaporates over time, we apply this property to our algorithm. For this, it is necessary to update progressively the pheromone rate of each option. So, at the end of each iteration, when all ants have constructed solutions, we reduce the pheromone rate of every element of the initial set of couples. Let γ be the evaporation rate and ρ_{min} the minimum threshold of pheromone. We multiply by ($1-\gamma$) the pheromone rate of each duo in order to reduce it. If the result is less than ρ_{min}, then we set it to ρ_{min}. Since the real ants secrete continuously pheromone during all the time required to collect food, there must be more pheromone on the most used path. It corresponds logically to the current shortest path. To reflect this in our algorithm, we add the following value to the pheromone rate of every option belonging to the best solution of the current iteration:

$$\frac{1}{1 + |Obj_{best} - Obj_{bestcour}|}$$

where Obj_{best} is the value of the best solution found since the beginning until the end of the current iteration, and $Obj_{bestcour}$ is the best solution found in the present iteration.

After each increase, we verify if the new rate exceeds ρ_{max}. If so, then we set it to ρ_{max}.

Numerical Results

We perform our algorithm in a computer DELL PRECISION T3500 with an Intel Xeon 5 GHz processor. At first, we look for the best values of the parameters, and after that, we test the performance of our algorithm.

Setting the Parameters

The evaporation rate γ, the pheromone exponent α, and the visibility exponent β are very important because they have an impact on the numerical results. Thus, to find the best values of these parameters, we have tested twenty different instances. For each of them, we consider ten values between 0.1 and 1 of every parameter. And then, we select the values which give best results at most cases. We obtain the following results: $\rho = 0.2$, $\alpha = 0.3$, and $\beta = 0.9$.

After performing several tests, we fix the number of ants to *10* and the number of iterations to *50*. ρ_{min} is set equal to *1* and ρ_{max} to *10*.

Comparison with CPLEX

CPLEX is an optimization software package which is generally used to do integer linear programming. Usually, it gives optimal results; thus, to know the quality of the solutions obtained by CSP-ANT, we calculate the percentages of deviation using the following formula:

$$dev = \frac{Obj_{(CSP-ANT)} - Obj_{(CPLEX)}}{Obj_{(CPLEX)}} \times 100$$

where $Obj_{(CSP\text{-}ANT)}$ is the value of the solution obtained by CSP-ANT and $Obj_{(CPLEX)}$ is the value of the optimal solution found by CPLEX.

All algorithms are coded in C++ language, and we use the version 12.5 of CPLEX. In Table 1, we report the percentages of deviation of several instances.

—means that the computer memory is insufficient to resolve this instance.

The numerical results show that CSP-ANT is very efficient and gives very good solutions which are very close to the optimal solutions. In addition to this, it is able to solve large instances which cannot be solved using CPLEX because it requires a lot of computer memory.

Table 1 Numerical results

N	N_p	$Obj_{(CPLEX)}$	$Obj_{(CSP\text{-}ANT)}$	dev (%)
10	100	1,624	1,624	0
15	100	2,475	2,487	0.48
20	100	3,334	3,371	1.11
25	100	4,257	4,363	2.49
30	100	5,265	5,510	4.65
35	100	6,361	6,510	2.34
40	100	7,514	7,814	3.99
45	100	8,757	9,352	6.79
50	100	10,081	10,538	4.53
55	100	11,506	11,791	2.47
60	100	12,953	13,229	2.06
65	100	14,480	14,641	1.11
70	100	16,058	16,957	5.59
75	100	17,706	18,318	3.45
80	100	19,402	19,976	2.95
85	100	21,162	21,500	1.59
90	100	23,826	24,397	2.39
95	100	26,293	26,812	1.97
100	100	28,346	28,871	1.85
200	3,500	–	33,592	–
300	3,500	–	51,727	–

Conclusion

In this chapter, we address the container storage problem in a port terminal. We improve widely the work that we did in [5] by minimizing simultaneously reshuffles and the total distance traveled by SC between quays and the container yard. We consider additional constraints such as the order in which containers are unloaded from vessels in order to avoid reshuffles at quays; we also determine the exact location assigned to every container. The major contributions of this chapter are the linear mathematical model and the effective ant colony-based algorithm (CSP-ANT). In fact the proposed algorithm is able to find sometimes the optimal results, and in most cases it gives very good solutions which are close to the optimal results. Its average percentage of deviation is equal to 2.73 %; this proves that CSP-ANT is more effective than the hybrid algorithm that we had proposed in [5], which has a percentage deviation equal to 10.22 %.

References

1. Sauri, S., Martin, E.: Space allocating strategies for improving import yard performance at marine terminals. Transport Res. Part E **47**(6), 1038–1057 (2011)
2. Kim, K.H., Kim, H.B.: Segregating space allocation models for container inventories in port container terminals. Int. J. Prod. Econ. **59**(1–3), 415–423 (1999)

3. Cao, J., Shi, Q., Der-Horng, L.: A decision support method for truck scheduling and storage allocation problem at container. Tsinghua Sci. Technol. **13**(Suppl 1), 211–216 (2008)
4. Mingzhu, Y., Xiangtong, Q.: Storage space allocation models for inbound containers in an automatic container terminal. Eur. J. Oper. Res. **226**(1), 32–45 (2013)
5. Moussi, R., Ndiaye, N.F., Yassine, A.: Hybrid genetic simulated annealing algorithm (HGSAA) to solve storage container problem in port. Intelligent Information and Database Systems (ACIIDS), Lecture Notes in Computer Science, vol. 7197, pp. 301–310. (2012)
6. Dorigo, M., Maniezzo, V., Colorni, A.: The ant system: optimization by a colony of cooperating agents. IEEE Trans. Syst. Man Cybernet. Part B **26**(1), 1–13 (1996)

FEM Post-processing in Identifying Critical Points in an Image

I.C. Cimpan

Abstract Separatrix segmentation is a data-driven method involving detection of ridges and valleys, which combines advantages of more widely used edge and region based techniques. Identifying saddle points is a vital step because ascending and descending slope lines are generated from the saddle points to define the separatrices that will generate the ridges (ascending slope lines) and valleys (descending slope lines). In our laboratory [2], we identified an important source of separatrix segmentation errors which were traced to the detection of an excess of saddle points in rectangular pixel images. The goal of this work is to compute a rigorous notion of critical point regardless of pixel shape (rectangular, hexagonal, triangular) and regardless of the number of neighbours. We want a well-defined discrete analogue of the Hessian test to determine if a critical point is a local minimum, maximum or saddle (Discrete Hessian Approach) This solution will consider FEM (Finite Element Method) Post-Processing techniques to estimate an image gradient and calculate critical points using rigorous mathematics.

Keywords Separatrix segmentation • FEM post-processing • Critical points

Flow of Ideas

Flow of Ideas: Segmentation

- Image segmentation (what it means, why we need it)
- Segmentation methods (vast literature in classifying them). Watershed and separatrix segmentation:
  - Identifying critical points in the image
  - Constructing separatrices starting from saddle points

I.C. Cimpan (✉)
SMSAS, University of Kent, Canterbury, Kent, UK
e-mail: icc4@kent.ac.uk

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.38_8

- Issues in defining critical points: different number of saddles due to different grids and different choice of nearest neighbourhood
- Aim: correct number of critical points regardless of pixel shape (rectangular, hexagonal, triangular) and regardless of the number of neighbours
  - We want to apply definitions accepted in continuous domain to a discrete domain in order to find the correct number of critical points in an image (Discrete Hessian Approach)

Flow of Ideas: Discrete Hessian Approach

- Mesh the image into triangulated elements [1]
- Construct basis functions on each node (vertex) of the image
- Estimate the gradient using basis functions and distributional derivative
- Calculate the zeroes of the estimated gradient
- Calculate the derivative of the estimated gradient and apply "Hessian test"
- Results in two-dimensional space and images

Introduction

Image Segmentation and Segmentation Methods

Image segmentation means partitioning the image into objects with respect to some characteristics like colour, intensity and texture. There is a vast usage of image segmentation, from which we mention: to identify/study anatomical structures, diagnosis, locate tumours and other pathologies.

The process of segmenting an image is difficult and time consuming and can be: manual, semi-automatic and automatic.

There are many segmentation methods including: region growing, classifiers artificial neural networks, Markov random field models, deformable models, atlas guided approaches, watershed methods and level set methods.

Watershed/Separatrix Segmentation

Low level cues derived from an image can be employed by data-driven segmentation, without any further information (e.g. models or templates) being necessary. Contrary to model based segmentation, which is concerned with a particular set of objects properties and types, data-driven segmentation focuses on generic-object detection.

The concepts of separatrices can be traced back to the work of [1] and of Maxwell in the nineteenth century. Cayley analysed the relation between a local minimum, saddle points and local maxima to evaluate water catchment areas in hydrology. He defined a watershed as a ridge line that “passes from summit to summit, through a single intervening knot (saddle point)”. Separatrix theory thus can be easily related to watershed segmentation which has been used widely in recent years [2–4]. The “multiple-saddle point problem” that appears in rectangular images using an 8-neighbourhood was identified by Rosin in [5]. He showed that four (4)-neighbour connectivity type detected insufficient saddle points; eight (8)-neighbour connectivity type detected too many saddle points; and six (6)-neighbour connectivity type indicated a more appropriate number of saddle points. In the figures below are shown examples of critical points in different lattices, rectangular and hexagonal, and different chosen neighbourhood (Fig. 1a) and separatrices in an example image (Fig. 1b) (see [6]).

Having identified the source of separatrix segmentation errors to be an excess of saddle points in the images, we want a well-defined discrete analogue of the Hessian test to determinate if a critical point is a local minimum, maximum or saddle (Discrete Hessian Approach). Our aim is to have a stronger definition for identifying critical points that can be applied to any pixel grid regardless of pixel shape (squared, rectangular, hexagonal or triangular) and regardless of number of pixels considered in the nearest neighbourhood (you have the option of 4-neighbours or 8-neighbours in rectangular and squared pixels). The goal is to compute a rigorous notion of critical point regardless of pixel shape (rectangular, hexagonal, triangular) and regardless of the number of neighbours. This method will use FEM (Finite Element Method) Post-Processing techniques and estimate a gradient based on hat functions.

Material and Methods

Separatrix segmentation is a data-driven method involving detection of ridges and valleys, which combines advantages of more widely used edge and region based techniques. Identifying precise critical points is a vital step because ascending and descending slope lines are generated from the saddle points towards maxima and minima to define the separatrices that will generate the ridges (ascending slope lines) and valleys (descending slope lines) (Fig. 2a).

Two separatrices running uphill to maxima from one saddle point correspond to the watersheds used in watershed segmentation.

The method to identify extremum and saddle points relates to the nearest neighbourhood only. The relationship between the grey value of a hexagonal pixel and the grey values of its neighbours (whether each difference is positive, negative or zero) is analysed and according to the number of sign changes critical points are identified. If we have no sign changes (all positive or all negative), then the point is an extremum, a maximum if all signs are negative, a minimum if all are positive. More than four sign changes identifies a saddle point.

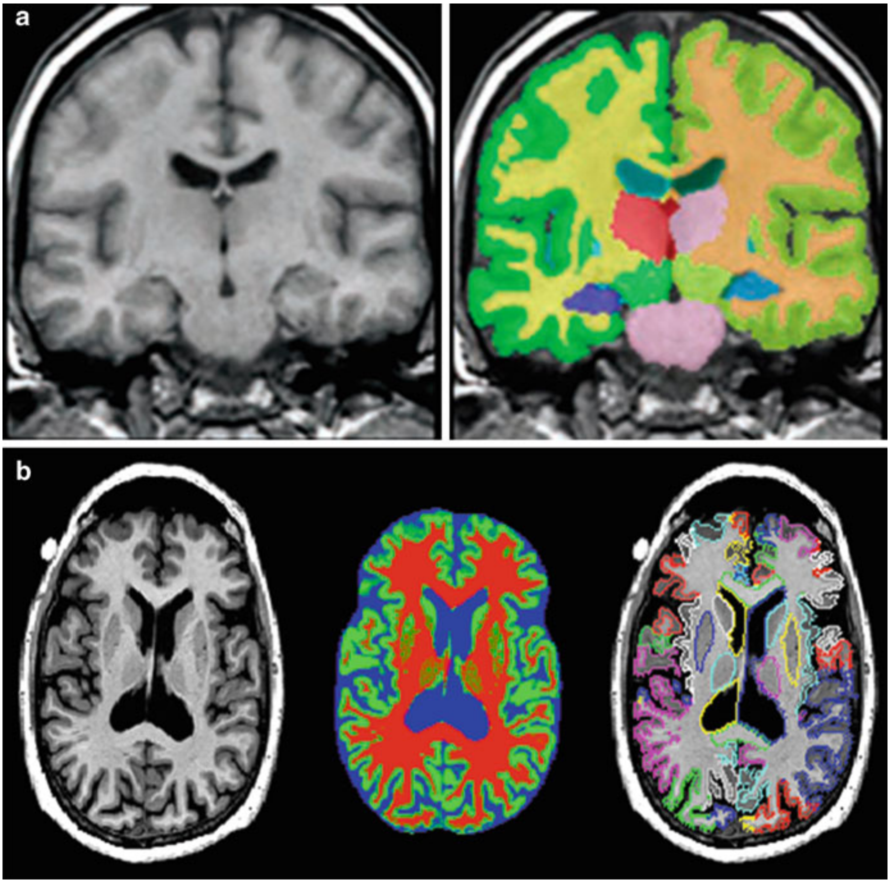


Fig. 1 Critical points and separatrices. (**a**) (*Left*) critical points in the hexagonal image and (*right*) critical points in the rectangular image (8-neighbourhood). Comparison: two different grids showing different number of critical points. (28) Minima in the hexagonal image and (37) minima in the original image. (26) Minima from the hexagonal image were in the same position or shifted with a distance less than the maximal hexagon diameter. (**b**) Separatrices on a hexagonal mesh

As we refine the mesh, we get more critical points using this definition of sign change, up to infinite number of critical points. Our interest is, by refining the mesh, to get, in a discrete domain, the critical points that are very close (or exact) to the critical points in the continuous domain.

We consider a function $u : \Omega \in \mathbb{R}^2 \to \mathbb{R}$ generated by an MRI scan to be a pixelated image whose value over a pixel grid $\mathcal{T}$ is a piecewise constant $u_h \in \mathbb{P}^0(\mathcal{T})$.

The goal is to compute some notion of critical point of function u. We want to use the "Hessian test" to determinate if a critical point is a local minimum, maximum or saddle (we want to use a well-defined discrete analogue of the Hessian test). Considering for simplicity the notation: $u_h \in \mathbb{P}^0(\mathcal{T})$ as a piecewise constant over

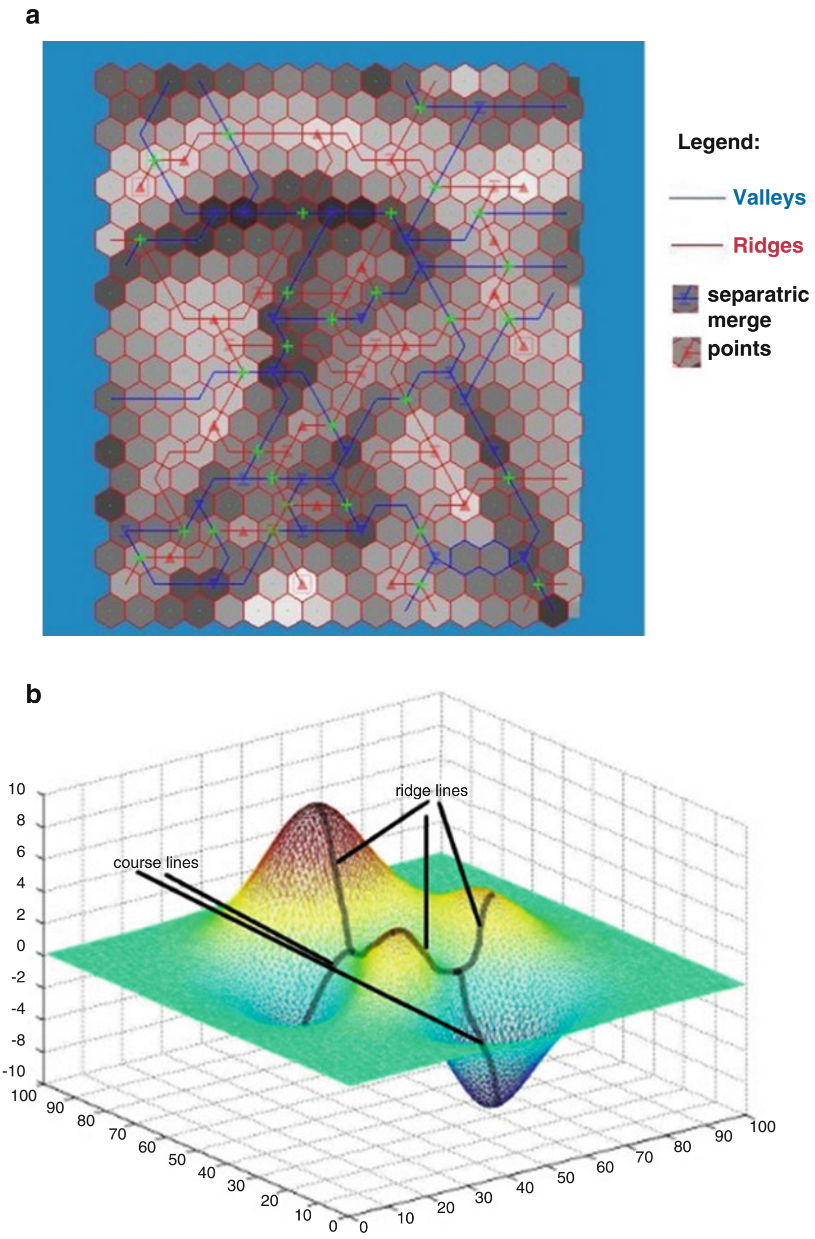


Fig. 2 Separatrices. (**a**) Ridges and valleys

the pixels, in the pixelated image $\mathcal{T}$, we use the distributional derivative to derive a formula for ∇u_h:

$$\begin{aligned} < \nabla u_h | \psi >_{H^{-1} \times H_0^1} = &- < u_h | \nabla \psi >_{L_2 \times L_2} \text{ as } u_h \in L_2(\Omega) \\ &= - \int_\Omega u_h \nabla \psi \\ &= - \sum_{\Omega_k \in \mathcal{T}} \int_{\Omega_k} u_h \nabla \psi \end{aligned} \tag{1}$$

for any $\psi \in \mathcal{C}_0^\infty(\Omega)$
where:

Ω_k= one element in the domain $\mathcal{T}$

$\psi \in \mathcal{C}_0^\infty(\Omega)$

$H_0^1 = \{\psi \in L_2(\Omega) | \nabla \psi \in L_2(\Omega) \text{ and } \psi|_{\partial\Omega} = 0\}$

Let V_h be a finite dimensional subspace of H_0^1. For all $\psi \in \mathbb{V}_h$ we have:

$$< G[u_h], \psi >_{L_2 \times L_2} = < \nabla u_h | \psi >_{H^{-1} \times H_0^1} \tag{2}$$

$$G[u_h] = \sum_{\mathfrak{n} \in \mathfrak{N}} d_\mathfrak{n} \psi_\mathfrak{n}(x) \tag{3}$$

where:

$d_\mathfrak{n} = n$ dimensional vector with constant elements (the unknowns)

$\psi_n(x) = \begin{cases} 1, & \text{if } x = n \text{ (on vertex)} \\ 0, & \text{on any other vertices} \end{cases}$

$G[u_h]$ will capture the distributional derivative of u_h over the finite element space $\mathbb{V}_h$.

Considering Eq. (2), we transform equation (1)
$< \nabla u_h | \psi > = - \sum_{\Omega_k \in \mathcal{T}} \int_{\Omega_k} u_h \nabla \psi$ into:

$$< G[u_h], \psi > = - \sum_{\Omega_k \in \mathcal{T}} \int_{\Omega_k} u_h \nabla \psi \tag{4}$$

From Eqs. (3) and (4) results:

$$\int\limits_\Omega G[u_h] \psi = \int\limits_\Omega (\sum_{j \in \mathbb{N}_v} d_j \psi_j) \psi_i = \sum_{j \in \mathbb{N}_v} \int\limits_\Omega d_j \psi_j \psi_i =$$

$$\sum_{j\in\mathbb{N}_v}\sum_{\Omega_k\in\mathcal{T}}\int\limits_{\Omega_k} d_j\psi_j\psi_i = \sum_{j\in\mathbb{N}_v} d_j \sum_{\Omega_k\in\mathcal{T}}\int\limits_{\Omega_k}\psi_j\psi_i =$$
$$\sum_{j\in\mathbb{N}_v} d_j\int\limits_{\Omega}\psi_j\psi_i = -\sum_{\Omega_k\in\mathcal{T}}\int_{\Omega_k} u_h\nabla\psi_i \tag{5}$$

with $i, n = \overline{1, \mathfrak{N}_v}$ and $k = \overline{1, \mathfrak{N}_e}$
and where:

$\mathfrak{N}_v$ = number of vertices in $\mathcal{T}$
$\mathfrak{N}_e$ = number of elements in $\mathcal{T}$

Equation (5) becomes:

$$\sum_{\Omega_k\in\mathcal{T}}\sum_{\mathfrak{n}\in\mathfrak{N}_v} d_{\mathfrak{n}} M_{ni}^k = -b_i, \tag{6}$$

where we have used the following notation:

$M_{ni}^k = \int_{\Omega_k}\psi_n\psi_i$ where $M_{ni} = \sum\limits_{\Omega_k\in\mathcal{T}} M_{ni}^k$ is called mass matrix
$b_i = -\int_{\Omega} u_h\nabla\psi_i$ is called load vector.

with $i, n = \overline{1, \mathfrak{N}_v}$ and $k = \overline{1, \mathfrak{N}_e}$
and where:

$\mathfrak{N}_v$ = number of vertices in $\mathcal{T}$
$\mathfrak{N}_e$ = number of elements in $\mathcal{T}$

So we need to calculate M_{ni} and b_i from the input data in order to find $d_{\mathfrak{n}}$ which are the unknowns when trying to calculate the estimated gradient $G[u_h]$. Enforced boundary conditions were used on the boundary. The estimated gradient values were replaced with the exact gradient values in order to avoid high oscillations.

Once we know the estimated gradient $G[u_h]$ we can calculate the zeroes x_0 of $G[u_h]$ and the derivative $DG[u_h](x_0)$

$$DG[u_h](x_0) = \sum_{\mathfrak{n}\in\mathfrak{N}_v} d_{\mathfrak{n}}\psi'_{\mathfrak{n}}(x_0) \tag{7}$$

$$G[u_h](x_0) = \sum\nolimits_{\mathfrak{n}\in\mathfrak{N}_v} d_{\mathfrak{n}}\psi_{\mathfrak{n}}(x_0)$$

Having now an estimated second derivative $DG[u_h]$ we can apply the well-known *Hessian Test* defined on a continuous domain.

In conclusion we have u_h approximating u, $G[u_h]$ to estimate ∇u and $DG[u_h]$ to estimate D^2u. We can analyse how closed are these approximations to the exact data. We want the error to converge to zero as the mesh size h is refined. The convergence results will appear in a forthcoming paper.

Results on Two-Dimensional Functions and Images

Results on Two-Dimensional Functions

Experimental tests on two-dimensional functions show that the estimated gradient $G[u_h]$ resembles the exact gradient closely on well-behaved functions. One significant result is plotted in Fig. 3 for the function $\sin(2\pi x)\sin(2\pi y)$: estimated gradient versus exact gradient in Fig. 3a and estimated critical points plotted on the mesh of exact function.

Initial Results on Images

Initial results on images show that, on sharp edges, the estimated gradients present oscillations that lead to inaccurate critical points. While the estimated gradients for two-dimensional functions can be adjusted by refining the mesh, on images we need to consider smoothing algorithms in order to avoid high oscillations on sharp edges. Figures 4 and 5 show two examples of cropped images: the original image and estimated critical points on the image mesh.

Conclusions and Further Work

The solution $G[u_h]$ presented for two-dimensional functions give proper experimental results on well-behaved functions. Figure 3 shows quality results for the estimated gradient and estimated critical points on two-dimensional functions. Furthermore, we can show mathematically and experimentally that the estimated gradient is converging towards the exact gradient, the derivative of the estimated gradient converging towards the exact second derivative and the estimated critical points converging towards the exact critical points.

Challenges appear when applying this solution on MRI images, because even the high-resolution MRI are not smooth enough, so we need additional smoothing algorithms. Different solutions like smoothing and resampling the image can be considered in a future work. Also, it is important to implement accurate boundary conditions on images, for this solution.

Further Work Next steps of this work require implementation of new boundary conditions for images: a polynomial interpolant of degree n_x, where n_x is the

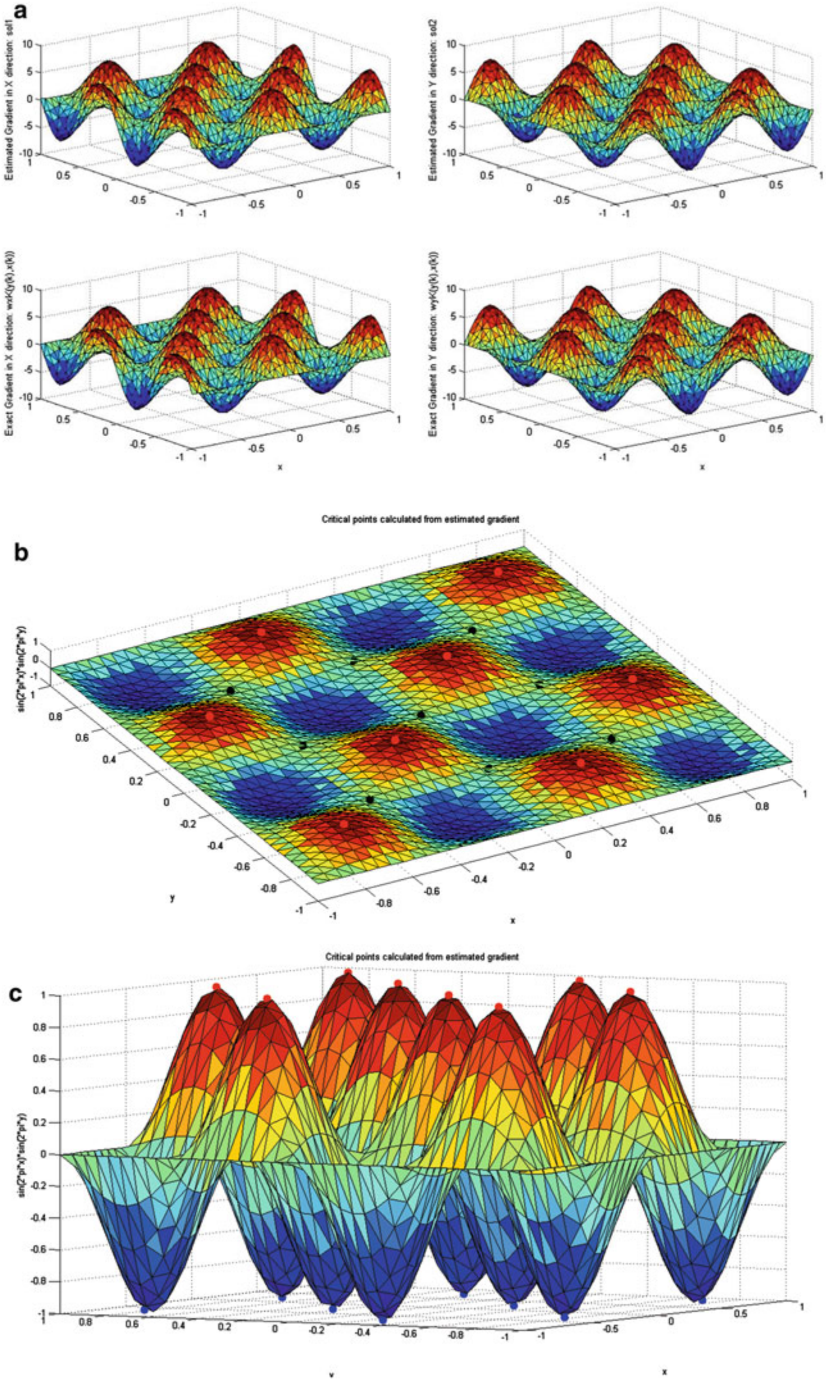


Fig. 3 Estimated gradient $G[u_h]$ versus exact gradient and critical points using $G[u_h]$ definition for function $\sin(2\pi x)\sin(2\pi y)$. (**a**) Estimated gradient $G[u_h]$ and exact gradient. (**b**, **c**) Critical points calculated from zeroes of $G[u_h]$

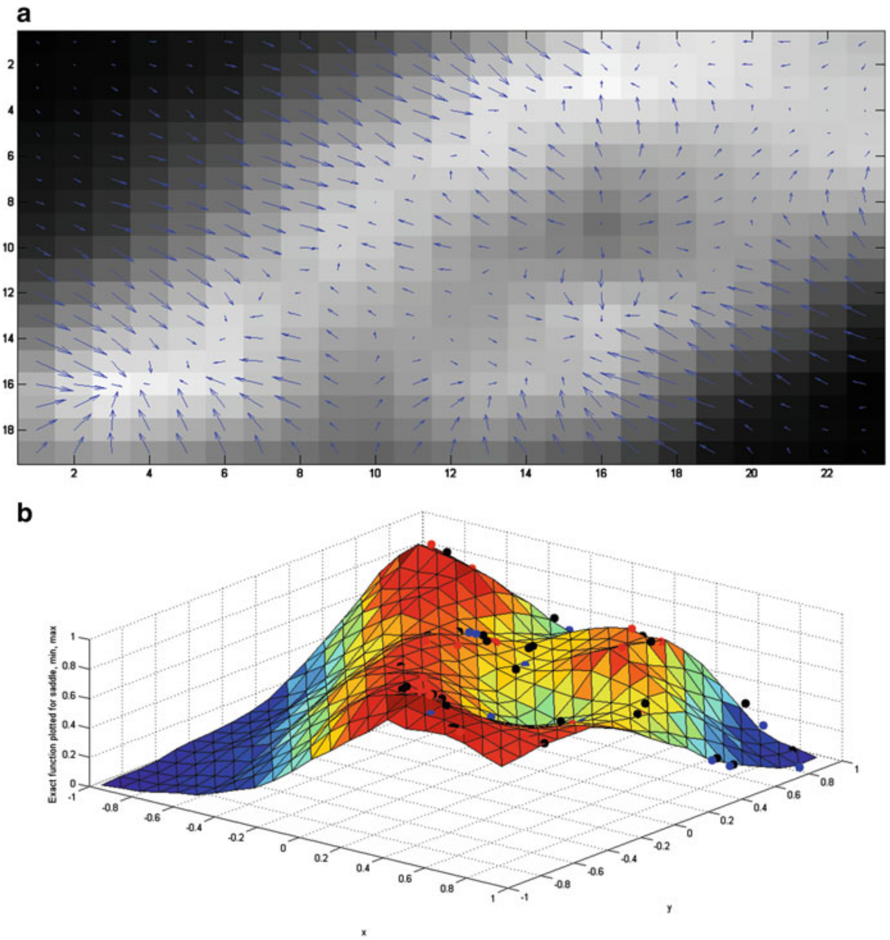


Fig. 4 Example 1 MRI for critical points using $G[u_h]$ definition on image mesh. (**a**) Original image (cropped). (**b**) Critical points of the original image

number of nodes in x direction. We also need a rigorous mathematical proof for the computational results above. Once we achieve accurate critical points on images, we can construct separatrices to enable automatic segmentation.

Acknowledgements We would like to acknowledge the helpful conversations with Tristan Pryer.

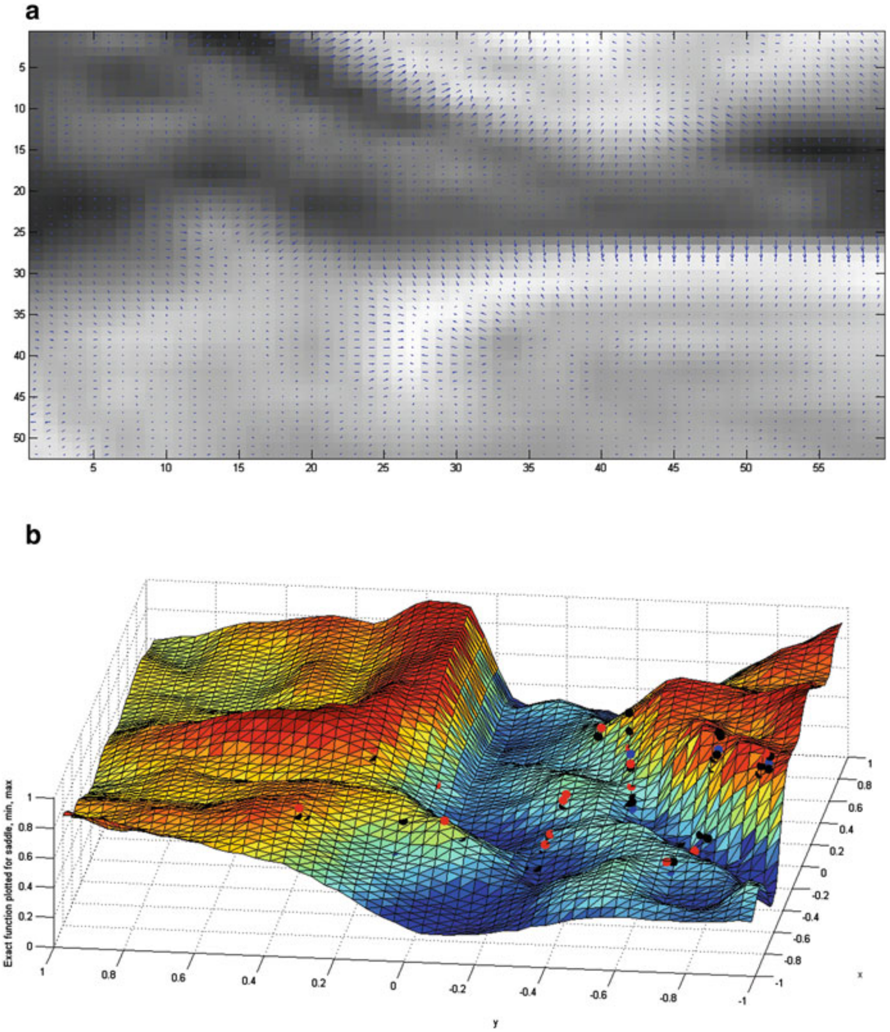


Fig. 5 Example 2 MRI for critical points using $G[u_h]$ definition on image mesh. (**a**) Original image (cropped). (**b**) Critical points of the original image

References

1. Cayley, A.: On contour and slope lines. Lond. Edinb. Dublin Philos. Mag. J. Sci. **18**, 264–268 (1859)
2. Meyer, F., Beucher, S.: Morphological segmentation. J. Vis. Commun. Image Represent. **1**, 21–46 (1990)
3. Digabel, H., and Lantuejoul, C. Iterative Algorithms, Actes du Second Symposium Europeen d'Analyse Quantitative des Microstructures en Sciences des Materiaux, Biologie et Medecine, Caen, 4-7 October 1977, J.-L. Chermant, Ed., Riederer Verlag, Stuttgart, pp. 85–99, (1978)

4. Gonzales, R.: Image segmentation. In: Digital Image Processing, International Edition, Digital image processing, Pearson Education India (2002)
5. Rosin, P.: Early image representation by slope districts. J. Vis. Commun. Image Represent. **6**, 228–243 (1995)
6. Cimpan, I.C.: Improved algorithms for converting rectangular pixels to hexagons. M.Sc. dissertation, University of Kent (2010)

Global and Local Segmentation of Images by Geometry Preserving Variational Models and Their Algorithms

Jack Spencer and Ke Chen

Abstract The imaging science as a research field is increasingly used in many disciplines (Mathematics, Statistics, Physics, Chemistry, Biology, Medicine, Engineering, Psychology, Computer and Information Science, etc.) because the imaging technology is being developed in a fast pace (with cost down and resolution up). The advance in imaging brings unprecedented challenges and demands on better image analysis techniques based on optimisation, geometry and nonlinear partial differential equations, beyond the traditional filtering-based linear techniques (FFT, wavelets, Wiener filters, etc.). Of course, in addition to modelling and analysis, there is an urgent need for advanced, accurate and fast computational algorithms.

In this paper we shall first discuss variational models that are frequently used for detecting global features in an image, i.e. all objects and their boundaries. These include the Chan-Vese (IEEE Trans Image Process 10(2):266–277, 2001) model of the Mumford and Shah (Commun Pure Appl Math 42:577–685, 1989) type and other related models. We then present a review on newer models that are designed to incorporate geometric constraints and detect local features in an image, i.e. local objects and their boundaries. In our first ever attempt, we compare six of such local selection models. Various test results are given to illustrate the models presented. Some open challenges are also highlighted.

Keywords Image processing • Selective segmentation • Level set function • Convex relaxation

Introduction

Image segmentation is a fundamental task in artificial intelligence for computer vision and imaging applications. Automatic segmentation is necessary and only accurate segmentation is useful. Among all the methods proposed for segmentation

J. Spencer (✉) • K. Chen
Centre for Mathematical Imaging Techniques and Department of
Mathematical Sciences, University of Liverpool, Liverpool, UK
e-mail: cmit@liv.ac.uk; k.chen@liv.ac.uk; www.liv.ac.uk/cmit

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.49_9

in the vast literature, variational methods are capable of achieving these two requirements [1–3].

Given an image $z = z(x, y),\ (x, y) \in \Omega \subset \mathbb{R}^2$, the segmentation model by Mumford-Shah [4] (MS) solves

$$\min_{u,K} F_{MS}(u, K) = \alpha \int_{\Omega\setminus K} |\nabla u|^2 dxdy + \mathcal{H}^1(K) \\ + \lambda \int_{\Omega} (u - z)^2 dxdy \tag{1}$$

for the reconstruction of an automatically segmented image u and the features' boundary set K, with $\mathcal{H}^1(K)$ the length of K. It is known that computing K numerically is a nontrivial matter and there are no direct solvers yet. A special case by the MS model, where u is a piecewise constant, can be solved with the help of level set functions as proposed in the Chan-Vese model [5]. The image is reconstructed as a cartoon of the original where each region, Ω_i, consists of homogeneous intensity (with $i = 1, \ldots, L$), separated by an edge set Γ, a closed subset of Ω. The two-phase example ($L = 2$) is of particular interest with $\Omega_1 = \text{in}(\Gamma)$ and $\Omega_2 = \text{out}(\Gamma)$,

$$PC(\Gamma, c_1, c_2) = \underbrace{\text{Length}(\Gamma)}_{regularisation} \\ + \underbrace{\lambda \int_{\text{in}(\Gamma)} (z - c_1)^2\, d\Omega\, + \lambda \int_{\text{out}(\Gamma)} (z - c_2)^2\, d\Omega}_{\text{fitting terms}}\,. \tag{2}$$

The above models have been widely used and extended to include new functionalities. Here we emphasise that most of such studies serve the original aims of MS, i.e. to identify all objects or boundaries present in a given image z.

However, in many practical applications, identifying all objects is expensive, challenging and above all not needed. Below we focus our attention on the so-called local and selective models (or interactive segmentation models) where only certain objects of the image are desired.

Selective Segmentation with Active Contours

The variational models discussed here are based on the work of two earlier papers: Le Guyader and Gout (2008) presented "Geodesic active contour under geometrical conditions" [6] that aimed to stop a developing contour on edges in the vicinity of a set of given markers, and Badshah and Chen [7] incorporated this idea to the global intensity fitting idea of Chan-Vese [5], with the model [7] given as

$$F(\Gamma, c_1, c_2) = \mu \int_\Gamma d(x) g(|\nabla z|)\, ds + \lambda_1 \int_{\mathrm{in}(\Gamma)} (z - c_1)^2\, d\Omega$$
$$+ \lambda_2 \int_{\mathrm{out}(\Gamma)} (z - c_2)^2\, d\Omega.$$

The details of the functions $d(x)$ and $g(x)$ are discussed later. This formulation is altered to extend the domain of the integrals to the entire domain, by using the level set approach [5], based on the work of Osher and Sethian [8]:

$$F^{LS}(\phi, c_1, c_2) = \mu \int_\Omega \delta_\epsilon(\phi) g |\nabla \phi|\, d\Omega$$
$$+ \lambda_1 \int_\Omega (z - c_1)^2 H(\phi)\, d\Omega$$
$$+ \lambda_2 \int_\Omega (z - c_2)^2 \big(1 - H(\phi)\big)\, d\Omega$$

The functional is minimised, using a regularised Heaviside and delta function, that will also be discussed later, by finding the Gateaux derivatives of F^{LS}. The Euler-Lagrange equation for ϕ can be considered as the steady state of the following PDE:

$$\frac{\partial \phi}{\partial t} = \mu \delta_\epsilon(\phi) \nabla \cdot \left(\frac{g \nabla \phi}{|\nabla \phi|} \right) + \alpha W(x) |\nabla \phi|$$
$$- \delta_\epsilon(\phi) \Big(\lambda_1 (z - c_1)^2 - \lambda_2 (z - c_2)^2 \Big),$$

with Neumann boundary conditions. Minimisation for c_1 and c_2:

$$c_1(\phi) = \frac{\int_\Omega z H_\epsilon(\phi)\, d\Omega}{\int_\Omega H_\epsilon(\phi)\, d\Omega},$$
$$c_2(\phi) = \frac{\int_\Omega z \big(1 - H_\epsilon(\phi)\big)\, d\Omega}{\int_\Omega \big(1 - H_\epsilon(\phi)\big)\, d\Omega}. \quad (3)$$

Model 1: Dual Level Set Model

The first model we look at is a dual level set model [9] introduced by Rada et al., motivated by the fact that the fitting term in Badshah and Chen [7] performs a global segmentation, and so new terms are needed to perform local segmentation. In the following, Γ_G is a contour that finds the global result and Γ_L finds the local result. This results in three additional terms: one for the regularisation of the local contour and three for local fitting. The variational model:

$$\begin{aligned} F_A(\Gamma, c_1, c_2) = \mu_1 \int_{\Gamma_L} & d(x,y) g(|\nabla z|)\, ds \\ & + \mu_2 \int_{\Gamma_G} g(|\nabla z|)\, ds \\ & + \lambda_{1G} \int_{\text{in}(\Gamma_G)} (z - c_1)^2 \, d\Omega \\ & + \lambda_{2G} \int_{\text{out}(\Gamma_G)} (z - c_2)^2 \, d\Omega \\ & + \lambda_1 \int_{\text{in}(\Gamma_L)} (z - c_1)^2 \, d\Omega \\ & + \lambda_2 \int_{\text{out}(\Gamma_L) \cap \text{in}(\Gamma_G)} (z - c_1)^2 \, d\Omega \\ & + \lambda_3 \int_{\text{out}(\Gamma_L) \cap \text{out}(\Gamma_G)} (z - c_2)^2 \, d\Omega . \end{aligned} \tag{4}$$

Rada-Chen, which we will refer to as the dual model, uses the level set formulation of [5], to represent inside and outside each contour. This gives an equivalent of (4):

$$\begin{aligned} F_D^{LS}(\phi, c_1, c_2) = \; & \mu_1 \int_\Omega d(x,y) g(|\nabla z|) \delta_\epsilon(\phi_L)) |\nabla \phi_L| \, d\Omega \\ + \mu_2 & \int_\Omega g(|\nabla z|) \delta_\epsilon(\phi_G) |\nabla \phi_G| \, d\Omega \\ + \lambda_{1G} & \int_\Omega (z - c_1)^2 H(\phi_G) \, d\Omega \\ + \lambda_{2G} & \int_\Omega (z - c_2)^2 \big(1 - H(\phi_G)\big) \, d\Omega \\ + \lambda_1 & \int_\Omega (z - c_1)^2 H(\phi_L) \, d\Omega \\ + \lambda_2 & \int_\Omega (z - c_2)^2 \big(1 - H(\phi_L) H(\phi_G) \, d\Omega \\ + \lambda_3 & \int_\Omega (z - c_2)^2 \big(1 - H(\phi_L)\big)\big(1 - H(\phi_G)\big) \, d\Omega . \end{aligned} \tag{5}$$

This functional is minimised using the Gateaux derivatives to derive the following PDEs for the global contour:

$$\frac{\partial \phi_G}{\partial t} = \mu_2 \delta_\epsilon(\phi_G) \nabla \cdot \left(\frac{g \nabla \phi_G}{|\nabla \phi_G|} \right) - \alpha g |\nabla \phi_G|$$

$$\begin{aligned}
&- \delta_\epsilon(\phi_G)\Big(\lambda_{1G}(z-c_1)^2 - \lambda_{2G}(z-c_2)^2\\
&+ \lambda_2(z-c_1)^2\big(1-H(\phi_L)\big)\\
&- \lambda_3(z-c_2)^2\big(1-H(\phi_L)\big)\Big),
\end{aligned}$$

and the local contour:

$$\begin{aligned}
\frac{\partial\phi_L}{\partial t} &= \mu_1\delta_\epsilon(\phi_L)\nabla\cdot\left(\frac{W\nabla\phi_L}{|\nabla\phi_L|}\right) - \alpha g|\nabla\phi_L|\\
&- \delta_\epsilon(\phi_L)\Big(\lambda_1(z-c_1)^2 - \lambda_2(z-c_2)^2 H(\phi_G)\\
&+ \lambda_2(z-c_1)^2\big(1-H(\phi_G)\big)\\
&- \lambda_3(z-c_2)^2\big(1-H(\phi_G)\big)\Big),
\end{aligned}$$

with Neumann boundary conditions. Minimisation for c_1 and c_2 follows, where H_ϵ^G, H_ϵ^L denote $H_\epsilon(\phi_G)$ and $H_\epsilon(\phi_L)$, respectively:

$$\begin{aligned}
I_1 &= \lambda_{1G}\int_\Omega zH_\epsilon^G d\Omega + \lambda_1\int_\Omega zH_\epsilon^L d\Omega\\
&\quad + \lambda_2\int_\Omega z\big(1-H_\epsilon^L\big)H_\epsilon^G d\Omega,\\
I_2 &= \lambda_{1G}\int_\Omega H_\epsilon^G d\Omega + \lambda_1\int_\Omega H_\epsilon^L d\Omega\\
&\quad + \lambda_2\int_\Omega \big(1-H_\epsilon^L\big)H_\epsilon^G d\Omega,\\
I_3 &= \lambda_{2G}\int_\Omega z\big(1-H_\epsilon^G\big)\, d\Omega\\
&\quad + \lambda_3\int_\Omega z\big(1-H_\epsilon^L\big)\big(1-H_\epsilon^G\big)\, d\Omega\\
I_4 &= \lambda_{2G}\int_\Omega \big(1-H_\epsilon^G\big)\, d\Omega,\\
&\quad + \lambda_3\int_\Omega \big(1-H_\epsilon^L\big)\big(1-H_\epsilon^G\big)\, d\Omega.
\end{aligned}$$

Then

$$c_1(\phi) = \frac{I_1}{I_2}, \qquad c_2(\phi) = \frac{I_3}{I_4}.$$

Model 2: Coefficient of Variation Model

Badshah et al. introduced a new model [10], linked to their previous model discussed earlier. The fitting term uses the variance given by

$$Var(z) = \frac{1}{N}\sum_{i,j}\left(z_{i,j} - Mean(z)\right)^2,$$

where $z_{i,j}$ denotes the image intensity at (i, j) and $Mean(z)$ is the mean intensity. The coefficient of variation (CoV) is defined as

$$CoV^2 = \frac{Var(z)}{(Mean(z))^2}.$$

This value is larger in areas of the image where there are edges and is employed in this model as a fitting term:

$$\begin{aligned} F_{CoV}(\Gamma, c_1, c_2) = \mu \int_\Gamma & d(x, y)g(|\nabla z|)\, ds \\ & + \lambda_1 \int_{\text{in}(\Gamma)} \frac{(z - c_1)^2}{c_1^2} d\Omega \\ & + \lambda_2 \int_{\text{out}(\Gamma)} \frac{(z - c_2)^2}{c_2^2}\, d\Omega. \end{aligned}$$

This is adjusted as before with the level set formulation:

$$\begin{aligned} F^{LS}_{CoV}(\phi, c_1, c_2) = \mu \int_\Omega & d(x, y)g(|\nabla z|)\delta_\epsilon(\phi)|\nabla\phi|\, d\Omega \\ & + \lambda_1 \int_\Omega \frac{(z - c_1)^2}{c_1^2} H(\phi)\, d\Omega \\ & + \lambda_2 \int_\Omega \frac{(z - c_2)^2}{c_2^2}\left(1 - H(\phi)\right) d\Omega. \end{aligned}$$

The PDE is derived by minimising this functional, to give

$$\begin{aligned} \frac{\partial\phi}{\partial t} =& \mu\delta_\epsilon(\phi)\nabla\cdot\left(\frac{W(x, y)\nabla\phi}{|\nabla\phi|}\right) + \alpha W(x, y)|\nabla\phi| \\ & - \delta_\epsilon(\phi)\left(\lambda_1\frac{(z - c_1)^2}{c_1^2} - \lambda_2\frac{(z - c_2)^2}{c_2^2}\right), \end{aligned}$$

with Neumann boundary conditions and where $W(x, y) = d(x, y)g(|\nabla z|)$. Minimisation for c_1 and c_2:

$$c_1(\phi) = \frac{\int_\Omega z^2 H_\epsilon(\phi)\, d\Omega}{\int_\Omega z\big(1 - H_\epsilon(\phi)\big)\, d\Omega},$$

$$c_2(\phi) = \frac{\int_\Omega z^2\big(1 - H_\epsilon(\phi)\big)\, d\Omega}{\int_\Omega z\big(1 - H_\epsilon(\phi)\big)\, d\Omega}.$$

Model 3: Area Model

This model, introduced by Rada et al., is again based on [7], whilst the growth of the contour is limited by the introduction of two new terms. Given a number of markers within the object of choice, the size of the object inside and outside the object respectively can be approximated by A_1 and A_2. This stops the contour evolving past the desired solution, by penalising the growth of the contour. The model is given as

$$\begin{aligned} F_A(\Gamma, c_1, c_2) = {} & \mu \int_\Gamma g(|\nabla z|)\, ds + \lambda_1 \int_{\text{in}(\Gamma)} (z - c_1)^2\, d\Omega \\ & + \lambda_2 \int_{\text{out}(\Gamma)} (z - c_2)^2\, d\Omega \\ & + \nu \Big(\int_{\text{in}(\Gamma)} d\xi d\eta - A_1 \Big)^2 \\ & + \nu \Big(\int_{\text{out}(\Gamma)} d\xi d\eta - A_2 \Big)^2. \end{aligned}$$

Adjusted, it gives the level set formulation:

$$\begin{aligned} F_A^{LS}(\phi, c_1, c_2) = {} & \mu \int_\Omega g(|\nabla z|)\delta_\epsilon(\phi(x, y))|\nabla\phi(x, y)|\, d\Omega \\ & + \lambda_1 \int_\Omega (z - c_1)^2 H(\phi(x, y)) d\Omega \\ & + \lambda_2 \int_\Omega (z - c_2)^2 \big(1 - H(\phi(x, y))\big) d\Omega \\ & + \nu \Big(\int_\Omega H(\phi(\xi, \eta)\, d\xi d\eta - A_1 \Big)^2 \\ & + \nu \Big(\int_\Omega \big(1 - H(\phi(\xi, \eta)\big)\, d\xi d\eta - A_2 \Big)^2. \end{aligned}$$

The PDE is then:

$$\frac{\partial\phi}{\partial t} = \mu\delta_\epsilon(\phi)\nabla\cdot\left(\frac{g(|\nabla z|)\nabla\phi}{|\nabla\phi|}\right) - \alpha g(|\nabla z|)|\nabla\phi| \\ - \delta_\epsilon(\phi)\Big(\lambda_1(z-c_1)^2 - \lambda_2(z-c_2)^2\Big) \\ + \nu\delta_\epsilon(\phi)\Big(\int_\Omega H\,d\Omega - A_1\Big)^2 \\ + \nu\delta_\epsilon(\phi)\Big(\int_\Omega \big(1-H\big)\,d\Omega - A_2\Big)^2,$$

with Neumann boundary conditions. Minimisation for c_1 and c_2 is as before (3).

Model 4: P-Model

This model is linked to the area model of the previous section. Given the same set of markers, a polygon P is formed. A new term, P_d, is then introduced that assigns each pixel a value based on its distance from P, i.e. 0 inside P and increasing as (i, j) gets further from P. This term, in a similar way to the area model [11], penalises the growth of the contour whilst providing additional information about the object, such as its location and boundary. The model is defined:

$$F_P(\Gamma, c_1, c_2) = \mu\int_\Gamma g(|\nabla z|)\,ds + \theta\int_{\text{in}(\Gamma)} P_d\,d\Omega \\ + \lambda_1\int_{\text{in}(\Gamma)}(z-c_1)^2\,d\Omega \\ + \lambda_2\int_{\text{out}(\Gamma)}(z-c_2)^2\,d\Omega.$$

Adjusted, it gives the level set formulation:

$$F_A^{LS}(\phi, c_1, c_2) = \mu\int_\Omega \delta_\epsilon(\phi)g|\nabla\phi|\,d\Omega \\ + \theta\int_\Omega P_d H(\phi)\,d\Omega \\ + \lambda_1\int_\Omega (z-c_1)^2 H(\phi)\,d\Omega \\ + \lambda_2\int_\Omega (z-c_2)^2\big(1-H(\phi)\big)\,d\Omega.$$

The functional is minimised, giving the PDE:

$$\begin{aligned}\frac{\partial\phi}{\partial t} = &\mu\delta_\epsilon(\phi)\nabla\cdot\left(\frac{g\nabla\phi}{|\nabla\phi|}\right)\\ &-\delta_\epsilon(\phi)\Big(\lambda_1(z-c_1)^2-\lambda_2(z-c_2)^2+\theta P_d\Big),\end{aligned}$$

with Neumann boundary conditions. Minimisation for c_1 and c_2 is as before (3).

Model 5: Local Information Model

This model develops the Badshah-Chen model, using the geometrical constraints of [6, 7] but only including pixels in a neighbourhood of the zero level set used in the data fitting. Results can be improved by only using information from this fixed narrow band, but the width is problem dependent and therefore causes problems with reliability. Zhang et al. [12] present a variable band that adjusts adaptively as the curve evolves, a local information (LI) model. The model generalises the Badshah-Chen model, with the addition of a local fitting energy function:

$$b(\phi,\gamma_{in},\gamma_{out}) = H(\phi+\gamma_{in})\big(1-H(\phi-\gamma_{out})\big),$$

that characterises the domain Ω_γ which is a narrow band region surrounding the boundary Γ. Inside Ω_γ, $b=1$ and outside $b=0$. The level set formulation of the model is given as

$$\begin{aligned}F_{JP}^{LS}(\phi,c_1,c_2) = &\mu\int_\Omega\delta_\epsilon(\phi)W|\nabla\phi|\,d\Omega\\ &+\lambda_1\int_\Omega(z-c_1)^2bH(\phi)\,d\Omega\\ &+\lambda_2\int_\Omega(z-c_2)^2b\big(1-H(\phi)\big)\,d\Omega,\end{aligned}$$

where $W=g(|\nabla z|)d(x)$. Zhang et al. discuss the details of how the width of the band is selected automatically in [12]. As before, the following PDE is derived:

$$\begin{aligned}\frac{\partial\phi}{\partial t} =&\mu\delta_\epsilon(\phi)\nabla\cdot\left(\frac{g\nabla\phi}{|\nabla\phi|}\right)\\ &-\lambda_1\Big(\delta(\phi)b+\frac{\partial b}{\partial\phi}H(\phi)\Big)(z-c_1)^2\\ &+\lambda_2\Big(\delta(\phi)b-\big(1-H(\phi)\big)\frac{\partial b}{\partial\phi}\Big)(z-c_2)^2,\end{aligned}$$

with Neumann boundary conditions. Minimisation for c_1 and c_2:

$$c_1(\phi) = \frac{\int_\Omega z H_\epsilon(\phi) b \, d\Omega}{\int_\Omega H_\epsilon(\phi) b \, d\Omega},$$

$$c_2(\phi) = \frac{\int_\Omega z\big(1 - H_\epsilon(\phi)\big) b \, d\Omega}{\int_\Omega \big(1 - H_\epsilon(\phi)\big) b \, d\Omega}.$$

Model 6: Interactive Segmentation Model

The interactive segmentation (IS) model employs the idea of Chan et al. [13] and Bresson et al. [14], to reformulate nonconvex models as convex ones. Two-phase segmentation involves a binary constraint, i.e. 1 inside and 0 outside the object. By allowing a function u to take intermediate values, this constraint is relaxed. When computing the global minimiser, u^*, thresholding the function at any value $p \in (0, 1)$ gives the contour of the object. Nguyen et al. formulate the problem as follows:

$$\min_{0 \leq u \leq 1} \left(\int_\Omega g_b(x) |\nabla u| \, d\Omega + \lambda \int_\Omega h_r(x) u \, d\Omega \right). \tag{6}$$

The fitting term, $h_r(x)$, is given based on user input; the foreground and background regions are marked on a subjective basis. A probability map $P(x)$ is used, based on the geodesic distances to the foreground and background regions, denoted by $D_F(x)$ and $D_B(x)$, respectively. This idea is based on the work of Sapiro et al. [15]. The estimate of the probability a pixel x belongs to the foreground is given as

$$P(x) = \frac{D_B(x)}{D_F(x) + D_B(x)}.$$

The model also uses foreground and background Gaussian mixture models (GMMs) introduced in [16]. Denote by $Pr(x|F)$ and $Pr(x|B)$ the probabilities that pixel x fits the foreground and background GMMs, respectively. The normalised log likelihood that x belongs to the foreground and background is

$$P_F(x) = \frac{-\log P_r(x|F)}{-\log Pr(x|F) - \log Pr(x|B)},$$

$$P_B(x) = \frac{-\log P_r(x|B)}{-\log Pr(x|F) - \log Pr(x|B)}.$$

The fitting term is then given as

$$h_r(x) = \alpha(P_B(x) - P_F(x)) + (1 - \alpha)(1 - 2P(x)),$$

where α is an automatically selected trade-off parameter based on the distance between the foreground and background GMMs.

Implementation

We will briefly discuss some of the details of implementing each of these models, which use similar ideas. Each functional is adjusted with the level set method, which incorporates the Heaviside function, H, into the formulation. In order to compute the first-order optimality conditions given by the derivatives of the functionals, H is replaced by an analytic approximation H_ϵ. In the papers discussed and the tests conducted, there are two choices for this approximation:

$$H_\epsilon^1(x) = \begin{cases} 0, & x < -\epsilon \\ \frac{1}{2}\left[1 + \frac{x}{\epsilon} + \frac{1}{\pi}\sin(\frac{\pi x}{\epsilon})\right], & |x| \le \epsilon, \\ 1, & x > \epsilon, \end{cases}$$

and

$$H_\epsilon^2(x) = \frac{1}{2}\left(1 + \frac{2}{\pi}\arctan\left(\frac{x}{\epsilon}\right)\right). \tag{7}$$

Some of the models use a distance function to keep the contour in the vicinity of a set of given markers. The distance function is chosen as

$$d(x) = \prod_{i=1}^{m}\left(1 - \exp\left(-\frac{|x - \hat{x}_i|^2}{2\sigma^2}\right)\right),$$

with [12] using a variable σ that alters the effect of d based on the positions of the markers. The edge detector in each model is given as

$$g(|\nabla z|) = \frac{1}{1 + |\nabla z|^2}. \tag{8}$$

Nguyen et al. introduce an alternative edge detector, where the probability map $P_F(x)$ is used in the following way. They apply the edge detector (8) to $P_F(x)$ and the image z, given as g_c and g_e, respectively. The weighting edge term is then given by

$$g_b(x) = \beta \cdot g_c(x) + (1 - \beta) \cdot g_e(x),$$

where β is computed in a similar way to α, based on the GMM map. This term picks up weaker edges than (8) alone.

Models 1–4 use an additive operator-splitting method (AOS), proposed by Tai et al. [17] and Weickert [18], involving the spatial reduction of a two-dimensional problem into two one-dimensional ones. Model 5 uses a banded time-marching scheme, and Model 6 uses a Split Bregman solver, first used in the convex segmentation context by Goldstein et al. [19].

Results

In this section we discuss the results for each model for three difficult test problems, shown in Fig. 1. These are challenging for a number of reasons. Each object contains intensity inhomogeneity and has other objects in close proximity, without a clear edge separating them. In these cases especially, selective segmentation models can be sensitive to both initialisation and parameter selection. For this reason it is important to examine the robustness of each model, by checking their performance under such variation. As a measure of accuracy, we use a region overlap measure, called the Tannimoto Coefficient (TC) [20], with A as the region segmented and B as the ground truth region, and TC is given by

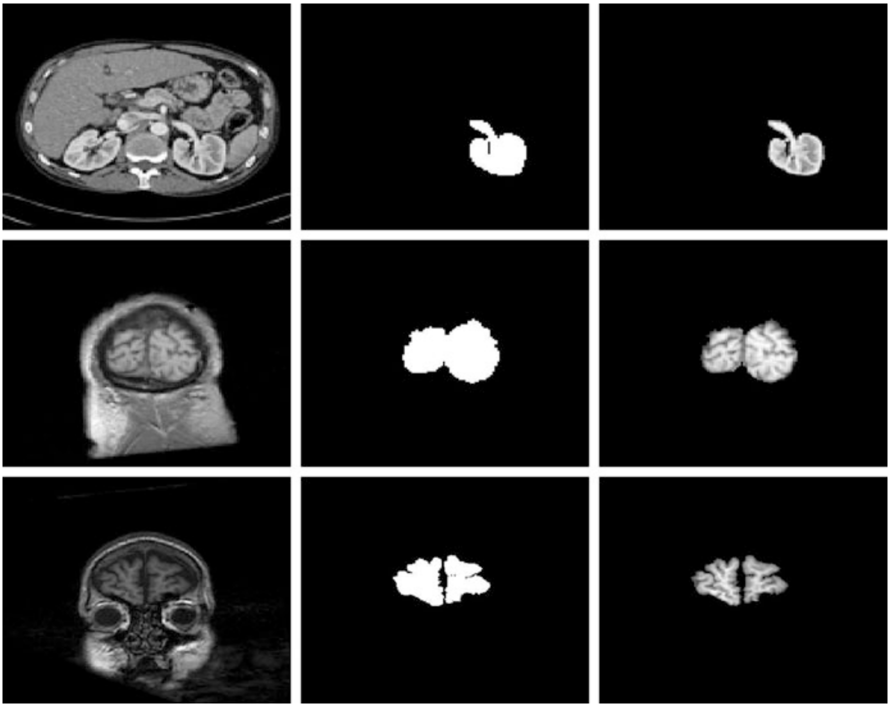


Fig. 1 Rows 1–3 show Sets 1–3 respectively. On the *left* is the given image, the *middle* is ground truth segmentation, and the *right* is the desired object

Table 1 Results for dual model

Set 1				
	s_1	s_2	s_3	s_4
M1	37.88	37.19	39.39	34.60
M2	**93.10**	92.42	93.01	59.96
Set 2				
	s_1	s_2	s_3	s_4
M1	81.14	**83.42**	80.17	82.47
M2	**81.61**	80.51	81.03	79.46
Set 3				
	s_1	s_2	s_3	s_4
M1	85.17	**86.60**	80.74	79.83
M2	**85.65**	84.65	80.57	78.52

$$TC = \frac{N(A \cap B)}{N(A \cup B)},$$

where $N(\cdot)$ indicates the number of pixels in the enclosed region. In each model the initial contour, Γ_0, is initialised as a polygon given by markers provided by the user. Ideally, success would not be too dependent on the choice of this polygon. We use two different marker sets for each test problem; M1 is a naive input within the chosen object and M2 is a refined selection that considers image information. We then compare the performance of each model for M1 and M2, under a variation of at least one parameter. In each table, where there is a visually successful segmentation the best result for each set, in terms of TC, is given in bold.

Model 1: Dual Model

The results for the dual model [9] by Rada and Chen are presented in Table 1. The parameters used in [9] are in most cases fixed. However, results can vary based on the choice of the regularisation of the Heaviside function for each level set. Specifically, both the choice of regularisation (7) and the choice of ϵ for each can be adjusted. Here, for ϕ_L, H^1_ϵ is used and, for ϕ_G, H^2_ϵ is used, and the choice of ϵ is tested. In Table 1 $s_1 - s_4$ refers to permutations of the choices of this parameter. We can see that a good result is achieved in all but one case, that is, Set 1 for M1. This is especially the case for Set 1, M2 which has exceptionally good results. One weakness of the model is that it is hard to predict what choice of s is appropriate; intuition does not play a role so that even a user experienced with the model cannot predict reliably what will work. Compounding this drawback is that the dual level sets require the computation of ϕ_G and ϕ_L at each iteration, increasing time taken to compute a solution. That being said, the results in these challenging test sets are consistently good and excellent for some.

Table 2 Results for CoV model

Set 1						
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
M1	29.71	28.76	23.43	26.29	34.67	34.29
M2	62.17	62.86	70.67	79.24	**87.90**	87.59

Table 3 Results for area model

Set 1						
	s_1^1	s_2^1	s_3^1	s_4^1	s_5^1	s_6^1
M1	60.65	61.22	57.22	66.02	**88.64**	82.32
M2	66.58	66.80	**87.09**	44.95	45.60	72.73
Set 2						
	s_1^2	s_2^2	s_3^2	s_4^2	s_5^2	s_6^2
M1	80.62	80.28	80.39	81.10	**81.40**	80.99
M2	50.28	34.71	34.64	70.43	70.35	69.82

Model 2: CoV Model

The results for the CoV model [10] are shown in Table 2. In [10] the parameter that was adjusted in their tests was the regularisation parameter, μ, and is varied here within the range set by Badshah et al. for their tests. The results were poor for Sets 2 and 3, so they have not been included. It is possible that the model is not capable of dealing with the intensity difficulties within the object selected in these two images. For Set 1 successful results were obtained for M2, which was typical for this model. Successful results were very dependent on initialisation, with very poor results when markers are not near the boundary of the object.

Model 3: Area Model

In Table 3 results for Sets 1 and 2 are presented. The results obtained for Set 3 were substandard for all parameters tested, indicating it was too challenging a problem for this model. In [11], the parameters that were adjusted for each problem were μ and ν, the regularisation and area constraint parameters, respectively. In Table 3, $s_1 - s_6$ refer to some permutation of these two parameters within the range given in [11]. These were not the same for each problem set, so s^1 and s^2 refer to Sets 1 and 2, respectively. A positive of this model is that, for Sets 1 and 2, there is a successful result for the naive M1 marker set, suggesting that little user knowledge is required for initialisation. There are also good results for a range of parameters that can be selected in a predictable way.

Table 4 Results for P-model

Set 1					
	θ_1^1	θ_2^1	θ_3^1	θ_4^1	θ_5^1
M1	87.62	**88.00**	87.81	67.62	**88.00**
M2	66.81	66.90	67.09	67.37	67.75
Set 2					
	θ_1^2	θ_2^2	θ_3^2	θ_4^2	θ_5^2
M1	69.92	68.89	67.95	65.88	65.06
M2	**82.23**	82.19	80.28	79.49	78.56
Set 3					
	θ_1^3	θ_2^3	θ_3^3	θ_4^3	θ_5^3
M1	64.74	**80.10**	78.14	72.93	69.26
M2	65.59	66.59	66.76	67.56	69.13

Model 4: P-Model

The results for the P-model are presented in Table 4, varying the area constraint parameter, θ, for the different marker sets. The range of θ is different for each test and is referred to by $\theta^1 - \theta^3$ for Sets 1–3, respectively. They show that the model achieves at least 80 % for each test, which is good. They also show that in Sets 1 and 3, this is achieved by the naive marker set M1. However, in those cases an improved marker set fails for the same parameter range. This highlights one drawback of the model. That is, it is sensitive to the area constraint. The selection of this parameter, θ, is dependent on the size and location of the polygon selected by the user and the size and shape of the desired object. In this case, the boundary of the polygon given by M2 is much closer to the boundary of the object than for M1. This means that for any new marker set, a new range of θ for which a successful segmentation will be achieved applies and therefore has to be discovered; consequently it is quite difficult to use intuitively. In the cases where there was a successful segmentation for M1, the corresponding TC percentage was given, although successful results were acquired for M2 as well.

Model 5: LI Model

The results for the LI model [12] for different marker sets are shown in Table 5. In [12] the parameter that is generally adjusted for their examples is the regularisation parameter, μ. The results show that a good result is achieved for Sets 1 and 2, whilst an acceptable result is achieved for Set 3. However, these results are all for the refined marker set M2; in Sets 1 and 3, we can see that for the simple M1 the LI model fails completely across the varied parameter, and Set 2 has a poor result. This indicates that the model is dependent on the initialisation, requiring some understanding of what image features will make an initialisation fail, i.e. for

Table 5 Results for LI model

Set 1					
	μ_1	μ_2	μ_3	μ_4	μ_5
M1	20.19	20.19	20.19	20.19	18.29
M2	88.40	**88.40**	88.21	82.13	74.19
Set 2					
	μ_1	μ_2	μ_3	μ_4	μ_5
M1	70.59	70.59	70.59	69.65	68.90
M2	81.20	81.20	**81.20**	81.10	78.50
Set 3					
	μ_1	μ_2	μ_3	μ_4	μ_5
M1	15.01	15.01	14.94	14.94	04.39
M2	75.33	75.33	75.84	**76.54**	53.74

Sets 2 and 3 the structure of the brain tissue involves intensity inhomogeneity that can be interpreted by the model as a series of objects, so that a simple initialisation can pick up some internal structure as the final result. An adjusted initialisation, nearer the boundary, produces a successful segmentation. The results are consistent across the range of the varied regularisation parameter, which indicates a strength of the model.

Model 6: IS Model

When testing the IS model [21], Nguyen et al. fixed the fitting parameter as $\lambda = 100$. Although this can be varied, with a smaller λ yielding a smoother contour, it is kept fixed in these tests. Instead, we look at four different foreground and background seed regions, I_n, for $n = 1, ..., 4$. We try a simple user input, I_1, which marks the object in a straightforward way. As n increases we incorporate more image information, such as difficult boundaries or awkward object shape. An example of this, in the case of Set 3, is shown in Fig. 2. This is an attempt to establish the level of detail required for user input in challenging cases. Table 6 shows the results for the IS model for the four different initialisations. They show that for the most basic user input, the model tends to fail. Slight refinement of the input produces consistently good results and in the case of Set 1 an excellent result. Whilst the TC percentage does not always increase as n increases, the speed of the Split Bregman solver means that repeat attempts can be made with intuitive adjustments of the input, until there is a successful result.

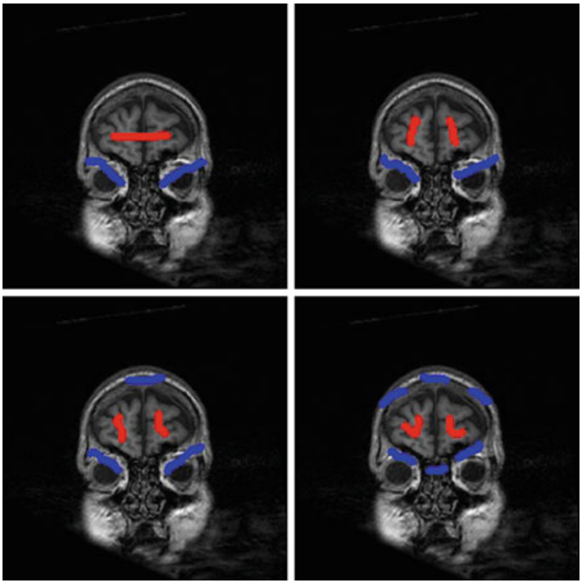


Fig. 2 These are the user initialisations for Set 3, from I_1 on the *left* to I_4 on the *right*. Similar progressions are used for Sets 1 and 2. *Blue* and *red* represent background and foreground regions, respectively

Table 6 Results for IS model

Set 1			
I_1	I_2	I_3	I_4
62.50	89.38	89.60	**92.40**
Set 2			
I_1	I_2	I_3	I_4
63.04	**82.94**	74.84	76.35
Set 3			
I_1	I_2	I_3	I_4
47.73	61.87	**88.23**	86.91

Future Developments

It is possible to explore the benefits of extending models 1–4 to the convex framework. In the nonconvex setting, the solutions computed are sometimes local minima, which are incorrect. Early results suggest some success by incorporating the work of [13] and [14], especially for the P-model. The IS model demonstrates the advantages of a fitting term that does not model the image as being piecewise constant. Incorporating and developing such ideas will be imperative to the future improvement of selective segmentation models. Examples of recent models that tackle intensity inhomogeneity within objects are region-scalable fitting [22], where Li et al. introduce a fitting term that incorporates intensity information in a local region at a controllable scale. This model uses the level set framework of models 1–5. A recent model, by Chen et al. [23], estimates a bias field estimator that adjusts the piecewise constant image model which can extend the type of image where

Table 7 Comparative results for all models

Model	Set 1	Set 2	Set 3
Dual	**93.10**	**83.42**	86.60
CoV	87.90	58.50	48.72
Area	87.09	81.40	61.30
P	88.00	82.83	80.10
LI	88.40	81.20	76.54
IS	92.40	82.94	**88.23**

selection of objects is viable. Other image information, rather than intensity, can be utilised, such as texture and shape. Klodt and Cremers use moment constraints [24] as a shape prior for segmentation, which can also aid the accurate selection of objects. These models [23, 24] use the convex relaxation framework of model 6.

Conclusions

We have examined the results for all six models, for three difficult test problems. In Table 7, a summary of the best results is presented. The CoV model wasn't effective for two test problems, perhaps indicative that it is not suitable for segmentation of the brain, due to the intensity structure of the desired object. The area model is of similar quality for two test sets, but fails in the third. Despite this, it is a model very capable of selectively segmenting difficult images, where other models would fail. The results show that the most accurate models are the dual model [9] of Rada and Chen and the IS model [21] of Nguyen et al., with the former best in Sets 1 and 2 and the latter best in Set 3. However, for the dual model there are some drawbacks on the choice of regularising the Heavisides for each level set in the functional (5), which is difficult to do intuitively. The interactive user input, together with the fast Split Bregman solver that allows repeat attempts, makes the IS model the most appropriate for selective segmentation. Alternatives, which are very reliable for all three problems, are the P-model and the LI model [12]. Each has good results, robust to variation, that can be used with a reasonable level of naivety on the part of the user.

Acknowledgements This paper reviews a joint CMIT work with many colleagues: Jianping Zhang, Lavdie Rada, Noor Badshah and Haider Ali. Other CMIT collaborators in imaging include T. F. Chan, R. H. Chan, B. Yu, C. Brito, L. Sun, F. L. Yang, N. Chumchob, M. Hintermuller, Y. Q. Dong and X. C. Tai.

References

1. Chan, T.F., Shen, J.H.: Image Processing and Analysis—Variational, PDE, Wavelet, and Stochastic Methods. SIAM, Philadelphia (2005)
2. Mitiche, A., Ben-Ayed, I.: Variational and Level Set Methods in Image Segmentation. Springer Topics in Signal Processing. Springer, New York (2010)
3. Vese, L.A.: Variational Methods in Image Processing. Chapman and Hall/CRC, Boca Raton (2013)
4. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. **42**, 577–685 (1989)
5. Chan, T.F., Vese, L.: Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
6. Le Guyader, C., Gout, C.: Geodesic active contour under geometrical conditions theory and 3D applications. Numer. Algorithms **48**, 105–133 (2008)
7. Badshah, N., Chen, K.: Image selective segmentation under geometrical constraints using an active contour approach. Commun. Comput. Phys. **7**(4), 759–778 (2009)
8. Osher, S., Sethian, J.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. **79**(1), 12–49 (1988)
9. Rada, L., Chen, K.: A new variational model with dual level set functions for selective segmentation. Commun. Comput. Phys. **12**(1), 261–283 (2012)
10. Badshah, N., Chen, K., Ali, H., Murtaza, G.: Coefficient of variation based image selective segmentation model using active contours. East Asian J. Appl. Math. **2**, 150–169 (2012)
11. Rada, L., Chen, K.: Improved selective segmentation model using one level-set. J. Algorithms Comput. Technol. **7**(4), 509–540 (2013)
12. Zhang, J., Chen, K., Yu, B., Gould, D.A.: A local information based variational model for selective image segmentation. Inverse Problems Imaging **8**(1), 293–320 (2014)
13. Chan, T.F., Esedoglu, S., Nikilova, M.: Algorithms for finding global minimizers of image segmentation and denoising models. SIAM J. Appl. Math. **66**, 1932–1648 (2006)
14. Bresson, X., Esedoglu, S., Vandergheynst, P., Thiran, J.P., Osher, S.: Fast global minimization of the active contour/snake model. J. Math. Imaging Vision **28**, 151–167 (2007)
15. Bai, X., Sapiro, G.: A geodesic framework for fast interactive image and video segmentation and matting. In: IEEE ICCV, pp. 1–8 (2007)
16. Yang, C., Duraiswami, R., Gumerov, N., Davis, L.: Improved fast gauss transform and efficient kernel density estimation. In: IEEE ICCV, pp. 664–671 (2003)
17. Lu, T., Neittaanmaki, P., Tai, X.C.: A parallel splitting-up method for partial differential equations and its applications to navier-stokes equations. RAIRO Math. Modell. Numer. Anal. **26**(6), 673–708 (1992)
18. Weickert, J.: Efficient and reliable schemes for nonlinear diffusion filtering. IEEE Trans. Image Process. **7**, 398–410 (1998)
19. Goldstein, T., Bresson, X., Osher, S.: Geometric applications of the split bregman method: Segmentation and surface reconstruction. J. Sci. Comput. **45**(1–3), 272–293 (2010)
20. Crum, W.R., Camara, O., Hill, D.L.G.: Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans. Med. Imaging **25**(11), 1451–1461 (2006)
21. Nguyen, T., Cai, J., Zhang, J., Zheng, J.: Robust interactive image segmentation using convex active contours. IEEE Trans. Image Process. **21**, 3734–3743 (2012)
22. Li, C.M., Kao, C.Y., Gore, J.C., Ding, Z.H.: Minimization of region-scalable fitting energy for image segmentation. IEEE Trans. Image Process. **17**(10), 1940–1949 (2008)
23. Chen, D., Yang, M., Cohen, L.D.: Global minimum for a variant mumford-shah model with application to medical image segmentation. Comput. Methods Biomech. Biomed. Eng. **1**(1), 48–60 (2013)
24. Klodt, M., Cremers, D.: A convex framework for image segmentation with moment constraints. In: IEEE International Conference on Computer Vision (ICCV) (2011)

Part II
Pure Mathematics

Positive and Negative Interval Type-2 Generalized Fuzzy Number as a Linguistic Variable in Interval Type-2 Fuzzy Entropy Weight for MCDM Problem

Nurnadiah Zamri and Lazim Abdullah

Abstract Generalized fuzzy number is an extended method of fuzzy number. Generalized fuzzy number has received significant attention from researchers in many areas. However, most of the generalized fuzzy number is defined only on one side which is in positive generalized fuzzy number. Therefore, the aim of this paper is to introduce a new generalized fuzzy number which considers positive and negative side in the concept of interval type-2 fuzzy set (IT2FS). Then, a new linguistic variable is established from the concept of new generalized fuzzy number. This new linguistic variable is applied into the interval type-2 entropy weight for multi-criteria decision-making (MCDM) method. Interval type-2 entropy weight is chosen as the weight in MCDM because the determination of this weight in the existing project delivery decision-making model relies on experts' knowledge and experience excessively. An aggregation process in MCDM method is modified in line with the new linguistic variable. An illustrative example is used in order to check the efficiency of the new method. This approach offers a practical, effective, and simple way to produce a comprehensive judgment.

Keywords Interval type-2 fuzzy set (IT2FS) • Generalized fuzzy number • Interval type-2 entropy weight • Multi-criteria decision-making (MCDM)

N. Zamri (✉)
Faculty of Informatics and Computing, University Sultan Zainal Abidin, Besut Campus, 22200 Besut, Terengganu, Malaysia
e-mail: nadiahzamri@unisza.edu.my

L. Abdullah
Department of Mathematics, Faculty of Science and Technology, University Malaysia Terengganu, Kuala Terengganu, Terengganu 21030, Malaysia
e-mail: lazim_m@umt.edu.my

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.30_10

Introduction

A fuzzy number is a fuzzy set on real numbers [1] or we can say that a fuzzy number is a fuzzy subset in support real number which is both "normal" and "convex" [2]. Fuzzy number is linked to the development of fuzzy sets from [3]. Beníteza et al. [4] stated that fuzzy number is an extremely suitable methodology that embraces adequately subjective knowledge and objective knowledge.

In 1985, Chen [5] proposed the concept of generalized fuzzy numbers, one of the extend method from the idea of Zadeh [3]. Generalized fuzzy number is believed can point out many cases it is not to possible to restrict the membership function to the normal form [6]. Since then, tremendous efforts are spent and significant advances are made on the development of numerous methodologies for comparing of generalized fuzzy numbers. For example, Cho et al. [7] introduced the notion of generalized fuzzy numbers and generalized fuzzy mappings and give Fubini theorem for integrals of generalized fuzzy mappings. Chen and Chen [8] proposed a fuzzy risk analysis on the ranking of generalized trapezoidal fuzzy numbers (GTFNs). Chen and Chen [9] also established a fuzzy risk analysis based on the ranking of generalized fuzzy numbers with different heights and different spreads. Farhadinia and Ban [10] extended a similarity measure of GTFNs to similarity measures of generalized trapezoidal intuitionistic fuzzy numbers (GTIFNs) and generalized interval-valued trapezoidal fuzzy numbers (GIVTFNs) such that the initial properties are to be preserved.

Since that, various authors have discussed on generalized fuzzy number method in decision making field. For example, Wei et al. [11] developed a generalized triangular fuzzy correlated averaging (GTFCA) operator. The GTFCA operator has been successfully applied to the multiple attribute decision-making problems with triangular fuzzy information. Su et al. [12] extended the induced generalized ordered weighted average (IGOWA) operator to a new aggregation operator called induced generalized intuitionistic fuzzy ordered weighted averaging (IG-IFOWA) operator for multi-attribute group decision-making. Yu et al. [13] distributed a new fuzzy multi-criteria decision-making (MCDM) approach for generalized fuzzy numbers based on the proposed ranking fuzzy numbers.

Generalized fuzzy number stated only on a positive range fuzzy set and neglecting the negative side. Banking on the premises of the statement that every matter has two sides such as negative and positive and bad and good [14]. It has proven by Ying and Yang's theories; where it is becoming one rotundity when both side turn into complementary. Specifically the objective of this paper is to propose a new generalized fuzzy number which considers both sides which are positive and negative fuzzy numbers in terms of IT2FSs. This new generalized fuzzy number is implemented as a linguistic number and applied into the interval type-2 entropy weight for MCDM method. Moreover, a modified distance measure is performed in tandem with the new fuzzy linguistic variable.

In recent years, there are lots of papers discussed on an interval type-2 fuzzy TOPSIS, but too little attention has been paid focused on a new positive and negative generalized fuzzy number as the linguistic variable in interval type-2 fuzzy entropy weight for interval type-2 fuzzy TOPSIS manner. For example, Chen and Lee [15] presented an interval type-2 fuzzy TOPSIS (IT2 FT) method to handle fuzzy multiple attribute group decision-making problems based on IT2FSs where the weights of the attributes and ratings for alternatives were in interval type-2 fuzzy linguistic variable terms. In another example, Chen and Lee [16] presented a new method to handle fuzzy multiple attribute group decision-making problems based on the ranking values and the arithmetic operations of IT2FSs.

For these reasons, this new method gives broad space to consider uncertain and vague because it uses IT2FS rather than fuzzy set. This approach is seen to provide a new perspective in fuzzy type-2 decision-making environment. They offer a practical, effective, and low-risk computation to produce a comprehensive judgment.

Preliminaries

Interval Type-2 Fuzzy Sets

This section briefly reviews some definitions of type-2 fuzzy sets and IT2FSs from Mendel et al. [17].

Definition 2.1 [17]. A type-2 fuzzy set $\tilde{\tilde{A}}$ in the universe of discourse X can be represented by a type-2 membership function $\mu_{\tilde{\tilde{A}}}$, shown as follows:

$$\tilde{\tilde{A}} = \left\{ \left((x,u)\,, \mu_{\tilde{\tilde{A}}}(x,u) \right) | \forall x \in X, \forall u \in J_x \subseteq [0,1]\,, 0 \le \mu_{\tilde{\tilde{A}}}(x,u) \le 1 \right\}, \quad (1)$$

where J_x denotes an interval in [0, 1]. Moreover, the type-2 fuzzy set $\tilde{\tilde{A}}$ also can be represented as follows:

$$\tilde{\tilde{A}} = \int\limits_{x \in X} \int\limits_{u \in J_x} \mu_{\tilde{\tilde{A}}}(x,u) \,/\, (x,u)\,, \quad (2)$$

where $J_x \subseteq [0,1]$ and $\iint$ denotes the union over all admissible x and u.

Definition 2.2 [17]. Let $\tilde{\tilde{A}}$ be a type-2 fuzzy set in the universe of discourse X represented by the type-2 membership function $\mu_{\tilde{\tilde{A}}}$. If all $\mu_{\tilde{\tilde{A}}} = 1$, then A is called an IT2FSs. An IT2FS $\tilde{\tilde{A}}$ can be regarded as a special case of a type-2 fuzzy set, represented as follows:

$$\tilde{\tilde{A}} = \int_{x \in X} \int_{u \in J_x} 1/\,(x,u), \quad (3)$$

where $J_x \subseteq [0,1]$.

Entropy Weight

Definition 2.2 [18]. Entropy weight is a parameter that describes how much different alternatives approach one another in respect to a certain attribute [18]. Conversely, low information entropy is a sign of a highly organized system. In information theory, the entropy value can be calculated as in Eq. (4)

$$H\left(p_1, p_2, \ldots, p_n\right) = -\sum_{j=1}^{n} p_j \ln p_j, \tag{4}$$

where H is the level of entropy and p_j is the probability of occurrence of event.

Szmidt and Kacprzyk [19] proposed a new entropy method for IFS. In their paper, they proposed the IF entropy as a ratio of distances between (F, F_{near}) and (F, F_{far}). The expression is given as follows:

$$E_{SK}(F) = \frac{(F, F_{near})}{\left(F, F_{far}\right)} \tag{5}$$

where (F, F_{near}) is the distance from F to the nearer point F_{near} among positive ideal point and negative ideal point and (F, F_{far}) is the distance from F to the farther point F_{far} among positive ideal point and negative ideal point. De Luca and Termini [20] have already proposed the axioms of entropy for FSs. Szmidt and Kacprzyk [19] then expressed IF entropy in the following definition:

Generalized Fuzzy Number

Let $\tilde{A}$ be a generalized trapezoidal fuzzy number, $\tilde{A} = \left(a_1, a_2, a_3, a_4; w_{\tilde{A}}\right)$, where a_1, a_2, a_3, and a_4 are real values, $w_{\tilde{A}}$ is the maximum membership value of the generalized trapezoidal fuzzy number $\tilde{A}$, and $w_{\tilde{A}} \in [0, 1]$. If $-1 \leq a_1 \leq a_2 \leq a_3 \leq a_4 \leq 1$, then $\tilde{A}$ is called a standardized generalized fuzzy number. If $w_{\tilde{A}} = 1$, then $\tilde{A}$ becomes a traditional fuzzy number and it can be represented as $\tilde{A} = (a_1, a_2, a_3, a_4)$. If $a_2 = a_3$, then $\tilde{A}$ is a triangular fuzzy number. If $a_1 = a_2 = a_3 = a_4$, then $\tilde{A}$ is crisp values.

According to Wang and Luo [21], the membership function $\mu_{\tilde{A}}$ of a generalized fuzzy number $\tilde{A}$ is defined as follows:

$$\mu_A(x) = \begin{cases} f_{\tilde{A}}^{L}(x), & a_1 \leq x \leq a_2, \\ w_{\tilde{A}}, & a_2 \leq x \leq a_3, \\ f_{\tilde{A}}^{R}(x), & a_3 \leq x \leq a_4, \\ 0, & otherwise, \end{cases} \tag{6}$$

where $f_{\tilde{A}}^{L}(x)$ and $f_{\tilde{A}}^{R}(x)$ are continuous mapping functions and $w_{\tilde{A}} \in [0, 1]$.

Some examples of generalized fuzzy numbers, shown as follows:

The combination between three basic concepts (“interval type-2 fuzzy sets,” “entropy weight,” and “generalized fuzzy number”) will produced a new linguistic variable shown in section “Development of PNIT2GFN.” This new linguistic variable is applied into an interval type-2 entropy weight in MCDM method in section “Decision-Making Based on the Interval Type-2 Entropy Weight.” Then, a numerical example from Yu et al. [13] is presented to test the effectiveness of the new model (shown in section “Numerical Example”).

An Algorithm

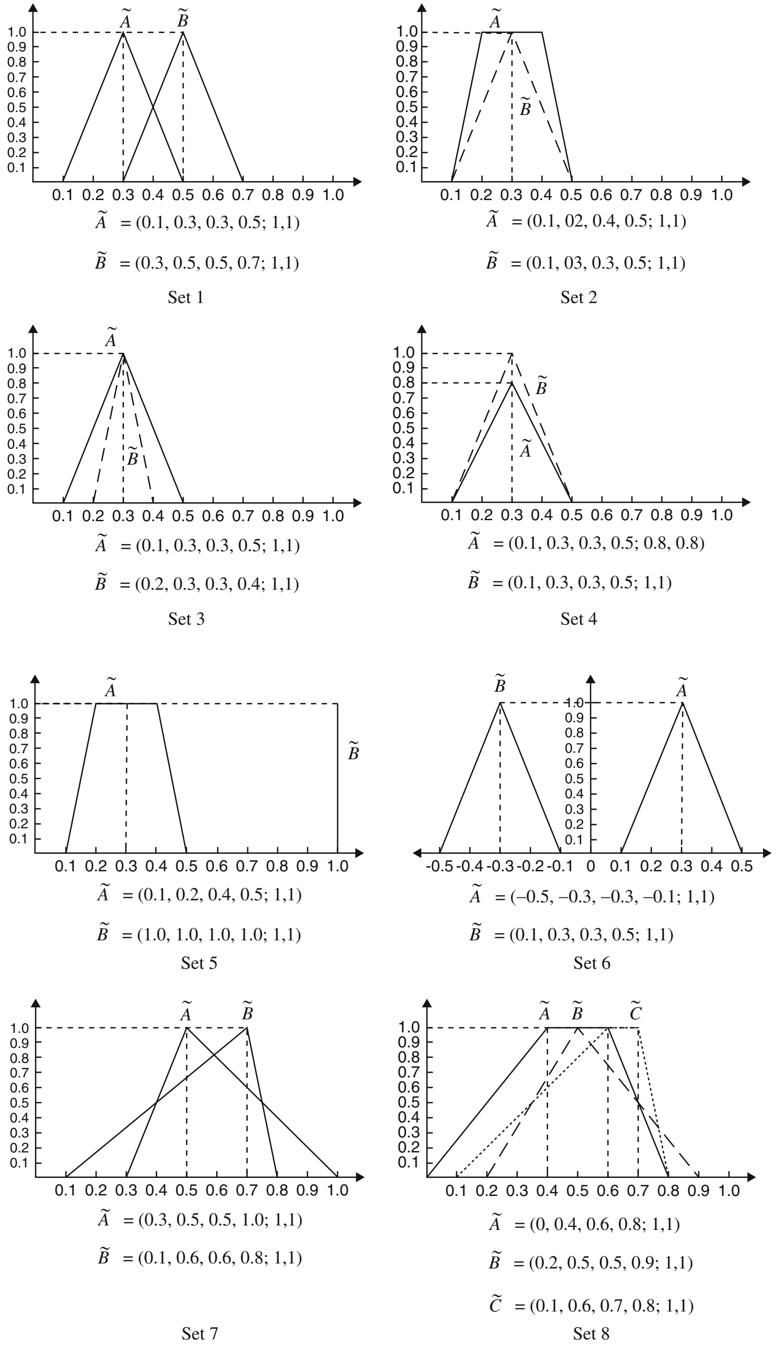
This section focuses on the developing of a new generalized fuzzy number in section “Development of PNIT2GFN.” Then, the new linguistic variable is applied into the interval type-2 entropy weight in MCDM method in section “Decision-Making Based on the Interval Type-2 Entropy Weight.”

Development of PNIT2GFN

This section critically develops a new positive and negative generalized fuzzy number with IT2FS. From this new generalized fuzzy number, a new linguistic variable can be formed for MCDM method (will be discussed in section “Decision-Making Based on the Interval Type-2 Entropy Weight”).

A linguistic data is a variable whose value is naturally language phase in dealing with too complex situation to be described properly in conventional quantitative expressions [22]. A linguistic variable is a variable whose values are words or sentences in a natural or artificial language [23]. Linguistic variable is the most important component in MCDM method. In today’s highly competitive environment, an effective linguistic variable proofs to be an advantage and also a vital requirement in any entity. In the practical decision-making process, sometimes, because of the time pressure and lack of knowledge or data or the decision-makers (DMs) have limited attention and information processing capacities, the DMs cannot provide their preference with single exact value, a margin error, or some possibility distribution on the possible values, but several possible values [24]. Thus, to overcome all the above difficulties, a new linguistic variable is developed using the generalized fuzzy number considering both positive and negative aspects in IT2FS concepts. The whole of this process is shown as follows:

Figure 2 shows the flow of PNIT2GFN. It starts with Fig. 2a, where Zadeh [23] came out with the idea of fuzzy numbers. Figure 2b shows the generalized fuzzy number and Fig. 2c shows the interval type-2 trapezoidal fuzzy number. The combination between “Fig. 2b and c ” and the idea from “example of Set 6 in Fig. 1” will distribute Fig. 2d. The flow for distribution of PNIT2GFN is shown in Fig. 2.



$\tilde{A}$ = (0.1, 0.3, 0.3, 0.5; 1,1)

$\tilde{B}$ = (0.3, 0.5, 0.5, 0.7; 1,1)

Set 1

$\tilde{A}$ = (0.1, 02, 0.4, 0.5; 1,1)

$\tilde{B}$ = (0.1, 03, 0.3, 0.5; 1,1)

Set 2

$\tilde{A}$ = (0.1, 0.3, 0.3, 0.5; 1,1)

$\tilde{B}$ = (0.2, 0.3, 0.3, 0.4; 1,1)

Set 3

$\tilde{A}$ = (0.1, 0.3, 0.3, 0.5; 0.8, 0.8)

$\tilde{B}$ = (0.1, 0.3, 0.3, 0.5; 1,1)

Set 4

$\tilde{A}$ = (0.1, 0.2, 0.4, 0.5; 1,1)

$\tilde{B}$ = (1.0, 1.0, 1.0, 1.0; 1,1)

Set 5

$\tilde{A}$ = (−0.5, −0.3, −0.3, −0.1; 1,1)

$\tilde{B}$ = (0.1, 0.3, 0.3, 0.5; 1,1)

Set 6

$\tilde{A}$ = (0.3, 0.5, 0.5, 1.0; 1,1)

$\tilde{B}$ = (0.1, 0.6, 0.6, 0.8; 1,1)

Set 7

$\tilde{A}$ = (0, 0.4, 0.6, 0.8; 1,1)

$\tilde{B}$ = (0.2, 0.5, 0.5, 0.9; 1,1)

$\tilde{C}$ = (0.1, 0.6, 0.7, 0.8; 1,1)

Set 8

Fig. 1 Eight sets of generalized fuzzy number [8, 9]

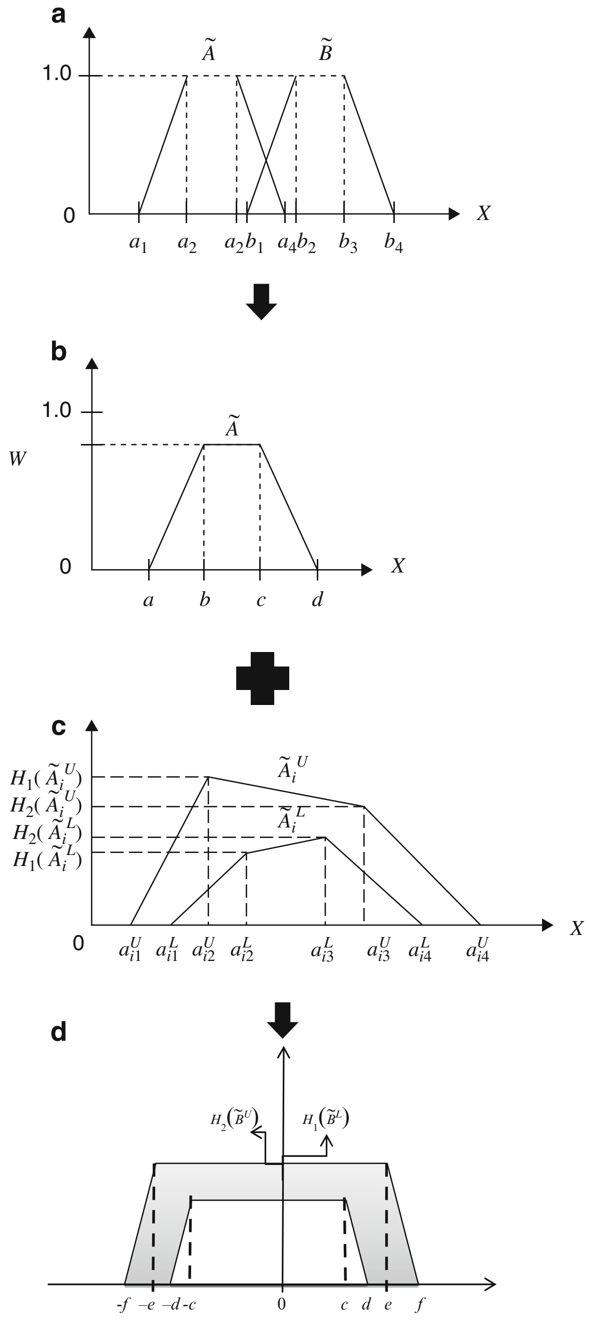


Fig. 2 The flow of positive and negative interval type-2 generalized fuzzy number (PNIT2GFN). (**a**) Two trapezoidal fuzzy numbers [23], (**b**) generalized fuzzy number [5, 25, 26], (**c**) interval type-2 trapezoidal fuzzy number [15], (**d**) PNIT2GFN

The PNIT2GFN in Fig. 2d can be represented by $\tilde{\tilde{B}}$, where $\tilde{\tilde{B}} = \left(\tilde{B}^L, \tilde{B}^U\right) = \left(\left(-d, -c, c, d; H_1\left(\tilde{B}^L\right)\right), \left(-f, -e, e, f; H_2\left(\tilde{B}^U\right)\right)\right)$. The values of $-d, -c, c, d$ and $-f, -e, e, f$ are the reference points of the PNIT2GFN $\tilde{\tilde{B}}$, and then $H_1\left(\tilde{B}^L\right) = w_1\left(\tilde{B}^L\right)$ and $H_2\left(\tilde{B}^U\right) = w_2\left(\tilde{B}^U\right)$ denote the second membership value in the membership function. The membership function $\mu_{\tilde{\tilde{B}}}(x)$ of a PNIT2GFN $\tilde{\tilde{B}}$ is defined as follows:

$$\mu_{\overset{\approx}{B}}(x)^U = \begin{pmatrix} f^{Negative}_{\underset{B}{\sim}}(x) & -d \leq x \leq -c \\ f^{Positive}_{\underset{B}{\sim}}(x) & c \leq x \leq d \\ 0 & otherwise \end{pmatrix} \tag{7}$$

$$\mu_{\overset{\approx}{B}}(x)^U = H_2\left(\tilde{B}^U\right) \quad \forall H_2\left(\tilde{B}^U\right) \in J_x \subseteq [0, 1] \tag{8}$$

and

$$\mu_{\overset{\approx}{B}}(x)^L = \begin{cases} f^{Negative}_{\tilde{B}}(x) & -f \leq x \leq -e \\ f^{Positive}_{\tilde{B}}(x) & e \leq x \leq f \\ 0 & otherwise \end{cases} \tag{9}$$

$$\mu_{\overset{\approx}{B}}(x)^L = H_1\left(\tilde{B}^L\right) \quad \forall H_1\left(\tilde{B}^L\right) \in J_x \subseteq [0, 1] \tag{10}$$

where

$$\mu_{\overset{\approx}{B}}(x) = \left[\mu_{\overset{\approx}{B}}(x)^L, \mu_{\overset{\approx}{B}}(x)^U\right]. \tag{11}$$

Then, the arithmetic operations of the PNIT2GFN are presented as follows:

Proposition 1. The addition operation between two PNIT2GFN $\tilde{\tilde{B}}_1 = \left(\tilde{B}_1, \tilde{B}_1\right) = \left(\left(-d_1, -c_1, c_1, d_1; H_1\left(\tilde{B}^L\right)\right), \left(-f_1, -e_1, e_1, f_1; H_2\left(\tilde{B}^U\right)\right)\right)$ and $\tilde{\tilde{B}}_2 = \left(\tilde{B}_2, \tilde{B}_2\right) = \left(\left(-d_2, -c_2, c_2, d_2; H_2\left(\tilde{B}^L\right)\right), \left(-f_2, -e_2, e_2, f_2; H_2\left(\tilde{B}^U\right)\right)\right)$ is defined as follows:

$$\begin{aligned} \tilde{\tilde{B}}_1 \oplus \tilde{\tilde{B}}_2 &= \left(\left(-d_1, -c_1, c_1, d_1; H_1\left(\tilde{B}^L\right)\right), \left(-f_1, -e_1, e_1, f_1; H_2\left(\tilde{B}^U\right)\right)\right) \oplus \\ &\quad \left(\left(-d_2, -c_2, c_2, d_2; H_1\left(\tilde{B}^L\right)\right), \left(-f_2, -e_2, e_2, f_2; H_2\left(\tilde{B}^U\right)\right)\right) \\ &= \begin{pmatrix} \left(-d_1 \oplus -d_2, -c_1 \oplus -c_2, c_1 \oplus c_2, d_1 \oplus d_2\right), \\ \left(-f_1 \oplus -f_2, -e_1 \oplus -e_2, e_1 \oplus e_2, f_1 \oplus f_2\right); \\ \left(\min\left(H_1\left(\tilde{B}^L\right), H_1\left(B^L\right)\right)\right)\left(\min\left(H_2\left(\tilde{B}^U\right), H_2\left(\tilde{B}^U\right)\right)\right) \end{pmatrix}. \end{aligned} \tag{12}$$

Proposition 2. The subtraction operation between two PNIT2GFN $\tilde{\tilde{B}}_1 = \left(\tilde{B}_1, \tilde{B}_1\right) = \left(\left(-d_1, -c_1, c_1, d_1; H_1\left(\tilde{B}^L\right)\right), \left(-f_1, -e_1, e_1, f_1; H_2\left(\tilde{B}^U\right)\right)\right)$ and $\tilde{\tilde{B}}_2 = \left(\tilde{B}_2, \tilde{B}_2\right) = \left(\left(-d_2, -c_2, c_2, d_2; H_2\left(\tilde{B}^L\right)\right), \left(-f_2, -e_2, e_2, f_2; H_2\left(\tilde{B}^U\right)\right)\right)$ is defined as follows:

$$\begin{aligned}\tilde{\tilde{B}}_1 - \tilde{\tilde{B}}_2 &= \begin{pmatrix}\left(-d_1, -c_1, c_1, d_1; H_1\left(\tilde{B}^L\right)\right), \\ \left(-f_1, -e_1, e_1, f_1; H_2\left(\tilde{B}^U\right)\right)\end{pmatrix} - \\ &\quad \begin{pmatrix}\left(-d_2, -c_2, c_2, d_2; H_1\left(\tilde{B}^L\right)\right), \\ \left(-f_2, -e_2, e_2, f_2; H_2\left(\tilde{B}^U\right)\right)\end{pmatrix} \\ &= \begin{pmatrix}\left(\left(-d_1\right) - \left(-d_2\right), \left(-c_1\right) - \left(-c_2\right), c_1 - c_2, d_1 - d_2\right), \\ \left(\left(-f_1\right) - \left(-f_2\right), \left(-e_1\right) - \left(-e_2\right), e_1 - e_2, f_1 - f_2\right); \\ \left(\min\left(H_1\left(\tilde{B}^L\right), H_1\left(B^L\right)\right)\right)\left(\min\left(H_2\left(\tilde{B}^U\right), H_2\left(\tilde{B}^U\right)\right)\right)\end{pmatrix}.\end{aligned} \tag{13}$$

Proposition 3. The multiplication operation between two PNIT2GFN $\tilde{\tilde{B}}_1 = \left(\tilde{B}_1, \tilde{B}_1\right) = \begin{pmatrix}\left(-d_1, -c_1, c_1, d_1; H_1\left(\tilde{B}^L\right)\right), \\ \left(-f_1, -e_1, e_1, f_1; H_2\left(\tilde{B}^U\right)\right)\end{pmatrix}$ and $\tilde{\tilde{B}}_2 = \left(\tilde{B}_2, \tilde{B}_2\right) = \begin{pmatrix}\left(-d_2, -c_2, c_2, d_2; H_2\left(\tilde{B}^L\right)\right), \\ \left(-f_2, -e_2, e_2, f_2; H_2\left(\tilde{B}^U\right)\right)\end{pmatrix}$ is defined as follows:

$$\begin{aligned}\tilde{\tilde{B}}_1 \otimes \tilde{\tilde{B}}_2 &= \begin{pmatrix}\left(-d_1, -c_1, c_1, d_1; H_1\left(\tilde{B}^L\right)\right), \\ \left(-f_1, -e_1, e_1, f_1; H_2\left(\tilde{B}^U\right)\right)\end{pmatrix} \otimes \\ &\quad \begin{pmatrix}\left(-d_2, -c_2, c_2, d_2; H_1\left(\tilde{B}^L\right)\right), \\ \left(-f_2, -e_2, e_2, f_2; H_2\left(\tilde{B}^U\right)\right)\end{pmatrix} \\ &= \begin{pmatrix}\left(-d_1 \otimes -d_2, -c_1 \otimes -c_2, c_1 \otimes c_2, d_1 \otimes d_2\right), \\ \left(-f_1 \otimes -f_2, -e_1 \otimes -e_2, e_1 \otimes e_2, f_1 \otimes f_2\right); \\ \left(\min\left(H_1\left(\tilde{B}^L\right), H_1\left(B^L\right)\right)\right)\left(\min\left(H_2\left(\tilde{B}^U\right), H_2\left(\tilde{B}^U\right)\right)\right)\end{pmatrix}.\end{aligned} \tag{14}$$

Proposition 4. The arithmetic operation between the PNIT2GFN $\tilde{\tilde{B}}_1 = \left(\tilde{B}_1, \tilde{B}_1\right) = \left(\left(-d_1, -c_1, c_1, d_1; H_1\left(\tilde{B}^L\right)\right), \left(-f_1, -e_1, e_1, f_1; H_2\left(\tilde{B}^U\right)\right)\right)$ and the crisp value k is defined as follows:

$$k\tilde{\tilde{B}}_1 = \left(k\tilde{B}_1, k\tilde{B}_1\right) = \begin{pmatrix}\left(-d_1 \times k, -c_1 \times k, c_1 \times k, d_1 \times k; H_1\left(\tilde{B}^L\right)\right), \\ \left(-f_1 \times k, -e_1 \times k, e_1 \times k, f_1 \times k; H_2\left(\tilde{B}^U\right)\right)\end{pmatrix} \tag{15}$$

$$\begin{aligned}\frac{\tilde{\tilde{B}}_1}{k} &= \left(\tilde{B}_1 \times \frac{1}{k}, \tilde{B}_1 \times \frac{1}{k}\right) \\ &= \begin{pmatrix}\left(-d_1 \times \frac{1}{k}, -c_1 \times \frac{1}{k}, c_1 \times \frac{1}{k}, d_1 \times \frac{1}{k}; H_1\left(\tilde{B}^L\right)\right), \\ \left(-f_1 \times \frac{1}{k}, -e_1 \times \frac{1}{k}, e_1 \times \frac{1}{k}, f_1 \times \frac{1}{k}; H_2\left(\tilde{B}^U\right)\right)\end{pmatrix}.\end{aligned} \tag{16}$$

From all the formulas and arithmetic operations of PNIT2GFN, we can define the new linguistic variable for PNIT2GFN as follows:

Table 1 The new linguistic variables for the ratings of attributes

Linguistic variables for the ratings of the vehicles	
Very low (VL)	((−0.2, −0.15, 0.15, 0.2; 1), (−0.2, −0.15, 0.15, 0.2; 0.9))
Low (L)	((−0.4, −0.2, 0.2, 0.4; 1), (−0.4, −0.2, 0.2, 0.4; 0.9))
Medium (M)	((−0.6, −0.4, 0.4, 0.6; 1), (−0.6, −0.4, 0.4, 0.6; 0.9))
High (H)	((−0.8, −0.6, 0.6, 0.8; 1), (−0.8, −0.6, 0.6, 0.8; 0.9))
Very high (VH)	((−1, −0.8, 0.8, 1; 1), (−1, −0.8, 0.8, 1; 0.9))

Here, this study uses five basic linguistic terms as “very low” (VL), “low” (L), “medium” (M), “high” (H), and “very high” (VH). Thus, the new linguistic variable of PNIT2GFN shown in Table 1.

This new linguistic variable is still new. Therefore, we applied this new linguistic variable in interval type-2 entropy weight for MCDM method (shown in section “Decision-Making Based on the Interval Type-2 Entropy Weight”). Then, we test this new linguistic variable by using a numerical example which is Yu et al. [13]. Further calculations will be shown in Section “Numerical Example.”

Decision-Making Based on the Interval Type-2 Entropy Weight

In this section, we focus on handling method for a new linguistic variable of PNIT2GFN in MCDM method. The new linguistic variable is applied in interval type-2 entropy weight in Step 2. In Step 4, the distance measure is lightly modified in line with the new linguistic variable. Therefore, the full six steps of MCDM method are shown as follows:

Step 1: **Establish a decision matrix**

Establish a decision matrix where all the values are dedicated by the experts.

$$D = \begin{array}{c} \\ f_1 \\ f_1 \\ \vdots \\ f_1 \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_j \end{array} \\ \begin{bmatrix} \tilde{\tilde{f}}_{11} & \tilde{\tilde{f}}_{12} & \cdots & \tilde{\tilde{f}}_{1j} \\ \tilde{\tilde{f}}_{21} & \tilde{\tilde{f}}_{22} & \cdots & \tilde{\tilde{f}}_{1j} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\tilde{f}}_{i1} & \tilde{\tilde{f}}_{i2} & \cdots & \tilde{\tilde{f}}_{ij} \end{bmatrix} \end{array} \tag{17}$$

where $f_1, f_2, \ldots, f_i$ represents the alternative and $C_1, C_2, \ldots, C_j$ represents the criteria. Each entry value is considered as IT2FSs values, which is denoted as $\tilde{\tilde{f}}_{ij}$.

Then, $\tilde{\tilde{f}}_{ij} = \tilde{\tilde{B}}_{ij} = \left(\tilde{B}^L_{ij}, \tilde{B}^U_{ij}\right) = \left(\begin{array}{c} \left(-d_{ij}, -c_{ij}, c_{ij}, d_{ij}; H_1\left(\tilde{B}^L_{ij}\right)\right), \\ \left(-f_{ij}, -e_{ij}, e_{ij}, f_{ij}; H_2\left(\tilde{B}^U_{ij}\right)\right) \end{array} \right)$.

Step 2: **Calculate the entropy-based objective weighting**

The information about weight $\tilde{\tilde{w}}_j$ of the criterion C_j $(j = 1, 2, \ldots, n)$ is completely unknown, we establish an interval type-2 entropy weight for determining the criteria weight in line with the new linguistic variable.

In this step, entropy-based objective weighting method is used. In order to determine the objective weights by entropy measures, the decision matrix needs to be normalized for each criterion C_j $(j = 1, 2, \ldots, n)$ to obtain the projection value for each criterion $\tilde{\tilde{p}}_{ij}$.

$$\tilde{\tilde{p}}_{ij} = \left(\tilde{p}_{ij}{}^{L}, \tilde{p}_{ij}{}^{U}\right) = \left(\left(\frac{\tilde{s}_{ij}}{\sum_{i,j=1}^{m} \tilde{s}_{ij}}\right)^{L}, \left(\frac{\tilde{s}_{ij}}{\sum_{i,j=1}^{m} \tilde{s}_{ij}}\right)^{U}\right)$$

where

$$\tilde{s}_{ij}^{L} = \frac{\sqrt{\sum_{i=1}^{n} \begin{bmatrix} \left((-d)_{ij}^{near} - (-d_{ij})\right)^2 + \left((-c)_{ij}^{near} - (-c_{ij})\right)^2 \\ +\left(c_{ij}^{near} - c_{ij}\right)^2 + \left(d_{ij}^{near} - d_{ij}\right)^2 \end{bmatrix}}}{\sqrt{\sum_{i=1}^{n} \begin{bmatrix} \left((-d)_{ij}^{far} - (-d_{ij})\right)^2 + \left((-c)_{ij}^{far} - (-c_{ij})\right)^2 \\ +\left(c_{ij}^{far} - c_{ij}\right)^2 + \left(d_{ij}^{far} - d_{ij}\right)^2 \end{bmatrix}}} \tag{18a}$$

and

$$\tilde{s}_{ij}^{U} = \frac{\sqrt{\sum_{i=1}^{n} \begin{bmatrix} \left((-f)_{ij}^{near} - (-f_{ij})\right)^2 + \left((-e)_{ij}^{near} - (-e_{ij})\right)^2 \\ +\left(f_{ij}^{near} - f_{ij}\right)^2 + \left(e_{ij}^{near} - e_{ij}\right)^2 \end{bmatrix}}}{\sqrt{\sum_{i=1}^{n} \begin{bmatrix} \left((-f)_{ij}^{far} - (-f_{ij})\right)^2 + \left((-e)_{ij}^{far} - (-e_{ij})\right)^2 \\ +\left(f_{ij}^{far} - f_{ij}\right)^2 + \left(e_{ij}^{far} - e_{ij}\right)^2 \end{bmatrix}}} \tag{18b}$$

then, $\tilde{\tilde{s}}_{ij} = \left(\tilde{s}_{ij}^{L}, \tilde{s}_{ij}^{U}\right)$.

After normalizing the decision matrix, calculate the entropy values $\tilde{\tilde{E}}_j$ as

$$\tilde{\tilde{E}}_{ij} = \left(\tilde{E}_{ij}^{L}, \tilde{E}_{ij}^{U}\right) = \left(\left(-k\sum_{j=1}^{n} \tilde{p}_{ij} \ln \tilde{p}_{ij}\right)^{L}, \left(-k\sum_{j=1}^{n} \tilde{p}_{ij} \ln \tilde{p}_{ij}\right)^{U}\right). \tag{19}$$

k is a constant; let $k = (\ln(m))^{-1}$.

Then, the degree of divergence $\tilde{\tilde{\mu}}_j$ of the intrinsic information of each criterion C_j $(j = 1, 2, \ldots, n)$ may be calculated as

$$\tilde{\tilde{\mu}}_{ij} = \left(\tilde{\mu}^L_{ij}, \tilde{\mu}^U_{ij}\right) = \left(\left(1 - \tilde{E}_{ij}\right)^L, \left(1 - \tilde{E}_{ij}\right)^U\right). \tag{20}$$

The value $\tilde{\tilde{\mu}}_j$ represents the inherent contrast intensity of C_j. The higher the $\tilde{\tilde{\mu}}_j$ is, the more important the criterion C_j is for the problem. Then, the objective weight for each criterion can be calculated.

$$\tilde{\tilde{w}}_{ij} = \left(\tilde{w}^L_{ij}, \tilde{w}^U_{ij}\right) = \left(\left(\frac{\tilde{\mu}_{ij}}{\sum_{i,j=1}^{n} \tilde{\mu}_{ij}}\right)^L, \left(\left(\frac{\tilde{\mu}_{ij}}{\sum_{i,j=1}^{n} \tilde{\mu}_{ij}}\right)^U\right)\right). \tag{21}$$

Step 3: **Weighted decision matrix**

Construct the weighted decision matrix $\overline{Y}_w$,

$$\overline{Y}_{w_{ij}} = \left(\tilde{\tilde{v}}_{ij}\right)_{m \times n} = \begin{array}{c} \\ f_1 \\ f_2 \\ \vdots \\ f_i \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \cdots & x_j \end{array} \\ \begin{bmatrix} \tilde{\tilde{v}}_{11} & \tilde{\tilde{v}}_{12} & \cdots & \tilde{\tilde{v}}_{1j} \\ \tilde{\tilde{v}}_{21} & \tilde{\tilde{v}}_{22} & \cdots & \tilde{\tilde{v}}_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{\tilde{v}}_{i1} & \tilde{\tilde{v}}_{i2} & \cdots & \tilde{\tilde{v}}_{ij} \end{bmatrix} \end{array}, \tag{22}$$

where $\tilde{\tilde{v}}_{ij} = \left(\tilde{v}^L_{ij}, \tilde{v}^U_{ij}\right)$, $\tilde{v}^L_{ij} = \tilde{w}^L_{ij} \otimes \tilde{B}^L_{ij}$, and $\tilde{v}^U_{ij} = \tilde{w}^U_{ij} \otimes \tilde{B}^U_{ij}$, $1 \le i \le m$, and $1 \le j \le n$.

Step 4: **Calculate the distance measure**

Calculate the new distance measures using the n-dimensional Euclidean distance. The separation of each alternative from the ideal solution is given as

$$D^*_j = \sqrt{\sum_{i,j=1}^{n} \left(\tilde{\tilde{v}}^*_{ij} - \tilde{\tilde{v}}^L_{ij}\right)^2 + \left(\sum_{i,j=1}^{n} H_1\left(\tilde{B}^L_{ij}\right)\right) + \left(\sum_{i,j=1}^{n} H_2\left(\tilde{B}^L_{ij}\right)\right)}$$
$$+ \sqrt{\sum_{i,j=1}^{n} \left(\tilde{\tilde{v}}^*_{ij} - \tilde{\tilde{v}}^U_{ij}\right)^2 + \left(\sum_{i,j=1}^{n} H_1\left(\tilde{B}^U_{ij}\right)\right) + \left(\sum_{i,j=1}^{n} H_2\left(\tilde{B}^U_{ij}\right)\right)}, \; j = 1, \ldots, J. \tag{23a}$$

Similarly, the separation from the negative ideal solution is given as

$$D_j^- = \sqrt{\sum_{i,j=1}^{n} \left(\tilde{\tilde{v}}_{ij}^- - \tilde{\tilde{v}}_{ij}^L\right)^2 + \left(\sum_{i,j=1}^{n} H_1\left(\tilde{B}_{ij}^L\right)\right) + \left(\sum_{i,j=1}^{n} H_2\left(\tilde{B}_{ij}^L\right)\right)}$$
$$+ \sqrt{\sum_{i,j=1}^{n} \left(\tilde{\tilde{v}}_{ij}^- - \tilde{\tilde{v}}_{ij}^U\right)^2 + \left(\sum_{i,j=1}^{n} H_1\left(\tilde{B}_{ij}^U\right)\right) + \left(\sum_{i,j=1}^{n} H_2\left(\tilde{B}_{ij}^U\right)\right)}, \quad j = 1, \ldots, J. \tag{23b}$$

Step 5: **Relative closeness**

Calculate the relative closeness to the ideal solution. The relative closeness of the alternative f_i is defined as

$$P_j = \frac{D_j^-}{D_j^* + D_j^-}, \quad 1 \leq j \leq n. \tag{24}$$

Step 6: **Rank the values**

Sort the values of P_j in a descending sequence, where $1 \leq j \leq n$. The larger the value of P_j, the higher the preference of the alternatives f_i, where $1 \leq i \leq n$.

In this MCDM framework, we introduced a new linguistic variable of PNIT2GFN. This new linguistic variable is applied into the interval type-2 entropy weight to capture the completely unknown information about criteria weights. Then, follow by the modification of the distance measures to capture the new linguistic variable. This new linguistic variable is hoped to be one of the standard scales in solving the decision-making problems.

Numerical Example

In this section, we give a numerical example to test the ability of the proposed method in handling MCDM problems. Example in this section refers to MCDM problem used in Yu et al. [13].

Assume that there are three committee members consist of D$_1$, D$_2$, and D$_{3,}$ to assess the suitability of the potential furniture supplies. A committee is formed to select the best among three furniture suppliers, A1, A2, and A3, based on nine criteria: price of product (C1), cost of transportation (C2), promotion (C3), quality of product (C4), delivery time (C5), store image (C6), origin (C7), ergonomic (C8), and customization (C9). These three committees used the linguistic terms shown in Table 1 to represent the evaluating values of the alternatives with respect to different attributes, respectively. Based on the proposed method, the new linguistic terms shown in Table 1 can be represented in Table 2.

Table 2 Linguistic of decision matrix

Criteria	Alternatives	Decision-makers		
		D_1	D_2	D_3
C_1	A_1	H	VH	H
	A_2	VH	VH	H
	A_3	H	M	M
C_2	A_1	H	H	H
	A_2	H	VH	H
	A_3	H	H	VH
C_3	A_1	H	H	H
	A_2	H	VH	H
	A_3	H	H	H
C_4	A_1	H	VH	H
	A_2	H	H	VH
	A_3	H	H	H
C_5	A_1	H	H	M
	A_2	VH	H	M
	A_3	H	H	H
C_6	A_1	H	M	H
	A_2	VH	H	M
	A_3	H	H	H
C_7	A_1	M	H	H
	A_2	H	H	H
	A_3	M	M	H
C_8	A_1	H	H	VH
	A_2	VH	VH	VH
	A_3	H	H	H
C_9	A_1	M	M	L
	A_2	H	H	H
	A_3	M	H	L

Step 1: **Establish a decision matrix**

Establish a decision matrix where all the values are dedicated by the committee members.

Step 2: **Calculate the entropy-based objective weighting**

Use interval type-2 entropy weight formulas (Eq. (19)) to calculate the entropy value in the decision matrix. Therefore, the entropy value is represented in Table 3.

Then, the maximal entropy value is shown in Table 4:

The weight of attributes is calculated using the weight formula (Eq. (21)). The result is shown in Table 5.

Step 3: **Weighted decision matrix**

Next, Eq. (22) is applied respectively to yield the interval type-2 fuzzy weighted normalize decision matrix.

Step 4: **Calculate the distance measure**

Table 3 Interval type-2 entropy for each criterion

	A_1	A_2	A_3
C_1	1.634889	1.555493	2.005027
C_2	1.732051	1.634889	1.634889
C_3	1.732051	1.634889	1.732051
C_4	1.634889	1.634889	1.732051
C_5	1.852682	1.732051	1.732051
C_6	1.852682	1.732051	1.732051
C_7	1.852682	1.732051	2.005027
C_8	1.634889	1.489803	1.732051
C_9	2.460931	1.732051	2.201398

Table 4 Maximal entropy value

	A_1	A_2	A_3
C_1	0.664338	0.898064	0.910797
C_2	0.703819	0.943904	0.742659
C_3	0.703819	0.943904	0.786796
C_4	0.664338	0.943904	0.786796
C_5	0.752838	1	0.786796
C_6	0.752838	1	0.786796
C_7	0.752838	1	0.910797
C_8	0.664338	0.860138	0.786796
C_9	1	1	1

Table 5 Entropy-based weights for each attribute

	$\tilde{\tilde{a}}_j$	$\tilde{\tilde{T}}_j$	$\tilde{\tilde{w}}_j$
C_1	0.8244	2.473199	0.071001
C_2	0.796794	2.390383	0.08501
C_3	0.811506	2.434519	0.077425
C_4	0.798346	2.395037	0.084197
C_5	0.846545	2.539634	0.060424
C_6	0.846545	2.539634	0.060424
C_7	0.887878	2.663635	0.042093
C_8	0.770424	2.311272	0.099329
C_9	1	3	0

Based on Eqs. (23a) and (23b), we can calculate the new distance measures using the n-dimensional Euclidean distance.

Step 5: **Relative closeness**

Based on Eq. (24), we can calculate the relative closeness to the ideal solution. The relative closeness of the alternative f_i is defined as Table 6.

Step 6: **Rank the values**

We can calculate the relative degree of closeness $C(x_j)$ of each alternative x_j with respect to the positive ideal solution x^+, where $1 \leq j \leq 3$, shown in Table 7.

Table 6 Ideal solutions

	$d^{+}(x_j)$	$d^{-}(x_j)$
A$_1$	1.622526	0.047437
A$_2$	1.776002	0.056858
A$_3$	1.562626	0.043947

Table 7 Final ranking order

	$C(x_j)$
A$_1$	0.027679
A$_2$	0.030229
A$_3$	0.026653

The preferred order of the alternatives A_1, A_2, and A_3 is: $A_2 > A_1 > A_3$. That is, the best alternative among A_1, A_2, and A_3 is A_2. The ranking order of the proposed method is consistent with Yu et al.'s example.

Conclusion

This paper has critically introduced a new generalized fuzzy number or can be known as "PNIT2GFN." This new PNIT2GFN considered both sides which are positive and negative side. It is due to the statement that "every matter has two sides such as negative and positive, bad and good, and etc." [14]. Besides, the concept of second membership function in IT2FS is successfully applied into the PNIT2GFN. The concept of IT2FS is believed to threat the imprecise sources, information ambiguity, and uncertain factors [27, 28]. From this new PNIT2GFN, we developed a new linguistic variable. The linguistic variable is applied into the interval type-2 entropy weight for MCDM method. An aggregation process in MCDM method is modified in line with the new linguistic variable. An illustrative example from Yu et al. [13] was tested in order to check the efficiency of the new method. The efficiency of using new method is proven with a straightforward computation in illustrative examples. This approach is seen to provide a new perspective in type-2 decision-making area. They offer a practical, effective, and simple way to produce a comprehensive judgment. This research can further be extended by using non-symmetrical interval triangular and trapezoidal T2FS.

Acknowledgment This research is supported by Cluster Grant (Dana Penyelidikan Khas), Universiti Sultan Zainal Abidin. This support is gratefully acknowledged.

References

1. Giachetti, R.E., Young, R.E.: A parametric representation of fuzzy numbers and their arithmetic operators. Fuzzy Sets Syst. **91**, 185–202 (1997)
2. Cheng, C.-H.: A new approach for ranking fuzzy numbers by distance method. Fuzzy Sets Syst. **95**, 307–317 (1998)
3. Zadeh, L.A.: Fuzzy sets. Inf. Control. **8**, 338–353 (1965)
4. Beníteza, J.M., Martínb, J.C., Román, C.: Using fuzzy number for measuring quality of service in the hotel industry. Tour. Manag. **28**, 544–555 (2007)
5. Chen, S.H.: Operations on fuzzy numbers with function principal Tamkang. J. Manag. Sci. **6**, 13–25 (1986)
6. Kaur, A., Kumar, A.: A new approach for solving fuzzy transportation problems using generalized trapezoidal fuzzy numbers. Appl. Soft Comput. **12**, 1201–1213 (2012)
7. Cho, S.-J., Lee, B.-S., Lee, G.-M., Kim, D.-S., Song, Y.-O.: Fubini theorem for generalized fuzzy number-valued integrals. Fuzzy Sets Syst. **105**, 177–179 (1999)
8. Chen, S.J., Chen, S.M.: Fuzzy risk analysis based on the ranking of generalized trapezoidal fuzzy numbers. Appl. Intell. **26**(1), 1–11 (2007)
9. Chen, S.M., Chen, J.H.: Fuzzy risk analysis based on ranking generalized fuzzy numbers with different heights and different spreads. Expert Syst. Appl. **36**(3), 6833–6842 (2009)
10. Farhadinia, B., Banb, A.I.: Developing new similarity measures of generalized intuitionistic fuzzy numbers and generalized interval-valued fuzzy numbers from similarity measures of generalized fuzzy numbers. Math. Comput. Model. **57**, 812–825 (2013)
11. Wei, G., Zhao, X., Lin, R., Wang, H.: Generalized triangular fuzzy correlated averaging operator and their application to multiple attribute decision making. Appl. Math. Model. **36**(7), 2975–2982 (2012)
12. Su, Z.-X., Xia, G.-P., Chen, M.-Y., Wang, L.: Induced generalized intuitionistic fuzzy OWA operator for multi-attribute group decision making. Expert Syst. Appl. **39**, 1902–1910 (2012)
13. Yu, V.F., Chi, H.T.X., Dat, L.Q., Phuc, P.N.K., Shen, C.-W.: Ranking generalized fuzzy numbers in fuzzy decision making based on the left and right transfer coefficients and areas. Appl. Math. Model. **37**, 8106–8117 (2013)
14. Imran, C.T., Syibrah, M.N., Mohd Lazim, A.: New condition for conflicting bifuzzy sets based on intuitionistic evaluation. World Acad. Sci. Eng. Technol. **19**, 451–455 (2008)
15. Chen, S.-M., Lee L-W, L.-W.: Fuzzy multiple attributes group decision-making based on the ranking values and the arithmetic operations of interval type-2 fuzzy sets. Expert Syst. Appl. **37**, 824–833 (2010)
16. Chen, S.-M., Lee, L.-W.: Fuzzy multiple attributes group decision-making based on the interval type-2 TOPSIS method. J. Expert Syst. Appl. **37**, 2790–2798 (2010)
17. Mendel, J.M., John, R.I., Liu, F.L.: Interval type-2 fuzzy logical systems made simple. IEEE Trans. Fuzzy Syst. **14**(6), 808–821 (2006)
18. Liu, H., Kong, F.: A new MADM algorithm based on fuzzy subjective and objective integrated weights. Int. J. Inf. Syst. Sci. **1**, 420–427 (2005)
19. Szmidt, E., Kacprzyk, J.: Entropy for intuitionistic fuzzy sets. Fuzzy Sets Syst. **118**(3), 467–477 (2001)
20. De Luca, A., Termini, S.: A definition of non-probabilistic entropy in the setting of fuzzy sets theory. J. Inf. Control **20**, 301–312 (1972)
21. Wang, Y.M., Luo, Y.: Area ranking of fuzzy numbers based on positive and negative ideal points. Comput. Math. Appl. **58**(10), 1769–1779 (2009)
22. Zhang, S.F., Liu, S.Y.: A GRA-based intuitionistic fuzzy multi-criteria group decision making method for personnel selection. Experts Syst. Appl. **38**, 11401–11405 (2011)
23. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning. Inf. Sci. **8**, 199–249 (1975)
24. Peng, D.-H., Gao, G.-Y., Gao, Z.-F.: Generalized hesitant fuzzy synergetic weighted distance measures and their application to multiple criteria decision-making. Appl. Math. Model. **37**(8), 5837–5850 (2013)

25. Chen S.H.: Ranking generalized fuzzy number with graded mean integration. In: Proceedings of the 8th International Fuzzy Systems Association World Congress, Taipei, Taiwan, Republic of China, vol. 2, pp. 899–902, 1999
26. Wei, S.H., Chen, S.M.: Fuzzy risk analysis based on interval-valued fuzzy numbers. Expert Syst. Appl. **36**(2), 2285–2299 (2009)
27. Zamali, T., Abu Osman, M.T., Mohd Lazim, A.: Equilibrium linguistic computation method for fuzzy group decision-making. Malaysian J. Math. Sci. **6**(2), 225–242 (2012)
28. Zamali, T., Lazim, M.A., Abu Osman, M.T.: Sustainable decision-making model for municipal solid-waste management: bifuzzy approach. J. Sustain. Sci. Manag. **7**(1), 56–68 (2012)

The Number of Complex Roots of a Univariate Polynomial Within a Rectangle

Ganhewalage Jayantha Lanel and Charles Ching-An Cheng

Abstract Let $f(z)$ be a nonzero complex univariate polynomial and let R be a rectangle in the complex plane. The number of complex roots of $f(z)$ inside R is given by the winding number of $f(z)$ on R if $f(z)$ has no roots on the boundary of R. In this paper the result is extended to all rectangles R even when $f(z)$ has roots on the boundary of R under the condition that $f(z)$ is square-free. It can also be used to formulate an algorithm that isolates the complex roots of any polynomial.

Keywords Polynomial • Real root isolation • Complex root isolation

Throughout $f(z)$ will be a polynomial with complex coefficients and R a rectangle in the complex plane. The *real part* and *imaginary part* of f are polynomials $f_1, f_2 \in \mathbf{R}[x, y]$ such that $f(z) = f(x + iy) = f_1(x, y) + if_2(x, y)$. We shall identify a complex point P with its coordinates (x, y) in the plane. A complex point P is *axial* if $f_1(P)$ or $f_2(P)$ is zero but not both. Geometrically, this means that $f(P)$ lies on an axis but is not the origin.

Let O be the origin and W_1, W_2 two nonzero points in the complex plane. Then the *pseudo argument change* from W_1 to W_2 is $\frac{\pi}{4}n$ where n is the least number of times the ray $\overrightarrow{OP}$ approaches or leaves a half-axis when the nonzero point P is revolved about O counterclockwise from the quadrant or the half-axis containing W_1 to that containing W_2. We shall denote this value by $\Delta parg(W_1, W_2)$. For example, if W_1, W_2 are in the same quadrant (or on the same half-axis), then $\Delta parg(W_1, W_2) = 0$. If W_1 is on the positive x-axis and W_2 in the second quadrant not on any axis, then $\Delta parg(W_1, W_2) = \frac{3\pi}{4}$ and $\Delta parg(W_2, W_1) = \frac{5\pi}{4}$.

G.J. Lanel (✉)
Department of Mathematics, University of Sri Jayewardenepura, Nugegoda, Sri Lanka
e-mail: ghjlanel@sjp.ac.lk

C.C.-A. Cheng
Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309, USA
e-mail: cheng@oakland.edu

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.03_11

Let N be a *standard* polygon, i.e., a polygon whose edges are either horizontal or vertical. If an edge E of N contains infinitely many axial points, then f_1 or f_2 must be identically zero on E, i.e., $f(E)$ lies on an axis. Let P be an axial point on the boundary ∂N of the polygon N. Then there exist points A, B on the adjacent edge(s) to P such that A, P, B are oriented counterclockwise on ∂N and each of $\overline{AP}$ and $\overline{BP}$ either contains no axial points except P or contains infinite many axial points (thus must consist of axial points only). If $\overline{AP}$ and $\overline{BP}$ do not both contain infinitely many axial points, then P is said to be *critical* and A, B *isolate* P. More explicitly, if P is a non-corner point, then it is critical if and only if there exist points $A, B \in \partial N$ on opposites of P such that $\overline{AB}$ contains no axial point except P. If P is a corner point of R then it is critical if not both $\overline{AP}$ and $\overline{BP}$ consists of axial points.

Define the *pseudo argument change* attributed to a critical point P relative to ∂N by

$$\Delta_{\partial N} parg(P) = \begin{cases} \Delta parg(f(A), f(B)), \text{if } \Delta parg(f(A), f(B)) \leq \pi/2 \\ -\Delta parg(f(B), f(A)), \text{otherwise.} \end{cases}$$

Hence, if one of $f(\overline{AP})$ and $f(\overline{PB})$ is and the other is not on an axis, then $|\Delta_{\partial N} parg(P)| = \pi/4$. If both $f(\overline{AP})$ and $f(\overline{PB})$ are not on an axis, then $|\Delta_{\partial N} parg(P)| = \pi/2$ or 0. Therefore, $\Delta_{\partial N} parg(P) = 0, \pm\pi/4$ or $\pm\pi/2$. It is not hard to see that it is independent of the choice of A and B.

Since the pseudo argument changes attributed to critical points count the "number" of times f crosses an axis, the result below follows directly from the Argument Principle (see [3]).

Proposition 1. *Suppose $f(z) \in \mathbf{C}[z]$, N is a standard polygon and suppose P_i, $i = 1, \ldots, s$, are all the critical points of f on ∂N. If there is no root of f on ∂N, then the number of roots of f in the interior of N is given by*

$$\frac{1}{2\pi} \sum_{i=1}^{s} \Delta_{\partial N} parg(P_i).$$

Proof. Assume $P_1, \ldots, P_s$ are oriented counterclockwise on ∂N. We will choose points $T_1, \ldots, T_s$ on ∂N such that T_i is between P_i and P_{i+1}, subscripts modulo s. If $f(P_i)$ and $f(P_{i+1})$ are on "neighboring" half-axes, i.e., $|\Delta_{\partial N} \arg(f(P_i), f(P_{i+1}))| = \pi/2$, then (P_i, P_{i+1}) are of *type I*, and we use the Intermediate Value Theorem to choose T_i such that $|\Delta_{\partial N} \arg(f(P_i), f(T_i))| = \pi/4$ and $|\Delta_{\partial N} \arg(f(T_i), f(P_{i+1}))| = \pi/4$. Otherwise, $f(P_i)$ and $f(P_{i+1})$ are on the same half-axis, i.e., $\Delta_{\partial N} \arg(f(P_i), f(P_{i+1})) = 0$. Since there are no critical points between P_i and P_{i+1}, $f_{\partial N}(P_i P_{i+1}) = \{f(P) | P \text{ is between } P_i \text{ and } P_{i+1} \text{ on } \partial N\}$ is either entirely on an axis or none of its points except $f(P_i), f(P_{i+1})$ are on the axis. In the first case, we say that (P_i, P_{i+1}) is of *type II* and in the second case of *type III*. In both cases, we choose T_i to be any point on ∂N between P_i and P_{i+1}

and so $\Delta_{\partial N} \arg(f(P_i), f(T_i)) + \Delta_{\partial N} \arg(f(T_i), f(P_{i+1})) = 0$. It is not difficult to see that if (P_{j-1}, P_j) and (P_k, P_{k+1}) are of type I or II and (P_i, P_{i+1}) is of type III for $i = j \ldots k-1$, then

$$\sum_{i=j}^{k} \bigl(\Delta_{\partial N} \arg(f(T_{i-1}), f(P_i)) + \Delta_{\partial N} \arg(f(P_i), f(T_i))\bigr)$$

$$= \sum_{i=j}^{k} \Delta_{\partial N} parg(P_i).$$

where $j \leq k$ and subscripts modulo s.

Therefore, using the principle of argument, the number of roots of f in the interior of N is given by

$$\frac{1}{2\pi} \sum_{i=1}^{s} \bigl(\Delta_{\partial N} \arg(f(T_{i-1}), f(P_i)) + \Delta_{\partial N}\ \arg(f(P_i), f(T_i))\bigr)$$

$$= \frac{1}{2\pi} \sum_{i=1}^{s} \Delta_{\partial N} parg(P_i).$$

Remark. Suppose $f(z) \in \mathbf{C}[z]$ and E is a straight line in the complex plane. Using the rotation of axes, it is easy to see that if $f(z)$ vanishes on infinitely many points of E, then $f(z)$ vanishes at all points of the line. Using this observation, one can remove the hypothesis from Proposition 1 that N must be standard.

Suppose Z is a root of f on ∂R. If Z is on an edge E and not a corner point of the rectangle R, then there exist points A, B of E on opposite sides of Z such that either there are no axial points on $\overline{AB}$ or $f(\overline{AB})$ lies on an axis. If Z is a corner point of R, then there exist points A, B such that each of $\overline{AZ}$ and $\overline{BZ}$ satisfies the following: either all points except Z are axial or none of them are. For convenience we shall say that A, B *isolate* Z and assume that A, Z, B are oriented counterclockwise on ∂R. Define the *pseudo argument change* attributed to Z relative to ∂R by $\Delta_{\partial R} parg(Z) = -\Delta parg(f(B), f(A))$. It is not hard to see that this is independent of the choice of A, B.

Main Theorem. Suppose $f(z) \in \mathbf{C}[z]$ is square-free and suppose R is a rectangle in the complex plane. If Z_i, $i = 1, \ldots, m$, are the roots of f on ∂R and P_i, $i = 1, \ldots, n$, the critical points of f on ∂R, then the number of roots of f in the interior of R is given by

$$\frac{1}{2\pi} \left(\sum_{i=1}^{m} \Delta_{\partial R} parg(Z_i) + \sum_{i=1}^{n} \Delta_{\partial R} parg(P_i) \right).$$

Remarks. 1. The above result can be used to symbolically computing the number of roots of a given polynomial within a rectangle. One first computes the largest square-free divisor f of the given polynomial. Then its real part f_1 and the imaginary part f_2 are computed. After computing $\gcd(f_1, f_2)$ and the cofactors of f_1, f_2, an existing real root isolation procedure can be used on them to find the isolating intervals of the roots and critical points on the boundary of the rectangle (see [2]). Next, bisection is used to make these intervals disjoint. Then the endpoints of the intervals can be used as those that isolate the roots and the critical points. Finally $\Delta_{\partial R} parg(Z_i)$ and $\Delta_{\partial R} parg(P_i)$ can be computed based on the signs of the real and imaginary parts of $f(A_i), f(B_i)$ where $\{A_i, B_i\}$ isolates Z_i or P_i.

2. The hypothesis that f be square-free is needed for the Theorem. For instance, $f(z) = z^2$ is not square-free with real part $f_1(x, y) = x^2 - y^2$ and imaginary part $f_2(x, y) = 2xy$. Let R be the rectangle $[-1, 1] \times [-1, 0]$. Then $(0, 0)$ is the only zero on ∂R and the critical points of f on ∂R are $(-1, 0), (-1, -1), (0, -1), (1, -1), (1, 0)$. It is not difficult to check that $\Delta_R parg(0, 0) = 0$, $\Delta_R parg(-1, 0) = \Delta_R parg(1, 0) = \pi/4$ and $\Delta_R parg(-1, -1) = \Delta_R parg(0, -1) = \Delta_R parg(1, -1) = \pi/2$ so the result of Main Theorem does not hold. In fact it can be shown that $f(z) = z^n$ has $4n$ critical points on the boundary of a small rectangle containing $(0, 0)$ each attributing pseudo argument change of $\pi/2$. So if the Main Theorem is true for the same rectangle R in this case, then the pseudo argument change attributed by $(0, 0)$ would have to be $n\pi$.

Let R and T be standard rectangles such that the interior of T contains a root Z of f on ∂R. We say that T is *small* relative to R if $\partial R \cap \partial T = \{A, B\}$ is contained in the edge(s) of R and A, B isolate Z. The *inside boundary* of T (relative to R) is the set of all points of ∂T which lie inside R or on ∂R.

In order to prove the Main Theorem, we need the following Lemma whose proof will be postponed until the end of the paper.

Lemma 2. *Suppose R is a standard rectangle in the complex plane, $f(z) \in \mathbf{C}[z]$ is square-free and $Z = x_0 + iy_0$ is a root of f on ∂R. Then there exists a rectangle T such that the following holds:*

1. *T is arbitrarily small (in length/width) and is small relative to R such that Z is the only root of f contained in the interior of T.*
2. *Either both f_1, f_2 are strictly monotonic or one of them is in the upper-lower cone and the other in the left-right cone formed by the lines $y - y_0 = \pm(x - x_0)$.*
3. *The pseudo argument change attributed by each critical point on ∂T is $\pi/2$.*

Proof of Main Theorem. By (1) of Lemma 2, there exists disjoint rectangles T_i small relative to R and each Z_i is the only root of f contained in the interior of T_i such that (2)–(3) hold. Let $N = R - \cup_{i=1}^{m} T_i$ and let $Q_{i,j}$, $j = 1, \ldots, \ell_i$ be the critical points on the inside boundary of T_i. Each of these $Q_{i,j}$ is also a critical point on ∂N. Using (1) and (2) of Lemma 2, one sees that there are at most three critical points on the inside boundary of each T_i. By (3) of Lemma 2, $\sum_j \Delta_{\partial N} parg(Q_{i,j})$ is between 0 and $-3\pi/2$. Hence, $\Delta_{\partial R} parg(Z_i) = \sum_j \Delta_{\partial N} parg(Q_{i,j})$. Then the number of roots of f in the interior of R is the same as those in the interior of N which, by Proposition 1, is

$$\frac{1}{2\pi}\left(\sum_{i=1}^{m}\sum_{j=1}^{\ell_i}\Delta_{\partial N} parg(Q_{ij}) + \sum_{i=1}^{n}\Delta_{\partial N} parg(P_i)\right)$$
$$= \frac{1}{2\pi}\left(\sum_{i=1}^{m}\Delta_{\partial R} parg(Z_i) + \sum_{i=1}^{n}\Delta_{\partial R} parg(P_i)\right)$$

since $\Delta_{\partial R} parg(P_i) = \Delta_{\partial N} parg(P_i)$.

Proof of Lemma 2. We first prove that if $Z = x_0 + iy_0$ is a root of f, then (x_0, y_0) is a non-singular point for both f_1 and f_2. Suppose $\partial f_1/\partial x(x_0, y_0) = \partial f_1/\partial y(x_0, y_0) = 0$. Then, by Cauchy-Riemann, $\partial f_2/\partial y(x_0, y_0) = \partial f_1/\partial x(x_0, y_0) = 0$ and $\partial f_2/\partial x(x_0, y_0) = -\partial f_1/\partial y(x_0, y_0) = 0$. Hence, $f'(z_0) = 0$, contradicting the fact that f is square-free. The same for f_2 can be proved similarly.

Without loss of generality we may translate the axes so that $(x_0, y_0) = (0, 0)$, i.e., Z is the origin. Therefore, only two cases need to be considered: case (i) both partials of f_1, f_2 are nonzero at $(0, 0)$; case (ii) one of f_1, f_2 has its x-partial vanish at $(0, 0)$ and y-partial nonzero at $(0, 0)$, and the other has its x-partial nonzero at $(0, 0)$ and y-partial vanish at $(0, 0)$.

In the first case, by the Implicit Function Theorem, there exists $S_1 = [-a, a]\times\mathbf{R}$ and $S_2 = \mathbf{R}\times[-b, b]$ for positive real a, b over which f_1 is a strictly monotonic. We may choose each S_i such that only one root of f, i.e., $(0, 0)$, is in it. Let $T_1 = S_1 \cap S_2$. Then f_1 is strictly monotonic in this arbitrarily small T_1. Similarly there exists arbitrarily small rectangle T_2 over which f_2 is strictly monotonic. Let $T = T_1 \cap T_2$. Then f_1, f_2 are strictly monotonic over T.

In the second case, we may assume that $\frac{\partial f_1}{\partial x}(0, 0) = 0$, $\frac{\partial f_1}{\partial y}(0, 0) \neq 0$, $\frac{\partial f_2}{\partial x}(0, 0) \neq 0$, $\frac{\partial f_2}{\partial y}(0, 0) = 0$. Since $\frac{\partial f_1}{\partial y}(0, 0) \neq 0$ and $\frac{\partial f_2}{\partial x}(0, 0) \neq 0$, by the Implicit Function Theorem, there exist functions $y = \phi_1(x)$ and $x = \phi_2(y)$ defined on $(-h_1, h_1)$ and $(-h_2, h_2)$ that agree with f_1, f_2 around $(0, 0)$. Since $\frac{\partial f_1}{\partial x}(0, 0) = 0$ and $\frac{\partial f_2}{\partial y}(0, 0) = 0$, we have $\phi_1'(0) = 0$ and $\phi_2'(0) = 0$. Hence, there exists arbitrarily small positive δ_1, δ_2 such that $|\phi_1(x)| < |x|$ for $-\delta_1 < x < \delta_1$ and $|\phi_2(y)| < |y|$ for $-\delta_2 < y < \delta_2$. Let $S_1 = [-\delta_1, \delta_1] \times [-\delta_1, \delta_1]$, $S_2 = [-\delta_2, \delta_2] \times [-\delta_2, \delta_2]$ and $T = S_1 \cap S_2$. Then T is arbitrarily small, small relative to R. So one of f_1, f_2 is in the upper-lower cone and the other in the left-right cone created by the lines $y = \pm x$. This proves (1) and (2).

From (1) and (2), one sees that there are exactly four critical points on ∂T. Using Proposition 1, one sees that the sum of pseudo argument changes of the critical points on ∂T must be 2π. Since each pseudo argument change of a critical point is between $-\pi/2$ and $\pi/2$, $\Delta_{\partial T} parg(P) = \pi/2$ for each critical point P on ∂T.

Remark. Isolation of polynomial roots is an important area in Computer Algebra (see [1]). Isolation of complex roots is discussed first in [4] although the algorithm does not solve the problem completely, e.g., in case when all coefficients are integral. Using the Main Theorem, one can formulate a symbolic algorithm for isolating complex roots of a complex polynomial as follows. Using a root bound theorem, find a rectangle in which all roots of the given polynomial lie. Subdivide the rectangle into two sub-rectangles and count the number of roots in the interiors and boundaries of the sub-rectangles. Repeat these on all sub-rectangles created, keeping track of the number of roots in each sub-rectangle and outputting it when it has exactly one root in it. In any case, the hypothesis that the given polynomial must be square-free cannot be dropped. Wilf [5] avoids the problem by stopping the algorithm when a fixed number of subdivisions found roots on the boundary.

References

1. Collins, G.E.: Infallible Calculation of Polynomial Zeros to Specified Precision. Mathematical Software, pp. 35–68. Academic, New York (1977)
2. Johnson, J.R.: Algorithms for polynomial real root isolation. In: Quantifier Elimination and Cylindrical Algebraic Decomposition, pp. 269–299. Springer, New York (1998)
3. Marden, M.: Geometry of Polynomials. American Mathematical Society, Providence (1966)
4. Pinkert, J.: An exact method for finding the roots of a complex polynomial. ACM Trans. Math. Softw. **2**(4), 351–363 (1976)
5. Wilf, H.: A global bisection algorithm for computing the zeros of polynomials in the complex plane. J. ACM **25**(3), 415–420 (1978)

Proof of Fermat's Last Theorem for n = 3 Using Tschirnhaus Transformation

B.B.U. Perera and R.A.D. Piyadasa

Abstract This paper gives a proof on Fermat's last theorem (FLT) for n = 3 by firstly reducing the Fermat's equation to a cubic equation of one variable and then using Tschirnhaus transformation to reduce it to a depressed cubic. By showing that this last equation has nonrational roots, it was concluded that the Fermat's equation cannot have integer solutions.

Keywords Fermat's last theorem • Tschirnhaus transformation • Integer roots

Introduction

Ever since Pierre de Fermat (1637) left an unfinished conjecture that the equation

$$x^n + y^n = z^n \tag{1}$$

(which is the so-called Fermat's last theorem or FLT) cannot have integer solutions for exponent n > 2, there has been many attempts to prove that the statement is true. But until Andrew Wiles gave a 100-page long proof in 1995 which took him 7 years, it was intractable. In fact, once it was in the Guinness Book of World Records for the most difficult mathematical problems. Recently, there were many shorter proofs for the theorem (see [1] and [2]) and for the case n = 3 (see [3–6]), in particular.

This paper attempts to prove the FLT for the case $n = 3$ using a different and much direct and easier approach.

B.B.U. Perera (✉) • R.A.D. Piyadasa
Department of Mathematics, University of Kelaniya, Dalugama, Sri Lanka
e-mail: upeksha@kln.ac.lk; piyadasa94@gmail.com

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.34_12

Method

Consider the Fermat's equation for $n = 3$:

$$x^3 + y^3 = z^3, \quad (x, y) = 1 \tag{2}$$

where (x, y) denotes the greatest common divisor of x and y; hence $(x, y) = 1$ meaning that x and y are co-prime.

Following a similar procedure as in [6], consider (2) in the form

$$(z - y)\left[(z - y)^2 + 3zy\right] = x^3$$

and $(z, y) = (3, z) = 1$. Hence we have $(z–y)$ and $\left[(z - y)^2 + 3zy\right]$ are co-prime. Hence $(z–y)$ is a cube, say $z–y = u^3$, where u is a factor of x.

Considering (2) again, in the form

$$(z - x)\left[(z - x)^2 + 3zx\right] = y^3$$

and $(z, x) = (3, z) = 1$. Hence we have $(z–x)$ and $\left[(z - x)^2 + 3zx\right]$ are co-prime. Hence $(z–x)$ also is a cube, say $z–x = h^3$, where h is a factor of y and $(u, h) = 1$.

Thus, we have

$$z - x = h^3, \quad \text{and} \quad z - y = u^3$$

where $(u, h) = 1$.

Thus, $x = z - h^3$, and $y = z - u^3$.

Substituting for x and y in (2) we get

$$z^3 = \left(z - u^3\right)^3 + \left(z - h^3\right)^3$$

or

$$z^3 - 3z^2\left(u^3 + h^3\right) + 3z\left(u^6 + h^6\right) - \left(u^9 + h^9\right) = 0 \tag{3}$$

This is a cubic equation in one variable z. Now if we can show that (3) has no integer roots, then the proof follows.

Using Tschirnhaus transformation [7] we can remove the z^2 term thus reducing (3) to a depressed cubic (or monic trinomial).

Let $z = t + u^3 + h^3$. This transforms (3) into

$$t^3 - 6u^3h^3t - 3u^3h^3\left(u^3 + h^3\right) = 0 \tag{4}$$

This equation is in the form

$$t^3 - 6vwt - v^3 - w^3 = 0 \tag{5}$$

where $v^3 + w^3 = 3u^3h^3\left(u^3 + h^3\right)$ and $vw = 2u^3h^3$. By using the method of Tartaglia and Cardano for finding roots of a cubic equation, roots of (5) can be written as

$$v + w, \quad v\omega + w\omega^2, \quad v\omega^2 + w\omega$$

where ω is the cube root of unity.

Now, v^3, w^3 are the roots of the equation

$$X^2 - 3u^3h^3\left(u^3 + h^3\right)X + 8u^9h^9 = 0 \tag{6}$$

Roots of (6) are

$$X = u^3h^3\left(\frac{3\left(u^3 + h^3\right) \pm \sqrt{9h^6 - 14u^3h^3 + 9u^6}}{2}\right) \tag{7}$$

By observing that the expression inside the square root can be written as

$$9h^6 - 14u^3h^3 + 9u^6 = \left(3h^3 - \frac{7}{3}u^3\right)^2 + \frac{32}{9}u^6 > 0$$

which is never zero or a perfect square for all nonzero u and h, we can see that the two roots v^3, w^3 of (6) are irrational.

Now, suppose that $v + w = k$ is an integer. Then

$$(v + w)^3 = k^3$$
$$v^3 + w^3 = k^3 - 3vw\,(v + w) = k^3 - 6ku^3h^3$$

which is an integer since k, u, h are all integers. This contradicts the previous result that v^3, w^3 are irrational.

Therefore, the only real root of $v + w$ is also irrational so that t and in turn z also are irrational meaning that the FLT for $n = 3$ cannot possibly have integer solutions.

References

1. Piyadasa, R.A.D., Shadini, A.M.D.M., Jayasinghe, W.J.M.L.P.: Simple analytical proofs of three Fermat's theorems. Can. J. Comput. Math. Nat. Sci. Eng. Med. 2(3) (2011) [Online]. http://www.ampublisher.com/Mar%202011/CMNSEM-1103-014.pdf. Accessed 19 Sept 2013

2. Tschirnhaus, E.W.: A method for removing all intermediate terms from a given equation. ACM SIGSAM Bull. 37(1) (2003) (Translated by R. F. Green) [Online]. http://www.sigsam.org/bulletin/articles/143/tschirnhaus.pdf. Accessed 19 Sept 2013
3. Piyadasa, R.A.D.: Mean value theorem and Fermat's last theorem for n = 3. In: Proceedings of the Annual Research Symposium 2007, Faculty of Graduate Studies, University of Kelaniya [Online]. http://www.kln.ac.lk/uokr/ARS2007/4.6.pdf. Accessed 15 Nov 2013
4. Piyadasa, R.A.D.: Simple and analytical proof of Fermat's last theorem for n = 3. In: Proceedings of the Annual Research Symposium 2008, Faculty of Graduate Studies, University of Kelaniya [Online]. http://www.kln.ac.lk/uokr/ARS2008/4.20.pdf. Accessed 15 Nov 2013]
5. Piyadasa, R.A.D.: Method of infinite descent and proof of Fermat's last theorem for n = 3. J. Comput. Math. Nat. Sci. Eng. Med. 1(6) (2010). http://www.ampublisher.com/September%202010/CMNSEM-1009-012.pdf. Accessed 19 Sept 2013
6. Piyadasa, R.A.D.: A simple and short analytical proof of Fermat's last theorem. Can. J. Comput. Math. Nat. Sci. Eng. Med. 2(3) (2011) [Online]. http://www.ampublisher.com/Mar%202011/CMNSEM-1103-015.pdf. Accessed 19 Sept 2013
7. Weisstein, E.W.: Vieta's substitution. From MathWorld—a Wolfram web resource [Online]. http://mathworld.wolfram.com/VietasSubstitution.html. Accessed 19 Sept 2013

Part III
Computational Geometry

Geometrical Problems Related to Crystals, Fullerenes, and Nanoparticle Structure

Mikhail M. Bouniaev, Nikolai P. Dolbilin, Oleg R. Musin, and Alexey S. Tarasov

Abstract This paper focuses on three groups of geometrical problems, closely *related* to material sciences in general and particularly to crystal/quasicrystal structures along with their formations and fullerenes. Some new results in mathematics are presented and discussed, for example, in section one, new estimates of minimum radius of local identity that guarantee that a Delone set is a point regular set. New results related to locally rigid packings are discussed in section two. One of the goals of the paper is to establish some internal (mathematically) and external (applications to material science) connections between research agendas of various studies in geometry and material sciences.

Keywords Component • Crystalline structures • Delone set • Tammes problem • Packings • Irreducible graph • Contact graph fullerenes

Introduction

This paper focuses on three groups of geometrical problems related to material sciences in general and particularly to crystal/quasicrystal structures and their formations and fullerenes. The first group of problems connects local and global descriptions of geometric structures related to crystals and quasicrystals; the second one focuses on the local structure, namely, on packings and contact graphs; the

M.M. Bouniaev (✉) • O.R. Musin
University of Texas at Brownsville, Brownsville, TX, USA
e-mail: Mikhail.Bouniaev@utb.edu; Oleg.Musin@utb.edu

N.P. Dolbilin
Steklov Mathematical Institute of Russian Academy of Sciences,
Moscow State University, Moscow, Russia
e-mail: Dolbilin@mi.ras.ru

A.S. Tarasov
Kharkevich Institute for Information Transmission Problems of Russian
Academy of Sciences, Moscow, Russia
e-mail: Tarasov.Alexey@gmail.com

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.06_13

last group is related to "proper" placing of a given set of points on n-dimensional sphere. Some new results in mathematics are presented and discussed, for example, in section one, new estimates of minimum radius of local identity that guarantee that a Delone set is a regular point set. New results related to locally rigid packings are discussed in section two, where we focus on packings of congruent N circles on spheres (the Tammes problem) and flat square tori. Based on the concept of irreducible (or locally rigid) contact graphs, we solved the Tammes problem for $N = 13$. Moreover, recently we have found a complete list of all irreducible contact graphs with $N < 12$ vertices. Toroidal packings are interesting for two practical reasons—periodical packings of the plane by circles and the problem of super resolution of images. We classified all locally optimal spherical arrangements up to $N = 11$.

For packings on tori, we have found optimal arrangements for $N = 6$, 7, and 8. Interestingly, for case $N = 7$, there are three different optimal arrangements. Our proofs are based on computer enumerations of spherical and toroidal irreducible contact graphs. One of the goals of the paper is to establish some internal (mathematically) and external (applications to material science) connections between research agendas of various studies in geometry and material sciences.

How Small a "Small domain" Could Be to Describe a Global Structure

It is worth mentioning that any of the problems identified above could be also described as optimization problems in geometry. We'll start with the development of mathematical concepts motivated by the challenge to find the smallest possible domain in a given discrete structure that contains a sufficient condition to provide a global order of the entire structure in crystals or quasicrystals.

Since crystallization is a process resulting from a mutual interaction of just nearby atoms, it was believed (L. Pauling, R. Feynman, et al.) that the long-range order, first of all, the 3D periodicity of atomic structures of crystals, is determined by local rules restricting the arrangement of nearby atoms.

The concept of crystalline structures was first developed in the nineteenth century. Fundamental discoveries of Rene Just Hauy were based on the presentation of a crystalline structure as a periodic lattice. Later on, Evgraf Fedorov introduced a concept of a regular point system with a crystal as the union of such regular point systems. Fedorov intuitively understood and argued that this new definition does not cancel the lattice periodicity but generalizes this idea. The following definitions will explain the concept.

Definition 1.1. Subset X of R^d is called a *Delone set* with parameters r and R (or (r, R)-*set*) where r and R are some positive numbers if (r-condition) any open ball $B^o(r)$ of radius r has at most one point from X and (R-condition) if any closed ball B(R) of radius R has at least one point from X.

Remark. The definition of a Delone set requires the existence of numbers r and R with specified properties. However, for the sake of shortening theorem statements and proofs, we included these two parameters into the definition of a Delone set as a characteristic of the set in the assumption that they exist. It is clear that the choice of either r or R is not unique.

Let us denote by B(x, Q) a closed ball of radius Q(Q > 0) centered at the point x, B^o(x, Q) an open ball, and by X_x(Q) the intersection of X and B(x, Q).

Definition 1.2. Given a Delone set X and a point x from X, the set X_x(Q) is called Q-*neighborhood* of the point x in X.

The following easy-to-prove statements add an additional illustration of the concept of a Delone set.

Statement 1.1. If X is a Delone set with parameters r, R in $\mathbf{R}^d$, then a family of balls {B(x, r), x belongs to X} is a packing of balls in $\mathbf{R}^d$, and a family of balls {B(x, R), x belongs to X} is a covering of **R** by balls.

Statement 1.2. If X is a Delone set in $\mathbf{R}^\mathbf{d}$, then the dimension of the affine hull Aff(X_x(2R)) is d for any point x that belongs to X.

Definition 1.3. Delone set X is called a *regular* point set if for any two points x and y from X, there is a symmetry s of the X such that s(x) = y.

Statement 1.3. Delone set X is a regular point set if and only if there is a crystallographic group G such that X is a G-orbit of some point x; in other words, X = G(x) = {g(x)|g belongs to G}.

Definition 1.4. We say that subset X of $\mathbf{R}^d$ is a *crystal* if X is the G-orbit of some finite set X_0, i.e., X is the union of orbits of several points with respect to the same crystallographic group G.

Definition 1.5. Let Iso(d) be the complete group of all isometries of Euclidean d-space $\mathbf{R}^d$. A subgroup G of the group Iso(d) is called *crystallographic* if:

1. It is discrete (i.e., an orbit of any point of space is discrete).
2. Its fundamental domain is compact.

Remark. From the well-known Schoenflies–Bieberbach theorem (which was the answer to Hilbert's question stated in his XVIII problem), it follows that *any space group contains a translational subgroup with a finite index.* This theorem explains why under Definition 1.4, the periodicity of crystals in all dimensions is not an additional requirement but a direct corollary from the Schoenflies–Bieberbach theorem.

Remark. A mathematical model of an ideal crystal uses two concepts: the Delone set (which is *of local nature*) and the crystallographic group (which is *of global nature*).

The main reason of the local theory is to develop a methodology of how to establish crystallographic symmetry in a crystalline structure from the pairwise identity of local neighborhoods around each atom. Before the 1970s, there were neither formal statements that used mathematical language and concepts nor rigorously proven results in this regard until B. Delone and R. Galiulin formulated the problem and Delone's students N. Dolbilin and M. Stogrin developed a mathematically sound *local theory of crystals* [1–5]. The core of the local theory is a statement that a local identity of a point set or of another structure within certain radius **R** implies a global regularity of the structure. Before we proceed, we would like to give a formal definition of local identity.

Definition 1.6. Given a Delone set, we say that the Q-neighborhood of point x in X is identical to the Q-neighborhood of point x′, if there is a space isometry g such that $g(x) = x'$ and $g(X_x(Q)) = X_{x'}(Q)$.

It is clear that with a given Q, the relation to be identical is an equivalence relation on a set of all Q-neighborhoods in X. Therefore, the set of all Q-neighborhoods in X could be presented as a union of equivalence classes $\Psi_i(Q)$. Let N(Q) stand for the cardinal number of the set of the equivalence classes of Q-neighborhoods.

Definition 1.7. Set X is said to be of *finite type*, if for any $Q > 0$ there are just finitely many classes of Q-neighborhoods, i.e., N(Q) is a finite number.

Statement 1.5. Given a Delone set of finite type, function N(Q) is defined for all $Q > 0$. Moreover, N(Q) is a positive, continuous from the left, integer-valued, monotonically nondecreasing step function.

Statement 1.6. A Delone set X is a crystal if and only if N(Q) is a bounded function, i.e., there is a number N_0 such that $N(Q) < N_0$ for all $Q > 0$. max N(Q) is the number of orbits in crystal X.

Remark. According to the last statement and Definition 1.3 of the regular point set, a regular point set is a crystal such that $N(Q) = 1$ for all positive Q.

The requirement to be a set of finite type is a strong requirement. All crystals (see the definition above) are sets of finite type. However, an inverse statement is false. Moreover, recently, there were found sets of finite type such that for all $Q > Q_0$, each class $\Psi_i(Q)$ occurs in X in infinitely many orientations. This property is present in vertex sets of some mosaics (Pinwheel tilings by J. Conway and C. Radin). The Pinwheel mosaics are not only non-periodic, but each of its finite fragments occurs in infinitely many orientations. This is an unexpected property, for example, contrarily to this property, in well-known non-periodic Penrose patterns, each local fragment occurs only in a finite number of orientations.

Statement 1.7. A Delone set is of finite type if and only if N(2R) is finite.

In order to formulate the *local theorem*, we introduce the following definition.

Definition 1.7. Given Delone set X, a point x, and a neighborhood $X_x(Q)$, the *symmetry group* $S_x(Q)$ of the neighborhood $X_x(Q)$ is a group of all isometries (rotations around the x) which leave $X_x(Q)$ invariant and maps x onto x.

Remark. It should be emphasized that the symmetry group contains not all symmetries of $X_x(Q)$ but only those which leave the center fixed. Thus, there can be such X, Q, x, and x′ where $X_x(Q)$ and $X_{x'}(Q)$ are the same, but symmetry groups $S_x(Q)$ and $S_{x'}(Q)$ may be different. The difference can be a result of the requirement for a rotation to keep the central point either x or x′ fixed. It should also be emphasized that the symmetry group $S_x(Q)$ can only decrease while the radius Q increases.

Theorem 1.1 (Local Theorem [1]). A Delone set X is a regular set if and only if there is a positive number Q such that the two conditions hold:

1. $N(Q+2R)=1$.
2. $S_x(Q)=S_x(Q+2R)$ for any x from X.

Remark. For condition (1), all (Q + 2R)-neighborhoods are pairwise congruent. Therefore, in condition (2), it suffices to require that two groups are equal for at least one point x. For all the rest points x′ from X, the equality will follow from condition (1).

Proof. We are going to prove two lemmas first.

Lemma 1 *Let x and y be two arbitrary points from an* (r,R)*-point set X. There is a finite sequence of points* $x_1=x$, x_2, … , $x_n=y$ *that belong to X, such that* $|x_i-x_{i+1}|<2R$ *for any i,* $0<i<n$.

We are going to call such sequence ($x_1=x$, x_2, … , $x_n=y$ and $|x_ix_{i+1}|<2R$) 2R-*sequence*.

Let us take a segment [xy] and assume that its length |xy| is not less than 2R (otherwise x, y is only one-step 2R-sequence). Let us draw a ball B_1 with the diameter $[x_1y]$. The radius $r(B_1)$ is equal or larger than R. Otherwise, there would be a ball of radius greater than R but free of points that belong to X that contradicts the definition of a Delone set and selection of R. Let x_2 be a point from X that belongs to the ball B_1. Since $[x_1y]$ is a diameter of B_1, for any other point x_2 from B_1, the distance $|x_2y|<|x_1y|$ and $|x_1x_2|<2R$. If $|x_2y|<2R$, then x_1, x_2,y is the required sequence.

If $|x_2y|$ is equal or greater than 2R, we draw a new ball B_2 with the diameter $|x_2y|$. For the previous argument applied to the ball B_2, there is at least one more point x_3 such that$|x_3y|<|x_2y|$ and $|x_2x_3|<2R$.

We can proceed this way until we get point x_{n-1} such that $|x_{n-1}y|<2R$. Such point does exist because the sequence $x_1=x$, x_2, … , x_{n-1} gets closer to y ($|x_{i-1}x_i|<|x_ix_{i+1}|$). Therefore, the defined 2R-sequence has to be finite because in the ball $B_y(|xy|)$, centered at point y with radius |xy|, there is only finite subset of the Delone set. Lemma is proved.

Lemma 2 *Let a Delone set X fulfill both conditions of the local theorem. Assume that an isometry g exists, such that* $g(x)=g(x')$ *and* $g(X_x(Q))=X_{x'}(Q)$. *Then* $g(X_x(Q+2R))=X_{x'}(Q+2R)$.

For condition (1), all $(Q + 2R)$-neighborhoods are congruent. Therefore, there is an isometry f such that $f(X_x(Q + 2R)) = X_{x'}(Q + 2R)$.

Let us put $s = f^{-1}$ g, where g performs first, followed by f^{-1}. It is easy to prove that $s(x) = x$ and $s(X_x(Q)) = X_x(Q)$. Therefore, the isometry s is the symmetry of the Q-neighborhood $X_x(Q)$, i.e., the s belongs to the symmetry group $S_x(Q)$.

For condition (2), s is also a symmetry of a greater neighborhood $S_x(Q + 2R)$ (i.e., $s(X_x(Q + 2R)) = X_x(Q + 2R)$). Since $s = f^{-1}$ g, it follows that $g = f$ s.

Since $g(X_x(Q + 2R)) = f(s\ (X_x(Q + 2R))) = f\ (X_x(Q + 2R)) = X_{x'}(Q + 2R)$, $g(X_x(Q + 2R)) = X_{x'}(Q + 2R)$, and the proof is over. □

From these two lemmas, the theorem follows directly.

Let us take two points x and x′ from X, and let g be an isometry which moves the $(Q + 2R)$-neighborhood of x onto $(Q + 2R)$-neighborhood of $x'{:}g(X_x(Q + 2R)) = X_{x'}(Q + 2R)$.

We'll prove that g is a symmetry of the entire set X. Let us take an arbitrary point y from X, and let us connect it to x by 2R-sequence $x_1 = x, x_2, \dots, x_n = y$. By Lemma 1, such sequence exists.

Since $|x_1x_2| < 2R$, the Q-neighborhood $X_{x2}(Q)$ of point x_2 belongs to the $(Q + 2R)$-neighborhoodX_{x1} $(Q + 2R)$ of the x_1.

Therefore, $g(X_{x2}(Q)) = X_{x'2}(Q)$ where $g(x_2) = x'_2$. By Lemma 2, $g(X_{x2}(Q + 2R)) = X_{x'2}(Q + 2R)$. Next point x_3 from 2R-sequence belongs to $X_{x2}(Q + 2R)$. Therefore, its Q-neighborhood $X_{x3}(Q)$ belongs to $X_{x2}(Q + 2R)$. It follows that $g(X_{x3}(Q)) = X_{x'3}(Q)$, where $g(x_3) = x'_3$. By Lemma 2, $g(X_{x3}(Q + 2R)) = X_{x'3}$ $(Q + 2R)$. Going along the 2R-sequence, we will obtain y such that $g(y) = y'$, where y′ is a point from X.

We have proved that the g-image of any point y from X is a point y′ from X. By the same argument, one can prove that any point from X has a g-preimage in X. Therefore, g is symmetry of X. In other words, the symmetry group of X operates on the set X.

Remark. The local theorem is true for all dimensions and for hyperbolic and sphere spaces too.

For dimension $d = 3$ (the case of a real crystal), the following progress has been made.

Lemma 3 (Shtogrin). Let X be a Delone set. If $N(2R) = 1$ (all $2R$-neighborhoods are identical), then the symmetry group of the $2R$-neighborhood contains no n-fold axis with $n > 6$.

The proofs of the theorems below follow from this lemma and the local theorem.

Theorem 1.2 (Shtogrin, Dolbilin). For any (r, R) Delone set X, $N(10R) = 1$ implies that X is a regular point set.

Even in this case of the 3-dimensional space, regardless of this progress, the value 10R seems to be significantly overestimated. Probably, in dimension 3, radius 4R may be sufficient for the regularity of a Delone set. For dimension 2, there exists the following theorem.

Theorem 1.3. For any Delone set X in plane, if $N(4R) = 1$, then X is a regular point set. Moreover, for any number $0 < a < 4R$, there is a Delone set X that is not a regular point set, and $N(4R - a) = 1$ for this set.

Summarizing the current status of the local theory and answering the question, "how small a 'small domain' could be to describe a global structure," we can say that it cannot be smaller than 4R, and we know now that it could be as small as 10R. There are also calculations showing that the identity of 8R-neighborhoods (i.e., $N(8R) = 1$) also implies the regularity.

For dimensions $d > 3$, it follows from the Local Theorem 1.1 that existence of constant $c = c(d, R/r)$ such that $N(cR) = 1$ implies the regularity of point set X in $\mathbf{R}^d$. Intuitively, it is quite clear that dependence of constant cof fraction R/r is not really necessary. For dimension $d = 4$, it has been shown recently (N. Dolbilin, not published) that the constant $c(d, R/r)$ can be dropped from the ratio R/r.

We can identify the following agenda for the future research related to the local theory:

1. To improve (decrease) the existing upper bounds for Q for dimension 3.
2. To drop $c(d, R/r)$ as a parameter in front of ratio R/r in the theorems' statements for dimensions greater than 3 (where d is a dimension of the space and (r, R) Delone set parameters).
3. To draw a line between crystals and quasicrystals in terms of local conditions. This task has become relevant and important due to two discoveries. In the 1970s, R. Penrose found patterns in plane which contain repetitive arbitrary large identical patches. On the other hand, these patches have the fivefold symmetry that can't occur in any structure with crystallographic symmetry. Later on (1982), Shechtman obtained an alloy in his laboratory that had a sound fivefold symmetry. This discovery is an argument for its non-crystallographic structure (Nobel Prize in Chemistry, 2011).
4. A long-term goal is to fill the gap between the mathematical local theory and empirical concepts of self-assembly that occurs during the formation of a natural crystal/quasicrystal on the nanoscale. This goal will require a close cooperation of mathematicians and specialists in crystallography and structural chemistry. In the presentation, we will discuss some points of the local theory and challenges ahead.

Focus on the Local Structure in Connection to Fullerenes and Related Problems in Computational and Discrete Geometry

In this section, we will focus mostly on the local structure of a geometric model of matter with carbon molecules as our inspiration and motivation. In the last few decades, the study of nanoparticles in general and fullerenes and carbon tubes in particular has been accompanied by geometric observations, computations, and utilization of recent and classical results (like the Euler theorem or group theory studies).

With the current research of the subject by chemists and physical chemists and interesting geometry, like the fivefold local symmetry of icosahedra, a noticeable trend is developing to connect and relate results and problems in physical chemistry to the cutting-edge research in discrete and computational geometry.

Though sphere packing, lattices, and polyhedral in general, and Platonic solids in particular, have been studied by geometers for hundreds of years, these studies have recently received a strong impetus because they arise in a number of diverse problems in physics, chemistry, biology, nanotechnology, and a variety of other disciplines. A lot of interesting mathematical results and its applications can be found in [6–9].

We will focus on some specific areas of research in geometry with the goal to establish an association between well-known (mostly still open or partially solved) problems in geometry and nanoscience studies. We will also establish some related to nanoscience terminological equivalencies between various problems in mathematics, like the Tammes problem [10], and well-distributed points on the sphere problem [7].

Though some researchers define fullerene as any molecule composed entirely of carbon in the form of a hollow sphere, ellipsoid, or tube, we will adopt "more geometric" definition [11].

Definition 2.1. A fullerene as a closed cage molecule containing only hexagonal and pentagonal faces.

It follows from Euler's theorem that a fullerene must contain exactly 12 pentagonal faces but can differ in the number of hexagonal faces. Theoretically, fullerene C_{20} is feasible with only 12 pentagonal faces and no hexagonal faces. In 1970, Osawa [12] suggested that an icosahedral C_{60} molecule might be chemically stable. Experimental work of Kroto, Smalley, and coworkers in mid-1980s established the stability of C_{60} molecule in the gas phase [13].

Though theoretically C_{60} is the smallest stable carbon molecule, the number of vertices (carbon atoms) in the fullerene can be very large, C_{70}, C_{80}, C_{90}, and others. Thus, C_{70} molecule contains 37 faces, 25 hexagons, and 12 pentagons with a carbon atom at the vertices of each polygon and a bond along each polygon edge (Fig. 1).

Number N_c carbon atoms in the icosahedral fullerene can be found by the formula

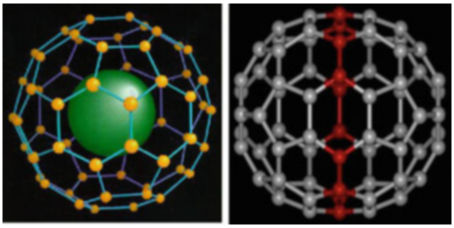


Fig. 1 C_{60} and C_{70}

$$N_c = 20\left(n_2 + nm + m^2\right) \tag{1}$$

where n and m are integers that specify 20 equilateral triangles that an icosahedron consists of. When $(n, m) = (1, 0)$, we get 20 carbon atoms and 12 pentagonal faces. We obtain C_{60}, when $(n, m) = (1, 1)$.

The diameter of a corresponding icosahedron is given by

$$d_i = 5 * \mathrm{sqrt}(3) * a_{c-c}\left(n_2 + nm + m^2\right) \tag{2}$$

where $a_{c–c}$ is the nearest neighbor carbon–carbon distance on the fullerene.

Using formula (1), we can claim that theoretically huge molecules as C_{80}, C_{140}, $C_{180}, \ldots, C_{740}, \ldots$ exist.

Experiments demonstrate that some atoms can be placed inside fullerenes (even inside C_{60}). We also know that C_{60} via nucleophilic addition reactions may undergo multiple attachments of primary amines [7]. Experimental data indicates that there exist a plethora of problems in discrete and computational geometry that could be attributed to two or three types and broadly described as:

1. What (how many/much) might be inside a cage-type molecule or spheres associated with it, for instance, inscribed or subscribed spheres.
2. How to place circular areas (cups) on these spheres to maximize/minimize certain values.
3. What is the configuration of projections of carbon atoms on these spheres. This question is particularly interesting because of the observation that the distance between various pairs of atoms in fullerene (C_{60} is an example) is not exactly the same. What is the relation between an optimal configuration, as defined in [7], and the projections on associated spheres.

The first problem is naturally related to the packing problem in geometry.

Definition 2.2. Sphere packings where all spheres are constrained by their neighbors to stay in one location are called locally rigid (or jammed).

Definition 2.3. Let P be any circle packing of a sphere. Let us connect centers of any two tangent circles in P by edge. Then we obtain a planar graph which is called a contact graph.

Definition 2.4. A contact graph is called irreducible if P is locally rigid (Figs. 2 and 3).

There are several connections between geometric problem (2) and other sphere packing problems. Note that in physics in most cases minimum energy configurations of particles are also locally rigid [4].

Problem (2) is also directly related to the Tammes problem. W. Habicht, K. Schutte, B.L. van der Waerden, and L. Danzer applied irreducible contact graphs for the kissing number and the Tammes problem. Formally, the Tammes problem can be stated as follows:

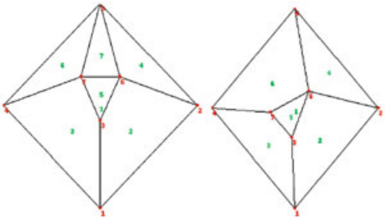


Fig. 2 Irreducible graphs for n = 7

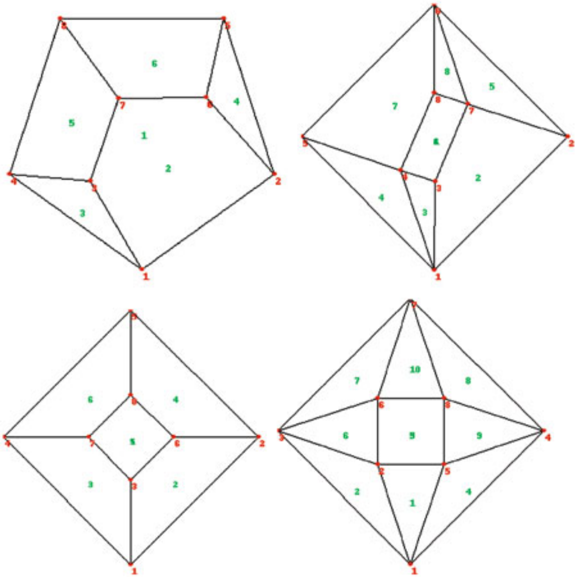


Fig. 3 Irreducible graphs for n = 8

How should N congruent non-overlapping spherical caps be packed on the surface of a unit sphere so that the angular diameter of spherical caps will be as large as possible [10]?

The Tammes problem is presently solved only for several values of N: for N = 3, 4, 6, and 12 by L. Fejes Toth (1943); for N = 5, 7, 8, and 9 by Schutte and van der Waerden (1951); for N = 10 and 11 by Danzer (1963); and for N = 24 by Robinson in 1961.

We have recently solved the Tammes problem for the case N = 13 [14]. The optimal arrangement of 13 circles on the unit sphere was conjectured more than 60 years ago [15]. Our proof is based on a computer enumeration of irreducible contact maximum graphs with 13 vertices. Namely, we applied the following method.

Definition 2.5. Let X be a finite set in the unit sphere in 3-space. (We call such sets *spherical subsets*.) The contact graph CG(X) is the graph with vertices in X and edges (x, y), where x and y are from X and such that dist(x, y) = F(X), where F(X) denotes the minimum angular distance between distinct points in X.

Definition 2.6. Let X be a spherical subset with |X| = N. We say that CG(X) is maximal if F(X) is maxima as possible. We denote this graph by G(N).

It is not hard to prove the following statements:

Lemma 2.1. Let CG(X) be a maximal graph G(N). Then for N > 5, this graph is irreducible.

Lemma 2.2. Let the graph CG(X) be irreducible. Then (1) CG(X) is a planar graph; (2) degrees of its vertices can take only the values 0 (isolated vertex), 3, 4, and 5; (3) all faces of CG(X) are equilateral convex polygons of side length F(x); (4) all faces of CG(X) are polygons with at most 2π/F(x) vertices.

Theorem 2.1. The spherical arrangement of 13 points P(13) is the best possible and the maximal arrangement unique up to isometry (Figs. 4 and 5).

Here we provide a sketch of our computerized proof. For more details, see http://dcs.isa.ru/taras/tammes13/ . The proof consists of two parts: (1) create list L(13) of all graphs with 13 vertices that satisfy Lemma 2.1; (2) using linear approximations and linear programming, remove from the list L(13) all graphs that do not satisfy the geometric properties of G(13) (see [14, Propositions 3.6–3.11].

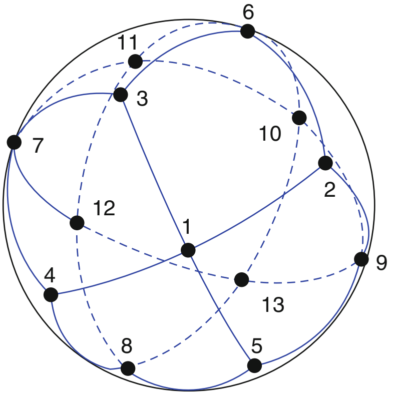


Fig. 4 The spherical arrangement of 13 points P(13)

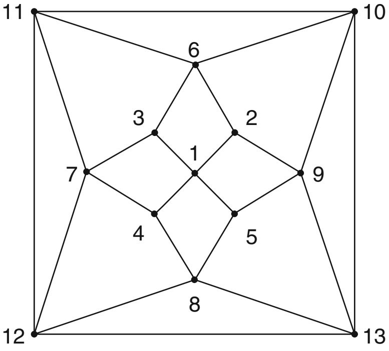


Fig. 5 The contact graph of P(13)

To create list L(13), we use the plantri program [http://cs.anu.edu.au/~bdm/plantri/] developed by Gunnar Brinkmann and Brendan McKay. This program is the isomorphic-free generator of planar graphs, including triangulations, quadrangulations, and convex polytopes.

Using this method, we have recently found a complete list of all irreducible contact graphs with $N < 12$ vertices. Particularly, we created lists L(N) of all planar graphs with $N < 12$ vertices that satisfy Lemma 2.1 and then using linear approximations and linear programming removed from this list all the graphs that do not satisfy the geometric properties of irreducible graphs.

That gave us enumeration of all locally rigid arrangements where around a central point, we have $N < 12$ points.

Problem (3) is related to the "best" configuration of points on sphere with the following explanations below what "the best" means [7].

Definition 2.7. Given an N-point configuration

$$\omega_{\mathrm{N}} = \{\mathrm{X}_1, \ldots, \mathrm{X}_N\}$$

on sphere S^2, we define its generalized energy as

$$\mathrm{E}_{\alpha}\left(\omega_{\mathrm{N}}\right) = \Sigma_{\mathrm{i}\neq\mathrm{j}} \left|\mathrm{X}_{\mathrm{i}}\text{-}\mathrm{X}_{\mathrm{j}}\right|^{\alpha}.$$

The question is how to maximize $\mathrm{E}_{\alpha}(\omega_{\mathrm{N}})$ when $\alpha > 0$ and minimize when $\alpha < 0$ and how to minimize logarithmic energy

$$\mathrm{E}_0\left(\omega_{\mathrm{N}}\right) = \Sigma_{\mathrm{i}\neq\mathrm{j}} \log\left(1/\left|\mathrm{X}_{\mathrm{i}}\text{-}\mathrm{X}_{\mathrm{j}}\right|\right)$$

or maximize the product

$$\mathrm{P}\left(\omega_{\mathrm{N}}\right) = \exp\left(-\mathrm{E}_0\left(\omega_{\mathrm{N}}\right)\right), \quad \text{when } \alpha < 0.$$

Definition 2.8. A collection of points that minimizes the logarithmic energy is called an optimal configuration. The points are referred to as logarithmic points.

If $\alpha = 1$, then with the exception of small N, it is a long-standing open problem in discrete geometry (L. FejesToth—1956); for small negative numbers α, the problem is equivalent to the Tammes problem and could be stated as "how to maximize the minimum distance between any pairs of given numbers of points on sphere." The study of fullerenes sparks an interest in the Thompson problem, the case when $\alpha = -1$ (Fekete Points). The problem for case $\alpha = 0$ (logarithmic points) was formulated by L.L. Whyte in 1952.

In conclusion, we consider periodic planar circle packings with the maximal circle radius, i.e., packings of congruent circles on a square flat torus [16]. This problem is interesting due to another practical reason—the problem of super resolution of images. We have found optimal arrangements for $N = 6$, 7, and 8 circles.

Surprisingly, for the case N = 7, there are three different optimal arrangements. Our method is based on a computer enumeration of toroidal irreducible contact graphs.

Acknowledgments The work has been supported in part by the grant of RF Government N 11.G34.31.0053 (Delaunay Laboratory of Yaroslavl University) and RFBR grant 11-01-00633-a and USA NSF grant DMS-1101688.

References

1. Delone, B.N., Dolbilin, N.P., Stogrin, M.I., Galiulin, R.V.: A local criterion for regularity of a system of points. Sov. Math. Dokl. **17**, 319–322 (1976)
2. Dolbilin, N.P.: Local properties of discrete regular systems. Sov. Math. Dokl. **230**(3), 516–519 (1976)
3. Dolbilin, N., Lagarias, J.C., Senechal, M.: Multiregular point systems. Discret. Comput. Geom. **20**, 477–498 (1998)
4. Dolbilin, N.: Which clusters can form a crystal? In: Volume Voronoi's Impact on Modern Science, Book 2, Kyiv, pp. 96–104, 1998
5. Dolbilin, N., Schatschneider, D.: The local theorem for tilings, Quasicrystals and discrete geometry (Toronto, ON, 1995). Fields Inst. Monogr., **10** (Amer. Math. Soc., Providence, RI, 1998), pp. 193–199
6. Atiyah, M., Sutcliffe, P.: Polyhedra in physics, chemistry and geometry. Milan J. Math. **71**(1), 33–58 (2003)
7. Dragnev, P.D., Saff, E.B.: Riesz spherical potentials with external fields and minimal energy points separation. Potential Anal. **26**, 139–162 (2007)
8. Conway, J.H., Sloane, N.J.A.: Sphere packings, lattices and groups. Springer-Verlag, New York (1988)
9. Gromov, M.: Crystals, proteins, stability and isoperimetry. Bull. Am. Math. Soc. **48**, 229–257 (2011)
10. Tammes, P.M.L.: On the origin of number and arrangement of the places of exit on pollen grains. Dissertation, Groningen (1930)
11. Dresselhaus, M.S., Dresselhaus, G., Eklund, P.C.: Science of Fullerenes and Carbon Nanotubes: Their Properties and Applications. Academic Press, New York (1996)
12. Osawa, E.: Kagaku (Kyoto) **25**, 854 (1970)
13. Kroto, H.W., Heath, J.R., O'Brein, S.C., Curl, R.F., Smalley, R.E.: Nature (London) **318**, 162 (1982)
14. Musin, O.R., Tarasov, A.S.: The strong thirteen spheres problem. Discret. Comput. Geom. **48**, 128–141 (2012)
15. Tóth, F.: Lagerungen in der Ebene, auf der Kugel und im Raum. Springer, Berlin (1953)
16. Musin, O.R., Nikitenko, A.V.: Optimal packings of congruent circles on a square flat torus. Arxiv e-prints, arXiv:1212.0649 (2012)

Tomographic Inversion Using NURBS and MCMC

Zenith Purisha and Samuli Siltanen

Abstract A new approach in tomographic inversion using nonuniform rational B-splines (NURBS) combined with Markov Chain Monte Carlo (MCMC) is discussed. Low dimension of parameters is the benefit in using NURBS, but the resulting inverse problem is nonlinear. MCMC comes forth to tackle this problem. Another advantage is that the result will be directly in CAD software so that it will be convenient for optimizing the shape. A numerical example with simple simulated data, a simple homogeneous simple shape with attenuation one inside the curve and zero outside the curve, is given. The result is compared with filtered back projection and Tikhonov regularization. The potential drawback of the proposed method is heavy computation.

Keywords Tomographic • NURBS • Bayesian inversion • MCMC

Introduction

Tomography is a useful way to study unknown structures of the object. In tomography, the measurement data of the object are collected from various directions. One example is X-ray tomography, based on the absorption of X-rays as they pass through the different parts of an object. Another example is an electron microscopy (EM), which uses a beam of electrons to create an image of a specimen and produce a magnified image.

In some applications of tomography, the dataset is limited [2, 4–6, 12]. One example is in the medical environment, where it is important to avoid a high X-ray dose to the patient. This situation leads to the production of only sparse data. In EM, the specimen cannot be tilted in all directions, leading to a limited-angle problem, and making the reconstruction task is very ill posed (i.e., extremely sensitive to measurement noise and modeling error). Therefore, a process of introducing additional information in order to solve this problem is needed. Therefore, some

Z. Purisha (✉) • S. Siltanen
Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
e-mail: zenith.purisha@helsinki.fi; samuli.siltanen@helsinki.fi

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.17_14

additional information needs to be introduced for making the recovery process reliable and robust against noise. Such information is called *a priori* knowledge. Commonly, a penalty for complexity term emerges on the basis of the information.

Here, we propose a new approach for tomographic reconstruction from sparse data. The unknown parameter is modeled using nonuniform rational B-splines (NURBS) [9]. In 2D, the curve is determined by a parameter vector called the knot vector and some points called control points. There are advantages in using NURBS:

- The number of parameters is small, because we only need to recover the control points, which are very few in number compared to the points of the curve. Having fewer parameters leads to more robust algorithms.
- By working in NURBS, the results are readily in a form used by computer-aided design (CAD) software, because NURBS is the building block of CAD systems, and the computer numerical control (CNC) machines of industrial production devices use NURBS to describe the shapes of objects to be manufactured. Therefore, it is convenient in creating, modifying, analyzing, or optimizing. Most modern factories use CNC machines, and relatively cheap (less than USD 5000) devices are available for small-scale production. Having the result of reconstruction immediately in an industrially producible form can save time in research and development and enable new kinds of production possibilities for start-up companies located anywhere in the world, including developing countries.

The difficulty in the proposed method is that the linear inverse problem of tomography becomes nonlinear (observing only the control points). Nonlinear inverse problem have more complex relationships between the data and model. This is why we use the flexible computational approach called Bayesian inversion [3, 4, 12].

Statistical (Bayesian) inversion provides an effective way to complement insufficient and incomplete measurements with *a priori* knowledge of the unknown. In any particular application there is typically some understanding about the types of objects one is looking for. Markov Chain Monte Carlo (MCMC) is one of the most popular techniques for sampling from probability distributions and is based on the construction of a Markov Chain that has the desired distribution as its equilibrium distribution [10, 13]. With the increasing availability of computer power, Monte Carlo techniques are being increasingly used. Monte Carlo methods are especially useful for simulating systems with many coupled degrees of freedom.

This paper presents the first feasibility study on the new approach. In this preliminary analysis, we use simulated data and a relatively simple problem, but the reconstruction works successfully. This approach is quite promising for solving more complex problems.

NURBS

NURBS is a parametric representation for describing and modeling curves and surfaces which are basically piecewise polynomial functions. NURBS curve and surface is a standard system in CAD because NURBS model is powerful and flexible.

There are 3 important things in NURBS, those are control points, knots, and basis functions.

1. Control points ($\mathbf{p}_i$)
   Control points are a set of points by which the positions can determine the shape of NURBS curves. The curve can be managed easier by connecting the control points by the line sequentially, called *control polygon*. The shape of control polygon will be followed by the curve.
2. Knot vector
   A knot vector is a set of coordinates in the parametric space. In one dimension a knot vector is written

   $$\mathbf{t} = \{t_1, t_2, \ldots, t_{n+p+2}\},$$

   where $t_i \in \mathrm{R}$ is the ith knot, $i = 1, 2, \ldots, n + p + 2$ is the knot index, $n + 1$ is the number of basis function, and p is the degree of polynomial function.
   The knot vector gives information how width the each interval of knot affects the shape of the curve by the changing of control points. Inserting and removing knots are possible to handle the curve in the proper space.
   Basically, there are two types of knot vector:

   - *uniform* if the knots are equally spaced in the parametric space.
   - *nonuniform* knot vectors may have either spaced or multiple internal knot elements.

   If the first and last knot vector elements appear $p+1$ times, then it is called *open* knot vector; otherwise, it is called *periodic* knot vector. The knot vector has to be monotonically increasing, $t_i \leq t_{i+1}$.
   Formally, an open *uniform* knot vector is given by

   $$\begin{aligned} t_i &= 0, \quad 1 \leq i \leq \mathbf{k} \\ t_i &= i - \mathbf{k}, \quad \mathbf{k} + 1 \leq i \leq n + 1 \\ t_i &= n - k + 2, \quad n + 2 \leq i \leq n + \mathbf{k} + 1, \end{aligned}$$

   where $\mathbf{k} = p + 1$.

3. Basis function
   Basis function, $N_{i,p}(t)$, is a function that gives information how strongly the ith control point, $\mathbf{p}_i$, attracts the curve in specific interval, where i is the index of control point.
   The basis functions have the following form:

$$N_{i,0}(t) = \begin{cases} 1 & \text{if } \mathrm{t_i \leq t < t_{i+1}}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} N_{i,p}(t) = \frac{t - t_i}{t_{i+p} - t_i} N_{i,p-1}(t) + \\ \frac{t_{i+p+1} - t}{t_{i+p+1} - t_{i+1}} N_{1+i,p-1}(t). \end{aligned} \tag{1}$$

The general form of NURBS curve can be written as follows:

$$\mathcal{S}(t) = \sum_{i=0}^{n} p_i^h N_{i,p}(t),$$

where the $p_i^h s$ are the four-dimensional homogenous control polygon vertices for the nonrational four-dimensional B-spline curve.

From (1), the four-dimensional space is projected back into three-dimensional space by dividing with the homogeneous coordinate which yields the rational B-spline curve as follows:

$$\begin{aligned} \mathcal{S}(t) &= \frac{\sum_{i=0}^{n} \mathbf{p}_i N_{i,p}(t)\omega_i}{\sum_{i=0}^{n} N_{i,p}(t)\omega_i} \\ &= \sum_{i=0}^{n} \mathbf{p}_i R_{i,p}(t), \end{aligned} \tag{2}$$

where $\mathbf{p}_i$s are the three-dimensional control points for the rational B-spline curve, ω_i are the weights, and

$$R_{i,p}(t) = \frac{\omega_i N_{i,p}(t)}{\sum_{i=0}^{n} \omega_i N_{i,p}(t)}, \tag{3}$$

are the rational B-spline basis function. The $\omega_i \geq 0$ for all i. The weight can be one of the controllers in attracting the curve to the control points. A curve with all weights which set to 1 has the same shape as if all weights set to 10. The shape of different curves will change if the weights of control points are different, while other elements are fixed as in [8, 11].

Basically, the shape of NURBS curve is defined by the knot vectors and location of the control points and the weights. Most designers assume that the knot vectors are fixed and only allow to modify the control points and the weights. In this simulation data, the weights are assumed to be equal. The open uniform knot vector is chosen as a common knot vector used in CAD.

NURBS has several important qualities which make it powerful for modeling. NURBS provides the flexibility to design many variety of shapes (standard analytic shapes and free-form shape of curves and surfaces) by manipulating the control points and the weights. The amount of information or parameters required for a NURBS representation is much smaller than the amount of information required by other common representations. Those conditions bring to the evaluation which is reasonably fast and computationally stable. Invariant under-scaling, rotation, translation, and shear as well as parallel and perspective projection are also interesting and important properties of the NURBS curve.

Tomographic Measurement Model

In this preliminary result we measure the simple shape. We build homogeneous simple bottle shape from the closed NURBS curve.

To avoid inverse crime [7], we produce the synthetic phantom using NURBS with ten parameters and knot vector

$$[0\ 0\ 0\ \frac{1}{10}\ \frac{2}{10}\ \frac{3}{10}\ \frac{4}{10}\ \frac{5}{10}\ \frac{6}{10}\ \frac{7}{10}\ \frac{8}{10}\ \frac{9}{10}\ 1\ 1\ 1],$$

and equal weight for all control points. In the inversion, eight parameters are recovered with different open knot vectors.

From the NURBS curve, we set the X-ray attenuation becomes one inside the curve and zero outside the curve.

Consider the operator $\mathcal{B}$ as Fig. 1 which has the following form:

$$\mathcal{B}(\mathbf{p}) = \begin{cases} 1 \text{ if pixel inside the NURBS curve,} \\ 0 \text{ if pixel outside the NURBS curve.} \end{cases}$$

The vector $\mathbf{f}$ comes from the mapping of the parameter $\mathbf{p}$ applied to operator $\mathcal{B}$. Our proposed shape is as Fig. 2.

The object is measured with pixel size 64 × 64 using parallel-beam geometry; as an example see Fig. 3.

From the source, the wave will be penetrated through the matter and the detector will record the projection images from different directions. All X-ray imaging are based on the absorption of X-ray as they pass through the different parts of the object.

Fig. 1 Mapping from **p** to **f**

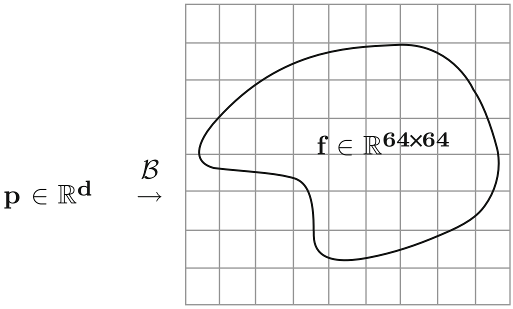
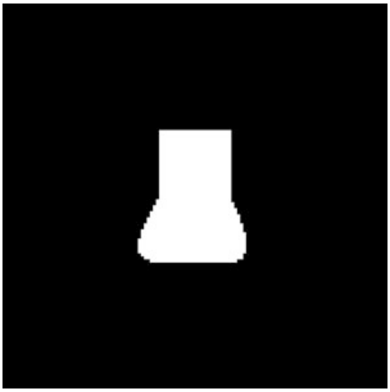

$\mathbf{p} \in \mathbb{R}^{\mathbf{d}}$
$\xrightarrow{\mathcal{B}}$
$\mathbf{f} \in \mathbb{R}^{\mathbf{64 \times 64}}$

Fig. 2 Homogeneous simple shape NURBS



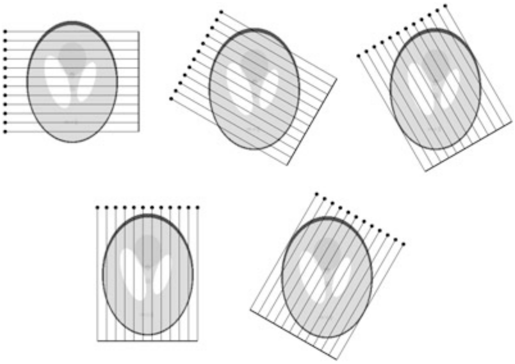


Fig. 3 Parallel-beam X-ray measurement geometry. There are 5 different directions (angles) and 11 lines. *Black dots* show the locations of the X-ray source at different times of measurement. The *thick line* represents the detector measuring the intensity of the X-rays after passing through the target. High attenuation is shown here as *darker shade* of *gray* and low attenuation as lighter shade

In this simulated data, we collected projection of the images of the object from 18 directions (i.e., the measured angles are $0°, 10°, 20°, \ldots, 170°$), using 95 lines for each direction.

Bayesian Inversion

Here, we construct the measurement model as follows and consider the indirect measurement

$$m = A(\mathbf{f}) + \varepsilon,$$

where $m \in \mathbb{R}^k$ is the measurement data, A is an operator of projection image from Radon transform where $\mathbf{f}$ is the quantity of interest, and ε is the error of the measurement.

The inverse problem is to find $\mathbf{f}$ which depends on $\mathbf{p}$, the parameters (control points) of the NURBS curve.

We use the probability theory to model our lack of information in the inverse problem. MCMC method can be used to generate the parameters according to the conditional probability

$$\pi(\mathbf{p} \mid m) = \frac{\pi(\mathbf{p})\pi(m \mid \mathbf{p})}{\pi(m)},$$

called the *posterior distribution* and with the *likelihood function*

$$\pi(m \mid \mathbf{p}) = C \exp(-\frac{1}{2\sigma^2} \|A(\mathcal{B}\mathbf{p}) - m\|_2^2).$$

Assume that the angle of each parameter is not less than $(i - 1)45$ and not more than $(i + 1)45$, where i is the index of parameter and the radius of each parameter from the central point of the object is not less than 0 and not more than 15. This information becomes our *prior* information and it guarantees that the behavior of the parameters will not switch each other during the computation. We formulate the angle condition as follows:

$$\mathbf{A}(\theta_i) = \begin{cases} 1 - \frac{|\theta_i - \theta_i'|}{45} & \text{for } \theta_i' - 45 \leq \theta_i \leq \theta_i' + 45 \\ \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta_i' = 45i$.

The radius terms are as follows:

$$\mathbf{R}(r_i) = \begin{cases} 1 - \frac{|r_i - 1|}{15} & \text{for } 0 \le r_i \le 15 \\ 0, & \text{otherwise.} \end{cases}$$

Hence,

$$\pi(\mathbf{p}) = \mathbf{A}(\theta_i) \cdot \mathbf{R}(r_i)$$

as our *prior* distribution.

To get a useful answer to our inverse problem, we need to draw a representative estimate from the posterior probability distribution. We study the *conditional mean estimate* (CM) defined as the integral

$$\mathbf{p}^{\text{CM}} := \int_{\mathbb{R}^N} \mathbf{p}\pi(\mathbf{p}\,|\,m)\,dp. \tag{4}$$

However, the integration in (4) is over a high-dimensional space, and standard numerical integration quadratures are ineffective. We resort instead to MCMC methods, whose basic idea is to generate a random sequence $p^{(1)}, p^{(2)}, \ldots, p^{(N)}$ of samples with the property that

$$\mathbf{p}^{\text{CM}} \approx \frac{1}{N} \sum_{k=1}^{N} p^{(k)}, \tag{5}$$

and denote

$$\mathbf{p}_N^{\text{CM}} = \frac{1}{N} \sum_{k=1}^{N} p^{(k)}. \tag{6}$$

The sequence $p^{(1)}, p^{(2)}, \ldots, p^{(N)}$ of vectors can, of course, be analyzed more thoroughly than just by taking their average. In recent years, statisticians have been increasingly drawn to MCMC methods to simulate nonstandard multivariate distributions. The Gibbs sampling algorithm is one of the best known of these methods, but a considerable amount of attention is now being devoted to the Metropolis–Hasting algorithm [1, 14]. We use the Metropolis–Hastings algorithm to get the parameter sequences. The algorithm takes the form:

1. Set $n = 1$ and initialize $\mathbf{p}^{(1)}$, where $\mathbf{p}^{(1)}$ depends on $\boldsymbol{\theta}^{(1)}$ and $\mathbf{r}^{(1)}$.
2. Draw a random integer k from 1 to number of control points.
3. Set $\theta := \theta^k + \epsilon_k$ and $r := r^k + \epsilon_k$. Set $p^{(k)} = (r\cos\theta, r\sin\theta)$ then $\mathbf{p}$ will contain the proposed $p^{(k)}$.
4. If $\pi(\mathbf{p}|m) \ge \pi(\mathbf{p}^{(n)}|m)$ then set $\mathbf{p}^{(n+1)} := \mathbf{p}$.

5. If $\pi(\mathbf{p}|m) < \pi(\mathbf{p}^{(n)}|m)$, draw a random number s from uniform distribution on $[0, 1]$.
   If $s \leq \frac{\pi(\mathbf{p}|m)}{\pi(\mathbf{p}^{(n)}|m)}$, then set $\mathbf{p}^{(n+1)} := \mathbf{p}$; else set $\mathbf{p}^{(n+1)} := \mathbf{p}^{(n)}$.
6. If $n = N$ then stop; else set $n := n + 1$ and go to 2^{nd} step.

Applying the CM estimate parameter, $\mathbf{p}_N^{\text{CM}}$, to the NURBS curve (2), we denote by $\mathcal{S}_N^{\text{CM}}$. Then apply it to operator $\mathcal{B}$, we get the reconstruction of the shape,

$$\mathbf{f}_N^{\text{CM}} = \mathcal{B}(\mathbf{p}_N^{\text{CM}}).$$

Let us make a remark concerning the convergence properties of our tomographic reconstruction algorithm. The posterior probability is a compactly supported probability density in $\mathbb{R}^n$, and consequently its mean value is well defined and unique. By the basic theory of Monte Carlo integration, our computation will produce an approximation to the mean value with accuracy increasing when the chain becomes longer. It is another question whether the recovered shape is close to the actual measured object. The closeness of the shapes is related to the number of knots and control points used in the NURBS model and on the smoothness properties of the boundary of the measured object. Precise analysis of the reconstruction error is outside the scope of this paper, which is essentially an initial feasibility study for the general approach.

Computational Result

We present a numerical example to demonstrate our proposed method with eight parameters of interest. The knot vector is fixed in each iteration which becomes the open uniform knot vector degree 2:

$$[0\ 0\ 0\ \frac{1}{7}\ \frac{2}{7}\ \frac{3}{7}\ \frac{4}{7}\ \frac{5}{7}\ \frac{6}{7}\ 1\ 1\ 1].$$

The object is measured by Radon transform using 0.1 % noise. Using Metropolis–Hasting algorithm, we get the average of the parameters's chains for the reconstruction of the curve for some iterations as in Fig. 4.

The iteration stops when $N = 1{,}000{,}000$. The average of the control points, $\mathbf{p}_N^{\text{CM}}$, is shown in Table 1. By applying the control points to the NURBS curve (2), then we get the curve $\mathcal{S}_N^{\text{CM}}$ as in Fig. 5. The effect choosing the open knot vector and the first control point as the last control point yields the kink in this point as in Fig. 4. Finally, using the operator $\mathcal{B}$, we get the shape reconstruction as in Fig. 6.

The reconstruction using filtered back projection and Tikhonov regularization is also given. In filtered back projection, the object is recovered by using `iradon` command in MATLAB as it is shown in Fig. 7. The Tikhonov regularization is discussed in [7]. See Fig. 8. Both of the reconstruction use pixel size 64 × 64.

Computation time for all methods is recorded as in Table 2.

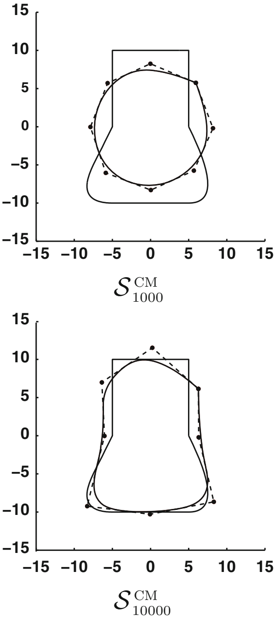
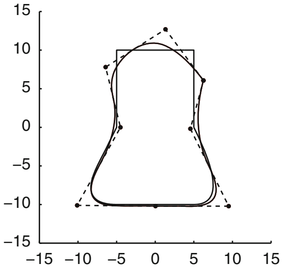


Fig. 4 The *thin black line* is the target curve. The *thick black line* is the reconstruction of the NURBS curve, $\mathcal{S}^{\mathrm{CM}}_{N}$. The *black circle* markers are control points, $\mathbf{p}^{\mathrm{CM}}_{N}$

Table 1 Eight control points, $\mathbf{p}^{\mathrm{CM}}_{1000000}$, are obtained from the NURBS-MCMC reconstruction

x	y
6.24	6.07
1.33	12.7
−6.24	7.80
−4.50	0
−10.1	−10.1
0.03	−10.2
9.50	−10.2
4.55	−0.18

Fig. 5 The *thin black line* is the target curve. The *thick black line* is the reconstruction of the NURBS curve, $\mathcal{S}^{\mathrm{CM}}_{1000000}$. The *black circle* markers are control points, $\mathbf{p}^{\mathrm{CM}}_{1000000}$



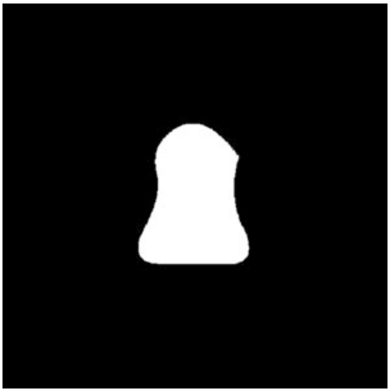


Fig. 6 Final reconstruction using NURBS and MCMC in image size 512 × 512

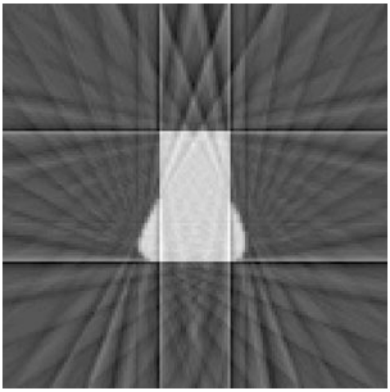


Fig. 7 Reconstruction using filtered back projection with error 0.1 %

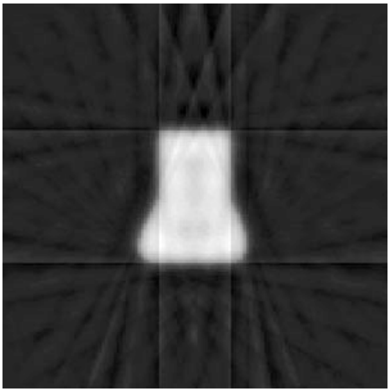


Fig. 8 Reconstruction using Tikhonov Regularization with error 0.1 %

Table 2 Consuming time for all reconstruction methods

FBP	Tikhonov regularization	NURBS-MCMC
1.5 s	905 s	17,000 s

Discussion and Conclusion

In Fig. 6 we see the reconstruction using our proposed method in image size 512 × 512. Changing the resolution in this case does not matter because our reconstruction is in vector graphics form. Only with 16 numbers (Table 1), this NURBS-MCMC reconstruction can recover the data successfully and it is automatically in CAD format or CNC machine.

In filtered back projection and Tikhonov reconstruction, we need 4,096 numbers to give us the information of the object. Also we can see from Figs. 7 and 8 that it is almost impossible to represent the shape of the object because there are so many artifacts appearing. Because of this, these reconstructions cannot represent the result directly in CNC machine. It is very different with NURBS-MCMC reconstruction, which only has exactly two colors, black and white; hence, the shape of the object is obvious.

The proposed method, NURBS-MCMC, is quite promising to be applied in computational tomography inversion. The potential drawback of this method is heavy in computation as we can see in Table 2, but this drawback can be solved by using parallel computing.

Acknowledgements This work was supported by the Academy of Finland through the Finnish Centre of Excellence in Inverse Problems Research 2012–2017, decision number 250215.

References

1. Chib, S., Greenberg, E.: Understanding the Metropolis-Hastings algorithm. Am. Stat. **49**(4), 327–335 (1995)
2. Hämäläinen, K., et al.: Sparse tomography. SIAM J. Sci. Comput. **35**(3), B644–B665 (2013)
3. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer, New York (2004)
4. Kolehmainen, V., Siltanen, S., et al.: Statistical inversion for medical X-ray tomography with few radiographs: II. Application to dental radiology. Phys. Med. Biol. **48**, 1465–1490 (2003)
5. Mohammad-Djafari, A.: A Bayesian approach to shape reconstruction of a compact object from a few number of projections. arXiv preprint physics/0111120 (2001)
6. Mohammad-Djafari, A., Sauer, K.: Shape reconstruction in X-ray tomography from a small number of projections using deformable models. In: Maximum Entropy and Bayesian Methods, pp. 183–198. Springer, Netherlands (1998)
7. Mueller, J., Siltanen, S.: Linear and Nonlinear Inverse Problems with Practical Applications. SIAM, Computational Science and Engineering (2012)
8. Piegl, L., Tiller, W.: The NURBS Book. Monographs in Visual Communication. Springer, Berlin/Heidelberg (1997)

9. Renken, F., and Subbaraya, G., *Nurbs-based solutions to inverse problems in droplet shape prediction*, Computer methods in applied mechanics and engineering, 190 (2000), pp. 1391–1406.
10. Robert, Christian, and George Casella. Monte Carlo statistical methods. Springer Science & Business Media, 2013.
11. Rogers, D. F., *An Introduction to NURBS : with historical perspective*, vol. 1 of Academic Press, Morgan Kaufmann, 2001.
12. Siltanen, Samuli, et al. Statistical inversion for medical x-ray tomography with few radiographs: I. General theory. Physics in medicine and biology 48.10 (2003): 1437.
13. Tierney, Luke. Markov chains for exploring posterior distributions. the Annals of Statistics (1994): 1701-1728.
14. Tierney, Luke. "A note on Metropolis-Hastings kernels for general state spaces." Annals of Applied Probability (1998): 1-9.

Solving Fuzzy Differential Equation Using Fourth-Order Four-Stage Improved Runge–Kutta Method

Faranak Rabiei, Fudziah Ismail, and Saeid Emadi

Abstract In this paper the fuzzy improved Runge–Kutta method of order four for solving first-order fuzzy differential equations is proposed. The scheme is two step in nature and is based on the fourth-order improved Runge–Kutta method for solving first-order ordinary differential equations. The numerical examples are tested to illustrate the efficiency of method.

Keywords Fuzzy improved Runge–Kutta method • Fuzzy differential equations • Two-step methods • Improved Runge–Kutta method

Introduction

Fuzzy differential equations (FDEs) are used for modeling the problems in science and engineering. Most of the problems require the solution of FDEs which satisfied fuzzy initial conditions. The concept of fuzzy derivative was first introduced by Chang and Zadeh [1], and later Dubois and Prade [2] proposed the extension principle for solving FDEs. It is difficult to find the exact solution of FDEs; therefore, several numerical methods were developed to address this problem. Abbasbandy and Allahviranloo [3] developed numerical algorithm for solving FDEs based on Seikkala's work [4]. Ahmad and Hasan [5] presented a new fuzzy version of Euler's method for solving FDEs with fuzzy initial values. In this paper the improved Runge–Kutta method of order four with four stages given by Rabiei et al. in [6] is developed for solving first-order fuzzy initial value problems.

F. Rabiei (✉) • F. Ismail
Department of Mathematics, Universiti Putra Malaysia, Selangor, Malaysia

Institute for Mathematical Research, Universiti Putra Malaysia, Selangor, Malaysia
e-mail: Faranak_rabiei@upm.edu.my; Fudziah@upm.edu.my

S. Emadi
Olympia College, Kuala Lumpur OCKL, Kuala Lumpur, Malaysia
e-mail: S.emadi60@gmail.com

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.18_15

In sections "Preliminaries" and "Fuzzy Initial Value Problems," some basic definitions and theorem on FDEs are given. In the next section, the fuzzy improved Runge–Kutta method of order four with four stages (FIRK4-4) is proposed and numerical examples to illustrate the efficiency of a new method are given in the section "Numerical Example."

Preliminaries

The fuzzy set is a generalization of a classical set that allows membership function to take any value in the unit interval [0, 1]. The formal definition of a fuzzy set is as follows:

Definition 1 (See [1]) Let Ω be a universal set. A fuzzy set A in Ω is defined by a membership function $A(t)$ that maps every element in Ω to the unit interval [0, 1]. A fuzzy set A in Ω may also be presented as a set of ordered pairs of a generic element t and its membership value, as shown in the following equation:

$$A = \left\{ (t, A(t)) \middle| t \in \Omega \right\}$$

Definition 2 (See [1]) Let A be a fuzzy set defined in Ω. The support of A is the crisp set of all elements in Ω such that the membership function of A is nonzero, that is,

$$\sup\ p(A) = \left\{ t \in \Omega \middle| A(t) > 0 \right\}.$$

Definition 3 (See [7]) Let A be a fuzzy set defined in Ω by membership function $A(t)$: $\Omega \to [0, 1]$. Let us denote by $\mathbb{R}_{\mathbb{F}}$ the class of fuzzy subsets of the real axes (i.e., A: $\mathbb{R} \to [0, 1]$) satisfying the following properties:

1. $\forall A \in \mathbb{R}_{\mathbb{F}}$, A is normal, that is, there exists $t_0 \in \mathbb{R}$ such that $A(t_0) = 1$.
2. $\forall A \in \mathbb{R}_{\mathbb{F}}$, A is convex, that is, for all $t, y \in \mathbb{R}$ and $0 \leq \lambda \leq 1$, it holds that

$$A\left(\lambda t + (1-\lambda)\, y\right) \geq \min\left(A(t), A(y)\right).$$

3. $\forall A \in \mathbb{R}_{\mathbb{F}}$, A is upper semicontinuous on $\mathbb{R}$, that is, for any $t_0 \in \mathbb{R}$, it holds that $A(t_0) \geq \lim_{t \to t_0^{\pm}} A(t)$.
4. $[A]^0 = \mathrm{cl}\{t \in \mathbb{R} \mid A(t) \geq 0\}$ is a compact, where $\mathrm{cl}(U)$ denotes the closure of subset U.

Then $\mathbb{R}_{\mathbb{F}}$ is called the space of fuzzy members. Obviously $\mathbb{R} \subset \mathbb{R}_{\mathbb{F}}$.

Definition 4 (See [1]) Let A be a fuzzy set defined in $\mathbb{R}_{\mathbb{F}}$. The r cut of A is the crisp set $[A]^r$ that contains all elements in $\mathbb{R}$ such that the membership values of A is greater than or equal to r, that is,

$$[A]^r = \left\{ t \in \mathbb{R} \middle| A(t) \geq r \right\}, \quad r \in (0,1],$$
$$[A]^0 = \text{cl} \left\{ t \in \mathbb{R} \middle| A(t) > 0 \right\}.$$

Definition 5 (See [7]) Let D: $\mathbb{R}_{\mathbb{F}} \times \mathbb{R}_{\mathbb{F}} \to \mathbb{R}_+ \cup \{0\}$, $D(u, v) = \text{Sup}_{r\in[0\ 1]}$ max $\{|u_1(r) - v_1(r)|, |u_2(r) - v_2(r)|\}$ be the Hausdorff distance between fuzzy numbers, where $[u]_r = [u_1(r), u_2(r)]$, $[v]_r = [v_1(r), v_2(r)]$. The following properties are well known:

$$D\left(u + w, v + w\right) = D\left(u, v\right), \quad \forall u, v, w \ \in \mathbb{R}_{\mathbb{F}},$$
$$D(k.u, k.v) = |k|\, D\left(u, v\right), \quad \forall k \ \in \ \mathbb{R}, \ u, v \ \in \mathbb{R}_{\mathbb{F}},$$
$$D\left(u + v, w + e\right) = D\left(u, w\right) + D\left(v, e\right), \quad \forall u, v, w, e \ \in \mathbb{R}_{\mathbb{F}}.$$

where $(\mathbb{R}_{\mathbb{F}}, D)$ is a complete metric space.

Definition 6 (See [4]) A function f: $\mathbb{R} \to \mathbb{R}_{\mathbb{F}}$ is said to be fuzzy continuous function if f exists for any fixed arbitrary $t_0 \in \mathbb{R}$ and $\varepsilon > 0$, $\delta > 0$ such that $|t - t_0| < \delta \Rightarrow D[f(t), f(t_0)] < \varepsilon$.

Definition 7 (See [7]) Let x, $y \in \mathbb{R}_{\mathbb{F}}$, if there exists $z \in \mathbb{R}_{\mathbb{F}}$ such that $x = y + z$, then z is called H-difference of x, y and it is denoted by $x \ominus y$. (Note that $x \ominus y \neq x + (-1)y = x - y$.

Definition 8 (See [7]) Let f: $(a, b) \to \mathbb{R}_{\mathbb{F}}$ and $t_0 \in (a, b)$. We say that f is H-differentiable (differentiability in the sense of Hukuhara) at t_0, if there exists an element $f'(t_0) \in \mathbb{R}_{\mathbb{F}}$, such that:

1. For all $h > 0$ sufficiently near to zero, $\exists\, f(t_0 + h) \ominus f(t_0)$, $\exists\, f(t_0) \ominus f(t_0 - h)$ and the limits (in the metric D),

$$\lim_{h\to 0+} \frac{f\left(t_0 + h\right) \ominus f\left(t_0\right)}{h} = \lim_{h\to 0+} \frac{f\left(t_0\right) \ominus f\left(t_0 - h\right)}{h} = f'\left(t_0\right)$$

   f is called (1)-differentiable at t_0 or
2. For all $h < 0$ sufficiently near to zero, $\exists\, f(t_0 + h) \ominus f(t_0)$, $\exists\, f(t_0) \ominus f(t_0 - h)$ and the limits,

$$\lim_{h\to 0^-} \frac{f\left(t_0 + h\right) \ominus f\left(t_0\right)}{h} = \lim_{h\to 0^-} \frac{f\left(t_0\right) \ominus f\left(t_0 - h\right)}{h} = f'\left(t_0\right)$$

   f is called (2)-differentiable at t_0.

Theorem (See [7, 8]) Let f: $(a, b) \to \mathbb{R}_{\mathbb{F}}$ be a function denoted by $f(t) = (f_1(t, r), f_2(t, r))$, for each $r \in [0,1]$. Then,

1. If f is (1)-differentiable, then $f_1(t, r)$ and $f_2(t, r)$ are differentiable functions and $f'(t) = \left(f_1'\left(t, r\right), f_2'\left(t, r\right)\right)$.

2. If f is (2)-differentiable, then $f_1(t, r)$ and $f_2(t, r)$ are differentiable functions and $f'(t) = \left(f_2'(t, r), f_1'(t, r)\right)$.

Fuzzy Initial Value Problems

Consider the fuzzy initial value problem

$$y'(x) = f(t, y(t)), \quad y(t_0) = y_0, \quad t \in [t_0 T]$$

where f is a fuzzy function with r-level sets of initial value

$$[y_0]^r = [y_1(0; r), y_2(0; r)], \quad r \in [0, 1].$$

We have $y(t, y) = [y_1(t; r), y_2(t; r)]$ and $f(t, y) = [f_1(t, y), f_2(t, y)]$ where

$$\begin{aligned} f_1(t, y) &= F[t, y_1(t; r), y_2(t; r)], \\ f_2(t, y) &= G[t, y_1(t; r), y_2(t; r)]. \end{aligned}$$

By using the extension principle, when $y(t)$ is fuzzy number we have the membership function

$$f(t, y(t))(s) = \sup\left\{y(t)(\tau) \,\middle|\, s = f(t, \tau)\right\}, \quad s \in \mathbb{R}.$$

It follows that

$$[f(t, y)]^r = [f_1(t, y; r), f_2(t, y; r)], \quad r \in [0, 1],$$

where

$$\begin{aligned} f_1(t, y; r) &= \min\left\{f(t, u) \,\middle|\, u \in [y_1(r), y_2(r)]\right\}, \\ f_2(t, y; r) &= \max\left\{f(t, u) \,\middle|\, u \in [y_1(r), y_2(r)]\right\}. \end{aligned}$$

Throughout this paper we also consider fuzzy function which is continuous in metric space D. Then the continuity of $f(t, y(t); r)$ guarantees the existence of the definition of $f(t, y(t); r)$ for $t \in [t_0, T]$ and $r \in [0, 1]$. Therefore, the functions F and G are defined (see [9]).

Fuzzy Improved Runge–Kutta Method of Order Four with Four Stages

Based on the construction of the improved Runge–Kutta method by Rabiei et al. [6], the improved Runge–Kutta method of order four with four stages (IRK4-4) is given by

$$y_{n+1} = y_n + h\left(b_1 k_1 - b_{-1} k_{-1} + \sum_{i=2}^{4} b_i\left(k_i - k_{-i}\right)\right),$$

for $1 \leq n \leq N-1$, where

$$\begin{aligned} k_1 &= f\left(t_n, y_n\right), \quad k_{-1} = f\left(t_{n-1}, y_{n-1}\right), \\ k_i &= f\left(t_n + c_i h, y_n + h\sum_{j=1}^{i-1} a_{ij} k_j\right) \quad 2 \leq i \leq 4, \\ k_{-i} &= f\left(t_{n-1} + c_i h, y_{n-1} + h\sum_{j=1}^{i-1} a_{ij} k_{-j}\right) \quad 2 \leq i \leq 4. \end{aligned}$$

$c_2, \ldots, c_4 \in [0,1]$ and f depends on both t and y, while k_i and k_{-i} depend on the values of k_j and k_{-j} for $j = 1, \ldots, i-1$. In each step we only need to evaluate the values of $k_1, k_2, \ldots,$ while $k_{-1}, k_{-2}, \ldots$ are calculated from the previous step. Based on IRK4-4 methods we proposed the FIRK method of order four with four stages in which the coefficients of method are given in Table 1. Let the exact solution $[Y(t)]^r = [Y_1(t; r), Y_2(t; r)]$ where are approximated by $[y(t)]^r = [y_1(t; r), y_2(t; r)]$. We define

$[k_i(t, y(t; r))]^r = [k_{i1}(t, y(t; r)), k_{i2}(t, y(t; r))]$, for $i = 1, \ldots, 4$. Note that the values of $k_{-i1}((t_{n-1}, y(t_{n-1}; r))$ and $k_{-i2}(t_{n-1}, y(t_{n-1}; r))$, in each step, are replaced by $k_{i1}(t_n, y(t_n; r))$ and $k_{i2}(t_n, y(t_n; r))$, $i = 1, \ldots, 4$ from the previous step; therefore, there is no need to evaluate them again.

Table 1 Table of coefficients for IRK4-4

$c_1 = 0$				
$c_2 = \frac{1}{5}$	$a_{21} = \frac{1}{5}$			
$c_3 = \frac{3}{5}$	$a_{31} = 0$	$a_{32} = \frac{3}{5}$		
$c_4 = \frac{4}{5}$	$a_{41} = \frac{2}{15}$	$a_{42} = \frac{4}{25}$	$a_{43} = \frac{38}{75}$	
$b_{-1} = \frac{19}{288}$	$b_1 = \frac{307}{288}$	$b_2 = \frac{-25}{144}$	$b_3 = \frac{25}{144}$	$b_4 = \frac{125}{288}$

The fuzzy improved Runge–Kutta method of order four with four stages (FIRK4-4) is given by

$$\begin{aligned} y_1(t_{n+1};r) = y_1(t_n;r) + h\Bigg(& b_1 k_{11}(t_n, y(t_n;r)) - b_{-1}k_{-11}(t_{n-1}, y(t_{n-1};r)) \\ &+ \sum_{i=2}^{4} b_i \{k_{i1}(t_n, y(t_n;r)) - k_{i1}(t_{n-1}, y(t_{n-1};r))\}\Bigg), \end{aligned}$$

$$\begin{aligned} y_2(t_{n+1};r) = y_2(t_n;r) + h\Bigg(& b_1 k_{12}(t_n, y(t_n;r)) - b_{-1}k_{-12}(t_{n-1}, y(t_{n-1};r)) \\ &+ \sum_{i=2}^{4} b_i \{k_{i2}(t_n, y(t_n;r)) - k_{i2}(t_{n-1}, y(t_{n-1};r))\}\Bigg). \end{aligned}$$

where

$$\begin{aligned} k_{11}(t_n, y(t_n;r)) &= \min\left\{f(t_n, u)\,\Big|\, u \in [y_1(t_n;r), y_2(t_n;r)]\right\}, \\ k_{12}(t_n, y(t_n;r)) &= \max\left\{f(t_n, u)\,\Big|\, u \in [y_1(t_n;r), y_2(t_n;r)]\right\}, \\ k_{21}(t_n, y(t_n;r)) &= \min\left\{f(t_n + c_2h, u)\,\Big|\, u \in [z_{11}(t_n, y(t_n;r)), z_{12}(t_n, y(t_n;r))]\right\}, \\ k_{22}(t_n, y(t_n;r)) &= \max\left\{f(t_n + c_2h, u)\,\Big|\, u \in [z_{11}(t_n, y(t_n;r)), z_{12}(t_n, y(t_n;r))]\right\}, \\ k_{31}(t_n, y(t_n;r)) &= \min\left\{f(t_n + c_3h, u)\,\Big|\, u \in [z_{21}(t_n, y(t_n;r)), z_{22}(t_n, y(t_n;r))]\right\}, \\ k_{32}(t_n, y(t_n;r)) &= \max\left\{f(t_n + c_3h, u)\,\Big|\, u \in [z_{21}(t_n, y(t_n;r)), z_{22}(t_n, y(t_n;r))]\right\}, \\ k_{41}(t_n, y(t_n;r)) &= \min\left\{f(t_n + c_4h, u)\,\Big|\, u \in [z_{31}(t_n, y(t_n;r)), z_{32}(t_n, y(t_n;r))]\right\}, \\ k_{42}(t_n, y(t_n;r)) &= \max\left\{f(t_n + c_4h, u)\,\Big|\, u \in [z_{31}(t_n, y(t_n;r)), z_{32}(t_n, y(t_n;r))]\right\}. \end{aligned}$$

and

$$\begin{aligned} z_{11}(t_n, y(t_n;r)) &= y_1(t_n;r) + ha_{21}k_{11}(t_n, y(t_n;r)), \\ z_{12}(t_n, y(t_n;r)) &= y_2(t_n;r) + ha_{21}k_{12}(t_n, y(t_n;r)), \\ z_{21}(t_n, y(t_n;r)) &= y_1(t_n;r) + h\sum_{j=1}^{2} a_{3j}k_{j1}(t_n, y(t_n;r)), \\ z_{22}(t_n, y(t_n;r)) &= y_2(t_n;r) + h\sum_{j=1}^{2} a_{3j}k_{j2}(t_n, y(t_n;r)), \\ z_{31}(t_n, y(t_n;r)) &= y_1(t_n;r) + h\sum_{j=1}^{3} a_{4j}k_{j1}(t_n, y(t_n;r)), \\ z_{32}(t_n, y(t_n;r)) &= y_2(t_n;r) + h\sum_{j=1}^{3} a_{4j}k_{j2}(t_n, y(t_n;r)). \end{aligned}$$

Numerical Example

In this section, we solved the fuzzy initial value problems to show the efficiency and accuracy of the proposed methods. The exact solution $[Y(t)]^r = [Y_1(t;\ r), Y_2(t;\ r)]$ is used to estimate the global error as well as to approximate the starting values of $[y(t_1)]^r = [y_1(t_1;\ r), y_2(t_1;\ r)]$ at the first step.

We define the

$$\text{error}\left(t_i, y\left(t_i; r\right)\right) = \left|y\left(t_i; r\right) - Y\left(t_i; r\right)\right|$$

We tested the following problems and the numerical results of FIRK4-4 are given in Tables 2, 3, 4, and 5 and Figs. 1 and 2.

Problem 1 (See [5])

Table 2 Numerical results of y_1 at $tN = 1, N = 10$ for Problem 1

r	FIRK 4-4	Exact	Error FIRK4-4
0	−0.5000001348	−0.5000000000	1.34×10^{-7}
0.1	−0.4743417769	−0.4743416490	1.27×10^{-7}
0.2	−0.4472137161	−0.4472135955	1.20×10^{-7}
0.3	−0.4183301262	−0.4183300132	1.12×10^{-7}
0.4	−0.38729843895	−0.3872983346	1.04×10^{-7}
0.5	−0.3535534859	−0.3535533906	9.52×10^{-8}
0.6	−0.3162278514	−0.3162277660	8.52×10^{-8}
0.7	−0.2738613526	−0.2738612788	7.38×10^{-8}
0.8	−0.2236068582	−0.2236067978	6.02×10^{-8}
0.9	−0.1581139257	−0.1581138830	4.26×10^{-8}
1.0	0.0	0.0	0.0

Table 3 Numerical results of y_2 at $tN = 1, N = 10$ for Problem 1

r	FIRK 4-4	Exact	Error FIRK4-4
0	0.5000001348	0.5000000000	1.34×10^{-7}
0.1	0.4743417769	0.4743416490	1.27×10^{-7}
0.2	0.4472137161	0.4472135955	1.20×10^{-7}
0.3	0.4183301262	0.4183300132	1.12×10^{-7}
0.4	0.38729843895	0.3872983346	1.04×10^{-7}
0.5	0.3535534859	0.3535533906	9.52×10^{-8}
0.6	0.3162278514	0.3162277660	8.52×10^{-8}
0.7	0.2738613526	0.2738612788	7.38×10^{-8}
0.8	0.2236068582	0.2236067978	6.02×10^{-8}
0.9	0.1581139257	0.1581138830	4.26×10^{-8}
1.0	0.0	0.0	0.0

Table 4 Numerical results of $y_1 = y_3$ with $h = 0.05$, $r = 1$ for Problem 2

t	FIRK4-4	Exact	Error FIRK4-4
0	1,000	1,000	0.0
0.5	863.029510248510	863.029510151307	1.06×10^{-7}
1.0	745.137917180233	745.137917003690	1.83×10^{-7}
1.5	643.667682659973	643.667682428077	2.36×10^{-7}
2.0	556.331442427529	556.331442159126	2.71×10^{-7}
2.5	481.160443818906	481.160443528665	2.91×10^{-7}
3.0	416.460165712376	416.460165411589	3.01×10^{-7}

Table 5 Numerical results of $y_2 = y_4$ with $h = 0.05$, $r = 1$ for Problem 2

t	FIRK4-4	Exact	Error FIRK4-4
0	0.0	0.0	0.0
0.5	138.404191948935	138.404192056939	1.06×10^{-7}
1.0	255.814845307619	255.814845503778	2.03×10^{-7}
1.5	355.181724667715	355.181724925378	2.62×10^{-7}
2.0	439.043335806680	439.043336104905	3.01×10^{-7}
2.5	509.584216431359	509.584216753848	3.24×10^{-7}
3.0	568.684246681833	568.684247016039	3.34×10^{-7}

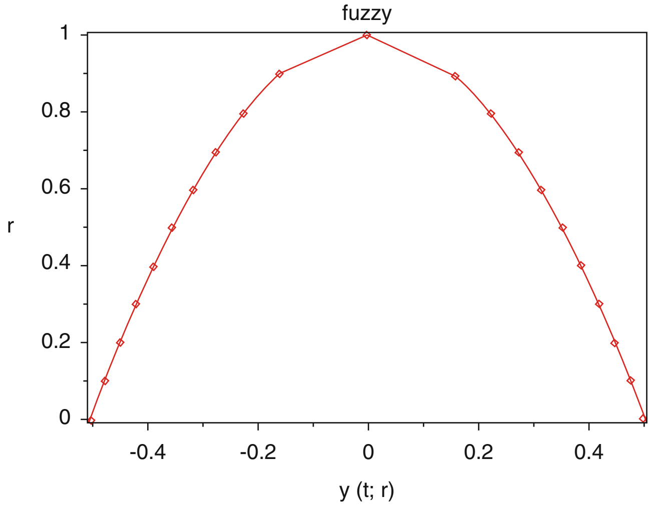


Fig. 1 The approximated solution of $y_1(t)$ and $y_2(t)$ (*solid line*) and exact solution (*points*) with $h = 0.1$, $t \in [0\ 1]$ for Problem 1

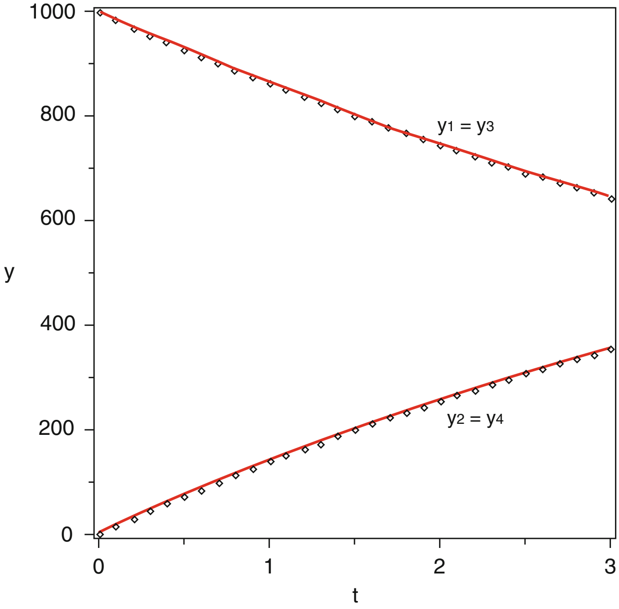


Fig. 2 The approximated solution of $y_1(t)$, $y_2(t)$, $y_3(t)$, and $y_4(t)$ (*solid line*) and exact solution (*points*) with $h = 0.1$, $r = 1$ for Problem 2

$$y'(t) = y(t)\,(1 - 2t)\,, \quad t \geq 0,$$

$$y(0) = \left[-\frac{\sqrt{1-r}}{2}, \tfrac{\sqrt{1-r}}{2}\right].$$

The exact solution is given by $Y\,(t; r) = \left\lfloor -\frac{\sqrt{1-r}}{2} \mathrm{e}^{t-t^2}, \frac{\sqrt{1-r}}{2} \mathrm{e}^{t-t^2} \right\rfloor$.

Problem 2 (Radioactivity Decay Model, see [10])

$$y'(t) = Ay(t) + f,$$

where

$$y'(t)=\begin{bmatrix} y_1'(t)\\ y_2'(t)\\ y_3'(t)\\ y_4'(t)\end{bmatrix},\quad y=\begin{bmatrix} y_1(t)\\ y_2(t)\\ y_3(t)\\ y_4(t)\end{bmatrix},$$

$$A=\begin{bmatrix} 0 & 0 & -0.4+0.1r & 0\\ 0.2+0.1r & 0 & 0 & -0.04+0.01r\\ -0.2-0.1r & 0 & 0 & 0\\ 0 & -0.02-0.01r & 0.4-0.1r & 0\end{bmatrix},$$

$$f=\begin{bmatrix} 4.9+5r\\ 0\\ 5.1-0.1r\\ 0\end{bmatrix},\quad y(0)=\begin{bmatrix} 995+5r\\ 0\\ 1,005-5r\\ 0\end{bmatrix}.$$

The exact solutions for $r=1$ are given by

$$Y_1\,(t;r)=Y_3\,(t;r)=\frac{50}{3}+\frac{2,950}{3}\mathrm{e}^{\frac{-3}{10}t},$$

$$Y_2\,(t;r)=Y_4\,(t;r)=\frac{500}{3}-\frac{29,500}{27}\mathrm{e}^{\frac{-3}{10}t}+\frac{2,500}{27}\mathrm{e}^{\frac{-3}{100}t}.$$

Conclusion

For tested Problems 1 and 2, the approximated solution by FIRK4-4, exact solution, and maximum global error are given in Tables 2, 3, 4, and 5. The numerical results show that FIRK4-4 with four stages gives high error accuracy. Also Figs. 1 and 2 show the curve of approximated solution compared with the exact solution, and we can see that the approximated solution by FIRK4-4 almost tends to the exact solution which indicates the accuracy of method.

In this paper we developed the fuzzy improved Runge–Kutta methods for solving first-order FDEs. The scheme is two step in nature and is based on the improved Runge–Kutta method for solving ordinary differential equations. The method of order four with four stages is proposed. Numerical results show that fuzzy improved Runge–Kutta methods with high error accuracy are efficient for solving first-order FDEs.

Acknowledgment This research has been financed by University Putra Malaysia using Universiti Putra grant with project number GP-IPM/2013/9405100 at the Institute For Mathematical Research, Universiti Putra Malaysia.

References

1. Chang, S.L., Zadeh, L.A.: On fuzzy mapping and control. IEEE Trans. Syst. Man Cybern. **2**, 30–34 (1972)
2. Dubois, D., Prade, H.: Towards fuzzy differential calculus part 3: differentiation. Fuzzy Set. Syst. **8**, 225–233 (1982)
3. Abbasbandy, S., Allahviranloo, T.: Numerical solution of fuzzy differential equations by Taylor method. J. Comput. Methods Appl. Math. **2**, 113–124 (2002)
4. Seikkala, S.: On the fuzzy initial value problem. Fuzzy Set. Syst. **24**, 319–330 (1987)
5. Ahmad, M.Z., Hasan, M.K.: Anew fuzzy version of Euler's method for solving differential equations with fuzzy initial values. Sains Malays. **40**(6), 651–657 (2011)
6. Rabiei, F., Ismail, F.: Improved Runge–Kutta methods for solving ordinary differential equations. Sains Malays. **42**(11), 1679–1687 (2013)
7. Akbarzadeh Ghanaie, Z., Mohseni Moghadam, M.: Solving fuzzy differential equations by Runge–Kutta method. J. Math. Comput. Sci. **2**(2), 208–221 (2011)
8. Chalco-Cano, Y., Roman-Flores, H.: On new solution of fuzzy differential equations. Chaos Solitons Fractals **38**, 112–119 (2008)
9. Friedman, M., Ma, M., Kandel, A.: Numerical solutions of fuzzy differential equations. Fuzzy Set. Syst. **105**, 133–138 (1999)
10. Solaymani Fard, O., Ghal-eh, N.: Numerical solution for linear system of first order fuzzy differential equations with fuzzy constant coefficients. Inform. Sci. **181**, 4765–4779 (2011)

Effect of Bird Strike on Compressor Blade

A. Ajin Kisho, G. Dinesh Kumar, John Mathai, and Vickram Vickram

Abstract Certification requirement demands for civil and military aircraft to withstand the impact of foreign object damage at critical flight conditions. Experimental tests for conducting bird impact analysis are costly and time-consuming, and thus an accurate solution for designing a component against bird impact is important. Bird impact on aircraft is a soft body impact; it requires the density of a fluid, viscosity, and shape of bird projectile and length to diameter ratio should be precisely selected. This paper investigates the effect and influence of all such parameters due to bird impact. The initial degradation and failure of individual compressor blades struck by a bird were investigated. Subsequent damage to other fan blades and engine components is also evaluated. Results will be compared in terms of pressure profile, and stagnation pressure at the center of the impact and the bird trajectory after the impact. The bird strike velocity varied from 190 to 250 m/s. A numerical model of this problem has been developed with the finite, non-nuclear element program LS-DYNA. This paper presents the bird strike analysis using Lagrangian, Arbitrary Lagrangian Eulerian (ALE) method, and Smooth Particle Hydrodynamic (SPH) technique in LS-DYNA. Throughout the study, the most influencing parameters have been identified and peak pressures and forces are compared to those results available in the literature.

Introduction

Bird strikes present a significant safety and financial threat to aircraft worldwide [1]. The bird strikes were estimated to cost commercial aviation over one billion dollars worldwide during 1999–2000. Despite the efforts provided to avoid collisions between birds and aircraft and to produce bird-proof aircraft, bird strike causes every year damages of millions of US-dollars and unacceptable losses in human lives. Bird strike is a major threat to aircraft because the collision with a bird during flight can lead to structural damage.

A.A. Kisho (✉) • G.D. Kumar • J. Mathai • V. Vickram
Department of Aeronautical Engineering, Hindustan University, Chennai, Tamil Nadu, India
e-mail: kishoaero@gmail.com

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.40_16

Table 1 Number of strikes in different phase

Phase	Number of strikes	Damage (%)
Landing/landing roll	1, 351	3
Approach	1, 130	7
Take-off	996	5
Climb	433	10
Parked/ground checks	53	13
Descent	30	10
Taxi	30	3

Although exterior aircraft structures are exposed to various foreign object damage like runway debris or tire rubber impact, about 90 % of all incidences today are reported to be caused by bird strike.

History

Since 1988, such incidents have claimed the life of over 195 people [1]. In United States, more than 50,000 incidents of bird strikes were reported between 1990 and 2003. In India, there were 74 reports of bird hits from military and civilian airports in 1997, 45 in 1998, and 39 in 1999. There were 26 more reports till September 2000 and there was no loss of life. The number of vultures have declined in North India and there were very few incidents of vulture hits on aircraft. The bird strikes of civil and military aircraft are reported into a National Wildlife Strike Database. Table 1 shows the number of strikes in different phases.

Need for Simulation

The use of computer simulation to simulate the bird impact on new structural components serves as a major tool for the development of new components by minimizing the number of empirical testing [2]. It allows the impact of different structural and material parameters to be studied before the actual fabrication of the prototypes, thus reducing time and cost incurred in empirical testing. In view of this, before being introduced into operational work, the aircraft components must be certified for a certain level of bird impact resistance. After a bird strike, the aircraft must be able to safely land. Full-scale tests with real flesh-and-bones birds are mandatory for the homologation of new structures.

Bird strike tests are expensive, difficult to perform, and little repeatable [3]. The analytical and numerical schemes have been implemented to support the development of new structures and hence to reduce time and cost.

FAA Regulation on Bird Strike

Federal Aviation Regulation (FAR) 33.76 requires that engines be capable of withstanding impact with birds ranging from 0.8 to 8 lb [3], without sustaining damage that poses a fatal threat to passengers and crew.

To meet these standards, jet engine manufacturers depend heavily on a multiple costly physical tests requiring the destruction of full-scale engines by simulating bird strikes using birds or bird surrogates in accordance with FAA guidelines. The failure rate of aircraft engines has reached extremely low. This means that many flight crews will never have to face an engine failure during their career than occurring in flight simulator. However, simulators are not fully representative of engine failures because accelerations due to a failed engine, noise caused by an engine stall, vibrations in the event of a blade rupture are hard to simulate.

The engine must continue to produce at least 75 % thrust for 5 m after ingesting a bird. Fan integrity tests must demonstrate that the engine does not catch fire or disintegrate after being struck by a single 4 lb bird.

Methodology

In the early stages of bird-strike simulations, the bird was represented by a pressure pulse on the structure. This was based on the assumption that, since a bird is mainly made of water, it could be represented by a jet of fluid [4]. Since then, many progresses have been made. Table 2 refers the bird weight requirements as per standards. The blade property, bird surrogates, bird model, and bird shape were obtained from literature.

Blade Material

The material property of blade is taken as Titanium Ti-64 and following bilinear material property [5] (Table 3).

Table 2 Bird weight requirements

Component	Bird weight (lb)	Regulation far 25
Windshield	4	775
Wing leading edge	4	571
Empennage	8	571 and 571
Engine	4	Section 33

Table 3 Various parameters of blade. Bird Properties considered for the analysis

Material property	Symbol	Value
Density	ρ	4,420 kg/m^3
Young's modulus	E	119.35 GPa
Hardening modulus	E_h	0.959 GPa
Yield stress	σ_y	1,311 MPa
Poisson's ratio	υ	0.3

Theory of Bird Strike

A bird undergoing impact at very high velocity behaves as a highly deformable projectile where the yield stress is much lower than the sustained stress [6]. Accordingly, the impact can be categorized as a hydrodynamic impact. That, and the fact that the density of flesh is generally close to the density of water, makes it possible for a bird to be imagined as a lump of water hitting a target.

The bird strike event is divided into two stages, the initial shock (time at impact) and steady flow. The pressure of the initial shock is called Hugoniot pressure and is given by equation [4]; the pressure of the steady flow (stagnation pressure) is calculated according to Bernoulli and is given by equation

$$P_{sh} = \rho \upsilon_{sh} \upsilon_{im} \tag{1}$$

$$P_{stag} = \frac{1}{2} \rho \upsilon_{im}^2 \tag{2}$$

Equation gives the stagnation pressure for an incompressible fluid; however, if the fluid is compressible, its value will increase with respect to its porosity, z. Airoldi gives a useful expression to calculate the modified stagnation pressure

$$P_{stag\ z} = \frac{1}{1-z} P_{stag} \tag{3}$$

Analytically, those two pressures are important since the Hugoniot pressure gives the maximum possible value for the impact at its beginning and the stagnation pressure gives the expected reading when the flow stabilizes. It is important that the pressure is not dependent on the size of the projectile since the mass is not a variable in the pressure equations. This implies that the pressure results are the same regardless of the projectiles, provided that they share the same impact velocity. The force and energy of a bigger projectile is proportionally larger and will cause more damage.

The values of the variables needed to calculate the stagnation pressure are easily available. On the contrary, the Hugoniot pressure depends on the impact velocity and the shock velocity, which depends on the impact velocity. Moreover, the equation changes even if the porosity is included or not, or if the fluid considered is water or a substitute. The equations given below apply to a projectile with an amount of air mixed in, also called porosity, since experience has shown that porosity has a non-negligible effect on the overall results and is closer to the behavior of a bird upon impact.

$$\rho_1 \upsilon_{sh} = \rho_2 (\upsilon_{sh} - \upsilon_{im}) \tag{4}$$

$$P_1 - P_2 = \rho \upsilon_{sh} \upsilon_{im} \tag{5}$$

$$\frac{\rho_1}{\rho_2} = (1 - z)(\frac{P_2}{A} + 1)\frac{-1}{B} + z(1 - q) \tag{6}$$

With, $A = \frac{\rho_1 C_0^2}{4k-1}$,
$B = 4k - 1$ and

$$\frac{\rho_2}{\rho_1} = \frac{1}{1-q} \tag{7}$$

$z(1 - q)$ is the contribution of the air mixed and it is negligible.

Steps in Creating the Bird Model

The bird model is obtained theoretically and experimentally. It describes the theory of the bird strike and provides a sample of the available experimental data [7]. Then a demonstration is given as to how to evaluate a bird model based on the following criteria:

1. Pressure profile at the center of impact
2. Mass loss
3. Impulse profile at the center of impact
4. Radial pressure distribution
5. Shape of the sustained deformations
6. Solution time

The three modeling methods mentioned earlier are presented in addition to a brief parametric study [14] of the factors influencing the fluid–structure interaction [14]. They are compared and evaluated with respect to the theoretical and experimental information available. The experimental data which are generally used as a reference are evaluated with respect to the theory, demonstrating that although useful, they should be referred to with care.

Bird Surrogates

The impact of a real bird is representative of that impact itself [7]. Considering the impact of a real bird, not only the weight and the physical properties of the bird, but also parameters such as the species, the age, and the size are relevant because of the influence on the impact loads.

The jelly projectile model is generally accepted as a substitute of the real bird [2]. At high impact velocity, a bird impacting a rigid or deformable structure like an aircraft behaves like a fluid. Hence, a hydrodynamic material model is a reasonable approximation.

Bird Shape

The shape of the bird is important since both the impact force and pressure are considered. In addition, the shape of the bird becomes important [8]; when it is necessary to obtain specific load conditions, a blunt cylinder shape is used as a bird surrogate for the bird strike on the compressor blade, while the rugby-ball shape is recommendable to reproduce the impact loads of a real bird (Fig. 1).

Several representative bird geometries have been proposed for both impact experiments utilizing bird surrogates and finite element models of high velocity bird impact [8]. One of the most common is a cylinder with rounded hemispherical ends. The ratio of total length to diameter for the bird geometry was 2.

It is necessary to define both a suitable material model and a pressure–volume state equation for the bird. For majority of the interaction, the blade interacts with the gelatin-like fluid and only interacts with the solid bird for a brief instant at the moment of impact.

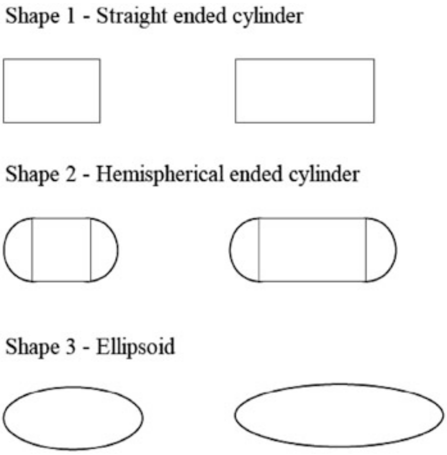

Shape 1 - Straight ended cylinder
Shape 2 - Hemispherical ended cylinder
Shape 3 - Ellipsoid

Fig. 1 Various shapes for bird

Table 4 Bird parameters

S. no.	Parameters	Assumptions or values chosen
1	Bird mass	1.82 kg
2	Bird geometry	Cylinder with hemispherical ends
3	Bird density	950 kg/m^3
4	Bird material	Viscous hydrodynamic fluid

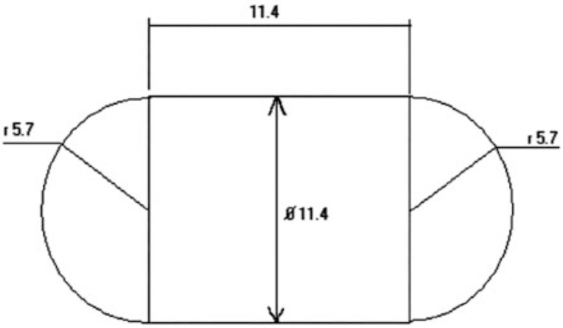


Fig. 2 Hemispherical bird model with dimensions

Bird Model

The bird models were defined by a hemispherical-ended cylinder and were based on a standard volume of 1.917E + 06 mm^3 and length 228 mm (Fig. 2). Table 4 shows the bird parameters for the analysis.

Contact Algorithm

The fluid–structure interaction in the bird impact simulation is the contact algorithm [9]. It helps to prevent penetrations and calculates reaction forces. The contact algorithm helps for large deformations and splitting of the projectile, sliding of the bird material over the target surface and the creation of multiple contact interfaces due to possible fracture and penetration of the structure. Friction is another aspect, whereas the study advises that best results compared to experimental results can be obtained with zero friction.

Equation of State

Real birds and artificial gelatin birds are mostly made up of water. Therefore, a hydrodynamic response can be considered as a valid approximation for a constitutive model for bird strike analyses [10]. An equation of state (EOS) describes the pressure–volume relationship with parameters of water at room temperature.

The research works focus on the EOS of the bird. Initially, a polynomial EOS with the parameters of the water at room temperature was used. Subsequently, these parameters were modified to keep into account the porosity of the jelly used in the tests. Hydrodynamic pressure–volume relation can be defined by an EOS in the form of a third-degree polynomial [14].

$$\mathrm{P} = \mathrm{C}_0 + \mathrm{C}_1\mu + \mathrm{C}_2\mu^2 + \mathrm{C}_3\mu^3 \tag{8}$$

$$\mu = (\rho) / (\rho_0) - 1 \tag{9}$$

where, μ is relative density. For a material such as water which exhibits the linear Hugoniot relation between shock velocity and particle velocity, the EOS can be expressed in terms of the following coefficients

$$\mathrm{C}_0 = 0 \tag{10}$$

$$\mathrm{C}_1 = \rho_0\mathrm{C}_2 \tag{11}$$

$$\mathrm{C}_2 = (2\mathrm{k} - 1)\,\mathrm{C}_1 \tag{12}$$

$$\mathrm{C}_3 = (\mathrm{k} - 1)\,(3\mathrm{k} - 1)\,\mathrm{C}_1 \tag{13}$$

Numerical Analysis

In recent years, explicit FE codes have been used to develop high efficiency bird-proof structures. These codes adopted various finite element approaches to model the impact phenomena: the Lagrangian approach, Eulerian or Arbitrary Lagrangian Eulerian (ALE) approach, and Smooth Particle Hydrodynamics (SPH).

Lagrangian Approach

The Lagrangian modeling method is the standard approach for most structural finite element analyses with the nodes of the Lagrangian mesh being associated to the material [7, 11].

The major problem associated with the Lagrangian bird impact models is the severe mesh deformation [11] (Fig. 3). Large deformations of the elements may lead to inaccurate results, severe hour glassing, reduced time steps, and error termination, which have to be reduced.

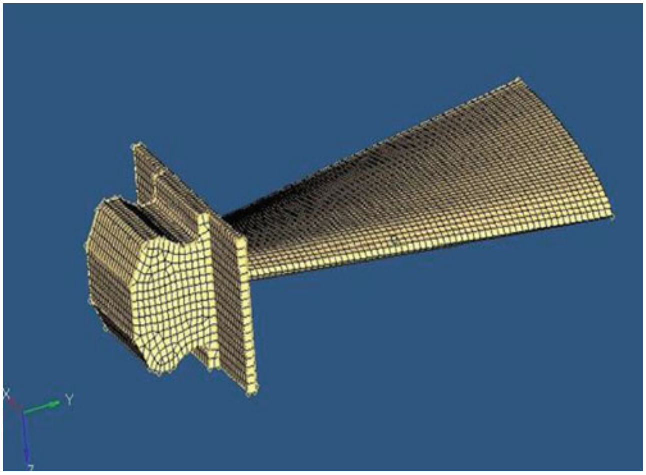


Fig. 3 Meshed blade

The Lagrangian modeling method divides a volume into a large number of small geometries called elements. Because those geometries are simple in shape, it is possible to know the state of the solid by using mathematical relations. However, when the deformations are large, it becomes increasingly difficult to calculate the state and stresses in the elements because the time step, based on the aspect ratio, keeps on decreasing. Moreover, the accuracy of the results obtained decreases. Also, since in this method the material moves with the mesh, if the material undergoes large deformations, the mesh will also suffer some deformation and this leads to results which are inaccurate and numerical instabilities. The mesh size was found to have the most influence on the result. The number of elements obtained in this approach is 21,380 elements.

Arbitrary Lagrangian Eulerian Approach

A better alternative is the Eulerian modeling technique, where the mesh remains fixed in space and the material flows through the mesh. Because the mesh is fixed, mesh deformations do not occur and the explicit time step has no influence [12]. Stability problems do not occur due to excessive element deformation. The initial idea of ALE modeling is taken from the Eulerian formulation for fluid flow where a material moves through a fixed mesh. The main difference is that, here, the mesh is allowed to deform and move so as to follow the flow of fluid. This represents a major improvement with respect to the Eulerian mesh because it decreases the size of the required mesh to a certain extent.

In a bird strike simulation typically, only the impact is modeled as a fluid-like body with Eulerian elements and the target as a solid structure with Lagrangian elements. Because the mesh in the classical Eulerian technique is fixed in space, the computational domain, a coupled Eulerian–Lagrangian approach, is used for this fluid–structure interaction problem like bird strike problems. The computational domain for the structural analyses with the classical Eulerian technique is large, leading to high computational cost due to the high number of elements and the cost-intensive calculation of element volume fractions and interactions. The element size of the Eulerian mesh has to be very small in order to achieve accurate results.

In the classical Eulerian approach, the surrounding Eulerian box is not fixed in space but can be moved if needed. The initial number of elements for the Eulerian domain can be reduced, leading to computational time savings. However, due to the wide spreading of the bird material, the lateral expansion of the Eulerian box is significant and the size of the Eulerian elements is increased considerably. ALE, which is multi-material Eulerian method, the material flows through a mesh, in which each element is allowed to take two or more material (Fig. 4). In Eulerian mesh, there is dissipation of mass of elements.

Smooth Particle Hydrodynamic Approach

SPH is particle method which is applicable to wide range of physics like Crash, Mechanics, and fracture models in Brittle and Ductile materials of solids. It is treated to be very easy for representing the physics, which makes SPH very extraordinary method. Due to the reason that SPH is a very simple method, many problems are hardly reproduced with classical methods. The fluid is represented as set of particles moving with some flow velocity. And compared to ALE method,

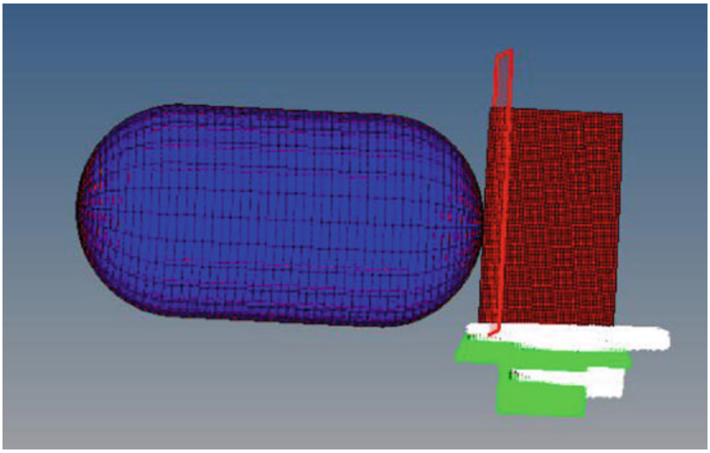


Fig. 4 ALE approach model

from computation point of view, SPH is more economical with its capabilities to calculate only on particles. Because of the large deformation of a bird, this theory is also applicable to bird strike analysis in spite of the much lower velocity [11].

The SPH method uses the Lagrangian formulation for the equations of motion, but instead of a grid, it uses interpolation formula, called kernel functions, to calculate an estimation of the field variables at any point (Fig. 5). The kernel function is active only over a given neighborhood for each node, called support domain. The method is said to be mesh-free because there is no predefined grid of nodes restraining which nodes can interact together. The number of elements obtained in this approach is 9,000 elements.

In practice, the SPH method uses fewer elements than the ALE method, avoids the material interface problems associated with it, and has a shorter solution time. It also follows the flow of the bird much more accurately than the previous methods, especially in the case of secondary bird strike (if the bird is deflected to another structural component).

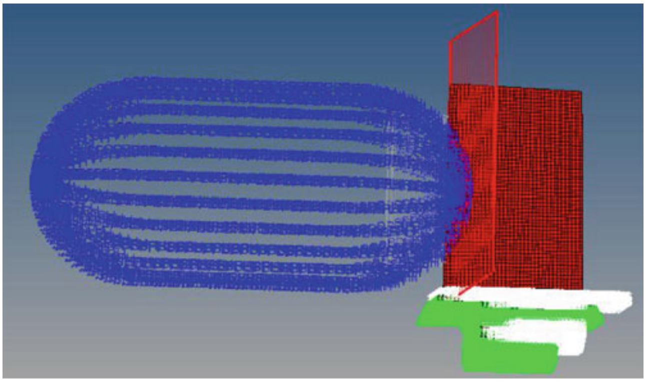


Fig. 5 Meshed using SPH method

Similarly to the Lagrangian mesh, the size of the SPH particles, or the amount of particles used, has an influence on the fluid–structure interaction, and hence the final results. The particles are evenly distributed, which is important because for the time being, the initial dimension of the support domain is the same for all the particles.

Results and Discussion

Lagrangian Approach

The deformation of the blade and the bird for this case are shown in Figs. 6 and 7. The figure shows the evolution of the bird deformation during impact [13]. The severity of contact at the initial stages of impact will lead to plastic deformation in the blade.

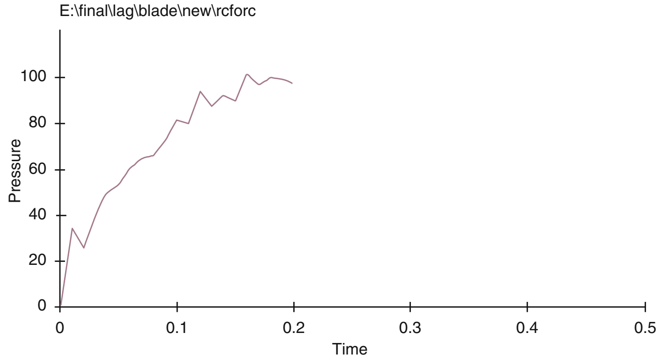
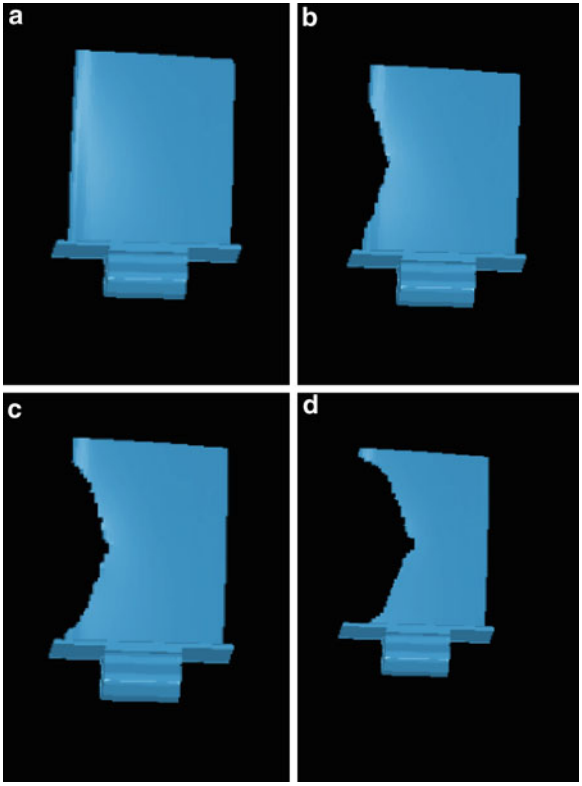


Fig. 6 Variation of pressure with time by Lagrangian approach

Fig. 7 Deformation of compressor blade by Lagrangian approach (**a**) T = 0, (**b**) T = 0.05, (**c**) T = 0.1, (**d**) T = 0.2



The bird reaches its maximum impact force of 140 KN at about 0.16 ms after initial contact. There are two dominant peaks for the hemispherical-ended bird, the first reaches its maximum force of 38 KN at about 0.035 ms and the second reaches its second peak of 123 KN at about 0.16 ms.

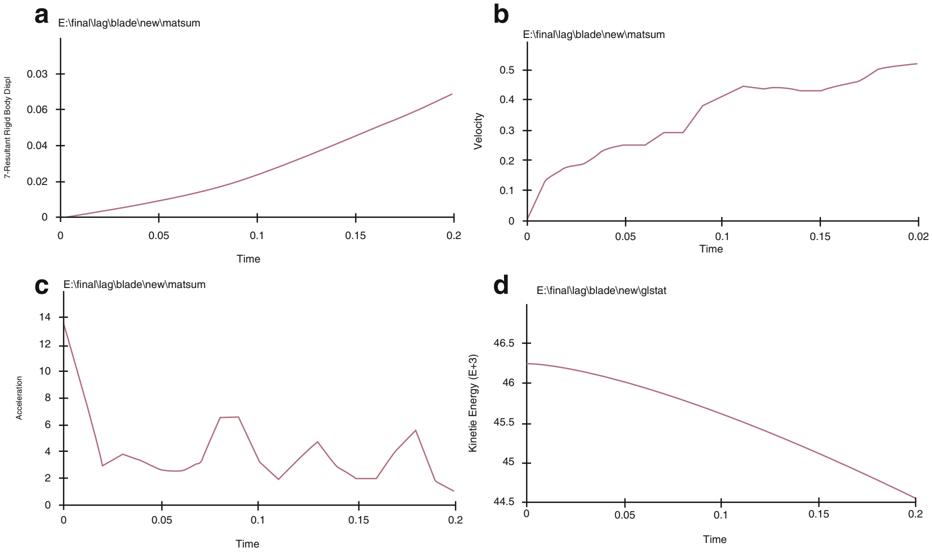


Fig. 8 Variation of (**a**) displacement, (**b**) velocity, (**c**) acceleration, (**d**) kinetic energy vs. time by Lagrangian approach

The kinetic energy of the bird, blade, and the system, which is composed of both the bird and blade, have been plotted in Fig. 8. The kinetic energy is transmitted from the bird to the compressor blade. But, the total kinetic energy decreases, as it gets transformed into deformation energy of the compressor blade, as can be seen from Fig. 8. The deformation energy of the blade is found to be much larger than that of the bird, due to the inelastic fluid-like properties of the bird elements.

ALE Approach

At the beginning of the analysis, the highly denser material is concentrated in one part of the mesh, but as the analysis progresses, the fluid is allowed to flow throughout the model. Some finite elements analysis software even makes it possible to only model the fluid. At each time step, the position of the material is evaluated with respect to the nodes. The coupling with a solid structure is done by tracking the relative displacements between the coupled Lagrangian nodes and the bird. However, mesh distortion can become an issue with the ALE method if the elements' volume becomes negative, and it is often difficult to track material.

The interaction between the bird and the structure is controlled by the *constrained-Lagrange–in-solid card in LS-DYNA. When using the ALE method, this card is critical to obtain good results. It is important to allow coupling only between the bird and the structure, otherwise the gap of air can interfere. Also, the minimum volume fraction required for an element to be computed should be high enough so that the pressure rise is instantaneous once the bird strikes.

The impact force obtained by ALE approach is 0.080 MN. This result cannot be compared with the Lagrange approach because the geometrical models in both cases are not same. The variables used in the ALE cards for this case will be used as reference to create an ALE model that fits the geometrical dimensions of a bird strike analysis (Fig. 9). The maximum impact force obtained by this approach is 0.012 MN.

Finally, the most important parameter is the penalty factor which governs the interaction between the fluid and the structure. Damping should be adjusted so that the pressure remains positive at all times. There are two coupling options: one can either adjust the penalty factor to a constant value, or use a load curve which will increase the stiffness linearly according to the penetration. Both options have been considered in the simulations (Figs. 10 and 11).

Contact	Type	Peak pressure value (MPa)
*Constraint Lagrange in solid	Lagrange approach	101.98
*Control ALE	ALE approach	124.28

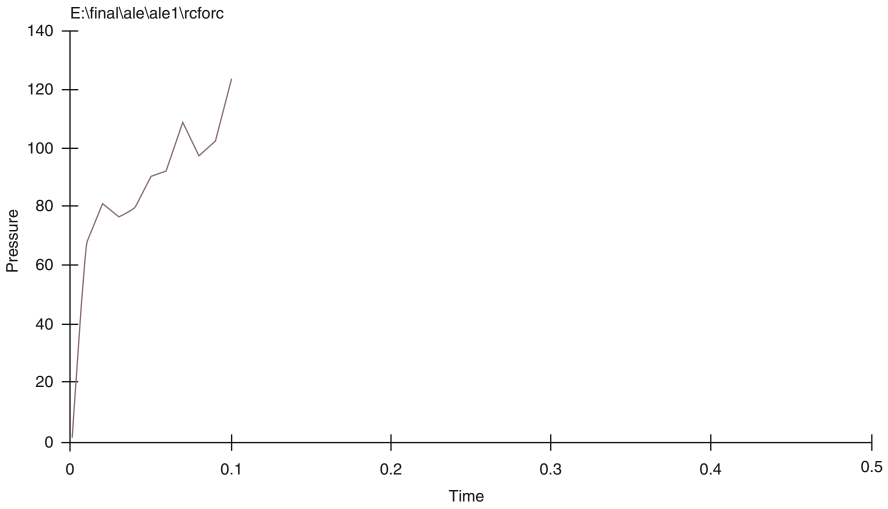


Fig. 9 Variation of pressure with time by ALE approach

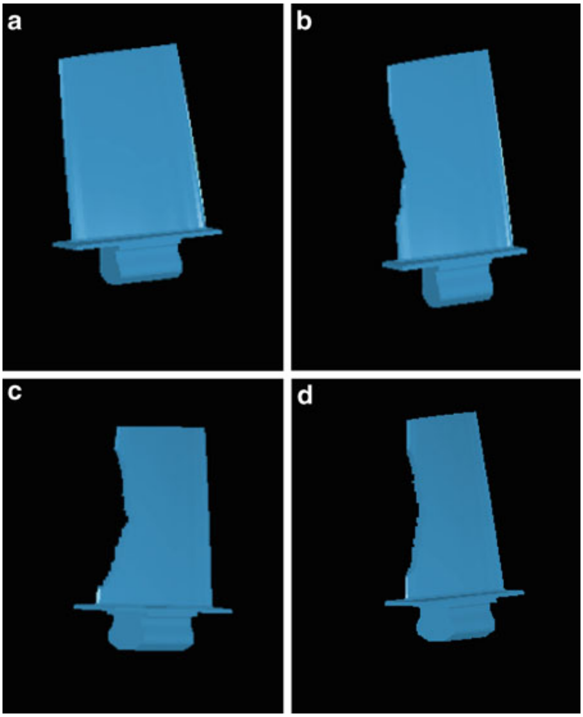


Fig. 10 Deformation of compressor blade by ALE approach (**a**) T = 0, (**b**) T = 0.03, (**c**) T = 0.06, (**d**) T = 0.1

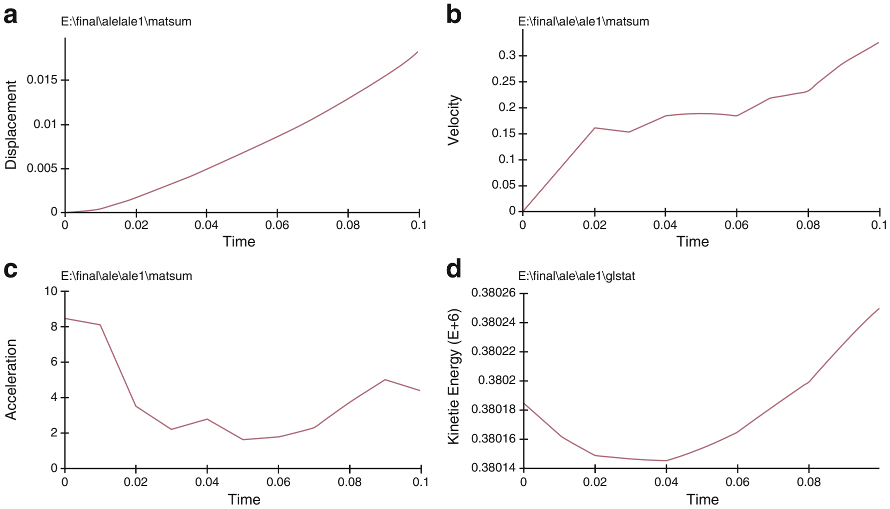


Fig. 11 Variation of (**a**) displacement, (**b**) velocity, (**c**) acceleration, (**d**) kinetic energy vs. time by ALE approach

Conclusions

The importance of compressor blade and its influence on the aircraft engine, the understanding of bird strike problem, and its proper analysis now have even more significance. While manufacturers have high priority to certification, it is also important to have an accurate numerical model to assess the bird impact resistance on the compressor blades. To validate the present finite element formulation, comparison with the experimental data for a bird striking a compressor blade can be carried out. The Lagrangian bird, modeled using hydrodynamic constitution law, is found to be appropriate for the bird strike analysis.

To understand the importance of bird geometry modeling, the frequently used bird and configurations have been examined. It is found that the initial contact area between the bird and compressor blade in the early phase of the impact has a significant effect on the peak impact force value.

The impact force profile is also found to be highly dependent on the deformation of the compressor blade. The maximum impact force increases with larger birds, but for the maximum plastic strain of the blade due to the reduced density of the larger bird.

References

1. Mao, R.H., Meguid, S.A., Ng, T.N.: Finite element modeling of a bird striking an engine fan blade. J. Aircr. **44**(2), 583–596 (2007)
2. Alberto, C.: Robust bird strike modeling using LS DYNA. Machine design at University of Puerto Rico-Mayaguez Campus, Technical Communicator at Ferreyros SAA, Peru
3. Tho, C.-H., Smith, M.R.: Bird strike simulation for BA609 spinner and rotor controls. In: 9th International LSDYNA Users Conference, 2006
4. Mithun, N., Mahesh, G.S.: Finite element modelling for bird strike analysis and review of existing numerical methods. Int. J. Eng. Res. Technol. **1**(8), (2012)
5. McCallum, S.C., Constantinou, C.: Base systems: the influence of bird shape in bird strike analysis. In: 5th European LS DYNA User's Conference
6. Trivikram, et al.: Effect of ply number & orientation of composite structure in bird strike analysis. In: EASI Engineering, 4th ANSA & μETA International Conference
7. Ryabov, A.A., et al.: Fan blade bird strike analysis using Lagrangian, SPH and ALE approaches. In: 6th European LS-DYNA User's Conference
8. Lavoie, M.A., et al.: Validation of available approaches for numerical bird strike modelling tools. Int. Rev. Mech. Eng. **xx**, (2007)
9. Chalipat, S., Shankapal, S.R.: Characterization of bird impact properties using finite element code. Coventry University Postgraduate Study Centre, MSRSAS, Bangalore (2002)
10. Anil Kumar, P.V.: Bird strike analysis of typical gas turbine stator blade. Automotive Engineering Centre, MSRSAS, Bangalore (2003)
11. Airoldi, Cacchione, B.: Modeling of impact forces and pressures in Lagrangian bird strike analyses. Int. J. Impact Eng. **32**, 1651–1677 (2006)

12. Dobyns, A., et al.: Bird strike analysis and test of a spinning S-92 tail rotor. In: American Helicopter Society 57th Annual Forum, Washington, 2001
13. Monaghan, J.J., Gingold, R.A.: Shock simulation by the particle method SPH. J. Comput. Phys. **52**, 374–389 (1983)
14. Chuan, K.C.: Finite element analysis of bird strikes on composite and glass panels. Department of Mechanical Engineering, National University of Singapore (2005/2006)

Part IV
Statistics

Asymptotic Density Crossing Points of Self-Normalized Sums and Normal

Thorsten Dickhaus and Helmut Finner

Abstract We define generalized self-normalized sums as t-type statistics with flexible norming sequence in the denominator. It will be shown how Edgeworth expansions can be utilized to provide a full characterization of asymptotic crossing points (ACPs) between the density of such generalized self-normalized sums and the standard normal density. Although the proof of our main ACP theorem is self-contained, we also draw connections to related expansions for the cumulative distribution function of generalized self-normalized sums that we have derived in previous work.

Keywords Edgeworth expansion • Large deviations • Likelihood ratio • Student's t

Introduction

Density crossing points (CPs) are important objects in statistical theory and practice. To mention only one specific application, the CPs between two probability density functions (pdfs) f and g (say) determine level sets of the likelihood ratio f/g. Such level sets in turn characterize rejection regions of likelihood ratio tests or decision boundaries of Bayes classifiers, respectively. For further possible applications, we defer the reader to Appendix B in [1]. In [2], a systematic approach toward characterizing (asymptotic) crossing points (ACPs) between the standard normal density φ and densities of standardized sums of independent and identically distributed (iid) random variables fulfilling the conditions of the central limit theorem has been worked out. Moreover, the authors derived the (asymptotic) density crossing

T. Dickhaus (✉)
Institute for Statistics, University of Bremen, Bremen, Germany
e-mail: dickhaus@uni-bremen.de

H. Finner
Institute of Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research, Düsseldorf, Germany
e-mail: finner@ddz.uni-duesseldorf.de

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.02_17

points between φ and the Lebesgue densities f_ν of t-distributions with ν degrees of freedom (with ν tending to infinity). According to [3], f_ν is given by

$$f_\nu(z) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}} \left(1+\frac{z^2}{\nu}\right)^{-\nu/2-1/2}.$$

Recall that f_ν is the density of the self-normalized sum

$$S_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}}, \tag{1}$$

assuming that $n = \nu + 1$, the random variables $X_i : 1 \leq i \leq n$ are iid normally distributed with mean μ and variance $\sigma^2 \in (0,\infty)$, and $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$.

The major results in Section 3 of [2] and Appendix A of [1] were that there exist exactly two CPs of φ and f_ν for any $\nu > 0$ and that the positive solution z_ν (say) of the equation $\varphi(z) = f_\nu(z)$ converges monotonically to its limiting value $\lim_{\nu\to\infty} z_\nu = \sqrt{1+\sqrt{2}} \approx 1.553773974$. In the present work, we generalize the latter findings by providing a theory of the ACPs of the pdfs of (generalized) self-normalized sums and φ. Our results are of quite general character. Especially, no normal distribution has to be assumed for the initial random variables X_i, and the norming sequence in the denominator of the self-normalized sum can take different forms.

Notation and Preliminaries

Throughout the work, $(X_i)_{i\in\mathbb{N}}$ denotes a sequence of iid random variables with values in $\mathbb{R}$ and $\mathbb{E}[X_1^2] < \infty$. Further moment conditions will be imposed where necessary. Moreover, we require that the following condition holds true throughout the remainder.

Condition 1. The distribution of X_1 has a nondegenerate absolutely continuous component with respect to the Lebesgue measure.

Condition 1 implies Cramér's condition; see p. 45 in [4].

We denote the expectation of X_1 by $\mu = \mathbb{E}[X_1]$ and its variance by $\sigma^2 = \mathbb{E}[(X_1 - \mu)^2]$. In [5, 6] and [7], the statistic

$$Y_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}}$$

has been investigated with respect to Edgeworth expansions. Peter Hall's result in [7] was that

$$F_{Y_n}(y) = \Phi(y) + \sum_{i=1}^{k} n^{-i/2} P_i(y)\varphi(y) + o(n^{-k/2}) \tag{2}$$

uniformly in $y \in \mathbb{R}$ for any $k \geq 1$ for which the $(k+2)$-nd moment of X_1 exists ("minimal moment condition"), provided that Condition 1 is fulfilled. The first four polynomials on the right-hand side of (2) are given by

$$\begin{aligned}
P_1(y) &= \frac{\alpha_3}{6}(2y^2+1),\\
P_2(y) &= -y\left\{\frac{\alpha_3^2}{18}(y^4+2y^2-3) - \frac{\kappa}{12}(y^2-3) + \frac{1}{4}(y^2+3)\right\},\\
P_3(y) &= \left(-\frac{1}{36}y^6 + \frac{5}{48} + \frac{5}{24}y^4 + \frac{5}{8}y^2\right)\alpha_3\alpha_4\\
&\quad + \left(-\frac{1}{40} - \frac{1}{20}y^4 - \frac{1}{5}y^2\right)\alpha_5\\
&\quad + \left(-\frac{35}{216}y^4 + \frac{1}{162}y^8 - \frac{175}{432}y^2 + \frac{7}{324}y^6 - \frac{35}{432}\right)\alpha_3^3\\
&\quad + \left(-\frac{1}{4}y^4 - \frac{1}{8}y^2 - \frac{1}{16} + \frac{1}{6}y^6\right)\alpha_3,\\
P_4(y) &= \left(-\frac{1}{1944}y^{11} + \frac{25}{108}y^5 - \frac{5}{216}y^3 - \frac{5}{1944}y^9 + \frac{5}{108}y^7 - \frac{35}{72}y\right)\alpha_3^4\\
&\quad + \left(\frac{1}{216}y^9 - \frac{5}{12}y^5 + \frac{1}{6}y^3 + \frac{29}{24}y - \frac{1}{18}y^7\right)\alpha_3^2\alpha_4\\
&\quad + \left(-\frac{1}{6}y + \frac{1}{36}y^3 - \frac{1}{36}y^9 + \frac{11}{36}y^5 + \frac{1}{12}y^7\right)\alpha_3^2\\
&\quad + \left(-\frac{1}{12}y^3 - \frac{1}{2}y + \frac{2}{15}y^5 + \frac{1}{60}y^7\right)\alpha_3\alpha_5\\
&\quad + \left(-\frac{37}{96}y - \frac{1}{288}y^7 + \frac{7}{96}y^5 - \frac{11}{96}y^3\right)\alpha_4^2\\
&\quad + \left(\frac{1}{4}y + \frac{1}{24}y^7 + \frac{1}{24}y^3 - \frac{1}{4}y^5\right)\alpha_4\\
&\quad + \left(\frac{1}{18}y^3 + \frac{1}{6}y - \frac{1}{45}y^5\right)\alpha_6 - \frac{1}{8}y^7 + \frac{3}{8}y^5,
\end{aligned}$$

where

$$\alpha_\ell = \frac{\mathbb{E}[(X_1-\mu)^\ell]}{\sigma^\ell},\ \ell = 3,\ldots,6, \text{ and } \kappa = \alpha_4 - 3, \tag{3}$$

assuming that the respective minimal moment condition holds true. The functional κ is usually referred to as the excess kurtosis of X_1. The polynomials P_1 and P_2 are nowadays well known and can be found in various statistics textbooks. Two different methods for deriving the approximation polynomials can be found in [6] in an elementary way and in the textbook [4] under a broad scope of the so-called smooth function models. Computational or algorithmic, respectively, methods for the derivation of the P_i's in such smooth function models have been discussed in [8]. The method of [6] has recently been worked up and described in a more straightforward way; see [9].

In Section 11.4.1 of [10], an Edgeworth expansion for the self-normalized sum S_n as in (1) with normalization constant $(n-1)^{-1}$ in the denominator is given. The approximation polynomials in this case differ from the P_i's from above. This shows that the norming sequence in the denominator of the self-normalized sum is of importance for the asymptotic behavior of these t-type statistics. Therefore, we investigate a more general statistic in this paper, namely,

$$T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{a_n \sum_{i=1}^n (X_i - \bar{X}_n)^2}}$$

which differs from Y_n with respect to the norming sequence a_n. However, it will be assumed that $\lim_{n\to\infty} n a_n = 1$.

In the remainder, we refer to S_n as a "Studentized sum," to Y_n as a "standard self-normalized sum," and to T_n as a "generalized self-normalized sum." In order to study pdfs, it seems tempting to take formal derivatives on both sides of (2). However, since the derivative operator can induce singularities and diminish smoothness, an extra assumption is necessary to carry the expansion (2) over to the pdf case. As pointed out in Sect. 2.8 of [4], the assumption of a bounded density of $\bar{X}_n$ for some $n > 1$ is crucial and will therefore appear in our main theorem.

ACP Theorem for Generalized Self-Normalized Sums

Since the polynomials P_i, $1 \leq i \leq 4$, refer to the Edgeworth expansion of the standard self-normalized sum Y_n, let us first express the generalized self-normalized sum T_n in terms of Y_n. Setting $b_n = \sqrt{n a_n}$, we obtain $T_n = \sqrt{(n a_n)^{-1}} Y_n = Y_n / b_n$ and, consequently,

$$F_{T_n}(t) = \mathbb{P}(T_n \leq t) = \mathbb{P}(Y_n \leq b_n t) = F_{Y_n}(b_n t). \tag{4}$$

Equation (4) is the starting point for our considerations with respect to ACPs of the pdf of T_n and the standard normal density φ. The following lemma provides asymptotic expansions for b_n.

Lemma 1. *Let $b_n = \sqrt{n a_n}$, where a_n is given by*

$$a_n = \frac{1}{n(1 - \sum_{j=1}^{M} C_j n^{-j/2})} \tag{5}$$

for some integer $M \geq 4$ and real, given constants $C_1, \ldots, C_M$ (implying $\lim_{n\to\infty} n a_n = 1$). Then, we have that $\lim_{n\to\infty} b_n^j = 1$ for all $j \geq 1$, and asymptotic expansions for b_n and $b_n^2 = n a_n$ are given by

$$\begin{aligned}
b_n &= 1 + \frac{C_1}{2\sqrt{n}} + \frac{C_2 + 3C_1^2/4}{2n} + \frac{C_3/2 + C_1C_2 + 9C_1^3/16 - C_1/4\left(C_2 + C_1^2\right)}{n^{3/2}} \\
&\quad + \frac{1}{n^2}\left[\frac{C_4}{2} + C_1C_3 + \frac{C_2^2}{2} + \frac{3C_2C_1^2}{2} - \frac{C_1\left(C_3 + 2\,C_1C_2 + C_1^3\right)}{8}\right. \\
&\quad - \frac{\left(C_2 + C_1^2\right)^2}{8} - \frac{C_1\left(C_3/2 + C_1C_2 + C_1^3/2\right)}{4} + \frac{59}{128}C_1^4 \\
&\quad \left. + \frac{3C_1^2\left(C_2 + C_1^2\right)}{16}\right] + O\left(n^{-5/2}\right), \\
b_n^2 &= 1 + \frac{C_1}{\sqrt{n}} + \frac{C_2 + C_1^2}{n} + \frac{C_3 + C_1C_2 + C_1\left(C_2 + C_1^2\right)}{n^{3/2}} \\
&\quad + \frac{1}{n^2}\left[C_4 + C_1C_3 + C_2\left(C_2 + C_1^2\right) + C_1\left(C_3 + 2\,C_1C_2 + C_1^3\right)\right] + O\left(n^{-5/2}\right).
\end{aligned}$$

Proof. The lemma is a simple application of the Taylor expansion for the square root function. ■

Remark 1. Actually, the expansion for b_n^2 follows straightforwardly from that for b_n by working out the square. In an analogous manner, expansions for b_n^j for any integer $j \geq 3$ can be obtained.

Theorem 1. *Let Condition 1 be fulfilled, assume that $\mathbb{E}[X_1^4]$ is finite, and let α_3 and κ be defined as in (3). Assume that $\bar{X}_n$ has a bounded density for some $n > 1$. Let the norming sequence $\{a_n\}_{n\in\mathbb{N}}$ in the denominator of T_n be as in (5). Then, the ACP behavior of f_{T_n} and φ can be characterized as follows.*

(i) If $\alpha_3 \neq 0$ or $C_1 \neq 0$, we obtain ACPs between f_{T_n} and φ as solutions of the equation

$$\frac{C_1}{2} + \frac{\alpha_3}{2}t - \frac{C_1}{2}t^2 - \frac{\alpha_3}{3}t^3 = 0.$$

(ii) *If* $\alpha_3 = C_1 = 0$ *and* ($\kappa \neq 3$ *or* $C_2 \neq 3$)*, ACPs between* f_{T_n} *and* φ *are given as solutions of the equation* $B_4 t^4 + B_2 t^2 + B_0 = 0$*, where the real constants* B_4*,* B_2 *and* B_0 *are defined by*

$$B_4 = \frac{1}{4}\left(1 - \frac{\kappa}{3}\right), \; B_2 = \frac{\kappa - C_2}{2},$$
$$B_0 = \frac{1}{2}\left(C_2 - \frac{\kappa + 3}{2}\right).$$

(iii) *In case of* $\alpha_3 = C_1 = 0$ *and* $\kappa = C_2 = 3$*, assume that* α_5 *is finite. Then, if* $\alpha_5 \neq 0$ *or* $C_3 \neq 0$*, we obtain ACPs between* f_{T_n} *and* φ *by solving the equation*

$$\frac{C_3}{2} - \frac{3\alpha_5}{8}t - \frac{C_3}{2}t^2 + \frac{\alpha_5}{20}t^5 = 0.$$

(iv) *If the assumptions of cases (i)–(iii) are not fulfilled and* α_6 *is finite, ACPs are given as solutions of the equation*

$$\left(\frac{\alpha_6}{45} - \frac{3}{2}\right)t^6 + \left(\frac{41}{4} - \frac{\alpha_6}{6}\right)t^4 + \left(3 - \frac{C_4}{2}\right)t^2 + \frac{\alpha_6}{6} + \frac{1}{2}\left(C_4 - \frac{45}{2}\right) = 0.$$

Proof. Due to the bounded density assumption regarding $\bar{X}_n$, we take formal derivatives on both sides of (4), resulting in

$$f_{T_n}(t) = b_n \varphi(b_n t) + \sum_{i=1}^{k} n^{-i/2} \frac{d}{dt}\left[P_i(b_n t)\varphi(b_n t)\right] + o(n^{-k/2}),$$
$$\frac{f_{T_n}(t)}{\varphi(t)} = b_n \frac{\varphi(b_n t)}{\varphi(t)} + \sum_{i=1}^{k} \frac{n^{-i/2}}{\varphi(t)} \frac{d}{dt}\left[P_i(b_n t)\varphi(b_n t)\right] + o(n^{-k/2}). \quad (6)$$

Plugging in $\varphi(t) = \exp(-t^2/2)/\sqrt{2\pi}$, we obtain that (6) is equivalent to

$$\frac{f_{T_n}(t)}{\varphi(t)} = \exp\left(\frac{t^2}{2}(1 - b_n^2)\right)$$
$$\times \left[b_n + \sum_{i=1}^{k} n^{-i/2}\left(\frac{d}{dt}P_i(b_n t) - b_n^2 t P_i(b_n t)\right)\right] + o(n^{-k/2}). \quad (7)$$

Making use of the Taylor expansion $\exp(x) = \sum_{\ell=0}^{m} x^\ell/\ell! + O(x^{m+1})$, $x \to 0$, (7) can equivalently be expressed by

$$\begin{aligned}\frac{f_{T_n}(t)}{\varphi(t)} &= \sum_{\ell=0}^{m_k} \frac{\left(t^2(1-b_n^2)\right)^{\ell}}{2^{\ell}\ell!} \\ &\quad \times \left[b_n + \sum_{i=1}^{k} n^{-i/2}\left(\frac{d}{dt}P_i(b_n t) - b_n^2 t P_i(b_n t)\right)\right] + o(n^{-k/2})\end{aligned} \tag{8}$$

for a suitable integer m_k. Now, we subdivide the proof into the four different cases.

First, notice that $P_1(b_n t) = \alpha_3(2b_n^2t^2+1)/6$ and $(d/dt)P_1(b_n t) = (2b_n^2\alpha_3 t)/3$. With $k = m_k = 1$, (8) consequently becomes

$$\begin{aligned}\frac{f_{T_n}(t)}{\varphi(t)} &= \left[1 + \frac{t^2}{2}(1-b_n^2)\right] \\ &\quad \times \left[b_n + n^{-1/2}\left(\frac{2}{3}b_n^2\alpha_3 t - b_n^2 t\left(\frac{\alpha_3}{6}(2b_n^2t^2+1)\right)\right)\right] + o(n^{-1/2}).\end{aligned} \tag{9}$$

Utilizing the first-order expansions (up to the $n^{-1/2}$ terms) for b_n and b_n^2 that we have reported in Lemma 1 in (9), we get

$$\sqrt{n}\left(\frac{f_{T_n}(t)}{\varphi(t)} - 1\right) = \frac{C_1}{2} + \frac{\alpha_3}{2}t - \frac{C_1}{2}t^2 - \frac{\alpha_3}{3}t^3 + o(1). \tag{10}$$

Hence, the assertion under part (i) of the theorem follows.

For $\alpha_3 = C_1 = 0$, we get that the summand corresponding to $i = 1$ in (8) vanishes. Therefore, we have to utilize $P_2(b_n t) = (\kappa/12 - 1/4)b_n^3t^3 - (\kappa+3)b_n t/4$ and $(d/dt)P_2(b_n t) = (\kappa-3)b_n^3t^2/4 - (\kappa+3)b_n/4$ in (8) and we obtain

$$\begin{aligned}n\left(\frac{f_{T_n}(t)}{\varphi(t)} - 1\right) &= \left(\frac{1}{4} - \frac{\kappa}{12}\right)b_n^5t^4 + \frac{nt^2}{2} \\ &\quad \times \left[\frac{\kappa b_n^3}{n} + (1-b_n^2)(b_n - \frac{(\kappa+3)b_n}{4n})\right] \\ &\quad + n\left(b_n - \frac{(\kappa+3)b_n}{4n} - 1\right) + o(1) \\ &= \frac{1}{4}\left(1 - \frac{\kappa}{3}\right)b_n^5t^4 \\ &\quad + \left(\frac{\kappa b_n^3}{2} - \frac{b_n n(b_n^2-1)}{2} + \frac{\kappa+3}{8}(b_n^3 - b_n)\right)t^2 \\ &\quad + \frac{1}{2}\left(2n(b_n-1) - \frac{(\kappa+3)b_n}{2}\right) + o(1).\end{aligned}$$

Making use of the second-order expansions for b_n and b_n^2 including the n^{-1} terms, we obtain

$$n\left(\frac{f_{T_n}(t)}{\varphi(t)}-1\right)=\frac{1}{4}\left(1-\frac{\kappa}{3}\right)t^4+\frac{\kappa-C_2}{2}t^2+\frac{1}{2}\left(C_2-\frac{\kappa+3}{2}\right)+o(1)$$

and hence the assertion of part (ii).

If $\alpha_3 = C_1 = 0$, $\kappa = C_2 = 3$, and $\alpha_5 < \infty$, we obtain $P_3(b_n t) = \alpha_5/5(-b_n^4 t^4/4 - b_n^2 t^2 - 1/8)$ and $(d/dt)P_3(b_n t) = \alpha_5/5(-b_n^4 t^3 - 2b_n^2 t)$. Utilizing these representations and the expansions for the norming sequences up to the $n^{-3/2}$ terms in (8) leads to

$$\lim_{n\to\infty}\left\{n^{3/2}\left(\frac{f_{T_n}(t)}{\varphi(t)}-1\right)\right\}=\frac{C_3}{2}-\frac{3\alpha_5}{8}t-\frac{C_3}{2}t^2+\frac{\alpha_5}{20}t^5,$$

and the assertion of part (iii) follows.

Finally, in case of $\alpha_3 = C_1 = 0$, $\kappa = C_2 = 3$, $\alpha_5 = C_3 = 0$ and $\alpha_6 < \infty$, we have

$$P_4(t)=t^5\left(\frac{3}{2}-\frac{\alpha_6}{45}\right)+\frac{t^3}{2}\left(\frac{\alpha_6}{9}-\frac{31}{4}\right)+\frac{t}{2}\left(\frac{\alpha_6}{3}-\frac{99}{4}\right),$$
$$\frac{d}{dt}P_4(t)=t^4\left(\frac{15}{2}-\frac{\alpha_6}{9}\right)+\frac{t^2}{2}\left(\frac{\alpha_6}{3}-\frac{93}{4}\right)+\frac{1}{2}\left(\frac{\alpha_6}{3}-\frac{99}{4}\right).$$

Plugging these expressions, together with the approximations for b_n and its powers up to the n^{-2} terms into (8), we obtain the target equation

$$\lim_{n\to\infty}\left\{n^{2}\left(\frac{f_{T_n}(t)}{\varphi(t)}-1\right)\right\}=\left(\frac{\alpha_6}{45}-\frac{3}{2}\right)t^6+\left(\frac{41}{4}-\frac{\alpha_6}{6}\right)t^4+\left(3-\frac{C_4}{2}\right)t^2$$
$$+\frac{\alpha_6}{6}+\frac{1}{2}\left(C_4-\frac{45}{2}\right),$$

completing the proof. ■

Remark 2. The Studentized sum S_n is a special case of the generalized self-normalized sum T_n for $C_1 = 0$, $C_2 = 1$, and $C_j = 0$ for all $j > 2$, and the standard self-normalized sum Y_n can be generated from T_n by setting $C_j = 0$ for all $j \geq 1$. Therefore, the ACP results in Theorem 1 can directly be carried over to Studentized and standard self-normalized sums. In this sense, Theorem 1 gives a comprehensive answer to the ACP question for t-type statistics.

Example 1. For three commonly used normalization sequences, we get in case of $\alpha_3 = \kappa = 0$ (as encountered if the X_i's are normally distributed) the following ACPs according to case (ii) of Theorem 1.

(a) For $a_n = 1/n$, we have $B_4 = 1/4$, $B_2 = 0$, $B_0 = -3/4$; hence, two ACPs are given by $\pm 3^{1/4}$.

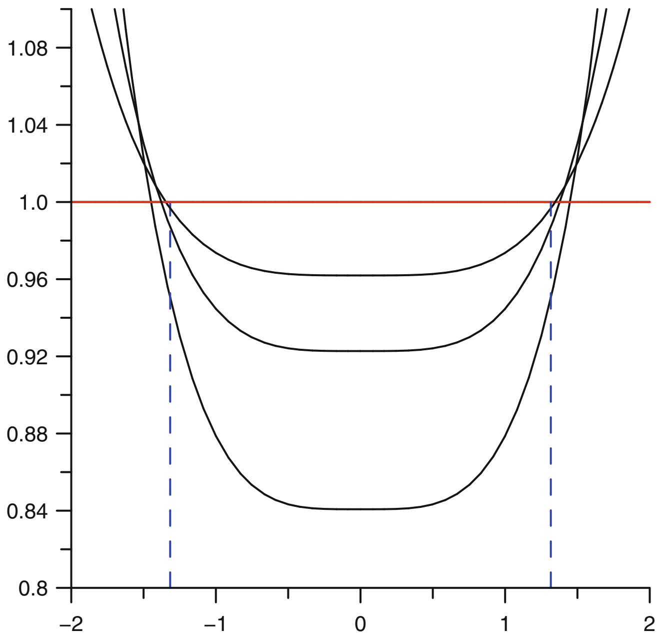


Fig. 1 Likelihood ratio $\lambda_n(z) = \sqrt{(n-1)/n}\, f_{n-1}(\sqrt{(n-1)/n} \cdot z)/\varphi(z)$ for $n = 5, 10, 20$, and $-2 \le z \le 2$, where f_{n-1} denotes the pdf of Student's t distribution with $n-1$ degrees of freedom. The curves can be identified by noticing that $\lambda_n(0)$ is increasing in n. Asymptotic crossing points (ACPs) are $\pm 3^{1/4}$ as pointed out in case (a) of Example 1 and indicated by the two *dashed vertical lines*

(b) For $a_n = 1/(n-1)$, we have $B_4 = 1/4$, $B_2 = -1/2$, $B_0 = -1/4$; hence, two ACPs are $\pm\sqrt{1+\sqrt{2}}$.

(c) For $a_n = (n-3)^{-1}$, we have $B_4 = 1/4$, $B_2 = -3/2$, $B_0 = 3/4$; hence, four ACPs are given by $\pm\sqrt{3 \pm \sqrt{6}}$. If the X_i's are normally distributed, this case corresponds to the standardized t-distribution with $\mathrm{Var}(T_n) = 1$ for all $n > 3$.

We may recall that cases (b) and (c) are the examples originally studied in [2]. For a plausibility check of case (a), we derived Fig. 1. In the case that the X_i's are normally distributed, the standard self-normalized sum Y_n has exactly the density $f_{Y_n}(z) = \sqrt{(n-1)/n}\, f_{n-1}(\sqrt{(n-1)/n} \cdot z)$ leading to the graphical representation of the likelihood ratio $\lambda_n(z)$ in Fig. 1. In case of non-normal X_i's, a closed-form representation of f_{Y_n} is hardly available, cf., e.g., [11].

Remark 3. Let us analyze the number of ACPs in case (ii) of Theorem 1. For $\kappa \neq 3$, the solutions of the equation $B_4 t^4 + B_2 t^2 + B_0 = 0$ can be expressed by

$$t = \pm \frac{\sqrt{(\kappa-3)(3(\kappa-3) \pm \sqrt{D(\kappa, C_2)})}}{\kappa - 3},$$

where $D(\kappa, C_2) = 3(3C_2^2 - 4C_2\kappa + 2\kappa^2 - 6C_2 + 9) > 0$. This representation shows that there exist either two or four ACPs, depending on the signs of $(\kappa - 3)$ $(3(\kappa - 3) \pm \sqrt{D(\kappa, C_2)})$.

In case of $\kappa = 3$ and $C_2 \neq 3$, we obtain $B_4 = 0$ and $B_2 = -B_0 \neq 0$; hence, there exist exactly two ACPs, namely, ±1.

Concluding Remarks

Although the classical Edgeworth expansion for Y_n (where the norming sequence is given by $a_n = 1/n$) reported in (2), together with asymptotic expansions for b_n and b_n^j for $j \geq 2$, is sufficient for the CP results in section "ACP Theorem for Generalized Self-Normalized Sums," we conclude this paper by relating our findings to a formal expansion for the generalized self-normalized sum T_n that we have derived in [9]. In the latter article, we computed the polynomials $\tilde{P}_i$ (say) appearing in an expansion of the form

$$F_{T_n}(t) = \Phi(t) + \sum_{i=1}^{k} n^{-i/2} \tilde{P}_i(t)\varphi(t) + o(n^{-k/2}). \tag{11}$$

Expansion (11) holds uniformly in $t \in \mathbb{R}$ under the same conditions as required for the expansion in (2). However, the coefficients of the polynomials $\tilde{P}_i$ in (11) depend not only on the cumulants of X_1 but additionally on the constants C_j appearing in the formal representation (5) of the norming sequence a_n. For instance, the first two $\tilde{P}_i$ are given by

$$\begin{aligned}\tilde{P}_1(t) &= \frac{\alpha_3 t^2}{3} + \frac{\alpha_3}{6} + \frac{C_1 t}{2},\\ \tilde{P}_2(t) &= \frac{3tC_1^2}{8} + \frac{\alpha_4 t^3}{12} + \frac{\alpha_3^2 t}{6} - \frac{t^3 C_1^2}{8} - \frac{\alpha_3^2 t^3}{9}\\ &\quad - \frac{\alpha_3^2 t^5}{18} + \frac{\alpha_3 C_1 t^2}{4} + \frac{tC_2}{2} - \frac{t^3}{2} - \frac{\alpha_3 C_1 t^4}{6} - \frac{\alpha_4 t}{4}.\end{aligned}$$

Taking formal derivatives in (11), we obtain

$$f_{T_n}(t) = \varphi(t) + \sum_{i=1}^{k} n^{-i/2} \left[\frac{d}{dt}\tilde{P}_i(t)\varphi(t) + \tilde{P}_i(t)\frac{d}{dt}\varphi(t)\right] + o(n^{-k/2})$$

or, equivalently,

$$\frac{f_{T_n}(t)}{\varphi(t)} - 1 = \sum_{i=1}^{k} n^{-i/2} \left[\frac{d}{dt}\tilde{P}_i(t) - t\tilde{P}_i(t)\right] + o(n^{-k/2}).$$

Therefore, the CP results in Theorem 1 can also be derived by solving the equations

$$\frac{d}{dt}\tilde{P}_i(t) - t\tilde{P}_i(t) = 0, \ i = 1, \ldots, 4.$$

We double-checked our results from Theorem 1 by making use of this connection.

References

1. Finner, H., Roters, M., Dickhaus, T.: Characterizing density crossing points, extended online version (2007). Available via http://www.helmut-finner.de/Density_Crossing_Points.pdf
2. Finner, H., Roters, M., Dickhaus, T.: Characterizing density crossing points. Am. Stat. **61**(1), 28–33 (2007)
3. Student: The probable error of a mean. Biometrika **6**, 1–25 (1908)
4. Hall, P.: The Bootstrap and Edgeworth Expansion. Springer Series in Statistics. Springer, New York (1992)
5. Hsu, P.: The approximate distributions of the mean and variance of a sample of independent variables. Ann. Math. Stat. **16**, 1–29 (1945)
6. Chung, K.-L.: The approximate distribution of Student's statistic. Ann. Math. Stat. **17**, 447–465 (1946)
7. Hall, P.: Edgeworth expansion for Student's t statistic under minimal moment conditions. Ann. Probab. **15**, 920–931 (1987)
8. Kabaila, P.: A method for the computer calculation of edgeworth expansions for smooth function models. J. Comput. Graph. Stat. **2**(2), 199–207 (1993)
9. Finner, H., Dickhaus, T.: Edgeworth expansion for normalized sums: Chung's 1946 Method Revisited. Stat. Probab. Lett. **80**(23–24), 1875–1880 (2010)
10. Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses, 3rd edn. Springer Texts in Statistics. Springer, New York (2005)
11. Sansing, R., Owen, D.: The density of the t-statistic for non-normal distributions. Commun. Stat. **3**, 139–155 (1974)

Exponential Ratio-Cum-Exponential Dual to Ratio Estimator in Double Sampling

Diganta Kalita, B.K. Singh, and Sanjib Choudhury

Abstract A class of exponential ratio-cum-exponential dual to ratio estimators for estimating a finite population mean in double sampling scheme is proposed. The expressions for bias and mean squared error (MSE) of the proposed estimator have been derived for two different cases. An asymptotic expression for MSE is obtained. Empirical studies are carried out to illustrate the performance of the constructed estimator over other estimators.

Keywords Exponential ratio-cum-dual to ratio estimator • Bias • Mean squared error • Auxiliary information • Double sampling

Introduction

It is well established that the use of auxiliary variable x at the estimation stage improves the precision of an estimate of the population mean of a character y under study. When the correlation between study variable y and auxiliary variable x is highly positive, the classical ratio estimator [1] is considered to be most practicable. The product estimator of [2] and then by [3] is employed quite effectively in the case of high negative correlation between study variable y and auxiliary variable x. Further, if the relation between y and x is a straight line passing through the

D. Kalita (✉)
Department of Statistics, North Lakhimpur College (Autonomous), North Lakhimpur, Assam 787001, India
e-mail: dkalita.nl@gmail.com

B.K. Singh
Department of Mathematics, North Eastern Regional Institute of Science and Technology, Nirjuli, Itanagar 791109, India
e-mail: bksinghnerist@gmail.com

S. Choudhury
Department of Mathematics, National Institute of Technology Nagaland, Nagaland 797103, India
e-mail: sanjibchy07@gmail.com

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.05_18

neighbourhood of the origin and the variance of y about this line is proportional to auxiliary variable x, the ratio estimator is as good as regression estimator.

Let us consider a finite population $U = \{u_1, u_2, \ldots, u_N\}$ of size N units and the value of the variables on the ith unit u_i $(i = 1, 2, \ldots, N)$, be (y_i, x_i). Let $\overline{Y} = \sum_{i=1}^{N} \frac{y_i}{N}$ and $\overline{X} = \sum_{i=1}^{N} \frac{x_i}{N}$ be the population means of the study variable y and the auxiliary variable x, respectively.

For estimating the population mean $\overline{Y}$ of y, a simple random sample of size n is drawn without replacement from the population U. Reference [4] proposed an exponential ratio and product estimators, respectively, as

$$\overline{y}_{\text{Re}} = \overline{y}\ \exp\left(\frac{\overline{X} - \overline{x}}{\overline{X} + \overline{x}}\right)$$

and

$$\text{and}\ \overline{y}_{Pe} = \overline{y} \exp\left(\frac{\overline{x} - \overline{X}}{\overline{x} + \overline{X}}\right).$$

If the population mean $\overline{X}$ of the auxiliary variable x is not known before start of the survey, a first-phase sample of size n_1 is drawn from the population, on which only the auxiliary variable x is observed. Then, a second-phase sample of size n is drawn, on which both study variable y and auxiliary variable x are observed. Let $\overline{x}_1 = \sum_{i=1}^{n_1} \frac{x_i}{n_1}$ denote the sample mean of size n_1 based on the first-phase sample and $\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$ and $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ denote the sample means of variables y and x, respectively, obtained from the second-phase sample of size n. The double sampling version of ratio and product estimators of population mean $\overline{Y}$ are respectively given by

$$\overline{y}_R^d = \overline{y}\frac{\overline{x}_1}{\overline{x}} \quad \text{and} \quad \overline{y}_p^d = \overline{y}\frac{\overline{x}}{\overline{x}_1}.$$

Reference [5] suggested an exponential ratio and product estimators for $\overline{Y}$ in double sampling as

$$\overline{y}_{\text{Re}}^d = \overline{y}\ \exp\left(\frac{\overline{x}_1 - \overline{x}}{\overline{x}_1 + \overline{x}}\right)\ \text{and}$$

$$\overline{y}_{Pe}^d = \overline{y}\ \exp\left(\frac{\overline{x} - \overline{x}_1}{\overline{x} + \overline{x}_1}\right), \text{respectively.}$$

Let $x_i^{*d} = (1 + g')\,\overline{x}_1 - g'x_i$, $i = 1, 2, \ldots, N$, where $g' = n/(n_1 - n)$. Then $\overline{x}^{*d} = (1 + g')\,\overline{x}_1 - g'\overline{x}$ is an unbiased estimator of $\overline{X}$ and $corr\left(\overline{y}, \overline{x}^{*d}\right) = -ve$. Using this transformation, Ref. [6] suggested dual to ratio estimator in double sampling as

$$\overline{y}_R^{*d} = \overline{y}\frac{\overline{x}^{*d}}{\overline{x}_1}.$$

Utilizing the transformation x_i^{*d}, the exponential ratio and product estimators of [5] in double sampling are converted to exponential dual to ratio and product estimators, respectively, as

$$\overline{y}_{\text{Re}}^{*d} = \overline{y}\ \exp\left(\frac{\overline{x}^{*d} - \overline{x}_1}{\overline{x}^{*d} + \overline{x}_1}\right)$$

$$\text{and}\quad \overline{y}_{Pe}^{*d} = \overline{y}\ \exp\left(\frac{\overline{x}_1 - \overline{x}^{*d}}{\overline{x}_1 + \overline{x}^{*d}}\right).$$

In this paper, we have studied the properties of the class of estimators of the linear combination of exponential ratio and exponential dual to ratio estimators in double sampling. Numerical illustrations are given in the support of the present study.

The Proposed Estimator

We suggest the following exponential ratio-cum-exponential dual to ratio estimators for $\overline{Y}$ in double sampling as

$$t = \overline{y}\left\{\alpha\ \exp\left(\frac{\overline{x}_1 - \overline{x}}{\overline{x}_1 + \overline{x}}\right) + \beta\ \exp\left(\frac{\overline{x}^{*d} - \overline{x}_1}{\overline{x}^{*d} + \overline{x}_1}\right)\right\} \tag{1}$$

where α and β are unknown constants such that $\alpha + \beta = 1$.

Remarks

1. For $(\alpha, \beta) = (1, 0)$, the estimator t reduces to the 'exponential ratio estimator in double sampling' $\left(\overline{y}_{\text{Re}}^{d}\right)$ with its properties.
2. For $(\alpha, \beta) = (0, 1)$, the estimator t reduces to the 'exponential dual to ratio estimator' $\left(\overline{y}_{\text{Re}}^{*d}\right)$ in double sampling with its properties.

For the bias and MSE of the proposed estimators, the following two cases are considered:

Case I: When the second-phase sample of size n is a subsample of the first-phase sample of size n_1

Case II: When the second-phase sample of size n is drawn independently of the first-phase sample of size n_1

Case I: To obtain bias (B) and MSE (M) of the proposed estimator t, we write

$$\begin{aligned} &e_0 = \left(\overline{y} - \overline{Y}\right)/\overline{Y}, e_1 = (\overline{x} - \overline{X})/\overline{X}, \\ &e'_1 = \left(\overline{x}_1 - \overline{X}\right)/\overline{X} \ \textit{and} \end{aligned}$$

$$\left.\begin{gathered} E\left(e_0\right) = E\left(e_1\right) = E\left(e'_1\right) = 0 \\ E\left(e_0^2\right) = fC_y^2,\ \ E\left(e_1^2\right) = fC_x^2,\ \ E\left(e'^2_1\right) = f_1 C_x^2 \\ E\left(e_0 e_1\right) = fC_x^2 C_{yx},\ \ E\left(e_0 e'_1\right) = f_1 C_x^2 C_{yx}, \\ E\left(e_1 e'_1\right) = f_1 C_x^2 \end{gathered}\right\} \tag{2}$$

where

$$f = \frac{1}{n} - \frac{1}{N},\ \ f_1 = \frac{1}{n_1} - \frac{1}{N},\ \ C_y = \frac{S_y}{\overline{Y}},\ \ C_x = \frac{S_x}{\overline{X}},$$

$$C_{yx} = \rho_{yx}\frac{C_y}{C_x},\ \ \rho_{yx} = \frac{S_{yx}}{S_y S_x}$$

$$\begin{gathered} S_y^2 = \tfrac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2,\ \ S_x^2 = \tfrac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2 \text{ and} \\ S_{yx} = \tfrac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)\left(x_i - \overline{X}\right). \end{gathered}$$

Expressing t in terms of e's and retaining terms up to second powers of e's, we have

$$\begin{aligned} &\left(t - \overline{Y}\right) \cong \overline{Y}\Big[e_0 + \frac{1}{2}\left\{\alpha\left(e'_1 - e_1 + e_0 e'_1 - e_0 e_1\right) + \right\} g'\left(1 - \alpha\right) \\ &\left(e'_1 - e_1 + e_0 e'_1 - e_0 e_1 + e_1 e'_1 - e'^2_1\right) - \frac{1}{4}\Big\{\alpha\left(e_1 e'_1 + e'^2_1 - e_1^2\right) \\ &- g'^2\left(1 - \alpha\right)\left(e_1 e'_1 - e'^2_1 - e_1^2\right)\Big\} + \frac{1}{8}\Big\{\ \alpha\left(e'^2_1 + e_1^2\right) \\ &+ \left(1 - \alpha\right) g'^2 \left(e'^2_1 + e_1^2\right)\Big\}\Big]. \end{aligned} \tag{3}$$

Therefore, the bias of the estimator t can be obtained by using the results of (2) in (3) as

$$B(t)_I = \overline{Y}\left[\frac{1}{8} f^* \left\{\alpha + g'^2 (1-\alpha)\right\} - \frac{1}{4}\left\{\alpha f^{**} + (1-\alpha) f g'^2\right\} \right.$$
$$\left. - \frac{1}{2}\left\{\alpha + (1-\alpha) g'\right\} f_2 C_{yx}\right] C_x^2$$

where $f_2 = \frac{1}{n} - \frac{1}{n_1}$, $f^* = f + f_1$, $f^{**} = f_1 - f_2$.

From (3), we can write

$$\left(t - \overline{Y}\right) \cong \overline{Y}\left[e_0 + \frac{1}{2}\left\{\alpha\left(e_1' - e_1\right) + g'(1-\alpha)\left(e_1' - e_1\right)\right\}\right] \tag{4}$$

Squaring both the sides of (4), taking expectations of its terms and using the results of (2), we get the MSE of t to the first-degree approximation as

$$M(t)_I = \overline{Y}^2\left[f C_y^2 + \frac{1}{4}\left\{\alpha + g'(1-\alpha)\right\}^2 f_2 C_x^2 - \right.$$
$$\left. \left\{\alpha + g'(1-\alpha)\right\} f_2 \rho_{yx} C_y C_x\right]. \tag{5}$$

Optimization of (5) with respect to α gives its optimum value as

$$\alpha = \frac{2\rho_{yx} C_y}{(1-g') C_x} - \frac{g'}{1-g'} = \alpha_{opt.I}. \tag{6}$$

Substituting the value of $\alpha_{opt.\,I}$ from (6) in (5), we get the asymptotic optimum MSE of t as

$$opt.M(t)_I = \overline{Y}^2 C_y^2\left(f - f_2 \rho_{yx}^2\right). \tag{7}$$

Efficiency Comparisons

Efficiency comparisons of asymptotic optimum MSE of proposed class of estimators t

1. *With sample mean per unit estimator* $\overline{y}$

   The MSE of sample mean per unit estimator is given by

$$M(\overline{y}) = \overline{Y}^2 f C_y^2 \tag{8}$$

From (7) and (8), it is found that the proposed class of estimators is more efficient than $\overline{y}$, since

$$M\left(\overline{y}\right) - opt.M(t)_I = \overline{Y}^2 f_2 C_y^2 \rho_{yx}^2 > 0.$$

2. *With usual ratio estimator in double sampling*
To compare with the usual ratio estimator $\overline{y}_R^d$ in double sampling, we write the MSE of $\overline{y}_R^d$ to the first degree of approximation as

$$M\left(\overline{y}_R^d\right)_I = \overline{Y}^2 \left(f C_y^2 + f_2 C_x^2 - 2 f_2 \rho_{yx} C_y C_x\right) \tag{9}$$

We note from (7) and (9) that the proposed estimator has smaller MSE than that of the usual ratio estimator $\overline{y}_R^d$ in double sampling, since

$$M\left(\overline{y}_R^d\right)_I - opt.M(t)_I = \overline{Y}^2 f_2 \left(C_x - \rho_{yx} C_y\right)^2 > 0.$$

3. *With exponential ratio estimator in double sampling*
The MSE of exponential ratio estimator in double sampling to the first degree of approximation is given as

$$M\left(\overline{y}_{\mathrm{Re}}^d\right)_I = \overline{Y}^2 \left(f C_y^2 + \frac{1}{4} f_2 C_x^2 - f_2 \rho_{yx} C_y C_x\right) \tag{10}$$

From (7) and (10), it is found that the proposed estimator is superior to estimator $\overline{y}_{\mathrm{Re}}^d$, since $M\left(\overline{y}_{\mathrm{Re}}^d\right)_I - opt.M(t)_I = \overline{Y}^2 f_2 \left(\frac{1}{2} C_x - \rho_{yx} C_y\right)^2 > 0$.
4. *With dual to ratio estimator in double sampling*
The MSE of dual to ratio estimator in double sampling to the first degree of approximation is given as

$$M\left(\overline{y}^{*d}_{\mathrm{R}}\right)_I = \overline{Y}^2 \left(f C_y^2 + g'^2 f_2 C_x^2 - 2 g' f_2 \rho_{yx} C_y C_x\right) \tag{11}$$

From (7) and (11), it is obtained that the proposed estimator is better than $\overline{y}_{\mathrm{R}}^{*d}$, since

$$M\left(\overline{y}_{\mathrm{R}}^{*d}\right)_I - opt.M(t)_I = \overline{Y}^2 f_2 \left(\rho_{yx} C_y - g' C_x\right)^2 > 0.$$

5. *With exponential dual to ratio estimator in double sampling*
The MSE of exponential dual to ratio estimator in double sampling is given by

$$M\left(\overline{y}_{\mathrm{Re}}^{*d}\right)_I = \overline{Y}^2 \left(f C_y^2 + \frac{1}{4} g'^2 f_2 C_x^2 - g' f_2 \rho_{yx} C_y C_x\right). \tag{12}$$

From (7) and (12), it is found that the proposed estimator is more efficient than $\overline{y}^{*d}_{\mathrm{Re}}$, since

$$M\left(\overline{y}^{*d}_{\mathrm{Re}}\right)_I - opt.M(t)_I = \overline{Y}^2 f_2\left(\frac{1}{2}g'C_x + \rho_{yx}C_y\right)^2 > 0.$$

From the above results, it is found that the proposed class of estimators has shown better efficiency over others in case of its optimality.

Case II

To obtain bias and MSE of the proposed estimator t, we have

$$\left.\begin{array}{l} E\left(e_0\right) = E\left(e_1\right) = E\left(e'_1\right) = 0 \\ E\left(e_0^2\right) = fC_y^2,\ \ E\left(e_1^2\right) = fC_x^2,\ \ E\left(e'^2_1\right) = f_1C_x^2, \\ E\left(e_0e_1\right) = fC_x^2C_{yx},\ \ E\left(e_0e'_1\right) = E\left(e_1e'_1\right) = 0 \end{array}\right\}. \tag{13}$$

Taking expectations in (3) and using the results of (13), we get the approximate bias of t to the first-degree approximation as

$$B(t)_{II} = \left[\frac{1}{8}\left\{\alpha + g'^2\left(1-\alpha\right)\right\} f^* - \frac{1}{2}\left\{\alpha + g'\left(1-\alpha\right)\right\}\right.$$
$$\left. fC_{yx} - \frac{1}{2}g'\left(1-\alpha\right) f_1 - \frac{1}{4}\left\{g'^2\left(1-\alpha\right) f^*C_x^2 - \alpha f_2\right\}\right] C_x^2.$$

Squaring both the sides of (4), taking expectations of its terms and using the results of (13), we get the MSE of t to the first-degree approximation as

$$M(t)_{II} = \overline{Y}^2\Big[fC_y^2 + \frac{1}{4}\left\{\alpha + g'\left(1-\alpha\right)\right\}^2 f^*C_x^2 -$$
$$\left\{\alpha + g'\left(1-\alpha\right)\right\} f\rho_{yx}C_yC_x\Big]. \tag{14}$$

Optimizing (14) with respect to α, we get its optimum value as

$$\alpha = \frac{2f\rho_{yx}C_y}{f^*\left(1-g'\right)C_x} - \frac{g'}{1-g'} = \alpha_{opt.II} \quad \text{(say)}. \tag{15}$$

Substituting the value of $\alpha_{opt.\,II}$ from (15) in (14), we get the asymptotic optimum MSE of t as

$$opt.M(t)_{II} = \overline{Y}^2 fC_y^2\left(1 - \frac{f\rho_{yx}^2}{f^*}\right). \tag{16}$$

Efficiency Comparisons

Efficiency comparisons of asymptotic optimum MSE of the proposed class of estimators t

1. *With sample mean per unit estimator* $\overline{y}$
   From (8) and (16), it is found that the proposed estimator is more efficient than $\overline{y}$, since

$$M\left(\overline{y}\right) - opt.M(t)_{II} = \overline{Y}^2 \frac{\left(fC_y\rho_{yx}\right)^2}{f^*} > 0.$$

2. *With usual ratio estimator in double sampling*
   The MSE of ratio estimator in double sampling is given by

$$M\left(\overline{y}_R^d\right)_{II} = \overline{Y}^2\left(fC_y^2 + f^*C_x^2 - 2f\rho_{yx}C_yC_x\right) \tag{17}$$

   From (16) and (17), it is obtained that the proposed estimator is better than $\overline{y}_R^{(d)}$, since

$$M\left(\overline{y}_R^d\right)_{II} - opt.M(t)_{II} = \frac{\overline{Y}^2}{f^*}\left(f^*C_x - f\rho_{yx}C_y\right)^2 > 0.$$

3. *With exponential ratio estimator in double sampling*
   The MSE of exponential ratio estimator in double sampling is given by

$$M\left(\overline{y}_{\mathrm{Re}}^d\right)_{II} = \overline{Y}^2\left(fC_y^2 + \frac{1}{4}f^*C_x^2 - f\rho_{yx}C_yC_x\right). \tag{18}$$

   We note from (16) and (18) that the proposed estimator has smaller MSE than that of $\overline{y}_{\mathrm{Re}}^d$, since

$$M\left(\overline{y}_{\mathrm{Re}}^d\right)_{II} - opt.M(t)_{II} = \frac{\overline{Y}^2}{f^*}\left(\frac{1}{2}f^*C_x - f\rho_{yx}C_y\right)^2 > 0.$$

4. *With dual to ratio estimator in double sampling*
   The MSE of the dual to ratio estimator in double sampling is given by

$$M\left(\overline{y}_{\mathrm{R}}^{*d}\right)_{II} = \overline{Y}^2\left(fC_y^2 + g'^2f^*C_x^2 - 2g'f\rho_{yx}C_yC_x\right). \tag{19}$$

From (16) and (19), we have

$$M\left(\overline{y}_{\mathrm{R}}^{*d}\right)_{II} - opt.M(t)_{II} = \frac{\overline{Y}^2}{f^*}\left(g' f^* C_x - f\rho_{yx} C_y\right)^2 > 0.$$

5. *With exponential dual to ratio estimator in double sampling*
The MSE of the exponential dual to ratio estimator in double sampling is given by

$$M\left(\overline{y}_{\mathrm{Re}}^{*d}\right)_{II} = \overline{Y}^2\left(fC_y^2 + \frac{1}{4}g'^2 f^* C_x^2 - g' f\rho_{yx} C_y C_x\right). \tag{20}$$

From (16) and (20), we found that the estimator t is more efficient than $\overline{y}_{\mathrm{Re}}^{*d}$, since

$$M\left(\overline{y}_{\mathrm{Re}}^{*d}\right)_{II} - opt.M(t)_{II} = \frac{\overline{Y}^2}{f^*}\left(\frac{1}{2}g' f^* C_x - f\rho_{yx} C_y\right)^2 > 0.$$

From the above results, it is found that the proposed class of estimators has shown better efficiency over others in case of its optimality.

Empirical Study

To examine the merits of the proposed estimator, we have considered four natural populations data sets. The descriptions of the populations are given below:

Population I: Source: ([7], p. 228)

X: Fixed capital, Y: Output, $N = 80$, n $= 10$, $n_1 = 30$, $\overline{Y} = 5182.64$, $\rho_{yx} = 0.9413$, $C_Y = 0.3542$, $C_X = 0.7507$

Population II: Source: ([7], p. 228)

X: Number of workers, Y: Output, $N = 80$, $n = 10$, $n_1 = 30$, $\overline{Y} = 5182.64$, $\rho_{yx} = 0.9150$, $C_Y = 0.3542$, $C_X = 0.9484$

Population III: Source: [8]

X: Number of agricultural labourers for 1961, Y: Number of agricultural labourers for 1971, $N = 278$, $n = 30$, $n_1 = 70$, $\overline{Y} = 39.0680$, $\rho_{yx} = 0.7213$, $C_Y = 1.4451$, $C_X = 1.6198$

Population IV: Source: ([6], p. 324)

X: Population of village, Y: Number of cultivators in the village, $N = 487$, $n = 20$, $n_1 = 95$, $\overline{Y} = 449.846$, $\rho_{yx} = 0.881815$, $C_y = 0.8871$, $C_X = 0.7696$

To observe the relative performance of different estimators of $\overline{Y}$, we have computed the percentage relative efficiencies of the proposed estimator t, conventional ratio, dual to ratio, exponential ratio and exponential dual to ratio estimators in double sampling and sample mean per unit estimator $\overline{y}$ with respect to usual unbiased estimator $\overline{y}$. The findings are presented in Tables 1 and 2.

Table 1 Percentage relative efficiencies of different estimators w.r.t. $\overline{y}$ for Case I

Estimators	$\overline{y}$	$\overline{y}_R^d$	$\overline{y}_R^{*d}$	$\overline{y}_{Re}^d$	$\overline{y}_{Re}^{*d}$	t
Population I	100.00	72.36	297.97	297.97	220.31	307.77
Population II	100.00	36.64	200.42	200.42	245.05	276.16
Population III	100.00	130.04	147.96	146.35	137.98	149.98
Population IV	100.00	277.79	141.21	190.45	118.62	277.92

Table 2 Percentage relative efficiencies of different estimators w.r.t. $\overline{y}$ for Case II

Estimators	$\overline{y}$	$\overline{y}_R^d$	$\overline{y}_R^{*d}$	$\overline{y}_{Re}^d$	$\overline{y}_{Re}^{*d}$	t
Population I	100.00	38.89	252.94	252.89	285.64	351.68
Population II	100.00	20.09	130.02	130.02	303.23	308.85
Population III	100.00	91.66	133.69	161.68	157.77	162.00
Population IV	100.00	281.21	152.68	219.11	123.19	294.83

Conclusions

From Tables 1 and 2, it is evident that the proposed estimator t is more efficient than all other estimators considered in this paper. Thus, the uses of the proposed estimators are preferable over other estimators.

References

1. Cochran, W.G.: Sampling theory when the sampling-units are of unequal sizes. J. Am. Stat. Assoc. **37**, 199–212 (1942)
2. Robson, D.S.: Applications of multivariate polykays to the theory of unbiased ratio-type estimation. J. Am. Stat. Assoc. **52**, 511–522 (1957)
3. Murthy, M.N.: Product method of estimation. Sankhya A **26**, 69–74 (1964)
4. Bahl, S., Tuteja, R.K.: Ratio and product type exponential estimator. Inf. Optim. Sci. **12**, 159–163 (1991)
5. Singh, H.P., Vishwakarma, G.K.: Modified exponential ratio and product estimators for finite population mean in double sampling. Aust. J. Stat. **36**(3), 217–225 (2007)
6. Kumar, M., Bahl, S.: A class of dual to ratio estimator in double sampling. Stat. Pap. **47**, 319–326 (2006)
7. Murthy, M.N.: Sampling theory and methods. Statistical Publishing Soc, Calcutta, India (1967)
8. Das, A.K.: Contribution to the theory of sampling strategies based on auxiliary information. Unpublished Ph.D. thesis, Bidhan Chandra Krishi Vishwavidyalaya, India (1988)

Analysis of Performance of Indices for Indian Mutual Funds

Rahul Ritesh

Abstract The objective of the study is to analyze performance of the equity diversified mutual fund on the basis of risk-adjusted return over the last 3 years. Future predictions have been made based on this obtained data, which is then matched with the actual values of average return and the return by the buy-and-hold strategy of the investor over the next 1, 2, and 3 years using regression techniques. To get better results, we use bootstrapping and then check the results again. The indices used are Sharpe ratio, Treynor ratio, coefficient of variation, and information ratio, with RBI treasury bill rate as the risk-free rate. The results obtained indicate that most of the indices do not work well in Indian markets, and so there is a need to change the formulae to suit our needs.

Keywords Indian mutual funds • Sharpe ratio • Treynor ratio • Risk-adjusted return • Performance of mutual fund indices • Information ratio

Introduction

Mutual fund is one of the greatest innovations seen in modern financial markets serving the interest of the investors, both with huge and small investment. The introduction of theory of portfolio management by Harry Markowitz in 1952 has revolutionized the concept of managing portfolio of securities including stocks and bonds which has continuously seen an improvement of the fund market and has yielded higher returns than its benchmark. The concept of eliminating unsystematic risk by diversifying portfolio by way of inclusion of negatively correlated securities has gained ground and resulted in beating the market portfolio.

In India, the first mutual fund was introduced in the year 1963 when the Unit Trust of India was formed as a joint initiative of Government of India and Reserve Bank of India. Subsequently the entry of private sector in the industry and the introduction of more comprehensive regulation by Securities and Exchange Board

R. Ritesh (✉)
Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati, India
e-mail: r.ritesh@alumni.iitg.ernet.in

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.21_19

of India in 1996 paved the way for the proper development of the industry. The Indian mutual fund has given a spectacular return in the recent years which had fascinated the interest of many investors. Following the entry of foreign players, the market was flooded with a variety of schemes, and also the increasing investor's awareness about this financial instrument has definitely given a great threat to the growth of this industry.

With so many funds in the market, it has become very essential to identify the truly worthy ones for measuring the performance of these funds, and to realize their strategy to form a higher earning portfolio, a lot of theories have been developed, and many significant researches have been carried out in different parts of the world. But over the time, it has been modified by including other factors such as skewness and kurtosis, and many other ratios such as information, appraisal, Sortino, and Omega-Sharpe have been evolved during the time based on the similar idea of risk-adjusted return and taking the beta of the portfolio into account.

Ironically, the glorious past performance of any fund house does not give us a guarantee for higher returns in the future. It is very evident that during the boom phase, everyone goes up and makes huge profit, but the real testament of their managerial capability can be observed during the bear market. Future prediction of the fund performance is very difficult, but some researchers believe that Treynor ratio gives a better reflection of the future performance of any fund but not a guarantee whatsoever. So analysts all over the world calculate all kinds of popular ratios, and some of them such as Morningstar, Lipper, and Value Research have even developed their own ratios, and it is over the investor to choose the one which suits them the best.

Literature Reviews

- Carhart [1], Grinblatt and Titman [2], and Stutzer [3], among others, provided evidence that confirmed the predictability of future performance by past performance, i.e., returns persists.
- Elton et al. [4] reported that the past carries the information about the future performance of the funds, and they highlighted that both the 1- and 3-year alphas convey the future information.
- Treynor and Majuy [5], on the contrary, observed that the fund managers who can beat the benchmark for the long run surprisingly still have significant probability of underperforming (extending to a long period of time).
- A sample of 50 Indian mutual funds was analyzed over 26 months by Bhattacharjee and Roy [6]. They found that in short term, the mutual funds were able to generate above normal return.
- Deb et al. [7] observed that, on an average, fund managers have not been able to beat their style benchmark.

Data and Methodologies

The data used in this study cover the period from April 1994 to April 2012 which also cover the different bull and bear phase of the Indian capital market. The monthly closing net asset value (NAV) data of 242 equity diversified mutual funds have been collected from the website of AMFI (Association of Mutual Funds in India) and also from the smctradeonline.com. For the final analysis, only 169 mutual funds have been taken into account for which the data was available for at least 72 months, i.e., the funds which was launched on or before March 2008. The risk-adjusted return for each fund has been calculated against its respective benchmark and the monthly closing price data of each index is collected from the websites of BSE (Bombay Stock Exchange) and NSE (National Stock Exchange). The risk-free rate of return has been taken as the return on the treasury bills. The monthly data is available on the website of RBI (Reserve Bank of India) under the Database of Indian Economy.

We now define a few variables that have been used in our calculations:

Net Asset Value

NAV is a mutual fund's price per share. It is also used for defining exchange-traded fund's share price. In both cases, the per share amount of the fund is calculated by dividing the total value of all the securities in its portfolio, less any liabilities, divided by the number of fund shares outstanding. In the context of mutual funds, NAV per share is computed once a day based on the closing market prices of the securities in the fund's portfolio. All mutual funds' buy and sell orders are processed at the NAV of the trade date. However, investors must wait until the following day to get the trade price. Mutual funds pay out virtually all of their income and capital gains. As a result, changes in NAV are not the best gauge of mutual fund performance, which is best measured by annual total return.

Calculation of Returns

The monthly return of each fund has been calculated from the monthly closing NAV by using the following formula:

$$R_t = \frac{NAV_{t+1} - NAV_t}{NAV_t}$$

where $R_t =$ return of asset over month t
$NAV_{t+1} =$ closing NAV of the month $t + 1$
$NAV_t =$ closing NAV of the month t

Similarly, the monthly return of all the indices has been calculated using the following formula:

$$R_{B_t} = \frac{P_{t+1} - P_t}{P_t}$$

R_{B_t} = return of benchmark over the month t
P_{t+1} = closing price of month $t + 1$
P_t = closing price of month t

Also the annualized return of the treasury bill has been converted into the monthly return for the further calculation of the risk-adjusted return.

Indices for Calculating Mutual Fund Performance

Using the monthly returns of each mutual fund, the following ratios are calculated for the purpose of evaluating risk-adjusted performance:

1. *Sharpe ratio*
   The most commonly used measure of risk-adjusted performance is the Sharpe ratio [8] which measures the fund's excess return per unit of its risk. It can be expressed as following:

$$\frac{E\left[R_p - R_f\right]}{\sigma} = \frac{E\left[R_p - R_f\right]}{\sqrt{var\left[R_p - R_f\right]}}$$

   where R_p = return of the portfolio
   R_f = risk-free rate of return
   σ = standard deviation of the excess portfolio return

   This ratio is based on the trade-off between risk and return. A high Sharpe ratio means that the fund delivers a huge return for its level of volatility. It also allows a direct comparison of the risk-adjusted performance of any two mutual funds, regardless of their volatilities and their correlation with the benchmark. The principal advantage of this ratio is that it is directly computable from any observed series of the returns without need for additional information surrounding the source of return.
2. *Treynor ratio*
   We know that security market line represents the expected total return of every security or portfolio i as a linear function of the return of the market portfolio m:

$$E_i = R_f + \beta_i\left[E_m - R_f\right]$$

where $E_i = E[R_i]$ is the unconditional continuous expected return of the portfolio i, R_f denotes the continuous expected return on the risk-free security, and $\beta_i = cov(R_i, R_m)$ is the beta of the portfolio i.

This equilibrium relation corresponds to the market model $r_{it} = \alpha_i + \beta_i r_m + \epsilon_{it}$, where $r_i = R_i - R_f$ denotes the excess return on the portfolio i.

The Treynor ratio is the ratio of Jensen's alpha over the stock beta and can be expressed as following:

$$TR = \frac{\alpha}{\beta}$$

3. *Information ratio*

The information ratio [9] is a measure of the risk-adjusted return of a financial security (or asset or portfolio). It is also known as appraisal ratio and is defined as expected active return divided by tracking error, where active return is the difference between the return of the security and the return of a selected benchmark index, and tracking error is the standard deviation of the active return. It can be expressed as following:

$$IR = \frac{\alpha}{\sigma(\epsilon)}$$

4. *Coefficient of variation*

In probability theory and statistics, the coefficient of variation (CV) is a normalized measure of dispersion of a probability distribution. It is also known as unitized risk or the variation coefficient. The absolute value of the CV is sometimes known as relative standard deviation (RSD), which is expressed as a percentage. It is defined as the difference of the return of the portfolio per unit risk of the portfolio and the return of the respective benchmark per unit the market risk.

$$\frac{R_p}{\sigma_p} - \frac{R_b}{\sigma_b}$$

where R_p = return of the portfolio
R_b = return of the benchmark
σ_p = standard deviation of the return of portfolio
σ_b = standard deviation of the return of benchmark

Initial Observations

The performance measures such as Sharpe, IR, Treynor, and CV for each scheme have been calculated using the data of the past 36 months, and this is rolled up to

March 2008. The future performance measures were computed for the next 1, 2, and 3 years using the following indicators:

1. $R_p - R_f$
2. $R_p - R_b$
3. $Buy - and - hold\ return\ over\ its\ benchmark$

All the mutual funds of the same type were taken together and regressed for a period of 36 months. The ratios, then obtained, were matched with the next 1-, 2-, and 3-year data to see the effectiveness of the indices by using regression method. The results obtained have been described using the following factors:

1. p-value: The p-value or calculated probability is the estimated probability of rejecting the null hypothesis of a study question when that hypothesis is true.

   The null hypothesis is usually a hypothesis of "no difference," e.g., no difference between blood pressures in group A and group B.

   We refer to statistically significant as $p < 0.05$, i.e., less than five in a hundred chance of being wrong. A regression which has a *p-value* more than 0.05 is considered to be nonsignificant for our evaluation.
2. R^2: The coefficient of determination, R^2, is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. R^2 is most often seen as a number between 0 and 1.0, used to describe how well a regression line fits a set of data. An R^2 near 1.0 indicates that a regression line fits the data well, while an R^2 closer to 0 indicates a regression line does not fit the data very well. It is the proportion of variability in a dataset that is accounted for by the statistical model.
3. F-statistic: An F-test is any statistical test in which the test statistic has a continuous probability distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a dataset, in order to identify the model that best fits the population from which the data were sampled. The data obtained from the F-test is taken as F-statistic. In our case, we have taken the F-statistic to be the p-value.

Classification of Mutual Funds

We have classified the observed mutual funds into the following three categories:

1. *Indian Joint Venture AMCs*: These include mutual funds like Birla Sun Life Asset Management Co. Ltd., DSP Merrill Lynch Fund Managers Limited, and HDFC Asset Management Company Ltd.
2. *Private Indian AMCs*: These are mutual funds floated by private Indian companies. These include MFs like Cholamandalam Asset Management Co. Ltd., Kotak Mahindra Asset Management Co. Ltd., Reliance Capital Asset Management Ltd., Sahara Asset Management Co. Pvt. Ltd., and Tata Asset Management Private Ltd.

Table 1 Performance of Sharpe ratio

		Private Indian	Indian JV	Foreign AMC
$R_p - R_f$	Fut_1Y	−ve	−ve	−ve
$R_p - R_f$	Fut_2Y	−ve	−ve	−ve
$R_p - R_f$	Fut_3Y	−ve	−ve	−ve
$R_p - R_m$	Fut_1Y	**+ve**	**+ve**	**+ve**
$R_p - R_m$	Fut_2Y	−ve	−ve	−ve
$R_p - R_m$	Fut_3Y	−ve	−ve	−ve
BAHR	Fut_1Y	−ve	−ve	−ve
BAHR	Fut_2Y	−ve	−ve	−ve
BAHR	Fut_3Y	−ve	−ve	−ve

Table 2 Performance of information ratio

		Private Indian	Indian JV	Foreign AMC
$R_p - R_f$	Fut_1Y	−ve	−ve	−ve
$R_p - R_f$	Fut_2Y	−ve	−ve	−ve
$R_p - R_f$	Fut_3Y	−ve	−ve	−ve
$R_p - R_m$	Fut_1Y	**+ve**	**+ve**	**+ve**
$R_p - R_m$	Fut_2Y	−ve	−ve	**F_NS**
$R_p - R_m$	Fut_3Y	−ve	−ve	−ve
BAHR	Fut_1Y	−ve	−ve	−ve
BAHR	Fut_2Y	−ve	−ve	−ve
BAHR	Fut_3Y	−ve	−ve	−ve

3. *Foreign AMCs*: These are mutual funds floated by companies which have a major stake held by the foreign investors. A few of them are ABN AMRO Asset Management (I) Ltd., HSBC Asset Management (India) Private Ltd., ING Investment Management (India) Pvt. Ltd., and Morgan Stanley Investment Management Pvt. Ltd.

The calculations have been done for the above three categories separately. In case the F-statistic for a certain class of mutual funds is greater than 5 %, we take it as nonsignificant, which is represented in the following Tables 1, 2, 3, and 4 as **F_NS**. Also, +**ve** and –ve signifies that the correlation is positive and negative, respectively, with p-value being less than 0.05.

Summary of Observations

Since there are a lot of nonsignificant results, we now try to reanalyze the data after using bootstrapping.

Table 3 Performance of Treynor ratio

		Private Indian	Indian JV	Foreign AMC
$R_p - R_f$	Fut_1Y	**F_NS**	**F_NS**	**F_NS**
$R_p - R_f$	Fut_2Y	**F_NS**	**F_NS**	**F_NS**
$R_p - R_f$	Fut_3Y	**F_NS**	**F_NS**	**F_NS**
$R_p - R_m$	Fut_1Y	**F_NS**	**F_NS**	**F_NS**
$R_p - R_m$	Fut_2Y	**F_NS**	**F_NS**	**F_NS**
$R_p - R_m$	Fut_3Y	**F_NS**	**F_NS**	**F_NS**
BAHR	Fut_1Y	**F_NS**	**F_NS**	**F_NS**
BAHR	Fut_2Y	**F_NS**	**F_NS**	**F_NS**
BAHR	Fut_3Y	**F_NS**	**F_NS**	**F_NS**

Table 4 Performance of coefficient of variation

		Private Indian	Indian JV	Foreign AMC
$R_p - R_f$	Fut_1Y	**+ve**	**F_NS**	**+ve**
$R_p - R_f$	Fut_2Y	**+ve**	**F_NS**	**+ve**
$R_p - R_f$	Fut_3Y	**+ve**	**+ve**	**+ve**
$R_p - R_m$	Fut_1Y	**+ve**	**F_NS**	**+ve**
$R_p - R_m$	Fut_2Y	**+ve**	**F_NS**	**+ve**
$R_p - R_m$	Fut_3Y	**+ve**	**F_NS**	**+ve**
BAHR	Fut_1Y	**+ve**	**F_NS**	**+ve**
BAHR	Fut_2Y	**+ve**	**F_NS**	**F_NS**
BAHR	Fut_3Y	**+ve**	**F_NS**	**+ve**

Bootstrapping of Samples

Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution. One standard choice for an approximating distribution is the empirical distribution of the observed data. In the case where a set of observations can be assumed to be from an independent and identically distributed population, this can be implemented by constructing a number of resamples of the observed dataset (and of equal size to the observed dataset), each of which is obtained by random sampling with replacement from the original dataset.

It may also be used for constructing hypothesis tests. It is often used as an alternative to inference based on parametric assumptions when those assumptions are in doubt or where parametric inference is impossible or requires very complicated formulae for the calculation of standard errors.

Confidence Intervals

There are different types of bootstrap confidence intervals available in the literature. We have implemented only two types:

1. *Bootstrap-t confidence interval*: We can construct a $(1-\alpha)\,\%$ confidence interval as $\left[\overline{X} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}},\ \overline{X} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$ when X_is are *i.i.d.* sample from $\mathcal{N}\left(\mu,\ \sigma^2\right)$. But the problem will occur if we are not sampling from normal distribution, but rather some other distribution. In that case, the following bootstrap confidence interval can be constructed: Let $\overline{X}_{boot,\ b}$ and $s_{boot,\ b}$ be the sample mean and standard deviation of the bth resample, $b = 1, \ldots, B$. Define

$$t_{boot,b} = \frac{\overline{X} - \overline{X}_{boot,b}}{\frac{s_{boot,b}}{\sqrt{n}}}.$$

Notice that $t_{boot,b}$ is defined in the same way as t except for two changes: first, $\overline{X}$ and s in t are replaced by $\overline{X}_{boot,b}$ and $s_{boot,b}$, and second, μ in t is replaced by $\overline{X}$ in $t_{boot,b}$. The last point is a bit subtle, and you should stop to think about it. A resample is taken using the original sample as the population. Thus, for the resample, the population mean is $\overline{X}$!.

Because the resamples are independent of each other, the collection $t_{boot,1}$, $t_{boot,2}$, … can be treated as a random sample from the distribution of the t-statistic. After B values of $t_{boot,b}$ have been calculated, one from each resample, we find the $100\,(1-\alpha)$ and $100\left(1-\frac{\alpha}{2}\right)$ percentiles of this collection of $t_{boot,b}$ values. Call these percentiles t_L and t_U. More specifically, we find t_U and t_L as we described earlier. We sort all the B values from smallest to largest. Then we calculate the $B\frac{\alpha}{2}$ and round to the nearest integer. Suppose the result is K_L. Then the K_Lth sorted value of $t_{boot,b}$ is t_L. Similarly, let K_U be $B\,\left(1-\frac{\alpha}{2}\right)$ rounded to the nearest integer and then t_U is the K_Uth sorted value of $t_{boot,b}$. Finally we can make the bootstrap confidence interval for μ as $\left[\overline{X} + t_L\frac{s}{\sqrt{n}},\ \overline{X} + t_U\frac{s}{\sqrt{n}}\right]$. We get two advantages through bootstrap:

- We do not need to know the population distribution.
- We do not need to calculate the distribution of t-statistic using probability theory.

2. *Bootstrap percentile confidence interval*: The percentile bootstrap proceeds in a similar way to the basic bootstrap, using percentiles of the bootstrap distribution, but with a different formula:

$$\left(\theta^*_{(\alpha)};\ \theta^*_{(1-\alpha)}\right)$$

where $\theta^*_{(1-\alpha)}$ denotes the $(1-\alpha)$ percentile of the bootstrapped coefficient θ^*.

Advantages and Disadvantages of Bootstrapping

- *Advantages*: A great advantage of bootstrap is its simplicity. It is a straightforward way to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution, such as percentile points, proportions, odds ratio, and correlation coefficients. Moreover, it is an appropriate way to control and check the stability of the results.
- *Disadvantages*: Although bootstrapping is (under some conditions) asymptotically consistent, it does not provide general finite-sample guarantees. Furthermore, it has a tendency to be overly optimistic. The apparent simplicity may conceal the fact that important assumptions are being made when undertaking the bootstrap analysis (e.g., independence of samples) where these would be more formally stated in other approaches.

Other Important Variables

1. *Skewness*: It is a measure of the asymmetry of the probability distribution of a real-valued random variable. The skewness value can be positive or negative or even undefined. Qualitatively, a negative skew indicates that the tail on the left side of the probability density function is longer than the right side, and the bulk of the values (including the median) lie to the right of the mean. A positive skew indicates that the tail on the right side is longer than the left side, and the bulk of the values lie to the left of the mean. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, typically (but not necessarily) implying a symmetric distribution.
2. *Kurtosis*: It is a measure of the "peakedness" of the probability distribution of a real-valued random variable. In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution, and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population.
3. *Fat-tailed distribution*: A fat-tailed distribution is a probability distribution that has the property, along with the heavy-tailed distributions, that exhibits extremely large skewness or kurtosis. This comparison is often made relative to the ubiquitous normal distribution, which itself is an example of an exceptionally thin tail distribution, or to the exponential distribution.
4. *Quantiles*: Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. Dividing ordered data into q essentially equal-sized data subsets is the motivation for q-quantiles; the quantiles are the data values marking the boundaries between consecutive subsets. Put another way, the *k*th q-*quantile* for a random variable is the value x such that the probability that the random variable will be less than x is at most

$^{k}\!/_{q}$ and the probability that the random variable will be more than x is at most $\frac{q-k}{q} = 1 - \frac{k}{q}$. There are q − 1 of the q-*quantiles*, one for each integer k satisfying $0 < k < q$.

Results

We sample the data using bootstrap and then take the 95 % confidence interval of the result. Once this is done, the results are then calculated using the methods explained before. The results are calculated separately for all the mutual funds, both separately and category wise.

We provide the following Tables 5, 6, 7, and 8 to summarize the results:

On comparing the results obtained with the one without bootstrapping, we see that the results have shown a slight improvement on bootstrapping. This is because the extreme aberrations have been removed by taking only the 95 % confidence interval. However, the change in the dataset is very minimal to reflect any significant change in the results. In a few cases dealing with the Indian Private Venture AMCs,

Table 5 Performance of Sharpe ratio with bootstrapped data

		Private Indian	Indian JV	Foreign AMC
$R_p - R_f$	Fut_1Y	−ve	−ve	−ve
$R_p - R_f$	Fut_2Y	−ve	−ve	−ve
$R_p - R_f$	Fut_3Y	−ve	−ve	−ve
$R_p - R_m$	Fut_1Y	**+ve**	**+ve**	**+ve**
$R_p - R_m$	Fut_2Y	−ve	−ve	−ve
$R_p - R_m$	Fut_3Y	−ve	−ve	−ve
BAHR	Fut_1Y	−ve	−ve	−ve
BAHR	Fut_2Y	−ve	−ve	−ve
BAHR	Fut_3Y	−ve	−ve	−ve

Table 6 Performance of information ratio with bootstrapped data

		Private Indian	Indian JV	Foreign AMC
$R_p - R_f$	Fut_1Y	−ve	−ve	−ve
$R_p - R_f$	Fut_2Y	−ve	−ve	−ve
$R_p - R_f$	Fut_3Y	−ve	−ve	−ve
$R_p - R_m$	Fut_1Y	**+ve**	**+ve**	**+ve**
$R_p - R_m$	Fut_2Y	−ve	−ve	**F_NS**
$R_p - R_m$	Fut_3Y	−ve	−ve	−ve
BAHR	Fut_1Y	−ve	−ve	−ve
BAHR	Fut_2Y	−ve	−ve	−ve
BAHR	Fut_3Y	−ve	−ve	−ve

Table 7 Performance of Treynor ratio with bootstrapped data

		Private Indian	Indian JV	Foreign AMC
$R_p - R_f$	Fut_1Y	**F_NS**	**F_NS**	+**ve**
$R_p - R_f$	Fut_2Y	**F_NS**	**F_NS**	+**ve**
$R_p - R_f$	Fut_3Y	**F_NS**	**F_NS**	+**ve**
$R_p - R_m$	Fut_1Y	**F_NS**	**F_NS**	+**ve**
$R_p - R_m$	Fut_2Y	**F_NS**	**F_NS**	+**ve**
$R_p - R_m$	Fut_3Y	**F_NS**	**F_NS**	+**ve**
BAHR	Fut_1Y	**F_NS**	**F_NS**	**F_NS**
BAHR	Fut_2Y	**F_NS**	**F_NS**	**F_NS**
BAHR	Fut_3Y	**F_NS**	**F_NS**	+**ve**

Table 8 Performance of coefficient of variation with bootstrapped data

		Private Indian	Indian JV	Foreign AMC
$R_p - R_f$	Fut_1Y	+**ve**	**F_NS**	+**ve**
$R_p - R_f$	Fut_2Y	+**ve**	**F_NS**	+**ve**
$R_p - R_f$	Fut_3Y	+**ve**	+**ve**	**F_NS**
$R_p - R_m$	Fut_1Y	+**ve**	**F_NS**	**F_NS**
$R_p - R_m$	Fut_2Y	+**ve**	**F_NS**	**F_NS**
$R_p - R_m$	Fut_3Y	+**ve**	**F_NS**	+**ve**
BAHR	Fut_1Y	+**ve**	**F_NS**	+**ve**
BAHR	Fut_2Y	+**ve**	**F_NS**	**F_NS**
BAHR	Fut_3Y	+**ve**	**F_NS**	+**ve**

we see a fat-tailed distribution, signifying that the returns are not very close to each other and have a significant deviation from the mean.

Also, it can be seen that the Treynor ratio, which was giving nonsignificant results for all the MFs, is now giving some significant results for the Foreign AMCs. The results obtained in our case are in sharp contrast with the popular belief that Treynor ratio is one of the best estimators available.

As per Rao, who evaluated the performance of Indian mutual funds in bear market through return performance index, risk-return analysis, Treynor ratio, Sharpe ratio, Sharpe measure, Jensen measure, and Fama measure, the medium-term debt funds performed extremely well during that period while InfoTech equity funds suffered the most. Some fund managers were able to diversify the risk and maximize the return, and even few managed to give returns greater than the risk-free return during bear market, but the debt funds were the ultimate winners.

Deb, Banerjee, and Chakrabarti, in 2007, found very little evidence of market timing skill of fund managers though their stock selection ability was decent enough to produce good returns from the market.

Conclusion

This study of the persistence of mutual funds in India was aimed at providing an insight to the ways in which the return of a mutual fund can be properly evaluated. A study of this kind is important because of lack of extensive studies for the Indian market. Most of the devised indices have been developed mostly for foreign markets, and hence it is of more importance to see if the Indian markets follow a similar trend.

In order to accomplish our aim, we take up the most prominent indices, viz., Treynor ratio, Sharpe ratio, coefficient of variation, and information ratio. These are applied to the data obtained for the mutual funds in India and then fitted accordingly. The positive/negative correlation is then measured along with the F-statistic and R^2 data in order to measure the accuracy of the results obtained. While the results are satisfactory for some of the indices, the same obtained for Treynor ratio is very deviating, for it does not at all fit in the Indian context.

Moreover, as expected, the bootstrapped data provides more accurate results, thereby keeping the dataset much closer to the actual one. However, with the results varying a great deal for almost all the indices, there needs to be made an adjustment to all the formulae used for them to be applied to the Indian markets.

Acknowledgments I must thank Prof. Rajen K. Sinha and Dr. Siddhartha P. Chakrabarty for providing me with all the support and motivation for carrying out the research.

References

1. Carhart, M.M.: On persistence in mutual fund performance. J. Financ. **52**(1), 57–82 (1997)
2. Grinblatt, M., Titman, S.: The persistence of mutual fund performance. J. Financ. **47**(5), 1977–1984 (1992)
3. Stutzer, M.: Fund managers may cause their benchmarks to be priced 'risks'. J. Invest. Manag. **1**(3), 1–13 (2003)
4. Elton, E.J., Gruber, M.J., Blake, C.R.: The persistence of risk-adjusted mutual fund performance. J. Bus. **69**(2), 133–157 (1996)
5. Treynor, J.L., Majuy, K.: Can mutual funds outguess the market? Harv. Bus. Rev. **43**, 131–136 (1966)
6. Bhattacharjee, K., Roy, B.: Fund performance measurement without benchmark—a case of select Indian mutual funds. In: 10th Capital Markets Conference, Indian Institute of Capital Markets Paper, December 2006
7. Deb, S.G., Banerjee, A., Chakrabarti, B.B.: Performance of Indian equity mutual funds vis-a vis their style benchmarks: an empirical exploration. In: 10th Capital Markets Conference, Indian Institute of Capital Markets, December 2006
8. Sharpe, W.F.: The Sharpe ratio. J. Portf. Manag. **21**(1), 49–58 (1994)
9. Goodwin, T.H.: The information ratio. Financ. Anal. J. **54**(4), 34–43 (1998)

Counting Regular Expressions in Degenerated Sequences Through Lazy Markov Chain Embedding

G. Nuel and V. Delos

Abstract Nowadays, Next- Generation Sequencing (NGS) produces huge number of reads which are combined using multiple alignment techniques to produce sequences. During this process, many sequencing errors are corrected, but the resulting sequences nevertheless contain a marginal level of uncertainty in the form of $\sim 0.1\,\%$ or less of degenerated positions (like the letter "N" corresponding to any nucleotide).

A previous work Nuel (Pattern Recognition in Bioinformatics. Springer, New York, 2009) showed that these degenerated letters might lead to erroneous counts when performing pattern matching on these sequences. An algorithm based on Deterministic Finite Automata (DFA) and Markov Chain Embedding (MCE) was suggested to deal with this problem.

In this paper, we introduce a new version of this algorithm which uses Nondeterministic Finite Automata (NFA) rather than DFA to perform what we call "lazy MCE.". This new approach proves itself much faster than the previous one and we illustrate its usefulness on two NGS datasets and a selection of regular expressions.

A software implementing this algorithm is available: `countmotif`, http://www.math-info.univ-paris5.fr/~delosvin/index.php?choix=4.

Keywords Nondeterministic finite automaton • Deterministic finite automaton • Next-generation sequencing • Pattern matching • Lazy determinization

G. Nuel (✉) • V. Delos
Institute of Mathematics (INSMI), National Center for French Research (CNRS),
Department of Applied Mathematics (MAP5), Université Paris Descartes,
Sorbonne Paris Cité, Paris 75006, France
e-mail: nuel@math.cnrs.fr

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.28_20

Introduction

Over the last decade, sequencing technologies have made constant progress both in term of throughput and reduced costs. Nowadays, Next -Generation Sequencing (NGS) techniques produce huge amount of data, usually in the form of a large number of short reads with a high error rate (e.g.,x: ∼1 %, see [2]). Using redundancy of the reads and highly computational algorithms, one can produce from these datasets longer sequenced fragments with much lower error rate [8]. However, even after such a correction, it is not unusual that a low proportion of positions remain only partially determined (i.e., degenerated).

The IUPAC alphabet (see Table 1) is especially designed to represent these degenerated position in DNA (or RNA) sequences. We can see in Fig. 1 an example of a sequence which contains numerous of such degenerated positions. When processing sequences though analysis, it is convenient to simply ignore these degenerated position. But such an approach, might have unexpected consequence depending on the analysis performed. For example, if we consider the problem of matching the regular expression G[AG]T in sSequence W36091, we can see that

Table 1 IUPAC alphabet [4]

Symbol	Meaning
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
R	Purine (A or G)
Y	Pyrimidine (C or T)
M	C or A
K	T or G
W	T or A
S	C or G
B	Not A
D	Not C
H	Not G
V	Not G, not U
N	Any base

```
>W36091
GAATTCTACTGCATCTTGCCAAGGTATTATGCCTCCNATTCCAATA
ATCGGAATRTCTAGNTNTNAAGCCAGATYAGRTACAGGACTAAGTG
CTASTGGNATAGNCCCTGGGTCANAGNATCBCCCAAGATNATNCCT
GANGA
```

Fig. 1 EMBL sequence W36091. Degenerated positions are highlighted with a *grey background*,; possible matching positions of G[AG]T are *underlined*

we have a total of 4 matching positions on non degenerated letters, and possible 8 additional matching positions on one or more degenerated letters. As a consequence, the count of G[AG]T ranges from 4 to 12 in W36091.

In Nuel [7], we previously suggested to deal with this problem using Deterministic Finite Automata (DFA) and Markov Chain Embedding (MCE). The method proved itself to be reliable but quite tedious to put in practice, especially when working with highly complex regular expressions which corresponding DFA have a large amount of states (ex.g.,: 10,000 or more). This problem is well known in the context of pattern matching where Nondeterministic Finite Automata (NFA) are always preferred to DFA for dealing with regular expression [3].

In this paper, we want to take advantage of an NFA to introduce a new lazy MCE approach able to deal with pattern matching over degenerated sequences in a much more efficient way than the MCE of [7].

The paper is organized as follows: in section "Pattern Matching, NFA, and Lazy Determinization," we briefly recall some element of the pattern matching theory and describe in details NFA and lazy NFA pattern matching algorithms; in section "Lazy Markov Chain Embedding," we first recall the principle of MCE using DFA before to introducinge our new algorithm; in section "Illustration on NGS Data," we illustrate the usefulness of our algorithm on two NGS dataset, before to concludinge in section "Conclusions.".

Pattern Matching, NFA, and Lazy Determinization

In this section, we briefly recall some well- known elements of the pattern matching theory. For more detail, please consult a reference textbook like [3].

Pattern Matching

Let us consider a regular expression R over the finite alphabet Σ. The purpose of pattern matching is to find (or count) all matching positions of R in a text $X_{1:n} \in \Sigma^n$. More formally, it means that we want to determine the set $\mathcal{I}(R, X_{1:n}) = \{1 \leqslant i \leqslant n, X_{1:i} \in \Sigma^* R\}$ where * denotes the Kleene's closure operator. The number $N(R, X_{1:n})$ of matching occurrences is hence defined as the cardinal of $\mathcal{I}(R, X_{1:n})$ or simply as

$$N(R, X_{1:n}) = \sum_{i=1}^{n} \mathbb{1}_{X_{1:i} \in \Sigma^* R} \tag{1}$$

where $\mathbb{1}$ denotes the indicator function.

NFA

In order to perform pattern matching of regular expression R, the classical approach consists first in building a Nondeterministic Finite Automaton (NFA) corresponding to R, i.e., which recognizes the regular language $\Sigma^* R$. Let $(\Sigma, \mathcal{Q}, \mathcal{S}, \mathcal{F}, \delta)$ being such an NFA where $\mathcal{Q}$ is the finite state space, $\mathcal{S} \subset \mathcal{Q}$ the (non empty) set of starting states, $\mathcal{F} \subset \mathcal{Q}$ the (non empty) set of final states, and $\delta : \mathcal{Q} \times \Sigma \to \mathcal{P}(\mathcal{Q})$ (set of parts of $\mathcal{Q}$) being the transition function. A finite sequence $X_{1:n} \in \Sigma^n$ is recognized by the NFA if and only if thereit exists a sequence $q_0, q_1, \dots, q_n \in \mathcal{Q}$ such as the following: i) $q_0 \in \mathcal{S}$; ii) $q_n \in \mathcal{F}$; and iii) for all $1 \leqslant i \leqslant n$, $q_i \in \delta(q_{i-1}, X_i)$.

We know from the classical pattern matching theory that for any regular expression R, it exists an NFA which recognizes exactly $\Sigma^* R$. The Glushkov algorithm provides a way to build such an NFA from any regular expression. Please note that finding the smallest possible NFA having this property is an NP-hard problem in general, but that efficient reduction heuristics are available [1]. For example, we can see on Fig. 2 an NFA corresponding to our toy-example regular expression G[AG]T over the DNA alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

Once an NFA corresponding to R is available, all matching positions can be reported using Algorithm 1.

Algorithm 1 NFA matching algorithm of regular expression R

```
Input: $(\Sigma, \mathcal{Q}, \mathcal{S}, \mathcal{F}, \delta)$ a NFA that recognizes $\Sigma^* R$, and a sequence $X = X_{1:n}$
  matchCount = 0
  setOfStates = $\mathcal{S}$
  for $i = 1 \dots n$ do
    update setOfStates = $\delta$(setOfStates, $X_i$)
    if setOfStates $\cap \mathcal{F} \neq \emptyset$ then
      report matching position $i$
      matchCount = matchCount + 1
    end if
  end for
Output: all matching positions and number matchCount of occurrences.
```

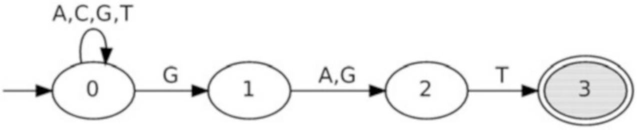


Fig. 2 NFA corresponding to the regular expression G[AG]T over the DNA alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. $\mathcal{Q} = \{0, 1, 2, 3\}$, $\mathcal{S} = \{0\}$, $\mathcal{F} = \{3\}$, $\delta(0, \mathtt{A}) = \{0, 1\}$, $\delta(0, \mathtt{C}) = \delta(0, \mathtt{G}) = \delta(0, \mathtt{T})\{0\}$, $\delta(1, \mathtt{A}) = \delta(1, \mathtt{G}) = 2$, $\delta(2, \mathtt{T}) = 3$

Algorithm 2 Lazy determinization NFA matching algorithm of regular expression R

```
Input: $(\Sigma, \mathcal{Q}, \mathcal{S}, \mathcal{F}, \delta)$ a NFA that recognizes $\Sigma^* R$, a sequence $X = X_{1:n}$, and a lookup table $\mathcal{L}$
  matchCount $= 0$
  setOfStates $= \mathcal{S}$
  for $i = 1 \dots n$ do
    if (setOfStates, $X_i$) $\in \mathcal{L}$ then
      setOfStates $= \mathcal{L}$(setOfStates, $X_i$)
    else
      compute setOfStates $= \delta$(setOfStates, $x$) and store it in $\mathcal{L}$
    end if
    if size($\mathcal{L}$) $>$ maxSize then
      flush $\mathcal{L}$
    end if
    if setOfStates $\cap \mathcal{F} \neq \emptyset$ then
      report matching position $i$
      matchCount $=$ matchCount $+ 1$
    end if
  end for
Output: all matching positions and number matchCount of occurrences.
```

Lazy Determinization

Using our NFA matching algorithm, we have a complexity of $\mathcal{O}(|\mathcal{Q}| \times n)$ for matching a regular expression which NFA has $|\mathcal{Q}|$ in a sequence of size n. This can be slow when considering large NFA (i.e., complex regular expression). The term $|\mathcal{Q}|$ in the complexity is due to the necessary computation at each position in X of $\delta(\mathcal{U}, a)$ for a set $\mathcal{U} \subset \mathcal{Q}$ and $a \in \Sigma$.

A very natural idea hence consists in keeping previously computed $\delta(\mathcal{U}, a)$ in a lookup table and usinge it to speed up computation. Since the size of this lookup table could be $2^{|\mathcal{Q}|} \times |\Sigma|$ in the worst case, the table is usually limited to a fixed size and flushed each time the limit size is reached. This results in Algorithm 2 which allows, for a modest additional memory cost, to achieve a complexity close to $\mathcal{O}(n)$ for solving the pattern matching problem.

If we consider the problem of matching the regular expression $R =$ G[AG]T on the sequence $X =$ AGGATGGGC, we start with setOfStates $= \{0\}$ and we get:

- setOfStates $= \delta(\{0\}, X_1 = \texttt{A}) = \{0\} = \mathcal{L}(\{0\}, \texttt{A})$;
- setOfStates $= \delta(\{0\}, X_2 = \texttt{G}) = \{0, 1\} = \mathcal{L}(\{0\}, \texttt{G})$;
- setOfStates $= \delta(\{0, 1\}, X_3 = \texttt{G}) = \{0, 1, 2\} = \mathcal{L}(\{0, 1\}, \texttt{G})$;
- setOfStates $= \delta(\{0, 1, 2\}, X_4 = \texttt{A}) = \{0, 3\} = \mathcal{L}(\{0, 1, 2\}, \texttt{A})$
- setOfStates $= \delta(\{0, 3\}, X_5 = \texttt{T}) = \{0\} = \mathcal{L}(\{0, 3\}, \texttt{T})$
- setOfStates $= \mathcal{L}(\{0\}, X_6 = \texttt{G}) = \{0, 1\}$ (in lookup table);
- setOfStates $= \mathcal{L}(\{0, 1\}, X_7 = \texttt{G}) = \{0, 1, 2\}$ (in lookup table);

Table 2 NFA and (minimal) DFA size for several DNA regular expressions (using IUPAC alphabet) and their corresponding counts in NGS datasets

Regular expression	NFA size	DFA size	DRR000019	DRR000020
GRT	4	5	88,645	3,876,732
GCTA	5	5	5,289	242,512
TTNGT	6	8	8,394	339,485
GCTAN{5, 10}TTNGT	30	251	114	4,936
GCTAN{10, 15}TTNGT	35	1,256	94	4,328
GCTAN{15, 20}TTNGT	40	6,289	65	2,749
GCTAN{20, 25}TTNGT	45	31,505	62	2,965
GCTAN{25, 30}TTNGT	50	157,836	72	2,893
GRTN{5, 10}GCTAN{5, 10}TTNGT	53	1,608	16	615
GRTN{10, 15}GCTAN{10, 15}TTNGT	63	17,848	15	714
GRTN{15, 20}GCTAN{15, 20}TTNGT	73	198,002	3	303

- setOfStates $= \delta(\{0, 1, 2\}, X_8 = \texttt{G}) = \{0, 1, 2\} = \mathcal{L}(\{0, 1, 2\}, \texttt{G})$;
- setOfStates $= \delta(\{0, 1, 2\}, X_9 = \texttt{T}) = \{0\} = \mathcal{L}(\{0, 1, 2\}, \texttt{T})$.

At the end of the algorithm, we have a total of matchCount $= 1$ matching position (position $i = 4$), the lookup table contains a total of 7 computed transitions, and saved 2 transition computations (for $i = 6$ and $i = 7$) by using the lookup table. Of course, the longer the considered sequence, the better gain from using a lookup table.

In Table 2, we can see the number of matching positions (ignoring all degenerated letters in the sequences) for various regular expression in two NGS datasets (see section "Illustration on NGS Data" for more details on these datasets). All these computations took a total of 0.5 s for DRR000019 (2.8 Mb) and 15 s for DRR000019 (124.9 Mb) using an Intel Xeon CPU E5-2609@2.40 GHz running Linux 3.2.0.

Lazy Markov Chain Embedding

Markov Chain Embedding

Let us now assume that our text of interest is a random sequence $X_{1:n}$ where the $X_i \in \Sigma$ are independent random variables with distribution $\pi_i(\cdot)$ defined by $\mathbb{P}(X_i = a) = p_i(a)$ for all $a \in \Sigma$. The number of occurrences $N = N(R, X_{1:n})$ is hence a random variable over $\{0, \ldots, n\}$. Obtaining the exact distribution of N is a well-known challenging problem that can be solved efficiently by MCE [5, 6].

The basic idea is that the random sequence $(\mathcal{U}_i)_{0 \leqslant i \leqslant n}$ defined by $\mathcal{U}_0 = \mathcal{S}$ and $\mathcal{U}_i = \delta(\mathcal{U}_{i-1}, X_i)$ for all $1 \leqslant i \leqslant n$ is a Markov sequence over the finite states of all possible configurations of NFA states encountered during the pattern matching process. In practice, we first build from the NFA corresponding to the problem a

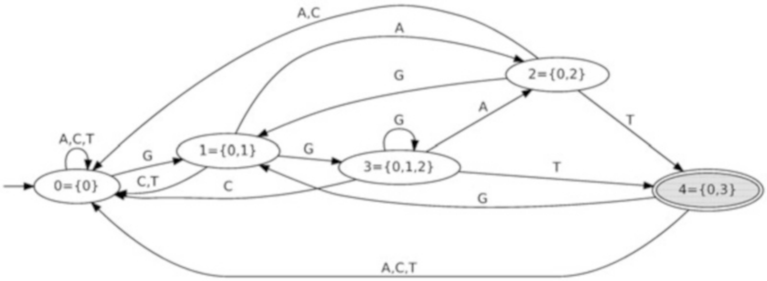


Fig. 3 (Minimal) DFA corresponding to the regular expression $\texttt{G[AG]T}$ over the DNA alphabet $\Sigma = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$. $\mathcal{Q} = \{0, 1, 2, 3, 4\}$, $s = 0$, $\mathcal{F} = \{4\}$, $\delta(0, \texttt{A}) = \delta(0, \texttt{C}) = \delta(0, \texttt{T}) = 0$, $\delta(0, \texttt{G}) = 1$, $\delta(1, \texttt{A}) = 2$, etc

Deterministic Finite Automaton (DFA) which encodes all possible configurations of NFA states and their transitions (this NFA can be seen as a complete lookup table as defined above). This DFA is then used to compute a transition probability table where we keep track of the number of matching states (i.e., configuration $\mathcal{U}$ of NFA state such as $\mathcal{U} \cap \mathcal{F} \neq \emptyset$).

For example, let us consider the case of the regular expression $R = \texttt{G[AG]T}$ in an i.i.d. random sequence of size $n = 10$ with $\pi_1(\texttt{A}) = \pi_1(\texttt{C}) = \pi_1(\texttt{G}) = \pi_1(\texttt{T}) = 0.25$. We first obtain from the NFA of Fig. 2 the DFA of Fig. 3, from which we obtain that $(\mathcal{U}_i)_{0 \leqslant i \leqslant n}$ is defined over the 5 states spaces $\{\{0\}, \{0, 1\}, \{0, 1, 2\}, \{0, 2\}, \{0, 3\}\}$ with the following transition matrix:

$$T(z) = \begin{pmatrix} 0.75 & 0.25 & 0.00 & 0.00 & 0.00 \\ 0.50 & 0.00 & 0.25 & 0.25 & 0.00 \\ 0.50 & 0.25 & 0.00 & 0.00 & 0.25z \\ 0.25 & 0.00 & 0.25 & 0.25 & 0.25z \\ 0.75 & 0.25 & 0.00 & 0.00 & 0.00 \end{pmatrix}$$

where z is a dummy variable used to keep track of the number of matching occurrences. The probability generating function (pgf) of N can therefore be obtained by:

$$g(z) = \sum_{k=0}^{n} \mathbb{P}(N = k) z^k = u T(z)^n v$$

where $u = (1\ 0\ 0\ 0\ 0)$ and $v = (1\ 1\ 1\ 1\ 1)^T$. For example, with $n = 10$ we get:

$$g(z) = 0.7645264 + 0.2210693z + 0.0142822z^2 + 0.0001221z^3$$

which indicates that we have 0 to 3 occurrences of our regular expression in our size $n = 10$ sequence, and that, (for example), $\mathbb{P}(N = 3) = 0.0001221$.

Our Contribution

If this MCE technique is very efficient for low- complexity regular expressions (i.e., DFA of reasonable size), it could be very slow when considering high-complexity expressions (e.g.,x: DFA of size 10,000 or more). We can see in Table 2 the minimal DFA size of various DNA regular expressions. Structured motifs (regular expression with one or several spacers) in particular lead to an exponential increase of complexity with their gap length. Note that if most DFA computations areis quite fast on a standard workstation (a couple of seconds in the worst case), the most complex expressions for which DFA size is greater thatn 100,000 might require a significant longer computational time (e.g.,x: 20 min or more).

Moreover, in the context of degenerated sequences, we expect only a small portion of the sequence (typically,: 10 or 1 or 0.1 %) of the sequence to be random (choice between at least two possible letters), the remaining position being completely deterministic. For example, in sequence W36091 (see Fig. 1), we have a total of 16 degenerated positions among a total of $n = 147$ positions.

For complex expressions, like with the lazy determinization introduced above, it is unlikely that all DFA states will be explored in the pattern matching process. Our idea is hence to adapt the lazy determinization approach to the context of MCE by providing the new lazy MCE approach described in Algorithm 3.

For simplification purpose, we assume there is a common background distribution π for all the letters in the sequence which hence naturally define the distribution of any degenerated letter in the sequence. However, the suggested method can easily adapt to heterogeneous background distributions.

For example, let us consider the problem of finding the matching position of $R = \texttt{G[AG]T}$ in $X = \texttt{TGGNATA}$ where there is only one degenerated position in $i = 4$. It means that $\mathcal{X}_i = \{X_i\}$ for all i except $\mathcal{X}_4 = \{\texttt{A},\texttt{C},\texttt{G},\texttt{T}\}$. It is easy to see that we have $\mathcal{M} = \{(\{0,1,2\},1.0)\}$ after position $i = 3$ when we meet the degenerated position. For simplification purpose, let us assume that $\pi_x = 0.25$ for all $x \in \Sigma$, for all $x \in \mathcal{X}_4 = \{\texttt{A},\texttt{C},\texttt{G},\texttt{T}\}$ we hence have $\texttt{prob} = 0.25$. After processing $\mathcal{X}_4$ we hence have: $\mathcal{M} = \{(\{0,2\},0.25),(\{0\},0.25),(\{0,1,2\},0.25),(\{0,3\},0.25z)\}$. After $\mathcal{X}_5 = \{\texttt{A}\}$ we get: $\mathcal{M} = \{(\{0\},0.50 + 0.25z),(\{0,2\},0.25)\}$. After $\mathcal{X}_6 = \{\texttt{T}\}$, $\mathcal{M} = \{(\{0\}, 0.50 + 0.25z),(\{0,3\},0.25z)\}$. And finally after $\mathcal{X}_7 = \{\texttt{T}\}$, we get $\mathcal{M} = \{(\{0\},0.50 + 0.50z)\}$. We hence conclude that we have 50 % of chance to see no matching position, and 50 % to see one matching position. One should note that despite the fact that two positions might match with the expression of interest ($i = 4$ and $i = 6$), it cannot match at these two positions at the same time. This is clearly pointed out by the fact that the degree of the resulting pgf is 1 and not 2.

For validation purpose, let us use this algorithm for $R = \texttt{G[AG]T}$ and $X = \texttt{NNNNNNNNNNN}$. We get: $g(z) = 0.764526 + 0.221069z + 0.0142822z^2 + 0.00012207z^3$ which is (up to numerical rounding) exactly the pgf computed using MCE in section "Markov Chain Embedding".

When dealing with long sequence, it is obviously faster to rely on Algorithm 3 only when necessary which means when dealing with degenerated positions and/or

Algorithm 3 Lazy Markov Chain Embedding algorithm for regular expression R

```
Input: $(\Sigma, \mathcal{Q}, \mathcal{S}, \mathcal{F}, \delta)$ an NFA that recognizes $\Sigma^* R$, a degenerated sequence $\mathcal{X} = \mathcal{X}_{1:n}$ where
  $\mathcal{X}_i \subset \Sigma$ and $\mathcal{X}_i \neq \emptyset$ for all i, a multimap $\mathcal{M}$ where keys are sets of states (i.e., DFA states) and
  values polynomials, and $\pi$ a probability distribution over $\Sigma$
  // initialization
  $\mathcal{M} = \{(\mathcal{S}, 1.0)\}$ and poly $\in \mathbb{R}[z]$
  // main loop
  for $i = 1 \dots n$ do
    set $\mathcal{M}'$ a new (empty) map
    for all $x \in \mathcal{X}_i$ do
      prob $= \pi_x / \sum_{y \in \mathcal{X}_i} \pi_y$
      for all $(\mathcal{U}, P)$ in $\mathcal{M}$ do
        if $(\mathcal{U}, x) \in \mathcal{L}$ then $\mathcal{V} = \mathcal{L}(\mathcal{U}, x)$ else compute $\mathcal{V} = \delta(\mathcal{U}, x)$ and store it in $\mathcal{L}$
        if $\mathcal{V} \cap \mathcal{F} \neq \emptyset$ then poly $= P \times$ prob $\times z$ else poly $= P \times$ prob
        if $(\mathcal{V}, Q)$ exists in $\mathcal{M}'$ then
          add $(\mathcal{V}, Q +$ poly$)$ in $\mathcal{M}'$
        else
          add $(\mathcal{V},$ poly$)$ in $\mathcal{M}'$
        end if
      end for
    end for
    update $\mathcal{M}$ with $\mathcal{M}'$
  end for
  // final computation
  poly=0
  for all $(\mathcal{U}, P)$ in $\mathcal{M}$ do
    poly=poly+P
  end for
Output: poly contains the pgf of the number of occurrences of $R$ in $\mathcal{X}$
```

when the map $\mathcal{M}$ contains more that one entry. For example, in the illustration of the previous paragraph, positions $i = 1, 2, 3$ could have been processed using simple NFA steps, and after degenerated position $i = 4$, we have to wait for position $i = 7$ to have only one entry in $\mathcal{M}$. The efficient algorithm will then alternate lazy NFA steps with lazy MCE steps keeping track of: (a) a minimum number of occurrence corresponding to the occurrence obtained through lazay NFA steps and; (b) a polynomial pgf corresponding toof the distribution of additional occurrences. When entering from NFA to MCE mode, one just needs to initialize the map $\mathcal{M} = \{(\mathcal{U}, 1.0)\}$ where $\mathcal{U}$ is the current NFA state. Once the map contains only one element $\mathcal{M} = \{(\mathcal{U}, P)\}$, we can return to NFA mode starting with state $\mathcal{U}$ and update the pgf using pgf = pgf $\times P$.

Using this approach on the sequence of Fig. 1, we obtain a minimum of 4 occurrences which can be combined with adjusted pgf polynomial to obtain the following overall pgf:

$$100 \times g(z) = 7.91016z^5 + 26.3672z^6 + 34.2773z^7$$
$$+22.2656z^8 + 7.71484z^9 + 1.36719z^{10} + 0.0976562z^{11}.$$

One should note that if four deterministic positions clearly appear in sequence W36091, it is less clear that the subsequence GRT will always match our regular expression R = G[AG]T. Also note that if we manually identified a total of 12 possible matching positions, the maximum number of matching positions is only 11. This is due to the incompatibility of matches in subsequence GGNAT.

Illustration on NGS Data

In order to illustrate our algorithm on NGS data, we considered the entries DRR000019 and DRR000020 from the NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov). The corresponding binary files DRR000019.sra (6.7 Mo) and DRR000020.sra (294 Mo) and were processed through the sra − toolkit tools (version 2.1.7) in order to export the reads into a classical FASTA file format containing a total of 11,052 sequences of a total size of 2,849,759 nucleotides for DRR000019, and 486,919 sequences for a size of 124,902,331 nucleotides for DRR000020. DRR000019.fasta contains a total of 2,067 degenerated positions (only N) which correspond to 0.073 % of degenerated letters. DRR000020.fasta contains a total of 85,772 degenerated positions (only N) which correspond to 0.069 % of degenerated letters.

We can see in Table 3 that despite the fact that only a small (apparently negligible) proportion of letters are degenerated, the consequence on regex counts could be significant. Ignoring degenerated positions in the sequence willth then often lead to underestimate the number of occurrences in the dataset which could be a problem. This appears clearly when comparing the counts of Table 2 (NFA counts ignoring degenerated positions) and the values of Table 3. Please note that

Table 3 Count results with maximum a posteriori (MAP) and 90 % confidence intervals

Regular expression	DRR000019	DRR000020
GRT	88,771 [88,754; 88,789]	3,884,031 [3,883,904; 3,884,157]
GCTA	5,295 [5,291; 5,300]	242,829 [242,798; 242,859]
TTNGT	8,390 [8,385; 8,396]	339,403 [339,373; 339,434]
GCTAN{5, 10}TTNGT	114 [114; 114]	4,936 [4,932; 4,940]
GCTAN{10, 15}TTNGT	93 [92; 94]	4,325 [4, 321; 4,329]
GCTAN{15, 20}TTNGT	67 [66; 68]	2,755 [2,751; 2,759]
GCTAN{20, 25}TTNGT	62 [62; 62]	2,964 [2,960; 2,968]
GCTAN{25, 30}TTNGT	71 [71; 71]	2,901 [2,898; 2,906]
GRTN{5, 10}GCTAN{5, 10}TTNGT	16 [16; 16]	620 [618; 622]
GRTN{10, 15}GCTAN{10, 15}TTNGT	15 [15; 16]	715 [714; 717]
GRTN{15, 20}GCTAN{15, 20}TTNGT	3 [3; 4]	305 [304; 306]

even in the case of confidence intervals centered on the MAP, the corresponding pgf could be much more complex. For example, with regex `GCTAN{5,10}TTNGT` on `DRR000019`, we get the following pgf:

$$g(z) = z^{114}\left(0.909322z^{0} + 0.088477z^{1}\right.$$
$$\left. + 0.00217688z^{2} + \ldots + 9.49115 \times 10^{-82} z^{27}\right).$$

In terms of running time, it takes less than 3 s to process all regular expressions of Table 3 on `DRR000019` using an Intel Xeon CPU E5-2609@2.40GHz running Linux 3.2.0. The same task on `DRR000020` takes 17 min. Note that this non linear increase (`DRR000020` is only roughly 40 times longer than `DRR000019`) is mainly due to the arrangement of degenerated position in the two dataset. Indeed, while both datasets have roughly the same rate of degenerated position (∼0.07 %), `DRR000019` contains much more run of "`N`" than `DRR000020`. For example, "`NNNNN`" appears 59 times in `DRR000019` and 4,486 times (roughly 80 times more) in `DRR000020`.

Conclusions

In this paper, we proposed a new algorithm to perform pattern matching of regular expression in sequences containing degenerated letters. Using NFA (linear size with the size regular expression) rather than DFA (exponentional size with the size regular expression), this algorithm is a dramatic improvement of the previous one [7]. Inspiring from the so- called "lazy determinization" used in pattern matching, we suggest the new concept of "lazy MCE" which only explores locally the combinatorics of possible occurrences. As a consequence, our new algorithm proves itself able to process NGS datasets with regular expression ranging for simple to highly complex in a reasonable time.

Our algorithm complexity is essentially linear with the dataset size but highly dependent on the content in degenerated letters of the considered dataset. When degenerated letters are rare enough (which is often the case in NGS datasets), the additional computational cost for processing lazy MCE steps remains reasonable in comparison with the cost of the classical deterministic lazy NFA steps. But the lazy MCE part can significantly increase the computational costs when dealing with too many degenerated positions, especially when considering highly complex regular expression.

However, for common DNA/RNA regular expression and NGS dataset, our algorithm will usually produce more reliable results than the simple NFA matching one with a running time of comparable order of magnitude.

Our algorithm is implemented in the `countmotif` program and freely available as part of the `fsaLib` package:

http://www.math-info.univ-paris5.fr/~delosvin/index.php?choix=4

Acknowledgements This work received the financial support of Sorbonne Paris-Cité in the context of the project "SA-Flex" (Structural Alphabet taking into account protein Flexibility).

References

1. Allauzen, C., Mohri, M.: A unified construction of the glushkov, follow, and antimirov automata. In: Královic, R., Urzyczyn, P. (eds.) Mathematical Foundations of Computer Science 2006. Lecture Notes in Computer Science, vol. 4162, pp. 110–121. Springer, Berlin/Heidelberg (2006)
2. Gilles, A., Meglécz, E., Pech, N., Ferreira, S.: Thibaut Malausa, and Jean-François Martin. Accuracy and quality assessment of 454 gs-flx titanium pyrosequencing. BMC Genomics **12**(1), 245 (2011)
3. Hopcroft, J.E., Motwani, R., Ullman, J.D.: Introduction to Automata Theory, Languages, and Computation, 3rd edn. Addison-Wesley, Boston (2006)
4. IUPAC: International Union of Pure and Applied Chemistry (2009). http://www.iupac.org
5. Lladser, M.E.: Mininal markov chain embeddings of pattern problems. In: Information Theory and Applications Workshop, pp. 251–255 (2007)
6. Nuel, G.: Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. J. Appl. Probab. **45**(1), 226–243 (2008)
7. Nuel, G.: Counting patterns in degenerated sequences. In: Pattern Recognition in Bioinformatics, pp. 222–232. Springer, New York (2009)
8. Zagordi, O., Klein, R., Däumer, M., Beerenwinkel, N.: Error correction of next-generation sequencing data and reliable estimation of hiv quasispecies. Nucleic Acids Res. **38**(21), 7400–7409 (2010)

Generalized Variance Estimations of Normal-Poisson Models

Célestin C. Kokonendji and Khoirin Nisa

Abstract This chapter presents three estimations of generalized variance (i.e., determinant of covariance matrix) of normal-Poisson models: maximum likelihood (ML) estimator, uniformly minimum variance unbiased (UMVU) estimator, and Bayesian estimator. First, the definition and some properties of normal-Poisson models are established. Then ML, UMVU, and Bayesian estimators for generalized variance are derived. Finally, a simulation study is carried out to assess the performance of the estimators based on their mean square error (MSE).

Keywords Covariance matrix • Determinant • Normal stable Tweedie • Maximum likelihood • UMVU • Bayesian estimator

Introduction

In multivariate analysis, generalized variance (i.e., determinant of covariance matrix) has important roles in the descriptive analysis and inferences. It is the measure of dispersion within multivariate data which explains the variability and the spread of observations. Its estimation usually based on the determinant of the sample covariance matrix. Many studies related to the generalized variance estimation have been done by some researchers; see, e.g., [1–3] under normality and non-normality hypotheses.

A normal-Poisson model is composed by distributions of random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)^{\mathrm{T}}$ with $k > 1$, where X_j is a univariate Poisson variable, and $(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k)$ given X_j are k-1 real independent Gaussian variables with variance X_j. It is a particular part of normal stable Tweedie (NST) models [4] with $p = 1$ where p is the power variance parameter of distributions within the Tweedie family. This model was introduced in [4] for the particular case of normal-Poisson with $j = 1$. Also, normal-Poisson is the only NST model which has a discrete component, and it is correlated to the continuous normal parts.

C.C. Kokonendji (✉) • K. Nisa
Laboratoire de Mathématiques de Besançon, University of Franche-Comté, Besançon, France
e-mail: celestin.kokonendji@univ-fcomte.fr; khoirin.nisa@univ-fcomte.fr

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.29_21

In literature, there is also a model known as Poisson-Gaussian [5–7] which is completely different from normal-Poisson. For any value of j, a normal-Poisson$_j$ model has only one Poisson component and k-1 normal (Gaussian) components, while a Poisson-Gaussian$_j$ model has j Poisson components and k-j Gaussian components. Poisson-Gaussian is also a particular case of simple quadratic natural exponential family (NEF) [5] with variance function $\mathbf{V}_F(\mathbf{m}) = \mathbf{Diag}_k(m_1, \ldots, m_j, 1, \ldots, 1)$, where $\mathbf{m} = (m_1, \ldots, m_k)$ is the mean vector and its generalized variance function is $\det\mathbf{V}_F(\mathbf{m}) = m_1, \ldots, m_j$. The estimations of generalized variance of Poisson-Gaussian can be seen in [8, 9].

Motivated by generalized variance estimations of Poisson-Gaussian, we present our study on multivariate normal-Poisson models and the estimations of their generalized variance using ML, UMVU, and Bayesian estimators.

Normal-Poisson Models

In this section, we establish the definition of normal-Poisson$_j$ models as generalization of normal-Poisson$_1$ model which was introduced in [4], and then we give some properties.

Definition 2.1 For a k-dimensional normal-Poisson random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)^{\mathrm{T}}$ with $k > 1$, it must hold that

1. X_j follows a univariate Poisson distribution.
2. $\left(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k\right) =: \mathbf{X}_j^c \Big| X_j$ are independent normal variables with mean 0 and variance X_j, i.e., $\mathbf{X}_j^c \Big| X_j \sim$ i.i.d. $N\left(0, X_j\right)$.

In order to satisfy the second condition, we need $X_{\mathrm{j}} > 0$, but in practice it is possible to have $x_j = 0$ in the Poisson sample. In this case, the corresponding normal components are degenerated as δ_0 which makes their values become 0s.

The NEF $F_t = F(\mu_t)$ of a k-dimensional normal-Poisson random vector $\mathbf{X}$ is generated by

$$\mu_t\left(d\mathbf{x}\right) = \frac{t^{x_j}\left(x_j!\right)^{-1}}{\left(2\pi x_j\right)^{(k-1)/2}} \exp\left(-t - \frac{1}{2x_j}\sum_{\ell \neq j} x_\ell^2\right) I_{x_j \in \mathsf{N}\backslash\{0\}} \delta_{x_j}\left(dx_j\right) \prod_{\ell \neq j} dx_\ell,$$

for a fixed power of convolution $t > 0$, where I_{A} is the indicator function of the set A and δ_{x_j} is the Dirac measure at x_j.. Since $t > 0$, then $\mu_{\mathrm{t}} := \mu^{*\mathrm{t}}$ is an infinitely divisible measure.

The cumulant function which is the log of the Laplace transform of μ_{t}, i.e., $\mathbf{K}_{\mu_t}\left(\boldsymbol{\theta}\right) = log \int_{\mathsf{R}^k} exp\left(\boldsymbol{\theta}^T\mathbf{x}\right)\mu_t\left(d\mathbf{x}\right)$, is given by

$$\mathbf{K}_{\mu_t}(\boldsymbol{\theta}) = texp\left(\theta_j + \frac{1}{2}\sum_{\ell \neq j}\theta_\ell^2\right). \tag{1}$$

The function $\mathbf{K}_{\mu_t}(\boldsymbol{\theta})$ in (1) is finite for all $\boldsymbol{\theta}$ in the canonical domain:

$$\boldsymbol{\Theta}(\mu_t) = \left\{\boldsymbol{\theta} \in R^k; \boldsymbol{\theta}^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_j^c := \theta_j + \sum_{\ell \neq j}\theta_\ell^2/2 < 0\right\}$$

with

$$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^T \quad \text{and} \quad \tilde{\boldsymbol{\theta}}_j^c := \left(\theta_1, \ldots, \theta_{j-1}, \theta_j = 1, \theta_{j+1}, \ldots, \theta_k\right)^T. \tag{2}$$

The probability distribution of normal-Poisson$_j$ is

$$P(\boldsymbol{\theta}; t)(d\mathbf{x}) = \exp\left\{\boldsymbol{\theta}^T\mathbf{x} - \mathbf{K}_{\mu_t}(\boldsymbol{\theta})\right\}\mu_t(d\mathbf{x})$$

which is a member of NEF $F(\mu_t) = \{P(\boldsymbol{\theta}; t); \boldsymbol{\theta} \in \boldsymbol{\Theta}(\mu_t)\}$.

From (1), we can calculate the first derivative of the cumulant function that produces a k-vector as the mean vector of F_{μ_t} and also its second derivative which is a $k \times k$ matrix that represents the covariance matrix. Using notations in (2), we obtain

$$\mathbf{K}'_{\mu_t}(\boldsymbol{\theta}) = \mathbf{K}_{\mu_t}(\boldsymbol{\theta}) \cdot \tilde{\boldsymbol{\theta}}_j^c \quad \text{and} \quad \mathbf{K}''_{\mu_t}(\boldsymbol{\theta}) = \mathbf{K}_{\mu_t}(\boldsymbol{\theta})\left\lfloor \tilde{\boldsymbol{\theta}}_j^c \tilde{\boldsymbol{\theta}}_j^{cT} + \mathbf{I}_k^{0_j}\right\rfloor$$

with $\mathbf{I}_k^{0_j} = \mathbf{Diag}_k\left(1, \ldots, 1, 0_j, 1, \ldots, 1\right)$.

The cumulant function presented in (1) and its derivatives are functions of the canonical parameter $\boldsymbol{\theta}$. For practical calculation, we need to use the mean parameterization:

$$P(\mathbf{m}; F_t) := P(\boldsymbol{\theta}(\mathbf{m}); \mu_t)$$

with $\boldsymbol{\theta}(\mathbf{m})$ is the solution in $\boldsymbol{\theta}$ of equation $\mathbf{m} = \mathbf{K}'_{\mu_t}(\boldsymbol{\theta})$.

The variance function of a normal-Poisson$_j$ model which is the variance-covariance matrix in term of mean parameterization is obtained through the second derivative of the cumulant function, i.e., $\mathbf{V}_{F_t}(\mathbf{m}) = \mathbf{K}''_{\mu_t}[\boldsymbol{\theta}(\mathbf{m})]$. Then we have

$$\mathbf{V}_{F_t}(\mathbf{m}) = \frac{1}{m_j}\mathbf{m}\mathbf{m}^{\mathrm{T}} + \mathbf{Diag}_k\left(m_j, \ldots, m_j, 0_j, m_j, \ldots, m_j\right) \tag{3}$$

with $m_j > 0$ and $m_\ell \in R,\ \ell \neq j$.

For $j = 1$, the covariance matrix of $\mathbf{X}$ can be expressed as below

$$V_{F_t}(m) = \left[\begin{array}{c|ccccc}
m_1 & m_2 & \cdots & m_j & \cdots & m_k \\
\hline
m_2 & m_1^{-1}m_2^2+m_1 & \cdots & m_1^{-1}m_2m_j & \cdots & m_1^{-1}m_2m_k \\
\vdots & \vdots & \ddots & \vdots & & \vdots \\
m_j & m_1^{-1}m_jm_2 & \cdots & m_1^{-1}m_j^2+m_1 & \cdots & m_1^{-1}m_jm_k \\
\vdots & \vdots & & \vdots & \ddots & \vdots \\
m_k & m_1^{-1}m_km_2 & \cdots & m_1^{-1}m_km_j & \cdots & m_1^{-1}m_k^2+m_1
\end{array}\right].$$

Indeed, for the covariance matrix above, one can use the following particular Schur representation of the determinant

$$\det\begin{pmatrix} \gamma & \mathbf{a}^T \\ \mathbf{a} & \mathbf{A} \end{pmatrix} = \gamma\det\left(\mathbf{A} - \gamma^{-1}\mathbf{a}\mathbf{a}^T\right) \tag{4}$$

with the non-null scalar $\gamma = m_1$, the vector $\mathbf{a} = (m_2, \cdots, m_k)^T$, and the $(k-1) \times (k-1)$ matrix $\mathbf{A} = \gamma^{-1}\mathbf{a}\mathbf{a}^T + m_1\mathbf{I}_{k-1}$, where $\mathbf{I}_j = \mathbf{Diag}_j\ (1, \cdots, 1)$ is the $j \times j$ unit matrix.

Consequently, the determinant of the covariance matrix $\mathbf{V}_{F_t}\ (\mathbf{m})$ for $j = 1$ is

$$\det\mathbf{V}_{F_t}\ (\mathbf{m}) = m_1^k$$

Then, it is trivial to show that for $j\in\{1, \ldots, k\}$, the generalized variance of normal-Poisson$_j$ model is given by

$$\det\mathbf{V}_{F_t}\ (\mathbf{m}) = m_j^k \tag{5}$$

with $m_j > 0, m_\ell \in R,\ \ell \neq j$. (5) expresses that the generalized variance of normal-Poisson models depends mainly on the mean of the Poisson component (and the dimension space $k > 1$) .

Among NST models, normal-gamma which is also known as gamma-Gaussian is the only model that has been characterized completely; see [5] or [10] for characterization by variance function and [11] for characterization by generalized variance function. For normal-Poisson models, here we give our result regarding to characterization by variance function and generalized variance. We state the results in the following theorems without proof.

Theorem 2.1 Let $k \in\{2, 3, \ldots\}$ and $t > 0$. If an NEF F_t satisfies (3), then, up to affinity, F_t is of normal-Poisson model.

Theorem 2.2 Let $F_t = F(\mu_t)$ be an infinitely divisible NEF on R^k such that

1. The canonical domain $\boldsymbol{\Theta}(\mu) = R^k$
2. $\det\mathbf{K}''_{\mu}\ (\boldsymbol{\theta}) = t\exp\left(k \cdot \boldsymbol{\theta}^T\tilde{\boldsymbol{\theta}}_j^c\right)$

for $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}_j^c$ given in (2). Then, up to affinity and power convolution, F_t is of normal-Poisson model.

All the technical details of proofs will be given in our article which is in preparation. In fact, the proof of Theorem 2.1 obtained by algebraic calculations and by using some properties of NEF is described in Proposition 2.1 below. An idea to proof Theorem 2.2 can be obtained using the infinite divisibility property of normal-Poisson for which this proof is the solution to the particular Monge–Ampère equation [12]: $\det\mathbf{K}''_\mu(\boldsymbol{\theta}) = texp\left(k \cdot \boldsymbol{\theta}^T\tilde{\boldsymbol{\theta}}_j^c\right)$. Gikhman and Skorokhod [13] showed that if μ is an infinitely divisible measure, then there exist a symmetric nonnegative definite $d \times d$ matrix $\boldsymbol{\Sigma}$ with rank k-1 and a positive measure ν on R^k such that

$$\mathbf{K}''_\mu(\boldsymbol{\theta}) = \boldsymbol{\Sigma} + \int_{\mathsf{R}^k} \mathbf{x}\mathbf{x}^T \exp\left(\boldsymbol{\theta}^T\mathbf{x}\right)\nu(d\mathbf{x}).$$

The Lévy–Khintchine formula of infinite divisibility distribution is also applied.

Proposition 2.1 Let μ and $\tilde{\mu}$ be two σ-finite positive measures on R^k such that $F = F(\mu)$, $\tilde{F} = F(\tilde{\mu})$, and $\mathbf{m} \in \mathbf{M_F}$.

1. If there exists $(\mathbf{d},c) \in \mathrm{R}^k\mathrm{x}\mathrm{R}$ such that $\tilde{\mu}(d\mathbf{x}) = exp\left\{\mathbf{d}^T\mathbf{x}\right\rangle + c\Big\}\,\mu(d\mathbf{x})$, then $F = \tilde{F} : \boldsymbol{\Theta}(\tilde{\mu}) = \boldsymbol{\Theta}(\mu) - \mathbf{d}$ and $K_{\tilde{\mu}}(\boldsymbol{\theta}) = K_\mu(\boldsymbol{\theta} + \mathbf{d}) + c$, for $\overline{\mathbf{m}} = \mathbf{m} \in \mathbf{M_F}$, $\mathbf{V}_{\tilde{F}}(\overline{\mathbf{m}}) = \mathbf{V}_F(\mathbf{m})$, and $\det\mathbf{V}_{\tilde{F}}(\overline{\mathbf{m}}) = \det\mathbf{V_F}(\mathbf{m})$.
2. If $\tilde{\mu} = \phi_*\mu$ is the image measure of μ by the affine transformation $\phi(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$, where $\mathbf{A}$ is a $k \times k$ nondegenerate matrix and $\mathbf{b} \in \mathrm{R}^k$, then $\boldsymbol{\Theta}(\tilde{\mu}) = \mathbf{A}^T\boldsymbol{\Theta}(\mu)$ and $K_{\tilde{\mu}}(\boldsymbol{\theta}) = K_\mu\left(\mathbf{A}^T\boldsymbol{\theta}\right) + \mathbf{b}^T\boldsymbol{\theta}$; for $\overline{\mathbf{m}} = \mathbf{A}\mathbf{m} + \mathbf{b} \in \phi(\mathbf{M_F})$, $\mathbf{V}_{\tilde{F}}(\overline{\mathbf{m}}) = \mathbf{A}\mathbf{V}_F\left(\phi^{-1}(\overline{\mathbf{m}})\right)\mathbf{A}^T$, and $\det\mathbf{V}_{\tilde{F}}(\overline{\mathbf{m}}) = (\det\mathbf{A})^2\det\mathbf{V}_F(\mathbf{m})$.
3. If $\tilde{\mu} = \mu^{*t}$ is the t-th convolution power of μ for $t > 0$, then $\boldsymbol{\Theta}(\tilde{\mu}) = \boldsymbol{\Theta}(\mu)$ and $K_{\tilde{\mu}}(\boldsymbol{\theta}) = tK_\mu(\boldsymbol{\theta})$; for $\overline{\mathbf{m}} = t\mathbf{m} \in t\mathbf{M}_F$, $\mathbf{V}_{\tilde{F}}(\overline{\mathbf{m}}) = t\mathbf{V}_F\Big(\phi t^{-1}(\overline{\mathbf{m}})$, and $\det\mathbf{V}_{\tilde{F}}(\overline{\mathbf{m}}) = t^k det\mathbf{V}_F(\mathbf{m})$.

Proposition 2.1 shows that the generalized variance function $\det\mathbf{V}_F(\mathbf{m})$ of F is invariant for any element of its generating measure (Part 1) and for the affine transformation $\phi(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ such that $\det\mathbf{A} = \pm 1$, particularly for a translation $\mathbf{x} \to \mathbf{x} + \mathbf{b}$ (Part 2).

A reformulation of Theorem 2.2, by changing the canonical parameterization into mean parameterization, is stated in the following theorem.

Theorem 2.3 Let $F_t = F(\mu_\mathrm{t})$ be an infinitely divisible NEF on R^k such that

1. $m_j > 0$ and $m_\ell \in \mathrm{R}$ with $\ell \neq j$
2. $\det\mathbf{V}_F(\mathbf{m}) = m_j^k$.

Then F_t is of normal-Poisson type.

Theorem 2.3 is equivalent to Theorem 2.2. The former is used for the estimation of generalized variance, and the latter is used for characterization by generalized variance.

Generalized Variance Estimations

Here we present three methods for generalized variance estimations of normal-Poisson models $P\left(\mathbf{m}; Ft\right) \in F_t = F\left(\mu_t\right)$, and then we report the result of our simulation study.

Consider $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be random vectors i.i.d. from $P(\mathbf{m}; F_t)$ of normal-Poisson models, and we denote $\overline{\mathbf{X}} = \left(\mathbf{X}_1 + \cdots + \mathbf{X}_n\right)/n = \left(\overline{X}_1, \cdots, \overline{X}_k\right)^T$ as the sample mean with positive j-th component $\overline{X}_j$. The followings are ML, UMVU, and Bayesian generalized variance estimators.

Maximum Likelihood Estimator

Proposition 3.1 The ML estimator of $\det \mathbf{V}_{F_t}\left(\mathbf{m}\right) = m_j^k$ is given by

$$T_{n,t} = \det \mathbf{V}_{F_t}\left(\overline{\mathbf{X}}\right) = \left(\overline{X}_j\right)^k. \tag{6}$$

Proof The ML estimator above is easily obtained by replacing m_j in (5) with its ML estimator $\overline{X}_j$. □

Uniformly Minimum Variance Unbiased Estimator

Proposition 3.2 The UMVU estimator of $\det \mathbf{V}_{F_t}\left(\mathbf{m}\right) = m_j^k$ is given by

$$U_{n,t} = n^{-k+1}\overline{X}_j\left(n\overline{X}_j - 1\right)\ldots\left(n\overline{X}_j - k + 1\right), \quad \text{if } n\overline{X}_j \geq k. \tag{7}$$

Proof This UMVU estimator is obtained using intrinsic moment formula of univariate Poisson distribution as follows:

$$\mathrm{E}\left[X\left(X-1\right)\ldots\left(X-k+1\right)\right] = m_j^k.$$

Letting $Y = n\overline{X}_j$ gives the result that (7) is the UMVU estimator of (5), because, by the completeness of NEFs, the unbiased estimation is unique. So, we deduced the desired result. ■

A deep discussion about ML and UMVU methods on generalized variance estimations can be seen in [9] for NEF and [4] for NST models.

Bayesian Estimator

Proposition 3.3 Under assumption of prior gamma distribution of m_j with parameter $\alpha > 0$ and $\beta > 0$, the Bayesian estimator of $\det \mathbf{V}_{F_t}(\mathbf{m}) = m_j^k$ is given by

$$B_{n,t,\alpha,\beta} = \left(\frac{\alpha + n\overline{X}_j}{\beta + n}\right)^k. \tag{8}$$

Proof Let $X_{1j}, \cdots, X_{nj}$ given m_j are Poisson(m_j) with probability mass function

$$P\left(X_{ij} = x_{ij} \middle| m_j\right) = \frac{m_j^{x_{ij}}}{x_{ij}!} e^{-m_j} = p\left(x_{ij} \middle| m_j\right).$$

Assuming that m_j follows gamma(α, β), then the prior probability distribution function of m_j is given by

$$f\left(m_j; \alpha, \beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} m_j^{\alpha-1} e^{-\beta m_j} \text{ for } m_j > 0 \text{ and } \alpha, \beta > 0$$

where $\Gamma(\alpha)$ is the gamma function: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$. Using the Bayes theorem, the posterior distribution of m_j given an observation sequence can be expressed as

$$f\left(m_j \middle| x_{ij}; \alpha, \beta\right) = \frac{p\left(x_{ij} \middle| m_j\right) f\left(m_j; \alpha, \beta\right)}{\int_{m_j > 0} p\left(x_{ij} \middle| m_j\right) f\left(m_j; \alpha, \beta\right) dm_j}$$
$$= \frac{(\beta + 1)^{\alpha + x_{ij}}}{\Gamma\left(\alpha + x_{ij}\right)} m_j^{\alpha + x_{ij} - 1} e^{-(\beta+1)m_j}$$

which is a gamma density with parameters $\alpha' = x_{ij} + \alpha$ and $\beta' = 1 + \beta$. Then with random sample $X_{1j}, \ldots, X_{nj}$, the posterior will be gamma $\left(\alpha + n\overline{X}_j, \beta + n\right)$. The Bayesian estimator of m_j is given by the mean of the posterior distribution, i.e., $\widehat{m}_b = \frac{\alpha + n\overline{X}_j}{\beta + n}$, and then this concludes the proof. ■

The choice of α and β depends on the information of m_j. Notice that for any positive value $c \in (0, \infty)$, if $\alpha = c\overline{X}_j$ and $\beta = c$, then the Bayesian estimator is the same as ML estimator. In practice, the parameter of prior distribution of

m_j must be known or can be assumed confidently before the generalized variance estimation. One can see, e.g., [14–16] for more details about Bayesian inference on m_j (univariate Poisson parameter).

Simulation Study

In order to look at the performances of ML, UMVU, and Bayesian estimators of the generalized variance, we have done a Monte Carlo simulation using R software [17]. We have generated $k = 2$, 4, 6, 8 dimensional data from multivariate normal-Poisson distribution $F(\mu_t)$ with $m_j = 1$. Fixing $j = 1$, we set several sample sizes n varied from 5 until 300, and we generated 1,000 samples for each sample size. For calculating the Bayesian estimator, in this simulation we assume that the parameters of prior distribution depend on sample mean of Poisson component, $\overline{X}_j$, and the dimension k. Then we set three different prior distributions: gamma $\left(\overline{X}_j, k\right)$, gamma $\left(\overline{X}_j, k/2\right)$, and gamma $\left(\overline{X}_j, k/3\right)$.

We report the results of the generalized variance estimations using the three methods in Table 1. From these values, we calculated the mean square error (MSE) of each method over 1,000 data sets using this following formula

$$MSE\left(\overset{\wedge}{GV}\right) = \frac{1}{1,000}\sum_{i=1}^{1,000}\left(\overset{\wedge}{GV}_i - m_j^k\right)^2$$

where $\overset{\wedge}{GV}$ is the estimate of m_j^k using each method.

From the values in Table 1, we can observe different performances of ML estimator ($T_{n,t}$), UMVU estimator ($U_{n,t}$), and Bayesian estimator ($B_{n,t,\alpha,\beta}$) of the generalized variance. The values of $T_{n,t}$ and $B_{n,t,\alpha,\beta}$ converge, while the values of $U_{n,t}$ do not, but $U_{n,t}$ which is the unbiased estimator always approximate the parameter ($m_1^k = 1$) and closer to the parameter than $T_{n,t}$ and $B_{n,t,\alpha,\beta}$ for small sample sizes $n \leq 25$. For all methods, the standard error of the estimates decreases when the sample size increases. The Bayesian estimator with gamma $\left(\overline{X}_j, k/2\right)$ prior distribution, i.e., $B_{n,t,\overline{X}_j,k/2}$, is exactly the same as $T_{n,t}$ for $k = 2$. This is because in this case, the Bayesian and ML estimators of m_1 are the same (i.e., $c = 1$).

The goodness of Bayesian estimator depends on the parameter of prior distribution, α and β. From our simulation, the result shows that smaller parameter β gives greater standard error to the estimations in small sample sizes, and the accuracy of $B_{n,t,\alpha,\beta}$ with respect to β varies with dimensions k. However, they are all asymptotically unbiased.

There are more important performance characterizations for an estimator than just being unbiased. The MSE is perhaps the most important of them. It captures the

Table 1 The expected values (with standard error) of $T_{n,t}$, $U_{n,t}$, and $B_{n,t,\alpha,\beta}$ with $m_1 = 1$ and $k \in \{2, 4, 6, 8\}$ (target values $m_1^k = 1$)

$k = 2n$	$T_{n,t}$	$U_{n,t}$	$B_{n,t,\overline{X}_j,k}$	$B_{n,t,\overline{X}_j,k/2}$	$B_{n,t,\overline{X}_j,k/3}$
k + 1	1.2790 (1.3826)	0.9533 (1.2050)	0.8186 (0.8849)	1.2790 (1.3826)	1.5221 (1.6454)
k + 5	1.1333 (0.8532)	0.9915 (0.8000)	0.8955 (0.6742)	1.1333 (0.8532)	1.2340 (0.9290)
k + 10	1.1121 (0.6295)	1.0276 (0.6056)	0.9589 (0.5428)	1.1121 (0.6295)	1.1714 (0.6631)
25	1.0357 (0.4256)	0.9959 (0.4175)	0.9604 (0.3946)	1.0357 (0.4256)	1.0628 (0.4367)
60	1.0090 (0.2526)	0.9924 (0.2505)	0.9767 (0.2445)	1.0090 (0.2526)	1.0201 (0.2553)
100	1.0086 (0.1988)	0.9986 (0.1979)	0.9890 (0.1950)	1.0086 (0.1988)	1.0153 (0.2002)
300	0.9995 (0.1141)	0.9962 (0.1140)	0.9929 (0.1134)	0.9995 (0.1141)	1.0017 (0.1144)
$k = 4n$	$T_{n,t}$	$U_{n,t}$	$B_{n,t,\overline{X}_j,k}$	$B_{n,t,\overline{X}_j,k/2}$	$B_{n,t,\overline{X}_j,k/3}$
k + 1	2.3823 (4.6248)	0.9460 (2.5689)	0.4706 (0.9135)	1.2859 (2.4964)	1.9190 (3.7254)
k + 5	1.6824 (2.4576)	0.9531 (1.6995)	0.5890 (0.8605)	1.1491 (1.6786)	1.4756 (2.1555)
k + 10	1.4664 (1.6345)	1.0027 (1.2456)	0.7072 (0.7882)	1.1328 (1.2626)	1.3430 (1.4969)
25	1.2711 (1.0895)	1.0169 (0.9327)	0.8212 (0.7039)	1.0930 (0.9368)	1.2079 (1.0353)
60	1.0978 (0.5682)	0.9961 (0.5288)	0.9060 (0.4689)	1.0287 (0.5324)	1.0741 (0.5559)
100	1.0589 (0.4209)	0.9983 (0.4028)	0.9419 (0.3744)	1.0180 (0.4046)	1.0451 (0.4154)
300	1.0273 (0.2305)	1.0071 (0.2271)	0.9874 (0.2215)	1.0138 (0.2275)	1.0228 (0.2295)
$k = 6n$	$T_{n,t}$	$U_{n,t}$	$B_{n,t,\overline{X}_j,k}$	$B_{n,t,\overline{X}_j,k/2}$	$B_{n,t,\overline{X}_j,k/3}$
k + 1	4.7738 (13.9827)	0.9995 (4.7073)	0.2593 (0.7594)	1.2514 (3.6655)	2.3548 (6.8972)
k + 5	2.9818 (6.2595)	0.9958 (2.7565)	0.3689 (0.7743)	1.1825 (2.4823)	1.8446 (3.8723)
k + 10	2.2232 (4.0454)	1.0124 (2.2131)	0.4733 (0.8612)	1.1406 (2.0756)	1.5778 (2.8709)
25	1.6399 (2.2478)	0.9555 (1.4833)	0.5708 (0.7824)	1.0513 (1.4410)	1.3076 (1.7923)
60	1.2479 (0.9978)	0.9827 (0.8226)	0.7778 (0.6220)	1.0283 (0.8222)	1.1319 (0.9051)
100	1.1830 (0.7646)	1.0235 (0.6800)	0.8853 (0.5722)	1.0517 (0.6798)	1.1151 (0.7207)
300	1.0530 (0.3758)	1.0022 (0.3608)	0.9539 (0.3404)	1.0119 (0.3612)	1.0322 (0.3684)
$k = 8n$	$T_{n,t}$	$U_{n,t}$	$B_{n,t,\overline{X}_j,k}$	$B_{n,t,\overline{X}_j,k/2}$	$B_{n,t,\overline{X}_j,k/3}$
k + 1	8.5935 (31.9230)	0.8677 (5.4574)	0.1232 (0.4576)	1.0535 (3.9134)	2.5038 (9.3010)
k + 5	4.7573 (12.5015)	0.8468 (3.0478)	0.1856 (0.4878)	1.0065 (2.6448)	1.9345 (5.0836)
k + 10	3.6816 (9.0892)	1.0394 (3.2258)	0.2994 (0.7392)	1.1394 (2.8130)	1.8789 (4.6387)
25	2.9055 (6.3150)	1.1341 (2.9623)	0.4314 (0.9377)	1.2129 (2.6362)	1.7675 (3.8416)
60	1.6201 (1.8804)	1.0511 (1.3062)	0.6794 (0.7885)	1.1035 (1.2807)	1.3059 (1.5156)
100	1.2890 (1.0907)	0.9850 (0.8667)	0.7541 (0.6381)	1.0199 (0.8630)	1.1308 (0.9569)
300	1.1056 (0.5378)	1.0086 (0.4968)	0.9199 (0.4474)	1.0213 (0.4967)	1.0578 (0.5145)

bias and the variance of the estimator. For this reason, we compare the quality of the estimators using MSE in Table 2 which are presented graphically in Figs. 1, 2, 3, and 4. From these figures, we conclude that all estimators become more similar when the sample size increases. For small sample sizes, $B_{n,t,\overline{X}_j,k}$ always has the smallest MSE, while $T_{n,t}$ always has the greatest MSE (except for $k = 2$). For $n \leq 25$, $U_{n,t}$ is preferable than $T_{n,t}$. In this situation, the difference between $U_{n,t}$ and $T_{n,t}$ increases when the dimension increases and also the difference between $T_{n,t}$ and $B_{n,t,\alpha,\beta}$.

Table 2 The mean square error of $T_{n,t}$, $U_{n,t}$, and $B_{n,t,\alpha,\beta}$ of Table 1

$\boldsymbol{k=2n}$	$MSE(T_{n,t})$	$MSE(U_{n,t})$	$MSE(B_{n,t,\overline{X}_j,k})$	$MSE(B_{n,t,\overline{X}_j,k/2})$	$MSE(B_{n,t,\overline{X}_j,k/3})$
k + 1	1.9894	1.4542	0.8159	1.9894	2.9800
k + 5	0.7458	0.6401	0.4654	0.7458	0.9179
k + 10	0.4088	0.3675	0.2963	0.4088	0.4690
25	0.1824	0.1743	0.1573	0.1824	0.1947
60	0.0639	0.0628	0.0603	0.0639	0.0656
100	0.0396	0.0391	0.0381	0.0396	0.0403
300	0.0130	0.0130	0.0129	0.0130	0.0131
$\boldsymbol{k=4n}$	$MSE(T_{n,t})$	$MSE(U_{n,t})$	$MSE(B_{n,t,\overline{X}_j,k})$	$MSE(B_{n,t,\overline{X}_j,k/2})$	$MSE(B_{n,t,\overline{X}_j,k/3})$
k + 1	23.2999	6.6019	1.1149	6.3136	14.7231
k + 5	6.5055	2.8904	0.9093	2.8398	4.8724
k + 10	2.8891	1.5514	0.7071	1.6118	2.3585
25	1.2604	0.8702	0.5274	0.8862	1.1151
60	0.3324	0.2797	0.2287	0.2843	0.3146
100	0.1806	0.1622	0.1435	0.1640	0.1746
300	0.0539	0.0516	0.0492	0.0519	0.0532
$\boldsymbol{k=6n}$	$MSE(T_{n,t})$	$MSE(U_{n,t})$	$MSE(B_{n,t,\overline{X}_j,k})$	$MSE(B_{n,t,\overline{X}_j,k/2})$	$MSE(B_{n,t,\overline{X}_j,k/3})$
k + 1	209.7568	22.1589	1.1254	13.4989	49.4073
k + 5	43.1085	7.5980	0.9979	6.1952	15.7078
k + 10	17.8618	4.8981	1.0191	4.3278	8.5761
25	5.4622	2.2020	0.7964	2.0790	3.3071
60	1.0571	0.6769	0.4362	0.6769	0.8366
100	0.6181	0.4629	0.3406	0.4647	0.5327
300	0.1440	0.1302	0.1180	0.1306	0.1368
$\boldsymbol{k=8n}$	$MSE(T_{n,t})$	$MSE(U_{n,t})$	$MSE(B_{n,t,\overline{X}_j,k})$	$MSE(B_{n,t,\overline{X}_j,k/2})$	$MSE(B_{n,t,\overline{X}_j,k/3})$
k + 1	1,076.7380	29.8009	0.9782	15.3177	88.7698
k + 5	170.4059	9.3124	0.9012	6.9951	26.7168
k + 10	89.8046	10.4076	1.0373	7.9326	22.2895
25	43.5105	8.7931	1.2025	6.9949	15.3466
60	3.9204	1.7088	0.7246	1.6509	2.3907
100	1.2732	0.7515	0.4676	0.7452	0.9327
300	0.3003	0.2469	0.2066	0.2472	0.2681

In this simulation, $B_{n,t,\overline{X}_j,k}$ is the best estimator because of its smallest MSE, but in general we cannot say that Bayesian estimator is much better than ML and UMVU estimators since it depends on the prior distribution parameters. In fact, one would prefer $U_{n,t}$ as it is the unbiased estimator with the minimum variance. However, if in practice we know the information about prior distribution of m_j, we can get a better estimate (in the sense of having a lower MSE) than $U_{n,t}$ by using $B_{n,t,\alpha,\beta}$.

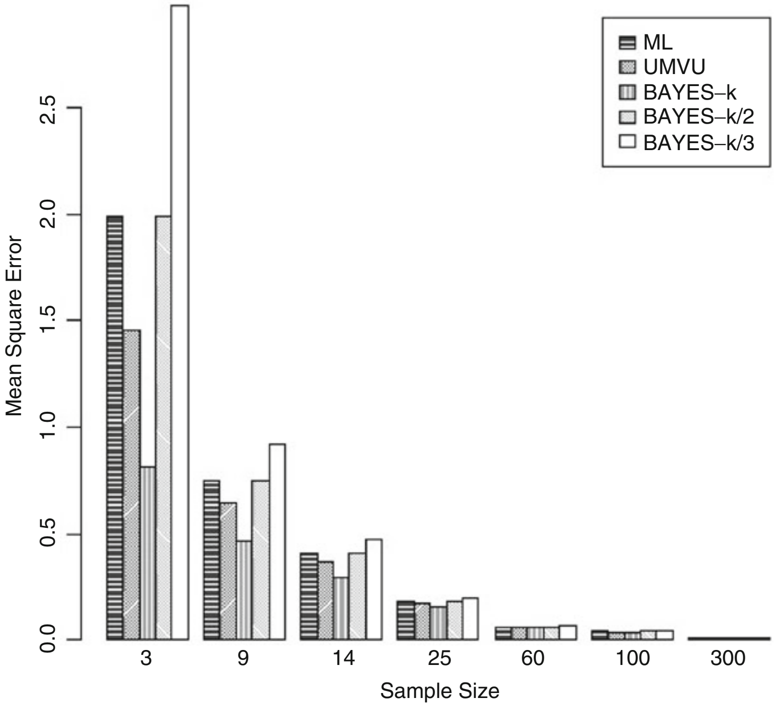


Fig. 1 MSE plot of $T_{n,t}$, $U_{n,t}$, $B_{n,t,\overline{x}_j,k}$,$B_{n,t,\overline{x}_j,k/2}$, and $B_{n,t,\overline{x}_j,k/3}$ for $k = 2$

Conclusion

In this chapter, we have established the definition and properties of normal-Poisson$_j$ models as a generalization of normal-Poisson$_1$ and showed that the generalized variance of normal-Poisson models depends mainly on the mean of the Poisson component. The estimations of generalized variance using ML, UMVU, and Bayesian estimators show that UMVU produces a better estimation than ML estimator, while compared to Bayesian estimator, UMVU is worse for some choice of prior distribution parameters, but it can be much better for other cases. However, all methods are consistent estimators, and they become more similar when the sample size increases.

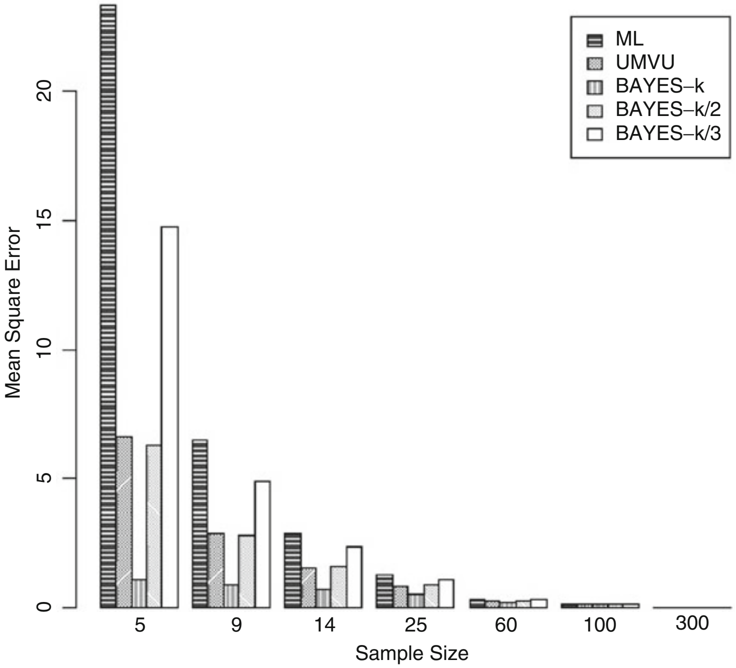


Fig. 2 MSE plot of $T_{n,t}$, $U_{n,t}$, $B_{n,t,\overline{x}_j,k}$,$B_{n,t,\overline{x}_j,k/2}$, and $B_{n,t,\overline{x}_j,k/3}$ for $k = 4$

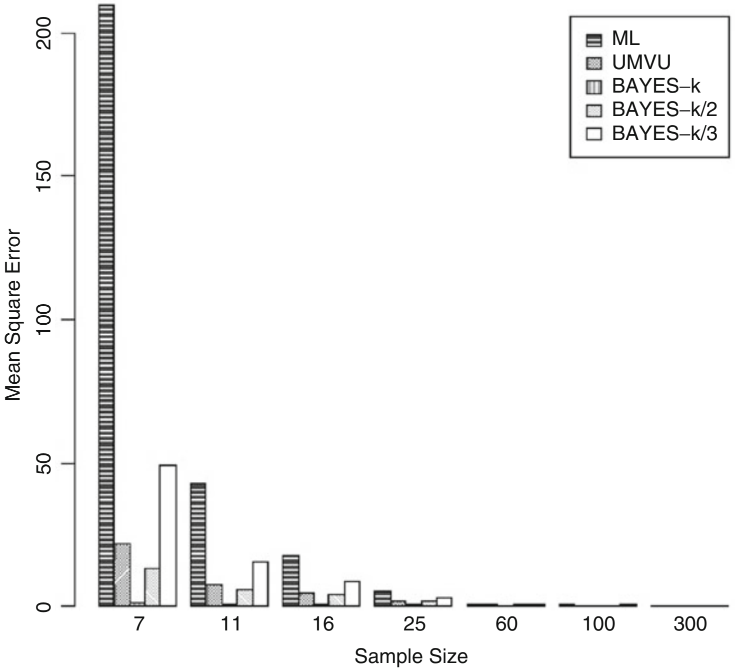


Fig. 3 MSE plot of $T_{n,t}$, $U_{n,t}$, $B_{n,t,\overline{x}_j,k}$,$B_{n,t,\overline{x}_j,k/2}$, and $B_{n,t,\overline{x}_j,k}$ for $k = 6$

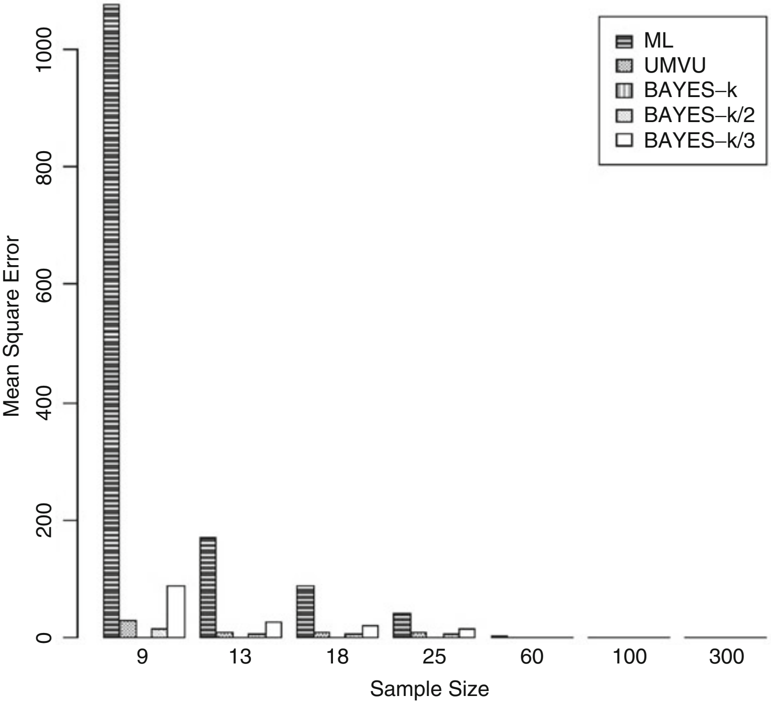


Fig. 4 MSE plot of $T_{n,t}$, $U_{n,t}$, $B_{n,t,\overline{x}_j,k}$,$B_{n,t,\overline{x}_j,k/2}$, and $B_{n,t,\overline{x}_j,k/3}$ for $k = 8$

References

1. Hassairi, A.: Generalized variance and exponential families. Ann. Stat. **27**(1), 374–385 (1999)
2. Kokonendji, C.C., Pommeret, D.: Estimateurs de la variance généralisée pour des familles exponentielles non gaussiennes. C. R. Acad. Sci. Ser. Math. **332**(4), 351–356 (2001)
3. Shorrock, R.W., Zidek, J.V.: An improved estimator of the generalized variance. Ann. Stat. **4**(3), 629–638 (1976)
4. Boubacar Maïnassara, Y., Kokonendji, C.C.: On normal stable Tweedie models and power generalized variance function of only one component. TEST **23**(3), 585–606 (2014)
5. Casalis, M.: The 2d + 4 simple quadratic natural exponential families on R^d. Ann. Stat. **24**(4), 1828–1854 (1996)
6. G. Letac, Le problem de la classification des familles exponentielles naturelles de $\mathbb{R}^d$ ayant une fonction variance quadratique, in Probability Measures on Groups IX, H. Heyer, Ed. Springer, Berlin, 1989, pp. 192–216.
7. Kokonendji, C.C., Masmoudi, A.: A characterization of Poisson-Gaussian families by generalized variance. Bernoulli **12**(2), 371–379 (2006)
8. Kokonendji, C.C., Seshadri, V.: On the determinant of the second derivative of a Laplace transform. Ann. Stat. **24**(4), 1813–1827 (1996)
9. Kokonendji, C.C., Pommeret, D.: Comparing UMVU and ML estimators of the generalized variance for natural exponential families. Statistics **41**(6), 547–558 (2007)
10. Kotz, S., Balakrishnan, N., Johnson, N.L.: Continuous Multivariate Distributions. Models and Application, vol. 1, 2nd edn. Wiley, New York (2000)

11. Kokonendji, C.C., Masmoudi, A.: On the Monge–Ampère equation for characterizing gamma-Gaussian model. Stat. Probab. Lett. **83**(7), 1692–1698 (2013)
12. Gutiérrez, C.E.: The Monge-Ampère Equation. Birkhäuser, Boston (2001). Boston: Imprint: Birkhäuser
13. Gikhman, I.I., Skorokhod, A.V.: The Theory of Stochastic Processes 2. Springer, New York (2004)
14. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis, 2nd edn. Springer, New York (1985)
15. Sultan, R., Ahmad, S.P.: Posterior estimates of Poisson distribution using R software. J. Mod. Appl. Stat. Methods **11**(2), 530–535 (2012)
16. Hogg, R.V.: Introduction to Mathematical Statistics, 7th edn. Pearson, Boston (2013)
17. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2009)

Vehicle Routing Problem with Uncertain Costs via a Multiple Ant Colony System

Nihat Engin Toklu, Luca Maria Gambardella, and Roberto Montemanni

Abstract We consider the capacitated vehicle routing problem (VRP) with uncertain travel costs, where the uncertainty represents the realistic factors like unfriendly weather conditions, traffic jams, etc. In this chapter, we present a multiple ant colony system approach in which ant colony optimization processes work concurrently to produce multiple solutions. Experimental results on different types of VRP instances (with clustered customers, randomly placed customers, and a mixed of the previous two), each instance having 200 customers, are finally discussed.

Keywords Vehicle routing problem • Robust optimization • Metaheuristic algorithms

Introduction

In the field of transportation, vehicle routing problem (VRP) is a well-known problem [2, 3, 10, 14, 17, 24]. A common variation of VRP is called the capacitated vehicle routing problem (CVRP), which is the variation we study here. In CVRP, we have a depot where a product is stored, a number of vehicles with a limited capacity, and a number of customers with various amounts of demands for stored products. The purpose is, by using the available vehicles, to distribute the products to the customers while minimizing the total cost of traveling. The important constraints of this problem are as follows: (a) The demand of each customer must be entirely satisfied by one visit of one vehicle; (b) a vehicle cannot carry more products than the dictated capacity, so the total demand on a vehicle's route must be less than or equal to the vehicle capacity; and (c) the route of a vehicle must start and end at the depot, without any intermediate visits to the depot.

To make CVRP clear, let us analyze the example in Fig. 1a. In this example, we have two vehicles available and three customers to satisfy. The cost of traveling

N.E. Toklu (✉) • L.M. Gambardella • R. Montemanni
Dalle Molle Institute for Artificial Intelligence (IDSIA—USI/SUPSI), Galleria 2,
Manno 6928, Switzerland
e-mail: engin@idsia.ch; luca@idsia.ch; roberto@idsia.ch

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.08_22

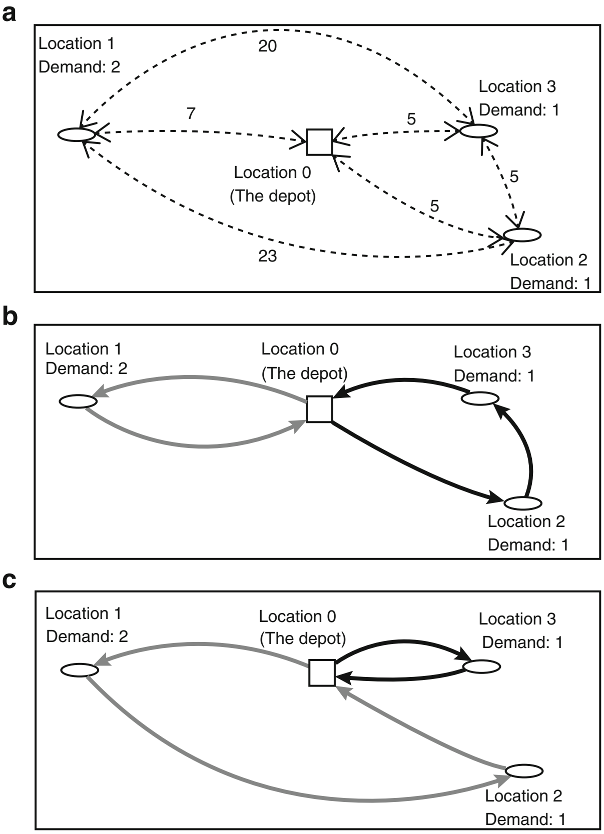


Fig. 1 (**a**) A CVRP example from [22]. Two vehicles are available, each with capacity 3. (**b**) A solution with total travel cost 29. (**c**) A solution with total travel cost 45

between the locations of the customers and the depot are given adjacent to the dotted arcs. Two solutions for this instance are given in Fig. 1b and 1c. Note that, because the capacity of a vehicle is three, a single vehicle cannot satisfy all the customers in one tour. Therefore, these solutions have to use two vehicle tours. Given that the objective is to minimize the total cost of traveling, it can be observed that the solution in Fig. 1b is more practical.

In a traditional optimization problem, the problem data would be known exactly. However, in the recent years, alternative perspectives have begun to draw attention, to make the optimization models more realistic. In these alternative perspectives, the common argument is that, in reality, the problem data cannot be known exactly because of factors that are difficult to predict. In the case of VRP, these factors would be unfriendly weather conditions, traffic jams, road constructions, etc., affecting the data of traveling cost/time between locations. Ignoring the uncertainty could lead to undesired situations. For example, a solution that looks optimal according

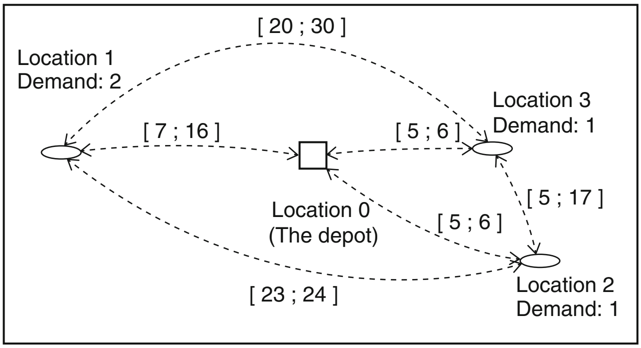


Fig. 2 A CVRPU counterpart of the instance shown in Fig. 1

to the mathematical model where the uncertainty is not considered could turn out to be far from optimal in reality. To avoid such undesirable situations, a perspective where the uncertainty is considered in the optimization model can be used. One such perspective is stochastic optimization [8], in which the goal is to do optimization when there is uncertainty, and the information about the uncertainty (probability distributions, etc.) is available. Another perspective is robust optimization [5–7, 16], where the probability distribution information about the uncertainty is not known and the uncertain data are usually expressed by intervals, instead of single numbers.

Now, let us look at a possible CVRPU counterpart of the instance in Fig. 1a, shown in Fig. 2, where the travel costs are intervals. An interesting effect of uncertainty can be observed here: when we consider the best-case values from the cost intervals, the solution shown in Fig. 1b is the more practical solution. But when we consider the worst-case values from the cost intervals, that solution becomes the less practical one.

In this chapter, we consider CVRP with uncertain travel costs (CVRPU). We assume that the probability distribution information is not available; therefore, we apply the perspective of robust optimization and put the travel costs as intervals into the optimization model.

An important concept in robust optimization is the degree of conservativeness, where conservativeness means the amount of pessimism in the assumptions during the optimization. For example, in the case of CVRP, a decision maker who is not conservative would assume that the travel on each path will go smoothly without any additional cost and do the optimization according to the lowest travel cost values possible from the travel cost intervals. On the other hand, a decision maker who is fully conservative would assume that the travel on each path will be problematic and do the optimization according to the highest travel cost values possible from the travel cost intervals. A partially conservative decision maker would assume that some paths will provide smooth travels and some paths will be problematic. Robust optimization methodologies which allow the decision maker to configure the degree of conservativeness are discussed in [4, 6].

Previous studies on VRP with robust optimization considerations are available in the literature. In [21], a mathematical programming approach is taken to solve a VRP where there is uncertainty in the customer demands. In [18], a robust mathematical model for VRP with deadlines is proposed, and, in [1], robust mathematical models for VRP with time window constraints are proposed. The methods listed above are exact methods, as they are designed to find the optimal values, given enough time. In the situations where the decision maker wants a near-optimal solution, within a limited amount of time, without demanding too much memory, metaheuristic approaches can be used. Within the category of metaheuristics, a previous study is [20], where a VRP with uncertain demands is solved using a particle swarm approach. Our approach that we propose here belongs to the category of metaheuristics. Differently from the study presented in [20], we consider that the uncertainty is in the travel costs, not in the demands.

The metaheuristic algorithm that we use in this study is an ant colony optimization (ACO; see [11, 12]) algorithm. ACO can be defined as a class of metaheuristic algorithms. An ACO algorithm simulates the behavior of the ants in the nature on a solution space. The inspiration of ACO is as follows. In the nature, the ants get out of their nests to reach a food source. In the beginning, various ants reach to the food source by using various paths and they mark the path they choose by leaving their pheromones. The ants that choose shorter paths can go back and forth more frequently to the food source, increasing their pheromones on their paths. The other ants that are influenced by the pheromones also get attracted to these shorter paths, leaving their own pheromones, thus increasing the total pheromones on shorter paths even more. Therefore, as the better solutions (shorter paths) get more pheromones, in the end, most of the ants gather around the best solution known so far. In this chapter, we use our robust multiple ant colony system approach, which was previously proposed in [23]. In this approach, multiple ant colonies work in parallel, each focused on a different conservativeness degree. Each colony working for the robust multiple ant colony system approach is a process of our robust ant colony system proposed in [22]. These colonies try to avoid the situation in which they get stuck working on dominated solutions, by communicating with each other and sharing their best solutions. In the end, the final best solutions of these ant colonies are collected in a solution pool. This solution pool contains solutions of different conservativeness degrees and allows the decision maker to analyze these alternative solutions and pick the best one according to the situation.

The goal of this experimental chapter is to confirm the practicality of RMACS by testing it on VRP instances with 200 customers, in which different policies of placing the customers are used (clustered, random, and mixed).

The structure of this chapter is as follows. In Sect. “Problem Definition,” a more formal problem definition is given. In Sect. “Methodology,” the approach that we use is explained. In Sect. “Experimental Results,” we present our experimental results. Finally, in Sect. “Conclusions,” the conclusions are drawn.

Problem Definition

Let us start defining CVRP and CVRPU in terms of a graph $G = (L, E)$, where L is the set of locations and E is the set of edges (i.e., paths between locations). The set of locations is expressed as $L = \{0, 1, 2, \ldots, |L|-1\}$ where 0 represents the location of the depot, 1 represents the location of the first customer, 2 represents the location of the second customer, and so on. The set of edges is expressed as $E = \{(i, j) \mid i, j \in L, i \neq j\}$, each edge representing the path between two locations. Each edge $(i, j) \in E$ has an associated traveling cost represented by c_{ij}. Also, note that the traveling costs are symmetric, which means $c_{ij} = c_{ji}$. Each customer $i \in (L \backslash \{0\})$ has a demand (i.e., expectation of delivery from the depot), and the amount of this demand is expressed by d_i. At the depot, there is no demand (i.e., $d_0 = 0$). The set of vehicles is expressed by V, and the number of vehicles is expressed by $|V|$. The capacity of each vehicle is Q.

Let us express a solution for CVRP and CVRPU by *sol*. The related definitions are as follows:

- *sol*[*v*] is the route (i.e., list of visits) of the vehicle $v \in V$, according to *sol*.
- |*sol*[*v*]| is the number of visits by the vehicle $v \in V$, according to *sol*.
- *sol*[*v*,*k*] is the *k*th visited location by the vehicle $v \in V$, according to *sol*.

By using these notations, let us now define the constraints of the problem. The first and the last visited locations of a vehicle *v* must be the depot (i.e., a vehicle *v* must start and end its journey at the depot):

$$sol\,[v, 1] = sol\,[v, |sol\,[v]|] = 0 \qquad \forall v \in V \tag{1}$$

The non-depot locations included in the route of a vehicle must consist of valid customers:

$$sol\,[v, k] \in (L \backslash \{0\}) \qquad \forall v \in V; k \in \{2, 3, \ldots, |sol\,[v]| - 1\} \tag{2}$$

A vehicle $v \in V$ must not visit the same customer twice:

$$\begin{aligned} & sol\,[v, k] \neq sol\,[v, k'] \\ & \quad \forall v \in V; k, k' \in \{2, 3, \ldots, |sol\,[v]| - 1\}; k \neq k' \end{aligned} \tag{3}$$

Two different vehicles $v, v' \in V$ must not have the same customers on their routes:

$$\begin{aligned} & sol\,[v, k] \neq sol\,[v', k'] \\ & \quad \forall v, v' \in V; v \neq v'; \\ & \quad k \in \{2, 3, \ldots, |sol\,[v]| - 1\}; k' \in \{2, 3, \ldots, |sol\,[v']| - 1\} \end{aligned} \tag{4}$$

The total demand of the customers of a vehicle route must not exceed the capacity of a vehicle:

$$\sum_{k \in \{2,3,\ldots,|sol[v]|-1\}} d_{sol[v,k]} \leq Q \qquad \forall v \in V \tag{5}$$

The cost of a solution sol is defined as follows:

$$\text{COST}(sol) = \sum_{v \in V} \sum_{k=2}^{|sol[v]|} c_{sol[v,k-1],sol[v,k]}$$

We are finally ready to make a complete definition of the deterministic CVRP:

$$\text{CVRP} \begin{cases} \text{minimize} & \text{COST}(sol) \\ \text{subject to} & (1), (2), (3), (4), \text{ and } (5) \end{cases}$$

In CVRPU, however, the costs c_{ij} are not single numbers, but they are intervals because of the uncertainty. Therefore, we need to define a different cost function, which evaluates the cost of a solution. Also, because our goal is to produce solutions with different conservativeness levels, the cost function should be configurable in terms of conservativeness. Bertsimas and Sim ([6, 7]) propose such an approach, where the conservativeness degree is configurable by using a parameter $\Gamma \geq 0$. Although in the studies [6, 7] it is shown that Γ can be set as a non-integer, let us focus on the implication of a Γ value when it is considered as an integer. In the case of our CVRP problem, the meaning of Γ is that, during the calculation of the solution cost, the travel costs of Γ number of edges are assumed to be equal to their highest values picked from their intervals, and the travel costs of the rest of the edges are assumed to be equal to their lowest values picked from their intervals. When using this approach, a nonconservative decision maker would set $\Gamma = 0$, so that all the edges would have their lowest costs, and a fully conservative decision maker would set Γ as the number of edges in the solution, so that all the edges would have their highest costs.

Now, let us formulate the evaluation approach of Bertsimas and Sim:

$$\begin{aligned} &\text{PERTURBEDCOST}\,(sol, \Gamma) = \\ &\quad \max \left\{ \sum_{v \in V} \sum_{k=2}^{sol[v]} \left[\underline{c}_{sol[v,k-1],sol[v,k]} \right.\right. \\ &\qquad\qquad + \gamma_{sol[v,k-1],sol[v,k]} \cdot \Big(\overline{c}_{sol[v,k-1],sol[v,k]} \\ &\qquad\qquad\qquad\qquad \left.\left. - \underline{c}_{sol[v,k-1],sol[v,k]} \Big) \right] \right\} \\ &\qquad\quad \text{s.t.} \sum_{(i,j) \in A} \gamma_{ij} \leq \Gamma \\ &\qquad\qquad\quad 0 \leq \gamma_{ij} \leq 1 \qquad \forall\, (i,j) \in E \end{aligned}$$

where γ_{ij} represents the assumption on the perturbation amount on the cost coefficient c_{ij}. According to this perturbation amount, the cost c_{ij} of the edge (i, j) is calculated as $\underline{c}_{ij} + \left(\gamma_{ij} \cdot \left(\overline{c}_{ij} - \underline{c}_{ij}\right)\right)$. By using the function PerturbedCost, we can now define CVRPU as

$$\text{CVRPU}\begin{cases} \text{minimize} & \text{PERTURBEDCOST}\,(sol, \Gamma) \\ \text{subject to} & (1), (2), (3), (4), \text{ and } (5) \end{cases}$$

Methodology

The methodology we use here is a heuristic approach called robust multiple ant colony system (RMACS) [23], in which multiple ant colonies work concurrently, each being focused on a different conservativeness degree.

In this section, we first explain the basic ant colony system. Then, we explain RMACS that executes multiple ant colony systems.

The Ant Colony System

On a VRP, the working of an ant colony optimization is as follows:

(a) First, artificial ants start "walking" on the graph $G = (L, E)$ to construct solutions.
(b) After the walking is finished, the ants mark their choices by leaving pheromones on the edges they used in their solutions, the amount of pheromone depending on the quality of the solution (measured by the function Cost and by the function PerturbedCost, in the case of CVRP and CVRPU, respectively).
(c) The next generation of artificial ants start walking to construct new solutions. While constructing their solutions, probabilistically, they become tempted to use the edges marked by the previous-generation ants.
(d) Like explained in (b), these new ants mark their choices.
(e) If the ending criterion is not met, we return to step (c) to repeat the process. After enough number of iterations, the artificial ants converge into a near-optimal solution, which has the highest concentration of pheromone.

The ant colony system is an ant colony optimization algorithm, which is elitist: only an ant that has found the best solution known so far leaves pheromones. The goal of this elitist behavior is to attract other ants into the best solution, so that they apply local searches around it.

We now give implementation details of the ant colony system, as previously discussed in [13]:

(a) An initial solution, *sol_init*, is generated by using the nearest neighbor heuristic (see [15]).

(b) The variable *sol_best*, which is to store the best-known solution, is initialized as *sol_init*.
(c) A new generation of ants is activated. In each generation, a certain number of ants exist. This number depends on a parameter, which is set as 10 in our study.
(d) Each ant of the active generation walks to construct a solution. Each constructed solution is evaluated by the function Cost in the case of CVRP and the function PerturbedCost in the case of CVRPU. If an ant was able to come up with a solution better than *sol_best*, that better solution is declared as the new *sol_best* and that ant leaves pheromones to attract the ants of the future generations toward its solution.
(e) If the finishing criterion is not met, we return back to step (c).

The way of generating a solution is as follows:

(a) We start considering the first vehicle. The depot (i.e., the location 0) is added into the solution as the first place to visit.
(b) Let us call a location which is not visited yet, and which does not violate the capacity of the current vehicle with its demand, a feasible location. As long as there are feasible locations left, they are picked and added onto the solution. If there are no more feasible locations, the depot is added onto the solution, which means that the vehicle's tour is completed and it is to return to the depot. After returning to the depot, if there are still locations that are not added onto the solution, a new vehicle is considered and step (b) is repeated.

During the process of constructing a solution, each ant is influenced by the pheromones left on the edges. The amount of pheromone on an edge $(i, j) \in E$ is denoted by τ_{ij}. Larger values for τ_{ij} cause the ants to be more attracted toward the edge (i, j). Let us now make the following definitions:

W: the set of ants, where $|W| = \Omega$.

N_w: the set of feasible locations for ant $w \in W$.

η_{ij}: heuristic closeness value between locations i and j, calculated by the inverse of the Euclidean distance between i and j, where $(i, j) \in E$.

β: the importance parameter for the factor of closeness, which affects the decisions of an ant.

α: a parameter within [0; 1], deciding the probabilities for an ant to do exploration (i.e., looking for a completely different decision, without getting affected by the pheromones) and exploitation (i.e., decision to completely follow the influence of the pheromones and to use the pieces of a solution with a high concentration of pheromone). Given these definitions, we can explain the behavior of an ant, which added the location i onto the solution recently and wants to visit another location j, as follows:

With probability α, the ant decides to do exploitation and visits a location j which maximizes $\tau_{ij}\ (\eta_{ij})^{\beta}$. On the other hand, with probability $1-\alpha$, the ant decides to do exploration, picking a location j probabilistically. In this case, the probability for ant w to pick a location j is as follows:

$$P_{ij}^{w} = \begin{cases} \dfrac{\tau_{ij} \cdot (\eta_{ij})^{\beta}}{\sum_{j' \in N_w} \tau_{ij'} \cdot (\eta_{ij'})^{\beta}} & \text{if } j \in N_w \\ 0 & \text{otherwise} \end{cases}$$

At the beginning, all the edges have an initial amount of pheromone, calculated as

$$\tau_0 = 1/\left(|L| \cdot \text{PERTURBEDCOST}\left(sol\_init, \Gamma\right)\right)$$

During the optimization process, two types of pheromone updates are done: local and global. The local update is the decreasing of pheromones on the edge (i, j) when an ant traverses from i to j. This decreasing is done to prevent the ants of the same generation from repeating the same solution. This decreasing is done as follows:

$$\tau_{ij} = (1 - \rho) \cdot \tau_{ij} + \rho \cdot \tau_0$$

where $\rho \in [0;\ 1]$ is a parameter configuring the amount of decrease in the pheromones. The global update is done when all the ants finish their walk. The purpose of the global update is to influence the ants of the next generation. This global update is formulated as

$$\tau_{ij} = (1 - \rho) \cdot \tau_{ij} + \rho/\text{PERTURBEDCOST}\left(sol\_best, \Gamma\right)$$

At the end of the walk of each ant, the classical local search called 3-opt (see [9, 15, 19] for the details about this heuristic) is applied on its generated solution, to increase its quality.

The RMACS Approach

Let us define $S^{\Gamma} = \{\Gamma_1,\ \Gamma_2, \ldots\}$ as the set of conservativeness levels which are interesting for the decision maker. The RMACS is an approach which executes $|S^{\Gamma}|$ number of ant colony systems concurrently. Each ant colony system focuses on a different conservativeness value within the set S^{Γ}. The purpose of these multiple colonies is to generate a solution pool, containing solutions of different conservativeness levels, so that the decision maker will analyze all these alternative solutions, see the effect of the uncertainty on the problem instance at hand, and pick the most practical solution.

In RMACS, an important feature is the sharing of solutions, with the help of a shared memory. Up to Δ number of generations, each ant colony of RMACS works by itself without any communication. Beginning with Δ-th generation, an ant colony does the following behaviors:

- Whenever an ant improves the best-known solution of its colony, that solution is uploaded into the shared memory.

- At each δ number of generations, an ant colony scans the shared memory to see if there are better solutions, than the ant colony's own best solution, when evaluated according to that colony's conservativeness degree. If such a better solution exists in the shared memory, the colony imports that better solution by forcing one of its ants to repeat its moves.

The usefulness of this solution sharing approach is that, by scanning the shared memory, the ant colonies stop working on dominated solutions, import better solutions from each other, and improve those solutions according to their own conservativeness degrees. In the end, a reliable solution pool without dominated solutions is generated.

Experimental Results

In this section, we present our results obtained from Homberger instances, downloadable from [25]. The Homberger instances that we used are:

- c1_2_1, in which the customers are clustered
- r1_2_1, in which the customers are randomly placed
- rc1_2_1, in which some customers are clustered and some customers are randomly placed

These considered instances originally have time window constraints as well. Since, in this study, we do not consider time window constraints, those constraints are ignored. Also, these instances were originally designed for the deterministic CVRP. We modified these instances in this study, so that the problem data are intervals to represent the uncertainty, not exactly known numbers. The modification of the problem data is done as follows. Let us define c'_{ij} as the cost of the edge (i, j) in the deterministic instance. The cost interval of the edge (i, j) in the CVRPU counterpart of this instance becomes $\left[\underset{-}{c}_{ij}; \overline{c}_{ij}\right] = \left[c'_{ij}; \mathrm{RND}\left(c'_{ij}, c'_{ij} \cdot UF\right)\right]$, where RND(a, b) means a random number between a and b, and UF is the uncertainty factor parameter, set as 1.5 in our experiments.

The experiments were done on a computer with Intel Core 2 Duo P9600 2.66 GHz processor with 4 GB of RAM. The RMACS was implemented in C programming language, with the parameter settings as $\alpha = 0.99$, $\beta = 1$, and $\rho = 0.1$, with execution time limit of 1,200 s. The parameters related to the solution sharing were set as $\Delta = 9{,}000$ and $\delta = 500$. The considered conservativeness degrees are $S^{\Gamma} = \{0, 10, 25, 50, 75, 100, 150, M\}$, where $M \geq (|L| + |V| - 1)$ is a number big enough to make all the cost assumptions equal to their highest values.

The results obtained are presented in Tables 1, 2, and 3. Each table is a solution pool found for an instance, in which each row represents a solution found with a different conservativeness degree. In these tables, Γ means the conservativeness degree of the ant colony during the optimization process, and ϒ means the scenario

Table 1 Results obtained for the instance *c1_2_1*

	$\Upsilon = 0$	$\Upsilon = 10$	$\Upsilon = 25$	$\Upsilon = 50$	$\Upsilon = 75$	$\Upsilon = 100$	$\Upsilon = 150$	$\Upsilon = \mathrm{M}$
$\Gamma = 0$	2,605.96	2,866.45	3,034.68	3,153.61	3,197.64	3,226.57	3,259.61	3,275.24
$\Gamma = 10$	2,681.62	2,820.88	2,928.86	3,019.85	3,069.50	3,101.97	3,140.08	3,155.45
$\Gamma = 25$	2,699.18	2,824.15	2,914.27	3,003.31	3,052.59	3,084.66	3,123.87	3,139.54
$\Gamma = 50$	2,732.36	2,848.15	2,930.63	2,999.28	3,039.24	3,066.04	3,098.75	3,114.38
$\Gamma = 75$	2,734.13	2,849.92	2,932.41	2,999.83	3,038.34	3,064.58	3,096.96	3,112.59
$\Gamma = 100$	2,734.13	2,849.92	2,932.41	2,999.83	3,038.34	3,064.58	3,096.96	3,112.59
$\Gamma = 150$	2,758.34	2,871.99	2,943.99	3,005.08	3,041.88	3,067.79	3,099.15	3,114.47
$\Gamma = \mathrm{M}$	2,734.13	2,849.92	2,932.41	2,999.83	3,038.34	3,064.58	3,096.96	3,112.59

Table 2 Results obtained for the instance *r1_2_1*

	$\Upsilon = 0$	$\Upsilon = 10$	$\Upsilon = 25$	$\Upsilon = 50$	$\Upsilon = 75$	$\Upsilon = 100$	$\Upsilon = 150$	$\Upsilon = \mathrm{M}$
$\Gamma = 0$	3,051.78	3,227.33	3,374.83	3,526.16	3,630.15	3,699.60	3,787.16	3,826.32
$\Gamma = 10$	3,100.14	3,198.38	3,308.91	3,439.13	3,527.06	3,592.27	3,676.18	3,711.87
$\Gamma = 25$	3,099.16	3,199.03	3,308.50	3,438.54	3,525.68	3,589.44	3,670.35	3,705.07
$\Gamma = 50$	3,099.16	3,199.03	3,308.50	3,438.54	3,525.68	3,589.44	3,670.35	3,705.07
$\Gamma = 75$	3,103.04	3,201.29	3,310.76	3,438.67	3,525.03	3,588.31	3,670.2	3,705.46
$\Gamma = 100$	3,106.31	3,205.29	3,314.76	3,441.95	3,526.48	3,586.82	3,666.77	3,700.49
$\Gamma = 150$	3,099.14	3,207.50	3,316.19	3,441.05	3,525.30	3,585.26	3,665.06	3,698.77
$\Gamma = \mathrm{M}$	3,099.49	3,208.58	3,317.27	3,442.13	3,525.64	3,585.20	3,664.99	3,698.71

Table 3 Results obtained for the instance *rc1_2_1*

	$\Upsilon = 0$	$\Upsilon = 10$	$\Upsilon = 25$	$\Upsilon = 50$	$\Upsilon = 75$	$\Upsilon = 100$	$\Upsilon = 150$	$\Upsilon = \mathrm{M}$
$\Gamma = 0$	2,959.80	3,174.16	3,359.28	3,523.84	3,617.19	3,672.79	3,734.40	3,759.04
$\Gamma = 10$	2,995.42	3,144.16	3,283.99	3,422.38	3,504.58	3,554.70	3,610.70	3,632.00
$\Gamma = 25$	3,000.15	3,145.25	3,277.53	3,410.51	3,492.55	3,542.90	3,598.97	3,620.67
$\Gamma = 50$	3,010.08	3,155.25	3,272.44	3,391.08	3,466.75	3,514.89	3,568.81	3,589.72
$\Gamma = 75$	3,012.80	3,163.98	3,289.29	3,409.08	3,481.80	3,528.47	3,582.41	3,602.34
$\Gamma = 100$	3,008.45	3,159.64	3,289.65	3,411.82	3,482.49	3,527.72	3,580.69	3,600.35
$\Gamma = 150$	3,009.94	3,154.56	3,286.90	3,409.16	3,484.08	3,530.92	3,585.13	3,606.17
$\Gamma = \mathrm{M}$	3,007.04	3,156.12	3,274.97	3,389.75	3,462.04	3,506.51	3,558.52	3,578.99

assumption during the final evaluation of the result after the optimization (i.e., what would be the cost of a solution *sol* if Υ number of edges would be perturbed toward their highest values, calculated by PerturbedCost(*sol*, Υ)).

In each solution pool table, it can be seen that there are many different solutions, each providing the various costs at various scenario assumptions, providing the decision maker a set of alternatives. The fact that similar results are found on different types of instances shows the practicality of RMACS: different types of customer placing policies can be handled.

To analyze the effects of the uncertainty on this problem, let us make a detailed analysis on one of these instances, r1_2_1, with the help of Fig. 3. In the figure, it

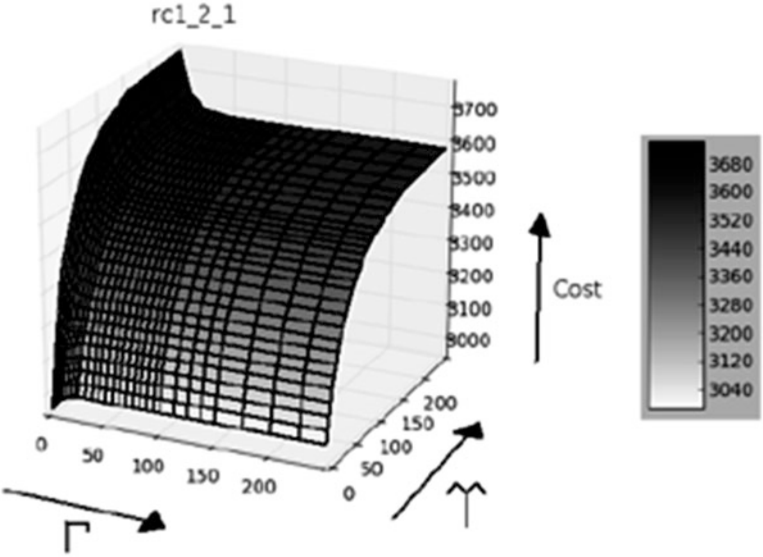


Fig. 3 The solution pool generated for the instance *rc1_2_1*

can be observed that the least conservative solution ($\Gamma = 0$) has the potential to be the cheapest solution at the best-case scenario, but also the most expensive solution at the worst-case scenario. Therefore, it is not a robust solution. A slight increase in the conservativeness degree to the values $\Gamma = 10$ and $\Gamma = 25$ seems to decrease the cost at the worst-case scenario significantly, providing much more robustness. In the region $\Gamma \geq 25$, the solutions are very similar to each other, behaving almost the same, meaning that configuring the degree of conservativeness provides a variety in possible behaviors (or provides the trade-off between the worst-case and the best-case cost) up to $\Gamma = 25$, in the case of this instance.

Conclusions

The robust multiple ant colony system, RMACS, was tested on instances with 200 customers, the instance types being the ones with the clustered customers, the ones with the randomly placed customers, and the mixed ones. It was seen that the RMACS approach was able to generate solution pools with alternative solutions of different conservativeness degrees, experimentally showing the practicality of RMACS in providing the decision maker an opportunity to analyze the effects of the uncertainty on the problem at hand.

Acknowledgment N. E. Toklu was supported by Hasler Stiftung through project 13002 "Metaheuristic approaches for robust optimization."

References

1. Agra, A., Christiansen, M., Figueiredo, R., Hvattum, L.M., Poss, M., Requejo, C.: The robust vehicle routing problem with time windows. Comput. Oper. Res. **40**(3), 856–866 (2013)
2. Baldacci, R., Christofides, N., Mingozzi, A.: An exact algorithm for the vehicle routing problem based on the set partitioning formulation with additional cuts. Math. Program. **115**(2), 351–385 (2008)
3. Baldacci, R., Toth, P., Vigo, D.: Exact algorithms for routing problems under vehicle capacity constraints. Ann. Oper. Res. **175**(1), 213–245 (2010)
4. Ben-Tal, A., Nemirovski, A.: Robust solutions of uncertain linear programs. Oper. Res. Lett. **25**(1), 1–13 (1999)
5. Ben-Tal, A., Nemirovski, A.: Robust solutions of linear programming problems contaminated with uncertain data. Math. Program. **88**(3), 411–424 (2000)
6. Bertsimas, D., Sim, M.: Robust discrete optimization and network flows. Math. Program. **98**(1), 49–71 (2003)
7. Bertsimas, D., Sim, M.: The price of robustness. Oper. Res. **52**(1), 35–53 (2004)
8. Birge, J.R., Louveaux, F.: Introduction to Stochastic Programming. Springer Verlag, New York (1997)
9. Croes, G.A.: A method for solving traveling-salesman problems. Oper. Res. **6**(6), 791–812 (1958)
10. Dantzig, G.B., Ramser, J.H.: The truck dispatching problem. Manage. Sci. **6**(1), 80–91 (1959)
11. Dorigo, M.: Learning and natural algorithms. PhD thesis, Politecnico di Milano (1992)
12. Dorigo, M., Maniezzo, V., Colorni, A.: Positive feedback as a search strategy. Technical report, Dipartimento di Elettronica, Politecnico di Milano (1991)
13. Gambardella, L.M., Taillard, E., Agazzi, G.: New ideas in optimization, chapter "MACS-VRPTW: a multiple ant colony system for vehicle routing problems with time windows," pp. 63–76. McGraw-Hill, New York (1999)
14. Golden, B.L., Raghavan, S., Wasil, E.A.: The vehicle routing problem: latest advances and new challenges, vol. 43. Springer, New York (2008)
15. Johnson, D.S., McGeoch, L.A.: The traveling salesman problem: a case study in local optimization. In: Aarts, E.H.L., Lenstra, J.K. (eds.) Local search in combinatorial optimization, pp. 215–310. Wiley, Chichester (1997)
16. Kouvelis, P., Yu, G.: Robust discrete optimization and its applications. Kluwer Academic, Dordrecht (1997)
17. Laporte, G.: The vehicle routing problem: an overview of exact and approximate algorithms. Eur. J. Oper. Res. **59**(3), 345–358 (1992)
18. Lee, C., Lee, K., Park, S.: Robust vehicle routing problem with deadlines and travel time/demand uncertainty. J. Oper. Res. Soc. **63**(9), 1294–1306 (2011)
19. Lin, S.: Computer solutions of the traveling salesman problem. Bell Sys. Tech. J. **44**(10), 2245–2269 (1965)
20. Moghaddam, B.F., Ruiz, R., Sadjadi, S.J.: Vehicle routing problem with uncertain demands: an advanced particle swarm algorithm. Comput. Ind. Eng. **62**, 306–317 (2012)
21. Sungur, I., Ordonez, F., Dessouky, M.: A robust optimization approach for the capacitated vehicle routing problem with demand uncertainty. IIE Trans. **40**(5), 509–523 (2008)
22. Toklu, N.E., Montemanni, R., Gambardella, L.M.: An ant colony system for the capacitated vehicle routing problem with uncertain travel costs. In: IEEE Symposium on Swarm Intelligence (SIS), pp. 32–39. (2013)
23. Toklu, N.E., Montemanni, R., Gambardella, L.M.: A robust multiple ant colony system for the capacitated vehicle routing problem. In: IEEE International Conference on Systems, Man, and Cybernetics (2013)
24. Toth, P., Vigo, D. (eds.): The Vehicle Routing Problem. SIAM, Philadelphia (2001)
25. Universidad de Malaga, Networking and Emerging Optimization Group. Capacitated VRP with Time Windows Instances. http://neo.lcc.uma.es/vrp/vrp-instances/capacitated-vrp-with-timewindows-instances/ (2013). Accessed 29 Nov 2013

The Granger Causality Effect between Cardiorespiratory Hemodynamic Signals

Samir Ghouali, Mohammed Feham, and Yassine Zakarya Ghouali

Abstract Granger causality (GC) is one of the most popular measures to reveal causality influence of time series based on the estimated linear regression model and has been widely applied in economics and neuroscience due to its reliability, clarity, and robustness.

Granger causality has recently received increasing attention to study causal interactions of neurophysiological data; in this chapter we have developed a model of causality between the respiratory, hemodynamic, and cardiac signals, more specifically, a study based on the Granger causality between three ECG leads, blood pressure, central venous pressure, pulmonary arterial pressure, respiratory impedance, and airway CO_2. We selected 187 patients of 250 for our study, taken from Montreal General Hospital/MF (Massachusetts General Hospital/Marquette Foundation) databases. These signals are ideal for understanding causality and coupling (unidirectional or bidirectional).

In this approach we aim to analyze and understand the interactions between the signals mentioned above, and identify the significance of this interaction. The originality of this chapter is the number of variables selected for the study. Unlike the majority of studies that are conducted only with two variables, our study is multidimensional. The main advantage of a multidimensional and multivariable model is to solve a myriad of problems which is not the case in the two-dimensional studies.

Keywords Direction of information • Cardiorespiratory hemodynamic signals • Granger causality • Multivariate study • (MGH/MF) waveform database

S. Ghouali (✉) • M. Feham
Faculty of Engineering Sciences, STIC Lab, University of Tlemcen, Tlemcen, Algeria
e-mail: ghtelec@yahoo.fr; m_feham@mail.univtlemcen.dz

Y.Z. Ghouali
Faculty of Economics, Business and Management Sciences, POLDEVA Lab,
University of Tlemcen, Tlemcen, Algeria
e-mail: ghouali.poldeva@yahoo.fr

K. Chen (ed.), *Forging Connections between Computational Mathematics and Computational Geometry*, Springer Proceedings in Mathematics & Statistics 124,
DOI 10.5176/2251-1911_CMCGS14.50_23

Introduction

Numerous studies in recent years have been devoted to the evaluation of causality; several applications of the latter are omnipresent in areas ranging from the economy [1, 2], climatology [3–5], directed information theory in networks [6], psychiatry [7], brain imaging field [8] and especially the analysis of biological systems, with a very special emphasis on the neural field [9–20], and the study of cardiac signals [21–29].

Although this one is not the universal definition of causality [30], it is commonly accepted that the notion of causality of two events describes why one event is caused by the other. According to this very general definition, in this chapter, we will look at the cardiovascular field. The importance of causality in this case appears in the spontaneous cardiovascular variability and complexity of cardiovascular regulation.

The detection and modeling of this causality depends strictly on all signals exploited to describe the observed interactions [31]. Characterization of the interdependence between the sensed signals is one of the most critical problems in cardiovascular pathophysiology [32]. From this idea, we can independently evaluate by what is called the strength of the relationship between two signals to a minimum.

The causality analysis has the ability to provide an original framework to identify the responsible mechanism for the spontaneous variations without the intervention of an artificial stimulus such as pharmacological intervention, an experiment on the patient or more severely surgery to obtain a specific causal relationship.

Causality is usually tested in time [33], frequency [34], and information domains [30]. In this chapter, methods assessing causality in time domain were chosen because they do not need to assume that the cardiovascular control mechanisms occur along specific temporal scales and the distribution of the statistic assessing causality under the null hypothesis of absence of a causal relationship between the two series is well known, thus allowing to easily keep under control the percentage of false causality detections.

In our case, we will look at the Granger causality because it perfectly studies multivariable models (several variables at once). The mathematical formulation of causality in measurable terms of predictability was given by Wiener [35]. Granger [1] introduced a specific notion of causality into time series analysis by evaluation of predictability autoregressive models.

The cardiovascular system is regulated by numerous control mechanisms acting to guarantee that the necessities of each physiological area are satisfied and that cardiovascular variables do not assume values incompatible with life. In the latter application, a large body of work has been developed to assess causality in both cardiovascular [24, 25] and cardiorespiratory [27] interactions.

Unfortunately, the world of physiological signals describing the behavior of a given system is not affected; it means that there's a restriction in the use of necessary physiological signals, which implies a certain limitation in reliability of causality. To achieve this, we can give a small example of a study of causality between heart period (HP) and systolic arterial pressure (SAP), it would be preferable to include breathing (R) in the test [36]. According to several studies, neglecting (R), this can lead us to erroneous results in the causal interpretation mentioned above, this observation suggests that (R) should be included in this causality, the validation of this hypothesis has been approved by work [40] and they were able to uncover the importance of (R) in the causation (HP) and (SAP).

This demonstration gave us the reflex to study a large number of physiological signals from the outset, to eliminate any gaps, ignorance, and negligence of other signals, multivariate system identification approaches permit the dynamic characterization of the causal interactions among cardiovascular regulatory mechanisms responsible for coupling the variability between signals (e.g., heart rate, arterial pressure, and respiratory signal) [38]. Multivariate characterization is not solely helpful both to derive information about the gain and phase of the relationship linking any signal pair but also to estimate causality (i.e., who drives whom) in multivariate recordings.

Due to cardiopulmonary anatomy, there are strong mechanical interactions between the mechanical activity of the heart and respiratory movements which implies a change in atrial and pulmonary receptors [39]. The approach was applied to three ECG leads, ART (arterial pressure), PAP (pulmonary arterial pressure), CVP (central venous pressure), respiratory impedance, and airways CO_2, taken from the MGH/MF (Massachusetts General Hospital/Foundation Marquette) database. These signals are ideal for understanding causality and coupling (unidirectional or bidirectional).

The goal of the chapter is to propose to study the direction of causality between the signals mentioned previously; our contribution in this chapter is based on the following points:

- 3D Analysis of cardiovascular signals.
- Study bivariate/multivariate between the cardiovascular, respiratory, and hemodynamic signals.
- Have the necessary information and details for our next work to develop telemedicine applications on smartphones and especially on the part intended to signal analysis.

The remainder of this chapter is organized as follows: In “Definitions of Some Variables Studied,” we give some definition of the variables. In “Data and Methodology,” we will establish the data used and the methodology to follow. Then, in “The Method and Findings,” we present the model used and the result obtained. And finally in “Discussions,” we lead an analysis, scientific discussion, and a projection of perspectives.

Definitions of Some Variables Studied

ECG leads: Lead systems allow you to look at the heart from different angles. Each different angle is called a lead. The different leads can be compared to radiographs taken from different angles.

ART: Blood pressure is the pressure of blood in the arteries, also referred to as blood pressure because this pressure is the force exerted by the blood against the walls of arteries, tends the wall of the artery.

CVP: Central venous pressure (CVP) also known as right atrial pressure (RAP) describes the pressure of blood in the thoracic vena cava near the right atrium of the heart; it reflects the amount of blood returning to heart and the heart's ability to pump blood into the arterial system.

PAP: Pulmonary arterial pressure measures the pressure in the pulmonary arteries, the latter carries blood from the right side of the heart to the lungs.

Data and Methodology

Data Analysis

The Massachusetts General Hospital/Marquette Foundation (MGH/MF) Waveform database is a comprehensive collection of electronic records of hemodynamic and electrocardiographic signals of stable and unstable patients in intensive care units, operating rooms, and cath labs heart. It is the result of collaboration between physicians, biomedical engineers, and nurses of the Massachusetts General Hospital [40], which includes three ECG leads, arterial pressure, pulmonary arterial pressure, central venous pressure, respiratory impedance, and airway CO_2. This multidimensional cardiac data collected from various parts of body can effectively imitate the signals from various body sensor nodes.

The original dataset contains total 250 sets of cardiac signals, each containing 12–86 min in most cases are about an hour of recording. We selected 187 patients were selected on 250 for our simulation and contain all the signals mentioned above unlike the rest that does not have the typical data to our studies, these signals include cardiac events such as extrasystole, premature supraventricular tachycardia, bradycardia, extrasystole and ventricular stimulation, which are manually annotated by clinical professionals.

Methodology

In the analysis of the causality relationship, the choice of the appropriate technique is an important theoretical and empirical question. Granger causality is the most appropriate technique to study the relationship between hemodynamic, cardiorespiratory, and electrocardiographic signals. The empirical strategy used in this chapter can be divided into three main stages. First, unit root tests in series are undertaken to determine the stationarity of the series. Second, the AIC (Ackaik) and SIC (Schwarz) criteria are used to determine the optimal lag used in the method of causation. Third and finally, we will test for multivariate causality proposed by Granger.

The Method and Findings

The Model Specification

The first causality was proposed and introduced by Wiener and Granger (Nobel 2003) and became a fundamental theory for the analysis of dynamic relationships between time series. Sims presented a slightly different specification of test by considering that future values help explain the present values.

In the remainder of this chapter, we will look at the multivariate Grange causality. Before beginning the multivariate Granger causality, it is necessary to move to the bivariate causality to see the difference and the limitation of the latter.

Bivariate Granger Causality Test

In this part, we will try to present the definitions of linear causality and discuss the following tests to identify the causal relationship between these two variables.

Granger causality test is designed to detect causal direction between two time series by examining a correlation between the current value of one variable and the past values of another variable. Based on Granger's definition of causality Y is strictly Granger causing X if the conditional distribution of X_t, given the past observation $X_{t-1}, X_{t-2}, \ldots$ and $Y_{t-1}, Y_{t-2}, \ldots$ differs from the conditional distribution of X_t, given the past observation $X_{t-1}, X_{t-2}, \ldots$ only.

Intuitively, Y is a Granger cause of X if adding past observations of Y to the information set increases the knowledge on the distribution of current values of X. More precisely, the linear Granger causality is conducted based on the following two-equation model:

$$x_t = a_1 + \sum_{i=1}^{k} \alpha_i x_{t-i} + \sum_{i=1}^{k} \beta_i y_{t-i} + e_{1t} \tag{1}$$

and

$$y_t = a_2 + \sum_{i=1}^{k} \mu_i x_{t-i} + \sum_{i=1}^{k} \Omega_i y_{t-i} + e_{2t} \tag{2}$$

Where all $\{x_t\}$ and $\{ y_t\}$ are stationary variables, e_{1t} and e_{2t} are the disturbances satisfying the regularity assumptions of the classical linear regression model and k is the optimal lag.

We say that the variable { $\mathrm{y_t}$} is not to Granger causality {$\mathrm{x_t}$} if and only if $\beta_\mathrm{i} = 0$ in (1), for any i $= 1, \ldots,$ k. To better explain this, the past values of { $\mathrm{y_t}$} do not provide any additional information on the performance of {$\mathrm{x_t}$}. In the same manner, and vice versa, {$\mathrm{x_t}$} does not Granger causality { $\mathrm{y_t}$} if and only if $\mu_\mathrm{i} = 0$ in (2), for any i $= 1, \ldots,$ k.

We can test the causal relationships between two variables {$\mathrm{x_t}$} and { $\mathrm{y_t}$} by checking the null hypothesis separately:

$H_0^1 : \beta_1 = \cdots = \beta_k = 0$ and $H_0^2 : \mu_1 = \cdots = \mu_k = 0$

1. If both Hypotheses H_0^1 and H_0^2 are accepted, there is no linear causal relationship between $\{x_t\}$ and $\{y_t\}$.
2. If hypothesis H_0^1 is accepted but hypothesis H_0^2 is rejected, then there exists linear causality running unidirectionally from $\{x_t\}$ to $\{y_t\}$.
3. If hypothesis H_0^1 is rejected but hypothesis H_0^2 is accepted, then there exists linear causality running unidirectionally from $\{y_t\}$ to $\{x_t\}$.
4. If both Hypotheses H_0^1 and H_0^2 are rejected, then there is feedback linear causal relationship between $\{x_t\}$ and $\{y_t\}$.

The purpose of the multidimensional study is that it allows us to quantitatively analyze and respond to a myriad of problems and suggestions which is not the case in bivariate studies (multivariate Granger causality: an estimation framework based on factorization of the spectral density matrix).

Multivariate Granger Causality

Consider a p dimensional multivariate stochastic process, $\mathrm{X}(\mathrm{t}) = [\mathrm{x_t}, \mathrm{y_t}, \ldots, \mathrm{z_t}]^\mathrm{T}$, where the rank of the original matrix X(t) is [p, 1], we can estimate the model as MVAR (multivariate autoregressive) result:

$$\left\{\begin{matrix} \mathrm{x}_1(\mathrm{t}) \\ \mathrm{x}_2(\mathrm{t}) \\ \vdots \\ \vdots \\ \mathrm{x}_\mathrm{p}(\mathrm{t}) \end{matrix}\right\} = \sum_{\mathrm{K}=1}^{\infty} \begin{bmatrix} \mathrm{A}_{11}(\mathrm{K}) & \mathrm{A}_{12}(\mathrm{K}) & \mathrm{A}_{13}(\mathrm{K}) \cdots \mathrm{A}_{1\mathrm{p}}(\mathrm{K}) \\ \mathrm{A}_{21}(\mathrm{K}) & \mathrm{A}_{22}(\mathrm{K}) & \mathrm{A}_{23}(\mathrm{K}) \cdots \mathrm{A}_{24}(\mathrm{K}) \\ \vdots & \vdots & \cdots\cdots\vdots \\ \vdots & \vdots & \cdots\cdots\vdots \\ \mathrm{A}_{\mathrm{p}1}(\mathrm{K}) & \mathrm{A}_{\mathrm{p}2}(\mathrm{K}) & \mathrm{A}_{\mathrm{p}3}(\mathrm{K}) \cdots \mathrm{A}_{\mathrm{PP}}(\mathrm{K}) \end{bmatrix} \left\{\begin{matrix} \mathrm{x}_1(\mathrm{t}-\mathrm{K}) \\ \mathrm{x}_2(\mathrm{t}-\mathrm{K}) \\ \vdots \\ \vdots \\ \mathrm{x}_\mathrm{p}(\mathrm{t}-\mathrm{K}) \end{matrix}\right\} + \left\{\begin{matrix} \varepsilon_1(\mathrm{t}) \\ \varepsilon_2(\mathrm{t}) \\ \vdots \\ \vdots \\ \varepsilon_\mathrm{P}(\mathrm{t}) \end{matrix}\right\} \tag{3}$$

where $\mathrm{A_{ij}}(\mathrm{K})$ is the coefficient at K Th lag and $_\mathrm{i}(\mathrm{t})$ is a corresponding error terms.

F-Test

Several statistics could be used to test the above hypotheses; one of the most commonly used statistics is the standard F-test. It is a statistical hypothesis test for testing the equality of two variances by taking the ratio of the two variances, it can be represented by:

$$F - test = \frac{\sigma_{\mathrm{x}}^{2}}{\sigma_{\mathrm{y}}^{2}} \tag{4}$$

Our Equations

First equation:

$$\begin{Bmatrix} ECG\,1\,(t) \\ ART\,(t) \\ CVP\,(t) \\ PAP\,(t) \\ RESP\,(t) \\ CO2\,(t) \end{Bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix}$$

$$+ \begin{bmatrix} A_{11,1} & A_{12,1} & A_{13,1} & A_{14,1} & A_{15,1} & A_{16,1} \\ A_{21,1} & A_{22,1} & A_{23,1} & A_{24,1} & A_{25,1} & A_{26,1} \\ A_{31,1} & A_{32,1} & A_{33,1} & A_{34,1} & A_{35,1} & A_{36,1} \\ A_{41,1} & A_{42,1} & A_{43,1} & A_{44,1} & A_{45,1} & A_{46,1} \\ A_{51,1} & A_{52,1} & A_{53,1} & A_{54,1} & A_{55,1} & A_{56,1} \\ A_{61,1} & A_{62,1} & A_{63,1} & A_{64,1} & A_{65,1} & A_{66,1} \end{bmatrix} \begin{bmatrix} ECG\,1\ (t-1) \\ ART\ (t-1) \\ CVP\ (t-1) \\ PAP\ (t-1) \\ RESP\ (t-1) \\ CO2\ (t-1) \end{bmatrix} + \cdots$$

$$+ \begin{bmatrix} A_{11,k} & A_{12,k} & A_{13,k} & A_{14,k} & A_{15,k} & A_{16,k} \\ A_{21,k} & A_{22,k} & A_{23,k} & A_{24,k} & A_{25,k} & A_{26,k} \\ A_{31,k} & A_{32,k} & A_{33,k} & A_{34,k} & A_{35,k} & A_{36,k} \\ A_{41,k} & A_{42,k} & A_{43,k} & A_{44,k} & A_{45,k} & A_{46,k} \\ A_{51,k} & A_{52,k} & A_{53,k} & A_{54,k} & A_{55,k} & A_{56,k} \\ A_{61,k} & A_{62,k} & A_{63,k} & A_{64,k} & A_{65,k} & A_{66,k} \end{bmatrix} \begin{bmatrix} ECG\,1\ (t-k) \\ ART\ (t-k) \\ CVP\ (t-k) \\ PAP\ (t-k) \\ RESP\ (t-k) \\ CO2\ (t-k) \end{bmatrix}$$

$$+ \begin{bmatrix} A_{11,p} & A_{12,p} & A_{13,p} & A_{14,p} & A_{15,p} & A_{16,p} \\ A_{21,p} & A_{22,p} & A_{23,p} & A_{24,p} & A_{25,p} & A_{26,p} \\ A_{31,p} & A_{32,p} & A_{33,p} & A_{34,p} & A_{35,p} & A_{36,p} \\ A_{41,p} & A_{42,p} & A_{43,p} & A_{44,p} & A_{45,p} & A_{46,p} \\ A_{51,p} & A_{52,p} & A_{53,p} & A_{54,p} & A_{55,p} & A_{56,p} \\ A_{61,p} & A_{62,p} & A_{63,p} & A_{64,p} & A_{65,p} & A_{66,p} \end{bmatrix} \begin{bmatrix} ECG\,1\ (t-p) \\ ART\ (t-p) \\ CVP\ (t-p) \\ PAP\ (t-p) \\ RESP\ (t-p) \\ CO2\ (t-p) \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix} \tag{5}$$

Second equation:

$$
\begin{aligned}
\left\{\begin{matrix} ECG\,2\,(t) \\ ART\ (t) \\ CVP\ (t) \\ PAP\ (t) \\ RESP\ (t) \\ CO2\,(t) \end{matrix}\right\} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} \\
+ \begin{bmatrix} A_{11,1} & A_{12,1} & A_{13,1} & A_{14,1} & A_{15,1} & A_{16,1} \\ A_{21,1} & A_{22,1} & A_{23,1} & A_{24,1} & A_{25,1} & A_{26,1} \\ A_{31,1} & A_{32,1} & A_{33,1} & A_{34,1} & A_{35,1} & A_{36,1} \\ A_{41,1} & A_{42,1} & A_{43,1} & A_{44,1} & A_{45,1} & A_{46,1} \\ A_{51,1} & A_{52,1} & A_{53,1} & A_{54,1} & A_{55,1} & A_{56,1} \\ A_{61,1} & A_{62,1} & A_{63,1} & A_{64,1} & A_{65,1} & A_{66,1} \end{bmatrix} \begin{bmatrix} ECG\,2\ (t-1) \\ ART\ \ (t-1) \\ CVP\ \ (t-1) \\ PAP\ \ (t-1) \\ RESP\ \ (t-1) \\ CO2\ \ (t-1) \end{bmatrix} + \cdots \\
+ \begin{bmatrix} A_{11,k} & A_{12,k} & A_{13,k} & A_{14,k} & A_{15,k} & A_{16,k} \\ A_{21,k} & A_{22,k} & A_{23,k} & A_{24,k} & A_{25,k} & A_{26,k} \\ A_{31,k} & A_{32,k} & A_{33,k} & A_{34,k} & A_{35,k} & A_{36,k} \\ A_{41,k} & A_{42,k} & A_{43,k} & A_{44,k} & A_{45,k} & A_{46,k} \\ A_{51,k} & A_{52,k} & A_{53,k} & A_{54,k} & A_{55,k} & A_{56,k} \\ A_{61,k} & A_{62,k} & A_{63,k} & A_{64,k} & A_{65,k} & A_{66,k} \end{bmatrix} \begin{bmatrix} ECG\,2\ (t-k) \\ ART\ \ (t-k) \\ CVP\ \ (t-k) \\ PAP\ \ (t-k) \\ RESP\ \ (t-k) \\ CO2\ \ (t-k) \end{bmatrix} \\
+ \begin{bmatrix} A_{11,p} & A_{12,p} & A_{13,p} & A_{14,p} & A_{15,p} & A_{16,p} \\ A_{21,p} & A_{22,p} & A_{23,p} & A_{24,p} & A_{25,p} & A_{26,p} \\ A_{31,p} & A_{32,p} & A_{33,p} & A_{34,p} & A_{35,p} & A_{36,p} \\ A_{41,p} & A_{42,p} & A_{43,p} & A_{44,p} & A_{45,p} & A_{46,p} \\ A_{51,p} & A_{52,p} & A_{53,p} & A_{54,p} & A_{55,p} & A_{56,p} \\ A_{61,p} & A_{62,p} & A_{63,p} & A_{64,p} & A_{65,p} & A_{66,p} \end{bmatrix} \begin{bmatrix} ECG\,2\ (t-p) \\ ART\ \ (t-p) \\ CVP\ \ (t-p) \\ PAP\ \ (t-p) \\ RESP\ \ (t-p) \\ CO2\ \ (t-p) \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix}
\end{aligned}
\tag{6}
$$

Third equation:

$$\left\{\begin{matrix} ECG\,3\,(t) \\ ART\,(t) \\ CVP\,(t) \\ PAP\,(t) \\ RESP\,(t) \\ CO2\,(t) \end{matrix}\right\} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix}$$

$$+ \begin{bmatrix} A_{11,1} & A_{12,1} & A_{13,1} & A_{14,1} & A_{15,1} & A_{16,1} \\ A_{21,1} & A_{22,1} & A_{23,1} & A_{24,1} & A_{25,1} & A_{26,1} \\ A_{31,1} & A_{32,1} & A_{33,1} & A_{34,1} & A_{35,1} & A_{36,1} \\ A_{41,1} & A_{42,1} & A_{43,1} & A_{44,1} & A_{45,1} & A_{46,1} \\ A_{51,1} & A_{52,1} & A_{53,1} & A_{54,1} & A_{55,1} & A_{56,1} \\ A_{61,1} & A_{62,1} & A_{63,1} & A_{64,1} & A_{65,1} & A_{66,1} \end{bmatrix} \begin{bmatrix} ECG\,3\,(t-1) \\ ART\,(t-1) \\ CVP\,(t-1) \\ PAP\,(t-1) \\ RESP\,(t-1) \\ CO2\,(t-1) \end{bmatrix} + \cdots$$

$$+ \begin{bmatrix} A_{11,k} & A_{12,k} & A_{13,k} & A_{14,k} & A_{15,k} & A_{16,k} \\ A_{21,k} & A_{22,k} & A_{23,k} & A_{24,k} & A_{25,k} & A_{26,k} \\ A_{31,k} & A_{32,k} & A_{33,k} & A_{34,k} & A_{35,k} & A_{36,k} \\ A_{41,k} & A_{42,k} & A_{43,k} & A_{44,k} & A_{45,k} & A_{46,k} \\ A_{51,k} & A_{52,k} & A_{53,k} & A_{54,k} & A_{55,k} & A_{56,k} \\ A_{61,k} & A_{62,k} & A_{63,k} & A_{64,k} & A_{65,k} & A_{66,k} \end{bmatrix} \begin{bmatrix} ECG\,3\,(t-k) \\ ART\,(t-k) \\ CVP\,(t-k) \\ PAP\,(t-k) \\ RESP\,(t-k) \\ CO2\,(t-k) \end{bmatrix}$$

$$+ \begin{bmatrix} A_{11,p} & A_{12,p} & A_{13,p} & A_{14,p} & A_{15,p} & A_{16,p} \\ A_{21,p} & A_{22,p} & A_{23,p} & A_{24,p} & A_{25,p} & A_{26,p} \\ A_{31,p} & A_{32,p} & A_{33,p} & A_{34,p} & A_{35,p} & A_{36,p} \\ A_{41,p} & A_{42,p} & A_{43,p} & A_{44,p} & A_{45,p} & A_{46,p} \\ A_{51,p} & A_{52,p} & A_{53,p} & A_{54,p} & A_{55,p} & A_{56,p} \\ A_{61,p} & A_{62,p} & A_{63,p} & A_{64,p} & A_{65,p} & A_{66,p} \end{bmatrix} \begin{bmatrix} ECG\,3\,(t-p) \\ ART\,(t-p) \\ CVP\,(t-p) \\ PAP\,(t-p) \\ RESP\,(t-p) \\ CO2\,(t-p) \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix} \tag{7}$$

We discussed earlier, we will try to make a causality test in three dimensions, after extensive research of the model, and it was found that could write our mathematical equations

Empirical Result

Before beginning our results, we must clarify some details and signs:

- Corresponds to the causal direction between three ECG leads to (ART, CVP, PAP, RESP, and CO_2).
- Corresponds to the causal direction between (ART, CVP, PAP, RESP, and CO_2) to three ECG leads.
- *: Indicates statistical significance at 1 %.
- A value above the sign, is the value of F-statistic, which is considered a measure of the correlation between the variables studied.
- A value above the sign, which is in brackets, corresponds to the value of the probability of causation.
- (*x*. *E x*): (*x* exponential *x*).
- *MGH Number*: corresponds to a given patient.

The displayed results are partial; we only look at a sample of 30 patients randomly chosen, all this in aim of showing the methodology followed for causality. In our chapter, the following tables summarize the results obtained and the rate of causality in each direction represented by the F-statistic seen previously and the corresponding probability.

MGH002

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	13.4646* → (8.E-100) ←	5.1146* → (5.E-27) ←	11.8814* → (1.E-85) ←	1.7138* → (0.0019) ←	1.6459* → (0.0038) ←
ECG2	12.3759* → (5.E-90) ←	4.8100* → (1.E-24) ←	12.040* → (6.E-87) ←	1.3941 (0.0399)	1.3370 (0.0631)
ECG3	5.80308* → (1.E-32) ←	3.13526* → (5.E-12) ←	7.27528* → (5.E-45) ←	0.83496 (0.7789)	0.83210 0.7835

MGH005

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	3.3758* → (1.E-13) ←	5.1716* → (2.E-27) ←	8.2914* → (8.E-54) ←	6.3103* → (7.E-37) ←	5.0512* → (2.E-26) ←
ECG2	3.5804* → (3.E-15) ←	3.5811* → (3.E-15) ←	3.8269* → (5.E-17) ←	5.2412* → (5.E-28) ←	3.6362* → (1.E-15) ←
ECG3	4.8860* → (3.E-25) ←	4.2643* → (2.E-20) ←	8.1746* → (9.E-53) ←	3.1918* → (2.E-12) ←	3.1083* → (8.E-12) ←

MGH010

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	1.2906 → (0.0897)	1.4608 → (0.0225)	1.3528 (0.0557)	1.1898 (0.1772)	1.3872 (0.0422)
ECG2	1.2241 → (0.1421)	1.4560 → (0.0235)	2.0112* → (6.E-05) ←	1.6922* (0.0024) ←	1.8152* → (0.0006) ←
ECG3	1.2631 → (0.1088)	1.3919 → (0.0406)	1.5093 → (0.0145)	1.6529* (0.0035) ←	1.8584* → (0.0004) ←

MGH016

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	5.81434* → (9.E-33) ←	9.07583* → (1.E-60) ←	8.58176* → (2.E-56) ←	8.94331* → (2.E-59) ←	6.14600* → (2.E-35) ←
ECG2	5.51416* → (3.E-30) ←	4.90756* → (2.E-25) ←	6.58831* → (3.E-39) ←	4.93960* → (1.E-25) ←	4.42447* → (1.E-21) ←
ECG3	3.86895* → (3.E-17) ←	4.25168* → (3.E-20) ←	4.58885* → (7.E-23) ←	4.40791* → (2.E-21) ←	3.79522* → (9.E-17) ←

MGH019

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	13.1779* (5 E-99) ←	63.8833* → (0.0000) ←	37.2080* → (0.0000) ←	15.3122* → (2.E-116) ←	4.91764* → (2.E-25) ←
ECG2	30.9990* (2.E-259) ←	17.460* → (5.E-136) ←	32.044* → (7.E-269) ←	21.700* → (1.E-174) ←	19.270* → (2.E-152) ←
ECG3	36.4516* (0.0000) ←	37.4346* → (0.0000) ←	31.4419* → (2.E-263) ←	18.0569* → (2.E-141) ←	17.6442* → (1.E-137) ←

MGH025

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	18.9841* → (6.E-150) ←	21.9002* → (2.E-176) ←	20.9682* → (5.E-168) ←	14.2303* → (1.E-106) ←	14.2677* → (5.E-107) ←
ECG2	16.8669* → (1.E-130) ←	18.2168* → (6.E-143) ←	17.6601* → (7.E-138) ←	14.084* → (8.E-105) ←	12.5292* → (2.E-91) ←
ECG3	21.9698* → (4.E-177) ←	37.8166* → (0.0000) ←	37.6249* → (0.0000) ←	34.7434* → (2.E-293) ←	30.9402* → (7.E-259) ←

MGH030

Lags 45	ART	CVP	PAP	RESP	CO$_2$
ECG1	6.80357* (3.E-40) ←	6.69719* (3.E-39) ←	16.6775* → (3.E-126) ←	10.6880* → (2.E-73) ←	11.7569* → (1.E-82) ←
ECG2	9.29697* → (5.E-60) ←	9.53410* (5.E-62) ←	8.45248* → (6 E-53) ←	8.85198* → (3 E-56) ←	19.4575* → (9.E-148) ←
ECG3	4.39468* (6.E-21) ←	3.27079* (1.E-12) ←	17.1573* → (2.E-130) ←	12.3559* → (6.E-88) ←	12.5787* → (7.E-90) ←

MGH038

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	34.1336* → (8.E-288) ←	42.0905* → (0.0000) ←	45.3102* → (0.0000) ←	29.5296* → (5.E-246) ←	31.7334* → (5.E-266) ←
ECG2	18.9646* → (1.E-149) ←	21.7894* → (2.E-175) ←	24.6819* → (7.E-202) ←	12.1942* → (2.E-88) ←	13.3511* → (9.E-99) ←
ECG3	23.5344* → (2.E-191) ←	29.4159* → (5.E-245) ←	30.9951* → (2.E-259) ←	19.4842* → (2.E-154) ←	23.2000* → (2.E-188) ←

MGH040

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	8.36020* → (2.E-54) ←	37.9324* → (0.0000) ←	8.91593* → (3.E-59) ←	274.813* → (0.0000) ←	297.264* → (0.0000) ←
ECG2	0.89056 (0.6825)	1.19442 (0.1723)	1.16756 → (0.2032)	1.71417* (0.0019) ←	1.44555 (0.0258)
ECG3	1.23840 (0.1292)	0.61661 (0.9809)	1.88425* → (0.0003) ←	1.16677 (0.2042)	1.14884 (0.2270)

MGH051

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	2.24987* → (3.E-06) ←	2.95403* (9.E-11) ←	3.94570* → (7.E-18) ←	1.22965 (0.1371)	1.77085* (0.0010) ←
ECG2	1.69168* → (0.0024) ←	0.47920 (0.9989)	1.72074* → (0.0017) ←	1.00992 (0.4536)	1.30522 (0.0803)
ECG3	2.16370* → (8.E-06) ←	0.41638 (0.9998)	2.28666* → (2.E-06) ←	0.78672 (0.8496)	1.57048* (0.0076) ←

MGH053

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	10.3098* (2.E-71) ←	5.50973* (3.E-30) ←	6.75286* → (1.E-40) ←	10.1978* (1.E-70) ←	10.8550* (2.E-76) ←
ECG2	4.73162* (6.E-24) ←	4.03205* (1.E-18) ←	5.24456* → (4.E-28) ←	1.77028* (0.0010) ←	4.92457* (2.E-25) ←
ECG3	21.9699* (4.E-77) ←	7.94890* → (8.E-51) ←	15.4340* → (1.E-117) ←	24.9747* → (1.E-204) ←	24.6144* (3.E-201) ←

MGH057

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	4.67572* (2.E-23) ←	1.21919 (0.1469)	3.39351* (8.E-14) ←	7.15210* (5.E-44) ←	3.57831* → (4.E-15) ←
ECG2	4.67572* (2.E-23) ←	2.18725* (6.E-06) ←	3.6724* (7.E-16) ←	3.73050* → (3.E-16) ←	1.46234 (0.0222)
ECG3	4.99836* (4.E-26) ←	1.43633 (0.0279)	5.08211* (9.E-27) ←	6.47808* → (3.E-38) ←	4.33791* (7.E-21) ←

MGH063

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	18.4866* → (2.E-145) ←	11.2375* → (8.E-80) ←	6.15746* → (1.E-35) ←	2.87877* → (3.E-10) ←	1.32860 (0.0673)
ECG2	1.05992 (0.3630)	1.06988 (0.3460)	1.74440* (0.0013) ←	1.58709* → (0.0069) ←	2.75945* (2.E-09) ←
ECG3	0.97387 (0.5226)	1.30668 (0.0795)	1.20308 (0.1631)	1.00263 → (0.4674)	0.44818 (0.9996)

MGH083

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	12.7866* → (1.E-93) ←	22.5566* → (2.E-182) ←	20.9038* → (2.E-67) ←	17.7196* → (2.E-138) ←	14.8111* → (6.E-112) ←
ECG2	7.61496* → (6.E-48) ←	9.96653* → (2.E-168) ←	9.84395* → (2.E-67) ←	13.3686* → (6.E-99) ←	9.15010* → (3.E-61) ←
ECG3	7.68454* → (1.E-48) ←	11.0423* → (5.E-78) ←	9.93059* → (3.E-68) ←	12.8800* → (2.E-94) ←	8.41489* → (7.E-55) ←

MGH089

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	73.1543* → (0.000) ←	95.5537* → (0.000) ←	31.9302* → (8.E-268) ←	31.0302* → (1.E-259) ←	23.8196* → (5.E-194) ←
ECG2	11.3719* → (5.E-81) ←	10.9364* → (4.E-77) ←	2.96698* (7.E-11) ←	5.51151* → (3.E-30) ←	2.26551* (2.E-06) ←
ECG3	24.1756* → (3.E-197) ←	33.800* → (9.E-285) ←	25.4177* (1.E-208) ←	25.3225* → (1.E-207) ←	18.7283* → (1.E-147) ←

MGH092

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	2.08356* (2.E-05) ←	4.54662* → (2.E-22) ←	6.58337* → (4.E-39) ←	4.53687* → (2.E-22) ←	5.11720* (5.E-27) ←
ECG2	1.80886* (0.0007) ←	6.06665* (8.E-35) ←	7.47223* (1.E-46) ←	5.26159* → (3.E-28) ←	3.28409* (5.E-13) ←
ECG3	1.81199* (0.0006) ←	10.0576* → (3.E-69) ←	21.9470* → (6.E-177) ←	11.5843* → (7.E-83) ←	2.75473* → (2.E-09) ←

MGH100

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	4.91054* (2.E-25) ←	16.0426* → (4.E-123) ←	3.79936* (8.E-17) ←	20.7232* → (9.E-166) ←	3.24029* → (1.E-12) ←
ECG2	0.79440 (0.8393)	7.89769* → (2.E-50) ←	1.16490 (0.2065)	8.52316* → (8.E-56) ←	1.21279 (0.1532)
ECG3	8.75542* (7.E-58) ←	15.9131* → (6.E-122) ←	8.65074* (6.E-57) ←	17.7816* → (6.E-139) ←	9.32829* (7.E-63) ←

MGH112

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	4.15145* (2.E-19) ←	3.49901* (1.E-14) ←	11.9535* → (3.E-86) ←	12.3790* → (5.E-50) ←	3.28641* (4.E-13) ←
ECG2	5.43397* (1.E-29) ←	3.44163* (3.E-14) ←	10.0673* (2.E-69) ←	8.05470* → (9.E-52) ←	18.3041* (1.E-143) ←
ECG3	34.8159* (6.E-294) ←	23.0383* → (7.E-187) ←	29.4004* → (9.E-245) ←	23.7304* → (3.E-193) ←	55.1399* → (0.0000) ←

MGH120

Lags 41	ART	CVP	PAP	RESP	CO$_2$
ECG1	3.10504* (1.E-10) ←	18.4799* → (4.E-130) ←	2.90835* (2.E-09) ←	1.37478 (0.0558)	2.97306* (6.E-10) ←
ECG2	1.91610* (0.0004) ←	8.35619* (9.E-49) ←	1.03304 (0.4125)	1.13946 (0.2494)	1.95470* (0.0003) ←
ECG3	5.81538* (2.E-29) ←	20.848* → (2.E-149) ←	5.41456* → (2.E-26) ←	2.19320* (2.E-05) ←	4.85186* (2.E-22) ←

MGH125

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	37.2982* → (0.0000) ←	59.1995* → (0.0000) ←	56.2916* → (0.0000) ←	41.9906* → (0.0000) ←	46.1347* → (0.0000) ←
ECG2	23.9073* → (8.E-195) ←	32.6219* (4.E-274) ←	34.3082* → (2.E-289) ←	27.3919* → (1.E-226) ←	26.0281* → (4.E-214) ←
ECG3	30.9119* → (1.E-258) ←	43.8797* → (0.0000) ←	44.4954* → (0.0000) ←	31.8713* → (3.E-267) ←	36.9114* → (0.0000) ←

MGH133

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	36.6303* (0.0000) ←	7.08807* → (2.E-43) ←	20.1787* → (8.E-161) ←	7.89099* → (2.E-50) ←	33.9250* → (7.E-286) ←
ECG2	26.6880* → (4.E-220) ←	4.67989* → (1.E-23) ←	18.5284* → (9.E-146) ←	5.10445* → (6.E-27) ←	26.7680* → (7.E-221) ←
ECG3	11.8091* (7.E-85) ←	5.58807* → (1.E-30) ←	11.1692* → (3.E-79) ←	6.56718* → (5.E-39) ←	10.2525* → (5.E-71) ←

MGH138

Lags 45	ART	CVP	PAP	RESP	CO$_2$
ECG1	3.12395* (1.E-11) ←	23.4754* → (7.E-187) ←	11.4000* → (1.E-79) ←	22.2300* → (9.E176) ←	3.46775* (4.E-14) ←
ECG2	14.7510* (4.E-109) ←	7.75526* → (4.E-48) ←	24.6622* → (2.E-197) ←	7.27460* → (4.E-44) ←	17.9733* (9.E-138) ←
ECG3	2.65184* (1.E-08) ←	30.2701* → (2.E-247) ←	8.61800* → (2.E-55) ←	22.8755* → (2.E-181) ←	2.22336* (5.E-06) ←

MGH140

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	10.6366* → (7.E-73) ←	16.0300* → (2.E-120) ←	2.64371* → (1.E-08) ←	5.33928* → (3.E-28) ←	1.78006* (0.0010) ←
ECG2	2.74211* → (3.E-09) ←	4.28183* → (5.E-120) ←	0.59752 (0.9852)	1.26172 (0.1202)	17.6785* → (4.E-135) ←
ECG3	1.10066 (0.2976)	1.47381 → (0.0211)	0.55523 → (0.9932)	1.40337 (0.0384)	53.3474* → (0.0000) ←

MGH145

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	23.3972* → (4.E-190) ←	37.4371* → (0.0000) ←	6.08542* → (5.E-35) ←	48.1748* → (0.0000) ←	33.5794* → (9.E-283) ←
ECG2	8.96339* (1.E-59) ←	29.8548* → (5.E-249) ←	1.67581* → (0.0028) ←	17.5093* → (2.E-136) ←	10.5899* → (5.E-74) ←
ECG3	4.27008* (2.E-20) ←	32.6098* → (5.E-274) ←	1.29463 → (0.0869)	28.2415* → (3.E-234) ←	11.3783* → (5.E-81) ←

MGH149

Lags 42	ART	CVP	PAP	RESP	CO_2
ECG1	11.4165* → (1.E-74) ←	60.0942* → (0.000) ←	50.2848* → (0.000) ←	25.8305* → (1.E-194) ←	0.85858 (0.7282)
ECG2	4.17901* (4.E-18) ←	17.564* → (2.E-125) ←	12.394* → (1.E-82) ←	19.6183* → (1.E-142) ←	2.0663* (6.E-05) ←
ECG3	5.30225* (3.E-26) ←	17.5014* → (5.E-125) ←	14.5339* → (2.E-100) ←	18.8263* → (5.E-136) ←	2.81741* (4.E-09) ←

MGH157

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	6.39863* → (1.E-37) ←	44.7048* → (0.0000) ←	42.9332* → (0.0000) ←	70.5331* → (0.0000) ←	35.5973* → (5.E-301) ←
ECG2	20.2817* → (1.E-161) ←	41.8912* → (0.0000) ←	80.1427* → (0.0000) ←	97.6992* → (0.0000) ←	33.1709* → (4.E-279) ←
ECG3	12.7252* → (4.E-93) ←	23.8653* → (2.E-194) ←	38.6965* → (0.0000) ←	42.1619* → (0.0000) ←	19.8184* → (2.E-157) ←

MGH164

Lags 45	ART	CVP	PAP	RESP	CO_2
ECG1	10.6244* (9.E-73) ←	38.4463* (0.000) ←	39.8908* (0.000) ←	3.08118* → (2.E-11) ←	0.46746 (0.9199)
ECG2	14.8957* → (2.E-110) ←	30.153* → (2.E-246) ←	25.192* → (3 E-202) ←	24.0169* → (9.E-192) ←	4.5827* → (2.E-22) ←
ECG3	41.7183* → (0.0000) ←	37.8544* → (0.0000) ←	30.4166* → (8.E-249) ←	62.3338* → (0.0000) ←	32.2493* → (4.E-265) ←

MGH172

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	29.6336* → (5.E-247) ←	103.428* → (0.0000) ←	98.9313* → (0.0000) ←	14.0733* → (3.E-105) ←	15.5381* → (1.E-118) ←
ECG2	5.59478* (6.E-31) ←	17.9636* → (1.E-140) ←	16.0905* → (1.E-123) ←	9.22566* → (6.E-62) ←	5.18213* (1.E-27) ←
ECG3	7.74507* (4.E-49) ←	5.19981* → (1.E-27) ←	2.82488* → (7.E-10) ←	9.66148* → (8.E-66) ←	12.2956* → (3.E-89) ←

MGH177

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	1.63370* → (0.0043) ←	1.95421* (0.0001) ←	1.63423* (0.0043) ←	1.62246* (0.0048) ←	5.68801* → (1.E-31) ←
ECG2	14.7421* → (2.E-111) ←	4.41430* → (2.E-21) ←	11.2246* → (7.E-80) ←	9.16150* → (2.E-61) ←	7.20844* → (2.E-44) ←
ECG3	0.96802 → (0.5340)	1.44906 (0.0250)	4.72939* (6.E-24) ←	1.05768 (0.3668)	1.80505* (0.0007) ←

MGH182

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	7.39968* (4 E-46) ←	9.08936* → (9 E-61) ←	9.27866* → (2.E-62) ←	9.78691* → (6 E-67) ←	6.90695* → (7 E-42) ←
ECG2	7.62345* → (5 E-48) ←	7.9467* → (8 E-51) ←	10.627* → (2 E-74) ←	9.12096* → (5 E-61) ←	8.97988* → (8 E-60) ←
ECG3	5.74363* → (4.E-32) ←	7.42181* → (3.E-46) ←	8.01550* → (2.E-51) ←	11.576* → (8.E-83) ←	9.77919* → (8.E-67) ←

MGH186

Lags 46	ART	CVP	PAP	RESP	CO2
ECG1	27.0971* → (7.E-224) ←	40.3905* → (0.0000) ←	41.2741* → (0.0000) ←	34.0191* → (9.E-287) ←	26.6640* → (6.E-220) ←
ECG2	4.67641* → (2.E-23) ←	5.41086* → (2.E-29) ←	6.47160* → (3.E-38) ←	5.78547* → (2.E-32) ←	4.84396* → (7.E-25) ←
ECG3	21.0929* → (4.E-169) ←	28.8383* → (9.E-240) ←	29.2925* → (7.E-244) ←	26.9987* → (5.E-223) ←	22.4236* → (3.E-181) ←

MGH191

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	12.5219* → (3.E-91) ←	1.99626* (7.E-05) ←	31.7273* → (5.E-226) ←	24.0795* → (2.E-196) ←	4.71440* (8.E-24) ←
ECG2	4.41125* (2.E-21) ←	1.80146* (0.0007) ←	10.9193* → (6.E-77) ←	7.43139* (2.E-46) ←	1.38848 (0.0418)
ECG3	8.60126* → (2.E-56) ←	1.30267 (0.0819)	30.7626* → (3.E-257) ←	22.7601* → (2.E-184) ←	1.02445 (0.4265)

MGH195

Lags 45	ART	CVP	PAP	RESP	CO$_2$
ECG1	18.5186* (1.E-145) ←	15.4394* → (1.E-117) ←	13.7019* → (6.E-102) ←	21.6494* → (3.E-174) ←	16.2000* → (1.E-124) ←
ECG2	41.2513* → (0.0000) ←	10.1139* → (8.E-70) ←	30.5860* → (1.E-255) ←	9.51562* → (2.E-64) ←	35.1716* → (3.E-297) ←
ECG3	8.53683* → (6.E-56) ←	10.1623* → (3.E-70) ←	8.12275* → (2.E-52) ←	24.5725* → (7.E-202) ←	9.66749* → (7.E-66) ←

MGH198

Lags 46	ART	CVP	PAP	RESP	CO$_2$
ECG1	11.4239* → (2.E-81) ←	60.0530* → (0.0000) ←	46.7532* → (0.0000) ←	28.9611* → (7.E-241) ←	7.84743* → (6.E-50) ←
ECG2	7.34840* → (1.E-45) ←	34.0190* → (9.E-287) ←	28.3777* → (1.E-235) ←	21.4622* → (2.E-172) ←	4.70172* → (1.E-23) ←
ECG3	8.90013* → (4.E-59) ←	43.8040* → (0.0000) ←	38.3062* → (0.0000) ←	23.9416* → (4.E-195) ←	3.79965* → (8.E-17) ←

MGH202

Lags 45	ART	CVP	PAP	RESP	CO_2
ECG1	17.3587* (3.E-132) ←	26.6135* → (7.E-215) ←	6.24928* (1.E-35) ←	9.14276* → (6.E-60) ←	0.76490 (0.8737)
ECG2	10.5459* (4.E-72) ←	15.500* → (1.E-115) ←	7.10172* → (1.E-42) ←	9.24585* → (7.E-61) ←	13.1002* → (9.E-95) ←
ECG3	85.9955* (0.0000) ←	67.0712* → (0.0000) ←	71.9574* → (0.0000) ←	3.19624* → (3.E-12) ←	13.3631* (8.E-97) ←

MGH227

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	17.3587* → (3.E-132) ←	26.6135* → (7.E-215) ←	6.24928* (1.E-35) ←	9.14276* → (6.E-60) ←	0.76490 (0.8737)
ECG2	7.11487* (1.E-43) ←	27.5168* → (1.E-227) ←	26.1840* → (1.E-215) ←	25.4145* → (1.E-208) ←	6.04228* (1.E-34) ←
ECG3	2.68736* (5.E-09) ←	32.9532* → (4.E-277) ←	36.3138* → (2.E-307) ←	33.3783* → (6.E-281) ←	2.52247* (6.E-08) ←

MGH229

Lags 46	ART	CVP	PAP	RESP	CO_2
ECG1	8.36020* (2.E-54) ←	37.9324* → (0.000) ←	8.91593* → (3.E-59) ←	274.813* → (0.000) ←	297.264* → (0.0000) ←
ECG2	0.89056 (0.6825)	1.1944 (0.1723)	1.16756 → (0.2032)	1.71417* (0.0019) ←	1.44555 (0.0258)
ECG3	1.23840 (0.1292)	0.61661 (0.9809)	1.88425* → (0.0003) ←	1.16677 (0.2042)	1.14884 (0.2270)

Discussions

In order to establish Granger causality on the multivariate analysis dealing with cardiac, respiratory, and hemodynamic signals, in contrast to existing work on the heart by taking generally only two signals (two variables) to maximum, our work focuses on 187 patients, but we displayed the results for 13, because the required sizes of the item force us to do that.

Our study was fulfilling on a case-by-case basis, we will accomplish that for two patients randomly just to see correspondence (Granger causality/Current status of patients).

MGH002: We evaluate after the measures that all blood pressures have a bilateral relationship with the cardiac signals and the rate of CO_2 and breathing as they only have a bilateral relationship only with ECG1 and this may be explained by a vicious circle generated by the heart ectopic that causes disorders of blood tensions which in it turn acted on the heart highlighting the disorder.

MGH019: Atrial fibrillation is a serious phenomenon, which is characterized by rapid acute contractions and disordering of the auricle of the heart, which can cause cardiac arrest. Due to this, there is a dangerous increase in blood pressure, the resulting physical stress causes hyperventilation, this pressure increase complicates the task of the heart that has been already weakened and disturbs the phenomenon of oxygen uptake and expulsion of CO_2 from pulmonary alveoli, which aggravates the situation of the heart that has a workload coupled with a deficiency of these drivers.

We will present a table that accounts the results in proportion to better understand the influence of signals studied. The table below contains all the results (186 patients).

ALL patients	ART (%)	CVP (%)	PAP (%)	RESP (%)	CO_2 (%)
ECG1	85.71 → 97.91 ←	73.46 → 81.63 ←	91.83 → 93.87 ←	83.67 → 93.87 ←	75.51 → 93.87 ←
ECG2	87.75 → 93.87 ←	75.51 → 79.59 ←	87.75 → 89.79 ←	85.71 → 97.95 ←	81.63 → 95.91 ←
ECG3	79.59 → 87.75 ←	69.38 → 77.55 ←	85.71 → 91.83 ←	83.67 → 89.79 ←	71.42 → 91.83 ←

Conclusion

Our methods are based and validated by the Granger causality. The mathematical search result obtained by this method could confirm the cardiorespiratory hemodynamic anatomy. The knowledge and the quantitative understanding of these interactions are critical in monitoring people at risk situations (awakening from anesthesia, age-related pathologies that followed pregnant women, etc.). So for our future telemedicine applications it is a real progress towards the complete analysis of signals received. Based on these results, and with the inclusion of all the interdependencies with these specific degrees protocols; that enable an excellent intervention.

References

1. Granger C.W.J.: Investigating causal relations by econometric and cross-spectral methods. Econometrica, 424–438 (1969)
2. Hong, Y., Liu, Y., Wang, S.: Granger causality in risk and detection of extreme risk spillover between financial markets. J. Econ. **150**(2), 271–287 (2009)
3. Mokhov, I. I., Smirnov, D.A.: El Nino-Southern oscillation drives North Atlantic oscillation as revealed with nonlinear techniques from climatic indices, Geophys (2006)
4. Triacca, U.: Is Granger causality analysis appropriate to investigate the relationships between atmospheric concentration of carbon dioxide and global surface air temperature? Theor. Appl. Climatol. **81**, 133–135 (2005)
5. Diks, C., Mudelsee, M.: Redundancies in the Earth's climatological time series. Phys. Lett. A **275**, 407–414 (2000)
6. Pierre-Olivier, A., Olivier, J.J. Michel.: On directed information theory and Granger causality graphs. J. Comput. Neurosci. 7–16 (2011)
7. Carolin Ligges, M., Ungureanu, M., Ligges, H., Witte: Understanding the time variant connectivity of the language network in developmental dyslexia: new insights using Granger causality. J. Neural. Transm. 529–543 (2010)
8. Xiang, Li., Kaiming, Li., Lei, G., Chulwoo, L., Tianming, L.: Fiber-centered granger causality analysis. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2011. Lect. Notes. Comput. Sci. **6892** 251–259 (2011)
9. Pereda, E., Quiroga, R.Q., Bhattacharya, J.: Nonlinear multivariate analysis of neurophysiological signals. Prog. Neurobiol. 1–37 (2005)
10. Yifan, Z., Steve, A., Billings, Hua-Liang, W., Ptolemaios, G., Sarrigiannis.: A parametric method to measure time-varying linear and nonlinear causality with applications to EEG data. IEEE. 1–7 (2013)
11. Florin, E., Gross, J., Pfeifer, J., Fink, G.R., Timmermann, L., Reliability of multivariate causality measures for neural data. J. Neurosci. Methods. 344–358, (2011)
12. Winfried, S., Vera, L., Iris-Tatjana, K., Nathan, W., Thomas, E.: Age-related changes in neural functional connectivity and its behavioral relevance. BMC Neurosci. 2–11 (2012)
13. Schad, A., Nawrath, J., Jachan, M., Henschel, K., Spindeler, L., Timmer, J., Schelter, B.: Approaches to the detection of direct directed interactions in neuronal networks. Springer Ser. Comput. Neurosci. **2**, 43–64 (2009)
14. Yang, C., Le Bouquin Jeannès, R., Faucon, G., Wendling, F.: Detecting causal interdependence in simulated neural signals based on pairwise and multivariate analysis. 32nd Annual International Conference of the IEEE EMBS, Buenos Aires, Argentina, August 31–September 4, 162–169 (2010)

15. Laura, A., Hovagim, B., Febo, C., Donatella, M., Maria, G.M., Fabrizio, De Vico, F., Alfredo, C., Serenella, Salinari, Fumikazu, Yoko, Y., Pablo, M., Andrzej, C., Andrea, T., Fabio, B.: Estimate of causality between independent cortical spatial patterns during movement volition in spinal cord injured patients. Brain Topogr. **19**(3), 107–123 (2007)
16. Niso, G., Bruña, R., Pereda, E., Gutiérrez, R., Bajo, R., Maestú, F., del-Pozo, F.: HERMES: towards an integrated toolbox to characterize functional and effective brain connectivity. Springer, Philadelphia (2013)
17. David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., Depaulis, A.: Identifying neural drivers with functional MRI: an electrophysiological validation. PLoS Biol. **6**, 2683–2697 (2008)
18. Paolo, Z., Gianna, M.T., Elisa, S., Anna, B., Fabrizio, B., Nivedita, A., Eugene, G.: The human brain pacemaker: synchronized infra-slow neurovascular coupling in patients undergoing non-pulsatile cardiopulmonary bypass. Neuroimage **72**, 10–19 (2013)
19. Ge, T., Kendrick, K., Feng, J.: A novel extended granger causal model approach demonstrates brain hemispheric differences during face recognition learning. PLoS Comput. Biol. **5**, e1000570 (2009)
20. Ge, T., Feng, J., Grabenhorst, F., Rolls, E.T.: Componential Granger causality and its application to identifying the source and mechanisms of the top-down biased activation that controls attention to affective vs sensory processing. Neuroimage **59**, 1846–1858 (2012)
21. Palu, S.M., Stefanovska, A.: Phys. Rev. E. **67** 055201R (2003)
22. Verdes, P.F.: Phys. Rev. E. **72**(2) 026222 (2005)
23. Faes, L., Porta, A., Cucino, R., Cerutti, S., Antolini, R., Nollo, G.: Causal transfer function analysis to describe closed loop interactions between cardiovascular and cardiorespiratory variability signals. Biol. Cybern. **90**, 390–399 (2004)
24. Faes, L., Widesott, L., Del Greco, M., Antolini, R., Nollo, G.: Causal cross-spectral analysis of heart rate and blood pressure variability for describing the impairment of the cardiovascular control in neurally mediated syncope. IEEE Trans. Biomed. Eng. **53**, 65–73 (2006)
25. Nollo, G., Faes, L., Porta, A., Antolini, R., Ravelli, F.: Exploring directionality in spontaneous heart period and systolic pressure variability interactions in humans: implications in the evaluation of baroreflex gain. Am. J. Physiol. Heart Circ. Physiol. **288**, 1777–1785 (2005)
26. Nollo, G., Faes, L., Porta, A., Pellegrini, B., Ravelli, F., Del Greco, M., Disertori, M., Antolini, R.: Evidence of unbalanced regulatory mechanism of heart rate and systolic pressure after acute myocardial infarction. Am. J. Physiol. Heart Circ. Physiol. **283**, 1200–1207 (2002)
27. Pereda, E., de La Cruz, D.M., De Vera, L., Gonzalez, J.J.: Comparing generalized and phase synchronization in cardiovascular and cardiorespiratory signals. IEEE Trans. Biomed. Eng. **52**, 578–583 (2005)
28. Giandomenico, N., Michela, M., Walter, M., Roberta C., Luca F.: Assessment of a prototype equipment for cuffless measurement of systolic and diastolic arterial blood pressure. J. Electrocardiol. **44**(2) (2010). doi: 10.1016/j.jelectrocard
29. Faes, L., Nollo, G., Porta, A.: Information based detection of nonlinear Granger causality in multivariate processes via a non-uniform embedding technique. Phys. Rev. E. **83**, 051112 (2011)
30. Hlavackova-Schindler, K., Palus, M., Vejmelka, M., Bhattacharya, J.: Causality detection based on information-heoretic approaches in time series analysis. Phys. Rep. **441**, 1–46 (2007)
31. Granger, C.W.J.: Testing for causality. A personal viewpoint. J. Econ. Dyn. Control **2**, 329–352 (1980)
32. Malliani, B.: A principles of cardiovascular neural regulation in health and disease. Kluwer, Norwell (2000)
33. Riedl, M., Suhrbier, A., Stepan, H., Kurths, J., Wessel, N.: Short-term couplings of the cardiovascualr system in pregnant women suffering from pre-eclampsia. Phil. Trans. Royal Soc. A. **368**, 2237–2250 (2010)
34. Kaminski, M., Ding, M., Truccolo, W.A., Bressler, S.: Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. Biol. Cybern. **85**, 145–157 (2001)

35. Wiener, N.: The theory of prediction. In: Beckenbach, E.F. (ed.) Modern mathematics for engineers. McGraw-Hill, New York (1956)
36. Laude, D., Elghozi, J.L., Girard, A., Bellard, F., Bouhaddi, M., Castiglioni, P., Cerutti, C., Cividjian, A., di Rienzo, M., Fortrat, J.O., Janssen, B., Karemaker, J.M., Leftheriotis, G., Parati, G., Persson, P.B., Porta, A., Quintin, L., Regnard, J., Rudiger, H., Stauss, H.M.: Comparison of various techniques used to estimate spontaneous baroreflex sensitivity (the EuroBaVar study). Am. J. Physiol. **286**, 226–231 (2004)
37. Luca, F., Giandomenico, N., Chon, K.I.H.: Assessment of granger causality by nonlinear model identification: application to short-term cardiovascular variability. Ann. Biomed. Eng. **36**(3), 381–395 (2008). doi:10.1007/s10439-008-9441-z
38. Xiao, X., Mullen, T.J., Mukkamala, R.: System identification: a multi signal approach for probing neural cardiovascular regulation. Physiol. Meas. **26**, 41–71 (2005)
39. http://www-timc.imag.fr/article886.html
40. Yelda, A., Tara, L., Alvarez, Suril, G., Paul, A., Taylor, Bharat, B.: Functional connectivity in vergence and saccade eye movement tasks assessed using granger causality analysis. 33rd Annual International Conference of the IEEE EMBS, Boston August 30–September 3. 8114–8117 (2011)