Gujarat Journal of Statistics and Data Science

Volume: 38 (New Series) **Issue: 1** **October 2022**

Published by

Gujarat Statistical Association

GUJARAT JOURNAL OF STATISTICS AND DATA SCIENCE
(Formerly GUJARAT STATISTICAL REVIEW)

Editorial Board

Founding Editor
Professor (Late) C. G. Khatri

Editor in Chief
Prof. (Retd.) D. K. Ghosh
UGC BSR Faculty Fellow
Marwadi University, Rajkot, Gujarat, India

Managing Editor

Prof. (Retd.) Bikas K. Sinha	ISI Kolkata

Editor

Prof. Ashis SenGupta	Emeritus Scientist, CSIR, Govt. of India; Augusta University, USA; ISI, Kolkata (Retd.); Middle East Technical University, Turkey

Associate Editors

Prof. Arnab Kumar Laha	IIM Ahmadabad
Prof. Chetna D. Bhavsar	Gujarat University, Ahmadabad
Prof. Debasis Kundu	IIT Kanpur
Prof. H. V. Kulkarni	Shivaji University, Kolhapur
Prof. K. Muralidharan	M. S. University Vadodara
Prof M. N. Patel	Gujarat University Ahmadabad
Prof. Mausumi Bose	ISI Kolkata
Prof. S. C. Bagui	University of West Florida, USA
Prof. Satheesh Kumar	University of Kerala, Trivandrum
Prof. Saumen Mandal	University of Manitoba, Canada
Prof. Thomas Mathew	University of Maryland, USA

Advisory Board

Prof. B. L. S. Prakasa Rao	C.R.R.A.I.M.S.C.S., Hyderabad, and Former Director & Emeritus Professor, I. S. I. Kolkata
Prof. (Retd) B. B. Jani	Gujarat University, Ahmedabad
Dr. Chirag Trivedi	R. J. T. College, Ahmedabad
Prof. Jyoti Divecha	S. P. University, V.V. Nagar
Prof. Kalpana Mahajan	Panjab University, Chandigarh
Prof. (Retd) P. C. Gupta	V. N. South Gujarat University, Surat
Prof. Piyush Kant Roy	B. H. U. Varanasi
Prof. Rakesh Srivastava	M. S. University Vadodara

Editorial Secretary

Dr. Parag Shah	Gujarat University

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Vol – 38 Issue – 1 October, 2022

CONTENTS PAGE

What is data science and what is big data?-An overview 1-6
B.L.S. Prakasa Rao

Univariate and bivariate discrete Laplace and generalized discrete Laplace distributions 7-18
Barry C. Arnold and Matthew A. Arvanitis

On Zero-inflated Generalized Hyper Poisson Distribution and its Properties 19-28
C. Satheesh Kumar and Rakhi Ramachandran

Construction of Bivariate Burr (Type VII) distributions via conditional specification: A brief survey 29-38
Indranil Ghosh

The simultaneous assessment of normality and homoscedasticity for bivariate grouped data 39-53
Ye Yang and Thomas Mathew

Heuristic tests for testing against ordered alternatives in heteroscedastic ANOVA 54-71
Anjana Mondal and Somesh Kumar

The Theory of k-NN Classification Rules: T.M. Cover and P.E. Hart (1967) Results and Extensions Revisited 72-86
Subhash Bagui and KL Mehra

Generalized inverses of some matrices associated with trees 87-94
Ravindra B. Bapat

Properties of the matrix $V + XTX'$ in linear statistical models 95-106
Stephen J., Augustyn Markiewiczm and Simo Puntanen

Optimal Designs in An Irregular Design Space 107-131
Ravindra Khattree

Agreement Between Raters for Hierarchical Designs with Repeated Measurements 132-150
Hairong Shi, Soumya Sahu, Domenic J. Reda, Bikas K. Sinha,
Dulal K. Bhaumik

Some Applications of Mixture Regression Modelling Techniques 151-165
Tapio Nummi, Jyrki Mottonen, Janne Salonen, Timothy E. O'Brien

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Vol – 38 Issue – 1 October, 2022

Gujarat Journal of Statistics and Data Science
(Formerly Gujarat Statistical Review)

Editor in Chief: message

The special volume of Gujarat Journal of Statistics and Data Science is being published as a memorial volume for Late Professor C. G. Khatri, Retired Professor and Head, Department of Statistics, Gujarat University, Gujarat, India. Dr. Khatri was one of the stalwart experts in the field of Multivariate analysis, eminent researcher in Matrix Algebra, specialised in generalized inverse of matrices. He had jointly published several research papers on generalized inverse of matrices and Multivariate distributions in various national and international journals, co - author with the legendary figures in the world of Statistics like Professor C. R. Rao and Late Professor S. K. Mitra. He was frequent visitor at Indian Statistical Institute, Kolkata, Delhi and at several universities in Canada and USA.

My memory goes back to 1984 when for the first time I met Professor C. G. Khatri at the Vice- Chancellor's office, South Gujarat University, Surat, while I was facing him for an interview for the post of Associate Professor. He asked a few questions on variance balanced design as he had published one paper on variance balanced design in 1982 and my Ph.D. thesis was containing one chapter on VB design. The next meeting with him was at the Department of Statistics, Gujarat University, where he was conducting twenty-one days course on Linear Models, funded by UGC, New Delhi. Twenty-one days of academic learning under his brilliant guidance made me a good research worker in Linear Models and Design of Experiments. He was pioneer in organising the annual conference of Gujarat Statistical Association annually at various colleges/universities in Gujarat and we used to meet him.

It gives me utmost heartiest pleasure in being associated with the noble task of publishing a journal as a memorial volume for a person who had earned name and fame as a Statistician in India and in the world. Since the new name of the journal is Gujarat Journal of Statistics and Data Science, hence we included the first paper on Data science authored by Professor B.L.S. Prakasa Rao. I was entrusted with the responsibility of Editor in Chief of Gujarat Journal of Statistics and Data Science and to edit this volume in the meeting of the Gujarat Statistical Association held on August 2021. I am grateful to the committee members of the Gujarat Statistical Association for having confidence and faith on me to handle and complete this mighty task. Moreover, with the full support from all office bearers of GSA and in particular the President, Dr N D Shah, we could restart the publication of this prestigious journal. I am really happy in bringing up the journal in time with research papers of the National and International reputed authors. Professor Bikas Kumar Sinha, Managing Editor and Professor Ashis SenGupta, Editor of this journal are extremely helpful at every stage with extreme courtesy for having the publication of this volume from beginning to end. We used to discuss various queries and problems with regard to this volume either by email or by mobile. I was able to solve the problems with their advices and suggestions. It has been a wonderful and truly rewarding experience to work with them. I am also extremely grateful to Professor B.L.S. Prakasa Rao, former Director and Emeritus Scientist of I S I Kolkata who suggested the new name of this journal as Gujarat Journal of Statistics and Data Science. One more person Dr. Parag Shah deserves special mention because

he has followed thru with all the editorial works and looked at stage by stage progress with extreme interest. He has been quite helpful in compiling the papers with manuscripts number and then making correspondence with authors, Editor, Managing Editor and Editor - in - Chief from time to time.

All the contributory authors and referees were seriously involved in their respective academic activities for which we have passed thru a lengthy editorial process over the last several months. I am extremely thankful to them for their academic and scientific interest in completing this task.

With these few words, I place the special volume of Gujarat Journal of Statistics and Data Science before our readers at large. We fondly hope they will not be disappointed with this volume.

Dilip Kumar Ghosh

Editor in Chief

Preface

My recollection of meeting Late Professor Chinubhai Ghelabhai *Khatri* [CGKhatri] goes back to 1970 during 57th Indian Science Congress at IIT, Kharagpur. As research scholars at CUDS, we [twin brothers] attended the same. Among others, Professors C R Rao and C G Khatri were both prominently visible in all sessions. After almost every presentation of young budding scholars, CGK would throw Qs like a roaring tiger but CRR was remarkably cool and would make observations in his characteristic smiling style. That was CGK – all along very much vocal and loud.

Over time … like many other seniors in our profession, Khatri became our Dear Chinu-da. We met him many times within India and abroad mainly at conferences. I used to meet him at ISI, Kolkata as well. He had a special liking for Sinha Brothers.

When DKG approached me, I did not give any second thought. I am delighted to have this opportunity to serve this Journal as its Managing Editor. With DKG as Chief Editor and ASG as the Editor and with a strong group of Advisory Board Members and Editorial Board Members, I am confident this revival of the journal will reflect on the passion of Late Prof. Khatri for the subject in terms of Teaching and Research in his favourite topics.

This first issue of GJSDS has been carefully crafted to reflect on the topics most liked by CGK. We fondly hope it will attract attention and satisfaction of our readers at large.

Bikas K Sinha

Kolkata

July 12, 2022

Forewords

It was the Summer (June) of 1979, I was a Ph.D. student and it was my very first presentation of a research paper in an international conference, Symposium on Variance Components, g-inverses and Applications, organized by Prof. C.R. Rao at The Ohio State University, Columbus, USA. Hardly had I completed, when someone from the front voiced very strongly that the original result, I had generalized was in fact established by him and not by CR Rao as I had stated. I politely answered that I had seen Prof. Rao's work and was ignorant about his result. Prof. C.R. Rao had his characteristic soft smile! Later I was told that the person was no other than Prof. C.G. Khatri! His passion with the research paper had left a lasting impression on me. Later in my research and teaching, I have used the classic book, *An Introduction to Multivariate Statistics*, by M. S. Srivastava, C. G. Khatri, North-Holland/New York, 1979, and I obviously came to greatly appreciate the profound impact Prof. Khatri has made in multivariate statistical inference and matrix theory. It was thus with reverence and pleasure that I accepted to be the Editor of this revival of Gujarat Statistical Review of which Prof. Khatri was the Founder-Editor. This new version will certainly progress in the footsteps of the earlier one as per the aims of Prof. Khatri. Moreover, we will also strive to fulfil our vision to enrich this journey with the emerging faces in theory and methodology, as well as with the challenges in real-data problems, which are surfacing in the arena of statistical science. With this opening issue, I invite senior and young researchers alike to submit their scholarly papers to our journal to take it to greater heights.

Ashis SenGupta

Editor

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

WHAT IS DATA SCIENCE AND WHAT IS BIG DATA?-AN OVERVIEW

B.L.S. Prakasa Rao [1]

1. *CR RAO Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad 500046. Email: blsprao@gmail.com*

Received: 15 Jun 2022 / Revised: 20 Jul 2022 / Accepted: 28 Jul 2022

Abstract

We give an overview of what is considered to be "Data Science" as opposed to the traditional "Statistics" program and explain some ideas concerning "Big Data" . Our overview is based on an excellent review article entitled "50 Years in Data Science" by David Donoho [4](*Journal of Computational and Graphical Statistics*, Vol. 26 (2017), 745-766) and some other recent articles on the web.

1 What is Data Science?

In a review of a book on Data Science, one reviewer commented "When trying to define Data Science, the ancient Buddhist parable of the blind men and the elephant springs to mind. The entrepreneur, the academic researcher and the university administrator approach the beast in turn and examine it but report something different. The entrepreneur sees the training of a generation of data workers to deal with large data sets as a *business opportunity*. The university administrator hears the siren's call of new programs and students, while the academic researcher is still trying to figure out where the science is. An exact definition of Data Science remains elusive" (cf. [11]).

The subject of Statistics is undergoing a change. Several Departments of Statistics are trying to rename their Department as the "Department of Data Science" and awarding degrees in Data Science programs. It is now what can be termed as " Data Science Moment" .

The question is whether Data Science is really different from Statistics. According to the Data Science Association (DSA) of USA "Data Science means the scientific study of creation, validation and transformation of data to create meaning. A "Data Scientist" is a professional who uses scientific methods to liberate and create meaning from raw data - somebody who can play with data, spot trends and learn truths few others know" .

According to American Statistical Association (ASA) "While there is not yet a consensus on what precisely constitutes Data Science, three professional communities are emerging as foundational to Data Science: (i) Database Management to enable the transformation, conglomeration and organization of data resources; (ii) Statistics and Machine Learning to convert data into knowledge; (iii) Distributed and Parallel systems to provide the computational infrastructure to carry out data analysis. Statistics and Machine Learning play a central role in Data Science" .

Data Science is experiencing rapid and unplanned growth spurred by the proliferation of complex and rich data in science industry and government. Some remark that the Data Science is the science for acquisition, management, analysis and inference from data (cf. PCMI Undergraduate faculty group (2018)[6]).

Data Science is in the intersection of Statistics, Computer Science and substantive application domains of these subjects. From Computer science comes Machine Learning and high performance

© 2022 Author(s). (https://www.thegsa.in/).

computing technologies for dealing with scale. From the subject of Statistics comes the tradition of exploratory data analysis, testing and visualization (cf. [10]).

New technology makes it possible to capture, annotate and store vast amounts of social media, logging and sensor data. Computing advances make it possible to analyze data in novel ways and at ever increasing scales (cf. [10]).

Data Science can be considered as a superset of the fields of statistics, machine learning and now AI (Artificial Intelligence) which adds some technology for scaling up to "Big Data" . The idea is to "learn from data" . Data Science has become a fourth approach to scientific discovery in addition to experimentation, modeling and computation. Data Science leading to scientific discovery and practice involves the collection, management, processing, analysis, visualization and interpretation of vasts amounts of heterogeneous data associated with a diverse scientific and interdisciplinary applications. Data Scientist is a professional who uses scientific methods to discover and create meaning from raw data.

Statistics also means the practice of science of collecting and analyzing numerical data in samples large as well as small. To a statistician, this definition of statistics seems already to encompass any thing that the definition of data scientist might think. However a lot of statistical work deals with inferences to be made from small samples as well as large samples.

The question is "Is Data Science just a "re-branding" of Statistics?" "Why do we need Data Science when we had Statistics for the last 100 years?" When physicists do mathematics, they do not say that they are doing number science. They are doing mathematics. If you are analyzing data, you are doing Statistics no matter by whatever name you call it. You can call it Data Science or Informatics or Big Data Analytics but it is still Statistics!!

Is Data Science a Science? There is a considerable debate about what the science of Data Science is. Data Science is inherently interdisciplinary. Working with data requires the mastery of a variety of skills and concepts including many traditionally associated with the fields of Statistics, Computer Science and Mathematics. The cycle of obtaining, cleaning and processing data, exploring data, defining questions, performing analysis and communicating the results lie at the core of Data Science experience (cf. PCMI Under graduate faculty group (2018)[6]).

"Data Scientist" means a professional who uses scientific methods to analyze and create meaning from raw data. He or she should have a solid foundation in machine learning, algorithms, modeling , statistics, mathematics and strong business acumen coupled with the ability to communicate findings to business leaders in a way that can influence how an organization approaches a business challenge.

Data Scientists employ models to understand the world and the subjects of mathematics and statistics provide the language for these models. A working data scientist requires a firm foundation in mathematics, statistics and computer science. A student of data science should have competencies in computational and statistical thinking, mathematical foundations, model building and assessment, algorithmic and software foundation, data cleaning and knowledge transfer and communication (cf. PCMI Under graduate faculty group (2018)[6]).

Data Scientist is a person who is better at statistics than any software engineer and better at software engineering than any statistician!!

2 Data Science and Big Data

Some people think that "Big Data" is different from Statistics but is a part of Data Science. This is not correct. Statisticians have been working with large data sets for several decades. As a mathematical statistician, I have been working in asymptotics dealing with large samples for decades. The concepts of sampling and sufficiency deal with large samples effectively and they were formulated by statisticians

dealing with large data sets.

Computer scientists seem to think that Data Science is concerned with really big data which traditional computing packages and other resources could not accommodate and data science trainees need additional skills to analyze big data sets.

More than 60 years ago, John Tukey [12], Professor of Statistics, Princeton university called for a reformulation of academic statistics, that is, statistics we teach for our undergraduate and master's students. His paper on "The Future of Data Analysis" was published in *The Annals of Mathematical Statistics, Vol. 33 (1962) 1-67*. He pointed out to the existence of an as yet unrecognized science whose subject of interest was "Data Analysis" . Data science involves the coupling of scientific discovery and practice which in turn consists of the collection, management, processing, analysis, visualization and interpretation of vasts amounts of heterogeneous data associated with a diverse array of scientific and inter-disciplinary applications.

Large data sets are encountered, for example, in meteorology, genomics, biological and environmental research. They are also present in other areas such as internet search, finance and business informatics. Data sets are big as they are gathered using sensor technologies. There are also examples of Big Data in areas which we can call Big Science and in Science for research. These include "Large Hadron Collision Experiment" which represent about 150 million sensors delivering data at 40 million times per second. There are nearly 600 million collisions per second. After filtering and not recording 99.999%, there are 100 collisions of interest per second.The Large Hadron Collider experiment generates more than a peta byte (1000 trillion bytes) of data per year. Astronomical data collected by Sloan Digital Sky Survey (SDSS) is an example of Big Data. Decoding human genome which took ten years to process earlier can now be done in a week or even in a few hours. This is also an example of a Big Data. Human genome data base is an example of a Big Data. A single human genome contains more than 3 billion base pairs. The 1000 Genomes project has 200 terabytes (200 trillion bytes) of data. Human brain data is an example of a Big Data. A single human brain scan consists of data on more than 200,000 voxel locations which could be measured repeatedly at 300 time points.

For Government, Big Data is present for climate simulation and analysis and for national security areas. For private sector companies, Big Data comes up from millions of back-end operations every day involving queries from customer transactions, from vendors etc. Big Data sizes are a constantly moving target. It involves increasing volume (amount of data), velocity (speed of data in and out) and variety (range of data types and sources). Big Data are high in volume, high in velocity and/or high in variety information assets. It requires new forms of processing to enable enhanced decision making, insight discovery and process optimization (cf. Prakasa Rao [7][8]).

During the last twenty years, several companies abroad are adopting to data-driven approach to conduct more targeted services to reduce risks and to improve performance. They are implementing specialized data analytics to collect, store, manage and analyze large data sets. For example, available financial data sources include stock prices, currency and derivative trades, transaction records, high-frequency trades, unstructured news and texts, consumer confidence and business sentiments from social media and internet among others. Analyzing these massive data sets help measuring firms risks as well as systemic risks. Analysis of such data requires people who are familiar with sophisticated statistical techniques such as portfolio management, stock regulation, proprietary trading, financial consulting and risk management.

Big Data are of various types and sizes. Massive amounts of data are hidden in social net works such as Google, Facebook, Linkedin, YouTube and Twitter. These data reveal numerous individual characteristics and have been exploited. Official statistics from the government is a Big Data (cf.[9]).

3 Exploratory Data Analysis

In his 1962 paper, John Tukey identified four driving forces behind "Data Science" which he called "Data Analysis" : (i) Formal theories of statistics; (ii) Developments in computer architecture and associated software; (iii) Large data sets; (iv) Quantification in different disciplines. Cleveland [3] wrote an article entitled "Data Science: An action plan for expanding the technical areas of the field of Statistics" . He proposed six focused points of activity for teaching data science: (i) Multidisciplinary investigations; (ii) Models and methods of data; (iii) Computing with Data; (iv) Teaching of techniques; (v) Evaluation of computer packages;(vi) Theory of statistics.

Over the last 50 years, many statisticians and data analysts developed computational environment for data analysis such as BMDP, SPSS, SAS and MINITAB and more recently packages such as R in the personal computer era. The programing language R is today the dominant quantitative programming environment.

Breiman [2] said "Statistics" starts with data. Think of the data as being generated by a black box in which a vector of input variables x (independent variables) go in one side and on the other side the response variables y come out. Inside the black box, nature functions to associate the predictor variable with the response variable. There are two goals to analyzing the data: (1) Prediction: To be able to predict what the responses are going to be to future input variables; (2) Inference: To infer how nature is associating the response variables to the input variables.

Let us see what is taught in today's Data Science programs at some of the best universities with statistics programs. (1) Research design and applications for Data and Analysis; (2) Exploring and analyzing Data; (3) Storing and Retrieving Data; (4) Applied Machine Learning; (5) Data Visualization and Communication.

Most of the departments of statistics do not teach about "Storing and Retrieving Data" . Machine learning is a rapidly growing field at the intersection of computer science and statistics concerned with finding patterns in data. It is responsible for advances in technology from personalized product recommendations to speech recognition in cell phones. Understanding of probability, statistics and linear algebra is important for "Applied Machine Learning" course. Students have to be taught tools such as "R" and "Python" for handling large data sets. Data Science course must involve a data analysis project with a large data set.

Master's program in Data Science should be a mixture of some material out of a Statistics master's program to learn techniques to analyze large data bases and some material from Computer Science master's program with inputs from statistics and machine learning.

The activities of Data Science can be classified into six parts: (1) Data gathering, Preparation and Exploration; (2) Data representation and transformation; (3) Computing with Data; (4) Data Modeling; (5) Data Visualization and Presentation; (6) Science about Data science.

Data gathering, Preparation and Exploration

Most of the effort devoted to Data Science is spent by diving into messy data to learn the basics of what is in them so that the data can be made ready for exploration. Data gathering includes traditional experimental design, survey sampling and other techniques as practiced by statisticians but also a variety of modern data gathering techniques from data resources such as Google and Wikipedia. Many data sets contain anomalies, mistakes and misprints. Any data driven project requires identifying such issues. The data might need grouping, smoothing and sub-setting. This process is some times called as "data cleaning" . Tukey had coined the term "Exploratory Data Analysis" . Every data scientist should devote time and effort to explore data to check its basic properties and to expose its unexpected features. This work adds crucial insights to every data driven project. Mahalanobis

discussed this issue and termed this process as "scrutiny of data" .

Data representation and transformation

A data scientist works with many different data sources. They will be in a wide range of formats and the data scientist should be able to handle them. Hardware and software constraints are a part of them. Data Scientists may find a transformation restructuring the original data into a new and more revealing form.

Data Processing

Every data scientist should know and use several languages for data analysis and data processing. These include R and Python and also specific languages for transforming and manipulating text and for managing complex computational aspects. Cluster and cloud computing and the ability to run massive numbers of jobs on such clusters is important aspect of modern computational scenario. Data scientists should be able to develop work flows which organize work to be split up over many jobs which can be run sequentially or across many machines in parallel.

Computing with data

R language transformed the practice of data analysis by creating a standard language which different analysts can all use to communicate and share algorithms and work flows. R language is used in many online presentations about data science initiatives.

Tidy data: Eighty percent of data analysis is spent on the process of cleaning and preparation of data. Wickham [13] developed a systematic way of thinking about "messy data formats that are commonly encountered in data analysis and shows how to transform each such format into a tidy format. He introduced a set of tools in R to transform the data into a universal "tidy" data format.

Data Modeling

(1) Generative modeling: Here one proposes a stochastic model that could have generated the data and derives methods to infer properties of the underlying generative mechanism. (2) Predictive modeling: Here one constructs methods which predict well over a given specific data set in line with machine learning ideas.

Data Visualization and Presentation

Data visualization starts with standard histograms, scatter plots and time series plots but in recent times, it can take more elaborate shapes. Data scientists decorate simple plots with additional colour or symbols to bring in important new factors. They create dashboards for monitoring data processing pipelines that access streaming or widely distributed data.

Science about Data Science

Data scientists are doing science about *Data science* which they identify commonly occurring analysis or process work flows.

Remark 1. *Each proposed notion of data science involves some enlargement of Statistics we teach and Machine Learning. Data Science is the science of learning from data. It studies the methods involved in the analysis and processing of data and proposes technology to improve methods in evidence based manner. Harvard University has started a journal recently named "The Harvard Data Science Review" . According to the Editor-in-Chief of the journal, the mission of the "Harvard Data Science Review" is to really help to define and shape what Data Science is exactly.*

Acknowledgment

This overview is based on several excellent articles mentioned in the list of references and other sources available on the web. Thanks are due to all the authors for their insights and explanations to understand the subject of Data Science and Big data. Work on this review was supported by INSA (Indian National Science Academy) under the INSA Senior Scientist program at the CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad, India.Thanks are due to Professor Dilip Kumar Ghosh, Editor for inviting me to submit an article for the first issue of this journal.

References

[1] B. Baumer. "A Data Science Course for Undergraduates: Thinking With Data." In: *The American Statistician* 69 (2015), pp. 333–342.

[2] L. Breiman. "Statistical Modeling: the Two Cultures." In: *Statistical Science* 16 (2001), pp. 199–231.

[3] W.S. Cleveland. "Data Science: An Action Plan for Expanding the Technical Areas of the Fields of Statistics." In: *International Statistical Review* 69 (2001), pp. 21–26.

[4] David Donoho. "50 Years of Data Science." In: *Journal of Computational and Graphical Statistics* 26 (2017), pp. 745–766.

[5] N.J. Horton, B.S. Baumer, and H. Wickham. "Taking a Chance in the Classroom: Setting the Stage for Data Science: Integration of Data Management Skills in Intro- ductory and Second Courses in Statistics." In: *Chance* 28 (2015), pp. 40–50.

[6] Park City Math Institute (PCMI). *Undergraduate Faculty Group (2018) Curriculum guide- lines for undergraduate programs in data science,* 2018.

[7] B.L.S. Prakasa Rao. *Brief Notes on Big Data: A Cursory Look, Lecture Notes*. 2015.

[8] B.L.S. Prakasa Rao. "Brief Notes on Big Data, Visleshana." In: *Computer Society of India, Special Interest Group - Big Data Analytics* 1 (3) (2017), pp. 9–12.

[9] Pyne. S., B.L.S. Prakasa Rao, and S.B. Rao. *Big Data Analytics*. Springer India, New Delhi, 20016.

[10] S.S. Skiena. *The Data Science Design Manual, Texts in Computer Science*. 2017.

[11] B. Steele, J. Chandler, and S. Reddy. "Review of Algorithms for Data Science." In: *Bull. Amer. Math. Soc.* 56 (2019), pp. 143–149.

[12] John W. Tukey. "The Future of Data Analysis." In: *The Annals of Mathematical Statistics* 33 (1962), pp. 1–67.

[13] H. Wickham. "Tidy Data." In: *Journal of Statistical Software* 59 (2014), pp. 1–23.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Univariate and Bivariate Discrete Laplace and Generalized Discrete Laplace Distributions

Barry C. Arnold[1] and Matthew A. Arvanitis[2]

1. *University of California, Riverside*
2. *USDA Forest Products Laboratory. Email: mathew@umbc.edu*

Received: 15 Jan 2022 / Revised: 22 Apr 2022 / Accepted: 05 Jun 2022

Abstract

Univariate discrete Laplace models are described and investigated. A more general mixture model is represented. Bivariate extensions of these models are discussed in some detail, with particular emphasis on associated parameter estimation strategies. Multivariate versions of the models are briefly introduced.

Keywords: Discrete Laplace, Asymmetric Discrete Laplace, Generalized Asymmetric Discrete Laplace, Bivariate Discrete Laplace, likelihood, method of moments.
AMS Subject Classification: 62E10

1 Introduction to univariate discrete Laplace models

Inusah and Kozubowski [4] define the univariate asymmetric discrete Laplace distribution to be one with characteristic function of the form:

$$\phi_X(t) = p_1 p_2 (1-(1-p_1)e^{it})^{-1}(1-(1-p_2)e^{-it})^{-1}, \tag{1}$$

where $p_1, p_2 \in (0,1)$.

An alternative description of the model is available and will be used. For it we begin with two independent geometric random variables, V_1 and V_2 with $V_i \sim geo(p_i)$, $i = 1, 2$. We then define

$$X = V_1 - V_2. \tag{2}$$

If X has a representation of this form, we write $X \sim ADL(p_1, p_2)$. Note that in this paper, geometric random variables are defined to have support $\{0, 1, 2, ...\}$, and can be thought of as representing the number of failures preceding the first success in series of Bernoulli trials.

The possible values of X are $\{..., -3. - 2. - 1, 0, 1, 2, 3, ...\}$.

For $x \geq 0$ we have

$$\begin{aligned} P(X = x) &= \sum_{v_2=0}^{\infty} P(V_1 = x + v_2, V_2 = v_2) \\ &= \sum_{v_2=0}^{\infty} p_1 q_1^{x+v_2} p_2 q_2^{v_2} \\ &= p_1 p_2 q_1^x (1 - q_1 q_2)^{-1} \end{aligned}$$

© 2022 Author(s). (https://www.thegsa.in/).

Analogously, for $x < 0$,

$$\begin{aligned} P(X = x) &= \sum_{v_1=0}^{\infty} P(V_1 = v_1, V_2 = v_1 - x) \\ &= \sum_{v_2=0}^{\infty} p_1 q_1^{v_1} p_2 q_2^{v_1 - x} \\ &= p_1 p_2 q_2^{-x} (1 - q_1 q_2)^{-1}. \end{aligned}$$

It is then not difficult to confirm that $\sum_{-\infty}^{\infty} P(X = x) = 1$

This distribution will be called the asymmetric discrete Laplace distribution (ADL). It is a discrete parallel to the asymmetric Laplace model $Y = U_1 - U_2$ where the U_i's are independent with $U_i \sim exp(\lambda_i), \quad i = 1, 2$.

A mixture alternative is suggested by consideration of available mixture representations of continuous asymmetric Laplace models. It includes an additional parameter for flexibility. See [5], [7], and [6] for relevant discussion on analogous asymmetric Laplace models. We thus will consider

$$Y = IV_1 + (1 - I)(-V_2), \tag{3}$$

where $V_i \sim geo(p_i), \ i = 1, 2$, and I is an independent Bernoulli random variable with $P(I = 1) = \pi$.

It is readily verified that the characteristic function of such a generalized asymmetric discrete Laplace (GADL) variable of the form (3) is given by

$$\phi_Y(t) = \frac{1 - \pi(1 - p_1)e^{it} - (1 - \pi)(1 - p_2)e^{-it}}{1 - (1 - p_1)e^{it} - (1 - p_2)e^{-it} + (1 - p_1)(1 - p_2)}. \tag{4}$$

The density of this GADL distribution is of the following form.

$$\begin{aligned} P(Y = y) &= \pi p_1 q_1^y, \quad y = 1, 2, ..., \\ &= \pi p_1 + (1 - \pi) p_2, \quad y = 0, \\ &= (1 - \pi) p_2 q_2^{-y}, \quad = -1, -2, -3, ..., \end{aligned} \tag{5}$$

where p_1, p_2 and π are parameters ranging over the interval $(0, 1)$.

From the characteristic function, or from the mixture representation (3) we find

$$E(Y) = \frac{\pi(1 - p_1)}{p_1} - \frac{(1 - \pi)(1 - p_2)}{p_2}$$

and

$$var(Y) = \left\{ \frac{\pi(1 - p_1)(2 - p_1)}{p_1^2} + \frac{(1 - \pi)(1 - p_2)(2 - p_2)}{p_2^2} \right\} - \left\{ \frac{\pi(1 - p_1)}{p_1} - \frac{(1 - \pi)(1 - p_2)}{p_2} \right\}^2.$$

In the case in which $X \sim ADL(p_1, p_2)$ the moments simplify to become

$$E(X) = (1 - p_1)/p_1 - (1 - p_2)/p_2$$

and

$$var(X) = (1 - p_1)/p_1^2 + (1 - p_2)/p_2^2.$$

2 Estimation for the ADL distribution

Maximum likelihood

Suppose we have a sample $X_1, X_2, \ldots, X_n$ from an asymmetric discrete Laplace distribution defined in (2). We wish to estimate the parameters using maximum likelihood.

Define

$$U = \sum_1^n X_i I(X_i \geq 0) \text{ and } V = -\sum_1^n X_i I(X_i < 0).$$

Taking partial derivatives of the log-likelihood and equating to 0, yields the following likelihood equations

$$\frac{n}{p_1} = \frac{n(1-p_2)}{p_1+p_2-p_1p_2} + \frac{U}{1-p_1}, \tag{6}$$

$$\frac{n}{p_2} = \frac{n(1-p_1)}{p_1+p_2-p_1p_2} + \frac{V}{1-p_2}. \tag{7}$$

In order to solve these equations it is convenient to rewrite them in terms of the means of the V_i's, thus we define $\mu_i = (1-p_i)/p_i, \quad i = 1, 2$. The likelihood equations then become

$$\mu_1 = \frac{\mu_1\mu_2}{1+\mu_1+\mu_2} + \frac{U}{n}, \tag{8}$$

$$\mu_2 = \frac{\mu_1\mu_2}{1+\mu_1+\mu_2} + \frac{V}{n}. \tag{9}$$

From these equations we deduce that

$$\mu_1 - \mu_2 = \frac{U}{n} - \frac{V}{n}. \tag{10}$$

Next, using (10), express μ_2 as a linear function of μ_1 and substitute this in equation (8). This upon rearranging is a quadratic equation in μ_1 which can be solved to yield the maximum likelihood estimate of μ_1, i.e.,

$$\widehat{\mu}_1 = \frac{U}{n} - \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{UV}{n^2}}, \tag{11}$$

and then

$$\widehat{\mu}_2 = \widehat{\mu}_1 - \frac{U}{n} + \frac{V}{n.}. \tag{12}$$

Method of moments

Suppose we have a sample $X_1, X_2, \ldots, X_n$ from an asymmetric discrete Laplace distribution represented in the form

$$X = V_1 - V_2,$$

where the V_i's are i.i.d. with $V_i \sim geo(p_i), \quad i = 1, 2$.

We will equate the first two moments of X to the corresponding sample moments. Elementary computations yield

$$E(X) = \frac{1-p_1}{p_1} - \frac{1-p_2}{p_2}$$

and

$$E(X^2) = \frac{1-p_1}{p_1^2} + \frac{1-p_2}{p_2^2} + (E(X))^2$$

Denote the first two sample moments, based on a sample of size n, by $M_1 = (1/n)\sum_{i=1}^n X_i$ and $M_2 = (1/n)\sum_{i=1}^n X_i^2$, and define $S^2 = M_2 - M_1^2$ and set up the equations $M_1 = E(X)$ and $S^2 = var(X)$ which may be solved to obtain method of moments estimates of the p_i's. Here, as in the maximum likelihood case, it is convenient to rewrite the equations in terms of the means of the V_i's (i.e., $\mu_i = (1-p_i)/p_i, \quad i = 1, 2$). The moment equations are of the form

$$M_1 = \mu_1 - \mu_2, \tag{13}$$

$$S^2 = \mu_1(1+\mu_1) + \mu_2(1+\mu_2). \tag{14}$$

From equation (13) we can write $\mu_2 = \mu_1 - M_1$ and substitute this into (16). This then can be rearranged into a quadratic function of μ_1 which is readily solved. In this way we obtain method of moments estimates of the following form.

$$\widetilde{\mu}_1 = \frac{1}{2}\left[M_1 - 1 + \sqrt{1 - M_1^2 + 2M_1 S^2}\right], \tag{15}$$

and

$$\widetilde{\mu}_2 = \widetilde{\mu}_1 - M_1. \tag{16}$$

Bayesian Method

Suppose we have a sample $X_1, X_2, ..., X_n$ from an asymmetric discrete Laplace distribution with density

$$f_X(x; p_1, p_2) = \left(\frac{p_1 p_2}{p_1 + p_2 - p_1 p_2}\right)(1-p_1)^{xI(x\geq 0)}(1-p_2)^{-xI(x<0)}$$

Define
$U = \sum_1^n X_i I(X_i > 0)$, $V = -\sum_1^n X_i I(X_i < 0)$ and $W = \sum_1^n I(X_i > 0)$
We wish to estimate the parameters from a Bayesian viewpoint.
The likelihood for the sample is given by

$$L(p_1, p_2) = \left(\frac{p_1 p_2}{p_1 + p_2 - p_1 p_2}\right)^n (1-p_1)^u (1-p_2)^v.$$

If we take independent beta priors for p_1 and p_2, i.e.,

$$\widetilde{p}_i \sim Beta(\alpha_i, \beta_i), \quad i = 1, 2,$$

then the posterior will be of the form

$$f(p_1, p_2 | \underline{X} = \underline{x}) \propto \left(\frac{p_1^{n+\alpha_1} p_2^{n+\alpha_2}}{(p_1 + p_2 - p_1 p_2)^n}\right)(1-p_1)^{u+\beta_1}(1-p_2)^{v+\beta_2}.$$

From this joint posterior density the usual Bayes estimates of p_1 and p_2, namely $E(p_i | \underline{X} = \underline{x}), \quad i = 1, 2$, will be obtained by numerical integration.

3 Estimation for the GADL distribution

Maximum likelihood

Suppose instead we have a sample $X_1, X_2, ..., X_n$ from a generalized asymmetric discrete Laplace distribution of the form

$$X = IW_1 + (1 - I)(-W_2)$$

where $I \sim Bernoulli(\pi)$ and the W_i's are independent with $W_i \sim geo(p_i)$, $i = 1, 2$.

We wish to estimate the parameters using maximum likelihood. Here too, we will define

$$U = \sum_1^n X_i I(X_i \geq 0) \text{ and } V = -\sum_1^n X_i I(X_i < 0),$$

and also define $N_0 = \sum_1^n I(X_i = 0)$, $N_1 = \sum_1^n I(X_i > 0)$ and $N_2 = \sum_1^n I(X_i < 0)$

Taking partial derivatives of the log-likelihood and equating to 0, yields the following likelihood equations

$$\frac{N_1}{\pi} = \frac{N_2}{1-\pi} - \frac{N_0(p_1 - p_2)}{\pi p_1 + (1-\pi)p_2} \tag{17}$$

$$\frac{N_1}{p_1} = \frac{U}{1-p_1} - \frac{N_0 \pi}{\pi p_1 + (1-\pi)p_2} \tag{18}$$

$$\frac{N_2}{p_2} - \frac{V}{1 = p_2} - \frac{N_0(1-\pi)}{\pi p_1 + (1-\pi)p_2} \tag{19}$$

Note these equations are particularly easy to solve if $N_0 = 0$. In other cases an iterative solution may be obtained. Note that if p_1 and p_2 are known, then equation (17) is equivalent to a quadratic equation in π. If p_1 and π are known, then equation (18) is equivalent to a linear equation in p_2. And, finally, if p_2 and π are known, then equation (19) is equivalent to a linear equation in p_1.

Method of moments

Suppose we have a sample $X_1, X_2, ..., X_n$ from a generalized asymmetric discrete Laplace distribution of the form

$$X = IW_1 + (1 - I)(-W_2)$$

where $I \sim Bernoulli(\pi)$ and the W_i's are independent with $W_i \sim geo(p_i)$, $i = 1, 2$.

We wish to estimate the parameters using the method of moments.

The first three moments about zero of X are:

$$\begin{aligned}
E(X) &= \pi\frac{1-p_1}{p_1} - (1-\pi)\frac{1-p_2}{p_2} \\
E(X^2) &= \pi\left\{2\left(\frac{1-p_1}{p_1}\right)^2 + \frac{1-p_1}{p_1}\right\} + (1-\pi)\left\{2\left(\frac{1-p_2}{p_2}\right)^2 + \frac{1-p_2}{p_2}\right\} \\
E(X^3) &= \pi\left\{6\left(\frac{1-p_1}{p_1}\right)^3 + 6\left(\frac{1-p_1}{p_1}\right)^2 + \left(\frac{1-p_1}{p_1}\right)\right\} \\
&\quad -(1-\pi)\left\{6\left(\frac{1-p_2}{p_2}\right)^3 + 6\left(\frac{1-p_2}{p_2}\right)^2 + \left(\frac{1-p_2}{p_2}\right)\right\}
\end{aligned}$$

If we denote the first three sample moments, based on a sample of size n, by $M_1 = (1/n)\sum_{i=1}^n X_i$, $M_2 = (1/n)\sum_{i=1}^n X_i^2$, and $M_3 = (1/n)\sum_{i=1}^n X_i^3$, and set up the equations $M_j = E(X^j)$, $j = 1, 2, 3$, then our method of moments estimates will be the solution to these three equations.

Note that, if p_1 and p_2 are known then the equation $M_3 = E(X^3)$ is a linear equation in π. Also, if p_2 and π are known, then the equation $M_2 = E(X^2)$ is equivalent to a quadratic equation in p_1, and finally, if π and p_1 are known then the equation $M_1 = E(X)$ is equivalent to a linear equation in p_2. Using these observations, an iterative scheme for identifying the method of moments estimates is readily set up.

Bayesian Method

Suppose we have a sample $X_1, X_2, ..., X_n$ from a generalized asymmetric discrete Laplace distribution of the form

$$X = IW_1 + (1 - I)(-W_2)$$

where $I \sim Bernoulli(\pi)$ and the W_i's are independent with $W_i \sim geo(p_i)$, $i = 1, 2$.

Recall that we defined

$$U = \sum_1^n X_i I(X_i \geq 0) \text{ and } V = -\sum_1^n X_i I(X_i < 0),$$

and also defined $N_0 = \sum_1^n I(X_i = 0)$, $N_1 = \sum_1^n I(X_i > 0)$ and $N_2 = \sum_1^n I(X_i < 0)$

We wish to estimate the parameters from a Bayesian viewpoint.

In this case the likelihood will be

$$L(\pi, p_1, p_2) = \pi^{n_1} p_1^{n_1}(1 - p_1)^u(1 - \pi)^{n_2} p_2^{n_2}(1 - p_2)^v[\pi p_1 + (1 - \pi)p_2]^{n_0}$$

A plausible prior with independent marginals will be of the form

$$\widetilde{\pi} \sim Beta(\tau_1, \tau_2), \text{ and } \widetilde{p}_i \sim Beta(\alpha_i, \beta_i), \quad i = 1, 2.$$

The corresponding joint posterior density will be

$$f(\pi, p_1, p_2|\underline{X} = \underline{x}) \propto \pi^{n_1+\tau_1-1} p_1^{n_1+\alpha_1-1}(1 - p_1)^{u+\beta_1-1}(1 - \pi)^{n_2+\tau_2-1} p_2^{n_2+\alpha_2-1}(1 - p_2)^{v+\beta_2-1} \times[\pi p_1 + (1 - \pi)p_2]^{n_0}. \tag{20}$$

If $n_0 = 0$ then the posterior density will have independent beta distributed marginals. If $n_0 > 0$ then a posterior density that is a mixture of distributions with independent marginals will be encountered.

4 Bivariate models

In [1], two bivariate asymmetric Laplace models are described. The first bivariate asymmetric Laplace model was introduced by [3] and we refer the reader to that source for detailed discussion of the model. Construction of the model begins with the components used in developing the general bivariate beta model introduced in [2]. Thus we begin with 8 independent gamma variables $U_1, U_2, ..., U_8$ with $U_j \sim \Gamma(\delta_j, 1)$, $j = 1, 2, ..., 8$.. We then define (X, Y) by

$$\begin{aligned} X &= \lambda_{11}^{-1}(U_1 + U_5 + U_7) - \lambda_{12}^{-1}(U_3 + U_6 + U_8), \\ Y &= \lambda_{21}^{-1}(U_2 + U_6 + U_7) - \lambda_{22}^{-1}(U_4 + U_5 + U_8), \end{aligned} \tag{21}$$

where it is assumed that the constraints, $\delta_1 + \delta_5 + \delta_7 = 1$, $\delta_3 + \delta_6 + \delta_8 = 1$, $\delta_2 + \delta_6 + \delta_7 = 1$, and $\delta_4 + \delta_5 + \delta_8 = 1$, have been imposed to ensure that the distribution has asymmetric Laplace

marginals. This model will be called the bivariate asymmetric Laplace model of the first kind and if (X, Y) is as defined in (21) we will write $(X, Y) \sim BAL(1)(\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}, \underline{\delta})$. Since there were four constraints on the δ_j's, this is an 8 parameter model. The marginal distributions depend only on the four λ parameters, thus:

$$X \sim AL(\lambda_{11}, \lambda_{12}), \quad Y \sim AL(\lambda_{21}, \lambda_{22}). \tag{22}$$

The second bivariate asymmetric Laplace model utilizes the closure under minimization property of the exponential distribution. For it we again begin with 8 independent random variables, $V_1, V_2, ..., V_8$ but this time we assume that they are exponentially distributed, thus $V_j \sim exp(\lambda_j), \quad j = 1, 2, ..., 8$. We then define

$$\begin{aligned} X &= min\{V_1, V_5, V_7\} - min\{V_3, V_6, V_8\}, \\ Y &= min\{V_2, V_6, V_7\} - min\{V_4, V_5, V_8\}. \end{aligned} \tag{23}$$

If (X, Y) has the structure shown in (23) then we will write $(X, Y) \sim BAL(II)(\underline{\lambda})$ and say that it has a bivariate asymmetric Laplace distribution of the second kind with parameter vector $\underline{\lambda}$. Note that both the first kind and the second kind bivariate asymmetric Laplace distributions have an 8 dimensional parameter space. The marginal distributions of the BAL(II) distribution are by construction of the asymmetric Laplace form. Thus:

$$X \sim AL(\lambda_1 + \lambda_5 + \lambda_7, \lambda_3 + \lambda_6 + \lambda_8), \tag{24}$$

$$Y \sim AL(\lambda_2 + \lambda_6 + \lambda_7, \lambda_4 + \lambda_5 + \lambda_8), \tag{25}$$

Discrete versions of the BAL(I)-(II) distributions will now be constructed using negative binomial and geometric components instead of gamma and exponential distributed components. But, before discussing such distributions, we will consider bivariate versions of the generalized asymmetric Laplace (GAL) distribution. This can be achieved by modifying the BAL(I) or the BAL(II) models by the introduction of two additional probability parameters.

The generalized version of the BAL(I) model may be defined as follows

$$\begin{aligned} X &= I_1 \lambda_{11}^{-1}(U_1 + U_5 + U_7) - (1 - I_1)\lambda_{12}^{-1}(U_3 + U_6 + U_8), \\ Y &= I_2 \lambda_{21}^{-1}(U_2 + U_6 + U_7) - (1 - I_2)\lambda_{22}^{-1}(U_4 + U_5 + U_8), \end{aligned} \tag{26}$$

where it is assumed that the constraints, $\delta_1 + \delta_5 + \delta_7 = 1$, $\delta_3 + \delta_6 + \delta_8 = 1$, $\delta_2 + \delta_6 + \delta_7 = 1$, and $\delta_4 + \delta_5 + \delta_8 = 1$, have been imposed, and where the I_j's are independent with $I_j \sim Bernoulli(p_j), \quad j = 1, 2$.

Means variances and covariance of the coordinates of this random vector are not difficult to evaluate, or could be evaluated by simulation.

The generalized version of the BAL(II) model may be defined as follows

$$\begin{aligned} X &= I_1[min\{(V_1, V_5, V_7\}] - (1 - I_1)[min\{V_3, V_6, V_8\}], \\ Y &= I_2[min\{(V_2, V_6, V_7\}] - (1 - I_2)[min\{V_4, V_5, V_8)\}], \end{aligned} \tag{27}$$

where the δ_j's are positive parameters, and where the I_j's are independent with $I_j \sim Bernoulli(p_j)$, $j = 1, 2$.

Means variances and covariance of the coordinates of this random vector are also not difficult to evaluate, or could be evaluated by simulation.

We now will define analogous discrete versions of these bivariate models. To construct the bivariate symmetric discrete Laplace model of the first kind (BDL(I)) we begin with a set of 8 independent random variables $U_1, U_2, ..., U_8$ with $U_i \sim neg.bin.(\delta_j, p)$, $j = 1, 2, ..., 8$. Note that all of the U_j's share a common value for p. This results in a construction of a bivariate distribution with symmetric discrete Laplace marginals. It will be seen that it is not possible to use this kind of construction to yield asymmetric marginals. To continue, we now define (X, Y) by

$$\begin{aligned} X &= (U_1 + U_5 + U_7) - (U_3 + U_6 + U_8), \\ Y &= (U_2 + U_6 + U_7) - (U_4 + U_5 + U_8), \end{aligned} \tag{28}$$

where it is assumed that the constraints, $\delta_1 + \delta_5 + \delta_7 = 1$, $\delta_3 + \delta_6 + \delta_8 = 1$, $\delta_2 + \delta_6 + \delta_7 = 1$, and $\delta_4 + \delta_5 + \delta_8 = 1$, have been imposed. This model will be called the bivariate discrete Laplace model of the first kind and if (X, Y) is as defined in (28) we will write $(X, Y) \sim BDL(1)(\underline{\delta})$. Since there were four constraints on the δ_j's, this is a 5 parameter model. The marginal distributions are differences of independent $geometric(p)$ variables and thus have discrete Laplace densities. Moments are obtainable from the representation (28).

The second bivariate asymmetric discrete Laplace model that we will consider will utilize the closure under minimization property of the geometric distribution. For it we again begin with 8 independent random variables, $V_1, V_2, ..., V_8$ but this time we assume that they are geometrically distributed, thus $V_j \sim geo(\tau_j)$, $j = 1, 2, ..., 8$, where $\tau_j \in (0, 1)$, $j = 1, 2, ..., 8$, We then define

$$\begin{aligned} X &= min\{V_1, V_5, V_7\} - min\{V_3, V_6, V_8\}, \\ Y &= min\{V_2, V_6, V_7\} - min\{V_4, V_5, V_8\}, \end{aligned} \tag{29}$$

using a construction parallel to that used in the construction of the BAL(II) model earlier described in this paper. If (X, Y) has the structure shown in (29) then we will write $(X, Y) \sim BADL(II)(\underline{\tau})$ and say that it has a bivariate asymmetric discrete Laplace distribution of the second kind with parameter vector $\underline{\tau}$. Note that the second kind bivariate asymmetric discrete Laplace distributions has an 8 dimensional parameter space. The marginal distributions of the BADL(II) distribution are of the asymmetric discrete Laplace form. Thus:

$$X \sim ADL(1 - (1 - \tau_1)(1 - \tau_5)(1 - \tau_7), 1 - (1 - \tau_3)(1 - \tau_6)(1 - \tau_8)), \tag{30}$$

$$Y \sim ADL(1 - (1 - \tau_2)(1 - \tau_6)(1 - \tau_7), 1 - (1 - \tau_4)(1 - \tau_5)(1 - \tau_8)), \tag{31}$$

The marginal moments of the BADL(II) distribution are thus readily identified. However $cov(X, Y)$ is quite complicated and will usually be approximated by simulation. using the definition (29).

Generalized versions of these bivariate discrete Laplace models can be formulated in a manner parallel to that used to generalize the bivariate asymmetric Laplace models.

The generalized version of the BDL(I) model may be defined as follows

$$X = I_1(U_1 + U_5 + U_7) - (1 - I_1)(U_3 + U_6 + U_8),$$
$$Y = I_2(U_2 + U_6 + U_7) - (1 - I_2)(U_4 + U_5 + U_8), \tag{32}$$

where it is assumed that the constraints, $\delta_1 + \delta_5 + \delta_7 = 1$, $\delta_3 + \delta_6 + \delta_8 = 1$, $\delta_2 + \delta_6 + \delta_7 = 1$, and $\delta_4 + \delta_5 + \delta_8 = 1$, have been imposed, and where the I_j's are independent with $I_j \sim Bernoulli(\pi_j)$, $j = 1, 2$, and the U_j's are independent with $U_j \sim neg.bin.(\delta_j, p)$, $j = 1, 2, 3..., 8$. This model will be called the generalized bivariate discrete Laplace model of the first kind and if (X, Y) is as defined in (32) we will write $(X, Y) \sim GBDL(1)(\pi_1, \pi_2, \underline{\delta})$. Since there were four constraints on the δ_j's, this is a 7 parameter model.

Means variances and covariance of the coordinates of this random vector are not difficult to evaluate, or could be evaluated by simulation.

The generalized version of the BADL(II) model may be defined as follows

$$X = I_1[min\{(V_1, V_5, V_7\}] - (1 - I_1)[min\{V_3, V_6, V_8\}],$$
$$Y = I_2[min\{V_2, V_6, V_7\}] - (1 - I_2)[min\{V_4, V_5, V_8)\}], \tag{33}$$

where the I_j's are independent with $I_j \sim Bernoulli(\pi_j)$, $j = 1, 2.$ and the V_j's are independent with $V_j \sim geo(\tau_j)$, $j = 1, 2, ..., 8$. This is a 10 parameter model.

Means variances and covariance of the coordinates of this random vector are also not difficult to evaluate, or could be evaluated by simulation.

Remark Even more general distribution than (32) can be constructed in which the I_i's are dependent indicators.

For this more general model we begin with a random vector (J_1, J_2, J_3, J_4) with 4 possible values $(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0)$ and $(0, 0, 0, 1)$ with associated probabilities π_1, π_2, π_3 and π_4, and then define $I_1 = max\{J_1, J_3\}$ and $I_2 = max\{J_2, J_3\}$.

The random variables (X, Y) are then defined as in (32) using the dependent I_i's just defined.

The same modification can be made to generalize (33).

Parameter estimation

None of the full models described in this paper are expected to be useful for practical purposes. Instead the authors expect that they will be used as a source for smaller, more manageable, submodels. For all four of the bivariate models discussed, only a small collection of specific cases may be described which have explicit discrete densities, and even these densities are rather complex. Therefore, unconventional methods for parameter estimation may be called for.

For illustrative purposes, we will consider parameter estimation for some submodels with significantly reduced parameter spaces. In describing submodels we adhere to the following conventions : (1) If $U \sim neg.bin(\delta, p)$ with $\delta = 0$ then $U = 0$ with probability 1. Consequently, such U's can be deleted from the description of the bivariate models BDL(I) and GBDL(I) as described in this paper.

(2) If $V \sim geo(\tau)$ with $\tau = 0$ then $V = \infty$ with probability 1. Consequently such V's can be deleted from the description of the bivariate models BADL(II) and GBADL(II) as described in this paper.

Some two-parameter submodels

In this subsection, we will consider parameter estimation for a pair of simple submodels of the $BDL(I)$ and $BADL(II)$ models.

Example 1. *To begin, consider the sub-model of the $BDL(I)$ model given by restricting its parameter space as follows:*

$$\begin{aligned} &p \in (0,1) \\ &\delta_1 = \delta_4 = \alpha \in (0,1) \\ &\delta_2 = \delta_3 = 1 \\ &\delta_5 = 1-\alpha \\ &\delta_6 = \delta_7 = \delta_8 = 0. \end{aligned} \tag{34}$$

We will call this model M1. We are adopting the convention that if $U \sim neg.bin.(\delta, p)$ with $\delta = 0$ then $U = 0$. The full parameter vector for this two parameter model is thus

$$(p, \underline{\delta}) = (p, (\alpha, 1, 1, \alpha, 1-\alpha, 0, 0, 0)).$$

Suppose that we have a sample of size n, i.e., $\{(X_j, Y_j) : j = 1, 2, ..., n\}$ and we wish to estimate the parameters p and α. Upon writing this model in terms of the U_i variables, thus

$$\begin{aligned} X &= (U_1 + U_5) - (U_3), \\ Y &= (U_2) - (U_4 + U_5), \end{aligned} \tag{35}$$

it may be verified that $E(X) = E(Y) = 0$, $var(X) = var(Y) = 2(1-p)/p^2$ and $cov(X,Y) = -(1-\alpha)(1-p)/p^2$. Consequently, if we define

$$T_1 = (1/n)[\sum_{i=1}^{n}[(X_i - \overline{X})^2 + (Y_i - \overline{Y})^2]$$

and

$$T_2 = (1/n)\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}),$$

we can set up the following moment equations

$$T_1 = E(T_1) = 4(1-p)/p^2, \tag{36}$$

$$T_2 = E(T_2) = -(1-\alpha)(1-p)/p^2. \tag{37}$$

These are readily solved to yield the following consistent moment estimates of the parameters of the model.

$$\widetilde{p} = [1 + \sqrt{1 + T_1}]/2$$

and

$$\widetilde{\alpha} = 1 + (4T_2/T_1).$$

Example 2. *This time we will consider the two-parameter sub-model of the $BADL(II)$ model given by restricting its parameter space as follows:*

$$\begin{aligned} \tau_2 = \tau_3 = \alpha \in (0,1) \\ \tau_1 = \tau_4 = \tau_5 = \beta \in (0,1) \\ \tau_6 = \tau_7 = \tau_8 = 0. \end{aligned} \tag{38}$$

We will call this model M2. We are adopting the convention that if $V \sim geo(\tau)$ with $\tau = 0$ then $V = \infty$. The full parameter vector for this two parameter model is thus

$$\underline{\tau} = (\beta, \alpha, \alpha, \beta, \beta, 0, 0, 0),$$

and the model, in terms of the V_i variables is given by

$$\begin{aligned} X = min\{V_1, V_5\} - V_3, \\ Y = V_2 - min\{V_4, V_5\}. \end{aligned} \tag{39}$$

Elementary computations yield

$$E(X) = -E(Y) = \frac{(1-\beta)^2}{1-(1-\beta)^2} - \frac{1-\alpha}{\alpha}$$

and

$$var(X) = var(Y) = \frac{(1-\beta)^2}{[1-(1-\beta)^2]^2} + \frac{1-\alpha}{\alpha^2}.$$

Let us define

$$T_1 = (1/2n)[\sum_{i=1}^{n}[X_i - Y_i]$$

and

$$T_2 = (1/2n)[\sum_{i=1}^{n}[(X_i - \overline{X})^2 + (Y_i - \overline{Y})^2].$$

We can then set up the following moment equations to be solved for (α, β).

$$T_1 = E(T_1) = \frac{(1-\beta)^2}{1-(1-\beta)^2} - \frac{1-\alpha}{\alpha}, \tag{40}$$

$$\tag{41}$$

$$T_2 = E(T_2) = \frac{(1-\beta)^2}{[1-(1-\beta)^2]^2} + \frac{1-\alpha}{\alpha^2}. \tag{42}$$

These can be solved by iterative substitution. Thus choose an initial value for α , perhaps $\alpha = 0.5$, and then with this value of alpha solve for β in equation (40). Then substitute this value of β in equation (42) and solve for α, then use equation (40) once more, etc.

Multi-parameter submodels

As the two-parameter models discussed in Section 4 would readily suggest, freeing more of the δ's in $BDL(I)$ or τ's in $BADL(II)$ can lead to more complex models with multiple tail dependencies. Due to the complexity of the models, more creative computer intensive approaches, will need to be applied for parameter estimation.

Remark Higher dimensional versions of the BDL(I), BADL(II), GBDL(I), GBADL(II) models are readily described. Since, for example, the completely general trivariate version of the asymmetric discrete Laplace model of Type II will involve 26 independent geometric components each with its own τ parameter, only submodels including just a limited number of components will be tractable and useful for modeling purposes. One very simple trivariate submodel model that might find application is the following.

$$\begin{aligned} X &= \min\{V_1, V_7\} - V_{19} \\ Y &= \min\{V_2, v_7\} - V_{19} \\ Z &= V_3 - V_{19} \end{aligned}$$

which involves only 5 of the 26 components in the general model, but still has a non-trivial dependence structure..

5 Conclusion

In this paper, two discrete Laplace models with new methods of construction are detailed. Standard parameter estimation techniques are exhibited for both. Further, new bivariate variants of this family of distributions are discussed, and parameter estimation techniques are outlined. We finished with a brief discussion of higher-dimensional models.

References

[1] Barry C Arnold. "Some bivariate and multivariate models involving independent gamma distributed components." In: *Methodology and Applications of Statistics*. Springer, 2021, pp. 115–128.

[2] Barry C Arnold and Hon Keung Tony Ng. "Flexible bivariate beta distributions." In: *Journal of Multivariate Analysis* 102.8 (2011), pp. 1194–1202.

[3] Matthew Alexander Arvanitis. *Likelihood-free estimation for some flexible families of distributions*. University of California, Riverside, 2018.

[4] Seidu Inusah and Tomasz J Kozubowski. "A discrete analogue of the Laplace distribution." In: *Journal of statistical planning and inference* 136.3 (2006), pp. 1090–1102.

[5] MC Jones. "On a characteristic property of distributions related to the Laplace." In: *South African statistical journal* 53.1 (2019), pp. 31–34.

[6] Tomasz J Kozubowski and Krzysztof Podgórski. "A multivariate and asymmetric generalization of Laplace distribution." In: *Computational Statistics* 15.4 (2000), pp. 531–540.

[7] Tomasz J Kozubowski and Krzysztof Podgórski. "Asymmetric Laplace distributions." In: *Mathematical Scientist* 25.1 (2000), pp. 37–46.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

On Zero-inflated Generalized Hyper Poisson Distribution and its Properties

C. Satheesh Kumar[1] and Rakhi Ramachandran[2]

1. *Department of Statistics, University of Kerala, Thiruvananthapuram - 695 581, India, email: drcsatheeshkumar@gmail.com*
2. *Department of Statistics, University of Kerala, Thiruvananthapuram - 695 581, India, email: rakhir3@gmail.com*

Received: 02 Apr 2022 / Revised: 23 May 2022 / Accepted: 10 Jun 2022

Abstract

A zero-inflated version of the generalized hyper-Poisson distribution is introduced and study some of its important statistical properties such as mean, variance, recursion relations for probabilities, raw moments and factorial moments. The estimation of the parameters of this distribution is considered and the distribution has been fitted to a well-known data set. Further a generalized likelihood ratio test procedure is applied for testing the significance of the inflation parameter.

Keywords: Confluent hypergeometric series, Count data modeling, Generalized likelihood ratio test, Model selection, Zero-inflated Hermite distribution.

1 Introduction

There are many situations in econometric, medical, engineering, manufacturing, public health, road safety, epidemiology, etc. where zero-inflation can be observed. To model zero-inflated count data, several zero-inflated models have been studied in the literature, among them zero-inflated Poisson distribution (ZIPD) of [6], zero-inflated Hermite distribution (ZIHD) of [4], zero-inflated hyper-Poisson distribution (ZIHPD) of [5] and zero-inflated modified hyper-Poisson distribution (ZIMHPD) of [3] are of special interest. The probability mass function (pmf) of the ZIMHPD is given by

$$f(x) = \begin{cases} \omega + (1-\omega)\frac{1}{\phi(1;\lambda;\alpha+\beta)}, & x = 0 \\ (1-\omega)\frac{1}{\phi(1;\lambda;\alpha+\beta)} \sum_{k=0}^{[\frac{x}{2}]} \frac{(x-k)!\alpha^{x-2k}\beta^k}{(\lambda)_{x-k}(x-2k)!k!}, & x = 1,2,\ldots \end{cases} \tag{1}$$

where $\omega \in [0,1)$, $\lambda > 0$, $\alpha > 0$, $\beta \geq 0$ and $\phi(a;b;\theta) = \sum_{r=0}^{\infty} \frac{(a)_r \theta^r}{(b)_r r!}$ is the confluent hypergeometric function with $(a)_r = a(a+1)(a+2)\ldots(a+r-1) = \frac{\Gamma(a+r)}{\Gamma(a)}$, for $r = 1,2,\ldots$ and $(a)_0 = 1$. For more details on confluent hypergeometric series, see [7] or [8]. Clearly when $\lambda = 1$, the ZIMHPD reduces to the ZIHD, when $\beta = 0$, the ZIMHPD reduces to the ZIHPD and when $\lambda = 1$ and $\beta = 0$, the ZIMHPD reduces to the ZIPD.

Through this paper we develop further a modified version of the ZIMHPD which we call the "zero-inflated generalized hyper-Poisson (ZIGHPD) distribution" and discuss some of its important statistical properties. In section 2, we present the definition of the ZIGHPD and obtain its probability generating function, expressions for its mean and variance, and recursion formulae for probabilities, raw moments and factorial moments. Further, the estimation of the parameters of the model is discussed

© 2022 Author(s). (https://www.thegsa.in/).

in section 3 and a test procedure is constructed in section 4. In section 5 both the procedures discussed in sections 3 and 4 are illustrated with its relevence with the help of a real life data set.

We need the following series representations in the sequel.

$$\sum_{x=0}^{\infty}\sum_{r=0}^{\infty} A(r,x) = \sum_{x=0}^{\infty}\sum_{r=0}^{x} A(r,x-r) \tag{2}$$

$$\sum_{x=0}^{\infty}\sum_{r=0}^{\infty} A(r,x) = \sum_{x=0}^{\infty}\sum_{r=0}^{[\frac{x}{m}]} A(r,x-rm). \tag{3}$$

2 Definition and Properties

We present the definition of the ZIGHP distribution and discuss some of its properties.

Definition 1. *A discrete random variable M is said to follow the "zero-inflated generalized hyper-Poisson distribution or in short ZIGHPD" with parameters ω, λ, α, β and γ if its p.m.f is*

$$\begin{aligned} g(m) &= P(M=m) \\ &= \begin{cases} \omega + (1-\omega)\frac{1}{\phi(1;\lambda;\alpha+\beta+\gamma)}, & m=0 \\ (1-\omega)\frac{1}{\phi(1;\lambda;\alpha+\beta+\gamma)}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-k-2j}}{(\lambda)_{m-k-2j}}\frac{\alpha^{m-2k-3j}}{(m-2k-3j)!}\frac{\beta^k}{k!}\frac{\gamma^j}{j!}, & m=1,2,\ldots \\ 0, otherwise \end{cases} \end{aligned} \tag{4}$$

in which $\omega \in [0,1)$, $\lambda > 0$, $\alpha > 0$, $\beta \geq 0$ and $\gamma \geq 0$.

Important special cases of the ZIGHPD includes the following cases.

1. when $\gamma = 0$, the ZIGHPD reduces to the zero-inflated modified hyper-Poisson distribution (ZIMHPD) distribution of [3].
2. when $\beta = \gamma = 0$ the ZIGHPD reduces to the ZIHPD of [5] with p.m.f. (1).
3. when $\gamma = 0$, $\lambda = 1$ the ZIGHPD distribution reduces to the zero-inflated Hermite (ZIH) distribution of [4].
4. when $\beta = \gamma = 0$, $\lambda = 1$ the ZIGHPD distribution reduces to the zero-inflated Poisson (ZIP) distribution of [6].
5. when $\omega = 0$, the distribution reduces to the generalized hyper-Poisson distribution (GHPD) distribution of [2].

Now we obtain the following results.

Result 2.1. *The probability generating function (p.g.f) $G(t)$ of the ZIGHPD with p.m.f (4) is the following.*

$$G(t) = \omega + (1-\omega)\,\frac{\phi(1;\lambda;\alpha t+\beta t^2+\gamma t^3)}{\phi(1;\lambda;\alpha+\beta+\gamma)}. \tag{5}$$

Proof. By definition, the p.g.f of the ZIGHPD having p.m.f (4) is given by

$$\begin{aligned}
G(t) &= \sum_{m=0}^{\infty} f(m)t^m \\
&= \left\{\omega + \frac{(1-\omega)}{\phi[1;\gamma;(\alpha+\beta+\gamma)]}\right\} + \frac{(1-\omega)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \sum_{m=1}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]} t^m \frac{(1)_{m-k-2j}}{(\lambda)_{m-k-2j}} \\
&\times \frac{\alpha^{m-2k-3j}}{(m-2k-3j)!}\frac{\beta^k}{k!}\frac{\gamma^j}{j!} \\
&= \omega + \frac{(1-\omega)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \sum_{m=0}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]} \frac{(1)_{m-k-2j}}{(\lambda)_{m-k-2j}} \frac{\alpha^{m-2k-3j}}{(m-2k-3j)!}\frac{\beta^k}{k!}\frac{\gamma^j}{j!} t^m \\
&= \omega + \frac{(1-\omega)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \sum_{m=0}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]} \frac{1}{(\lambda)_{m-k-2j}} \binom{m-k-2j}{m-3j-k} \\
&\times \binom{m-3j-k}{k} \alpha^{m-2k-3j}\beta^k\gamma^j t^m. \qquad (6)
\end{aligned}$$

Applying the series expansion (3), we get (6) as

$$\begin{aligned}
G(t) &= \omega + \frac{(1-\omega)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \sum_{m=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{[\frac{m}{2}]} \frac{1}{(\lambda)_{m+j-k}} \binom{j+m-k}{m-k}\binom{m-k}{k} \\
&\times \alpha^{m-2k}\beta^k\gamma^j t^{m+3j} \\
&= \omega + \frac{(1-\omega)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \sum_{m=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \frac{1}{(\lambda)_{m+j+k}} \binom{m+j+k}{m+k}\binom{m+k}{k} \\
&\times \alpha^{m}\beta^k\gamma^j t^{m+3j+2k}. \qquad (7)
\end{aligned}$$

Applying the series expansion (2) in the above, we get (7) as

$$\begin{aligned}
G(t) &= \omega + \frac{(1-\omega)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \sum_{j=0}^{\infty}\sum_{m=0}^{\infty}\sum_{k=0}^{m} \frac{1}{(\lambda)_{j+m}} \binom{j+k}{m}\binom{m}{k} \alpha^{m-k}\beta^k\gamma^j t^{m+3j+k} \qquad (8) \\
&= \omega + \frac{(1-\omega)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \sum_{j=0}^{\infty}\sum_{m=0}^{j}\sum_{k=0}^{m} \frac{1}{(\lambda)_{j}} \binom{j}{m}\binom{m}{k} \alpha^{m-k}\beta^k\gamma^{j-m} t^{3j-2m+k} \\
&= \omega + \frac{(1-\omega)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \sum_{j=0}^{\infty}\sum_{m=0}^{j} \frac{1}{(\lambda)_{j}} \binom{j}{m} \sum_{k=0}^{m}\binom{m}{k} (\alpha t)^{m-k}(\beta t^2)^k(\gamma t^3)^{j-m} \\
&= \omega + \frac{(1-\omega)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \sum_{j=0}^{\infty}\sum_{m=0}^{j} \frac{1}{(\lambda)_{j}} \binom{j}{m} (\alpha t+\beta t^2)^j (\gamma t^3)^{j-m}
\end{aligned}$$

$$\begin{aligned} G(t) &= \omega + \frac{(1-\omega)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \sum_{j=0}^{\infty} \frac{1}{(\lambda)_j} \sum_{m=0}^{j} \binom{j}{m} (\alpha t + \beta t^2)^j (\gamma t^3)^{j-m} \\ &= \omega + \frac{(1-\omega)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \sum_{j=0}^{\infty} \frac{1}{(\lambda)_j} (\alpha t + \beta t^2 + \gamma t^3)^j , \end{aligned}$$

which on simplification gives (5).

Result 2.2. *The mean and variance of the ZIGHPD with p.g.f* (5) *are*

$$Mean = \frac{(1-\omega)}{\lambda} (\alpha + 2\beta + 3\gamma) \Lambda_1$$

and

$$\begin{aligned} Variance &= \left\{ \left(\frac{2}{\lambda+1} \Lambda_2 - \frac{(1-\omega)}{\lambda} \Lambda_1^2 \right) (\alpha + 2\beta + 3\gamma)^2 + \Lambda_1 (\alpha + 4\beta + 9\gamma) \right\} \\ &\times \frac{(1-\omega)}{\lambda}, \end{aligned}$$

where $\Lambda_j = \frac{d_0}{d_j}$ *for* $j = 1, 2$, *in which* $d_0 = \phi^{-1}(1;\lambda;\alpha+\beta+\gamma)$ *and* $d_j = \phi^{-1}(1+j;\lambda+j;\alpha+\beta+\gamma)$.

Result 2.3. *For* $m \geq 0$ *a simple recursion formula for probabilities* $g(m) = g_m(\lambda^*)$ *of the ZIGHPD is the following, in which* $\lambda^* + j = (1+j, \lambda+j)$ *and* $\lambda^{(j)} = \frac{1+j}{\lambda+j}$.

$$g_1(\lambda^*) = \lambda^{(0)} d_1 \left(\alpha g_0(\lambda^* + 1) - \omega \right), \quad for\ m = 0 \tag{9}$$

$$g_2(\lambda^*) = \frac{\lambda^{(0)}}{2} d_1 \left(\alpha g_1(\lambda^* + 1) + 2\beta (g_0(\lambda^* + 1) - \omega) \right), \quad for\ m = 1 \tag{10}$$

$$g_3(\lambda^*) = \frac{\lambda^{(0)}}{3} d_1 \left(\alpha g_2(\lambda^* + 1) + 2\beta g_1(\lambda^* + 1) + 3\gamma (g_0(\lambda^* + 1) - \omega) \right), \quad for\ m = 2 \tag{11}$$

and

$$g_{m+1}(\lambda^*) = \frac{\lambda^{(0)}}{m+1} d_1 \left(\alpha g_m(\lambda^* + 1) + 2\beta g_{m-1}(\lambda^* + 1) + 3\gamma g_{m-2}(\lambda^* + 1) \right), \quad for\ m > 2. \tag{12}$$

Proof. The p.g.f of the ZIGHPD can be written as

$$\begin{aligned} G(t) &= \omega + (1-\omega) \frac{\phi(1;\lambda;\alpha t + \beta t^2 + \gamma t^3)}{\phi(1;\lambda;\alpha+\beta+\gamma)} \\ &= \sum_{m=0}^{\infty} t^m g_m(\lambda^*). \end{aligned} \tag{13}$$

On differentiating (13) with respect to t, we obtain the following.

$$\begin{aligned} \sum_{m=0}^{\infty} (m+1) g_{m+1}(\lambda^*) t^m &= (1-\omega) d_0 \lambda^{(0)} \left(\alpha + 2\beta t + 3\gamma t^2 \right) \\ &\times \phi(2;\lambda^* + 1; \alpha t + \beta t^2 + \gamma t^3). \end{aligned} \tag{14}$$

Also, from (13), we have

$$\sum_{m=0}^{\infty} g_m(\lambda^*+1)t^m = \omega + (1-\omega)d_0^*\lambda^{(0)}\phi(2;\lambda^*+1;\alpha t+\beta t^2+\gamma t^3). \tag{15}$$

Combining (14) and (15), we get

$$\sum_{m=0}^{\infty}(m+1)g_{m+1}(\lambda^*)t^m = \left\{\alpha\left(\sum_{m=0}^{\infty} g_m(\lambda^*+1)t^m - \omega\right) + 2\beta t \right.$$
$$\left.\left(\sum_{m=0}^{\infty} g_{m-1}(\lambda^*+1)t^m - \omega\right) + 3\gamma t^2\left(\sum_{m=0}^{\infty} g_{m-2}(\lambda^*+1)t^m - \omega\right)\right\} d_1\lambda^{(0)}. \tag{16}$$

Now, on equating the coefficients of t^0 on both sides of (16), we get (9), on equating the coefficients of t^1 on both sides of (16), we get (10), on equating the coefficients of t^2 on both sides of (16), we get (11) and on equating the coefficients of t^m for $m>2$ on both sides of (16), we get (12).

Result 2.4. *For $r \geq 0$, the recursion formula for raw moments $\mu_r(\lambda^*)$ of the ZIGHPD is*

$$\mu_{r+1}(\lambda^*) = \lambda^{(0)}d_1\left\{\sum_{k=0}^{r}\binom{r}{k}\mu_{r-k}(\lambda^*+1)(\alpha+2^{k+1}\beta+3^{k+1}\gamma) - \omega(\alpha+2^{k+1}\beta+3^{k+1}\gamma)\right\}. \tag{17}$$

Proof. For any $t \in \Re = (-\infty,\infty)$ and $i=\sqrt{-1}$, the characteristic function of the ZIGHPD is

$$\begin{aligned} H(t) &= G(e^{it}) \\ &= \omega + (1-\omega)\,\frac{\phi(1;\lambda;\alpha e^{it}+\beta e^{2it}+\gamma e^{3it})}{\phi(1;\lambda;\alpha+\beta+\gamma)} \\ &= \sum_{r=0}^{\infty}\mu_r(\lambda^{(*)})\frac{(it)^r}{r!}. \end{aligned} \tag{18}$$

Differentiating (18) with respect to t, we get

$$\begin{aligned} \sum_{r=0}^{\infty}\mu_{r+1}(\lambda^*)\frac{(it)^r}{r!} &= (1-\omega)d_0\lambda^{(0)}\left(\alpha e^{it}+2\beta e^{2it}+3\gamma e^{3it}\right) \\ &\times \phi(2;\lambda^*+1;\alpha e^{it}+\beta e^{2it}+\gamma e^{3it}). \end{aligned} \tag{19}$$

Also from (18), we have

$$\sum_{r=0}^{\infty}\mu_r(\lambda^*+1)\frac{(it)^r}{r!} = \omega + (1-\omega)d_0^*\phi(2;\lambda^*+1;\alpha e^{it}+\beta e^{2it}+\gamma e^{3it}). \tag{20}$$

Combining (19) and (20), we obtain

$$\begin{aligned} \sum_{r=0}^{\infty}\mu_{r+1}(\lambda^*)\frac{(it)^r}{r!} &= \lambda^{(0)}d_1\alpha\left(\sum_{r=0}^{\infty}\sum_{k=0}^{\infty}\mu_r(\lambda^*+1)\frac{(it)^k}{k!} - \omega\sum_{k=0}^{\infty}\frac{(it)^k}{k!}\right) + 2\lambda^{(0)}d_1\beta \\ &\times \left(\sum_{r=0}^{\infty}\sum_{k=0}^{\infty}\mu_r(\lambda^*+1)\frac{(2it)^k}{k!} - \omega\sum_{k=0}^{\infty}\frac{(2it)^k}{k!}\right) + 3\lambda^{(0)}d_1\gamma \\ &\times \left(\sum_{r=0}^{\infty}\sum_{k=0}^{\infty}\mu_r(\lambda^*+1)\frac{(3it)^k}{k!} - \omega\sum_{k=0}^{\infty}\frac{(3it)^k}{k!}\right), \end{aligned} \tag{21}$$

in the light of (2). On equating the coefficients of $\frac{(it)^r}{r!}$ in (21), we get (17).

Result 2.5. *For $r \geq 0$, the recursion formula for factorial moments $\mu_{[r]}(\lambda^*)$ of the ZIGHPD is*

$$\begin{aligned}\mu_{[r+1]}(\lambda^*) &= \lambda^{(0)} d_1 \left\{ \alpha \left(\mu_{[r]}(\lambda^*+1) - \omega \right) + 2\beta \left(\sum_{r=0}^{\infty} \mu_{[r-1]}(\lambda^*+1) - \omega \right) \right\} \\ &+ \left\{ \lambda^{(0)} d_1 3\gamma \left(\sum_{r=0}^{\infty} \mu_{[r-2]}(\lambda^*+1) - \omega \right) \right\}. \end{aligned} \tag{22}$$

Proof. The factorial moment generating function $F(t)$ of the ZIGHPD with p.g.f (5) is the following.

$$\begin{aligned} F(t) &= G(1+t) \\ &= \omega + (1-\omega) \frac{\phi[1;\lambda;\alpha(1+t)+\beta(1+t)^2+\gamma(1+t)^3]}{\phi(1;\lambda;\alpha+\beta+\gamma)} \\ &= \sum_{r=0}^{\infty} \mu_{[r]}(\lambda^*) \frac{t^r}{r!} \end{aligned} \tag{23}$$

Differentiating (23) with respect to t, we get

$$\begin{aligned} \sum_{r=0}^{\infty} \mu_{[r+1]}(\lambda^*) \frac{t^r}{r!} &= (1-\omega) d_0 \lambda^{(0)} \phi[2;\lambda^*;\alpha(1+t)+\beta(1+t)^2+\gamma(1+t)^3] \\ &\times \left(\alpha + 2\beta(1+t) + 3\gamma(1+t)^2 \right). \end{aligned} \tag{24}$$

Also, from (23) we have

$$\sum_{r=0}^{\infty} \mu_{[r]}(\lambda^*+1) \frac{t^r}{r!} = \omega + (1-\omega) d_0^* \, \phi[2;\lambda^*+1;\alpha(1+t)+\beta(1+t)^2+\gamma(1+t)^3]. \tag{25}$$

Now, combining (24) and (25) leads to

$$\sum_{r=0}^{\infty} \mu_{[r+1]}(\lambda^*) \frac{t^r}{r!} = d_1 \lambda^{(0)} \left(\sum_{r=0}^{\infty} \mu_{[r]}(\lambda^*+1) \frac{t^r}{r!} - \omega \right) \left(\alpha + 2\beta(1+t) + 3\gamma(1+t)^2 \right). \tag{26}$$

On simplifying (26) and equating the coefficients of $\frac{t^r}{r!}$ on both sides, we get (22).

3 Maximum Likelihood Estimation

Here we consider the estimation of the parameters ω, λ, α, β and γ of the ZIGHPD by the method of maximum likelihood. For any $y = 0, 1, 2, \ldots$, let $A(y)$ be the observed frequency of y events and let z be the highest value of y observed. Then the likelihood function of the sample is given by

$$L(\theta; y) = \prod_{y=0}^{z} [f(y)]^{A(y)},$$

where $f(y)$ is the p.m.f of the ZIGHPD given in (4).
Now $L(\theta; y)$ can be written as

$$L(\theta; y) = (f(0))^s \prod_{y=1}^{z} (f(y))^{A(y)},$$

where $s = A(0)$.

Then the log-likelihood function can be written as

$$\begin{aligned} \ln L(\theta; y) &= s \ln \left(\omega + \frac{(1-\omega)}{\phi(1; \lambda; \alpha+\beta+\gamma)} \right) + \sum_{m=1}^{z} A(m) \\ &\times \ln \left[\frac{(1-\omega)}{\phi(1; \lambda; \alpha+\beta+\gamma)} \sum_{j=0}^{[\frac{m}{3}]} \sum_{k=0}^{[\frac{m}{2}]} \frac{(1)_{m-k-2j}}{(\lambda)_{m-k-2j}} \frac{\alpha^{m-2k-3j}}{(m-2k-3j)!} \frac{\beta^k}{k!} \frac{\gamma^j}{j!} \right]. \end{aligned} \tag{27}$$

Assume that $\hat{\omega}$, $\hat{\lambda}$, $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ be the maximum likelihood estimators of the parameters ω, λ, α, β and $\hat{\gamma}$ of the ZIGHPD. Now, on differentiating the log-likelihood function (27) with respect to ω, λ, α, β and $\hat{\gamma}$ and equating to zero, we obtain the following likelihood equations:

$$\frac{\partial \ln L}{\partial \omega} = 0,$$

which implies

$$\frac{s\left[\phi(1; \lambda; \alpha+\beta+\gamma) - 1\right]}{\omega\phi(1; \lambda; \alpha+\beta+\gamma) + (1-\omega)} - \sum_{m=1}^{z} \frac{A(m)}{(1-\omega)} = 0, \tag{28}$$

$$\frac{\partial \ln L}{\partial \lambda} = 0,$$

which implies

$$\frac{s(1-\omega)}{\left[\omega\phi(1; \lambda; \alpha+\beta+\gamma) + (1-\omega)\right]\phi(1; \lambda; \alpha+\beta+\gamma)} \sum_{n=0}^{\infty} \frac{(\alpha+\beta+\gamma)^n}{(\lambda)_n} [\psi(\lambda) - \psi(\lambda+n)]$$

$$\sum_{m=1}^{z} A(m) \left\{ \frac{\sum_{n=0}^{\infty} \frac{(\alpha+\beta+\gamma)^n}{(\lambda)_n} [\psi(\lambda) - \psi(\lambda+n)]}{\phi(1; \lambda; \alpha+\beta+\gamma)} \right\} + \frac{\frac{1}{(\lambda)_{m-k-2j}} [\psi(\lambda) - \psi(\lambda+m-k-2j)]}{\sum_{j=0}^{[\frac{m}{3}]} \sum_{k=0}^{[\frac{m}{2}]} \frac{(1)_{m-k-2j}}{(\lambda)_{m-k-2j}} \frac{\alpha^{m-2k-3j-1}}{(m-2k-3j-1)!} \frac{\beta^k}{k!} \frac{\gamma^j}{j!}} = 0, \tag{29}$$

$$\frac{\partial \ln L}{\partial \alpha} = 0,$$

which implies

$$\frac{s(1-\omega)\phi(1; \lambda; \alpha+\beta+\gamma)}{\omega\phi(1; \lambda; \alpha+\beta+\gamma) + (1-\omega)} \phi(2; \lambda+1; \alpha+\beta+\gamma) - \sum_{m=1}^{z} A(m) \times$$

$$\left\{\frac{\phi(2;\lambda+1;\alpha+\beta+\gamma)}{\lambda\phi(1;\lambda;\alpha+\beta+\gamma)}-\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-k-2j}}{(\lambda)_{m-k-2j}}\frac{\alpha^{m-2k-3j-1}}{(m-2k-3j-1)!}\frac{\beta^k}{k!}\frac{\gamma^j}{j!}\right\}=0, \tag{30}$$

$$\frac{\partial \ln L}{\partial \beta}=0,$$

which implies

$$\frac{s(1-\omega)\phi(1;\lambda;\alpha+\beta+\gamma)}{\omega\phi(1;\lambda;\alpha+\beta+\gamma)+(1-\omega)}\phi(2;\lambda+1;\alpha+\beta+\gamma)-\sum_{m=1}^{z}A(m)\times$$

$$\left\{\frac{\phi(2;\lambda+1;\alpha+\beta+\gamma)}{\lambda\phi(1;\lambda;\alpha+\beta+\gamma)}-\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-k-2j}}{(\lambda)_{m-k-2j}}\frac{\alpha^{m-2k-3j}}{(m-2k-3j)!}\frac{\beta^{k-1}}{(k-1)!}\frac{\gamma^j}{j!}\right\}=0, \tag{31}$$

and

$$\frac{\partial \ln L}{\partial \gamma}=0,$$

which implies

$$\frac{s(1-\omega)\phi(1;\lambda;\alpha+\beta+\gamma)}{\omega\phi(1;\lambda;\alpha+\beta+\gamma)+(1-\omega)}\phi(2;\lambda+1;\alpha+\beta+\gamma)-\sum_{m=1}^{z}A(m)\times$$

$$\left\{\frac{\phi(2;\lambda+1;\alpha+\beta+\gamma)}{\lambda\phi(1;\lambda;\alpha+\beta+\gamma)}-\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-k-2j}}{(\lambda)_{m-k-2j}}\frac{\alpha^{m-2k-3j}}{(m-2k-3j)!}\frac{\beta^k}{k!}\frac{\gamma^{j-1}}{(j-1)!}\right\}=0. \tag{32}$$

On solving the likelihood equations (28), (29), (30), (31) and (32) with the help of some mathematical softwares, say *Mathematica*, one can obtain the maximum likelihood estimators of the parameters of the proposed distribution. In order to examine the existence of unique solutions to the above likelihood equations we have considered the second derivatives and observed that the second derivatives at the solution of the likelihood equations as negative and by utilizing the *Mathematica* software.

4 Testing

In order to test the significance of the inflation parameter in zero-inflated models, we can use any of the test procedures such as the generalized likelihood ratio test (GLRT), Wald test, Raos efficient score test (REST) etc. Here we adopt the GLRT procedure for testing the significance of the inflation parameter ω of the ZIGHPD.
The null hypothesis is given by

$$H_0 : \omega = 0 \; against \; the \; alternative \; hypothesis \; H_1 : \omega \neq 0.$$

The test statistic suggested in the case of GLRT is given by

$$-2\ln\psi = 2\left(\iota_1 - \iota_2\right), \tag{33}$$

where, $\iota_1 = \ln L(\hat{\theta}; y)$, where $\hat{\theta}$ is the maximum likelihood estimator for $\theta = (\omega, \lambda, \alpha, \beta, \gamma)$ with no restrictions, and $\iota_2 = InL(\hat{\theta}^*; y)$, in which $\hat{\theta}^*$ is the maximum likelihood estimator for θ under the null hypothesis H_0. The test statistic defined in (33) is asymptotically distributed as χ^2 with one degree of freedom.

5 Applications

Here we consider a real life biological data and illustrate the procedures of maximum likelihood estimation and GLRT which are discussed in sections 3 and 4. The data is based on the distribution of European Corn borer Pyrausta Naubilalis in field corn [1] and is presented in Table 1. We have fitted the ZIGHPD to the data set and considered the fitting of the models - ZIHP, zero-inflated alternative hyper-Poisson (ZIAHP), ZIP, zero-inflated generalized Poisson (ZIGP) and GHPD for comparison. For comparing the models we computed the values of χ^2, AIC, BIC and AICc. The numerical results obtained are presented in Tables 1. Based on the computed values of χ^2, AIC, BIC and AICc as presented in Table 1, one can observe that the ZIGHPD gives a better fit to the data set while all other models such as ZIHP, ZIAHP, ZIP, ZIGP and GHPD are not appropriate.

We have also calculated the values of the test statistic. The value of the test statistic for $\ln L(\hat{\theta}^*; w) = -236.84$ and $\ln L(\hat{\theta}; w) = -213.56$ is given by 46.56. The critical value of the test having 5% level of significance and degree of freedom one is 3.84, so that the null hypothesis is rejected in all the cases. Thus, we conclude that the additional parameter ω in the model is significant.

Table 1: Distribution of the spread of European Corn borer Pyrausta Naubilalis in field corn [1] and the expected frequencies computed using ZIHP, ZIAHP, ZIP, ZIGP, GHP and ZIGHP.

Count	Observed frequency	ZIHPD	ZIAHPD	ZIPD	ZIGPD	GHPD	ZIGHPD
0	206	253.6	265.73	252.15	300.3	252.8	202.4
1	143	143.4	148.4	151.43	160.34	162.7	140.5
2	128	125.2	144.2	140.6	95.5	134.4	120.8
3	107	107.9	127.1	118.35	104.6	126.048	99.8
4	71	81.7	80.361	75.78	80.81	80.4	80.4
5	36	27.682	7.29	23.9	34.2	10.5	38.5
6	32	20.4	4.6	7.1	6.25	7.4	39.4
7	17	13.7	2.3	5.4	4.12×10^{-4}	3.6	12.6
8	14	4.4	1.25	4.02	3.25×10^{-6}	1.12	19.2
9	7	2.3	0.5	1.25	1.96×10^{-9}	0.98	5.2
10	7	1.62	0.25	1.6	2.3×10^{-15}	0.87	7.9
11	2	0.002	0.0024	0.0015	1.06×10^{-21}	0.30	1.2
12	3	0.06	0.01	0.35	9.8×10^{-29}	0.15	1.8
13	3	0.027	0.006	0.021	7.5×10^{-45}	0.2	2.3
14	1	0.0001	0.00003	0.006	1.79×10^{-53}	0.46	2.2
15	1	0.0006	0.00009	0.0002	4.03×10^{-64}	0.07	1.1
16	1	0.0007	0.000006	0.00004	1.05×10^{-75}	0.002	1.7
17	2	0.008	0.000007	0.041	2.6×10^{-84}	0.00035	3.5
18	1	0.000004	0.000009	0.000008	4.86×10^{-95}	0.000001	1.5
Total	782	782	782	782	782	782	782
df		5	3	6	3	3	6
Estimates		γ=0.22 ω=0.79 θ=0.92	γ=12.09 ω=0.59 θ=7.0009	γ=3.95 ω=0.17	γ=3.28 ω=0.15 θ=0.0000003	γ=0.16 ω=0.31 θ=0.61	λ=0.64 ω=0.26 α=0.21 β=0.59 γ=0.025
χ^2-value		154.25	880.21	188.64	1193.01	41.13	7.48

P-value	0.0001	0.0001	0.0001	0.0001	0.0001	0.2787
AIC	1628.6	1733.5	910.6	2841.4	1322.5	840.25
BIC	1629.6	1734.3	911.8	2842.6	1323.3	841.25.5
AICc	1633.4	1739.6	915.3	2846.2	1327.2	845.7

Acknowledgements

The authors are highly thankful to the Editor in Chief, the Associate Editor and the anonymous referees for their fruitful comments, which immensely helped to improve the quality and presentation of the article.

Funding

The first author is particularly thankful to Department of Science and Technology, New Delhi (Government of India) for financial support (MTR/2017/000942).

References

[1] J Rodríguez Avi et al. "The 3 F 2 with complex parameters as generating function of discrete distribution." In: *Communications in Statistics-Theory and Methods* 37.19 (2008), pp. 3009–3022.

[2] C. S. Kumar and B. U. Nair. "A generalized hyper-Poisson distribution and its properties." In: *In Statistical Methods in Interdisciplinary Studies (S. Santhosh, Ed.), Department of Statistics, Maharajas College, Kochi, 71-78* (2011).

[3] C. S. Kumar and Rakhi Ramachandran. "A generalization to zero-inflated hyper-Poisson distribution: Properties and applications." In: *Communications in Statistics-Theory and Methods* (2022), pp. 1–14.

[4] C. S. Kumar and Rakhi Ramachandran. "On some aspects of a zero-inflated overdispersed model and its applications." In: *Journal of Applied Statistics* 47.3 (2020), pp. 506–523.

[5] C. S. Kumar and Rakhi Ramachandran. "On zero-inflated hyper-Poisson distribution and its applications." In: *Communications in Statistics-Simulation and Computation* 51.3 (2020), pp. 868–881.

[6] Diane Lambert. "Zero-inflated Poisson regression, with an application to defects in manufacturing." In: *Technometrics* 34.1 (1992), pp. 1–14.

[7] Arakaparampil M Mathai and Hans J Haubold. *Special functions for applied scientists*. Vol. 4. Springer, 2008.

[8] Lucy Joan Slater. *Generalized hypergeometric functions*. Cambridge university press, 1966.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Construction of Bivariate Burr (Type VII) distributions via conditional specification: A brief survey

Indranil Ghosh[1]

1. *University of North Carolina, Wilmington, North Carolina, email: ghoshi@uncw.edu*

Received: 26 May 2022 / Revised: 06 Jun 2022 / Accepted: 20 Jun 2022

Abstract

The univariate Burr distribution and its properties and applications have been studied quite extensively in the literature. Some generalizations, as well as multivariate extensions of it, have also been proposed for greater flexibility in modeling bivariate and multivariate data. In this paper, we construct generalized bivariate Burr (Type VII) distributions through conditional specification and discuss some of its properties. We conjecture at this point that an analogous development can be made in deriving other bivariate Burr family type distributions. We also provide some results which reveal the relationship between a Burr type distribution(s) with a beta type 1 distribution.

Keywords: Burr Type distributions, Bivariate Burr Type distributions, Conditional specification, Total positivity property, Dependence structures.

1 Introduction

In the last two decades, considerable amount of work has been done on introducing various bivariate non-normal models and then discussing their properties, fit and applications; for elaborate details, one may refer to the books by [5] and [2] and the references therein. One such model that has been studied not so extensively in the literature is the bivariate Burr type distributions; for pertinent details, see [7] in which the author primarily focuses on Burr Type XII (which is essentially a Pareto (Type IV) model) and its associated skewness and kurtosis measures. The Burr system of distributions was constructed in 1941 by Irving W. Burr. Since the corresponding density functions have a wide variety of shapes, this system is useful for approximating histograms, particularly when a simple mathematical structure for the fitted cumulative distribution function (c.d.f.) is required. Other applications include quantal response, approximation of distributions and development of non-normal control charts. A number of standard theoretical distributions are limiting forms of Burr distributions. The original motivation for developing the Burr system was to provide a method for fitting c.d.f.'s to frequency data. Burr choose to work with the c.d.f. $F(x)$ satisfying the differential equation

$$\frac{dy}{dx} = y\,(1-y)\,g\,(x,y)\,,$$

where $y = F(x)$. The function g must be positive for $0 \leq y \leq 1$ and x in the support of $F(x)$. The solutions $F(x)$ of Burr's differential equation can be classified by their functional forms, each of which gives rise to a family of c.d.f.s within the Burr system as different choices of $g(x, y)$ generate various solutions for $F(x)$. In this article, we focus on the development bivariate Burr (Type VII) distributions

© 2022 Author(s). (https://www.thegsa.in/).

via conditional specification. The flexibility of Burr Type distributions and its importance as a growth curve have attracted many researchers to study this distribution and various generalizations. Many of these generalizations were introduced in order to provide greater flexibility while modeling either skewed and/or heavy tailed data. We begin our discussion by considering the following of Burr (type-VII) distribution described in detail in [5]:

The pdf of a Burr (Type VII) distribution distribution is given by

$$f(x) = \frac{2k(\exp(2x))^k}{(1+\exp(2x))^{k+1}} \times I(-\infty < x < \infty), \quad k > 0. \tag{1}$$

Next, observe that the specification of joint distributions by means of conditional densities has received considerable attention in the literature. A study of bivariate distributions can not be complete without the knowledge of univariate distributions, which would naturally form the marginal or conditional distributions. There are various ways of conditional specifications by which one can identify or (classify) a family of bivariate distributions. We know that if we are given both the families of conditional densities, of X given Y and Y given X, then the information is more than enough to characterize the joint density of (X, Y). For a good reference on the theory and methodology related to conditional specification, the reader is suggested the book by [1] and the references cited therein.

The rest of the paper is organized as follows. In Section 2, we discuss briefly the idea of conditional specification of bivariate distributions. In Section 3, we discuss and derive the joint distribution of a bivariate Burr (Type VII) distributions via conditional specification from the information that both the conditionals (i.e., X given Y and Y given X) are in the same family of univariate Burr (Type VII) distribution. In Section 4, we discuss some useful structural properties of the derived distribution in Section 3. In Section 5, we discuss some useful distributional results which focuses on relationship between a Burr and Beta distribution of the first kind and provide some open questions for future works in this direction. Finally, in Section 6, we provide some concluding remarks.

2 Conditional specification of a bivariate distribution

There are several different conditional specifications through which one can identify or (classify) a family of bivariate distributions. If we are given both families of conditional densities, of X given Y and Y given X, then the information is more than enough to characterize the joint density of (X, Y). We focuses on cases in which the conditional densities are only assumed to be known to belong to specified parametric families. The models consequently derived are called conditionally specified models. Before proceeding further, we mention the following theorem which is very useful in the rest of our discussion. It is originally due to Aczel (see also [1]).

Theorem 2.1. *All solutions of the equation*

$$\sum_{i=1}^{r} f_i(x)\phi_i(y) = \sum_{j=1}^{s} g_j(y)\Psi_j(x), x \in S(X), y \in S(Y),$$

where ϕ_i (for $i = 1, 2, .., r$) and Ψ_j (for $j = 1, 2, .., s$) are given systems of mutually linearly independent functions, are of the form

$$\underline{f}(x) = C\underline{\Psi}(x),$$

and

$$\underline{g}(y) = D\underline{\phi}(y),$$

where $D = C'$.

3 Bivariate Burr(Type VII) distribution

Suppose, we want for each fixed x,

$$f(y|X=x) = 2k_1(x)\frac{(\exp(2y))^{k_1(x)}}{(1+\exp(2y))^{k_1(x)+1}} \times I(-\infty < y < \infty),$$

and, for each fixed y,

$$f(x|Y=y) = 2k_2(y)\frac{(\exp(2x))^{k_2(y)}}{(1+\exp(2x))^{k_2(y)+1}} \times I(-\infty < x < \infty),$$

where $k_1(x)$ is a function depending upon x and $k_2(y)$ is a function depending upon y and they are both unknown. Next we want to identify the class of all bivariate distributions for which both the conditionals are of Burr(Type VII) with the form mentioned above. For that we consider the following: Let $g(x)$ and $h(y)$ be the marginals of X and Y respectively. Then writing down the joint density of (X, Y) as a product of marginals and conditionals we can have

$$g(x)f\,(y|X=x) = h(y)f\,(x|Y=y)\,. \tag{2}$$

Equivalently, we can write

$$g(x)2k_1(x)\frac{(\exp(2y))^{k_1(x)}}{(1+\exp(2y))^{k_1(x)+1}} = h(y)2k_2(y)\frac{(\exp(2x))^{k_2(y)}}{(1+\exp(2x))^{k_2(y)+1}}. \tag{3}$$

Define

$$\theta_1(x) = g(x)k_1(x),$$

and

$$\theta_2(y) = h(y)k_2(y).$$

Then, we can write

$$\begin{aligned}&\exp\left[\log\theta_1(x) + k_1(x)2y - (k_1(x)+1)\log(1+\exp(2y))\right]\\ &= \exp\left[\log\theta_2(y) + k_2(y)2x - (k_2(y)+1)\log(1+\exp(2x))\right],\end{aligned} \tag{4}$$

which we can re-write equivalently as

$$\begin{aligned}&\log\left\{\theta_1(x)(1+\exp(2x))\right\} + k_1(x)\left\{2y - \log(1+\exp(2y))\right\}\\ &= \log\left\{\theta_2(y)(1+\exp(2y))\right\} + k_2(y)\left\{2x - \log(1+\exp(2x))\right\}.\end{aligned} \tag{5}$$

If we write

$$\begin{aligned} f_1(x) &= \log(\theta_1(x)(1+\exp(2x))), \\ f_2(x) &= k_1(x), \\ g_1(y) &= \log(\theta_2(y)(1+\exp(2y))), \\ g_2(y) &= k_2(y) \\ \phi_1(y) &= 1, \\ \phi_2(y) &= 2y - \log\left(1+e^{2y}\right), \\ \Psi_1(x) &= 1, \\ \Psi_2(x) &= 2x - \log\left(1+e^{2x}\right), \end{aligned} \tag{6}$$

then, equation (3) can be rewritten as

$$\sum_{i=1}^{2} f_i(x)\phi_i(y) = \sum_{i=1}^{2} g_i(x)\Psi_i(x).$$

This implies according to Aczel's theorem ([1]) that a general solution to the above equation, will be

$$\begin{aligned} f_1(x) &= a + b\left[2x - \log\left(1+\exp(2x)\right)\right], \\ f_2(x) &= c + d\left[2x - \log\left(1+\exp(2x)\right)\right], \\ g_1(y) &= a + c\left[2y - \log\left(1+\exp(2y)\right)\right], \\ g_2(y) &= b + d\left[2y - \log\left(1+\exp(2y)\right)\right], \end{aligned} \tag{7}$$

where a, b, c, d are unknown parameters. Note that $k_2(y) = b + d\left[2y - \log\left(1+\exp(2y)\right)\right]$, and $k_1(x) = c + d\left[2x - \log\left(1+\exp(2x)\right)\right]$.

Since, $f_2(x) = c + d\left[2x - \log\left(1+\exp(2x)\right)\right] = k_1(x) > 0$, for all $x \in \mathbb{R}$, and $-\log\left(1+\exp(2x)\right) < 0$, for all $x \in \mathbb{R}$, it follows that c should be positive $(c > 0)$ and $(d \leq 0)$. Similarly, from $g_2(y) = k_2(y) > 0$, we obtain that $b > 0$ and $d \leq 0$. The constant a is a normalizing constant and it can be evaluated from the condition $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$, or from the condition that the marginal densities when integrated over $(-\infty, \infty)$ is 1. To get the exact expression of the joint density $f(x,y)$, we need to simplify $\theta_1(x)$ and/ or $\theta_2(y)$. We have from (3)

$$f_1(x) = a + b\left[2x - \log\left(1+\exp(2x)\right)\right] = \log\left(\theta_1(x)\left(1+\exp(2x)\right)\right).$$

From this, we get

$$\theta_1(x) = \frac{\exp(a)\exp\left(2bx\right)}{\left[1+\exp\left(2bx\right)\right]^{b+1}}.$$

Using the expression of $\theta_1(x)$, we obtain the joint density of (X, Y) as

$$\begin{aligned} &f_{X,Y}(x,y) \\ &= \frac{\exp(a)\exp\left\{2bx + 2cy + 4dxy\right\}\left(1+\exp(2y)\right)^{d\log(1+\exp(2x))}}{\left(1+\exp(2x)\right)^{b+2dy+1}\left(1+\exp(2y)\right)^{c+1}} I(-\infty < (x,y) < \infty). \end{aligned} \tag{8}$$

The marginals densities can be obtained as follows. We have $\theta_1(x) = g(x)k_1(x)$. Next, using explicit expressions for $\theta_1(x)$ and $k_1(x)$ as obtained earlier, we get, the marginal density of X as

$$g_X(x) = \frac{\exp\{a+2bx\}}{(1+\exp(2x))^{b+1}\left[c+d\left\{2x-\log\left(1+\exp(2x)\right)\right\}\right]} I(-\infty < x < \infty). \tag{9}$$

We can study the rate of a random variable X, [see [4]] for pertinent details] with the following definition:

Following Klugman et al. (1998) , the rate of a continuous random variable X is given by

$$\begin{aligned}\tau_x &= -\lim_{x\to\infty}\frac{d\log g_X(x)}{dx}\\ &= -\lim_{x\to\infty}\left(-\frac{2\left(-bc+be^{2y}-2cdy-d\left(e^{2y}-c\right)\log\left(e^{2y}+1\right)+2de^{2y}y+d\right)}{\left(e^{2y}+1\right)\left(b+2dy-d\log\left(e^{2y}+1\right)\right)}\right)\\ &= 2.\end{aligned} \tag{10}$$

From the above, we can say that the tail of X is heavier. A similar behavior/pattern of the rate function can be observed for the other random variable Y.

Similarly, from $\theta_2(y) = h(y)\delta_2(y)$, we obtain the marginal density of Y as

$$h_Y(y) = \frac{\exp\{a+2cy\}}{(1+\exp(2y))^{c+1}\left[b+d\left\{2y-\log\left(1+\exp(2y)\right)\right\}\right]} I(-\infty < y < \infty). \tag{11}$$

Some representative bivariate density plots corresponding to the Eq. (9)) are provided in Figures $1-4$.

4 Structural properties of the bivariate Burr type (VII) distribution

In this section, we discuss some useful structural properties for the bivariate Burr (type-VII) distribution given in Eq. (9) beginning with the discussion on several useful properties for the marginals associated with it.

1. **Result** 1 The marginal densities corresponding to the bivariate density in Eq. (9) are log-concave. Therefore, the marginals are unimodal.

   Proof. Let us consider the marginal density of Y. From (11), one may obtain

$$\begin{aligned}\frac{\partial^2\log\left(h(y)\right)}{\partial y^2} &= -\left(\left(e^{2y}+1\right)^2\left(b+2dy-d\log\left(e^{2y}+1\right)\right)^2\right)^{-1}\\ &\times\Big(4(b^2(c+1)e^{2y}+bde^{2y}(4(c+1)y-1)\\ &-de^{2y}\log(e^{2y}+1)(2b(c+1)+d(4(c+1)y-1))\\ &+d^2(2e^{2y}y(2(c+1)y-1)-1)+(c+1)d^2e^{2y}\log^2(e^{2y}+1))\Big).\end{aligned} \tag{12}$$

   From (12), it is clear that for any choices of $(a,b,c,d)\in\mathbb{R}$ the second order derivative is <0. Hence, the proof. Observe that the tails of log-concave densities are necessarily sub-exponential. Similarly, the log-concavity for the distribution of X can be established.

2. **Result** 2 The hazard rate function for the marginal densities [Eq. (9) and (10)] are always decreasing for any choices of $a, b, c, d \in \mathbb{R}$, with $d \leq 0$.

   Proof. Let us consider the marginal density of Y.

   We have,

$$\begin{aligned}\eta(y) &= -\frac{h'(y)}{h(y)} \\ &= \frac{2\left(-bc + be^{2y} - 2cdy - d\left(e^{2y}-c\right)\log\left(e^{2y}+1\right) + 2de^{2y}y + d\right)}{\left(e^{2y}+1\right)\left(b + 2dy - d\log\left(e^{2y}+1\right)\right)}.\end{aligned}$$

   Consequently,

$$\begin{aligned}\eta'(y) &= -\left(\left(e^{2y}+1\right)^2\left(b+2dy-d\log\left(e^{2y}+1\right)\right)^2\right)^{-1} \\ &\quad\times\Big(4(b^2(c+1)e^{2y} + bde^{2y}(4(c+1)y-1) \qquad (13)\\ &\quad -de^{2y}\log(e^{2y}+1)(2b(c+1)+d(4(c+1)y-1)) \\ &\quad +d^2\left(2e^{2y}y(2(c+1)y-1)-1\right) + (c+1)d^2e^{2y}\log^2(e^{2y}+1)\right)\Big). \qquad (14)\end{aligned}$$

   Note that, for any choices of $(a, b, c, d) \in \mathbb{R}$, $\eta'(y) > 0$. Therefore, from Theorem (b) of [3], we can say that the hazard function decreasing. Hence, the proof.

3. **Total positivity of order 2 property:**

   Let t_{11}, t_{12}, t_{21} and t_{22} be real numbers with $0 < t_{11} < t_{12}$ and $0 < t_{21} < t_{22}$. (X, Y) has the total positivity of order two (TP_2) property iff for any such set of t_{ij}'s,

$$f_{X,Y}(t_{11}, t_{21})f_{X,Y}(t_{12}, t_{22}) - f_{X,Y}(t_{12}, t_{21})f_{X,Y}(t_{11}, t_{22}) \geq 0. \qquad (15)$$

   Theorem 4.1. *The bivariate Burr type (VII) distribution has the TP_2 property.*

   Proof. Let us consider different cases separately. If $0 < t_{11} < t_{21} < t_{12} < t_{22}$, then for the density function in (8), one can easily show that the condition in (12) is equivalent to

$$\begin{aligned}&\left[\{1+\exp(2t_{21}\}^{d\log(1+\exp(2t_{11}))-\log(1+\exp(2t_{12}))}\right] \times \left[\{1+\exp(2t_{22}\}^{d\log(1+\exp(2t_{21}))-\log(1+\exp(2t_{11}))}\right] \\ &\times \left[\{1+\exp(2t_{11}\}^{2d(t_{21}-t_{22})}\right] \times \left[\{1+\exp(2t_{12}\}^{2d(t_{22}-t_{21})}\right] \\ &\geq 0. \qquad (16)\end{aligned}$$

   Observe that (15) holds for every $0 < t_{11} < t_{12}$ and $0 < t_{21} < t_{22}$. Similarly, other cases can also be established.

   Furthermore, TP_2 is the most rigid dependence property. Consequently, the positive quadrant dependence property (alternatively, the TP_2 property) implies the following:

(a) X and Y are positive quadrant dependent,

(b) $X(Y)$ is a positive regression dependent of $Y(X)$,

(c) $X(Y)$ is a left tail decreasing in $Y(X)$,

(d) $P(X \leq x|Y = y)$ is non-increasing in y for all x,

(e) $P(Y \leq y|X = x)$ is non-increasing in x for all y,

(f) $P(Y > y|X > x)$ is non-decreasing in x for all y,

(g) $P(Y \leq y|X \leq x) \geq P(Y \leq y)P(X \leq x)$,

(h) $P(Y > y|X > x) \geq P(Y > y)P(X > x)$.

4. **Dependence structure:**

   Observe that the joint density in (8) can be re-written as

$$f(x, y) = C r_1(x) r_2(y) \exp\left(\vec{q}_1^T(x) M_2 \vec{q}_2(y)\right),$$

   where C = Normalizing constant, $r_1(x) = \frac{\exp(2bx)}{(1+\exp(2x))^{b+1}}$, $r_2(y) = \frac{\exp(2cy)}{(1+\exp(2y))^{c+1}}$, $\vec{q}_1(x) = (x \;\; \log(1+\exp(2x))$, $\vec{q}_2(y) = (1 \;\; \log(1+\exp(2y))$, and

$$M_2 = \begin{bmatrix} 4d & -2d \\ -2 & d \end{bmatrix}.$$

   Then, according to [1], based on the elements of the matrix M_2, we can have the following different scenarios:

   - One will observe positive correlation iff $4d^2 - 4d > 0$,
   - The distributions of X and Y will be independent iff $d = 0$.

     For this bivariate probability model, since $d \leq 0$, the quantity $4d^2 - 4d$ will always be non-negative. Consequently, the effective range of the correlation coefficient will be $[0, 1]$.

     Regarding bivariate aging pattern, since the bivariate density in Eq. (8), has the positive dependence property (alias, the TP_2 property), according to [6], the following holds: Positive dependence and positive one-dimensional aging implies positive bivariate aging.

5. Next, we propose the following result which will be helpful for us in characterizing the bivariate Burr (type VII) distribution.

Relation between Burr(Type VII) and Beta distribution

Lemma 4.1. *If a random variable $X \sim Burr(TypeVII)$ with parameter k, then $Y = \exp(2X) \sim Beta_2(k, 1)$, where $B_2(k, 1)$ means Beta distribution of the second kind.*

Proof. We have

$$f(x) = 2k \frac{(\exp(2x))^k}{(1+\exp(2x))^{k+1}} \times I(-\infty < x < \infty).$$

Let us consider $Y = \exp(2X)$, then $y \in (0, \infty)$. Then, the Jacobian of the above transformation is given by $|J(\frac{x}{y})| = \frac{1}{2y}$. Therefore, the density of Y is given by

$$f_Y(y) = \frac{2ky^k}{(1+y)^{k+1}}\left(\frac{1}{2y}\right) = \frac{y^{k-1}}{B(k,1)(1+y)^{k+1}} \times I(0 < y < \infty).$$

Hence $Y = \exp(2X) \sim Beta_2(k, 1)$.

6. Characterization via Beta distribution

Previously we mentioned that we want the class of all bivariate Burr(Type VII) densities which will have conditionals in the same family. Recalling that specifically we want

$$X|Y = y \sim Burr(TypeVII(k_2(y))),$$

and

$$Y|X = x \sim Burr(TypeVII(k_1(x))),$$

where the expressions for both the unknown parameters $k_2(y)$ and $k_1(x)$ is given earlier.

Next, using the relationship with Beta distribution we can write

- $E[\exp(2X)|Y = y] = \frac{k_2(y)}{(k_2(y)+1)}$.
- $E[\exp(2Y)|X = x] = \frac{k_1(x)}{(k_1(x)+1)}$.

Then, we have the following theorem.

Theorem 4.2. *If the joint density of two continuous random variables (X, Y) with $S(X) = (-\infty, \infty)$, $S(Y) = (-\infty, \infty)$ is of the form as in (8), with*

$X|Y = y \sim Burr\left(TypeVII(k_2(y))\right)$,
where $k_2(y) = b + d * (2y - \log(1 + \exp(2y)))$,
and $E[\exp(2Y)|X = x) = \frac{k_1(x)}{(k_1(x)+1)}$,
where $k_1(x) = c + d\left(2x - \log(1 + \exp(2x))\right)$.

Then, the above conditions uniquely determines the class of all Bivariate Burr(Type VII) distributions with parameters a, b, c, d with $a = \frac{b(1+\frac{c}{d}) - \log(\frac{c}{d})}{1+\frac{d}{c}}$.

5 Some other useful results and some open questions

First of all we note that all the members of the Burr Type distributions with one parameter(viz.(Type II,Type VII,Type $VIII$, Type X,Type XI)) can be transformed to a Beta distribution of both kind. We can illustrate few of them, for examples,

(a) If $X \sim Burr(TypeVII)$ with parameter k, then $Y = \exp(2X) \sim Beta_2(k, 1)$..

(b) If $X \sim Burr(Type(VIII))$ with parameter k, then $Y = \frac{\pi}{2}(tan^{-1}(\exp(x))) \sim Beta_1(k, 1)$.

(c) If $X \sim Burr(Type(II))$ with parameter k ,then $Y = 1 - \exp(-X) \sim Beta_2(k, 1)$.

Again, we mention the following straightforward result:

Result: If $U \sim Uniform(0,1)$ then $Z = U^{\frac{1}{k}} \sim Beta_1(k,1)$.

So, if we consider the transformation $Y = \psi(U^{\frac{1}{k}})$, then this transformation includes the class of all one parameter exponential family(OPEF, henceforth) densities as Uniform $(0,1)$ is a member of the OPEF and we consider here ψ to be monotonically increasing, invertible, differentiable and one-to-one function of the random variable U and we know that any one-to-one function of the members of the OPEF is itself a member of the OPEF. However we will consider the following:

- First consider all bivariate densities with conditionals in $\mathcal{F}_\psi$, and let us call that $\mathcal{F}_{C\psi}$, where C stands for conditional.
- Next we try to identify the class of all densities with conditionals in $\mathcal{F}_I$, where I stands for the identity function.
- Suppose that $F(x,y)\epsilon\mathcal{F}_{CI}$ and consider $(X',Y') = (\psi(X),\psi(Y))$.

  Then we claim that $(X',Y')\epsilon\mathcal{F}_{C\psi}$.

Consequently, the above compel us to consider the following questions as well which are as follows:

(i) Is it possible to find mixtures of $\mathcal{F}_{C\psi}$ distributions?

(ii) Is it possible to bring in the concept of hidden truncation in this context?

(iii) Can this family of densities be used efficiently in Stress-Strength analysis as an application?

6 Concluding remarks

The concept of conditional specification of bivariate distributions is not new but, except in normal and exponential families (see [1], it has not been well developed for other distributions in the literature. Computational difficulties, absence of analytically tractable forms of the marginal as well as bivariate densities, without a doubt discouraged further work in this direction. In this paper, we have addressed the characterization of a particular type of bivariate Burr distributions (Type VII) via conditional specification, where both the conditional distributions belonging to the same family. The results obtained in this paper can be equivalently mimicked in extending to a multivariate scenario albeit computational complexity. However, a daunting question will remain regarding its analytical tractability and associated inference which we plan to consider in a future research.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions that has helped tremendously to prepare this version of the manuscript.

References

[1] Barry C Arnold and José María Sarabia. "Conditional specification of statistical models: Classical models, new developments and challenges." In: *Journal of Multivariate Analysis* 188.C (2022).

[2] Narayanaswamy Balakrishnan and Chin Diew Lai. *Continuous bivariate distributions*. Springer Science & Business Media, 2009.

[3] Ronald E Glaser. "Bathtub and related failure rate characterizations." In: *Journal of the American Statistical Association* 75.371 (1980), pp. 667–672.

[4] Stuart A Klugman, Harry H Panjer, and Gordon E Willmot. *Loss models: from data to decisions*. Vol. 715. John Wiley & Sons, 2012.

[5] Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L Johnson. *Continuous multivariate distributions, Volume 1: Models and applications*. Vol. 1. John Wiley & Sons, 2004.

[6] Giovanna Nappo and Fabio L Spizzichino. "Revisiting Relations between Stochastic Ageing and Dependence for Exchangeable Lifetimes with an Extension for the IFRA/DFRA Property." In: *arXiv preprint arXiv:1905.10326* (2019).

[7] Robert N Rodriguez. "A guide to the Burr type XII distributions." In: *Biometrika* 64.1 (1977), pp. 129–134.

Appendix

In the figures below we provide some representative bivariate Burr (Type VII) type density (Eq. (8)) plots for various parameter choices.

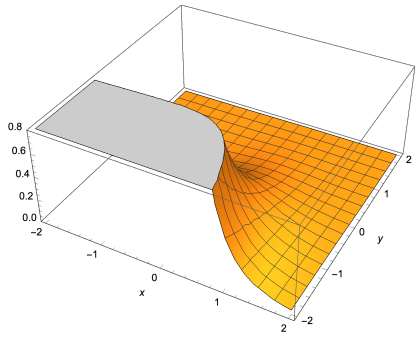


Figure 1: The Bivariate Burr (type VII) density for $a = 2.75, b = 3, c = 4.5, d = 2$

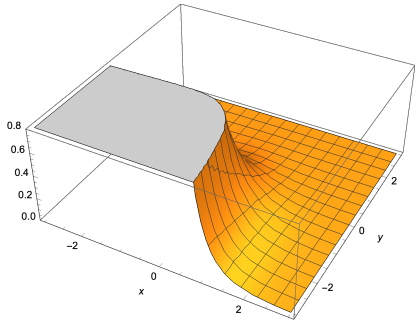


Figure 2: The Bivariate Burr (type VII) density $a = 1, b = 1, c = 1, d = 1$.

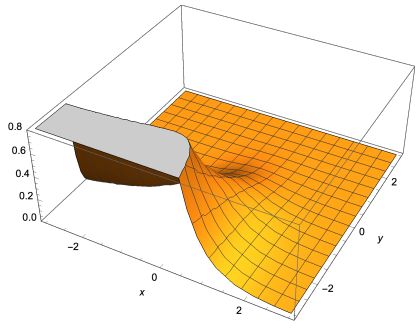


Figure 3: The Bivariate Burr (type VII) density $a = 0.5, b = 0.8, c = 0.7, d = 0.025$

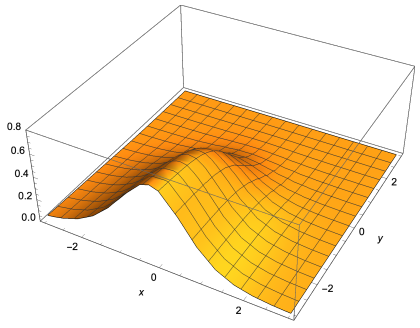


Figure 4: The Bivariate Burr (type VII) density $a = 0.5, b = 0.8, c = 0.7, d = 0.025$

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

The Simultaneous Assessment of Normality and Homoscedasticity for Bivariate Grouped Data

Ye Yang[1] and Thomas Mathew[2]

1. *Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250, USA*
2. *Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250, USA, Email: mathew@umbc.edu*

Received: 26 Dec 2021 / Revised: 30 Jun 2022 / Accepted: 10 Jul 2022

Abstract

For grouped bivariate data, the problem of simultaneously testing the assumptions of normality and homoscedasticity is investigated. Tests are derived under the assumption of a smooth alternative to the bivariate normal distribution, where the smooth alternative is specified using Legendre polynomials. Score statistics are obtained when the smooth alternatives are common across the different groups, and when they are different across the groups. Both balanced and unbalanced data situations are considered. For determining the order of the Legendre polynomials, a data-driven approach considered in the literature is recommended. The performance of the tests are assessed based on estimated Type I error probabilities. The results are illustrated using two examples.

Keywords: Bivariate normal distribution, Legendre polynomials, Score test, Smooth alternative.
AMS subject classification: 62F03, 62H15

1 Introduction

Background

In the analysis of variance, two crucial assumptions are normality and homoscedasticity. Model diagnostics are usually carried out in order to assess the validity of these assumptions. Parametric tests for homoscedasticity are usually based on the normality assumption. Furthermore, formal tests and graphical procedures for assessing normality usually assume homoscedasticity. In view of this, it is important to jointly test the validity of both, without making either assumption. For grouped data following a univariate linear regression model, joint testing of normality and homoscedasticity is developed in [21]. In a subsequent article, [22], the authors addressed the same problem in a one-way random effects model. The present article is an extension of the work reported in [21], when the data are bivariate. In other words, we have bivariate data falling into different groups, typically based on the different levels of a classification variable, and we want to test bivariate normality along with homoscedasticity. In particular, our methodology can be adopted to test bivariate normality along with homoscedasticity in a bivariate one-way fixed effects model. Following [21] and [22], the alternative to bivariate normality that we shall consider is specified by embedding the class of possible alternatives within the class of *smooth alternatives* introduced by [17]. This results in alternatives that depend on a finite number of parameters, the normal distribution being a special case. We can then develop the score test for testing bivariate normality and homoscedasticity, simultaneously. We start with the definition of a smooth alternative for the univariate case, as defined by [17].

© 2022 Author(s). (https://www.thegsa.in/).

Smooth alternatives

Suppose we want to test if a continuous random variable Y follows the density $f(y, \boldsymbol{\beta})$, where the vector $\boldsymbol{\beta}$ represents unknown parameters. An order k smooth alternative, denoted by $g_k(y; \boldsymbol{\theta}, \boldsymbol{\beta})$, is defined as

$$g_k(y; \boldsymbol{\theta}, \boldsymbol{\beta}) = C(\boldsymbol{\theta}, \boldsymbol{\beta}) \exp\left\{\sum_{i=1}^{k} \theta_i h_i(y; \boldsymbol{\beta})\right\} f(y; \boldsymbol{\beta}), \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)'$ is a vector of unknown parameters, $C(\boldsymbol{\theta}, \boldsymbol{\beta})$ is a normalizing constant, and the $\{h_i(y; \boldsymbol{\beta})\}$ are orthonormal polynomials, orthonormal with respect to $f(y; \boldsymbol{\beta})$. If all the θ_i's are zero, then the density in (1) reduces to the density $f(y; \boldsymbol{\beta})$. In other words, in order to assess the validity of the density $f(y; \boldsymbol{\beta})$, the hypothesis of interest is that of testing if all the θ_i's are zero. Thus the hypothesis now involves only k parameters, assuming that k is known (the choice of k will be addressed later).

Brief literature review

Score tests and their properties, under smooth alternatives, have been investigated extensively in the literature; a detailed review is available in the book by [18], and in the thesis by [20]. For various goodness of fit problems, theoretical investigations are available in [15], [11], [12], 1[13], [7], [6] and [10]. In particular, these authors develop *data driven smooth tests*; i.e., the order k in (1) is chosen using the data, based on a modified BIC criterion. The authors also note that the k so obtained converges to one in probability, the resulting score test is a consistent test against any alternative, and the asymptotic null distribution of the test statistic is a chisquare distribution having df = 1. The literature on the topic mostly addresses tests for the adequacy of a distribution, assuming the validity of other parametric assumptions. However, [14] address the simultaneous testing of independence and normality in a bivariate setup.

For testing multivariate normality when the covariance matrix is unstructured, several tests are available in the literature. It should be noted that typically, tests for multivariate normality actually test some property implied by multivariate normality. We refer to [16] for a review. Smooth tests have also been attempted for the assessment of multivariate normality; see [18], Section 6.4. As noted by these authors, and as we shall see, the problem of smooth tests for multivariate normality is rather challenging when the dimension gets large, and the authors concentrate mostly on the bivariate case. The article by [2] provides a thorough investigation of smooth tests for assessing bivariate normality. For jointly testing multivariate normality and homoscedasticity, we are aware of only two articles: [5] and [8], both in the context of multivariate models. In his work, [5] considers a set up where multivariate data are available from several groups, and it is desired to test multivariate normality along with the equality of the covariance matrices across the groups. The test is based on several Hotelling's T^2 type statistics from the different groups, and the associated F distributions resulting from the multivariate normality and homoscedasticity. In other words, the test is based on a property (namely, the F-distribution associated with the Hotelling's T^2 statistic) implied by multivariate normality. The later article by [8] extend the work of [5] to missing data situations. Furthermore, [9] provide an R code that can be used to carry out the test.

Summary of our contribution

Our set up is the same as that of [2] except that we are testing both bivariate normality and homoscedasticity when the data fall into groups. The model and the problem are formulated in the next section. Score tests are derived when we have smooth alternatives based on Legendre polynomials. Under the smooth alternative formulation, score statistics are derived under two scenarios: when there is a common smooth alternative across the different groups, and when the

smooth alternatives are different across the groups. For determining the order of the Legendre polynomials, a data-driven approach recommended in the literature will be used. The performance of the tests are assessed based on estimated Type I error probabilities. The results are illustrated using two examples.

2 The model and the testing problem

Suppose we have bivariate data grouped into a groups based on a classification variable. Let X_{lj}, $j = 1, 2, \ldots, n_l$, be a bivariate sample of size n_l from the lth group, $l = 1, 2, \ldots, a$. We assume that the mean vector and the covariance matrix of X_{lj} are $\boldsymbol{\mu}_l$ and Σ_l, respectively. We want to test if the X_{lj}'s come from bivariate normal populations, and if the Σ_l's are equal. Let's write

$$\begin{aligned}
&\boldsymbol{\mu}_l = (\mu_{l1}, \mu_{l2})', \ \Sigma_l = \begin{pmatrix} \sigma_{l11} & \sigma_{l12} \\ \sigma_{l12} & \sigma_{l22} \end{pmatrix}, l = 1, 2, \ldots, a, \\
&\boldsymbol{\sigma}_l = (\sigma_{l11}, \sigma_{l12}, \sigma_{l22})', \ \widetilde{\boldsymbol{\sigma}} = \begin{bmatrix} \boldsymbol{\sigma}_1 & \boldsymbol{\sigma}_2 & \cdots & \boldsymbol{\sigma}_a \end{bmatrix}, \\
&\text{and} \ \ \boldsymbol{\eta} = \widetilde{\boldsymbol{\sigma}} Q = \begin{bmatrix} \boldsymbol{\eta}_1 & \boldsymbol{\eta}_2 & \cdots & \boldsymbol{\eta}_a \end{bmatrix}, \qquad (2)
\end{aligned}$$

where Q is the $a \times a$ Helmert matrix; i.e., an orthogonal matrix with first column having all elements equal to $1/\sqrt{a}$. Thus $\widetilde{\boldsymbol{\sigma}}$ is a $3 \times a$ matrix containing all the variance and covariance parameters from the a groups. Furthermore, the $\boldsymbol{\sigma}_l$'s and $\boldsymbol{\eta}_l$'s are 3×1 vectors, $l = 1, 2, \ldots, a$, and the equality of the Σ_l's is equivalent to $\boldsymbol{\eta}_2 = \boldsymbol{\eta}_3 = \ldots = \boldsymbol{\eta}_a = \mathbf{0}$. We shall now introduce a smooth alternative to bivariate normality based on Legendre polynomials, and then derive the corresponding score test.

3 Smooth alternatives and score tests

We shall adopt the methodology in [2] to come up with the Legendre polynomials required to specify the smooth alternative to the bivariate normal distribution. Let b_i denote the i^{th} normalized Legendre polynomial on [0,1], $i = 1, 2, \ldots$, and let $b_0(u) = 1$ for $u \in [0, 1]$. Now define the function b_{ij} as:

$$b_{ij}(u_1, u_2) = b_i(u_1) b_j(u_2), \quad u_1, u_2 \in [0, 1].$$

The collection $\{b_{ij};\ i, j = 0, 1, 2, \ldots\}$ can be arranged as an ordered sequence, say $\tilde{\gamma}$, using the following rule: the function b_{ij} appears in $\tilde{\gamma}$ before the function b_{lk} if one of the following conditions is fulfilled:

i) $i + j < l + k$,

ii) $i + j = l + k$ and $\max(i, j) > \max(l, k)$,

iii) $i + j = l + k$, $\max(i, j) = \max(l, k)$ and $i > l$.

Denote the i^{th} element of $\tilde{\gamma}$ by $\tilde{\gamma}_i$. The first 20 elements are given by

$$\begin{aligned}
&\tilde{\gamma}_0 = b_{00}, \ \tilde{\gamma}_1 = b_{10}, \ \tilde{\gamma}_2 = b_{01}, \ \tilde{\gamma}_3 = b_{20}, \ \tilde{\gamma}_4 = b_{02}, \ \tilde{\gamma}_5 = b_{11}, \\
&\tilde{\gamma}_6 = b_{30}, \ \tilde{\gamma}_7 = b_{03}, \ \tilde{\gamma}_8 = b_{21}, \ \tilde{\gamma}_9 = b_{12}, \ \tilde{\gamma}_{10} = b_{40}, \ \tilde{\gamma}_{11} = b_{04}, \ \tilde{\gamma}_{12} = b_{31}, \ \tilde{\gamma}_{13} = b_{13}, \\
&\tilde{\gamma}_{14} = b_{22}, \ \tilde{\gamma}_{15} = b_{50}, \ \tilde{\gamma}_{16} = b_{05}, \ \tilde{\gamma}_{17} = b_{41}, \ \tilde{\gamma}_{18} = b_{14}, \ \tilde{\gamma}_{19} = b_{32}, \ \tilde{\gamma}_{20} = b_{23}, \qquad (3)
\end{aligned}$$

where we note that each $\tilde{\gamma}$ is a function of $u_1, u_2 \in [0, 1]$.

For $X = (X_1, X_2)' \sim N(\boldsymbol{\mu}, \Sigma)$, let $\boldsymbol{\mu} = (\mu_1, \mu_2)'$, $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$, and $\boldsymbol{\sigma} = (\sigma_{11}, \sigma_{12}, \sigma_{22})'$. Let M be a lower triangular matrix with positive diagonal elements such that $MM' = \Sigma^{-1}$. The matrix

M can be explicitly written as

$$M = M(\boldsymbol{\sigma}) = \begin{pmatrix} m_1 & 0 \\ m_2 & m_3 \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{\sigma_{22}}{|\Sigma|}} & 0 \\ \frac{-\sigma_{12}}{\sqrt{\sigma_{22}|\Sigma|}} & \frac{1}{\sqrt{\sigma_{22}}} \end{pmatrix}. \tag{4}$$

For $(y_1, y_2)' = M'(x - \boldsymbol{\mu})$, let $u_1 = \Phi(y_1)$ and $u_2 = \Phi(y_2)$, where $\Phi(.)$ is the standard normal cdf. With $\tilde{\gamma}_i$'s defined in (3), let $\gamma_i(x; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \tilde{\gamma}_i(u_1, u_2) = \tilde{\gamma}_i(\Phi(y_1), \Phi(y_2))$. It is now easy to verify that

$$\int_{R^2} \gamma_i(x; \boldsymbol{\mu}, \boldsymbol{\sigma}) f(x; \boldsymbol{\mu}, \boldsymbol{\sigma}) dx = 0, \;\; i = 1, 2, ...$$

$$\int_{R^2} \gamma_i(x; \boldsymbol{\mu}, \boldsymbol{\sigma}) \gamma_j(x; \boldsymbol{\mu}, \boldsymbol{\sigma}) f(x; \boldsymbol{\mu}, \boldsymbol{\sigma}) dx = \delta_{ij}, \;\; i, j = 1, 2, ...$$

where δ_{ij} is the Kronecker delta, and $f(x; \boldsymbol{\mu}, \boldsymbol{\sigma})$ denotes the density function of the bivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$. Thus the $\gamma_i(x; \boldsymbol{\mu}, \boldsymbol{\sigma})$, i = 1, 2, ..., form an orthonormal set of polynomials under the bivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$. For the bivariate normal density $f(x; \boldsymbol{\mu}, \boldsymbol{\sigma})$, a smooth alternative of order k, say $g_k(x, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma})$, can now be defined using the $\gamma_i(x; \boldsymbol{\mu}, \boldsymbol{\sigma})$'s:

$$g_k(x; \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = C(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \exp\left\{ \sum_{i=1}^{k} \theta_i \gamma_i(x; \boldsymbol{\mu}, \boldsymbol{\sigma}) \right\} f(x; \boldsymbol{\mu}, \boldsymbol{\sigma}), \tag{5}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)'$ and $C(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ is a normalizing constant.

The case of different θ_l-vectors across groups

Let $\boldsymbol{\theta}_l = (\theta_{l1}, \theta_{l2}, ..., \theta_{lk_l})'$ be a $k_l \times 1$ vector that will be used to define the smooth alternative for the bivariate data from the lth group. We recall that for l = 1, 2, ..., a, the bivariate data from the lth group is denoted by X_{lj}, j = 1, 2, ..., n_l. If L denotes the likelihood function for all the groups, then the log-likelihood function is given by

$$\ln L = \sum_{l=1}^{a} \left[\sum_{j=1}^{n_l} \ln C(\boldsymbol{\theta}_l, \boldsymbol{\mu}_l, \boldsymbol{\sigma}_l) + \sum_{j=1}^{n_l} \sum_{i=1}^{k_l} \theta_{li} \gamma_{lji} + \sum_{j=1}^{n_l} \ln f(x_{lj}; \boldsymbol{\mu}_l, \boldsymbol{\sigma}_l) \right], \tag{6}$$

where γ_{lji} is $\gamma_i(x; \boldsymbol{\mu}_l, \boldsymbol{\sigma}_l)$ realized at the j^{th} observation in the l^{th} group of the bivariate random vector $X = (X_1, X_2)'$; i.e., $\gamma_{lji} = \gamma_i(x_{lj}; \boldsymbol{\mu}_l, \boldsymbol{\sigma}_l)$. Let $\boldsymbol{\eta}*$ be the $3 \times (a-1)$ matrix defined as

$$\boldsymbol{\eta}* = \begin{bmatrix} \boldsymbol{\eta}_2 & \boldsymbol{\eta}_3 & \cdots & \boldsymbol{\eta}_a \end{bmatrix}, \tag{7}$$

where the $\boldsymbol{\eta}_l$'s are defined in (2). The null hypothesis of bivariate normality and homoscedasticity is given by

$$H_0 : \;\; \boldsymbol{\theta}_l = \mathbf{0} \;\; (l = 1, 2, ..., a), \quad \text{and} \quad \boldsymbol{\eta}_l = \mathbf{0} \;\; (l = 2, 3, ..., a). \tag{8}$$

Recall that the $\boldsymbol{\sigma}_l$'s in (6) are functions of $\boldsymbol{\eta}_l$, l = 1, 2, ..., a; see (2).

We now derive the score vector and its variance-covariance matrix. For this, let's define $M(\boldsymbol{\sigma}_l)$ similar to $M(\boldsymbol{\sigma})$ in (4) with $\boldsymbol{\sigma}$ replaced with $\boldsymbol{\sigma}_l$, and let

$$\tilde{M}(\boldsymbol{\sigma}_l) = \begin{pmatrix} m_{l1} & 0 & 0 & 0 & 0 \\ m_{l2} & m_{l3} & 0 & 0 & 0 \\ 0 & 0 & m_{l1}^2 & 0 & 0 \\ 0 & 0 & m_{l2}^2 & m_{l3}^2 & m_{l2}m_{l3} \\ 0 & 0 & 2m_{l1}m_{l2} & 0 & m_{l1}m_{l3} \end{pmatrix} = \begin{pmatrix} M(\boldsymbol{\sigma}_l) & 0 \\ 0 & M_1(\boldsymbol{\sigma}_l) \end{pmatrix}, \tag{9}$$

where $M(\boldsymbol{\sigma}_l)$ and $M_1(\boldsymbol{\sigma}_l)$ are 2×2 and 3×3 principal submatrices of $\tilde{M}(\boldsymbol{\sigma}_l)$. We note that

$$\frac{\partial \ln L}{\partial \theta_{lr}} = \sum_{j=1}^{n_l} \gamma_{ljr}, \quad l = 1, 2, ..., a; \ r = 1, 2, ..., k$$

Furthermore,

$$\partial_{(\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l)} \ln L = \begin{pmatrix} \frac{\partial \ln L}{\partial \mu_{l1}} \\ \frac{\partial \ln L}{\partial \mu_{l2}} \\ \frac{\partial \ln L}{\partial \sigma_{l11}} \\ \frac{\partial \ln L}{\partial \sigma_{l22}} \\ \frac{\partial \ln L}{\partial \sigma_{l12}} \end{pmatrix} = \tilde{M}(\boldsymbol{\sigma}_l) \sum_{j=1}^{n_l} \begin{pmatrix} Y_{1lj} \\ Y_{2lj} \\ \frac{1}{2}(Y_{1lj}^2 - 1) \\ \frac{1}{2}(Y_{2lj}^2 - 1) \\ Y_{1lj}Y_{2lj} \end{pmatrix} = \tilde{M}(\boldsymbol{\sigma}_l) \begin{pmatrix} \boldsymbol{t}_{l1} \\ \boldsymbol{t}_{l2} \end{pmatrix}, \tag{10}$$

where Y_{1lj} and Y_{2lj} are the components of $Y_{lj} = M(\boldsymbol{\sigma}_l)'(X_{lj} - \boldsymbol{\mu}_l) \sim N(0, I_2)$, and $\boldsymbol{t}_{l1}$ and $\boldsymbol{t}_{l2}$ are 2×1 and 3×1 vectors, respectively, given by

$$\boldsymbol{t}_{l1} = \sum_{j=1}^{n_l} \begin{pmatrix} Y_{1lj} \\ Y_{2lj} \end{pmatrix}, \ \boldsymbol{t}_{l2} = \sum_{j=1}^{n_l} \begin{pmatrix} \frac{1}{2}(Y_{1lj}^2 - 1) \\ \frac{1}{2}(Y_{2lj}^2 - 1) \\ Y_{1lj}Y_{2lj} \end{pmatrix}.$$

It is easily verified that the covariance matrix of $\sum_{j=1}^{n_l} \left(Y_{1lj}, Y_{2lj}, \frac{1}{2}(Y_{1lj}^2 - 1), \frac{1}{2}(Y_{2lj}^2 - 1), Y_{1lj}Y_{2lj}\right)'$, say $\tilde{I}_{ll}$, is

$$\tilde{I}_{ll} = n_l \text{diag}(I_2, 1/2, 1/2, 1). \tag{11}$$

We now give the expression for the partial derivative of $\ln L$ with respect to $\text{vec}(\boldsymbol{\eta})$. From (9) and (10), we note that

$$\frac{\partial \ln L}{\partial \text{vec}(\widetilde{\boldsymbol{\sigma}})} = \tilde{M}_1(\widetilde{\boldsymbol{\sigma}})(\boldsymbol{t}_{12}', \boldsymbol{t}_{22}', \cdots, \boldsymbol{t}_{a2}')',$$

where $\tilde{M}_1(\widetilde{\boldsymbol{\sigma}}) = \text{diag}\left[M_1(\boldsymbol{\sigma}_1), M_1(\boldsymbol{\sigma}_2), ..., M_1(\boldsymbol{\sigma}_a)\right]$, and $\widetilde{\boldsymbol{\sigma}}$ is the $3\times a$ matrix defined in (2). Then

$$\begin{aligned} \frac{\partial \ln L}{\partial \text{vec}(\boldsymbol{\eta})} &= \left[\frac{\partial \text{vec}(\widetilde{\boldsymbol{\sigma}})}{\partial \text{vec}(\boldsymbol{\eta})}\right]' \frac{\partial \ln L}{\partial \text{vec}(\widetilde{\boldsymbol{\sigma}})} \\ &= \left[\frac{\partial \text{vec}(\widetilde{\boldsymbol{\sigma}})}{\partial \text{vec}(\boldsymbol{\eta})}\right]' (\boldsymbol{t}_{12}', \boldsymbol{t}_{22}', \cdots, \boldsymbol{t}_{a2}')' \\ &= (Q' \otimes I_3)\tilde{M}_1(\widetilde{\boldsymbol{\sigma}})(\boldsymbol{t}_{12}', \boldsymbol{t}_{22}', \cdots, \boldsymbol{t}_{a2}')'. \end{aligned} \tag{12}$$

Note that the score vector for $\text{vec}(\boldsymbol{\eta})$ is a $3a\times 1$ vector; the first 3 elements correspond to $\boldsymbol{\eta}_1(3\times 1)$, the next 3 elements correspond to $\boldsymbol{\eta}_2(3\times 1)$, and so on. Thus the components of the score vector for testing our null hypothesis consist of

$$\begin{aligned} \frac{\partial \ln L}{\partial \theta_{lr}} &= \sum_{j=1}^{n_l} \gamma_{ljr}, \quad l = 1, 2, ..., a; \ r = 1, 2, ..., k_l, \\ \frac{\partial \ln L}{\partial \boldsymbol{\eta}_c} &= \sum_{l=1}^{a} q_{lc} M_1(\boldsymbol{\sigma}_l)\boldsymbol{t}_{l2}, \quad c = 2, 3, ..., a, \end{aligned} \tag{13}$$

where the second expression given above is obtained from (12).

Now we derive the elements of the covariance matrix among the scores. It is straightforward to see that

$$\text{Cov}\left(\frac{\partial \ln L}{\partial \theta_{lr}}, \frac{\partial \ln L}{\partial \theta_{ls}}\right) = n_l, \quad \text{for } r = s; \text{ and } = 0 \text{ for } r \neq s.$$

Let $I_{\theta\theta}$ denote the covariance matrix among $\frac{\partial \ln L}{\partial \theta_{lr}}$, $l = 1, 2, ..., a$; $r = 1, 2, ..., k_l$, and let $I_{\eta\eta}$ be similarly defined . Thus

$$I_{\theta\theta} = \text{diag}\left(n_1 I_{k_1}, n_2 I_{k_2}, ..., n_a I_{k_a}\right). \tag{14}$$

In view of (9)−(13) we also have,

$$\begin{aligned} I_{\eta\eta}(3a \times 3a) &= E\Big[\left(\frac{\partial \ln L}{\partial \text{vec}(\boldsymbol{\eta})}\right)\left(\frac{\partial \ln L}{\partial \text{vec}(\boldsymbol{\eta})}\right)'\Big] \\ &= (Q' \otimes I_3)\left(D_n \otimes \{M_1(\boldsymbol{\sigma})\ \text{diag}(1/2, 1/2, 1) M_1'(\boldsymbol{\sigma})\}\right)(Q \otimes I_3) \\ &= (Q' D_n Q) \otimes \{M_1(\boldsymbol{\sigma}) \text{diag}(1/2, 1/2, 1) M_1'(\boldsymbol{\sigma})\}, \end{aligned} \tag{15}$$

where $M_1(\boldsymbol{\sigma})$ denotes the value of $M_1(\boldsymbol{\sigma}_l)$ in (9) when the $\boldsymbol{\sigma}_l$'s are all equal, having a common value $\boldsymbol{\sigma}$ (i.e., under the null hypothesis), and

$$D_n = \text{diag}(n_1, n_2, ..., n_a). \tag{16}$$

Similar to $I_{\theta\theta}$ and $I_{\eta\eta}$, let $I_{\mu\mu}$, $I_{\mu\eta}$, $I_{\mu\theta}$ and $I_{\eta\theta}$ denote variances and covariances among the scores corresponding to the specified parameters (evaluated under the null hypothesis). We can see that $I_{\eta\mu} = \mathbf{0}$ and

$$I_{\mu\mu}(2a \times 2a) = D_n \otimes [M(\boldsymbol{\sigma}) M'(\boldsymbol{\sigma})], \tag{17}$$

where $M(\boldsymbol{\sigma})$ denotes the value of $M(\boldsymbol{\sigma}_l)$ in (9) when the $\boldsymbol{\sigma}_l$'s are all equal.

We next derive $I_{\mu\theta}$ and $I_{\eta\theta}$. For this, we shall first compute covariances among the components of $\partial_{(\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l)} \ln L$ in (10) and $\sum_{j=1}^{n_l} \gamma_{ljr}$ in (13) for $r = 1, 2, ..., k_l$, and for each fixed $l = 1, 2, ..., a$. Since $\partial_{(\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l)} \ln L = \tilde{M}(\boldsymbol{\sigma}_l) \begin{pmatrix} \boldsymbol{t}_{l1} \\ \boldsymbol{t}_{l2} \end{pmatrix}$, as noted in (10), where $\boldsymbol{t}_{l1}$ and $\boldsymbol{t}_{l2}$ are defined below equation (10), we have to compute the covariances between the components of the 5×1 vector $\begin{pmatrix} \boldsymbol{t}_{l1} \\ \boldsymbol{t}_{l2} \end{pmatrix}$ and $\sum_{j=1}^{n_l} \gamma_{ljr}$, and these covariances are free of any unknown parameters. In order to exhibit these covariances, let us write

$$\text{Cov}\left[\begin{pmatrix} \boldsymbol{t}_{l1} \\ \boldsymbol{t}_{l2} \end{pmatrix}, \sum_{j=1}^{n_l} \gamma_{ljr}\right] = n_l A_l = n_l \begin{pmatrix} A_l^{(1)} \\ A_l^{(2)} \end{pmatrix}, \tag{18}$$

where $A_l = ((a_{lij}))$ is a $5 \times k_l$ matrix, and the k_l's can be different, for $l = 1, 2, ..., a$, and $A_l^{(1)}$ and $A_l^{(2)}$ are the matrices consisting of the first two rows and the last three rows of A_l, respectively. The elements of the matrix A_l involve the constants c_i's and the e_i's defined below:

$$\begin{aligned} c_i &= \text{Cov}[b_i(\Phi(Z)), Z] = \int_{-\infty}^{\infty} b_i(\Phi(z)) z f(z) dz, \\ e_i &= \text{Cov}[b_i(\Phi(Z)), Z^2] = \int_{-\infty}^{\infty} b_i(\Phi(z)) z^2 f(z) dz, \end{aligned}$$

see [2] and [21]. It is easily verified that $c_i = 0$ for i even, and $e_i = 0$ for i odd (this follows from the expressions for the Legendre polynomials). The values of several non-zero c_is and e_is are given in [3], Appendix 6.2, and these are also reproduced in [21], Section 2.1. Since some of the c_i's and e_i's are

zero, only a few of the a_{lij}'s are non-zeros. The non-zero elements are given below, corresponding to $k_l = 20$ for all l. The a_{lij}'s do not depend on l, and below we have used the notation $a_{i,j}$:

$$a_{1,1} = a_{2,2} = c_1; \quad a_{1,6} = a_{2,7} = c_3; \quad a_{1,15} = a_{2,16} = c_5;$$
$$a_{3,3} = a_{4,4} = \frac{1}{2}e_2; \quad a_{3,10} = a_{4,11} = \frac{1}{2}e_4; \quad a_{5,5} = c_1^2; \quad a_{5,12} = a_{5,13} = c_1 c_3.$$

The above elements correspond respectively to

$$\tilde{\gamma}_1, \tilde{\gamma}_2, \quad \tilde{\gamma}_6, \tilde{\gamma}_7, \quad \tilde{\gamma}_{15}, \tilde{\gamma}_{16},$$

$$\tilde{\gamma}_3, \tilde{\gamma}_4, \quad \tilde{\gamma}_{10}, \tilde{\gamma}_{11}, \quad \tilde{\gamma}_5, \quad \tilde{\gamma}_{12}, \tilde{\gamma}_{13}$$

of the sequence $\tilde{\gamma}_1, \tilde{\gamma}_2, ...$; see (3). Thus we get

$$\text{Cov}\left(\partial_{(\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l)} \ln L, \sum_{j=1}^{n_l} \gamma_{ljr}\right) = n_l \tilde{M}(\boldsymbol{\sigma}_l) A_l.$$

Recall that $\frac{\partial \text{vec}(\tilde{\boldsymbol{\sigma}})}{\partial \text{vec}(\boldsymbol{\eta})} = Q' \otimes I_3$. Thus we finally have

$$\begin{aligned} I_{\mu,\theta} &= \text{diag}\left(n_1 A_1^{(1)}, n_2 A_2^{(1)}, ..., n_a A_a^{(1)}\right) = (D_n \otimes I_2) A^{(1)} \\ I_{\eta,\theta} &= (Q' \otimes I_3)\text{diag}\left(n_1 A_1^{(2)}, n_2 A_2^{(2)}, ..., n_a A_a^{(2)}\right) = (Q' D_n \otimes I_3) A^{(2)}, \end{aligned} \tag{19}$$

where $A^{(1)} = \text{diag}\left(A_1^{(1)}, A_2^{(1)}, ..., A_a^{(1)}\right)$ and $A^{(2)} = \text{diag}\left(A_1^{(2)}, A_2^{(2)}, ..., A_a^{(2)}\right)$ are $2a \times \tilde{k}$ and $3a \times \tilde{k}$ matrices, respectively. The matrices that form the blocks of the covariance matrix of the score vector are thus given by

$$\begin{aligned} I_{\theta\theta} &= \text{diag}\left(n_1 I_{k_1}, n_2 I_{k_2}, ..., n_a I_{k_a}\right), \\ I_{\eta\eta} &= (Q' D_n Q) \otimes \{M_1(\boldsymbol{\sigma})\text{diag}(1/2, 1/2, 1) M_1'(\boldsymbol{\sigma})\} \\ I_{\mu\mu} &= D_n \otimes [M(\boldsymbol{\sigma}) M'(\boldsymbol{\sigma})], \quad I_{\eta\mu} = 0 \\ I_{\mu,\theta} &= (D_n \otimes I_2) M(\boldsymbol{\sigma}) A^{(1)} \\ I_{\eta,\theta} &= (Q' D_n \otimes I_3) M_1(\boldsymbol{\sigma}) A^{(2)}, \end{aligned} \tag{20}$$

where we recall that $D_n = \text{diag}(n_1, n_2, ..., n_a)$.

The score statistic can now be constructed, and the associated asymptotic chisquare distribution has df $= \tilde{k} + 3(a-1)$, where $\tilde{k} = \sum_{l=1}^{a} k_l$, for k_l fixed. The df is simply the number of parameters involved in the null hypothesis (8). Estimates (MLEs) of $\boldsymbol{\mu}_l$ ($l = 1, 2, ..., a$) and an estimate of the common covariance matrix are necessary for computing the score statistic. The former are given by the sample mean vectors from each group, and the MLE of the common covariance matrix is obtained by pooling the samples from the a groups. As noted by [21] in the case of the univariate fixed effects model, the test statistic has a distribution that is free of any unknown parameters. Since the testing problem is invariant under the group of location-scale transformations, this property is to be expected. Thus the required percentiles can be estimated by Monte Carlo simulation when the sample sizes are not large enough to appeal to the large sample theory.

The case of common θ_l-vectors across groups

We now define the smooth alternative of order k using a common $k \times 1$ vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)'$ across the different groups. The log likelihood function is now given by

$$\ln L = \sum_{l=1}^{a} \left[\sum_{j=1}^{n_l} \ln C(\boldsymbol{\theta}, \boldsymbol{\mu}_l, \boldsymbol{\sigma}_l) + \sum_{i=1}^{k} \sum_{j=1}^{n_l} \theta_i \gamma_{lji} + \sum_{j=1}^{n_l} \ln f(x_{lj}; \boldsymbol{\mu}_l, \boldsymbol{\sigma}_l) \right], \tag{21}$$

where γ_{lji}'s are defined as in the previous sub-section. The null hypothesis of bivariate normality and homoscedasticity is given by

$$H_0 : \ \boldsymbol{\theta} = \mathbf{0} \ \text{ and } \ \boldsymbol{\eta}_l = \mathbf{0} \ (l = 2, 3, ..., a). \tag{22}$$

We shall use the same notations as in the previous sub-section, and the derivation of the score test is very similar with appropriate modifications to take into account the presence of the common $\boldsymbol{\theta}-$vector in the smooth alternative for the different groups. The scores with respect to $\boldsymbol{\theta}=(\theta_1, ..., \theta_k)'$ and $\boldsymbol{\eta}_c$ $(c = 2, 3, ..., a)$ are

$$\begin{aligned} \frac{\partial \ln L}{\partial \theta_r} &= \sum_{l=1}^{a} \sum_{j=1}^{n_l} \gamma_{ljr}, \quad r = 1, 2, ..., k \\ \frac{\partial \ln L}{\partial \boldsymbol{\eta}_c} &= \sum_{l=1}^{a} q_{lc} M_1(\boldsymbol{\sigma}_l) \boldsymbol{t}_{l2}, \quad c = 2, 3, ..., a, \end{aligned} \tag{23}$$

where the various quantities are defined in the previous sub-section. Let A_k be a $5 \times k$ matrix of constants, similar to the A_l matrix defined in (18), and let $A_k^{(1)}$ and $A_k^{(2)}$ be matrices consisting of the first two rows and the last three rows, respectively, of A_k. The matrices that form the blocks of the covariance matrix of the score vector are given by:

$$\begin{aligned} I_{\theta\theta} &= N I_k \\ I_{\eta\eta} &= (Q' D_n Q) \otimes \{M_1(\sigma) \ \text{diag}(1/2, 1/2, 1) M_1'(\sigma)\} \\ I_{\mu\mu} &= D_n \otimes [M(\boldsymbol{\sigma}) M'(\boldsymbol{\sigma})], \quad I_{\eta\mu} = 0 \\ I_{\mu\theta} &= (n_1, n_2, ..., n_a)' \otimes [M(\boldsymbol{\sigma}) A_k^{(1)}] \\ I_{\eta\theta} &= (Q' \otimes I_3)\{(n_1, n_2, ..., n_a)' \otimes [M_1(\boldsymbol{\sigma}) A_k^{(2)}]\}, \end{aligned} \tag{24}$$

where $N = \sum_{l=1}^{a} n_l$, and the other quantities are as defined in the previous sub-section. The score statistic can now be computed. For a fixed k, the large sample chisquare distribution associated with the score statistic now has df = $k + 3(a-1)$, which is the number of parameters involved in the null hypothesis (22).

We note that in all the cases of testing bivariate normality and homoscedasticity considered above, the distribution of the score statistic is free of any unknown parameters. Thus the required percentiles to carry out the test can be estimated by Monte Carlo simulation. Furthermore, a data driven choice can be developed for choosing the order of the polynomials in the smooth alternative, similar to what is noted in [21]. We shall now comment on the data driven choice of the orders. The type I error probabilities and powers that we have reported in the next section, and the illustrative examples given later, are all based on tests that use data driven choices of the orders.

In the paper by [2] on smooth tests for bivariate normality, two modified BIC selection rules were described for choosing the order of the polynomial under Legendre polynomial-based alternatives.

One choice is based on allowing the order k to vary from 1 to a specified upper bound $d(n)$, and a second choice is based on allowing k to vary from 5 to $d(n)$. Let k_0 denote a specified lower bound for k. [2] showed that the score statistic for bivariate normality converges in distribution to χ^2_1 when $k_0 = 1$, and χ^2_5 when $k_0 = 5$. Furthermore, the author provided results of a power study that showed that the test using $k_0 = 5$ is in general much more powerful than the test using $k_0 = 1$. Based on simulations [2] noted that as a result of the estimation of the five unknown parameters (in the mean vector and covariance matrix), the first 5 components of score are very small compared to the rest; as a consequence, the data driven procedure using $k_0 = 1$ usually does not choose k=2,3,4 or 5, which explains the difference between the powers for the two selection rules.

In the numerical results reported in the next section, and for the examples discussed later, we shall use data driven choices of the orders based on the guidelines given above.

4 Numerical results

In this section, we shall report estimated type I error probabilities in order to assess the performance of the different tests. In particular, we shall report estimated type I error probabilities when the tests are implemented using the large sample chisquare distribution. We note that when the chisquare distribution is not an accurate approximation for the distribution of the score statistic, the test can be implemented using estimated percentiles, in view of the fact that the null distribution of the score statistic is free of any unknown parameters.

For our numerical results, the choice of a (the number of groups), and that of the n_ls (the replications), are the same as those used in [21] in the univariate case. Thus we have chosen a = 3, 6, 9 and 18 groups. Regarding the choice of the n_ls, we have considered both balanced and unbalanced data situations, once again following [21]. In the case of balanced data, the n_ls have a common value, denoted by n, and the choices considered for the simulation are n = 15, 30, 60, and 90. Before we explain the choice of the n_ls in the unbalanced case, we want to point out that one of the possibilities we want to explore in the unbalanced case is to use the critical value corresponding to the balanced data situation with n replaced by the harmonic mean of the n_ls. Thus the choices for the unbalanced case are made as follows. For each value of a, we have chosen the n_ls in such a way that their harmonic mean, say $\tilde{n}$, coincides with one of the values of n in the balanced case, namely, n = 15, 30, 60, and 90. For a = 3, we chose the n_ls to be 10, 15 and 30, resulting in the harmonic mean $\tilde{n}$ = 15. For a=3, in order to have $\tilde{n}$ equal to the other choices of n in the balanced case, namely, $\tilde{n}$ = 30, 60 and 90, we simply multiplied the choices 10, 15 and 30 for the n_ls with 2, 4 and 6, respectively. The same strategy was adopted for a = 6, 9 and 18. Thus for a = 6, we first chose the n_ls to be 10, 10, 15, 15, 30, 30, yielding $\tilde{n}$ = 15. Such a choice of the n_ls was then multiplied by 2, 4 and 6, yielding $\tilde{n}$ = 30, 60 and 90, respectively. As already noted, in the unbalanced case we explored the possibility of using the critical value corresponding to the balanced data situation with $n = \tilde{n}$ for each group.

When we use a Legendre polynomial-based alternative with $k_0 = 1$, the df associated with the asymptotic chisquare distribution of the score statistic is $1+3(a-1) = 3a-2$ in the common $\boldsymbol{\theta}_l$ set up, and the df is $a+3(a-1) = 4a-3$ in the unequal $\boldsymbol{\theta}_l$ scenario. For the choice $k_0 = 5$, the corresponding dfs are, respectively, $5+3(a-1) = 3a+2$ and $5a+3(a-1) = 8a-3$. Throughout, we have implemented the data driven approach with $d(n) = 15$. Without loss of generality, we have chosen $\boldsymbol{\mu}_l = 0$ for each l. Furthermore, when homoscedasticity holds, we have assumed that the common value of the covariance matrices is the identity matrix. Throughout, we have used 10^4 simulations, and a 5% significance level. Note that we have a total of four scenarios for evaluating the type I error probabilities: smooth alternatives with common $\boldsymbol{\theta}_l$ or different $\boldsymbol{\theta}_l$, with $k_0 = 1$ and

				P (Type I error)		95th Percentile		P (Type I
a	n	df	$\chi^2_{(df,0.95)}$	Balanced	Unbalanced	Balanced	Unbalanced	error*)
3	15	7	14.067	0.048	0.049	13.975	13.975	0.050
3	30	7	14.067	0.047	0.043	13.931	13.712	0.045
3	60	7	14.067	0.049	0.049	14.015	14.010	0.050
3	90	7	14.067	0.049	0.050	14.009	14.067	0.051
6	15	16	26.296	0.060	0.063	27.020	27.347	0.054
6	30	16	26.296	0.058	0.056	26.888	26.771	0.049
6	60	16	26.296	0.055	0.056	26.656	26.716	0.051
6	90	16	26.296	0.053	0.053	26.596	26.502	0.049
9	15	25	37.652	0.078	0.073	40.038	39.836	0.048
9	30	25	37.652	0.065	0.059	38.893	38.560	0.047
9	60	25	37.652	0.059	0.057	38.319	38.150	0.048
9	90	25	37.652	0.055	0.054	38.075	38.029	0.050
18	15	52	69.832	0.101	0.089	75.231	73.666	0.041
18	30	52	69.832	0.075	0.068	72.323	72.173	0.049
18	60	52	69.832	0.065	0.057	71.562	70.696	0.043
18	90	52	69.832	0.058	0.055	70.718	70.363	0.048

Table 1: Type I error probabilities of the score tests and critical values of the score statistics under smooth alternatives with a common θ_l when k_0=1 and a 5% significance level; $\chi^2_{(df,0.95)}$ denotes the 95th percentile of the chisquare distribution with the specified df.

$k_0 = 5$ in each case. Table 1– Table 4 give the following numerical results: (i) the estimated type I error probabilities in the balanced and unbalanced data situations when the asymptotic chi-square distribution is used to carry out the test (these results appear under "P(Type I error)" in the tables), (ii) the 95th percentiles of the test statistics estimated by Monte Carlo simulation, and (iii) the estimated type I error probabilities in the unbalanced case when the estimated percentile for the balanced case is used with $n = \tilde{n}$ (these results appear under "P(Type I error*)" in the tables). In the tables, the type I error probabilities in the unbalanced case are reported against a single value of n. As already noted, this value of n is actually the harmonic mean (denoted by $\tilde{n}$) of the n_ls that have been chosen in the unbalanced case.

From the numerical results in Table 1 – Table 4 we draw the following conclusions. The asymptotic chisquare approximation can be very poor, especially when the number of groups is large and the within-group replication is small. This is clear from several of the estimated type I error probabilities, and is also evident from the discrepancy between the chi-square percentile and the corresponding percentile estimated by Monte Carlo simulation. In such scenarios, the test should be implemented using the critical value estimated using Monte Carlo simulation. Interestingly, the use of the estimated critical value corresponding to $\tilde{n}$ in the balanced case turns out to be quite accurate in the unbalanced case, at least in the scenarios considered for simulation.

5 Two examples

In order to illustrate the tests we have developed, we shall use two examples.

Example 1

In this example, measurements are available on male Egyptian skulls from 5 epochs: 4000BC, 3300BC, 1850BC, 200BC, and AD150. Thirty skulls were measured from each time period, resulting in a total

				P (Type I error)		95th Percentile		P (Type I
a	n	df	$\chi^2_{(df,0.95)}$	Balanced	Unbalanced	Balanced	Unbalanced	error*)
3	15	11	19.675	0.062	0.061	20.575	20.527	0.050
3	30	11	19.675	0.053	0.052	19.911	19.826	0.049
3	60	11	19.675	0.053	0.053	19.893	19.877	0.050
3	90	11	19.675	0.054	0.051	19.950	19.785	0.048
6	15	20	31.410	0.067	0.068	32.815	32.886	0.051
6	30	20	31.410	0.061	0.059	32.216	32.297	0.051
6	60	20	31.410	0.057	0.057	31.897	32.004	0.052
6	90	20	31.410	0.056	0.055	31.882	31.886	0.050
9	15	29	42.557	0.085	0.075	45.740	44.982	0.043
9	30	29	42.557	0.064	0.061	43.969	43.466	0.046
9	60	29	42.557	0.060	0.057	43.525	43.258	0.047
9	90	29	42.557	0.056	0.057	43.060	43.277	0.052
18	15	56	74.468	0.101	0.090	80.103	78.783	0.042
18	30	56	74.468	0.076	0.072	77.381	77.149	0.047
18	60	56	74.468	0.065	0.059	76.097	75.659	0.047
18	90	56	74.468	0.058	0.057	75.459	75.234	0.049

Table 2: Type I error probabilities of the score tests and critical values of the score statistics under smooth alternatives with a common $\boldsymbol{\theta}_l$ when k_0=5 and a 5% significance level; $\chi^2_{(df,0.95)}$ denotes the 95th percentile of the chisquare distribution with the specified df.

				P (Type I error)		95th Percentile		P (Type I
a	n	df	$\chi^2_{(df,0.95)}$	Balanced	Unbalanced	Balanced	Unbalanced	error*)
3	15	9	16.919	0.072	0.083	18.644	19.715	0.061
3	30	9	16.919	0.053	0.054	17.232	17.183	0.049
3	60	9	16.919	0.049	0.050	16.873	16.897	0.051
3	90	9	16.919	0.053	0.054	17.112	17.144	0.051
6	15	21	32.671	0.111	0.119	38.955	40.034	0.057
6	30	21	32.671	0.070	0.071	34.373	34.777	0.053
6	60	21	32.671	0.060	0.064	33.363	33.856	0.056
6	90	21	32.671	0.055	0.058	33.181	33.198	0.050
9	15	33	47.400	0.150	0.151	58.492	60.914	0.060
9	30	33	47.400	0.084	0.082	51.207	50.686	0.047
9	60	33	47.400	0.066	0.065	48.925	49.012	0.051
9	90	33	47.400	0.061	0.061	48.436	48.326	0.049
18	15	69	89.391	0.221	0.211	113.772	114.292	0.052
18	30	69	89.391	0.108	0.105	96.663	97.141	0.054
18	60	69	89.391	0.078	0.071	93.075	92.142	0.045
18	90	69	89.391	0.067	0.062	91.603	91.065	0.047

Table 3: Type I error probabilities of the score tests and critical values of the score statistics under smooth alternatives with different $\boldsymbol{\theta}_l$s when k_0=1 and a 5% significance level; $\chi^2_{(df,0.95)}$ denotes the 95th percentile of the chisquare distribution with the specified df.

				P (Type I error)		95th Percentile		P (Type I
a	n	df	$\chi^2_{(df,0.95)}$	Balanced	Unbalanced	Balanced	Unbalanced	error*)
3	15	21	32.671	0.112	0.114	39.189	39.446	0.052
3	30	21	32.671	0.080	0.081	35.481	35.531	0.051
3	60	21	32.671	0.066	0.071	34.268	34.489	0.052
3	90	21	32.671	0.066	0.066	34.209	34.117	0.050
6	15	45	61.656	0.158	0.147	73.450	74.349	0.054
6	30	45	61.656	0.106	0.102	68.560	68.024	0.047
6	60	45	61.656	0.085	0.086	65.268	65.891	0.054
6	90	45	61.656	0.078	0.075	65.047	64.531	0.046
9	15	69	89.391	0.195	0.178	107.889	108.431	0.052
9	30	69	89.391	0.124	0.121	98.986	98.403	0.048
9	60	69	89.391	0.092	0.091	95.538	95.000	0.047
9	90	69	89.391	0.076	0.081	93.077	93.332	0.052
18	15	141	169.711	0.273	0.248	205.796	203.253	0.044
18	30	141	169.711	0.154	0.148	187.207	186.519	0.047
18	60	141	169.711	0.107	0.099	179.333	177.904	0.045
18	90	141	169.711	0.088	0.083	175.911	176.244	0.051

Table 4: Type I error probabilities of the score tests and critical values of the score statistics under smooth alternatives with different $\boldsymbol{\theta}_l$s when k_0=5 and a 5% significance level; $\chi^2_{(df,0.95)}$ denotes the 95th percentile of the chisquare distribution with the specified df.

of 150 observations. Measurements were obtained on the following five variables:

EPOCH: Approximate Year of Skull Formation
MB: Maximal Breadth of Skull
BH: Basi-bregmatic Height of Skull
BL: Basialveolar Length of Skull
NH: Nasal Height of Skull

The data set is available in the R library "HSAUR", and was originally given in [4]. The main problem of interest was whether the measurements changed over time. Non-constant measurements of the skulls over time would indicate interbreeding with immigrant populations. Our interest is to test the validity of the normality and homoscedasticity assumptions.

Since the methodology we have developed is for bivariate data, we shall carry out our test procedures using the data on just two variables. We shall use the data on the pair of variables MB and BH for illustration. Thus we have balanced data with $a = 5$ groups, and $n = 30$ bivariate observations per group. The results of our analysis are presented in Table 5. We have used a data driven choice for the value of the order k, and this choice is also given in the table. In the case of different $\boldsymbol{\theta}_l$'s across the five groups, the table gives the values of the data driven choice of the k_l's for the different groups. While implementing the data driven choice, we varied k between specified lower and upper limits k_0 and $d(n)$ respectively; the choice of k_0 is given in the table, and $d(n)$ was chosen to be 15. The simulated critical value of the score statistic is given in each case, along with the value of the score statistic. A few of the score statistics are also presented as the sum of two terms: the first term corresponds to the score statistic for testing bivariate normality (i.e., $\boldsymbol{\theta}_l = 0$) and the second term corresponds to testing homoscedasticity. For example, in the case of a common $\boldsymbol{\theta}$ across the five groups, the value of the score statistic is presented as 15.563 (5.042+10.520). The value

	Data driven k	Estimated critical value	Score statistic	Decision on H_0
Common $\boldsymbol{\theta}_l$, k_0=5	5	28.244	15.563 (5.042+10.520)	Don't Reject
Common $\boldsymbol{\theta}_l$, k_0=1	1	22.956	10.588 (0.068+10.520)	Don't Reject
Different $\boldsymbol{\theta}_l$, k_0=5	5,5,6,5,5	58.008	44.190	Don't Reject
Different $\boldsymbol{\theta}_l$, k_0=1	1,1,1,1,1	28.717	13.780	Don't Reject

Table 5: Bivariate normality and homoscedasticity test for Example 1 using a 5% significance level (Variables: maximal breadth and basibregmatic height)

	Data driven k	Estimated critical value	Score statistic	Decision on H_0
Common $\boldsymbol{\theta}_l$, k_0=5	5	19.911	35.114 (11.904+23.21)	Reject
Common $\boldsymbol{\theta}_l$, k_0=1	1	13.931	24.999 (1.789+23.21)	Reject
Different $\boldsymbol{\theta}_l$, k_0=5	10,6,5	35.481	52.884	Reject
Different $\boldsymbol{\theta}_l$, k_0=1	1,1,3	17.232	33.436	Reject

Table 6: Bivariate normality and homoscedasticity test for Example 2 using a 5% significance level (Variables: Insulin Resistance and Relative Weight)

5.042 corresponds to the score statistic for testing bivariate normality, and the second value 10.520 corresponds to the score statistic for testing homoscedasticity. Such a decomposition is possible here since the covariance matrix of the score statistic is a block diagonal matrix. It can be verified that the estimated critical values given in Table 5 are all very close to the 95th percentiles of the chisquare distributions with the appropriate dfs. The decision concerning the null hypothesis is indicated in Table 5.

In this example, all tests failed to reject, and there is no evidence of either non-normality or heteroscedasticity. It should be noted that evidence of multivariate normality was indicated in the original data source, and our analysis agrees with this.

Example 2

Our second example is based on a data set taken from [19], who investigated the effect of several variables on chemical diabetes and overt diabetes. The data set is reproduced in [1] and in [20]. There were 90 subjects in the study, belonging to three groups: normal, those with chemical diabetes, and those with overt diabetes. Each group had 30 subjects. For each of the 90 subjects, data were obtained on the variables glucose intolerance, insulin response to oral glucose and insulin resistance, along with relative weight and fasting plasma glucose. For more details, we refer to [19].

We shall illustrate our tests using data on one pairs of variables: Insulin Resistance and Relative Weight. The results of the tests are presented in Table 6. We note that all the tests result in rejection of the null hypothesis. Looking at the individual components of the score statistics, it appears that the rejection is mainly caused by heteroscedasticity.

6 Discussion

This article develops score tests for the simultaneous assessment of normality and homoscedasticity for bivariate grouped data based on a class of smooth alternatives defined through Legendre polynomials. In our work we have considered two scenarios for the smooth alternatives: one option that specifies a

common non-normal distribution for the different groups, and a second option that specifies different non-normal distributions for the different groups. In a practical application, the experimenter/data analyst has to decide which alternative is appropriate.

Even though we have considered only Legendre polynomial based smooth alternatives, one can also specify smooth alternatives based on Hermite polynomials; see [18] where smooth alternatives based on Hermite polynomials are used for testing normality. In the doctoral dissertation by [20], smooth tests are derived for simultaneously testing normality and homoscedasticity when the alternative is specified using both Legendre polynomials and Hermite polynomials. The numerical results in [20] indicate that for certain alternatives, smooth tests derived using Hermite polynomials have lower power compared to the tests derived under Legendre polynomial based smooth alternatives. This has also been noted in [2] for testing bivariate normality against smooth alternatives. Thus, in the present work, we have not included tests based on Hermite polynomial-based alternatives. We have provided details concerning the numerical implementation of our proposed tests, and have also illustrated them using two examples.

References

[1] D F Andrews and A M Herzberg. *Data*. New York: Springer, 1985.

[2] M Bogdan. "Data driven smooth tests for bivariate normality." In: *Journal of Multivariate Analysis* 68 (1999), pp. 26–53.

[3] M Bogdan. "Data driven smooth tests for bivariate normality, preprint IM PWr 42/96, Technical University of Wroclaw." In: (1996).

[4] D J Hand et al. *A handbook of small datasets*. London: Chapman and Hall/CRC, 1994.

[5] D M Hawkins. "A new test for multivariate normality and homoscedasticity." In: *Technometrics* 23 (1981), pp. 105–110.

[6] T Inglot, W Kallenberg, and T Ledwina. "Data driven smooth tests for composite hypotheses." In: *The Annals of Statistics* 25 (1997), pp. 1222–1250.

[7] T Inglot and T Ledwina. "Asymptotic optimality of data-driven Neyman's tests for uniformity." In: *The Annals of Statistics* 24 (1996), pp. 1982–2019.

[8] M Jamshidian and S Jalal. "Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data." In: *Psychometrika* 75 (2010), pp. 649–674.

[9] M Jamshidian, S Jalal, and C Jansen. "MissMech: An R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR)." In: *Journal of Statistical Software* 56 (2014), pp. 1–31.

[10] A Janic and T Ledwina. "Data-driven smooth tests for a location-scale family revisited." In: *Journal of Statistical Theory and Practice* 3 (2009), pp. 645–664.

[11] W Kallenberg and T Ledwina. "Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests." In: *The Annals of Statistics* 23 (1995), pp. 1594–1608.

[12] W Kallenberg and T Ledwina. "Data-driven smooth tests for composite hypotheses: Comparison of powers." In: *Journal of Statistical Computation and Simulation* 59 (1997), pp. 101–121.

[13] W Kallenberg and T Ledwina. "Data-driven smooth tests when the hypothesis is composite." In: *Journal of the American Statistical Association* 92 (1997), pp. 1094–1104.

[14] W Kallenberg, T Ledwina, and E Rafajlowicz. "Testing bivariate independence and normality." In: *Sankhyã, Series A* 59 (1997), pp. 42–59.

[15] T Ledwina. "Data-driven version of Neyman's smooth test of fit." In: *Journal of the American Statistical Association* 89 (1994), pp. 1000–1005.

[16] C J Mecklin and D J Mundfrom. "An appraisal and bibliography of tests for multivariate normality." In: *International Statistical Review* 72 (2004), pp. 1231–1238.

[17] J Neyman. "Smooth test for goodness of fit." In: *Skandinaviska Aktuarietidskrift* 20 (1937), pp. 149–199.

[18] J C W Rayner, O Thas, and D J Best. *Smooth tests of goodness of fit: Using R*. New York: John Wiley, 2009.

[19] G M Reaven and R G Miller. "An attempt to define the nature of chemical diabetes using a multidimensional analysis." In: *Diabetologia* 16 (1979), pp. 17–24.

[20] Y Yang. *The simultaneous assessment of normality and homoscedasticity in some linear models*. Doctoral dissertation: University of Maryland Baltimore County, 2016.

[21] Y Yang and T Mathew. "The simultaneous assessment of normality and homoscedasticity in linear fixed effects models." In: *Journal of Statistical Theory and Practice* 12 (2018), pp. 66–81.

[22] Y Yang and T Mathew. "The simultaneous assessment of normality and homoscedasticity in one-way random effects models." In: *Statistics and Applications* 18 (2020), pp. 97–119.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Heuristic Tests for Testing Against Ordered Alternatives in Heteroscedastic ANOVA

Anjana Mondal [1] and Somesh Kumar [2]

1. *Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur - 721302, India. Email: mondal.anjana25@gmail.com*
2. *Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur - 721302, India. Email: smsh@maths.iitkgp.ac.in*

Received: 24 Feb 2022 / Revised: 12 May 2022 / Accepted: 10 Jun 2022

Abstract

In a heteroscedastic one-way ANOVA model, a heuristic Max-T test is proposed for testing the null hypothesis of equality of treatment effects, that is, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ against the alternative hypothesis of ordered effects; $H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$ (with at least one strict inequality). Asymptotic distribution of Max-T test statistic is derived and another test, called asymptotic Max-T(AMax-T) is proposed, based on the asymptotic distribution. Parametric bootstrap is used for the tests based on Max-T and AMax-T statistics to evaluate critical points and estimate the size and power functions. Simulation studies show that both Max-T and AMax-T tests achieve size. Powers of Max-T and AMax-T tests are observed to be almost the same. Under the departure from normality, the robustness of both the tests is studied by finding powers and sizes for five non-normal distributions (Laplace, exponential, Weibull, log-normal and t-distribution). It is observed that the tests are more robust for large sample sizes. For the proposed tests R packages are developed, which makes our tests user friendly in real situations. A practical example is provided to illustrate the application of these tests.

Keywords: ANOVA; heteroscedasticity; power; critical points; robustness.

1 Introduction

Traditional one-way ANOVA considers testing equality of treatment effects against an alternative hypothesis of at least one inequality with homogeneous error variances. However, in certain situations, it is more sensible to have an alternative hypothesis in which mean treatment effects have a natural ordering. Such applications arise in dose response trials, environmental studies, agricultural and industrial experiments (see for example, Bretz [6], Chen [8], Shan et al. [22] or Williams [28]). We can design more powerful tests for such ordered alternatives.

Specifically we are interested in testing the equality of treatment effects, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ against an ordered alternative hypothesis, $H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$, with at least one strict inequality, where μ_i is the i-th treatment effect for $i = 1, 2, \cdots, k$. There are many studies where ordered alternatives are considered. However, group variances are taken to be either homogeneous and unknown or heterogeneous and known. Bartholomew [1] derived the likelihood ratio test for testing H_0 against H_1 when variances are known. Shorack [24] extended Bartholomew's test to two-way ANOVA and general linear models and also proposed a rank analogue of these.

William [28] developed a test for testing against $M_0 \leq M_1 \leq \cdots \leq M_k$ with at least one strict inequality that uses maximum likelihood estimators of mean responses of k treatments $M_i, i = 1, 2, \cdots, k$

© 2022 Author(s). (https://www.thegsa.in/).

and control M_0. He considered equal but unknown variances and equal treatment replications. William [29] modified the test of [28] for unequal treatment replications. Marcus [16] carried out the power comparisons of Bartholomew's tests [1], [2] and William's tests [28], [29] by finding exact powers for $k = 3$ and simulated powers for $k = 4, 6$ and 8.

Further, Mudholkar and McDermott [17] proposed a class of tests by combining p-values for testing equality of treatment effects against different types of order restricted hypotheses. They have taken equal and known variance. Sirley [23] proposed non-parametric versions of Williams' tests [28] and [29]. Moreover, Bretz [6] extended William's tests [28] and [29] for general linear models, which allow for covariates and factorial treatments. Cabiolio and Peng [7] extended Jonckheere's [14] test for testing equality of treatment effects against ordered alternatives in a randomized block experiment with incomplete blocks.

In all the above studies, the error variances are considered equal or unequal but known. However, the homogeneous variance or known variance conditions are not suitable in many practical situations (see, for example, Bishop and Dudewicz [4], [5], Chen [8], [9], Noguchi et al. [18] or Pauly et al. [20, 19]).

Under the condition of heteroscedastic and unknown variances, Bishop and Dudewicz [4] proposed one-stage and two-stage procedures for one-way and two-way ANOVA models, and Bishop and Dudewicz [5] extended those for the r-way ANOVA model. Hasler [11] proposed multiple contrast tests using multiple degrees of freedom and sandwich estimation procedure for testing equality of means in heteroscedastic ANOVA. Chen [8] proposed one stage and two stage procedures for testing H_0 against natural alternatives in one-way ANOVA with unequal and unknown variances. Herberich et al. [12] proposed a test based on multiple comparison procedures for testing equality of treatment effects in heteroscedastic one-way ANOVA model with unequal sample sizes. All these works consider the hypothesis of at least one inequality against equality of mean effects.

Krishnamoorthy et al. [15] proposed a parametric bootstrap test for fixed effects one-way ANOVA model with heteroscedastic error variances. They compared this test with the existing tests; Welch's test [27], James's test [13] and generalized F(GF) test [26]. From their simulation studies, it is observed that the parametric bootstrap test performs better than others. They noted that the proposed test is also applicable in the random-effects one-way ANOVA model with heteroscedastic variances. In one-way ANOVA model when variances are unequal and data are unbalanced, Zhang [30] considered multiple comparisons $H_0 : \mu_i = \mu_j$ vs. $H_\alpha : \mu_i \neq \mu_j$, $i, j = 1, 2, \cdots, k$, $i \neq j$. He proposed a parametric bootstrap test modifying Tukey's range test.

In this paper, we consider the extension of One-Way ANOVA in two directions: testing against ordered alternatives and unknown and unequal variances. Precisely, we consider testing equality of treatment effects i.e., $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, against ordered alternatives $H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$ with at least one inequality. We also consider unequal replications of k treatments. For the testing problem, we propose a simultaneous test based on pair- wise comparisons among treatment effects, called the Max-T test. Under the null hypothesis, the asymptotic distribution of the Max-T test statistic is derived, and another test, the asymptotic Max-T (AMax-T) test, is proposed. For practical usefulness, the parametric bootstrap approach for both Max-T and AMax-T is used. From extensive simulation studies, it is observed that both the tests perform well when all the treatment effects are in strictly

increasing order or increasing order with at least one equality. As AMax-T is an asymptotic test, it achieves the required size for moderate to large sample sizes. But Max-T performs well for small samples also. R packages are developed for applying the tests in real-life situations.

The article is organized as follows. In Section 2, we introduce the Max-T test. An algorithm for finding critical values using a parametric bootstrap approach for this test is given. Asymptotic distribution of the Max-T test statistic is derived in Section 3, and a test based on the asymptotic distribution is introduced there. The parametric bootstrap approach is used to find the critical value of the asymptotic Max-T test. The corresponding algorithm is given in Section 3. Asymptotic accuracy of parametric bootstrap of Max-T is also established. The proof also establishes the accuracy of the parametric bootstrap of AMax-T test. An algorithm for finding the size and power of the tests is provided in Section 4, and the simulated results are tabulated and analyzed there. In Section 5, we study the robustness of the proposed Max-T and AMax-T tests, by calculating size and power of five non-normal distributions. The proposed methods are applied to a real data set in Section 6.

2 Max-T test

Consider the fixed effect one-way ANOVA model:

$$X_{ij} = \mu_i + \epsilon_{ij},\ j = 1, 2, \cdots, n_i,\ i = 1, 2, \cdots, k,$$

where the errors ϵ_{ij}'s are independent and normally distributed, say $N(0, \sigma_i^2)$, and the variances σ_i^2 are unknown and unequal. We are interested in testing $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ against $H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$ with at least one strict inequality.

Define

$$T_i = \frac{\bar{X}_{i+1.} - \bar{X}_{i.}}{\sqrt{\frac{S_{i+1}^2}{n_{i+1}} + \frac{S_i^2}{n_i}}},\ i = 1, 2, \cdots, k-1 \tag{1}$$

where $\bar{X}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ and $S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2$ are the sample mean and the sample variance respectively of the ith sample for $i = 1, 2, \cdots, k$.

We propose the test statistic for H_0 against H_1 as

$$T_M = \max \underline{\mathbf{T}}, \tag{2}$$

where $\underline{\mathbf{T}} = (T_1, T_2, \cdots, T_{k-1})^T$. The hypothesis H_0 is rejected at level α if $T_M > d$ such that $\sup_{H_0} P(T_M > d) = \alpha$. As the distribution of the test statistic T_M under H_0 is complicated, we use parametric bootstrap (PB) for finding critical values. In PB, samples are generated from parametric models using estimators of parameters in place of original parameters. We introduce PB version of Max-T statistic below.

Note that under H_0, μ_i's are equal, say, μ. The statistic (1) is independent of μ, so without loss of generality, we can take $\mu = 0$. Now, we generate bootstrap samples X_{ij}^* from $N(0, S_i^2)$. Define

$$T_i^* = \frac{\bar{X}_{i+1}^* - \bar{X}_i^*}{\sqrt{\frac{S_{i+1}^{*2}}{n_{i+1}} + \frac{S_i^{*2}}{n_i}}},\quad i = 1, 2, \cdots, k-1, \tag{3}$$

where $\bar{X}_i^* = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}^*$ and $S_i^{*2} = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(X_{ij}^* - \bar{X}_{i.}^*)^2$ are the sample mean and the sample variance of the bootstrap sample $X_{ij}^*, j = 1, 2, \cdots, n_i,\ i = 1, 2, \cdots, k$. PB version of T_M is defined as follows

$$T_M^* = \max \underline{\mathbf{T}}^*, \tag{4}$$

where $\underline{\mathbf{T}}^* = (T_1^*, T_2^*, \cdots, T_{k-1}^*)^T$.

H_0 is rejected at level α if $T_M > d^*$, where the critical value d^* is obtained from the following algorithm.

Algorithm 1.

1. From given k independent normally distributed heteroscedastic samples $X_{ij}, j = 1, 2, \cdots, n_i;\ i = 1, 2, \cdots, k$, calculate sample variances S_i^2 for $i = 1, 2, \cdots, k$.
2. Generate parametric bootstrap samples X_{ij}^* from $N(0, S_i^2)$, $j = 1, 2, \cdots, n_i,\ i = 1, 2, \cdots, k$.
3. Calculate the values of T_i^* for $i = 1, 2, \cdots, k-1$ as defined in Equation (3). Then calculate the observed statistic value of T_M^*, defined in Equation (4).
4. Repeat Steps 2 and 3 a large number of times, say $H = 5000$. Rearrange the H values of T_M^* in increasing order, say $T_{M(1)}^* \le T_{M(2)}^* \le \cdots \le T_{M(H)}^*$ and take $(1-\alpha)$-th quantile as bootstrap critical value, say d^*,
$$d^* := T_{M([(1-\alpha)H])}^*,$$
$T_{M(h)}$ is the h-th order statistic of H values of T_M.

3 Asymptotic Max-T test

In this section, we define asymptotic Max-T test (AMax-T), which is based on the asymptotic distribution of Max-T test. First we find the asymptotic distribution of the statistic $\underline{\mathbf{T}}$ under H_0.

Consider the asymptotic setup

$$\frac{n_i}{N} \to r_i \text{ as } \min_{1\le i\le k} n_i \to \infty \text{ and } N \to \infty,\ i = 1, 2, \cdots, k,\ N = \sum_{i=1}^{k} n_i. \tag{5}$$

We rewrite T_i as

$$T_i = \frac{Z_i}{\hat{b}_i},$$

where $Z_i = \sqrt{\frac{n_i n_{i+1}}{N}}(\bar{X}_{i+1.} - \bar{X}_{i.})$ and $\hat{b}_i = \sqrt{\frac{n_i}{N}S_{i+1}^2 + \frac{n_{i+1}}{N}S_i^2}$.
As samples variances $S_i^2 \xrightarrow{P} \sigma_i^2$ as $n_i \to \infty$ for $i = 1, 2, \cdots, k$, we get

$$\hat{b}_i \xrightarrow{P} \sqrt{r_i\sigma_{i+1}^2 + r_{i+1}\sigma_i^2} = b_i(\text{say}).$$

When the null hypothesis H_0 holds true, under the asymptotic set up (5), the distribution of $\underline{\mathbf{Z}} = (Z_1, Z_2, \cdots, Z_{k-1})^T$ is $N_{k-1}(\underline{\mathbf{0}}, \mathbf{\Sigma}_z)$, with $\mathbf{\Sigma}_z$ given by

$$\mathbf{\Sigma}_z = \begin{bmatrix} b_1^2 & e_1 & 0 & \dots & 0 & 0 \\ e_1 & b_2^2 & e_2 & \dots & 0 & 0 \\ 0 & e_2 & b_3^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & e_{k-2} & b_{k-1}^2 \end{bmatrix}$$

and $e_i = -\sqrt{r_i r_{i+2}}\sigma_{i+1}^2$.

Let $\hat{\mathbf{B}} = diag\left(\hat{b}_1, \hat{b}_2, \cdots, \hat{b}_{k-1}\right)$ and $\mathbf{B} = diag\left(b_1, b_2, \cdots, b_{k-1}\right)$. Then under (5), $\hat{\mathbf{B}} \xrightarrow{P} \mathbf{B}$. Hence under H_0, the asymptotic distribution of $\underline{\mathbf{T}} = \hat{\mathbf{B}}^{-1}\underline{\mathbf{Z}}$ is $N_{k-1}(\underline{\mathbf{0}}, \mathbf{D})$ using multivariate Slutsky's theorem and continuous mapping theorem, where $\mathbf{D} = \mathbf{B}^{-1}\mathbf{\Sigma}_z\mathbf{B}^{-1}$.

From the asymptotic distribution of the vector $\underline{\mathbf{T}}$ we define an asymptotic test, called AMax-T. The test rejects H_0 if $T_M > f^*$, where f^* can be found from the following algorithm.

Algorithm 2.

1. From a given sample X_{ij}, $j = 1, \cdots, n_i$, $i = 1, \cdots, k$, which are independent, normally distributed with heterogeneous variances generate sample variances S_i^2, $i = 1, 2, \cdots, k$.
2. Compute estimator of matrix $\mathbf{\Sigma}_z$, say, $\hat{\mathbf{\Sigma}}_z$ using $S_1^2, S_2^2, \cdots, S_k^2$ in place of $\sigma_1^2, \sigma_2^2, \cdots, \sigma_k^2$. Similarly calculate $\hat{\mathbf{B}}$ from $\mathbf{B}$ by replacing $\sigma_1^2, \sigma_2^2, \cdots, \sigma_k^2$. Then calculate $\hat{\mathbf{D}} = \hat{\mathbf{B}}^{-1}\hat{\mathbf{\Sigma}}_z\hat{\mathbf{B}}^{-1}$.
3. Generate $\underline{\hat{\mathbf{T}}}_A^*$ from $N_{k-1}(\underline{\mathbf{0}}, \hat{\mathbf{D}})$. Now take the estimated test statistic value $\hat{T}_{AM}^* = \max \underline{\hat{\mathbf{T}}}_A^*$.
4. Repeat Step 3 a large number of times, say $H = 5000$. Now, you have H number of $\hat{T}_{AM}^*$ values. Rearrange them in increasing order, say, $\hat{T}_{AM(1)}^*, \hat{T}_{AM(2)}^*, \cdots, \hat{T}_{AM(H)}^*$. Then take $(1-\alpha)$-th quantile as an critical value, say f^*,

$$f^* := \hat{T}_{AM([(1-\alpha)H])}^*.$$

In the next theorem, it is shown that the asymptotic conditional distribution of the bootstrap statistic $\underline{\mathbf{T}}^*$ given the sample $\underline{\mathbf{X}} = (X_{11}, X_{12}, \cdots, X_{kn_k})$ is the same as the asymptotic null distribution of $\underline{\mathbf{T}}$.

Theorem 3.1. *The asymptotic distribution of* $\underline{\mathbf{T}}^*$ *given* $\underline{\mathbf{X}}$ *is* $N_{k-1}(\underline{\mathbf{0}}, \mathbf{D})$.

Proof. Define

$$Z_i^* = \sqrt{\frac{n_i n_{i+1}}{N}}(\bar{X}_{i+1.}^* - \bar{X}_{i.}^*)$$

and

$$\hat{b}_i^{*2} = \frac{n_i}{N}S_{i+1}^{*2} + \frac{n_{i+1}}{N}S_i^{*2},\ i = 1, 2, \cdots, k-1.$$

It can be easily shown that,

$$E_*(Z_i^*) = 0,\ Var_*(Z_i^*) = \hat{b}_i^2,\ i = 1, 2, \cdots, k-1$$

and

$$Cov_*(Z_i^*, Z_{i+1}^*) = -\sqrt{r_i r_{i+2}} S_{i+1}^2 = \hat{e}_i(say),\ i = 1, 2, \cdots, k-2,$$

where E_*, Var_* and Cov_* denotes the conditional expectation, variance and covariance respectively. Therefore, given observations $\underline{\mathbf{X}} = (X_{11}, X_{12}, \cdots, X_{kn_k})$, the conditional distribution of $\underline{\mathbf{Z}}^* = (Z_1^*, Z_2^*, \cdots, Z_{k-1}^*)^T$ is $N_{k-1}(\underline{\mathbf{0}}, \mathbf{\Sigma}_z^*)$, where

$$\mathbf{\Sigma}_z^* = \begin{bmatrix} \hat{b}_1^2 & \hat{e}_1 & 0 & \dots & 0 & 0 \\ \hat{e}_1 & \hat{b}_2^2 & \hat{e}_2 & \dots & 0 & 0 \\ 0 & \hat{e}_2 & \hat{b}_3^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \hat{e}_{k-2} & \hat{b}_{k-1}^2 \end{bmatrix}.$$

By asymptotic set up (5)

$$\mathbf{\Sigma}_z^* \xrightarrow{P} \mathbf{\Sigma}_z. \tag{6}$$

Let $\mathbf{B}^* = diag\left(\hat{b}_1^*, \hat{b}_2^*, \cdots, \hat{b}_{k-1}^*\right)$.

By Bickel and Freedman [3], given the observations $\underline{\mathbf{X}}$, $S_i^{*2} \xrightarrow{P} \sigma_i^2$ in conditional probability, as $n_i \to \infty$ for $i = 1, 2, \cdots, k$.

Hence by asymptotic set up (5)

$$\mathbf{B}^* \xrightarrow{P} \mathbf{B} \text{ as } N, \min_{1 \le i \le k} n_i \to \infty. \tag{7}$$

Therefore, by Equations (6), (7), multivariate Slutsky's theorem and continuous mapping theorem, the conditional distribution of $\underline{\mathbf{T}}^* = \mathbf{B}^{*-1}\underline{\mathbf{Z}}^*$ converges to the distribution of $N_{k-1}(\underline{\mathbf{0}}, \mathbf{D})$ as $N, \min_{1 \le i \le k} n_i \to \infty$.

In the following theorem, the asymptotic accuracy of the parametric bootstrap test $\underline{\mathbf{T}}^*$ is established.

Theorem 3.2. *Let $F_{\underline{T}^*|\underline{X}}(\underline{t})$ denote the conditional distribution function of $\underline{\mathbf{T}}^*$ given the observations $\underline{\mathbf{X}}$ and $F_{\underline{T}|H_0}(\underline{t})$ denote the null distribution function of $\underline{\mathbf{T}}$. Then under the asymptotic set up (5),*

$$\sup_{\underline{t} \in \mathbb{R}^{k-1}} |F_{\underline{T}^*|\underline{X}}(\underline{t}) - F_{\underline{T}|H_0}(\underline{t})| \xrightarrow{P} 0.$$

Proof. Under H_0, $X_{ij} \sim N(\mu, \sigma_i^2)$, $j = 1, 2, \cdots, n_i$, $i = 1, 2, \cdots, k$.
Without loss of generality, we can choose $\mu = 0$.
Hence under H_0,

$$\underline{\mathbf{U}} = \left(\sqrt{N}\bar{X}_{1.}, \sqrt{N}\bar{X}_{2.}, \cdots, \sqrt{N}\bar{X}_{k.}\right)^T \sim N_k(\underline{\mathbf{0}}, \mathbf{\Sigma}_1)$$

and given $\underline{\mathbf{X}}$ conditionally

$$\underline{\mathbf{U}}^* = \left(\sqrt{N}\bar{X}_{1.}^*, \sqrt{N}\bar{X}_{2.}^*, \cdots, \sqrt{N}\bar{X}_{k.}^*\right)^T \sim N_k(\underline{\mathbf{0}}, \mathbf{\Sigma}_2)$$

where $\mathbf{\Sigma}_1 = diag\left(\frac{N\sigma_1^2}{n_1}, \frac{N\sigma_2^2}{n_2}, \cdots, \frac{N\sigma_k^2}{n_k}\right)$ and $\mathbf{\Sigma}_2 = diag\left(\frac{NS_1^2}{n_1}, \frac{NS_2^2}{n_2}, \cdots, \frac{NS_k^2}{n_k}\right)$.
Therefore, Kullback-Leibler divergence of $\underline{\mathbf{U}}^*$ from $\underline{\mathbf{U}}$ is

$$D_{KL}(\underline{\mathbf{U}}^* \mid\mid \underline{\mathbf{U}}) = \frac{1}{2}\left[\sum_{i=1}^{k} \frac{S_i^2}{\sigma_i^2} - k + \sum_{i=1}^{k} \log\left(\frac{\sigma_i^2}{S_i^2}\right)\right] \xrightarrow{P} 0 \text{ as } \min_{1 \le i \le k} n_i \to \infty. \tag{8}$$

Let

$$\underline{\mathbf{Y}} = (Y_1, Y_2, \cdots, Y_{k-1})^T = A\underline{\mathbf{U}},$$

where

$$A = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix},$$

$Y_i = \sqrt{N}\left(\bar{X}_{i+1.} - \bar{X}_{i.}\right), i = 1, 2, \cdots, k-1.$
We also consider

$$\underline{\mathbf{Y}}^* = \left(Y_1^*, Y_2^*, \cdots, Y_{k-1}^*\right)^T = A\underline{\mathbf{U}}^*,$$

where $Y_i^* = \sqrt{N}\left(\bar{Y}_{i+1.}^* - \bar{Y}_{i.}^*\right), i = 1, 2, \cdots, k-1.$

By Equation (8), $D_{KL}(\underline{\mathbf{Y}}^* \mid\mid \underline{\mathbf{Y}}) \xrightarrow{P} 0$ as $\min_{1\le i\le k} n_i \to \infty$.
This implies

$$\sup_{\underline{y}} |F_{\underline{Y}^*|\underline{X}}(\underline{y}) - F_{\underline{Y}|H_0}(\underline{y})| \xrightarrow{P} 0. \tag{9}$$

Next write $\underline{\mathbf{T}} = \hat{\mathbf{R}}\underline{\mathbf{Y}}$ and $\underline{\mathbf{T}}^* = \hat{\mathbf{R}}^*\underline{\mathbf{Y}}^*$, where $\hat{\mathbf{R}} = diag\left(\frac{\sqrt{n_1 n_2}}{N}\frac{1}{\hat{b}_1}, \cdots, \frac{\sqrt{n_{k-1} n_k}}{N}\frac{1}{\hat{b}_{k-1}}\right)$ and $\hat{\mathbf{R}}^* = diag\left(\frac{\sqrt{n_1 n_2}}{N}\frac{1}{\hat{b}_1^*}, \cdots, \frac{\sqrt{n_{k-1} n_k}}{N}\frac{1}{\hat{b}_{k-1}^*}\right)$.

Then under (5), $\hat{\mathbf{R}} \xrightarrow{P} \mathbf{R}$ and $\hat{\mathbf{R}}^* \xrightarrow{P} \mathbf{R}$, where $\mathbf{R} = diag\left(\frac{\sqrt{r_1 r_2}}{b_1}, \cdots, \frac{\sqrt{r_{k-1} r_k}}{b_{k-1}}\right)$. Now multivariate Slutsky theorem and Equation (9) yield the result.

4 Size and power studies

In this section simulation studies are carried out to analyze the size and power behaviour of the two tests, Max-T and AMax-T. The critical points have been calculated using parametric bootstrap as described in Algorithms 1 and 2. Sizes of Max-T and AMax-T tests are estimated using the following algorithm:

Algorithm 3.

1. Fix a configuration of μ, $(\sigma_1^2, \sigma_2^2, \cdots, \sigma_k^2)$ and $(n_1, n_2, \cdots, n_k)$.
2. Generate k number of samples of size n_i, $X_{i1}, X_{i2}, \cdots, X_{in_i}$ from $N(\mu, \sigma_i^2)$ for $i = 1, 2, \cdots, k$.
3. Using the observations from Step 2 calculate Max-T critical value d^* and AMax-T critical value f^* using Algorithms 1 and 2 respectively.
4. Using the data from Step 2 calculate test statistic value T_M.
5. Repeat Steps 2-4 a large number of times, say $B = 10,000$. Now, we have B number of values of the test statistic T_M, say T_{M_r}, $r = 1, 2, \cdots, B$, B number of critical values of Max-T, say d_r^*, $r = 1, 2, \cdots, B$ and B number of critical values of AMax-T, say f_r^*, $r = 1, 2, \cdots, B$.

6. Finally estimates of sizes of Max-T and AMax-T are:

$$\alpha_{MT} = \frac{\text{number of times } \ T_{M_r} \ \text{ more than } \ d_r^*}{B} = \frac{1}{B}\sum_{r=1}^{B} \mathbf{1}\{T_{M_r} > d_r^*\}$$

and

$$\alpha_{AMT} = \frac{\text{number of times } \ T_{M_r} \ \text{ more than } \ f_r^*}{B} = \frac{1}{B}\sum_{r=1}^{B} \mathbf{1}\{T_{M_r} > f_r^*\}$$

respectively.

Powers of the tests are also calculated using the above algorithm by fixing configurations of $(\mu_1, \mu_2, \cdots, \mu_k)$ in Step 1 as $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$ with at least one strict inequality.

For all the simulation studies, we have taken the nominal size as $\alpha = 0.05$. The outcomes of simulation studies are observed to be similar for nominal size $\alpha = 0.1$. When the number of groups is 4 or 5, the sizes of tests are calculated for various combinations of sample sizes and variances. For $k = 4$ and $k = 5$, size values are given in Tables 1 and 2 respectively. For $k = 4$, four different combinations of sample sizes and seven combinations of values of variances are chosen. The sample sizes are denoted by: $N_1 = (10, 15, 15, 20)$, $N_2 = (10, 30, 40, 10)$, $N_3 = (30, 40, 50, 60)$ and $N_4 = (70, 80, 60, 70)$. The variance configurations are $(0.1, 0.2, 0.3, 0.4)$, $(1, 1, 1, 1)$, $(1, 1, 1, 0.5)$, $(1, 1, 1, 2)$, $(1, 1, 1, 3)$ $(1, 2, 1, 3)$, $(3, 2, 1, 1)$.

In Table 2, for $k = 5$, four combinations of sample sizes and eight combinations of values of variances are chosen. The sample sizes are denoted by: $N_5 = (15, 10, 8, 10, 10)$, $N_6 = (10, 20, 20, 30, 20)$, $N_7 = (30, 40, 40, 40, 50)$ and $N_8 = (50, 60, 70, 40, 50)$. The variance configurations are $(0.1, 0.2, 0.3, 0.4, 0.5)$, $(1, 1, 1, 1, 1)$, $(1, 2, 2, 3, 4)$, $(1, 1, 1, 1, 0.5)$, $(1, 1, 1, 1, 2)$, $(1, 1, 1, 1, 3)$, $(1, 2, 3, 1, 1)$, $(3, 2, 1, 1, 1)$. Note that sample sizes are chosen as small, moderate, large and unbalanced. Similarly variances configurations are chosen as equal, unequal, small, or moderate.

Powers of Max-T and AMax-T are tabulated in Tables 3, 4, 5 and 6 for different combinations of parameters and sample sizes. Results for $k = 3$ are given in Tables 3 and 4. In both the tables two combination of variances are taken: equal (1,1,1) and variances (1,2,3). Sample sizes taken are moderate: $N_9 = (20, 30, 25)$ and large: $N_{10} = (60, 70, 50)$. In Table 3, μ_1, μ_2, μ_3 are taken to be in strictly increasing order, say $(\mu_1, \mu_2, \mu_3) = c(1, 1.1, 1.2)$, where c ranges from 1 to 5 with increment 0.2. The choice $c = 0$ gives the size value. In Table 4, two group means are taken to be equal and other unequal, that is, $(\mu_1, \mu_2, \mu_3) = c(1, 1, 1.5)$.
In Tables 5 and 6, powers and sizes are tabulated for $k = 4$. In Table 5, we take group means to be strictly ordered: $(\mu_1, \mu_2, \mu_3, \mu_4) = c(1, 1.1, 1.2, 1.3)$, equal variances: $(1, 1, 1, 1)$ and unequal variances: $(1, 2, 3, 4)$, moderate sample sizes: $N_{11} = (20, 25, 30, 20)$ and large sample sizes: $N_{12} = (40, 60, 50, 70)$. In Table 6, powers are given for monotonically increasing group means with two equal components, say $(\mu_1, \mu_2, \mu_3, \mu_4) = c(1, 1.2, 1.2, 1.5)$. Another variance configuration $(1, 1, 3, 4)$ is chosen here, all other configurations are the same as in Table 5.

Table 1: Estimated sizes of Max-T and AMax-T tests for $k = 4$ and nominal size 0.05

	N_1		N_2		N_3		N_4	
σ^2	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T
(0.1, 0.2, 0.3, 0.4)	0.0550	0.0608	0.0520	0.0734	0.0496	0.0562	0.0550	0.0582
(1, 1, 1, 1)	0.0510	0.0712	0.0568	0.0740	0.0544	0.0490	0.0558	0.0544
(1, 1, 1, 0.5)	0.0506	0.0720	0.0508	0.0618	0.0532	0.0556	0.0532	0.0560
(1, 1, 1, 2)	0.0490	0.0732	0.0552	0.0686	0.0492	0.0572	0.0532	0.0604
(1, 1, 1, 3)	0.0550	0.0702	0.0494	0.0734	0.0568	0.0602	0.0528	0.0576
(1, 2, 1, 3)	0.0504	0.0724	0.0524	0.0660	0.0582	0.0534	0.0516	0.0504
(3, 2, 1, 1)	0.0478	0.0666	0.0522	0.0670	0.0598	0.0600	0.0540	0.0492

Table 2: Estimated sizes of Max-T and AMax-T tests for $k = 5$ and nominal size 0.05

	N_5		N_6		N_7		N_8	
σ^2	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T
(0.1, 0.2, 0.3, 0.4, 0.5)	0.0530	0.0840	0.0528	0.0642	0.0570	0.0582	0.0538	0.0560
(1, 1, 1, 1, 1)	0.0478	0.0800	0.0510	0.0626	0.0532	0.0604	0.0490	0.0606
(1, 2, 2, 3, 4)	0.0482	0.0756	0.0504	0.0592	0.0502	0.0528	0.0548	0.0568
(1, 1, 1, 1, 0.5)	0.0526	0.0838	0.0504	0.0642	0.0468	0.0570	0.0498	0.0544
(1, 1, 1, 1, 2)	0.0514	0.0780	0.0518	0.0668	0.0566	0.0570	0.0540	0.0558
(1, 1, 1, 1, 3)	0.0526	0.0838	0.0550	0.0650	0.0500	0.0558	0.0492	0.0584
(1, 2, 3, 1, 1)	0.0490	0.0830	0.0502	0.0640	0.0536	0.0643	0.0478	0.0510
(3, 2, 1, 1, 1)	0.0510	0.0730	0.0550	0.0650	0.0560	0.0600	0.0512	0.0548

Table 3: Powers of Max-T and AMax-T tests for $k = 3$ and $\alpha = 0.05$ when $(\mu_1, \mu_2, \mu_3) = c(1, 1.1, 1.2)$

	$(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 1, 1)$				$(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 2, 3)$			
	N_9		N_{10}		N_9		N_{10}	
c	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T
0	0.0489	0.0583	0.0490	0.0628	0.0489	0.0545	0.0490	0.0550
1	0.1080	0.1174	0.1650	0.1628	0.0872	0.1030	0.1184	0.1320
1.2	0.1264	0.1392	0.2030	0.2054	0.1034	0.1118	0.1460	0.1460
1.4	0.1424	0.1552	0.2400	0.2636	0.1150	0.1162	0.1622	0.1770
1.6	0.1616	0.1812	0.2764	0.2924	0.1260	0.1380	0.1904	0.2008
1.8	0.1898	0.2074	0.3420	0.3458	0.1400	0.1488	0.2230	0.2174
2	0.2130	0.2408	0.3842	0.3862	0.1440	0.1658	0.2532	0.2540
2.2	0.2416	0.2530	0.4280	0.4602	0.1662	0.1794	0.2946	0.3008
2.4	0.2562	0.2868	0.5028	0.5172	0.1832	0.2006	0.3276	0.3428
2.6	0.3028	0.3262	0.5570	0.5580	0.2016	0.2272	0.3730	0.3758
2.8	0.3380	0.3484	0.6214	0.6314	0.2238	0.2364	0.3932	0.4274
3	0.3484	0.3814	0.6820	0.6882	0.2380	0.2548	0.4376	0.4586
3.2	0.3954	0.4142	0.7352	0.7484	0.2530	0.2868	0.4826	0.5022

Continued on next page

Table 3 – *Continued from previous page*

c	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T
3.4	0.4252	0.4702	0.7832	0.8034	0.2800	0.3032	0.5304	0.5540
3.6	0.4660	0.4952	0.8288	0.8394	0.3090	0.3232	0.5860	0.5952
3.8	0.5054	0.5456	0.8684	0.8672	0.3400	0.3464	0.6190	0.6312
4	0.5352	0.5766	0.9024	0.9058	0.3512	0.3788	0.6660	0.6852
4.2	0.5806	0.6032	0.9292	0.9364	0.3730	0.4006	0.7078	0.7222
4.4	0.6054	0.6448	0.9446	0.9474	0.3984	0.4134	0.7416	0.7602
4.6	0.6518	0.6812	0.9632	0.9680	0.4320	0.4530	0.7728	0.7962
4.8	0.6866	0.7100	0.9806	0.9828	0.4540	0.4782	0.8126	0.8278
5	0.7254	0.7418	0.9910	0.9842	0.4772	0.5160	0.8484	0.8552

Table 4: Powers of Max-T and AMax-T tests for $k = 3$ and $\alpha = 0.05$ when $(\mu_1, \mu_2, \mu_3) = c(1, 1, 1.5)$

	$(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 1, 1)$				$(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 2, 3)$			
	N_9		N_{10}		N_9		N_{10}	
c	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T
0	0.0489	0.0583	0.0490	0.0628	0.0489	0.0545	0.0490	0.0550
1	0.4626	0.4918	0.7852	0.7812	0.2422	0.2586	0.4088	0.4136
1.2	0.5972	0.6186	0.9060	0.9064	0.2996	0.3208	0.5450	0.5496
1.4	0.7123	0.7458	0.9686	0.9690	0.3812	0.3948	0.6572	0.6772
1.6	0.8366	0.8500	0.9912	0.9914	0.4662	0.5014	0.7808	0.7948
1.8	0.9128	0.9208	0.9982	0.9988	0.5454	0.5730	0.8668	0.8668
2	0.9558	0.9600	1	1	0.6430	0.6584	0.9258	0.9272
2.2	0.9834	0.98200	1	1	0.7272	0.7320	0.9654	0.9622
2.4	0.9944	0.9956	1	1	0.7968	0.8004	0.9816	0.9848
2.6	0.9978	0.9987	1	1	0.8422	0.8632	0.9936	0.9926
2.8	0.9987	0.9994	1	1	0.8946	0.9048	0.9970	0.9974
3	0.9992	1	1	1	0.9328	0.9450	0.9988	0.9992
3.2	1	1	1	1	0.9608	0.9590	1	1
3.4	1	1	1	1	0.9782	0.9768	1	1
3.6	1	1	1	1	0.9818	0.9880	1	1
3.8	1	1	1	1	0.9918	0.9942	1	1
4	1	1	1	1	0.9956	0.9975	1	1
4.2	1	1	1	1	0.9978	1	1	1
4.4	1	1	1	1	1	1	1	1
4.6	1	1	1	1	1	1	1	1
4.8	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1

Table 5: Powers of Max-T and AMax-T tests for $k = 4$ and $\alpha = 0.05$ when $(\mu_1, \mu_2, \mu_3, \mu_4) = c(1, 1.1, 1.2, 1.3)$

	$(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 1, 1, 1)$				$(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 2, 3, 4)$			
	N_{11}		N_{12}		N_{11}		N_{12}	
c	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T
0	0.0513	0.0592	0.0526	0.0530	0.0550	0.0622	0.0540	0.0515
1	0.1095	0.1320	0.1475	0.1680	0.0930	0.1020	0.1250	0.1270
1.2	0.1225	0.1435	0.2075	0.2165	0.0975	0.1125	0.1340	0.1360
1.4	0.1380	0.1750	0.2435	0.2525	0.1060	0.1355	0.1605	0.1625
1.6	0.1800	0.2015	0.2750	0.2990	0.1180	0.1315	0.1690	0.1755
1.8	0.2020	0.2160	0.3415	0.3540	0.1430	0.1460	0.1845	0.2290
2	0.2220	0.2445	0.3940	0.3980	0.1425	0.1755	0.2280	0.2545
2.2	0.2395	0.2735	0.4675	0.4550	0.1485	0.1804	0.2540	0.2800
2.4	0.2740	0.3115	0.5045	0.5355	0.1655	0.2045	0.2670	0.2875
2.6	0.3125	0.3300	0.5910	0.6220	0.1845	0.2150	0.3225	0.3320
2.8	0.3375	0.3770	0.6395	0.6635	0.2110	0.2390	0.3425	0.3675
3	0.3720	0.4015	0.7155	0.7250	0.2320	0.2535	0.3825	0.4405
3.2	0.4220	0.4535	0.7650	0.7680	0.2455	0.2600	0.4270	0.4440
3.4	0.4625	0.4915	0.8180	0.8270	0.2695	0.2910	0.5000	0.4990
3.6	0.5210	0.5575	0.8615	0.8800	0.2880	0.3225	0.5290	0.5420
3.8	0.5225	0.6005	0.9145	0.9150	0.3015	0.3485	0.5565	0.5890
4	0.5835	0.6310	0.9395	0.9440	0.3375	0.3530	0.5940	0.6190
4.2	0.6325	0.6785	0.9560	0.9645	0.3695	0.3880	0.6340	0.6605
4.4	0.6565	0.7240	0.9740	0.9830	0.3930	0.4165	0.6840	0.7205
4.6	0.7200	0.7565	0.9860	0.9863	0.4200	0.4230	0.7490	0.7545
4.8	0.7410	0.7910	0.9926	0.9930	0.4265	0.4585	0.7910	0.7855
5	0.7790	0.8195	0.9955	0.9965	0.4530	0.4815	0.8160	0.7995

Table 6: Powers of Max-T and AMax-T tests for $k = 3$ and $\alpha = 0.05$ when $(\mu_1, \mu_2, \mu_3) = c(1, 1.1, 1.2)$

	$(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 1, 1)$				$(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 2, 3)$			
	N_9		N_{10}		N_9		N_{10}	
c	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T
0	0.0489	0.0583	0.0490	0.0628	0.0489	0.0545	0.0490	0.0550
1	0.1080	0.1174	0.1650	0.1628	0.0872	0.1030	0.1184	0.1320
1.2	0.1264	0.1392	0.2030	0.2054	0.1034	0.1118	0.1460	0.1460
1.4	0.1424	0.1552	0.2400	0.2636	0.1150	0.1162	0.1622	0.1770
1.6	0.1616	0.1812	0.2764	0.2924	0.1260	0.1380	0.1904	0.2008
1.8	0.1898	0.2074	0.3420	0.3458	0.1400	0.1488	0.2230	0.2174
2	0.2130	0.2408	0.3842	0.3862	0.1440	0.1658	0.2532	0.2540
2.2	0.2416	0.2530	0.4280	0.4602	0.1662	0.1794	0.2946	0.3008
2.4	0.2562	0.2868	0.5028	0.5172	0.1832	0.2006	0.3276	0.3428
2.6	0.3028	0.3262	0.5570	0.5580	0.2016	0.2272	0.3730	0.3758
2.8	0.3380	0.3484	0.6214	0.6314	0.2238	0.2364	0.3932	0.4274
3	0.3484	0.3814	0.6820	0.6882	0.2380	0.2548	0.4376	0.4586
3.2	0.3954	0.4142	0.7352	0.7484	0.2530	0.2868	0.4826	0.5022
3.4	0.4252	0.4702	0.7832	0.8034	0.2800	0.3032	0.5304	0.5540
3.6	0.4660	0.4952	0.8288	0.8394	0.3090	0.3232	0.5860	0.5952
3.8	0.5054	0.5456	0.8684	0.8672	0.3400	0.3464	0.6190	0.6312
4	0.5352	0.5766	0.9024	0.9058	0.3512	0.3788	0.6660	0.6852

Continued on next page

Table 6 – *Continued from previous page*

c	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T	Max-T	AMax-T
4.2	0.5806	0.6032	0.9292	0.9364	0.3730	0.4006	0.7078	0.7222
4.4	0.6054	0.6448	0.9446	0.9474	0.3984	0.4134	0.7416	0.7602
4.6	0.6518	0.6812	0.9632	0.9680	0.4320	0.4530	0.7728	0.7962
4.8	0.6866	0.7100	0.9806	0.9828	0.4540	0.4782	0.8126	0.8278
5	0.7254	0.7418	0.9910	0.9842	0.4772	0.5160	0.8484	0.8552

Remark 1. From Tables 1 and 2, it can be noticed that the parametric bootstrap-based Max-T test achieves a size of 0.05 for all the combinations of sample sizes and variances those are considered here. It also achieves size for a very small sample as $(5, 5, 5, 5, 5)$ (not tabulated here). The asymptotic test AMax-T is liberal for small sample sizes. However as sample sizes increase, AMax-T, yields a size close to 0.05. This is due to the fact that AMax-T test is an asymptotic test. Similar observations can be made about the sizes of two tests from the first rows of Tables 3, 4, 5 and 6.

Remark 2. It is seen from Tables 3, 4, 5, and 6 that the powers of both tests increase as the value of c increases, and so the tests are good discriminators between group means.
If the variance of any of the groups increases then power decreases when all other parameters remain unchanged. Power also increases with the sample sizes. It is also seen that both Max-T and AMax-T perform equally well in terms of power. However, AMax-T is more liberal and so Max-T may be preferred. It is to be noted that both the tests perform well for strictly unequal μ_i's (see Table 3 and 5) and also for the combination of μ_i's with at least one equality (see Table 4 and 6).

Remark 3. For usage of our proposed tests, R packages have been developed. Packages are developed for arbitrary k. In the packages we only need data (excel data) and level of significance as input. The output is a vector, whose first component is the test statistic value and the second is the corresponding critical value of the test. If test statistic value is greater than critical value, the null hypothesis is rejected.

`OneWayMaxT` and `OneWayAMaxT` are respectively the packages for Max-T and AMax-T tests for any number of groups. The corresponding functions are `OneWayMaxTcrit` and `OneWayAMaxTcrit`. These are uploaded to the Github account. To use `OneWayMaxT` package, one needs to follow the steps as follows.

```
install.packages("devtools")
library(devtools)
install_github("AnjanaStat/OneWayMaxT")
library(OneWayMaxT)
OneWayMaxTcrit(data,alpha)
```

Similar steps are to be followed for the `OneWayAMaxT` package.

5 Robustness of Max-T and AMax-T tests

In this section, we investigate the robustness of Max-T and AMax-T tests under departure from normality by evaluating values of estimated sizes and powers. Donaldson [10] studied robustness of the classical ANOVA F-test when the normality of error distribution and homogeneous variance conditions are not satisfied. He checked the values of size and power of the test under deviation from the assumptions. Ramseyer and Tcheng [21] studied the robustness of the studentized range

test statistic under violations of the assumptions of normality and homogeneity of variances. They considered two non-normal distributions: U(0,1) and exponential with location zero and scale unity. Srivastava [25] also studied the robustness of t-test under the violation of normality.

In this section, we have calculated sizes and powers of Max-T and AMax-T tests for five non-normal distributions: t-distribution with degrees of freedom 3, Laplace distribution, exponential distribution with scale parameter unity, lognormal distribution and Weibull distribution with scale parameter 1 and shape parameter 2. Configurations of parameters are chosen to be same for normal and non-normal distributions. By generating data from respective non-normal distributions in Step 1 of Algorithm 3, sizes and powers are calculated. At first standardized non-normal random variable with mean 0 and variance 1, Y_{ij} is taken and then the final data with mean μ_i and variance σ_i^2 is taken as $X_{ij} = \mu_i + \sigma_i Y_{ij}$.

Results of robustness study are presented in Tables 7, 8, 9 and 10 for $k = 3$. The choices for variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ are taken as (1,1,1) and (1,2,3) and those of sample sizes (n_1, n_2, n_3) are chosen as (20,30,25) and (60,70,50). For power computation (μ_1, μ_2, μ_3) is taken as $c(1, 1.1, 1.2)$, where c ranges from 1 to 5, with increment 0.2. The choice $c = 0$ gives the size of the test. In the Tables N, T, L, Ex, Wei, LN correspond to normal, t, Laplace or double exponential, exponential, Weibull, and lognormal distributions respectively.

Table 7: Powers of Max-T when $(\mu_1, \mu_2, \mu_3) = c(1, 1.1, 1.2)$ and $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 1, 1)$

	$(n_1, n_2, n_3) = (20, 30, 25)$						$(n_1, n_2, n_3) = (60, 70, 50)$					
c	N	T	L	Ex	Wei	LN	N	T	L	Ex	Wei	LN
0	0.0489	0.0458	0.0470	0.0527	0.0580	0.0428	0.0490	0.0508	0.0560	0.0458	0.0558	0.0386
1	0.1080	0.1186	0.1072	0.1240	0.1102	0.1346	0.1650	0.1802	0.1716	0.1492	0.1612	0.1826
1.2	0.1264	0.1294	0.1288	0.1378	0.1304	0.1626	0.2030	0.2228	0.2002	0.2088	0.1902	0.2342
1.4	0.1424	0.1542	0.1498	0.1532	0.1526	0.1942	0.2400	0.2728	0.2482	0.2290	0.2354	0.2816
1.6	0.1616	0.1876	0.1612	0.1852	0.1668	0.2166	0.2764	0.3050	0.2862	0.2954	0.2772	0.3266
1.8	0.1898	0.2036	0.1892	0.2094	0.1898	0.2622	0.3420	0.3806	0.3360	0.3374	0.3342	0.3852
2	0.2130	0.2518	0.2140	0.2372	0.2240	0.2982	0.3842	0.4498	0.3924	0.3850	0.3734	0.4574
2.2	0.2416	0.2756	0.2454	0.2672	0.2344	0.3464	0.4280	0.4894	0.4560	0.4394	0.4416	0.5216
2.4	0.2562	0.3064	0.2658	0.3008	0.2654	0.3878	0.5028	0.5732	0.5060	0.5172	0.4910	0.5858
2.6	0.3028	0.3468	0.3062	0.3290	0.3006	0.4098	0.5570	0.6246	0.5804	0.5698	0.5508	0.6500
2.8	0.3380	0.3910	0.3438	0.3516	0.3342	0.4778	0.6214	0.6944	0.6246	0.6412	0.6192	0.7048
3	0.3484	0.4284	0.3686	0.3914	0.3726	0.4986	0.6820	0.7376	0.6908	0.6812	0.6850	0.7580
3.2	0.3954	0.4704	0.4088	0.4274	0.4092	0.5468	0.7352	0.7956	0.7380	0.7462	0.7370	0.8046
3.4	0.4252	0.5186	0.4442	0.4712	0.4326	0.5980	0.7832	0.8446	0.7850	0.7948	0.7752	0.8422
3.6	0.4660	0.5562	0.4776	0.5050	0.4776	0.6338	0.8288	0.8666	0.8406	0.8412	0.8238	0.8840
3.8	0.5054	0.5930	0.5160	0.5376	0.5090	0.6642	0.8684	0.8964	0.8812	0.8718	0.8668	0.9110
4	0.5352	0.6434	0.5636	0.5810	0.5492	0.7170	0.9024	0.9228	0.9134	0.9028	0.9012	0.9206
4.2	0.5806	0.6666	0.6040	0.6244	0.5870	0.7370	0.9292	0.9456	0.9314	0.9226	0.9322	0.9416
4.4	0.6054	0.6994	0.6342	0.6616	0.6214	0.7638	0.9446	0.9520	0.9504	0.9484	0.9448	0.9538
4.6	0.6518	0.7378	0.6642	0.6894	0.6648	0.8020	0.9632	0.9710	0.9684	0.9668	0.9668	0.9648
4.8	0.6866	0.7726	0.7002	0.7338	0.7072	0.8204	0.9806	0.9768	0.9730	0.9754	0.9778	0.9746
5	0.7254	0.7926	0.7294	0.7506	0.7230	0.8342	0.9910	0.9872	0.9820	0.9825	0.9892	0.9824

Table 8: Power of Max-T when $(\mu_1, \mu_2, \mu_3) = c(1, 1.1, 1.2)$ and $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 2, 3)$

	$(n_1, n_2, n_3) = (20, 30, 25)$						$(n_1, n_2, n_3) = (60, 70, 50)$					
c	N	T	L	Ex	Wei	LN	N	T	L	Ex	Wei	LN
0	0.0489	0.0438	0.0468	0.0380	0.0460	0.0212	0.0490	0.0478	0.0522	0.0384	0.0486	0.0226
1	0.0872	0.0908	0.0912	0.0756	0.0848	0.0654	0.1184	0.1250	0.1194	0.1026	0.1128	0.0922
1.2	0.1034	0.0942	0.0996	0.0802	0.0922	0.0762	0.1460	0.1614	0.1500	0.1160	0.1324	0.1082
1.4	0.1150	0.1286	0.1106	0.0930	0.1080	0.0958	0.1622	0.1768	0.1698	0.1356	0.1500	0.1468
1.6	0.1260	0.1406	0.1240	0.1070	0.1246	0.1082	0.1904	0.2128	0.1756	0.1688	0.1836	0.1568

Continued on next page

Table 8 – *Continued from previous page*

c	N	T	L	Ex	Wei	LN	N	T	L	Ex	Wei	LN
1.8	0.1400	0.1570	0.1410	0.1148	0.1324	0.1300	0.2230	0.2454	0.2244	0.1890	0.2124	0.2124
2	0.1440	0.1594	0.1498	0.1424	0.1362	0.1432	0.2532	0.2888	0.2666	0.2256	0.2378	0.2436
2.2	0.1662	0.1888	0.1754	0.1462	0.1630	0.1636	0.2934	0.3264	0.2982	0.2708	0.2728	0.2710
2.4	0.1832	0.2128	0.1856	0.1648	0.180	0.1906	0.3276	0.3670	0.3282	0.2930	0.3072	0.3324
2.6	0.2160	0.2234	0.2102	0.1812	0.1932	0.2074	0.3730	0.4202	0.3666	0.3432	0.3436	0.3818
2.8	0.2238	0.2570	0.2288	0.1962	0.2082	0.2162	0.3932	0.4474	0.4262	0.3844	0.3816	0.4132
3	0.2380	0.2786	0.2520	0.2102	0.2354	0.2644	0.4376	0.5072	0.4544	0.4208	0.4340	0.4700
3.2	0.2580	0.2998	0.2728	0.2432	0.2466	0.2960	0.4826	0.5464	0.4944	0.4624	0.4806	0.5322
3.4	0.2800	0.3230	0.2872	0.2644	0.2606	0.3258	0.5304	0.5940	0.5480	0.5246	0.5214	0.5756
3.6	0.3090	0.3620	0.3236	0.2936	0.2912	0.3612	0.5860	0.6558	0.5780	0.5610	0.5664	0.6332
3.8	0.3400	0.3844	0.3346	0.3116	0.3152	0.3856	0.6190	0.6932	0.6276	0.6092	0.6088	0.6770
4	0.3512	0.4076	0.3616	0.3402	0.3406	0.4256	0.6660	0.7286	0.6636	0.6672	0.6508	0.7278
4.2	0.3730	0.4512	0.3878	0.3654	0.3602	0.4614	0.7078	0.7516	0.7132	0.6968	0.6920	0.7688
4.4	0.3984	0.4954	0.4334	0.3978	0.3810	0.4972	0.7416	0.8034	0.7542	0.7562	0.7378	0.8034
4.6	0.4320	0.5130	0.4390	0.4228	0.4242	0.5338	0.7728	0.8324	0.7880	0.7840	0.7850	0.8298
4.8	0.4540	0.5494	0.4684	0.4530	0.4512	0.5648	0.8126	0.8582	0.8212	0.8186	0.8112	0.8632
5	0.5160	0.5600	0.5020	0.4922	0.4692	0.5908	0.8484	0.8870	0.8510	0.8544	0.8434	0.8800

Table 9: Power of AMax-T when $(\mu_1, \mu_2, \mu_3) = c(1, 1.1, 1.2)$ and $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 1, 1)$

	$(n_1, n_2, n_3) = (20, 30, 25)$						$(n_1, n_2, n_3) = (60, 70, 50)$					
c	N	T	L	Ex	L	LN	N	T	L	Ex	Wei	LN
0	0.0583	0.0594	0.0556	0.0624	0.0594	0.0526	0.0628	0.0498	0.0534	0.0504	0.0542	0.0444
1	0.1174	0.1342	0.1294	0.1414	0.1268	0.1598	0.1628	0.1824	0.1754	0.1596	0.1680	0.1858
1.2	0.1392	0.1582	0.1338	0.1584	0.1348	0.1950	0.2054	0.2236	0.2134	0.2062	0.2080	0.2390
1.4	0.1552	0.1788	0.1718	0.1756	0.1578	0.2220	0.2636	0.2706	0.2406	0.2420	0.2452	0.2906
1.6	0.1812	0.2068	0.1844	0.2040	0.1870	0.2556	0.2924	0.3266	0.2942	0.2920	0.2840	0.3456
1.8	0.2074	0.2430	0.2100	0.2290	0.2066	0.2942	0.3458	0.3874	0.3504	0.3498	0.3350	0.4012
2	0.2408	0.2706	0.2278	0.2702	0.2316	0.3202	0.3862	0.4480	0.4066	0.4092	0.3920	0.4798
2.2	0.2530	0.2954	0.2738	0.2898	0.2496	0.3826	0.4602	0.4992	0.4540	0.4666	0.4474	0.5320
2.4	0.2868	0.3484	0.3086	0.3324	0.2950	0.4168	0.5172	0.5708	0.5124	0.5208	0.5216	0.5962
2.6	0.3262	0.3792	0.3338	0.3562	0.3244	0.4644	0.5590	0.6308	0.5704	0.5852	0.5720	0.6670
2.8	0.3484	0.4152	0.3456	0.3818	0.3710	0.5002	0.6314	0.7068	0.6428	0.6332	0.6316	0.7178
3	0.3814	0.4602	0.3984	0.4270	0.3788	0.5512	0.6882	0.7464	0.6868	0.7118	0.6858	0.7664
3.2	0.4142	0.5076	0.4320	0.4968	0.4316	0.5868	0.7484	0.7934	0.7504	0.7548	0.7526	0.8094
3.4	0.4702	0.5322	0.4680	0.5370	0.4594	0.6218	0.8034	0.8390	0.7932	0.7928	0.7960	0.8496
3.6	0.4952	0.5958	0.5160	0.5804	0.4972	0.6590	0.8394	0.8776	0.8402	0.8446	0.8256	0.8860
3.8	0.5456	0.6246	0.5472	0.6218	0.5426	0.7042	0.8672	0.9056	0.8750	0.8762	0.8766	0.9078
4	0.5766	0.6714	0.6080	0.6588	0.5668	0.7356	0.9058	0.9282	0.9112	0.9130	0.9038	0.9238
4.2	0.6034	0.7014	0.6266	0.6732	0.6160	0.7514	0.9364	0.9488	0.9322	0.9344	0.9268	0.9460
4.4	0.6448	0.7434	0.6702	0.6856	0.6438	0.7874	0.9474	0.9584	0.9486	0.9522	0.9542	0.9550
4.6	0.6812	0.7752	0.7028	0.7186	0.6752	0.8196	0.9680	0.9676	0.9650	0.9664	0.9698	0.9676
4.8	0.7100	0.8120	0.7352	0.7554	0.7252	0.8448	0.9828	0.9824	0.9790	0.9802	0.9782	0.9756
5	0.7418	0.8350	0.7618	0.7824	0.7494	0.8650	0.9842	0.9865	0.9836	0.9852	0.9878	0.9772

Table 10: Power of AMax-T when $(\mu_1, \mu_2, \mu_3) = c(1, 1.1, 1.2)$ and $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 2, 3)$

	$(n_1, n_2, n_3) = (20, 30, 25)$						$(n_1, n_2, n_3) = (60, 70, 50)$					
c	N	T	L	Ex	Wei	LN	N	T	L	Ex	Wei	LN
0	0.0545	0.0554	0.0562	0.0510	0.0550	0.0288	0.0550	0.0518	0.0566	0.0492	0.0516	0.0266
1	0.1030	0.1032	0.1080	0.0838	0.0976	0.0832	0.1320	0.1430	0.1348	0.1102	0.1202	0.1024
1.2	0.1118	0.1208	0.1240	0.0926	0.1002	0.0860	0.1460	0.1568	0.1532	0.1220	0.1396	0.1230
1.4	0.1262	0.1340	0.1310	0.1008	0.1202	0.1030	0.1770	0.1908	0.1740	0.1478	0.1700	0.1462
1.6	0.1380	0.1538	0.1480	0.1292	0.1256	0.1254	0.2008	0.2068	0.2066	0.1720	0.1936	0.1710
1.8	0.1488	0.1728	0.1542	0.1320	0.1562	0.1364	0.2174	0.2542	0.2304	0.1862	0.2264	0.2176
2	0.1658	0.1892	0.1752	0.1588	0.1610	0.1672	0.2540	0.2980	0.2616	0.2270	0.2450	0.2390
2.2	0.1794	0.2040	0.2014	0.1626	0.1828	0.1796	0.3008	0.3380	0.2966	0.2724	0.2900	0.2936
2.4	0.2006	0.2248	0.2078	0.1896	0.1958	0.2144	0.3428	0.3852	0.3314	0.3200	0.3228	0.3240
2.6	0.2272	0.2584	0.2186	0.1978	0.2028	0.2384	0.3758	0.4258	0.3842	0.3576	0.3524	0.3662

Continued on next page

Table 10 – *Continued from previous page*

c	N	T	L	Ex	Wei	LN	N	T	L	Ex	Wei	LN
2.8	0.2364	0.2914	0.2478	0.2306	0.2266	0.2620	0.4274	0.4680	0.4212	0.3898	0.3984	0.4358
3	0.2548	0.3128	0.2714	0.2526	0.2490	0.2928	0.4586	0.5154	0.4728	0.4404	0.4426	0.4868
3.2	0.2868	0.3404	0.2788	0.2686	0.2578	0.3384	0.5022	0.5578	0.5056	0.4908	0.4794	0.5428
3.4	0.3032	0.3520	0.3158	0.2882	0.2956	0.3508	0.5540	0.6074	0.5458	0.5368	0.5500	0.6048
3.6	0.3232	0.3802	0.3354	0.3256	0.3292	0.4080	0.5952	0.6504	0.6042	0.5804	0.5748	0.6582
3.8	0.3464	0.4288	0.3770	0.3568	0.3506	0.4394	0.6312	0.6988	0.6438	0.6430	0.6432	0.6874
4	0.3788	0.4668	0.3916	0.3746	0.3700	0.4696	0.6852	0.7264	0.6914	0.6626	0.6748	0.7254
4.2	0.4066	0.4882	0.4184	0.4010	0.3982	0.5006	0.7222	0.7782	0.7238	0.7128	0.7176	0.7732
4.4	0.4134	0.4986	0.4518	0.4562	0.4186	0.5394	0.7602	0.8116	0.7594	0.7594	0.7442	0.8130
4.6	0.4530	0.5470	0.4788	0.4660	0.4536	0.5776	0.7962	0.8360	0.7982	0.7978	0.7866	0.8522
4.8	0.4782	0.5688	0.5024	0.4878	0.4752	0.6016	0.8278	0.8632	0.8258	0.8184	0.8332	0.8710
5	0.5160	0.6046	0.5316	0.5186	0.5110	0.6264	0.8552	0.8832	0.8620	0.8620	0.8518	0.8794

Remark 4. For equal and unequal group variances, there are marginal gains in powers for non-normal distributions for several values of c. The gains are higher for t and lognormal distributions in most cases. However, for Laplace, exponential and Weibull distributions, there is marginal loss of power for several values of c. Power differences range from 0 to 0.17 for moderate samples and from 0.0002 to 0.09 for large samples. Sizes of the tests are close to 0.05 for all distributions.

Remark 5. Similar observations are made for AMax-T test also. For moderate samples power difference ranges from 0 to 0.11 and for large samples from 0.0002 to 0.06. Size values based on all distributions are seen to be close to 0.05. However, tests are more conservative for exponential and lognormal distributions when variances are unequal. Thus both tests are quite robust.

6 A numerical example

In this section, we illustrate the proposed methods for testing the null hypothesis of equality of treatment effects against ordered alternatives under heteroscedasticity with the help of one example. We have considered the signal strengths of four cellphone operators - Airtel, Vi India, CellOne, Jio in October 2021 from different locations of India. The raw data is taken from Open Government Data (OGD) Platform-data.gov.in-India. The original data consists of signal strengths of 150,561 Airtel, 96,875 Vi India, 6787 CellOne and 327,540 Jio operators. Data are in dBm units. A value close to 0 indicates a greater noise level. The data is sorted independently in 4 different groups of Airtel, Vi India, CellOne and Jio operator. Independent random samples of sizes 55, 48, 50 and 50 are chosen from Vi India, Airtel, Jio and CellOne groups respectively, and are shown in Table 9.

We denote operators Vi India, Airtel, Jio and CellOne as respective groups 1, 2, 3, and 4. The p-values of goodness of fit test for testing whether each group is normally distributed with two unknown parameters (mean and variance) are obtained as follows: 0.777(CellOne), 0.128(Jio), 0.128(Vi India), 0.0924(Airtel). So we conclude that the data set in each group is normally distributed with unknown mean and variance. For checking the heteroscedastic variance condition, Bartlett's test is applied. As the observed p-value at level 0.05 is 0.027094, we conclude that the variances of the groups are different.

Our aim is to test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against $H_1 : \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ with at least one strict inequality, where μ_i is the mean strength of operator i, $i = 1, 2, 3, 4$. The value of the Max-T test statistic is evaluated to be 3.684883 and the critical value corresponding to sample sizes (55, 48, 50, 50) and sample variances $(80.89, 118.35, 263.84, 210.92)$ is obtained as 2.166031 at 5% level of

Vi India		Airtel		Jio		CellOne	
-99	-86	-99	-104	-99	-55	-88	-105
-92	-99	-112	-74	-106	-90	-51	-91
-106	-85	-98	-111	-77	-69	-86	-95
-79	-83	-102	-109	-69	-99	-81	-90
-100	-112	-109	-103	-90	-96	-61	-103
-112	-95	-98	-82	-109	-65	-91	-104
-76	-93	-84	-83	-59	-62	-65	-71
-107	-101	-91	-79	-90	-103	-86	-85
-80	-104	-98	-102	-90	-93	-81	-99
-92	-89	-105	-80	-109	-79	-100	-58
-88	-90	-105	-104	-105	-86	-83	-87
-84	-98	-96	-112	-90	-90	-64	-50
-107	-78	-84	-108	-90	-58	-79	-75
-94	-109	-98	-90	-108	-102	-75	-65
-96	-89	-102	-110	-89	-65	-67	-81
-94	-105	-95	-84	-90	-65	-63	-100
-104	-104	-71	-96	-112	-106	-88	-108
-96	-107	-98	-97	-59	-112	-53	-75
-96	-84	-99		-63	-92	-75	-75
-101	-92	-94		-90	-75	-89	-72
-95	-101	-102		-78		-85	
-102	-106	-94		-99		-93	
-100	-95	-109		-108		-89	
-93	-99	-110		-93		-94	
-105	-103	-82		-86		-81	
-98		-78		-86		-73	
-103		-102		-95		-93	
-95		-108		-77		-96	
-110		-104		-81		-93	
-87		-93		-70		-76	

Table 11: Signal strength

significance. AMax-T critical value corresponding to the same configurations obtained at 5% level of significance as 2.130859. Hence H_0 is rejected here. Therefore, signal strength values of the operator's Vi India, Airtel, Jio, and CellOne are in increasing order. As negative values indicate less background noise, according to the performance the order is: Vi India $\geq$ Airtel $\geq$ Jio $\geq$ CellOne with at least one strict inequality. It is further observed using Welch's t-test on Operators 1 and 2, that they have equal means. But still proposed Max-T and AMax-T tests reject H_0. This shows the good discrimination power of proposed tests.

Acknowledgements

We are thankful to Prof. Dr. Markus Pauly, TU Dortmund for suggesting the problem and some very useful discussions. We also thank a reviewer for suggesting some modifications.

References

[1] DJ Bartholomew. "A test of homogeneity for ordered alternatives." In: *Biometrika* 46.1/2 (1959), pp. 36–48.

[2] DJ Bartholomew. "A test of homogeneity for ordered alternatives. II." In: *Biometrika* 46.3/4 (1959), pp. 328–335.

[3] Peter J Bickel and David A Freedman. "Some asymptotic theory for the bootstrap." In: *The Annals of Statistics* 9.6 (1981), pp. 1196–1217.

[4] Thomas A Bishop and Edward J Dudewicz. "Exact analysis of variance with unequal variances: test procedures and tables." In: *Technometrics* 20.4 (1978), pp. 419–430.

[5] Thomas A Bishop and Edward J Dudewicz. "Heteroscedastic Anova." In: *Sankhya: The Indian Journal of Statistics, Series B* (1981), pp. 40–57.

[6] Frank Bretz. "An extension of the Williams trend test to general unbalanced linear models." In: *Computational Statistics and Data Analysis* 50.7 (2006), pp. 1735–1748.

[7] Paul Cabilio and Jianan Peng. "Multiple rank-based testing for ordered alternatives with incomplete data." In: *Statistics and Probability Letters* 78.16 (2008), pp. 2609–2613.

[8] Shun-Yi Chen. "One-sided range test for testing against an ordered alternative under heteroscedasticity." In: *Communications in Statistics-Simulation and Computation* 29.4 (2000), pp. 1255–1272.

[9] Shun-Yi Chen. "One-stage and two-stage statistical inference under heteroscedasticity." In: *Communications in Statistics-Simulation and Computation* 30.4 (2001), pp. 991–1009.

[10] Theodore S Donaldson. "Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio." In: *Journal of the American Statistical Association* 63.322 (1968), pp. 660–676.

[11] Mario Hasler. "Heteroscedasticity: multiple degrees of freedom vs. sandwich estimation." In: *Statistical Papers* 57.1 (2016), pp. 55–68.

[12] Esther Herberich, Johannes Sikorski, and Torsten Hothorn. "A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs." In: *PloS one* 5.3 (2010), e9788.

[13] GS James. "The comparison of several groups of observations when the ratios of the population variances are unknown." In: *Biometrika* 38.3/4 (1951), pp. 324–329.

[14] AR Jonckheere. "A test of significance for the relation between m rankings and k ranked categories." In: *British Journal of Statistical Psychology* (1954).

[15] Kalimuthu Krishnamoorthy, Fei Lu, and Thomas Mathew. "A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models." In: *Computational Statistics and Data Analysis* 51.12 (2007), pp. 5731–5742.

[16] Ruth Marcus. "The powers of some tests of the equality of normal means against an ordered alternative." In: *Biometrika* 63.1 (1976), pp. 177–183.

[17] G S Mudholkar and M P McDermott. "A class of tests for equality of ordered means." In: *Biometrika* 76.1 (1989), pp. 161–168.

[18] Kimihiro Noguchi et al. "Permutation tests are robust and powerful at 0.5% and 5% significance levels." In: *Behavior Research Methods* 53.6 (2021), pp. 2712–2724.

[19] Markus Pauly, Thomas Asendorf, and Frank Konietschke. "Permutation-based inference for the AUC: A unified approach for continuous and discontinuous data." In: *Biometrical Journal* 58.6 (2016), pp. 1319–1337.

[20] Markus Pauly, Edgar Brunner, and Frank Konietschke. "Asymptotic permutation tests in general factorial designs." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.2 (2015), pp. 461–473.

[21] Gary C Ramseyer and Tse-Kia Tcheng. "The robustness of the studentized range statistic to violations of the normality and homogeneity of variance assumptions." In: *American Educational Research Journal* 10.3 (1973), pp. 235–240.

[22] Guogen Shan, Daniel Young, and Le Kang. "A new powerful nonparametric rank test for ordered alternative problem." In: *PloS one* 9.11 (2014), e112924.

[23] Eryl Shirley. "A non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment." In: *Biometrics* (1977), pp. 386–389.

[24] Galen R Shorack. "Testing against ordered alternatives in model I analysis of variance: normal theory and nonparametric." In: *The Annals of Mathematical Statistics* 38.6 (1967), pp. 1740–1752.

[25] ABL Srivastava. "Effect of non-normality on the power function of t-test." In: *Biometrika* 45.3/4 (1958), pp. 421–430.

[26] Samaradasa Weerahandi. "ANOVA under unequal error variances." In: *Biometrics* (1995), pp. 589–599.

[27] Bernard Lewis Welch. "On the comparison of several mean values: an alternative approach." In: *Biometrika* 38.3/4 (1951), pp. 330–336.

[28] DA Williams. "A test for differences between treatment means when several dose levels are compared with a zero dose control." In: *Biometrics* (1971), pp. 103–117.

[29] DA Williams. "The comparison of several dose levels with a zero dose control." In: *Biometrics* (1972), pp. 519–531.

[30] Guoyi Zhang. "A parametric bootstrap approach for one-way ANOVA under unequal variances with unbalanced data." In: *Communications in Statistics-Simulation and Computation* 44.4 (2015), pp. 827–832.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

The Theory of k-NN Classification Rules: T.M. Cover and P.E. Hart (1967) Results and Extensions Revisited

Subhash Bagui [1] and KL Mehra [2]

1. *Department of Mathematics, The University of West Florida, Pensacola, FL 32514, Email: sbagui@uwf.edu*
2. *Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada, T6G 2G1, Email: kmehra@ualberta.ca*

Received: 11 May 2022 / Revised: 11 Jun 2022 / Accepted: 21 Jun 2022

Abstract

In the current data science era, k-NN rules are routinely used in machine learning problems without knowing their theoretical properties and asymptotic behavior. Cover and Hart [11] were the first to derive and examine the asymptotic risk of this rule. They studied this rule in a Bayesian framework and obtained (natural) upper and lower bounds on its asymptotic risk in terms of the optimal Bayes risk. The authors also showed that as k increases, the asymptotic risk decreases. This was a groundbreaking study, and we believe the NN Classification theory must be included in the data science curriculum and any multivariate analysis books in statistics. Keeping these in mind, this paper presents Cover and Hart ([11]) results and related material in such a way that their material is easily understood by students and future authors – and researchers - who can include this material in their multivariate statistics books. Using a combinatorial identity, we give a new proof of the result that the asymptotic risk of the k-NN rule decreases as k increases indefinitely. The rule has been widely tested on a variety of real-world data sets. For the benefit of readers, we present here the performance outcome of this rule on the well-known IRIS (Anderson [2]) data sets.

Keywords: Asymptotic risk, Bayes' risk, Nearest Neighbor Classification, Cover and Hart Inequality, Probability Error.
AMS subject classification: Primary 62H20, Secondary 62G07, 62G08

1 Introduction

The basic problem in classification/discrimination is to correctly classify an unknown observation Z,known to be from one of the s given populations (or classes) $\pi_1, \pi_2, \cdots, \pi_s$, to its right parent population [6]. The probability distributions of all these populations may be fully known to the statistician. In this event, the problem will fall in the category of parametric classification and Bayes procedure produces an optimal solution. On the other extreme, the probability distributions of these populations may be entirely unknown excepting, however, that an identified training sample is available from each one of them. The problem then falls under the domain of nonparametric classification, and quite naturally, no optimal solution exists in this case with respect to the unknown underlying distributions. Here is, in this latter type of scenarios, that one can profitably employ the NN type classification methodology.
The nearest neighbor (NN) type classification rule was first introduced by Fix and Hodges [3]. Their proposed rule was for only two ($s = 2$) populations and was based on the two available identified (training) samples - one from each of the two given populations - and is described for a fixed integer

© 2022 Author(s). (https://www.thegsa.in/).

k as follows: Their rule classifies Z to the population π_1 if $k_1/n_1 > k_2/n_2$, where $k_i = \#$ of observations from the population π_i $(i = 1, 2)$ among the k=k_1+k_2 observations nearest (determined using a distance function) to Z, and n_1, n_2 are the corresponding sample sizes with $n = n_1 + n_2$ denoting the total sample size. Cover and Hart ([11]) employed a slightly modified version of the above rule. Their procedure is based on an identified training sample from a mixture of s populations $\pi_1, \pi_2, \cdots, \pi_s$. For $s = 2$, the rule classifies Z to π_i $(i = 1, 2)$ if $k_i = \max\{k_1, k_2\}$; and for general s, it classifies Z to π_i $(i = 1, 2, \cdots, s)$ if $k_i = \max\{k_1, k_2, \cdots, k_s\}$ where $k_i = \#$ of observations from π_i $(i = 1, 2, \cdots, s)$ from among the k=k_1+k_2+$\cdots$+k_s observations nearest to Z.

In applying k-NN classification rule, in case of ties among two or more k_j's at the top, one may employ appropriate randomization in assigning $Z$to one of the tied classes or simply shift k to a different close enough workable value.

Preliminaries. Let (X_1, θ_1), (X_2, θ_2), ..., (X_n, θ_n) be independent identically distributed (i.i.d.) random variables (r.v.'s) taking values in $R^d \times \{1, 2, \cdots, s\}$ where $\{\theta_i\}$, $i = 1, 2, \cdots, n$ are i.i.d. r.v.'s with $P(\theta_i = j)$=ξ_j, $j = 1, 2, \cdots, s$, and $\sum_{j=1}^{s} \xi_j = 1$. The ξ_j's are called the prior probabilities associated with populations π_j, $j = 1, 2, \cdots, s$. If $\theta_i = j$, then X_i is identified with the population π_j, which is assumed to have a distinct probability density f_j, for each $j = 1, 2, \cdots, s$, with respect to Lebesgue measure μ. Let $(Z,\ \theta)$ be an observation taking value in $R^d \times \{1, 2, \cdots, s\}$ where only Z is observable. The object is to classify Z to one of the given s populations based on the information contained in the training sample. For the classification of the observation Z, it is desired to estimate θ by utilizing the information contained in the (identified) training sample$\{(X_i, \theta_i) : j = 1, 2, \cdots, n\}$. By identified sample, we mean that the value of each θ_i is assumed to be given or known 'a priori'. This is the Bayesian framework for which we also need a loss function. Let θ'_n be an estimate of θ and let $L(\theta, \theta'_n)$ denote the loss incurred in estimating θ with θ'_n. Then by Bayes theorem, the posterior probability of $(\theta = i)$ given $Z = z$ is

$$\eta_i(z) = P(\theta = i|Z = z) = \xi_i f_i(z) \Big/ \sum_{j=1}^{s} \xi_j f_j(z), i = 1, 2, \cdots, s. \tag{1}$$

If a classification rule δ assigns Z to the population π_j, then the conditional expected loss, i.e., conditional risk, given $Z = z$, is clearly

$$r_j(z) = E[L(\theta, \theta'_n)|Z = z] = \sum_{i=1}^{s} L(i, j)\eta_i(z). \tag{2}$$

The rule that minimizes this conditional risk $r_j(z)$ is called the Bayes rule. For the Bayes rule δ^*, therefore, the conditional (minimum) Bayes risk $r^*(z)$ is given by

$$r^*(z) = \min_j \{\sum_{i=1}^{s} L(i, j)\eta_i(z)\}; \tag{3}$$

and thereby the overall (minimum) Bayes risk R_s^* by

$$R_s^* = E[r^*(Z)] = \int r^*(z)f(z)dz, \tag{4}$$

where

$$f(z) = \sum_{j=1}^{s} \xi_j f_j(z) \tag{5}$$

denotes the marginal density of Z. In this article, we take $L(\theta, \theta'_n)$ as the $0-1$ loss function, i.e.,

$$L(\theta, \theta'_n) = \begin{cases} 0 & \text{if } \theta = \theta'_n \\ 1 & \text{if } \theta \neq \theta'_n. \end{cases} \tag{6}$$

Using this $0-1$ loss function, the expressions (3) and (4), the conditional Bayes risk and the overall Bayes risk can be expressed, respectively, as

$$\begin{aligned} r^*(z) &= \min\{\sum_{j\neq 1} \eta_j(z), \sum_{j\neq 2} \eta_j(z), \cdots, \sum_{j\neq s} \eta_j(z)\} \\ &= \min\{1-\eta_1(z), 1-\eta_2(z), \cdots, 1-\eta_s(z)\}, \end{aligned} \tag{7}$$

and

$$\begin{aligned} R^*_s = E[r^*(Z)] &= \int \min(1-\eta_1(z), 1-\eta_2(z), \cdots, 1-\eta_s(z)) f(z) dz \\ &= \int \min(\sum_{j\neq 1} \xi_j f_j(z), \sum_{j\neq 2} \xi_j f_j(z), \cdots, \sum_{j\neq s} \xi_j f_j(z)) dz. \end{aligned} \tag{8}$$

Nearest Neighbor (NN) classification rule

Let $\{(X_i, \theta_i)\}$, $i = 1, 2, \cdots, n$ be an identified training sample from a mixture of s populations with a mixture proportion $P(\theta_i = j) = \xi_j$, $j = 1, 2, \cdots, s$. We call $X'_{1n} \in \{X_1, X_2, \cdots, X_n\}$ the first nearest neighbor of Z if

$$\min_{i\in\{1,2,\cdots,n\}} \|X_i - Z\| = \|X'_{1n} - Z\|.$$

Let θ'_{1n} denote the class of X'_{1n}. Thus, the 1-NN rule classifies Z to the class of X'_{1n}, i.e., to the class θ'_{1n}. Let k_j denote the number of observations from the population π_j among $k = k_1 + k_2 + \cdots + k_s$ observations among $X_1, X_2, \cdots, X_n$ that are nearest to Z. The $k-$NN rule classifies Z to population π_i, if $k_i = \max\{k_1, k_2, \cdots, k_s\}$ with randomization in case of a tie at the top. The NN rule utilizes only the k nearest neighbors of Z , and the remaining $(n-k)$ observations in the (identified) training sample are ignored.

Suppose the proposed NN classification rule δ_n assigns Z to a category θ'_n. If θ is indeed the true category of Z, the NN classification rule incurs loss $L(\theta, \theta'_n)$. We define NN risk $R_s(\delta_n)$ by its expectation

$$R_s(\delta_n) = E[L(\theta, \theta'_n), \tag{9}$$

and the limiting (large sample), as $n \to \infty$, k-NN risk R_s for fixed k is given by

$$R_s = \lim_{n\to\infty} E[L(\theta, \theta'_n]. \tag{10}$$

The aim is to study the asymptotic (large sample) behavior of this k- NN risk. Let us first show that the 1-NN X'_{1n} converges to Z *a.s.*, as $n \to \infty$. This result will be needed later to obtain asymptotic (overall) NN risk.

Lemma 1.1. *Let Z and $(X_1, \theta), (X_2, \theta), \cdots, (X_n, \theta)$ be independent identically distributed random variables taking values on $R^d \times \{1, 2, \cdots, s\}$. Let X'_{1n} denote the 1-NN of Z derived from the sample $\{X_1, X_2, \cdots, X_n\}$. Then,*

$$X'_{1n} \xrightarrow{a.s.} Z, \text{ as } n \to \infty \tag{11}$$

Proof. For any $\varepsilon > 0$, we have in view of the independence of $Z$and X_i's that

$$\begin{aligned} P\left(\|X'_{1n} - Z\| > \varepsilon\right) &= P\left(\min_i \|X_i - Z\| > \varepsilon\right) \\ &= EP\left(\|X_1 - Z\| > \varepsilon, \|X_2 - Z\| > \varepsilon, \cdots, \|X_n - Z\| > \varepsilon | Z\right) \\ &= \left[E\{P\left(\|X_1 - Z\|\right) > \varepsilon | Z\}\right]^n \\ &= \left[1 - EP\left(\|X_1 - Z\| \leq \varepsilon | Z\right)\right]^n \to 0 \end{aligned} \tag{12}$$

as $n \to \infty$ by the Dominated Convergence theorem. Since the set $\{\|X'_{1n} - Z\| > \varepsilon\}$ is monotonically decreasing in n, it follows that

$$\lim_{n\to\infty} P\left[\bigcup_{l\geq n} \{\|X'_{1l} - Z\| > \varepsilon\}\right] = \lim_{n\to\infty} P\left(\|X'_{1n} - Z\| > \varepsilon\right) = 0$$

by convergence (12). Thus, by the standard *a.s.* convergence criterion, we can conclude that $X'_{1n} \xrightarrow{a.s.} Z$, as $n \to \infty$.

In Section 2, the upper and lower bounds on the asymptotic risk of the 1-NN rule in terms of optimal Bayes risk are given. In Section 3, we have given an alternative derivation of the asymptotic risk of the k-NN rule that provides a new proof for the decreasing nature of its asymptotic risk with increasing k using a combinatorial identity. The results of an empirical study are reported in Section 4. Finally, Section 5 contains some useful concluding remarks.

2 Asymptotic risk of 1-NN Classification Rule and its bounds when s = 2(Cover and Hart [11])

First, consider the case when $s = 2$. In order to classify Z, we estimate its class θ using information contained in the two (identified) training samples. Let 1-NN rule estimate θ by θ'_{1n}.
Thus, the conditional risk of 1-NN rule given $Z = z$and $X'_{1n} = x'_{1n}$ is given by

$$r(z; x'_{1n}) = E\left[L(\theta, \theta'_{1n}) | Z = z, X'_{1n} = x'_{1n}\right] = P\left(\theta \neq \theta'_{1n} | Z = z, X'_{1n} = x'_{1n}\right)$$

$$= P\left(\theta = 1 \cap \theta'_{1n} = 2 | Z = z, X'_{1n} = x'_{1n}\right) + P\left(\theta = 2 \cap \theta'_{1n} = 1 | Z = z, X'_{1n} = x'_{1n}\right). \tag{13}$$

Upon using the conditional independence of θ and θ'_{1n}, the equation (13) can be written as

$$\begin{aligned} r(z; x'_{1n}) &= P(\theta = 1 | Z = z) P(\theta'_{1n} = 2 | X'_{1n} = x'_{1n}) + P(\theta = 2 | Z = z) P(\theta'_{1n} = 1 | X'_{1n} = x'_{1n}) \\ &= \eta_1(z)\eta_2(x'_{1n}) + \eta_2(z)\eta_1(x'_{1n}), \end{aligned} \tag{14}$$

where η_i, $i = 1, 2$, is given in (1).
Now we state and prove a theorem on limiting 1-NN risk.

Theorem 2.1. *Suppose z is a continuity point of f_i, $i = 1, 2$. Then the limiting conditional risk given $Z = z$ and the unconditional risk of 1-NN Rule are given, respectively, by*

$$r(z) = \lim_{n\to\infty} r(z; x'_{1n}) = 2\eta_1(z)\eta_2(z). \tag{15}$$

and

$$R_2 = E[r(Z)] = 2E\left[\eta_1(Z)\eta_2(Z)\right], \tag{16}$$

where $\eta_2(z) = 1 - \eta_1(z)$, when $s = 2$.

Proof. Using Lemma 1.1, the continuity of f_i's at z and equation (14), we obtain

$$r(z) = \lim_{n\to\infty} r(z; x'_{1n}) = \eta_1(z) \cdot \lim_{n\to\infty} \eta_2(x'_{1n}) + \eta_2(z) \cdot \lim_{n\to\infty} \eta_1(x'_{1n})$$
$$= \eta_1(z)\eta_2(z) + \eta_2(z)\eta_1(z) = 2\eta_1(z)\eta_2(z). \tag{17}$$

By the Dominated Convergence theorem, therefore, we can conclude that

$$R_2 = \lim_{n\to\infty} E\left[r(Z; x'_{1n})\right] = E\left[r(Z)\right] = 2E\left[\eta_1(Z)\eta_2(Z)\right]. \tag{18}$$

This completes the proof.

Below we state a theorem giving lower and upper bounds on R_s with $s = 2$.

Theorem 2.2. *The asymptotic risk $R_2$of 1-NN Rule has the following upper and lower bounds:*

$$R_2^* \le R_2 \le 2R_2^*(1 - R_2^*),$$

where R_2^* is the (optimal) Bayes risk defined in (8) *for $s = 2$.*

Proof. In view of equation (16), we have

$$r(z) = \eta_1(z)\eta_2(z) + \eta_2(z)\eta_1(z) \ge \min\{\eta_1(z), \eta_2(z)\} = \min\{\eta_1(z), 1 - \eta_1(z)\}. \tag{19}$$

Now expectations on both sides of (19) and by equation (8) with $s = 2$, we have

$$R_2 = E\left[r(Z)\right] \ge \int \min\{\eta_1(z), 1 - \eta_1(z)\} f(z)dz = R_2^*. \tag{20}$$

In view of equation (7) for $s = 2$, the conditional Bayes risk is

$$r^*(z) = \min\{1 - \eta_1(z), 1 - \eta_2(z)\} = \min\{\eta_1(z), \eta_2(z)\} = \min\{\eta_1(z), 1 - \eta_1(z)\}$$

which is symmetric in η_1 and η_2, since $\eta_1 + \eta_2 = 1$. Thus, we may write

$$\eta_1(z)\eta_2(z) = \eta_1(z)(1 - \eta_1(z)) = r^*(z)(1 - r^*(z)), \tag{21}$$

so that by equations (18), (21) and Jensen's inequality, we have

$$R_2 = 2E\left[\eta_1(Z)(1 - \eta_1(Z))\right] = 2E\left[r^*(Z)(1 - r^*(Z))\right] = 2[E(r^*(Z)) - E(r^*(Z))^2]$$
$$\le 2[E(r^*(Z)) - (E(r^*(Z)))^2] = 2[R_2^* - R_2^{*2}] = 2R_2^*(1 - R_2^*). \tag{22}$$

Combining equations (20) and (22), we arrive at

$$R_2^* \le R_2 \le 2R_2^*(1 - R_2^*).$$

The proof is complete.

Example 1. *Suppose the r.v. Z has the beta density either $f_1(z) = 2z$ or $f_2 = 2(1-z)$, for $0 \le z \le 1$; and $\xi_1 = \xi_2 = 1/2$. From the equation* (18),

$$R_2 = 2E\left[\eta_1(Z)\eta_2(Z)\right] = 2\int \frac{\xi_1\xi_2 f_1(z)f_2(z)}{\xi_1 f_1(z) + \xi_2 f_2(z)}dz = 2\int_0^1 z(1 - z)dz = 2\frac{\Gamma(2)\Gamma(2)}{\Gamma(4)} = \frac{2}{6} = \frac{1}{3}.$$

From (8) *with $s = 2$,*

$$R_2^* = \int \min(\xi_1 f_1(z), \xi_2 f_2(\xi_2))dz = \int_0^1 [z \wedge (1 - z)]dz$$

$$= \int_{z<1-z} zdz + \int_{z>1-z} (1 - z)dz = \int_0^{1/2} zdz + \int_{1/2}^1 (1 - z)dz = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.$$

Note that $R_2^* = \frac{1}{4} \le R_2 = \frac{1}{3} \le 2R_2^*(1 - R_2^*) = \frac{3}{8} = 0.375$. Thus, the result of Theorem 2.2 is verified.

Now we extend the results of Theorem 2.2 to s populations in Theorem 2.3 below:

Theorem 2.3. *The asymptotic risk $R_s$of 1-NN Rule has the following upper and lower bounds:*

$$R_s^* \leq R_s \leq R_s^*(2 - \tfrac{s}{s-1}R_s^*),$$

where R_s^* is the (optimal) Bayes risk defined in (8) *for s populations.*

Proof. The conditional 1-NN rule risk under s populations is

$$r(z; x'_{1n}) = E\left[L(\theta, \theta'_{1n})|Z = z, X'_{1n} = x'_{1n}\right] = \sum_{i=1}^{s}\sum_{j\neq i}^{s} P\left[\theta = i \cap \theta'_{1n} = j|Z = z, X'_{1n} = x'_{1n}\right] \tag{23}$$

Since θ and θ'_{1n} are conditionally independent given $Z$and X'_{1n}, the equation (23) can be rewritten as

$$r(z; x'_{1n}) = \sum_{i=1}^{s}\sum_{j\neq i}^{s} P\left[\theta = i|z\right] P\left[\theta'_n = j|x'_{1n}\right] = \sum_{i=1}^{s}\sum_{j\neq i}^{s} \eta_i(z)\eta_j(x'_{1n}). \tag{24}$$

Since $X'_{1n} \to Z$ *a.s.* by Lemma 1.1 and f_j's are continuous *a.e.*, the posterior probability $\eta_j(x'_{1n}) \to \eta_j(z)$ *a.e.* Thus, the conditional risk given in equation (24) converges, as $n \to \infty$, to the asymptotic conditional risk

$$r(z) = \sum_{i=1}^{s}\sum_{j\neq i}^{s} \eta_i(z)\eta_j(z) = 1 - \sum_{j=1}^{s} \eta_j^2(z). \tag{25}$$

Let $\eta_l(z) = \max_j \{\eta_1(z), \eta_2(z), \cdots, \eta_s(z)\}$. Thus, the conditional Byes risk $r^*(z)$ given in (7) can be expressed as

$$r^*(z) = 1 - \eta_l(z), \tag{26}$$

so that applying Cauchy-Schwartz inequality to $\left[\sum_{j\neq l}^{s} \eta_j(z)\right]^2$ and using equation (26), we have

$$(s-1)\sum_{j\neq l}^{s} \eta_j^2(z) \geq \left[\sum_{j\neq l}^{s} \eta_j(z)\right]^2 = [1 - \eta_l(z)]^2 = [r^*(z)]^2. \tag{27}$$

Now adding the term $(s-1)\eta_l^2(z)$ on both sides in the above equation (27) and using $\eta_l(z) = 1 - r^*(z)$, we get

$$\sum_{j=1}^{s} \eta_j^2(z) \geq \frac{[r^*(z)]^2}{(s-1)} + [1 - r^*(z)]^2 = 1 - 2r^*(z) + \left(\tfrac{s}{s-1}\right)[r^*(z)]^2. \tag{28}$$

Combining the equations (25) and (28), we arrive at the following expression

$$r(z) \leq 2r^*(z) - \frac{s}{(s-1)}[r^*(z)]^2. \tag{29}$$

Now taking expectations on both sides of the equation (29) and applying Jensen's inequality, we obtain

$$\begin{aligned} R_s &\leq 2E[r^*(Z)] - \frac{s}{(s-1)}E[r^*(Z)]^2 \leq 2R_s^* - \frac{s}{(s-1)}[E(r^*(Z)]^2 \\ &= 2R_s^* - \frac{s}{(s-1)}R_s^{*2} = R_s^*(2 - \frac{s}{(s-1)}R_s^*), \end{aligned} \tag{30}$$

which gives the upper bound. To get a lower bound on R_s, note that – since the minimum is always less or equal to the mean and $\sum_{i=1}^{s} \eta_i(z) = 1$

$$\min\left\{\sum_{i\neq 1}\xi_i f_i(z), \sum_{i\neq 2}\xi_i f_i(z), \cdots, \sum_{i\neq s}\xi_i f_i(z)\right\} \leq \sum_{i=1}^{s}\eta_i(z)\sum_{j\neq i}^{s}\xi_j f_j(z). \tag{31}$$

Now dividing with $\sum_{i=1}^{s}\xi_i f_i(z)$ and then integrating with respect to z on both sides of the equation (31), the resulting equation at once yields (see equations (5), (8) and (25)) that

$$R_s^* \leq R_s. \tag{32}$$

The proof is complete in view of (30) and (32).

3 The k-NN Rule and its Asymptotic risk

We consider the case when $s = 2$. It is possible to show that kth NN of Z, which we denote by X'_{kn}, converges to Z. That is $X'_{kn} \xrightarrow{a.s.} Z$ as $n \to \infty$ for fixed k.

Lemma 3.1. *Let Z and $(X_1, \theta), (X_2, \theta), \cdots, (X_n, \theta)$ be independent identically distributed random variables taking values on $R^d \times \{1, 2, \cdots, s\}$. Let X'_{kn} denote the k-NN of Z derived from the sample $\{X_1, X_2, \cdots, X_n\}$. Then, $X'_{kn} \xrightarrow{a.s.} Z$, as $n \to \infty$.*

Proof. For any $\varepsilon > 0$, we have

$$\begin{aligned}
P(\|X'_{kn} - Z\| > \varepsilon) &= P(\|X'_{kn} - Z\| > \varepsilon, \|X'_{1n} - Z\| > \varepsilon) + P(\|X'_{kn} - Z\| > \varepsilon, \|X'_{1n} - Z\| \leq \varepsilon) \\
&\leq P(\|X_1 - Z\| > \varepsilon, \|X_2 - Z\| > \varepsilon, \cdots, \|X_n - Z\| > \varepsilon) \\
&\quad + P\left(\|X_{i_1} - Z\| > \varepsilon, \|X_{i_2} - Z\| > \varepsilon, \cdots, \left\|X_{i_{n-k+1}} - Z\right\| > \varepsilon\right) \\
&= [EP(\|X_1 - Z\|) > \varepsilon|Z]^n + [EP(\|X_{i_1} - Z\|) > \varepsilon|Z]^{n-k+1} \\
&= [1 - EP(\|X_1 - Z\| \leq \varepsilon|Z)]^n + [1 - EP(\|X_{i_1} - Z\| \leq \varepsilon|Z)]^{n-k+1} \\
&\to 0, \text{ as } n \to \infty.
\end{aligned} \tag{33}$$

Since the set $\{\|X'_{kn} - Z\| > \varepsilon\}$ is monotonically decreasing in n, it follows that

$$\lim_{n\to\infty} P\left[\bigcup_{j\geq n}\left\{\left\|X'_{kj} - Z\right\| > \varepsilon\right\}\right] = \lim_{n\to\infty} P(\|X'_{kn} - Z\| > \varepsilon) = 0$$

by (33). Thus, by the standard *a.s.* convergence criterion, we can conclude that $X'_{kn} \xrightarrow{a.s.} Z$, as $n \to \infty$. The proof of Lemma 3.1 is complete.

Let k-NN rule estimate θ by θ'_{kn}. Then the conditional risk of k-NN rule, given $Z = z$and $X'_{kn} = x'_{kn}$, is given by

$$r(z; x'_{kn}) = E[L(\theta, \theta'_{kn})|Z = z, X'_{kn} = x'_{kn}] = P(\theta \neq \theta'_{kn}|Z = z, X'_{kn} = x'_{kn})$$

$$= P(\theta = 1 \cap \theta'_{kn} = 2|Z = z, X'_{kn} = x'_{kn}) + P(\theta = 2 \cap \theta'_{kn} = 1|Z = z, X'_{kn} = x'_{kn}). \tag{34}$$

Upon using the conditional independence of θ and θ'_n, equation (34) can be written as

$$r(z;x'_{kn}) = P(\theta=1|Z=z)P(\theta'_{kn}=2|X'_{kn}=x'_{kn}) + P(\theta=2|Z=z)P(\theta'_{kn}=1|X'_{kn}=x'_{kn}), \quad (35)$$

where for $k=2m+1$,

$$P(\theta'_{kn}=2|X'_{kn}=x'_{kn}) = \sum_{i=0}^{m} \binom{2m+1}{i} \eta_1^i(x'_{kn})\eta_2^{2m+1-i}(x'_{kn}) \quad (36)$$

and

$$P(\theta'_{kn}=1|X'_{kn}=x'_{kn}) = \sum_{i=m+1}^{2m+1} \binom{2m+1}{i} \eta_1^i(x'_{kn})\eta_2^{2m+1-i}(x'_{kn}). \quad (37)$$

Thus, in view of equations (1), (36), and (37), the equation (35) can be rewritten as

$$r(z;x'_{kn}) = \eta_1(z)\sum_{i=0}^{m} \binom{2m+1}{i} \eta_1^i(x'_{kn})\eta_2^{2m+1-i}(x'_{kn}) + \eta_2(z)\sum_{i=m+1}^{2m+1} \binom{2m+1}{i} \eta_1^i(x'_{kn})\eta_2^{2m+1-i}(x'_{kn}); \quad (38)$$

similarly, when $k=2m+2$, the

$$P(\theta'_{kn}=2|X'_{kn}=x'_{kn}) = \sum_{i=0}^{m} \binom{2m+2}{i} \eta_1^i(x'_{kn})\eta_2^{2m+2-i}(x'_{kn}) + \tfrac{1}{2}\binom{2m+2}{m+1} \eta_1^{m+1}(x'_{kn})\eta_2^{m+1}(x'_{kn})$$

and

$$P(\theta'_{kn}=1|X'_{kn}=x'_{kn}) = \sum_{i=m+2}^{2m+2} \binom{2m+2}{i} \eta_1^i(x'_{kn})\eta_2^{2m+2-i}(x'_{kn}) + \tfrac{1}{2}\binom{2m+2}{m+1} \eta_1^{m+1}(x'_{kn})\eta_2^{m+1}(x'_{kn}).$$

The equations (35) in this case will have the form

$$\begin{aligned} r(z;x'_{kn}) &= \eta_1(z)[\sum_{i=0}^{m} \binom{2m+2}{i} \eta_1^i(x'_{kn})\eta_2^{2m+2-i}(x'_{kn}) + \tfrac{1}{2}\binom{2m+2}{m+1} \eta_1^{m+1}(x'_{kn})\eta_2^{m+1}(x'_{kn})] \\ &\quad +\eta_2(z)[\sum_{i=m+2}^{2m+2} \binom{2m+2}{i} \eta_1^i(x'_{kn})\eta_2^{2m+2-i}(x'_{kn}) + \tfrac{1}{2}\binom{2m+2}{m+1} \eta_1^{m+1}(x'_{kn})\eta_2^{m+1}(x'_{kn})]. \quad (39) \end{aligned}$$

We now state

Theorem 3.1. *Suppose z is a continuity point of f_i, $i=1,2$. Then for $s=2$, the limiting conditional risk given $Z=z$ and the unconditional risk of the k-NN rule are given, respectively, by*

(I) for $k=2m+1$,

$$r_k(z) = \lim_{n\to\infty} r(z;x'_{kn}), \quad (40)$$

with

$$r_k(z) = \eta_1(z)\sum_{i=0}^{m} \binom{2m+1}{i} \eta_1^i(z)\eta_2^{2m+1-i}(z) + \eta_2(z)\sum_{i=m+1}^{2m+1} \binom{2m+1}{i} \eta_1^i(z)\eta_2^{2m+1-i}(z) \quad (41)$$

and for $k = 2m+2$,

$$r_k(z) = \eta_1(z)[\sum_{i=0}^{m} \binom{2m+2}{i} \eta_1^i(z)\eta_2^{2m+2-i}(z) + \tfrac{1}{2}\binom{2m+2}{m+1}\eta_1^{m+1}(z)\eta_2^{m+1}(z)]$$
$$+\eta_2(z)[\sum_{i=m+2}^{2m+2} \binom{2m+2}{i} \eta_1^i(z)\eta_2^{2m+2-i}(z) + \tfrac{1}{2}\binom{2m+2}{m+1}\eta_1^{m+1}(z)\eta_2^{m+1}(z)]; \quad (42)$$

and

(II) $R_2(k) = E[r_k(Z)]$, *for* $s = 2$.

Proof. Since $X'_{kn} \to Z$ *a.s.*, as $n \to \infty$, so that conditionally $x'_{kn} \to z$ and f_j's *a.e.* continuous, the posterior probability $\eta_j(x'_{kn}) \to \eta_j(z)$ *a.e.* for $j = 1, 2$. Thus, the conditional risk as given in equation (24) converges, as $n \to \infty$, to asymptotic conditional risk, which for $k = 2m+1$ from (38), is given by the expression

$$r_k(z) = \eta_1(z)\sum_{i=0}^{m} \binom{2m+1}{i} \eta_1^i(z)\eta_2^{2m+1-i}(z) + \eta_2(z)\sum_{i=m+1}^{2m+1} \binom{2m+1}{i} \eta_1^i(z)\eta_2^{2m+1-i}(z) \quad (43)$$

and for $k = 2m+2$ from (39), by the expression

$$r_k(z) = \eta_1(z)[\sum_{i=0}^{m} \binom{2m+2}{i} \eta_1^i(z)\eta_2^{2m+2-i}(z) + \tfrac{1}{2}\binom{2m+2}{m+1}\eta_1^{m+1}(z)\eta_2^{m+1}(z)]$$
$$+\eta_2(z)[\sum_{i=m+2}^{2m+2} \binom{2m+2}{i} \eta_1^i(z)\eta_2^{2m+2-i}(z) + \tfrac{1}{2}\binom{2m+2}{m+1}\eta_1^{m+1}(z)\eta_2^{m+1}(z)] \quad (44)$$

This completes the first part(I) of the proof. The unconditional risk $R_2(k)$ is obtained by taking expectations with respect to Z, i.e., with respect to the distribution with density $f(z) = \xi_1 f_1(z) + \xi_2 f_2(z)$. This gives in part(II) the unconditional risk as $R_2(k) = E[r_k(Z)]$. The proof is complete.

The conditional risks given in (43) and (44) are, in fact, equal. This is proved in the following theorem.

Theorem 3.2. *The asymptotic conditional risk* $r_k(z)$ *has the following properties:*
(a) $r_{2m+1} = r_{2m+2}$ *for* $m = 0, 1, 2, \cdots$ *and (b)* $r_{2m+1} > r_{2m+3}$ *for* $m = 0, 1, 2, \cdots$
The above (a) and (b) imply that $r_{2m+2} > r_{2m+4}$ *for* $m = 0, 1, 2, \cdots$.

Proof. For notational simplicity, we will use $\eta_i(z) = \eta_i$ for $i = 1, 2$. For part (a), we will simplify the conditional risk $r_{2m+2}(z)$given in (44) in terms of $r_{2m+1}(z)$.
The conditional risk $r_{2m+2}(z)$from (44) can be written as

$$\begin{aligned} r_{2m+2}(z) &= \eta_1[\sum_{i=0}^{m} \binom{2m+2}{i} \eta_1^i\eta_2^{2m+2-i} + \tfrac{1}{2}\binom{2m+2}{m+1}\eta_1^{m+1}\eta_2^{m+1}] \\ &\quad +\eta_2[\sum_{i=m+2}^{2m+2} \binom{2m+2}{i} \eta_1^i\eta_2^{2m+2-i} + \tfrac{1}{2}\binom{2m+2}{m+1}\eta_1^{m+1}\eta_2^{m+1}] \quad (45) \\ &= \eta_1\sum_{i=0}^{m} \binom{2m+2}{i} \eta_1^i\eta_2^{2m+2-i} + \eta_2\sum_{i=m+2}^{2m+2} \binom{2m+2}{i} \eta_1^i\eta_2^{2m+2-i} + \tfrac{1}{2}\binom{2m+2}{m+1}\eta_1^{m+1}\eta_2^{m+1}(\eta_1+\eta_2) \\ &= \eta_1\sum_{i=0}^{m} \binom{2m+2}{i} \eta_1^i\eta_2^{2m+2-i} + \eta_2\sum_{i=m+2}^{2m+2} \binom{2m+2}{i} \eta_1^i\eta_2^{2m+2-i} + \tfrac{1}{2}\binom{2m+2}{m+1}\eta_1^{m+1}\eta_2^{m+1}. \quad (46) \end{aligned}$$

We will use the following combinatorial identities

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}, \tag{47}$$

and the one that it leads to, namely,

$$\binom{2m+2}{m+1} = \binom{2m+1}{m+1} + \binom{2m+1}{m} = 2\binom{2m+1}{m+1} = 2\binom{2m+1}{m} \tag{48}$$

in the above equation (45), so that

$$\begin{aligned}
r_{2m+2}(z) &= \eta_1[\sum_{i=0}^{m}[\binom{2m+1}{i} + \binom{2m+1}{i-1}]\eta_1^i\eta_2^{2m+2-i} + \tfrac{1}{2}[\binom{2m+1}{m+1} + \binom{2m+1}{m}]\eta_1^{m+1}\eta_2^{m+1}] \\
&\quad +\eta_2[\sum_{i=m+2}^{2m+2}[\binom{2m+1}{i} + \binom{2m+1}{i-1}]\eta_1^i\eta_2^{2m+2-i} + \tfrac{1}{2}[\binom{2m+1}{m+1} + \binom{2m+1}{m}]\eta_1^{m+1}\eta_2^{m+1}] \\
&= \eta_1[\sum_{i=0}^{m}\binom{2m+1}{i}\eta_1^i\eta_2^{2m+2-i} + \sum_{i=0}^{m-1}\binom{2m+1}{i}\eta_1^{i+1}\eta_2^{2m+1-i} + \binom{2m+1}{m}\eta_1^{m+1}\eta_2^{m+1}] \\
&\quad +\eta_2[\sum_{i=m+2}^{2m+1}\binom{2m+1}{i}\eta_1^i\eta_2^{2m+2-i} + \sum_{i=m+1}^{2m+1}\binom{2m+1}{i}\eta_1^{i+1}\eta_2^{2m+1-i} + \binom{2m+1}{m+1}\eta_1^{m+1}\eta_2^{m+1}] \\
&= \eta_1[\sum_{i=0}^{m}\binom{2m+1}{i}\eta_1^i\eta_2^{2m+2-i} + \sum_{i=0}^{m}\binom{2m+1}{i}\eta_1^{i+1}\eta_2^{2m+1-i}] \\
&\quad +\eta_2[\sum_{i=m+1}^{2m+1}\binom{2m+1}{i}\eta_1^i\eta_2^{2m+2-i} + \sum_{i=m+1}^{2m+1}\binom{2m+1}{i}\eta_1^{i+1}\eta_2^{2m+1-i}] \\
&= \eta_1[\sum_{i=0}^{m}\binom{2m+1}{i}\eta_1^i\eta_2^{2m+1-i}(\eta_2+\eta_1)] + \eta_2[\sum_{i=m+1}^{2m+1}\binom{2m+1}{i}\eta_1^i\eta_2^{2m+1-i}(\eta_2+\eta_1)] \\
&= \eta_1[\sum_{i=0}^{m}\binom{2m+1}{i}\eta_1^i\eta_2^{2m+1-i}] + \eta_2[\sum_{i=m+1}^{2m+1}\binom{2m+1}{i}\eta_1^i\eta_2^{2m+1-i}] = r_{2m+1}(z).
\end{aligned} \tag{49}$$

This completes the proof of part (a).
For the proof of part (b), we proceed as follows: From the equation (43), we express $r_{2m+3}(z)$ as

$$r_{2m+3}(z) = \eta_1\sum_{i=0}^{m+1}\binom{2m+3}{i}\eta_1^i\eta_2^{2m+3-i} + \eta_2\sum_{i=m+2}^{2m+3}\binom{2m+3}{i}\eta_1^i\eta_2^{2m+3-i}. \tag{50}$$

Now using the identity (47), we rewrite the equation (50) as

$$\begin{aligned} r_{2m+3}(z) &= \eta_1 \sum_{i=0}^{m+1}[\tbinom{2m+2}{i} + \tbinom{2m+2}{i-1}]\eta_1^i\eta_2^{2m+3-i} + \eta_2 \sum_{i=m+2}^{2m+3}[\tbinom{2m+2}{i} + \tbinom{2m+2}{i-1}]\eta_1^i\eta_2^{2m+3-i} \\ &= \eta_1[\sum_{i=0}^{m+1}\tbinom{2m+2}{i}\eta_1^i\eta_2^{2m+3-i} + \sum_{i=0}^{m}\tbinom{2m+2}{i}\eta_1^{i+1}\eta_2^{2m+2-i}] \\ &\quad +\eta_2[\sum_{i=m+2}^{2m+2}\tbinom{2m+2}{i}\eta_1^i\eta_2^{2m+3-i} + \sum_{i=m+1}^{2m+2}\tbinom{2m+2}{i}\eta_1^{i+1}\eta_2^{2m+2-i}] \\ &= \eta_1[\sum_{i=0}^{m}\tbinom{2m+2}{i}\eta_1^i\eta_2^{2m+2-i}(\eta_2+\eta_1) + \tbinom{2m+2}{m+1}\eta_1^{m+1}\eta_2^{m+2}] \\ &\quad +\eta_2[\sum_{i=m+2}^{2m+2}\tbinom{2m+2}{i}\eta_1^i\eta_2^{2m+2-i}(\eta_2+\eta_1) + \tbinom{2m+2}{m+1}\eta_1^{m+2}\eta_2^{m+1}], \end{aligned}$$

so that

$$r_{2m+3}(z) = \eta_1[\sum_{i=0}^{m}\tbinom{2m+2}{i}\eta_1^i\eta_2^{2m+2-i}] + \eta_2[\sum_{i=m+2}^{2m+2}\tbinom{2m+2}{i}\eta_1^i\eta_2^{2m+2-i}] + 2\tbinom{2m+2}{m+1}\eta_1^{m+2}\eta_2^{m+2}. \quad (51)$$

By simplifying the last term in (51), we get an upper bound on it as (since $\eta_1+\eta_2=1$ and $\eta_1\eta_2<\frac{1}{4}$ always)

$$2\tbinom{2m+2}{m+1}\eta_1^{m+2}\eta_2^{m+2} = 2\tbinom{2m+2}{m+1}\eta_1^{m+1}\eta_2^{m+1}(\eta_1\eta_2) < 2\tbinom{2m+2}{m+1}\eta_1^{m+1}\eta_2^{m+1}(\tfrac{1}{4}) = \tfrac{1}{2}\tbinom{2m+2}{m+1}\eta_1^{m+1}\eta_2^{m+1}(\eta_1+\eta_2). \quad (52)$$

In view of (51) and (52), we get the inequality

$$r_{2m+3}(z) < \eta_1[\sum_{i=0}^{m}\tbinom{2m+2}{i}\eta_1^i\eta_2^{2m+2-i}] + \eta_2[\sum_{i=m+2}^{2m+2}\tbinom{2m+2}{i}\eta_1^i\eta_2^{2m+2-i}] + \tfrac{1}{2}\tbinom{2m+2}{m+1}\eta_1^{m+1}\eta_2^{m+1}(\eta_1+\eta_2).$$

$$= \eta_1[\sum_{i=0}^{m}\tbinom{2m+2}{i}\eta_1^i\eta_2^{2m+2-i} + \tfrac{1}{2}\tbinom{2m+2}{m+1}\eta_1^{m+1}\eta_2^{m+1}] + \eta_2[\sum_{i=m+2}^{2m+2}\tbinom{2m+2}{i}\eta_1^i\eta_2^{2m+2-i} + \tfrac{1}{2}\tbinom{2m+2}{m+1}\eta_1^{m+1}\eta_2^{m+1}]$$

$$= r_{2m+2}(z) = r_{2m+1}(z),$$

by (49). Thus, the proof of part (b) is complete.
Using parts (a) and (b) in the theorem, we can infer part (c) $r_{2m+2} > r_{2m+4}$, $m = 0, 1, 2, \cdots$. This can be seen by applying in succession parts (a), then (b) and then (a) again as follows: $r_{2m+1} = r_{2m+2} > r_{2m+3} = r_{2m+4}$. This completes the proof of the theorem.

Theorem 3.3 can be obtained from Theorem 3.2 by taking expectations wrt to the random variable Z on both sides of the equality in part (a) and the inequality in part (b). The proofs given in Theorem 3.2 and Theorem 3.3 are new.

Theorem 3.3. *The unconditional asymptotic risk $R_2(k)$ has the following properties:*
(a) $R_2(2m+1) = R_2(2m+2)$ for $m = 0, 1, 2, \cdots$
(b) $R_2(2m+1) > R_2(2m+3)$ for $m = 0, 1, 2, \cdots$
In turn, the above (a) and (b) implies that $R_2(2m+2) > R_2(2m+4)$for $m = 0, 1, 2, \cdots$

	$\mathbf{P}_1$				
Feature↓	$k \rightarrow$				
	1	2	3	4	5
X_1	36.0	35.3	35.1	34.3	33.0
X_2	48.7	52.3	50.0	54.0	46.3
X_3	5.3	5.4	4.7	5.0	4.7
X_4	5.3	4.3	4.7	4.7	4.7
	$\mathbf{P}_2$				
X_1	31.3	34.7	36.0	38.8	38.0
X_2	52.7	51.7	47.8	46.7	47.0
X_3	4.7	6.0	5.3	5.0	4.7
X_4	3.7	4.0	3.3	3.3	3.3

Table 1: Error rate (%) using each feature of the IRIS data ($\mathbf{P}_1$ and $\mathbf{P}_2$)

Devroye [4] provided an improved upper bound on the asymptotic risk of the k-NN rule. This is stated in the following theorem.

Theorem 3.4. *(Devroye, [4]). The unconditional asymptotic risk $R_2(k)$, k odd, has the following bounds:*

(i) $R_2(k) \leq (1 + a_k)R^*$, $a_k = \frac{\alpha\sqrt{k}}{k-3.25} + (1 + \frac{\beta}{\sqrt{k-3}})$, $k \geq 5$ *where* $\alpha = 0.3399$ *and* $\beta = 0.9749$*;*

(ii) $R_2(k) \leq (1 + \sqrt{\frac{2}{k}})R^*$ *for all* $k \geq 1$*; and (iii)* $R_2(k) \leq (1 + \frac{1}{\sqrt{k}})R^*$ *for* $k \geq 3$.

Readers are referred to Devroye [4] for the proof of Theorem 3.4.

4 An Empirical Study (Bagui and Pal, [9])

The empirical behavior of the k-NN rule has been widely studied, and its performance is also well known and well appreciated in the literature. But for the sake of completeness of the paper, we present its performance outcome here on a very well-known data set, viz., Anderson's IRIS data (Johnson and Wichern [5]). This data has become a benchmark for comparing various clustering and classifier designs. The IRIS data include observations from three species of iris - π_1 :Iris setosa, π_2 :Iris versicolor, and π_3 :Iris virginica - on four features X_1 = Sepal length, X_2 = Sepal width, X_3 = Petal length, X_4 = Petal width for each of the three iris species. There are $n_1 = n_2 = n_3 = 50$ sample observations under each species π_i, $i = 1, 2, 3$. The data set has 3 classes (species) of 50 instances in 4-dimensional space each. The entire data set was partitioned into two disjoint sets – training and test sets. We kept 75 observations in the training set, with 25 instances being drawn randomly from each of the three species. Automatically, the test set will have the remaining 75 observations. We make a run on it using the k-NN rule. We call it a forward run. Then we swap the test set with the training set and make a run again using the k-NN rule. The latter one is called the reverse (backward) run. The experiment was done on two different partitions, $\mathbf{P}_1$ and $\mathbf{P}_2$. The error rates displayed in the following tables, the averages of the forward and backward runs, are taken from Bagui and Pal ([9]). For the convenience of researchers in this area, we include the IRIS data in the Appendix.

5 Concluding Remarks

In fact, Fix and Hodges [3] (henceforth F&H [3]) classify Z to the population π_1 if $k_1/n_1 > ck_2/n_2$ where the c depends on the classification costs and prior probabilities. Their classification rule is based

$k\downarrow$	$\mathbf{P}_1$		$\mathbf{P}_2$
1	4.0		5.3
2	2.0		18.7
3	6.7		3.3
4	6.7		4.7
5	4.0		3.3

Table 2: Error rate (%) using all features of the IRIS data (P_1 and P_2)

on nearest neighbor density estimators. They showed that the rule is consistent with the likelihood ratio approach. Silverman and Jones [1] revisited F&H [3] with a commentary on it for historical reasons and contemporary relevance. If $n_1 = n_2$ or $c = \hat{\xi}_2/\hat{\xi}_1 = [n_2/(n_1 + n_2)]/[n_1/(n_1 + n_2)] = n_2/n_1$ then F&H's ([3]) rule reduces to Cover and Hart's [11] k-NN rule. This article revisits Cover and Hart [11] (henceforth C&H [11]). It focuses on the asymptotic risk of both 1-NN and k-NN rules. C&H [11] obtains bounds on the asymptotic risk (R_2) of the 1-NN rule under a 2-population case which is given by $R_2^* \leq R_2 \leq 2R_2^*(1 - R_2^*)$, where R_2^* is the (optimal) Bayes' risk under two populations. Similarly, they also obtained bounds on the asymptotic risk (R_s) of the 1-NN rule under s-population ($s > 2$) case, $R_s^* \leq R_s \leq R_s^*(2 - \frac{s}{s-1}R_s^*)$. These famous bounds on NN classification error rates brought a lot of excitement to the classification and machine learning community at the beginning of the machine learning era. Hence C&H [11] was instrumental in popularizing the k-NN classification rule in the classification and machine learning community. For the k-NN rule in the $s = 2$ population case , we have given in this paper a different proof that as k increases, the asymptotic risk ($R_2(k)$) decreases; that is, a different proof for the inequality $R_2(2m + 1) > R_2(2m + 3)$, $m = 0, 1, 2, \cdots$. Devroye [4] (see Theorem 3.4) derived interesting (improved) upper bounds on the k-NN classification rule risk $R_2(k)$, in the $s = 2$ population case, for various ranges of values of k. C&H [11] results - and their extensions and improvements - are not commonly known in the statistical community as they are in the machine learning community. The main reason for revisiting C&H is [11] to represent and popularize these remarkable results and their extensions in the statistical community.

The k-NN classification methodology has also been studied in the extended context of correctly classifying a set of multiple observations [6], known to have come from one and the same class, among a set of s given classes by Bagui, Mehra and associates under varying scenarios. The interested readers are referred hereby to Bagui et al. ([9],[8]), Bagui and Mehra [7] and Bagui et al. [10]. For classifying a set of multiple observations as above to its correct parent category, Bagui et al. [8] employed a NN classification rule based on creating and utilizing as training samples all possible sub-samples of the original training sample. The asymptotic risk of this method was shown to have bounds similar to those of Cover and Hart [11] for a single observation. Bagui et al. [10] employed an NN-type majority vote classification rule and obtained lower and upper bounds on its risk in terms of the optimal Bayes risk under a separate training sampling scheme. Bagui and Mehra [7] had earlier also studied a Rank Nearest Neighbor (RNN)-type of classification rule for correctly classifying a set of multiple independent observations as above and derived suitable lower and upper bounds on its risk in terms of Bayes risk.

Acknowledgment

Dr. Bagui's research was partially supported by the DUP award, sponsored by Pace Foundation at the University of West Florida.

Appendix

Table 3: Data on Irises (Anderson (1939, [1]), Johnson and Wichern (2007, [10]), UCI Machine Learning Repository, URL: https://archive.ics.uci.edu/ml/datasets/iris)

No.	π_1 : **Iris setosa**				π_2 : **Iris versicolor**				π_3 : **Iris virginica**			
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
1	5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
2	4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
3	4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4	4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5	5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
6	5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
7	4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
8	5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
9	4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
10	4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
11	5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
12	4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
13	4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
14	4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
15	5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
16	5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
17	5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
18	5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
19	5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
20	5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
21	5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
22	5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
23	4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
24	5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
25	4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
26	5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
27	5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
28	5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
29	5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
30	4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
31	4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
32	5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
33	5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
34	5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
35	4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
36	5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
37	5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4

Continued on next page

Table 3 – continued from previous page

No.	π_1 : **Iris setosa**				π_2 : **Iris versicolor**				π_3 : **Iris virginica**			
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
38	4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
39	4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
40	5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
41	5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
42	4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
43	4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
44	5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
45	5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
46	4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
47	5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
48	4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
49	5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
50	5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

References

[1] Silverman BW et al. "An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation; Commentary on Fix and Hodges (1951); Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties, Authors: Evelyn Fix and J.L. Hodges." In: *Jr. International Statistical Review* 57(3) (1989), pp. 233–247.

[2] Anderson E. "The Irises of the Gasé Peninsula." In: *Bulletin of the American Iris Society* 59 (1939), pp. 2–5.

[3] Fix E and Hart JL. *Nonparametric discrimination: consistency properties*. Tech. rep. U.S. Air Force School of Aviation Medicine. Report No. 4. Randolph Field, Texas, 51.

[4] Devroye L. "On the Asymptotic Probability of Error in Nonparametric Discrimination." In: *The Annals of Statistics* 9(6) (1981), pp. 1320–1327.

[5] Johnson RA and Wichern DW. *Applied Multivariate Statistical Analysis*. 6th ed. New Jersey: Prentice Hall, 2007.

[6] Bagui SC. "Nearest Neighbor Classification Rules for Multiple Observations." PhD thesis. University of Alberta, 1989.

[7] Bagui SC and Mehra KL. "Classification of Multiple observations Using Multi-Stage Rank Nearest Neighbor Rule." In: *Journal of Statistical Planning and Inference* 76 (1999), pp. 163–183.

[8] Bagui SC, Mehra KL, and Rao MS. "A Nearest Neighbor Classification Rule for Multiple Observations Based on a sub-sample approach." In: *SankhyaA* 57(2) (1995), pp. 316–332.

[9] Bagui SC and Pal NR. "A multistage generalization of the rank nearest neighbor classification rule." In: *Pattern Recognition Letters* 16 (1995), pp. 601–614.

[10] Bagui SC et al. "Classification with Multiple Independent Measurements Under Separate Sampling Scheme." In: *Statistical Methodology* 3 (2006), pp. 234–252.

[11] Cover TM and Hart PE. "Nearest Neighbor Pattern Classification." In: *IEEE Transaction on Information Theory* 13 (1967), pp. 21–27.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Generalized inverses of some matrices associated with trees

Ravindra B. Bapat[1]

1. *Indian Statistical Institute, 7, S.J.S. Sansanwal Marg, New Delhi-110016, India, rbb@isid.ac.in*

Received: 04 May 2022 / Revised: 10 Jul 2022 / Accepted: 25 Jul 2022

Abstract

Moore-Penrose inverse of various matrices associated with a graph have been of considerable interest. In particular a formula for the Moore-Penrose inverse of the incidence matrix of a tree has been known for more then twenty-five years. We give a different approach to deriving the formula resulting in a more compact expression. Moore-Penrose inverse of the Laplacian and the edge-Laplacian follow as consequences. We then consider the distance matrix between the vertices and the edges of a tree and obtain a formula for its Moore-Penrose inverse.

Keywords: tree, incidence matrix, generalized inverse, Moore-Penrose inverse, vertex-edge incidence matrix
AMS Classification: 05C20, 05C50, 15A09

1 Introduction

The interplay between graphs and matrices is an active area of interest in the past few years and it is interesting from theoretical as well as practical considerations. The area falls under the broad topic of *Algebraic Graph Theory*. We refer to the books ([1],[4]) and [6] for more on this area.

There are various matrices associated with a graph, such as the incidence matrix. adjacency matrix, Laplacian matrix, distance matrix and so on. We may investigate these matrices from a linear algebraic or matrix theoretic point of view and inquire about their determinant, inverse, generalized inverse etc. It turns out that many of these have purely graph theoretic descriptions and lead to attractive properties.

Matrices associated with trees have particularly nice properties. The incidence matrix of a directed tree was considered in [2] where a formula for the Moore-Penrose inverse of the incidence matrix was given. This paper has inspired considerable research and now formulas for a variety of matrices are known. In this paper we give a different formula for the Moore-Penrose inverse which is much more concise than the original formula which was descriptive. We also prove some related results. In the next section we consider the distance matrix defined between vertices and edges of a tree and prove some properties. We give a formula for the Moore-Penrose inverse of the matrix.

For basic notions in graphs and matrices we refer to [1]. The transpose of the matrix A is denoted by A'. If A is a real $m \times n$ matrix then the $n \times m$ matrix G is called a generalized inverse of A if $AGA = A$. Furthermore G is called Moore-Penrose inverse of A, denoted A^+, if it satisfies $GAG = G, (AG)' = AG$ and $(GA)' = GA$. The Moore-Penrose inverse always exists and is unique. For generalized inverses ([3],[5]) are standard references.

2 Incidence matrix of a tree

Let T be a tree with vertex set $V(T) = \{1, \ldots, n\}$ and edge set $\{e_1, \ldots, e_{n-1}\}$. We assume the edges of T to be oriented and let Q be the vertex-edge incidence matrix, which is defined as follows. The rows

© 2022 Author(s). (https://www.thegsa.in/).

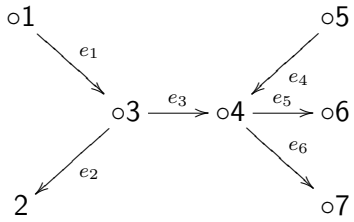
and the columns of Q are indexed by the vertices and the edges of T, respectively. The (e_j, i)-entry of Q is 1 if e_j originates at i, -1 if e_j terminates at i, and zero otherwise. The matrix Q is $n \times (n-1)$ and has rank $n-1$.

Let G be the $(n-1) \times n$ $0-1$ matrix defined as follows. Again the rows and the columns are indexed by $\{e_1, \ldots, e_{n-1}\}$ and $V(T)$ respectively. The (e_j, i)-entry of G is $1(0)$ if e_j is oriented away from (towards) i.

Let P be the $n \times n$ projection matrix defined as $P = I - \frac{1}{n}\mathbf{1}\mathbf{1}'$.

We remark that since a tree is a bipartite graph, the case of the incidence matrix of the unoriented, or undirected, tree can easily be reduced to a directed tree by a similarity transformation. Then the formulas for the unoriented case can easily deduced from those for the oriented case.

Example 1 Consider the tree



Then

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

In the next result we show that G is a left-inverse of Q, hence G is a generalized inverse of Q and that GP is the Moore-Penrose inverse of Q.

Theorem 2.1. *The following assertions are true:*

(i) $GQ = I$. Hence $QGQ = Q$ and G is a generalized inverse of Q.

(ii) The Moore-Penrose inverse of Q is given by $Q^+ = GP$.

Proof. (i) Let e_i and e_j be edges of T, where e_j has end-vertices u and v, and is directed from u to v. The (e_i, e_j)-element of GQ is given by

$$\sum_{w=1}^{n} g(e_i, w) q(w, e_j).$$

Note that $q(u, e_j) = 1, q(v, e_j) = -1$ and $q(w, e_j) = 0$ if $w \neq u, w \neq v$. If $e_i \neq e_j$, then either e_i is oriented away from both u and v, or is oriented towards both u and v. In the first case, $g(e_i, u) = g(e_i, v) = 1$, whereas in the second case $g(e_i, u) = g(e_i, v) = 0$. In either case it follows that the (e_i, e_j)-element of GQ is zero.

If $e_i = e_j$, then $g(e_i, u) = 1$ and $g(e_i, v) = 0$ and in this case, the (e_i, e_j)-element of GQ is 1. Thus $GQ = I$.

(ii) The Moore-Penrose inverse of Q is given by $G + X(I - QG) + (I - GQ)Y$ for some X and Y. Since $GQ = I$, then $Q^+ = G + X(I - QG)$ for some X. We have $\mathbf{1}'Q = 0$ and hence $Q'\mathbf{1} = 0$.

The matrix $I - QG$ has rank 1 and since $\mathbf{1}' = \mathbf{1}'(I - QG)$, $\mathbf{1}'$ forms a basis for its row space. Thus $X(I - QG) = u\mathbf{1}'$ for some u. We now determine u.

We have $Q^+\mathbf{1} = G\mathbf{1} + u\mathbf{1}'\mathbf{1} = G\mathbf{1} + nu$, and since $Q^+\mathbf{1} = 0$ we get $u = -\frac{G\mathbf{1}}{n}$. Thus $Q^+ = G - \frac{1}{n}G\mathbf{1}\mathbf{1}' = G\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) = GP$, and the proof is complete.

A formula for the Moore-Penrose inverse of Q was given in [2]. The formula in Theorem 2.1 is equivalent but more concise. We describe the formula for completeness. The rows and the columns of Q^+ are indexed by the edges and the vertices of T respectively. Let T_h and T_t be the components of $T \setminus e_j$, where T_h is closer to the head of e_j and T_t is closer to the tail. If $i \in T_h$ ($i \in T_t$) then the (e_j, i)-entry of Q^+ is $|V(T_t)|$ ($V(T_h)$. The sign of the entry is positive (negative) if e_j is oriented away from (towards) i. Using this description we see that Q^+ is given by

$$Q^+ = \frac{1}{7}\begin{bmatrix} 6. & -1 & -1 & -1 & -1 & -1 & .-1 \\ 1 & -6 & 1 & 1 & 1 & 1 & 1 \\ 4 & 4 & 4 & -3 & -3 & -3 & -3 \\ -1 & -1 & -1 & -1 & 6 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & -6 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & -6 \end{bmatrix}.$$

It can be checked that this formula is identical to the one given in 2.1,(ii).

Let T be a tree and let Q be the incidence matrix of T. The matrix $L = QQ'$ is the Laplacian matrix of T. It has the vertex degrees on the diagonal and the off-diagonal part is the negative of the adjacency matrix. Since $L = QQ'$, we have $L^+ = (Q')^+Q^+ = PG'GP$, and this is more concise than formulas in the literature ([9],[2])

The matrix $K = Q'Q$ is termed the edge-Laplacian of the tree. Since rank K = rank $Q = n - 1$, K is nonsingular. A formula for K^{-1} is given in [10]. Using Theorem 2.1 we get another formula, $K^{-1} = Q^+(Q')^+ = GPG'$.

In the next result we derive some further properties of G.

Let $\tau_i = 2 - \delta_i, i = 1, \ldots, n$; and let τ be the $n \times 1$ vector with components $\tau_1, \ldots, \tau_n$. Let D be the distance matrix of the tree. The vector τ features in many formulas involving the distance matrix of T. The distance matrix D of T is an $n \times n$ matrix with rows and columns indexed by the vertices and the (i, j)-entry equal to the distance (i.e., the number of edges in the path) between i and j, if $i \neq j$. If $i = j$, then the distance between i and j is set to be zero, The following formulas due to Graham and Pollak [8] and Graham and Lovász[7] are well-known and have had a lot of impact on research in this area.

Let T be a tree with n vertices. let D be the distance matrix of T and L the Laplacian of T. Then $\det D = (-1)^{n-1}(n-1)2^{n-2}$ and

$$D^{-1} = -\frac{1}{2}L + \frac{1}{2(n-1)}\tau\tau'.$$

As seen from the Graham and Pollak formula, the determinant of the distance matrix of a tree depends only on the number of vertices and not on the tree itself. This phenomenon has been shown to be true for many other parameters associated with a tree. In fact, the number of edges in a tree with n vertices is $n - 1$ and this is another example of the same phenomenon. In the next result we identify yet another such parameter.

Theorem 2.2. *The following assertions are true:*

(i) $G\tau = \mathbf{1}$.

(ii) $GD^{-1}G' = -\frac{1}{2}\left(I - \frac{1}{n-1}\mathbf{1}\mathbf{1}'\right)$.

(iii) For the edge e_j *of* T, *let* T_1 *and* T_2 *be the components of* $T \setminus \{e_j\}$. *Let* X *be the principal submatrix of* D^{-1} *indexed by the rows and the columns in* $V(T_1)$. *Then the sum of the elements of* X *is* $-\frac{n-2}{2(n-1)}$. *(Thus the sum neither depends on the tree, nor on the order of* X.*)*

Proof. (i) Let e_j be an edge of T and let T_1 and T_2 be the components of $T \setminus \{e_j\}$. We assume, without loss of generality, that $V(T_1) = \{1, \ldots, m\}$. We also assume that m is an end-vertex of e_j. The entry of $G\tau$ indexed by e_j is given by $\sum_{i=1}^{m} \tau_i = \sum_{i=1}^{m}(2 - \delta_i) = 2m - \sum_{i=1}^{m-1} \delta_i - (\delta_m - 1) - 1$. Note that $\delta_1, \ldots, \delta_{m-1}, \delta_m - 1$ are the degrees of all the vertices in T_1, and hence their sum is $2(m-1)$. It follows that $\sum_{i=1}^{m} \tau_i = 1$. Hence $G\tau = \mathbf{1}$.

(ii) We have

$$\begin{aligned} GD^{-1}G' &= G\left(-\frac{1}{2}L + \frac{1}{2(n-1)}\tau\tau'\right)G' \\ &= -\frac{1}{2}GLG' + \frac{1}{2(n-1)}G\tau\tau'G' \\ &= -\frac{1}{2}\left(I - \frac{1}{n-1}\mathbf{1}\mathbf{1}'\right), \end{aligned}$$

using (i) and the fact that $GLG' = GQQ'G' = I$.

(iii) The sum of the entries of X equals the (e_j, e_j)-element of $GD^{-1}G'$, which by (ii) equals $-\frac{1}{2}$ times a diagonal element of $\left(I - \frac{1}{n-1}\mathbf{1}\mathbf{1}'\right)$. The diagonal element is $1 - \frac{1}{n-1}$ and hence the sum of the elements in X is $-\frac{n-2}{2(n-1)}$.

We obtain a further property of the matrix G. The result may be compared to the known statement that $d(i,j) = \ell^+_{ii} + \ell^+_{jj} - 2\ell^+_{ij}$, see, for example, [[1]], Chapter 9.

Theorem 2.3. *Let* $H = G'G$ *and let* D *be the distance matrix of the tree* T. *Then*

$$d(i,j) = h_{ii} + h_{jj} - 2h_{ij}, \ \ i,j = 1, \ldots, n.$$

Proof. Let $i, j \in V(T), i \neq j$. Let $E^i(E^j)$ be the set of edges oriented away from $i(j)$. Let T' be the graph induced by the ij-path in T and let $F^i(F^j)$ be the set of edges in T', oriented away from $i(j)$. We make some simple observations:

(i) $F^i \cap F^j = \phi$

(ii) $F^i = E(T') \setminus F^j$

(iii) $E^i \setminus F^i = E^j \setminus F^j$

(iv) $h_{ii} = |E^i|, h_{jj} = |E^j|, h_{ij} = |E^i \setminus F^i|$

It follows from (i)-(iv) that

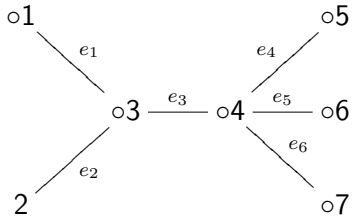
$$\begin{aligned} h_{ii} + h_{jj} - 2h_{ij} &= |E^i| + |E^j| - |E^i \setminus F^i| - |E^j \setminus F^j| \\ &= |F^i| + |E^i \setminus F^i| + |F^j| + |E^j \setminus F^j| - |E^i \setminus F^i| - |E^j \setminus F^j| \\ &= |F^i| + |F^j| \\ &= d(i,j), \end{aligned}$$

and the proof is complete.

3 Vertex-edge distance matrix of a tree

Let T be a tree with vertex set $V(T) = \{1, \ldots, n\}$ and edge set $\{e_1, \ldots, e_{n-1}\}$. If e_j is the edge with end-vertices u and v, then the distance between a vertex i and e_j is defined to be $d'(i, e_j) = \frac{1}{2}(d(i, u)+d(i, v))$. The vertex-edge distance matrix E of T is the $n \times (n-1)$ matrix with rows indexed by $1, \ldots, n$, columns indexed by $e_1, \ldots, e_{n-1}$ and with the (i, e_j)-entry equal to $d'(i, e_j)$.

Consider the tree



Then

$$E = \frac{1}{2}\begin{bmatrix} 1 & 3 & 3 & 5 & 5 & 5 \\ 3 & 1 & 3 & 5 & 5 & 5 \\ 1 & 1 & 1 & 3 & 3 & 3 \\ 3 & 3 & 1 & 1 & 1 & 1 \\ 5 & 5 & 3 & 1 & 3 & 3 \\ 5 & 5 & 3 & 3 & 1 & 3 \\ 5 & 5 & 3 & 3 & 3 & 1 \end{bmatrix}.$$

Lemma 3.1. *Let T be a tree with n vertices and let E be the vertex-edge incidence matrix of T. Then $E = \frac{1}{2}DM$, where D is the distance matrix of T and M is the $0-1$ vertex-edge incidence matrix of T.*

Proof. Let $i \in V(T)$ and let e_j be an edge of T with end-vertices u and v. Then $2d'(i, e_j) = d(i, u) + d(i, v)$. The (i, e_j)-element of DM is given by $\sum_{k=1}^{n} d(i, k)m(k, e_j) = d(i, u) + d(i, v)$ and hence the result is proved.

Lemma 3.2. *Let $\alpha_i = \sum_{j=1}^{n} d(i, j)$ and let $\beta_i = \sum_{j=1}^{n-1} d'(i, e_j)$. Then*

$$\alpha_i = \beta_i + \frac{n-1}{2}.$$

Proof. Since $D\tau = (n-1)\mathbf{1}$, then $D(2\mathbf{1} - \delta) = (n-1)\mathbf{1}$ and hence

$$D\delta = 2D\mathbf{1} - (n-1)\mathbf{1}. \tag{1}$$

Since $E = \frac{1}{2}DM$, we have $E\mathbf{1} = \frac{1}{2}DM\mathbf{1} = \frac{1}{2}D\delta = D\mathbf{1} - \frac{n-1}{2}\mathbf{1}$ by (1). Therefore $D\mathbf{1} = E\mathbf{1} + \frac{n-1}{2}\mathbf{1}$. Comparing the i-th element on both sides we get the result.

Lemma 3.3. *Let T be a tree with vertex set $V(T) = \{1, \ldots, n\}$, and let $V(T) = V_1 \cup V_2$ be the bipartition of $V(T)$. Let x be a vector of order $n \times 1$. Then*

$$\det[M, x] = \pm\left(\sum_{i \in V_1} x_i - \sum_{i \in V_2} x_i\right).$$

Proof. Orient each edge of T from V_1 to V_2. Let Q be the vertex-edge incidence matrix of T. Let S be the $n \times n$ diagonal matrix with its i-th diagonal entry equal to 1(respectively, -1) according as $i \in V_1$(respectively, $i \in V_2$.) Then note that $Q = SM$. Let Q_i(respectively, M_i) be the submatrix of Q(respectively, M) obtained by deleting the i-th row, $i = 1, \ldots, n$. Recall that $(-1)^i \det Q_i$ has the same sign for all i and equals ± 1. It follows that $(-1)^i \det M_i$ equals ± 1 and has the same sign if $i \in V_1$, and the opposite sign if $i \in V_2$. We get the result by expanding $\det[M, x]$ along the last column.

Corollary 3.1. *(i)* $\det[M,\tau] = \pm 2(|V_1| - |V_2|)$.
(ii) $\det[E,\mathbf{1}] = \pm\frac{1}{2}(|V_1| - |V_2|)$.

Proof. (i) We have

$$\begin{aligned}\det[M,\tau] &= \pm\left(\sum_{i\in V_1}\tau_i - \sum_{i\in V_2}\tau_i\right)\\ &= \pm\left(\sum_{i\in V_1}(2-\delta_i) - \sum_{i\in V_2}(2-\delta_i)\right)\\ &= \pm 2(|V_1| - |V_2|), \text{ since } \sum_{i\in V_1}\delta_i = \sum_{i\in V_2}\delta_i,\end{aligned}$$

and (i) is proved.

(ii) Note that $[E,\mathbf{1}] = D[\frac{1}{2}M, \frac{1}{n-1}\tau]$ and hence

$$\det[E,\mathbf{1}] = \frac{1}{2^{n-1}}\frac{1}{n-1}(\det D)(\det[M,\tau]).$$

Since $\det D = (-1)^{n-1}(n-1)2^{n-2}$ and by (i), $\det[M,\tau] = \pm(|V_1| - |V_2|)$ we get $\det[E,\mathbf{1}] = \pm\frac{1}{2}(|V_1| - |V_2|)$ and the proof is complete.

We now provide a formula for the Moore-Penrose inverse of the vertex-edge distance matrix. As before, let S be the signature matrix with its i-th diagonal entry equal to $1(-1)$ if $i \in V_1 (i \in V_2)$.

Theorem 3.2. *Let* $v = D^{-1}S\mathbf{1}$. *Then*

(i) $G = M^+D^{-1}$ *is a generalized inverse of* $2E$.

(ii) The Moore-Penrose inverse of E *is given by*

$$E^+ = \frac{1}{2}M^+D^{-1}\left(I - \frac{vv'}{v'v}\right).$$

Proof. (i) Let $F = 2E$. Since $GF = M^+D^{-1}DM = M^+M = I$, then G is a left-inverse, and hence a generalized inverse of F.

(ii) The Moore-Penrose inverse of F is given by $G + X(I - FG) + (I - GF)Y$ for some X and Y. Since $GF = I$, then $F^+ = G + X(I - FG)$ for some X. We have $v'DM = \mathbf{1}'SD^{-1}DM = \mathbf{1}'SM = 0$ and hence $F'v = 0$. The matrix $I - FG$ has rank 1 and since $v' = v'(I - FG)$, v' forms a basis for its row space. Thus $X(I - FG) = uv'$ for some u. We now determine u.

We have

$$F^+ = G + uv' = M^+D^{-1} + uv'.$$

Post-multiplying the preceding equation by v we get

$$F^+v = M^+D^{-1}v + uv'v. \tag{2}$$

Note that $F'v = 0$, and since F' and F^+ have the same null space, $F^+v = 0$. It follows from (2) that $u = -\frac{M^+D^{-1}v}{v'v}$ and hence

$$F^+ = M^+D^{-1} - \frac{1}{v'v}M^+D^{-1}vv' = M^+D^{-1}\left(I - \frac{vv'}{v'v}\right).$$

Hence

$$E^+ = \frac{1}{2}M^+D^{-1}\left(I - \frac{vv'}{v'v}\right)$$

and the proof is complete.

Acknowledgement

The author acknowledges the support of the Indian National Science Academy, New Delhi, under the INSA Senior Scientist scheme

References

[1] R.B. Bapat. *Graphs and Matrices*. Second. Hindustan Book Agency and Springer, 2014.

[2] R.B. Bapat. "Moore–Penrose inverse of the incidence matrix of a tree." In: *Linear and Multilinear Algebra* 42 (1997), pp. 159–167.

[3] Adi Ben-Israel and Thomas N.E. Greville. *Generalized Inverses. Theory and Applications*. Second. New York: Springer, 2003.

[4] Norman Bigs. *Algebraic graph theory*. Second. Cambridge: Cambridge University Press, 1993.

[5] S.L. Campbell and C.D. Meyer. *Generalized Inverses of Linear Transformation*. Pitman, 1979.

[6] Chris Godsil and Gordon Royle. *Algebraic graph theory*. New York: Springer–Verlag, 2001.

[7] R.L. Graham and L. Lovász. "Distance matrix polynomials of trees." In: *Adv. in Math.* 29(1) (1978), pp. 60–88.

[8] R.L. Graham and H.O. Pollak. "On the addressing problem for loop switching." In: *Bell. System Tech. J.* 50 (1971), pp. 2495–2591.

[9] S.J. Kirkland, M. Neumann, and B.L. Shader. "Distances in weighted trees and group inverse of Laplacian matrices." In: *SIAM J. Matrix Anal. Appl.* 18(4) (1997), pp. 827–841.

[10] R. Merris. "An edge version of the matrix-tree theorem and the Wiener index." In: *Linear and Multilinear Algebra* 25 (1989), pp. 291–296.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Optimal Designs In An Irregular Design Space

Ravindra Khattree [1]

1. *Department of Mathematics and Statistics,Co-Director, Center for Data Science and Big Data Analytics, and Participating Member, Center for Biomedical Research, Oakland University, Rochester, MI, 48309, USA; Email: khattree@oakland.edu; ORCID ID: http://orcid.org/0000-0002-9305-2365.*

Received: 28 Jan 2022 / Revised: 11 Jun 2022 / Accepted: 21 Jun 2022

Abstract

We provide two approaches to obtain optimum designs within an irregular design space. Basic idea is to obtain a design by eliminating design points so that the reduce set of design points still well approximates the design space. This is done by evaluating the eigen-structure of a specific matrix which is a function of corresponding two design matrices. Our criteria for optimization of eigen-structure are based on smallest eigenvalues and generalized antieigenvalues.

Keywords: Antieigenvalues, Eigen-structure, Eigenvalues, Generalized Antieigenvalue, Optimal Designs.
AMS Subject Classifications: 62K05, 62K20

1 Introduction

Professor C. G. Khatri's work on multivariate analysis and linear models had been a great source of inspiration for me especially during my student days and in subsequent years it has had a great influence on my work on multivariate and data analytics. While Professor Khatri and I had been in occasional communications since my days at the Indian Statistical Institute, I had the opportunity to meet him only once at the Joint Statistical meeting of the American Statistical Association at New Orleans in August 1988 just a few months before his untimely passing in 1989. However, I still had an opportunity to personally witness his genius, when I was informed sometimes in 1987 by Professor C. R. Rao that one of my work with Professor R. D. Gupta which was sent to Professor Rao as a first draft, had been further generalized within a span of few days by Professor C. G. Khatri. Our subsequent brief collaboration resulted in [4] which defines a new very general class of multivariate matrix distributions with certain attractive properties. With a feeling of admiration, I dedicate this work to Professor C. G. Khatri to honor his legacy.

Coming to the main theme of this article, consider an experiment in four quantitative factors $x_1 - x_4$, each taking values between 0 to 10. However one requires that for various levels of these factors, we must have, $x_1 + x_2 < 5.5, x_3 + x_4 > 5, x_2 + x_3 > 7$ and $x_1 + x_4 < 5$. Such constraints make the design space highly irregular. If one is looking for a design within this design space, either these constraints should be introduced for optimization of the objective function, which can make the problem highly complex and computationally intensive or alternatively, choose some heuristic approach to come up with a design which can be nearly optimal. Certainly, from a practical point of view, this second approach has much appeal and may be preferred. A way to accomplish this may be to generate a finite number of design points extensively covering this irregular design space and then choose a desired number of points within this region under some meaningful optimality criterion.

The objective of this work is to introduce two such criteria which can be readily adopted to obtain a design when the design space consists of a finite number of potential design points. The basic idea

© 2022 Author(s). (https://www.thegsa.in/).

is to rely on the eigen-structure of a function of design matrices and keep the eigen-structure of the chosen design of size n as close to the eigen-structure of the design corresponding to the finite design space consisting of N design points as described above.

In general the design space $\mathcal{D}$ may be quite large. That is okay and in fact preferred. However, for the example given above and for the sake of manageability of the data and tables, let us say that the levels of each x_i are allowed to be between 0 and 10 with an increment of 2. This results in $6^4 = 1296$ design points out of which only $N = 38$ points satisfy the imposed constraints as described above. These are given in Table 1 and a look at this table clearly shows that the design space is highly irregular. Suppose we wish to choose a design consisting of only $n = 18$ design points. How could we do that?

2 A Minimum Eigenvalue Based Formulation and A Few Matrix Results

Consider the standard linear model set up. Let $\mathcal{D}$ be a finite design space consisting of N design points out of which we want to select a design D of size n for a linear model with p linear parameters. We assume $n > p$. Let $\mathbf{X}$ be the $N \times p$ design matrix corresponding to design $\mathcal{D}$ and $\mathbf{X}_1$ be that of size $n \times p$ obtained by retaining n design points from $\mathcal{D}$ and discarding (without loss of generality) last $r = N - n$ rows of design matrix $\mathbf{X}$. Thus, $\mathbf{X}' = (\mathbf{X}_1{}' : \mathbf{X}_2{}')$. We assume that both $\mathbf{X}$ as well as $\mathbf{X}_1$ are of rank p. Also let $\mathbf{A} = \mathbf{X}'\mathbf{X}$ and $\mathbf{B} = \mathbf{X}_1'\mathbf{X}_1$. Clearly, $p \times p$ matrices $\mathbf{A}$ as well as $\mathbf{B}$ are also of rank p.

Design Point	x_1	x_2	x_3	x_4	Design Point	x_1	x_2	x_3	x_4
1	0	0	8	0	20	0	4	6	4
2	0	0	8	2	21	0	4	8	0
3	0	0	8	4	22	0	4	8	2
4	0	0	10	0	23	0	4	8	4
5	0	0	10	2	24	0	4	10	0
6	0	0	10	4	25	0	4	10	2
7	0	2	6	0	26	0	4	10	4
8	0	2	6	2	27	2	0	8	0
9	0	2	6	4	28	2	0	8	2
10	0	2	8	0	29	2	0	10	0
11	0	2	8	2	30	2	0	10	2
12	0	2	8	4	31	2	2	6	0
13	0	2	10	0	32	2	2	6	2
14	0	2	10	2	33	2	2	8	0
15	0	2	10	4	34	2	2	8	2
16	0	4	4	2	35	2	2	10	0
17	0	4	4	4	36	2	2	10	2
18	0	4	6	0	37	4	0	8	0
19	0	4	6	2	38	4	0	10	0

Table 1: Design Space $\mathcal{D}$ for a Generated Four Factor Design ($N = 38$).

How do you find a good design within the space $\mathcal{D}$? A meaningful approach may be to evaluate how different $\mathbf{B}$ is from $\mathbf{A}$. If the discarded design points were not that important then we expect $\mathbf{B}$ to resemble $\mathbf{A}$ in some meaningful sense – in our discussion in terms of their respective eigen-structures.To assess this, let $\mathbf{U}$ be the upper triangular square root matrix of $\mathbf{B}$ such that $\mathbf{B} = \mathbf{U}'\mathbf{U}$ and define,

$$\mathbf{G} = \mathbf{U}\mathbf{A}^{-1}\mathbf{U}'. \tag{1}$$

If $\mathbf{A}$ and $\mathbf{B}$ are similar in their eigen-structures then $\mathbf{G}$ must be nearly proportional to $\mathbf{I}_p$, the identity matrix of order $p \times p$, and thus eigenvalues of $\mathbf{G}$ must be close to each other. The following theorems (See [6]) show that the eigen-structure of $\mathbf{G}$ is pretty simple if $r \leq p$.

Theorem 2.1. *Let $\delta_1 \geq \delta_2 \geq ... \geq \delta_p$ be the ordered eigenvalues of $\mathbf{G}$. Also let $N - n = r \leq p$. Then $\delta_j = 1$ for $j = 1, 2, .., (p - r)$.*

Proof. Let $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$, where the order of $\mathbf{X}_1$ is $n \times p$ and the $r \times p$ matrix $\mathbf{X}_2$ contains the last r rows of $\mathbf{X}$. Then for $\mathbf{A}$ and $\mathbf{B}$ defined earlier, it is easily seen that,

$$\mathbf{A}^{-1} = (\mathbf{B} + \mathbf{X}_2'\mathbf{X}_2)^{-1} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{X}_2'(\mathbf{I}_r + \mathbf{X}_2\mathbf{B}^{-1}\mathbf{X}_2')^{-1}\mathbf{X}_2\mathbf{B}^{-1}.$$

Thus from (1),

$$\mathbf{G} = \mathbf{U}[\mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{X}_2'(\mathbf{I}_r + \mathbf{X}_2\mathbf{B}^{-1}\mathbf{X}_2')^{-1}\mathbf{X}_2\mathbf{B}^{-1}]\mathbf{U}' = \mathbf{I}_p - \mathbf{Z}$$

where, $\mathbf{Z} = \mathbf{U}'^{-1}\mathbf{X}_2'(\mathbf{I}_r + \mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2')^{-1}\mathbf{X}_2\mathbf{U}^{-1} = \mathbf{U}'^{-1}\mathbf{X}_2'(\mathbf{I}_r + \mathbf{X}_2(\mathbf{U}^{-1}\mathbf{U}'^{-1})\mathbf{X}_2')^{-1}\mathbf{X}_2\mathbf{U}^{-1} = \mathbf{W}'(\mathbf{I}_r + \mathbf{W}\mathbf{W}')^{-1}\mathbf{W}$, and where the $r \times p$ matrix $\mathbf{W} = \mathbf{X}_2\mathbf{U}^{-1}$ is of rank r. Therefore, the above matrix $\mathbf{Z}$ has the last $(p - r)$ eigenvalues as 0. Consequently, $\mathbf{G} = \mathbf{I}_p - \mathbf{Z}$ has the first $(p - r)$ eigenvalues as 1.

The above theorem suggests that as long as $r < p$, the effect of eliminating r design points manifests only on the last r eigenvalues of matrix $\mathbf{G}$. With $r = 1$, that is, when only one design point is eliminated from $\mathcal{D}$, this further simplifies as shown in next theorem.

Theorem 2.2. *Consider the matrix $\mathbf{G}$ when $\mathbf{X}_2 = \mathbf{x}'$ is the N^{th} row of matrix $\mathbf{X}$. Then the smallest eigenvalue of $\mathbf{G}$ is given by*

$$\delta_p = 1 - \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}. \tag{2}$$

Proof. In view of Theorem 1, it follows that only one of the eigenvalues of $\mathbf{G}$ matrix is not equal to 1 and thus $\delta_p = tr(\mathbf{G}) - (p - 1)$. But, $tr(\mathbf{G}) = tr[\mathbf{U}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{U}'] = tr[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{U}'\mathbf{U}]$
$= tr[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{B}] = tr[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X} - \mathbf{x}\mathbf{x}')] = tr(\mathbf{I}_p) - \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$. Thus, $\delta_p = tr(\mathbf{G}) - p + 1 = 1 - \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$. This proves (2).

The effect of eliminating one design point is therefore given by $1 - \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$. The quantity $\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$ can be recognized as the leverage of the i^{th} design point and is proportional to the variance function of $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is the least square estimator of the parameter vector $\boldsymbol{\beta}$.

In the context of eliminating r design points, while still keeping $r \leq p$, a meaningful function of last r eigenvalues of $\mathbf{G}$ may be used. With $\mathbf{X}' = (\mathbf{X}_1' : \mathbf{X}_2')$,, it is easy to check that, $tr(\mathbf{G}) = \delta_{p-r+1} + \delta_{p-r+2} + ... + \delta_p + (p - r) = tr[\mathbf{X}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'] = p - tr[\mathbf{X}_2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_2']$. Thus, a meaningful measure can also be $tr[\mathbf{X}_2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_2']$. A determinant may be an alternative possibility. One may also argue that if a design point was not relevant then its elimination must not make any

difference, in which case the smallest eigenvalue of $\mathbf{G}$ must equal 1. Thus a large value of the smallest eigenvalue of $\mathbf{G}$ or its departure from unity can also be a meaningful criterion for the elimination of a design point.

The following theorem will further help us devise a strategy to eliminate design points systematically.

Theorem 2.3. *With the set up as indicated earlier and with $r \leq p$,*

$$i.tr(\mathbf{G}) = p - \sum_{i=n+1}^{N} \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = p - \sum_{i=n+1}^{N}(1-\delta_{p,i}) = (p-r) + \sum_{i=n+1}^{N}\delta_{p,i}. \tag{3}$$

$$ii.det(\mathbf{G}) = \frac{det(\mathbf{X_1}'\mathbf{X_1})}{det(\mathbf{X}'\mathbf{X})} = \frac{1}{det(\mathbf{X_2}'\mathbf{X_2} - \mathbf{X_2}'(\mathbf{X_1}'\mathbf{X_1})^{-1}\mathbf{X_2})}. \tag{4}$$

Here $\mathbf{x}_i$ is the i^{th} design point, written as a column vector and $\delta_{p,i}$ is the smallest eigenvalue of $\mathbf{G}_i$. Further, the subscripted matrix $\mathbf{G}_i$ is same as the matrix $\mathbf{G}$ defined earlier when only the i^{th} design point $\mathbf{x}_i$ is removed from the design space $\mathcal{D}$.

Proof. It is easy to see that $p = tr[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})] = tr(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = tr(\mathbf{X_1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X_1}') + tr(\mathbf{X_2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X_2}') = tr(\mathbf{G}) + tr(\mathbf{X_2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X_2}') = tr(\mathbf{G}) + \sum_{i=n+1}^{N}\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$. This proves (3).

To prove (4), it suffices to observe that $det(\mathbf{G}) = det(\mathbf{A}^{-1}\mathbf{B}) = \frac{det(\mathrm{B})}{det(\mathrm{A})}$.

Thus $tr(\mathbf{G})$ is basically, apart from a constant $p-r$, the sum of the smallest $r = N-n$ eigenvalues, where with $i = 1, 2, \cdots, r$ each smallest eigenvalue corresponds to the instance i when i^{th} design point is eliminated, each time leaving the design with $N-1$ design points. Clearly when $r = p$, that constant $p-r$ vanishes. This will form a basis for the quick removal of relatively unimportant design points. In (4), we do notice that for a given $\mathcal{D}$, the denominator of the middle expression is constant and thus if one were to maximize the quantity in (4), the resulting design would be equivalent to a $d-$optimal design within $\mathcal{D}$.

Our various approaches are built around (1). Let D be a design and D^c be its complement within $\mathcal{D}$. There are ${}^{N}C_n$ possible choices for D and in principle one would want to choose the "best" design out of all these choices. However, under any chosen criterion for the best, with large values of N, this is neither recommended nor always computationally feasible. Referring to our rather relatively simple example in Table 1, with $N = 38$ and $n = 18$, we have ${}^{N}C_n = {}^{38}C_{18} = 33,578,000,610 \approx 3.3578 \times 10^{10}$ choices! We must thus devise some alternative strategy. In view of above results and criterion stated therein, the following approaches may be suggested.

Approach 1: With $r = 1$, eliminate one design point at a time, each time taking the matrix $\mathbf{X}$ to be the same $N \times p$ matrix and $\mathbf{X}_1$ to be an $(N-1) \times p$ matrix. Compute the respective smallest eigenvalues of matrices $\mathbf{G}_i$. Our design D corresponds to those n design points which have smallest values of the above and the remaining $N-n$ design points constitute D^c.

Approach 2: With $r = 1$, eliminate one design point at a time to get the list of smallest eigenvalues. After doing this, eliminate the design point with the highest value in the above list from the design space, making the design space smaller in size by one design point. Continue this process of elimination of successive design points from the reduced design spaces by using the largest of the smallest eigenvalues as the criterion. After $N-n$ eliminations, what is left is our chosen design.

Approach 3: This approach is a hybrid of the previous two and is based on Theorem 3, making it easier and faster to implement. Note that the Theorem 3 requires $r \leq p$ but $N-n$ will usually be larger than p. Thus the theorem cannot be directly applied. Instead, we eliminate in iterations,

($r_1 =$) p design points, which can be done in one shot by eliminating those p design points which have largest values of smallest eigenvalues. Since $N - n$ may not be divisible by p, for the last iteration, we must eliminate only t design points where $N - n = t(mod(p))$. Of course, in each iteration, the design space loses p (or fewer in the last iteration) design points.

We will illustrate our Approach 3 via two data sets. Results of Approach 1 will be self evident from the first list of smallest eigenvalues at the first step of our calculations of Approach 3. Approach 2 is straight forward but requires $N - n$ calculations of $\mathbf{X}$ each of sizes decreasing by one row and then subsequent calculations of $\mathbf{A}$ and $\mathbf{G}_i$ matrices. We will thus not elaborate on it. Our first data set is for a particular mixture design which we will call as the Parshvanath Mixture design of size 20, while the second illustration will be based on the data given in Table 1.

Design Point	x_1	x_2	x_3	x_4	Design Point	x_1	x_2	x_3	x_4
1	7	12	1	14	11	2	12	15	5
2	2	13	8	11	12	1	11	6	16
3	16	3	10	5	13	12	8	5	9
4	9	6	15	4	14	2	3	15	14
5	7	2	16	9	15	6	10	11	7
6	12	13	3	6	16	16	13	1	4
7	1	8	10	15	17	7	12	10	5
8	14	11	5	4	18	1	14	16	3
9	7	13	10	4	19	2	13	15	4
10	9	3	8	14	20	8	11	9	6

Table 2: Parshvanath Design (N= 20) as Space $\mathcal{D}$ for a Four Factor Mixture Design (All entries must be divided by 34 for mixture sum to be equal to 1).

Example 1 (Parshvanath Mixture Design) This mixture design along with many other larger mixture designs upto size 52, was introduced by Khattree in several publications ([8], [5], [9]) and was derived using the Parshvanath Yantram, a magic square inscribed at the entrance of ancient Parshvanath Jain Temple in Khajuraho, Madhya Pradesh, India. This is also shown to be a constant block-sum partially balanced incomplete block design when the mixture components are instead treated as quantitative factors (See [11], [9]). The mixture design is presented in Table 2, except that each entry must be further divided by 34 so that the sum of the mixture components is 1. However, for our purpose, that divisor is irrelevant. We take the design given in Table 2 as our design space $\mathcal{D}$ with $N = 20$ and suppose we want an optimal design D consisting of $n = 10$ design points under Approach 3. Also, suppose, our model requires only a linear mixture. Clearly, since this corresponds to a mixture experiment, the model will not have an intercept and thus the model is,

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon.$$

Thus $p = 4$. Clearly $N - n = 10$ design points must be removed and therefore only three iterations are needed to respectively eliminate ($p =$) 4, 4 and 2 design points from the design matrices $\mathbf{X}$ of the successively changing sizes of $20 \times 4, 16 \times 4$ and 12×4. Table 3 presents the values of smallest eigenvalues of various $\mathbf{G}_i$ matrices in increasing order after the first iteration. At this stage, four design points with largest magnitudes of the smallest eigenvalue, namely, {15, 20, 17, 9} must be eliminated thereby leaving us with the new reduced design space $\mathcal{D}_{(1)}$ of sixteen design points. It can be verified

that with the newly obtained matrix $\mathbf{G}_1$ using the $\mathcal{D}_{(1)}$ as design space, the design points {13, 2, 6 ,8} are eliminated at the second iteration and finally in the third iteration, another two design points namely, {4, 11} are removed. The final design chosen is thus given by design points {1, 3, 5, 7, 10, 12, 14, 16, 18, 19}. Let $\mathbf{X}_{opt}$ be the 10×4 design matrix of this chosen design and let $\mathbf{A}_{opt} = \mathbf{X}'_{opt}\mathbf{X}_{opt}$. The $det(\mathbf{A}_{opt})$, $tr(\mathbf{A}_{opt})$ and $tr(\mathbf{A}^{-1}_{opt})$ are respectively given as 1.3402×10^{11}, 4076 and 0.008915. In contrast, for the matrix $\mathbf{X}_{removed}$, the 10×4 design matrix consisting of ten eliminated design points, their values are 3.3016×10^{9}, 3404 and 0.037174 respectively. Being of equal size, the two designs can be compared. The first two measures are considerable larger in the former design compared to the latter one and the third is much smaller, thereby also confirming the superiority of the chosen design with respect to other measures.

Removed Design Pt	x_1	x_2	x_3	x_4	Smallest EValue	Removed Design Pt	x_1	x_2	x_3	x_4	Smallest EValue
3	16	3	10	5	0.66357	4	9	6	15	4	0.80903
16	16	13	1	4	0.70015	11	2	12	15	5	0.82743
12	1	11	6	16	0.71215	6	12	13	3	6	0.82838
14	2	3	15	14	0.71696	8	14	11	5	4	0.82865
1	7	12	1	14	0.71793	2	2	13	8	11	0.84261
18	1	14	16	3	0.71824	13	12	8	5	9	0.87460
5	7	2	16	9	0.72922	9	7	13	10	4	0.88898
10	9	3	8	14	0.75575	17	7	12	10	5	0.91718
7	1	8	10	15	0.79504	20	8	11	9	6	0.93673
19	2	13	15	4	0.79548	15	6	10	11	7	0.94193

Table 3: Parshvanath Design (N= 20) Smallest Eigenvalue Calculations (Iteration 1).

Note that under Approach 1, the criterion would eliminate the last ten design points of Table 3 and retain the first ten as our chosen design. The final result turns out to be the same even though values of smallest eigenvalues at various iterations would differ.

Example 2 (A Generated Design with Four Factors) The design space given in Table 1 and consisting of 38 design points is quite irregular. Suppose we wish to construct a suitable design consisting of only 18 design points chosen from this design space by using the above criterion and for the model with four factors and all first-order interactions,

$$y = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i=1}^{4} \sum_{j>i}^{4} \beta_{ij} x_i x_j + \epsilon.$$

There are eleven linear parameters in the model and thus $p = 11$. It is appropriate to standardize the variables $x_1 - x_4$ to have zero mean and unit standard deviation and we define the six interaction terms as corresponding products after this standardization. As earlier, the method will require removal of $r_1 = p = 11$ design points in first iterations and the remaining $r_2 = 9 < p$ design points in the second iteration. During the first iteration, the design points {11, 22, 14, 10, 12, 19, 2, 8, 27, 29, 25} have the highest values of the corresponding smallest eigenvalue and hence are eliminated. This leaves us with a reduced design space of $N_1 = 27$ design points. Another nine design points must be discarded from this reduced design space to arrive at the final design. After Iteration 2, these are found to be {20, 34, 33, 15, 23, 5, 13, 21, 16}. The finally chosen design consisting of design points {1, 3, 4, 6, 7, 9, 17, 18, 24, 26, 28, 30, 31, 32, 35, 36, 37, 38} is presented in Table 4. It can be shown that the sum of all smallest eigenvalues must add to $(n - p)$ which in our case for the resulting design is equal to 7.

Removed Design Pt	x_1	x_2	x_3	x_4	Smallest EValue	Removed Design Pt	x_1	x_2	x_3	x_4	Smallest EValue
26	0	4	10	4	0.06701	35	2	2	10	0	0.45077
24	0	4	10	0	0.13443	36	2	2	10	2	0.46571
17	0	4	4	4	0.16345	32	2	2	6	2	0.46894
4	0	0	10	0	0.29534	3	0	0	8	4	0.48267
6	0	0	10	4	0.29944	1	0	0	8	0	0.48580
18	0	4	6	0	0.31900	7	0	2	6	0	0.49796
37	4	0	8	0	0.36755	30	2	0	10	2	0.55784
38	4	0	10	0	0.36898	9	0	2	6	4	0.55932
31	2	2	6	0	0.44431	28	2	0	8	2	0.57148

Table 4: Chosen Design ($n = 18$) from Generated Design Space ($N = 38$) Smallest Eigenvalue Calculations.

3 A Criterion Based on Antieigenvalues

Let $\mathbf{C}_{p\times p}$ be a real symmetric positive definite matrix and consider the quadratic surface, $\mathbf{u}'\mathbf{C}^{-1}\mathbf{u} = \delta$ where δ is a known constant, in a $p-$ dimensional space. Since $\mathbf{C}$ is positive definite, this represents an ellipsoid and with an appropriate orthogonal rotation $\mathbf{v} = \mathbf{P}'\mathbf{u}$ where $\mathbf{C} = \mathbf{P}\Lambda\mathbf{P}'$ is the spectral decomposition of $\mathbf{A}$, the surface can be represented as,

$$\mathbf{v}'\Lambda^{-1}\mathbf{v} = \delta \text{ with } \Lambda = diagonal(\lambda_1, \lambda_2, ..., \lambda_p)$$

or

$$\frac{v_1^2}{\lambda_1} + \frac{v_2^2}{\lambda_2} + ... + \frac{v_p^2}{\lambda_p} = \delta \text{ where } \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0.$$

The eccentricities of certain two dimensional elliptical cross-sections of this ellipsoid can be quantified in decreasing order as $\sqrt{\frac{\lambda_1}{\lambda_p}} \geq \sqrt{\frac{\lambda_2}{\lambda_{p-1}}} \geq \sqrt{\frac{\lambda_3}{\lambda_{p-2}}} \geq \cdots$. The quantity $e_1 = \sqrt{\frac{\lambda_1}{\lambda_p}}$ is the eccentricity measured respectively via the two most elongated and most compressed directions and hence measures the extreme eccentricity. The next quantity $e_2 = \sqrt{\frac{\lambda_2}{\lambda_{p-1}}}$ represents the comparison of the next two most elongated and most compressed directions and similar comparisons continue for $[p/2]$ pairs where $[p/2]$ is the integer part of $p/2$. Clearly, whenever λ_i is considerably larger than $\lambda_{p-i+1}, i = 1, 2, \cdots, [p/2]$, e_i will also be large, indicating a particular cross section of the ellipsoid highly elongated. A one-to-one monotonically decreasing function of $e_i = \sqrt{\frac{\lambda_i}{\lambda_{p-i+1}}}$ is the i^{th} *antieigenvalue* of the matrix $\mathbf{C}$ namely,

$$\eta_i = \frac{2\sqrt{\lambda_i\lambda_{p-i+1}}}{\lambda_i + \lambda_{p-i+1}} = \frac{2}{e_i + e_i^{-1}}, i = 1, 2, \cdots, [p/2]. \tag{5}$$

The quantities $0 < \eta_1 \leq \eta_2 \leq ... \leq 1$ in (5) are ordered naturally by their magnitudes and each η_i defined above is a one-to-one monotonically decreasing function of corresponding e_i. Being a monotonic function of $e_i, i = 1, 2, \cdots, [p/2]$, these also measure the eccentricities. Accordingly, $\eta_i, i = 1, \cdots, [p/2]$, provide a way to measure the departure from sphericity of the corresponding matrix $\mathbf{C}$. Also, $\mathbf{C}$ and $\mathbf{C}^{-1}$ share the same set of antieigenvalues, a fact consistent with the parity of sphericity of the two matrices.

A single index of sphericity combining all antieigenvalues can be defined as the generalized antieigenvalue (See [10], [7])

$$\Delta = \prod_{i=1}^{[p/2]} \frac{2\sqrt{\lambda_i \lambda_{p-i+1}}}{\lambda_i + \lambda_{p-i+1}} = \prod_{i=1}^{[p/2]} \eta_i, \tag{6}$$

which is a function of all antieigenvalues and can be interpreted as an overall measure of eccentricity. One may alternatively use the $[p/2]^{th}$ root of Δ in (6) which would then be the geometric mean of all antieigenvalues. Venables in [17] uses the generalized antieigenvalue, although without using this specific nomenclature, as a test statistic to test the sphericity of a covariance matrix. Also see the classic textbook by Srivastava and Khatri [16] for the related discussion and [1], [12] for the extensive work where the generalized antieigenvalue is shown to be an appropriate measure to quantify the inefficiency of a least square estimator under misspecification of the model assumptions.

In our context, the matrix of interest is $\mathbf{G}$ defined in (1). Intuitively, if D is a good choice as a representative of the entire design space $\mathcal{D}$ then $\mathbf{X}_1'\mathbf{X}_1$ must be close enough, apart from a scaling constant, to $\mathbf{X}'\mathbf{X}$ in which case, we expect $\mathbf{G}$ to be nearly spherical. That it is so, can be assessed via individual antieigenvalues or the generalized antieigenvalue defined above. Considering the latter as the criterion, the problem then is to find the design D so that the generalized antieigenvalue of the corresponding $\mathbf{G}$ matrix is maximum among all choices of possible designs of same size.

Clearly, there are ${}^{N}C_{n}$ choices for possible designs and even for the moderate values of N and n, as in our examples, the number of choices is excessively large. We must thus devise some algorithm which even if not fully optimal, can provide some approximate solution. Approaches by several authors (See [3], [13], [14], [15], [2]) involve a variety of exchange algorithms to do so. Our approach uses the same idea but with the objective of optimizing the generalized antieigenvalue.

The basic exchange algorithm is given here. For a given initial partition $\{D, D^c\}$ of $\mathcal{D}$, initially let the $\mathbf{G}_0$ be as defined in (1). We exchange one design point of D with another design point of D^c and do so for all $n \times (N-n)$ pairs, each time computing the corresponding generalized antieigenvalue of corresponding $\mathbf{G}$ matrix, say $\mathbf{G}_i$, at i^{th} iteration. If the largest value in the $n \times (N-n)$ matrix of generalized antieigenvalues exceeds the largest obtained at the previous iteration, suggested exchange is initiated and the algorithm proceeds to the next iteration. Otherwise the solution has been obtained as the design D of immediate previous stage with largest value of generalized antieigenvalue. In other words, the algorithm continues until we cannot improve the largest value of the generalized antieigenvalue. With finite choices, it is ensured that the algorithm must converge but a global optimum cannot be assured. With practical limitations, it would be a good compromise to obtain a nearly optimal solution, especially when our choice of initial design at the beginning was carefully chosen. In our illustrative examples which follow, we have chosen the initial choice of designs as those finally obtained in the previous section.

Example 1 (Continued): We continue with the 20-point design space of Parshvanath and look for a design D which is nearly optimal under our generalized antieigenvalue criterion. Whether to choose $\mathbf{G}$ as defined in (1) or its inverse with square root matrix of $\mathbf{X}'\mathbf{X}$ appropriately defined, is immaterial since both choices lead to the same set of antieigenvalues. The exchange procedure as described above will be applied to obtain a design with 10 design points. As the starting point, we take the initial design as the one obtained in the previous section namely, $D = \{1, 3, 5, 7, 10, 12, 14, 16, 18, 19\}$. Accordingly, $D^c = \{2, 4, 6, 8, 9, 11, 13, 15, 17, 20\}$. The 10×10 matrix of generalized antieigenvalues (not shown) when the i^{th} design point of D is exchanged with the j^{th} design point of D^c is obtained and the corresponding choice (i_1, j_1) corresponding to the maximum value is identified. Specifically, design point 9 belonging to D is exchanged with the design point 10 belonging to D^c thereby updating the sets D and D^c. This and subsequent iterations are described in Table 5 with

Iteration	Design	Design Points	Max Gen AntiEvalue
Initial	D	1 3 5 7 **10** ↓ 12 14 16 18 19	0.94697
	D^c	2 4 6 8 **9** ↑ 11 13 15 17 20	
1	D	1 3 5 **7** ↓ 9 12 14 16 18 19	0.97713
	D^c	2 4 6 8 10 11 13 15 17 **20** ↑	
2	D	1 3 5 20 **9** ↓ 12 14 16 18 19	**0.98954**
	D^c	2 4 6 8 10 11 13 15 **17** ↑ 7	
3	D	1 3 5 20 17 12 14 16 18 19	0.98945
	D^c	2 4 6 8 10 11 13 15 9 7	

Table 5: Design Derivation ($n = 10$) from Parshvanath Design Space ($N = 20$) Antieigenvalue Calculations [Exchanged design points between D and D^c for the next iteration are highlighted in **bold**] and by ↓, ↑ s.

D					D^c				
Design Point	x_1	x_2	x_3	x_4	Design Point	x_1	x_2	x_3	x_4
1	7	12	1	14	2	2	13	8	11
3	16	3	10	5	4	9	6	15	4
5	7	2	16	9	6	12	13	3	6
9	7	13	10	4	7	1	8	10	15
12	1	11	6	16	8	14	11	5	4
14	2	3	15	14	10	9	3	8	14
16	16	13	1	4	11	2	12	15	5
18	1	14	16	3	13	12	8	5	9
19	2	13	15	4	15	6	10	11	7
20	8	11	9	6	17	7	12	10	5

Table 6: Designs $D(n = 10)$ and $D^c(N - n = 10)$ for the Parshvanath Design Space.

choices identified in bold along with the direction of exchange indicated by ↓ and ↑ arrows. Upon third iteration we observe that maximum of the generalized antieigenvalue does not further improve. We thus accept the choice of iteration 2 as the finally chosen design. See Table 6. It is easy to see that the finally obtained design under generalized antieigenvalue criterion is different by two design points from that obtained under minimum eigenvalue criterion (namely, the design points {9, 20} vs. {7, 10}). The very high value of maximum antieigenvalue, close to 1 is a good assurance about the quality of the chosen design.

Example 2 (Continued): Here we are looking for a design with 18 design points. For the model considered, we have $p = 11$ and thus computations are more intensive. Nonetheless, following the same approach with details shown in Table 7, we obtain the optimal design presented in Table 8, which corresponds to iteration 3. One can also see in the illustration that a design point once leaving D can again come back (Design point 28, in our case). The final design is different from the initial design again only in two design points namely, {23, 34} vs. {32, 36}. As earlier, maximum of generalized antieigenvalue exceeds 0.95, which is deemed quite satisfactory.

Iteration	Design	Design Points	Max Gen AntiEvalue
Initial	D	1 3 4 6 7 9 17 18 24 26 **28** ↓ 30 31 32 35 36 37 38	0.94697
	D^c	2 5 8 10 11 12 13 14 15 16 19 20 21 22 **23** ↑ 25 27 29 33 34	
1	D	1 3 4 6 7 9 17 18 24 26 23 30 31 32 35 **36** ↓ 37 38	0.95645
	D^c	2 5 8 10 11 12 13 14 15 16 19 20 21 22 28 25 27 29 33 **34** ↑	
2	D	1 3 4 6 7 9 17 18 24 26 23 30 31 **32** ↓ 35 34 37 38	0.95713
	D^c	2 5 8 10 11 12 13 14 15 16 19 20 21 22 **28** ↑ 25 27 29 33 36	
3	D	1 3 4 6 7 9 17 18 24 26 23 30 31 28 35 34 37 **38** ↓	**0.95826**
	D^c	2 5 8 10 11 12 13 14 15 16 19 20 21 22 32 25 27 29 33 **36** ↑	
4	D	1 3 4 6 7 9 17 18 24 26 23 30 31 28 35 34 37 36	0.95826
	D^c	2 5 8 10 11 12 13 14 15 16 19 20 21 22 32 25 27 29 33 38	

Table 7: Design Derivation ($n = 18$) from the Generated Design Space ($N = 38$) Antieigenvalue Calculations [Exchanged design points between D and D^c for the next iteration are highlighted in **bold**] and by $\downarrow, \uparrow s$.

4 Concluding Remarks

In this article we have introduced two new criteria to obtain the optimal designs based on the minimum eigenvalue and generalized antieigenvalue of a certain matrix. In practice both criteria can be easily implemented. It must be mentioned as an alternative interpretation that when we are looking at the eigen-structure of matrix $\mathbf{G}$, we are essentially considering the eigen-structure of $\mathbf{X}_1'\mathbf{X}_1$ *with respect to* $\mathbf{X}'\mathbf{X}$ in the sense of determinantal equation $det(\mathbf{X}_1'\mathbf{X}_1 - \lambda\mathbf{X}'\mathbf{X}) = 0$. Further, since both criteria rely on eigen-structures, the invariance with respect to orthogonal transformation of design points follows as long as the irregular design space is defined by linear constraints. However, the two criteria introduced here themselves are *not* equivalent since they respectively depend on eigenvalues and antieigenvalues and as shown in (5), each antieigenvalue is a function of a pair of eigenvalues.

One of the referees asked if instead of Approach 1 described above, is it possible to start with a random design with fixed number of design points and then interchange the points, using the Fedorov-type exchange algorithm satisfying the criteria introduced here? Answer to this query is a definite "Yes". However, exchange algorithms may require excessive computational resources especially when N and/or n are large. That is especially so if we start with an initial random design. Our Approach 1 (and also Approach 2) circumvents this problem. Approach 3 essentially relies on an exchange- type algorithm. However, as we have pointed out earlier, one can minimize the computational efforts by avoiding to initially start with a random design and instead choose the initial design carefully, such as that generated by using our Approach 1. Our examples show that with some wise-choices of the initial design, number of iterations to obtain the optimal designs can be substantially reduced.

The two data sets used here are only for the illustration purposes and thus purposely chosen to have the design spaces of manageable sizes so as to be included as tables in their entirety. However, approaches are valid and in fact, are more suitable when the design spaces are highly irregular for theoretical developments of the algorithms and computations and are still very vast consisting of thousands of potential design points. In experiments involving discrete choices and many other spatial experiments such irregular design spaces are quite common. Thus, it is of interest to develop efficient algorithms for the approaches introduced here.

The matrix $\mathbf{G}$ defined in (1) plays the pivotal role in the approach described in this work. However, there can be several variations to the criteria as well as to the algorithms. For example, the exchange algorithm can as well be applied to minimum eigenvalue criterion. Likewise, in Section 3, instead of generalized antieigenvalue, one could just use the smallest antieigenvalue as a criterion and maximize

D					D^c				
Design Point	x_1	x_2	x_3	x_4	Design Point	x_1	x_2	x_3	x_4
1	0	0	8	0	2	0	0	8	2
3	0	0	8	4	5	0	0	10	2
4	0	0	10	0	8	0	2	6	2
6	0	0	10	4	10	0	2	8	0
7	0	2	6	0	11	0	2	8	2
9	0	2	6	4	12	0	2	8	4
17	0	4	4	4	13	0	2	10	0
18	0	4	6	0	14	0	2	10	2
23	0	4	8	4	15	0	2	10	4
24	0	4	10	0	16	0	4	4	2
26	0	4	10	4	19	0	4	6	2
28	2	0	8	2	20	0	4	6	4
30	2	0	10	2	21	0	4	8	0
31	2	2	6	0	22	0	4	8	2
34	2	2	8	2	25	0	4	10	2
35	2	2	10	0	27	2	0	8	0
37	4	0	8	0	29	2	0	10	0
38	4	0	10	0	32	2	2	6	2
					33	2	2	8	0
					36	2	2	10	2

Table 8: Designs $D(n = 18)$ and $D^c(N - n = 20)$ for the Generated Four Factor Design Space.

it over all possible designs to obtain an optimal design. There is no reason, at least at this point, to believe that one criterion or algorithm may be superior to another and only an extensive empirical study could shed some light on such issues.

Acknowledgment

I wish to thank two referees and the editors for their helpful suggestions resulting in the improvements in the original draft.

References

[1] Peter Bloomfield and Geoffrey S Watson. "The inefficiency of least squares." In: *Biometrika* 62.1 (1975), pp. 121–128.

[2] Evangelos Evangelou and Zhengyuan Zhu. "Optimal predictive design augmentation for spatial generalised linear mixed models." In: *Journal of Statistical Planning and Inference* 142.12 (2012), pp. 3242–3253.

[3] V V Federov. *Theory of Optimal Experiments (Translation)*. Academic Press: New York, 1972.

[4] C G Khatri, Ravindra Khattree, and Rameshwar D Gupta. "On a class of orthogonal invariant and residual independent matrix distributions." In: *Sankhyā: The Indian Journal of Statistics, Series B* 53.1 (1991), pp. 1–10.

[5] Ravindra Khattree. "A class of designs for mixture experiments implied by the numerical configurations in Hindu/Jain Yantrams." In: *Journal of the Indian Statistical Association* 54.1-2 (2016), pp. 127–155.

[6] Ravindra Khattree. "A Note on Effects on the Eigenstructure of a Data Matrix when Deleting a Subset of Observations." In: *Journal of the Indian Society of Agricultural Statistics* 73.1 (2019a), pp. 11–17.

[7] Ravindra Khattree. "Antieigenvalues and antieigenvectors in statistics." In: *Journal of Statistical Planning and Inference* 114.1-2 (2003), pp. 131–144.

[8] Ravindra Khattree. "Mixture experiments in the interior: Yantram designs." In: *Journal of Statistical Theory and Practice* 9.4 (2015), pp. 797–822.

[9] Ravindra Khattree. "On construction of constant block-sum partially balanced incomplete block designs." In: *Communications in Statistics-Theory and Methods* 49.11 (2020), pp. 2585–2606.

[10] Ravindra Khattree. "On generalized antieigenvalue and antieigenmatrix of order r." In: *American Journal of Mathematical and Management Sciences* 22.1-2 (2002), pp. 89–98.

[11] Ravindra Khattree. "The Parshvanath yantram yields a constant-sum partially balanced incomplete block design." In: *Communications in Statistics-Theory and Methods* 48.4 (2019b), pp. 841–849.

[12] Martin Knott. "On the minimum efficiency of least squares." In: *Biometrika* 62.1 (1975), pp. 129–132.

[13] Toby J Mitchell. "An algorithm for the construction of "D-optimal" experimental designs." In: *Technometrics* 42.1 (2000), pp. 48–54.

[14] Toby J Mitchell. "Computer construction of "D-optimal" first-order designs." In: *Technometrics* 16.2 (1974), pp. 211–220.

[15] JA Royle. "Exchange algorithms for constructing large spatial designs." In: *Journal of Statistical Planning and Inference* 100.2 (2002), pp. 121–134.

[16] Muni Shanker Srivastava and C G Khatri. *An Introduction to Multivariate Statistics*. North-Holland, New York., 1979.

[17] W Venables. "Some implications of the union-intersection principle for tests of sphericity." In: *Journal of Multivariate Analysis* 6.1 (1976), pp. 185–190.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Properties of the Matrix $V + XTX'$ in Linear Statistical Models

Stephen J. Haslett[1], Augustyn Markiewicz[2] and Simo Puntanen[3]

1. *School of Mathematical and Computational Sciences & Research Centre for Hauora and Health, Massey University, New Zealand, email: s.j.haslett@massey.ac.nz; Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Australia; NIASRA, Faculty of Engineering and Information Sciences, University of Wollongong, Australia*
2. *Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Wojska Polskiego 28, PL-60637 Poznań, Poland, email: amark@up.poznan.pl*
3. *Faculty of Information Technology and Communication Sciences, FI-33014 Tampere University, Finland, email: simo.puntanen@tuni.fi*

Received: 27 Nov 2021 / Revised: 02 Feb 2022 / Accepted: 04 Apr 2022

Abstract

It is well known, due originally to C.R. Rao in early 1970s, that the best linear unbiased estimator, BLUE, of $\mathbf{X}\boldsymbol{\beta}$ in the linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ can be expressed in the form $\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}\mathbf{y}$, where $\mathbf{W}$ is a specific matrix of the form $\mathbf{W} = \mathbf{V}+\mathbf{X}\mathbf{T}\mathbf{X}'$ with $\mathbf{T}$ satisfying the column space condition $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$. We denote this class of matrices as $\mathcal{W}$. Choice of $\mathbf{T}$ as an identity matrix gives an obvious member $\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{X}' \in \mathcal{W}$. The matrices belonging to the class $\mathcal{W}$ have several interesting mathematical properties. In particular, the use of matrix $\mathbf{W} \in \mathcal{W}$ appears to be surprisingly handy and helpful tool when dealing with the linear statistical models. Our aim is to review and collect together some essential features of $\mathbf{W}$ and its use in linear statistical models. While doing this, we go through some related basic properties of the best linear unbiased estimation.

Keywords: Best linear unbiased estimator, BLUE, Column space, Generalized inverse, Löwner ordering,Linear sufficiency, Partitioned linear model.
MSC: 62J05, 62J10

1 Introduction: Basic Tools

We begin this article by introducing the notation and the basic mathematical tools that we are going to use; these matters will occupy the first two sections. In a nutshell, we slowly approach the problems what we meet if we want to use a particular kind of estimator in the linear model to catch the best linear unbiased estimator, BLUE, for the unknown parametric function. In our considerations the matrix of the type $\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}'$, where $\mathbf{X}'$ is the transpose of $\mathbf{X}$, will have the main role. But before the main goal, we need some basic tools and definitions.

In this article we consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ or shortly

$$\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}\,.$$

Here $\mathbf{y}$ is an n-dimensional observable response variable, and $\boldsymbol{\varepsilon}$ is an unobservable random error with a known covariance matrix $\mathrm{cov}(\boldsymbol{\varepsilon}) = \mathrm{cov}(\mathbf{y}) = \mathbf{V}$ (can be singular) and expectation $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \in \mathbb{R}^n$. The matrix $\mathbf{X}$ is a known $n \times p$ matrix, i.e., $\mathbf{X} \in \mathbb{R}^{n\times p}$, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of fixed (but unknown) parameters. We will denote $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ so that $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. Sometimes the covariance matrix is of the type $\sigma^2\mathbf{V}$, where σ^2 is an unknown positive constant.

© 2022 Author(s). (https://www.thegsa.in/).

By the partitioned linear model we mean that $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, or shortly denoted

$$\mathscr{M}_{12} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\} = \{\mathbf{y},\ \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\ \mathbf{V}\}\,.$$

In addition to the *full* model $\mathscr{M}_{12}$, we will consider the *small* models $\mathscr{M}_i = \{\mathbf{y},\ \mathbf{X}_i\boldsymbol{\beta}_i,\ \mathbf{V}\}$, $i = 1, 2$, and the *reduced* model

$$\mathscr{M}_{12\cdot 2} = \{\mathbf{M}_2\mathbf{y},\ \mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1,\ \mathbf{M}_2\mathbf{V}\mathbf{M}_2\}\,,$$

where $\mathbf{M}_2 = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_2}$, with $\mathbf{P}_{\mathbf{X}_2}$ being the orthogonal projector onto the column space of $\mathbf{X}_2$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix. Premultiplying the model $\mathscr{M}$ by an $f \times n$ matrix $\mathbf{F}$ yields the *transformed* model

$$\mathbf{F}\mathbf{y} = \mathbf{F}\mathbf{X}\boldsymbol{\beta} + \mathbf{F}\boldsymbol{\varepsilon}\,, \quad \text{or shortly } \mathscr{T} = \{\mathbf{F}\mathbf{y}, \mathbf{F}\mathbf{X}\boldsymbol{\beta}, \mathbf{F}\mathbf{V}\mathbf{F}'\}\,.$$

The reduced model $\mathscr{M}_{12\cdot 2}$ is of course one example of the transformed models. We will also shortly consider the linear model with *new* (unobserved, to be predicted) observations. This means that in addition to $\mathscr{M}$, we are dealing with a $q \times 1$ unobservable random vector $\mathbf{y}_*$ containing new observations. These new observations are assumed to come from $\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_*$, where $\mathbf{X}_*$ is a known $q \times p$ matrix, and $\boldsymbol{\varepsilon}_*$ is a q-dimensional random error vector whose (cross-)covariance matrix with $\mathbf{y}$ is known.

As for the notation: $\mathrm{r}(\mathbf{A})$, $\mathbf{A}^-$, $\mathbf{A}^+$, $\mathscr{C}(\mathbf{A})$, $\mathscr{N}(\mathbf{A})$, and $\mathscr{C}(\mathbf{A})^\perp$, denote, respectively, the rank, a generalized inverse, the (unique) Moore–Penrose inverse, the column space, the null space, and the orthogonal complement of the column space of the matrix $\mathbf{A}$. By $\mathbf{A}^\perp$ we denote any matrix satisfying $\mathscr{C}(\mathbf{A}^\perp) = \mathscr{C}(\mathbf{A})^\perp$. Furthermore, we will write $\mathbf{P}_\mathbf{A} = \mathbf{P}_{\mathscr{C}(\mathbf{A})} = \mathbf{A}\mathbf{A}^+ = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ to denote the orthogonal projector onto $\mathscr{C}(\mathbf{A})$. The orthogonal projector onto $\mathscr{C}(\mathbf{A})^\perp$ is denoted as $\mathbf{Q}_\mathbf{A} = \mathbf{I}_a - \mathbf{P}_\mathbf{A}$. We will shorten our notation as

$$\mathbf{H} = \mathbf{P}_\mathbf{X}\,, \quad \mathbf{M} = \mathbf{I}_n - \mathbf{P}_\mathbf{X}\,, \quad \mathbf{M}_i = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_i}\,,\ i = 1, 2\,.$$

One obvious choice for $\mathbf{X}^\perp$ is $\mathbf{M}$.

Next we recall some basic concepts when dealing with the best linear unbiased estimation. In particular we explore the problems when figuring out for which choice of matrix $\mathbf{N} \in \mathbb{R}^{n\times n}$ an estimator of the type

$$\mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{N}\mathbf{X})^-\mathbf{X}'\mathbf{N}\mathbf{y}$$

provides a representation for the best linear estimator, BLUE, of $\mathbf{X}\boldsymbol{\beta}$. Notice that the above representation can be interpreted to arise from solving $\mathbf{b}$ from

$$\mathbf{X}'\mathbf{N}\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{N}\mathbf{y}, \tag{1}$$

supposing that (1) is solvable for $\mathbf{b}$; this happens if and only if $\mathbf{X}'\mathbf{N}\mathbf{y} \in \mathscr{C}(\mathbf{X}'\mathbf{N}\mathbf{X})$. We will go through various particular choices of $\mathbf{N}$:

- the first and simplest case is $\mathbf{N} = \mathbf{I}_n$,
- then we take $\mathbf{N} = \mathbf{V}^{-1}$, and $\mathbf{N} = \mathbf{V}^+$, or $\mathbf{N} = \mathbf{V}^-$, i.e., $\mathbf{N} \in \{\mathbf{V}^-\}$,
- and so we slowly approach the most general case which is $\mathbf{N} = \mathbf{W}^- = (\mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}')^-$, where $\mathbf{W}$ belongs to a specific class $\mathcal{W}$, say.

A linear statistic $\mathbf{By}$ is said to be a linear unbiased estimator, LUE, for $\mathbf{K}\boldsymbol{\beta}$, where $\mathbf{K} \in \mathbb{R}^{k\times p}$, if its expectation is equal to $\mathbf{K}\boldsymbol{\beta}$, i.e.,

$$\mathrm{E}(\mathbf{By}) = \mathbf{BX}\boldsymbol{\beta} = \mathbf{K}\boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p, \quad \text{i.e., } \ \mathbf{BX} = \mathbf{K}\,.$$

When $\mathscr{C}(\mathbf{K}') \subseteq \mathscr{C}(\mathbf{X}')$ holds, $\mathbf{K}\boldsymbol{\beta}$ is said to be estimable. The LUE $\mathbf{By}$ is the best LUE, BLUE, of estimable $\mathbf{K}\boldsymbol{\beta}$ if $\mathbf{By}$ has the smallest covariance matrix in the Löwner sense among all linear unbiased estimators of $\mathbf{K}\boldsymbol{\beta}$:

$$\mathrm{cov}(\mathbf{By}) \le_{\mathsf{L}} \mathrm{cov}(\mathbf{B}_{\#}\mathbf{y}) \quad \text{for all } \mathbf{B}_{\#} : \mathbf{B}_{\#}\mathbf{X} = \mathbf{K}\,,$$

that is, $\mathrm{cov}(\mathbf{B}_{\#}\mathbf{y}) - \mathrm{cov}(\mathbf{By})$ is nonnegative definite for all $\mathbf{B}_{\#} : \mathbf{B}_{\#}\mathbf{X} = \mathbf{K}$.

Under the model $\mathscr{M}$, the ordinary least squares estimator, OLSE, for $\boldsymbol{\beta}$ is the solution minimizing the quantity $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ with respect to $\boldsymbol{\beta}$ yielding to the normal equation $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$. Thus, if $\mathbf{X}$ has full column rank, the OLSE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{+}\mathbf{y}$. Moreover, the OLSE of $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ is

$$\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} = \mathbf{X}\mathbf{X}^{+}\mathbf{y} = \mathbf{P}_{\mathbf{X}}\mathbf{y} = \mathbf{Hy} = \hat{\boldsymbol{\mu}}\,.$$

Obviously $\hat{\boldsymbol{\mu}} = \mathbf{Hy}$ is a LUE for $\mathbf{X}\boldsymbol{\beta}$; however, $\hat{\boldsymbol{\mu}}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$ only under specific conditions. Now the well-known simple version of the Gauss–Markov theorem says that under the model $\mathscr{M}_{\mathbf{I}} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n\}$, the OLSE of $\mathbf{X}\boldsymbol{\beta}$ is the BLUE of $\mathbf{X}\boldsymbol{\beta}$, or shortly

$$\hat{\boldsymbol{\mu}}(\mathscr{M}_{\mathbf{I}}) = \mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_{\mathbf{I}}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_{\mathbf{I}}) = \tilde{\boldsymbol{\mu}}(\mathscr{M}_{\mathbf{I}})\,. \tag{2}$$

Consider now the model $\mathscr{M}$ where $\mathbf{V}$ is positive definite, and suppose that $\mathbf{V}^{1/2}$ is the positive definite square root of $\mathbf{V}$. Premultiplying $\mathscr{M}$ by $\mathbf{V}^{-1/2}$ yields $\mathscr{M}_{\#} = \{\mathbf{V}^{-1/2}\mathbf{y},\ \mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta},\ \mathbf{I}_n\}$. In light of (2), the BLUE of $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}_{\#}$ equals the OLSE under $\mathscr{M}_{\#}$ and thus

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_{\#}) = \tilde{\boldsymbol{\mu}}(\mathscr{M}_{\#}) = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} =: \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}\mathbf{y},$$

where $\mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}$ is the orthogonal projector onto $\mathscr{C}(\mathbf{X})$ when the inner product matrix is $\mathbf{V}^{-1}$. It appears that

$$\mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}\mathbf{y} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_{\#})\,. \tag{3}$$

The result (3), sometimes referred to as the Aitken-approach, is well known in statistical textbooks; see Aitken (1935) [1].

It is clear that

$$\begin{aligned}\min_{\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}\mathbf{y})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}\mathbf{y}) \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)\,,\end{aligned}$$

where $\boldsymbol{\beta}_0$ is any solution to the generalized normal equation

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{4}$$

Equation (4) is always, i.e., for any $\mathbf{y} \in \mathbb{R}^n$, solvable for $\boldsymbol{\beta}$ and the general solution can be expressed, e.g., as

$$\begin{aligned}\boldsymbol{\beta}_0 &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + [\mathbf{I}_p - \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{+}]\mathbf{t} \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + (\mathbf{I}_p - \mathbf{P}_{\mathbf{X}'})\mathbf{t}\,,\end{aligned}$$

where $\mathbf{t} \in \mathbb{R}^p$ is free to vary. Thus

$$\mathbf{X}\boldsymbol{\beta}_0 = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}\mathbf{y}.$$

What about if $\mathbf{V}$ is singular? Can we straight away replace $\mathbf{V}^{-1}$ with $\mathbf{V}^{+}$ or even with an arbitrary generalized inverse $\mathbf{V}^{-}$? No, we better be careful: we need further information before such replacement can be done.

What happens if we try to use

$$\mathbf{P}_{\mathbf{X};\mathbf{V}^{+}}\mathbf{y} := \mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{+}\mathbf{y}$$

as a BLUE? First we observe that $\mathbf{P}_{\mathbf{X};\mathbf{V}^{+}}\mathbf{y}$ is a LUE for $\mathbf{X}\boldsymbol{\beta}$ if and only if

$$\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{+}\mathbf{X} = \mathbf{X}\,,$$

which, by Proposition 1.1 below, holds if and only if $\mathscr{C}(\mathbf{X}') \subseteq \mathscr{C}(\mathbf{X}'\mathbf{V}^{+}) = \mathscr{C}(\mathbf{X}'\mathbf{V})$, which further is equivalent to each of the following conditions:

$$\mathscr{C}(\mathbf{X}') = \mathscr{C}(\mathbf{X}'\mathbf{V})\,, \quad \mathrm{r}(\mathbf{X}) = \mathrm{r}(\mathbf{X}'\mathbf{V})\,, \quad \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})^{\perp} = \{\mathbf{0}\}\,. \tag{5}$$

Above we have used the rank rule of the matrix product

$$\mathrm{r}(\mathbf{AB}) = \mathrm{r}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{B})^{\perp}. \tag{6}$$

From (5) we observe that if $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{V})$, then $\mathbf{P}_{\mathbf{X};\mathbf{V}^{+}}\mathbf{y}$ is unbiased for $\mathbf{X}\boldsymbol{\beta}$. The model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, where $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{V})$, is often called a *weakly singular* linear model. We observe that under a weakly singular linear model the product $\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{+}$ is invariant for any choice of $(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}$ in view of the following Proposition, cf. [41, Lemma 2.2.4].

Proposition 1.1. *For nonnull matrices $\mathbf{A}$ and $\mathbf{C}$ the following holds:*

(a) $\mathbf{AB}^{-}\mathbf{C} = \mathbf{AB}^{+}\mathbf{C}$ *for all* $\mathbf{B}^{-} \iff \mathscr{C}(\mathbf{C}) \subseteq \mathscr{C}(\mathbf{B}) \,\&\, \mathscr{C}(\mathbf{A}') \subseteq \mathscr{C}(\mathbf{B}')$.

(b) $\mathbf{AA}^{-}\mathbf{C} = \mathbf{C}$ *for some (and hence for all)* $\mathbf{A}^{-} \iff \mathscr{C}(\mathbf{C}) \subseteq \mathscr{C}(\mathbf{A})$.

(c) $\mathbf{C}'\mathbf{A}^{-}\mathbf{A} = \mathbf{C}'$ *for some (and hence for all)* $\mathbf{A}^{-} \iff \mathscr{C}(\mathbf{C}) \subseteq \mathscr{C}(\mathbf{A}')$.

Things become a bit trickier when we consider an estimator like

$$\mathbf{P}_{\mathbf{X};\mathbf{V}^{-}}\mathbf{y} := \mathbf{X}(\mathbf{X}'\mathbf{V}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-}\mathbf{y},$$

where $\mathbf{V}^{-}$ is a given generalized inverse of $\mathbf{V}$. Supposing that $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{V})$ we see that the observed value of $\mathbf{P}_{\mathbf{X};\mathbf{V}^{-}}\mathbf{y}$ is invariant for $\mathbf{V}^{-}$ if and only if $\mathbf{y} \in \mathscr{C}(\mathbf{V})$. Do we know that this holds? The answer is *yes* in the case of a *consistent* linear model by which we mean such a model where the observed value of $\mathbf{y}$ belongs to $\mathscr{C}(\mathbf{X} : \mathbf{V})$:

$$\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{X}) \oplus \mathscr{C}(\mathbf{VM}) = \mathscr{C}(\mathbf{X}) \boxplus \mathscr{C}(\mathbf{MV})\,, \tag{7}$$

where "$\oplus$" refers to the direct sum and "$\boxplus$" to the direct sum of orthogonal subspaces. For decompositions in (7), see [37, Lemma 2.1]. The models we consider are assumed to be consistent in the sense of (7) and sometimes we use phrase "$\mathbf{y}$ belongs to $\mathscr{C}(\mathbf{X} : \mathbf{V})$ with probability 1, or shortly w.p. 1". For consistency, see, e.g., [8].

There is a related decomposition, see, e.g., [35, Th. 8]: for any conformable matrices $\mathbf{A}$ and $\mathbf{B}$ we have

$$\mathscr{C}(\mathbf{A} : \mathbf{B}) = \mathscr{C}(\mathbf{A} : \mathbf{Q}_{\mathbf{A}}\mathbf{B}), \text{ and thereby } \mathbf{P}_{(\mathbf{A}:\mathbf{B})} = \mathbf{P}_{\mathbf{A}} + \mathbf{P}_{\mathbf{Q}_{\mathbf{A}}\mathbf{B}}. \tag{8}$$

Thus if $\mathbf{X} = (\mathbf{X}_1 : \mathbf{X}_2)$ and

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P}_{(\mathbf{X}_1:\mathbf{X}_2)} = \mathbf{I}_n - (\mathbf{P}_{\mathbf{X}_2} + \mathbf{P}_{\mathbf{M}_2\mathbf{X}_1}) = \mathbf{M}_2\mathbf{Q}_{\mathbf{M}_2\mathbf{X}_1},$$

and by (6),

$$\mathrm{r}(\mathbf{M}_2\mathbf{X}_1) = \mathrm{r}(\mathbf{X}_1) - \dim \mathscr{C}(\mathbf{X}_1) \cap \mathscr{C}(\mathbf{X}_2).$$

As for the structure of the paper, in the next section we recall the fundamental BLUE equation which literally has a fundamental role for our considerations. In Section 3 we go through some mathematical properties of the so-called $\mathcal{W}$-matrices, i.e., the class $\mathcal{W}$, and in Section 4 we introduce some representations of the BLUEs. The use of the class $\mathcal{W}$ in the partitioned model is explored in Section 5. Sections 6 and 7 are devoted to particular properties of the perp-operator $\perp$ and for the linear sufficiency, respectively. In Section 8 we deal with the equality of the BLUEs under two models and in the last section we briefly discuss the model with new future observations. This paper is a review paper containing no essentially new results. However, we believe that our review provides a useful summary of the its area and thereby increases the insights and appreciation to the presented approach to best linear unbiased estimation.

2 The Fundamental BLUE Equation

In what follows, we frequently refer to the following Proposition, sometimes called the fundamental BLUE equation, see, e.g., Drygas [11, p. 55], Rao [38, p. 282], and Baksalary [2].

Proposition 2.1. *Consider the linear model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. *Then* $\mathbf{G}\mathbf{y}$ *is the* BLUE *for* $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ *if and only if* $\mathbf{G}$ *satisfies the equation*

$$\mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp}) = (\mathbf{X} : \mathbf{0}). \tag{9}$$

The corresponding condition for $\mathbf{B}\mathbf{y}$ *to be the* BLUE *of an estimable* $\mathbf{K}\boldsymbol{\beta}$ *is*

$$\mathbf{B}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp}) = (\mathbf{K} : \mathbf{0}). \tag{10}$$

Proposition 2.1 offers an extremely handy tool to check whether a given estimator is a BLUE. Moreover, it provides a convenient way to introduce various representations for the BLUE. Equation (9) is always solvable for $\mathbf{G}$ while (10) is solvable for $\mathbf{B}$ if and only if $\mathbf{K}\boldsymbol{\beta}$ is estimable. The solutions are unique if and only if $\mathscr{C}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp}) = \mathbb{R}^n$. As said, one choice for $\mathbf{X}^{\perp}$ is $\mathbf{M} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}}$. We can define the set $\{\mathbf{P}_{\boldsymbol{\mu}|\mathscr{M}}\}$ as follows:

$$\mathbf{G} \in \{\mathbf{P}_{\boldsymbol{\mu}|\mathscr{M}}\} \iff \mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0}).$$

If $\mathbf{G}_0$ is one particular solution for (9) then the general solution can be expressed as

$$\mathbf{G}_0 + \mathbf{E}(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}:\mathbf{V})}),$$

where $\mathbf{E} \in \mathbb{R}^{n\times n}$ is free to vary.

We see at once that under a weakly singular linear model we have

$$\mathbf{P}_{\mathbf{X};\mathbf{V}^+}(\mathbf{X} : \mathbf{V}\mathbf{M}) = \mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0}),$$

and it actually appears that

$$\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{+}\mathbf{y} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) \iff \mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{V})\,;$$

see [44, Cor. 1.1] and [34, p. 286]. Moreover, if $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{V})$, then for any $\mathbf{V}^{-}$,

$$\mathbf{P}_{\mathbf{X};\mathbf{V}^{-}}\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-}\mathbf{y} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})\,,$$

and $\mathbf{P}_{\mathbf{X};\mathbf{V}^{-}}\mathbf{y}$ is invariant for all generalized inverses involved assuming that the model is consistent.

Following Rao (1971) [40, Sec. 4] we can consider a matrix $\mathbf{W}$ defined as

$$\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{U}'\mathbf{X}' \in \mathrm{NND}_n\,,$$

where NND_n stands for the set of nonnegative definite (symmetric) $n \times n$ matrices and $\mathbf{U} \in \mathbb{R}^{p\times s}$ (for some s) is such that $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$; then we may denote $\mathbf{W} \in \mathcal{W}_{\geq}$. A more general class is such where

$$\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}' \in \mathbb{R}^{n\times n},$$

with $\mathbf{T}$ being any $p \times p$ matrix such that $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$. The set of such matrices $\mathbf{W}$ will be denoted as $\mathcal{W}$. Consider then the estimator

$$\mathbf{P}_{\mathbf{X};\mathbf{W}^{-}}\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}\mathbf{y}, \quad \text{where } \mathbf{W} \in \mathcal{W}\,.$$

Now by Proposition 1.1, $\mathbf{X}'\mathbf{W}^{-}\mathbf{X}$ is invariant for the choice of the generalized inverse of $\mathbf{W}$ if and only if

$$\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{W}') \quad \text{and} \quad \mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{W})\,. \tag{11}$$

It can be shown that

$$\mathscr{C}(\mathbf{W}') = \mathscr{C}(\mathbf{W})\,, \tag{12}$$

and thereby (11) holds, because assumption $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$ obviously implies $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{W})$. Using Proposition 1.1 we can conclude that $\mathbf{P}_{\mathbf{X};\mathbf{W}^{-}}\mathbf{y}$ is invariant for any choice of generalized inverses involved supposing that the model is consistent.

We further observe that $\mathbf{P}_{\mathbf{X};\mathbf{W}^{-}}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0})$ can be written as

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}(\mathbf{X} : \mathbf{W}\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,, \tag{13}$$

where the second part $\mathbf{P}_{\mathbf{X};\mathbf{W}^{-}}\mathbf{W}\mathbf{M} = \mathbf{0}$ holds in light of (11). By Proposition 1.1, the first part of (13),

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}\mathbf{X} = \mathbf{X} \tag{14}$$

holds if and only if $\mathscr{C}(\mathbf{X}') \subseteq \mathscr{C}[\mathbf{X}'(\mathbf{W}^{-})'\mathbf{X}]$, i.e.,

$$\mathrm{r}(\mathbf{X}) = \mathrm{r}[\mathbf{X}'(\mathbf{W}^{-})'\mathbf{X}] = \mathrm{r}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})\,. \tag{15}$$

The above equality holds in view of

$$\mathrm{r}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X}) = \mathrm{r}[\mathbf{X}'\mathbf{W}^{-}(\mathbf{X} : \mathbf{W}\mathbf{M})] = \mathrm{r}(\mathbf{X}'\mathbf{W}^{-}\mathbf{W}) = \mathrm{r}(\mathbf{X})\,,$$

where we have used the assumption $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{W}')$. Thereby, under the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$,

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}\mathbf{y} = \mathrm{BLUE}(\boldsymbol{\mu} \mid \mathscr{M})\,, \text{ i.e., } \mathbf{P}_{\mathbf{X};\mathbf{W}^{-}} \in \{\mathbf{P}_{\boldsymbol{\mu}|\mathscr{M}}\}\,. \tag{16}$$

There is one interesting approach to demonstrate the usefulness of matrix $\mathbf{W} \in \mathcal{W}_{\geq}$. Namely it is clear that

$$\mathbf{G}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{0}) \iff \mathbf{G}(\mathbf{X} : \mathbf{WM}) = (\mathbf{X} : \mathbf{0}),$$

where $\mathbf{W} = \mathbf{V} + \mathbf{XUU}'\mathbf{X}' \in \mathcal{W}_{\geq}$. Observing that $\mathscr{M}_{\mathbf{W}} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{W}\}$ is a weakly singular linear model we can conclude that

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}\mathbf{y} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_{\mathbf{W}}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}). \tag{17}$$

For (17) see also [9, Th. 10.1.3].

After this longish Introduction to Basic Tools and the Fundamental BLUE Equation, we will focus in more details on the properties of matrix which is of type $\mathbf{W}$ and its usage in BLUE-related matters. It appears to be surprisingly useful and powerful tool when dealing with linear statistical models.

3 Properties of the Class $\mathcal{W}$

For a given linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, let the set $\mathcal{W}$ of $n \times n$ matrices be defined as

$$\mathcal{W} = \left\{\mathbf{W} \in \mathbb{R}^{n\times n} : \mathbf{W} = \mathbf{V} + \mathbf{XTX}',\ \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})\right\}. \tag{18}$$

In (18), $\mathbf{T}$ can be any $p \times p$ matrix as long as $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$ is satisfied. It is clear that we can always choose $\mathbf{T} = \alpha^2\mathbf{I}_n$, where α is an arbitrary nonzero scalar. Moreover, $\mathbf{V} \in \mathcal{W}$ if and only if $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{V})$. If there is a need to emphasize that there is a particular model $\mathscr{M}$, say, under consideration we will use notation $\mathcal{W}(\mathscr{M})$. Sometimes we use the phrases like "$\mathbf{A}$ is a $\mathcal{W}$-matrix" indicating that $\mathbf{A} \in \mathcal{W}$.

Choosing $\mathbf{T}$ in (18) nonnegative definite, i.e., putting $\mathbf{T} = \mathbf{UU}'$ (for some $\mathbf{U}$), we get the set $\mathcal{W}_{\geq}$ of nonnegative definite matrices defined as

$$\mathcal{W}_{\geq} = \left\{\mathbf{W} \in \mathsf{NND}_n : \mathbf{W} = \mathbf{V} + \mathbf{XUU}'\mathbf{X}',\ \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})\right\}. \tag{19}$$

In (19), $\mathbf{U}$ can be any matrix comprising p rows so that $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$ is satisfied. Using $\mathcal{W}_{\geq}$ instead of $\mathcal{W}$ some considerations can become simpler, as can be expected.

Proposition 3.1 collects together some important properties of the class $\mathcal{W}$; see, e.g., [35, Prop. 12.1].

Proposition 3.1. *Let $\mathbf{V}$ be an $n \times n$ nonnegative definite matrix, let $\mathbf{X}$ be an $n \times p$ matrix, and define $\mathbf{W}$ as $\mathbf{W} = \mathbf{V} + \mathbf{XTX}'$, where $\mathbf{T}$ is a $p \times p$ matrix. Then the following statements are equivalent:*

(a) $\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{W})$,

(b) $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{W})$,

(c) $\mathrm{r}(\mathbf{X} : \mathbf{V}) = \mathrm{r}(\mathbf{W})$,

(d) $\mathbf{X}'\mathbf{W}^{-}\mathbf{X}$ *is invariant for any choice of* $\mathbf{W}^{-}$,

(e) $\mathscr{C}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})$ *is invariant for any choice of* $\mathbf{W}^{-}$,

(f) $\mathscr{C}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X}) = \mathscr{C}(\mathbf{X}')$ *for any choice of* $\mathbf{W}^{-}$,

(g) $\mathrm{r}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X}) = \mathrm{r}(\mathbf{X})$ *irrespective of the choice of* $\mathbf{W}^{-}$,

(h) $\mathrm{r}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})$ *is invariant with respect to the choice of* $\mathbf{W}^{-}$,

(i) $\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}\mathbf{X} = \mathbf{X}$ *for any choices of* $\mathbf{W}^{-}$ *and* $(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}$.

Moreover, each of these statements is equivalent to

(a') $\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{W}')$,

and hence to the statements (b')–(i') *obtained from* (b)–(i)*, by setting* $\mathbf{W}'$ *in place of* $\mathbf{W}$.

Shortly said, given the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, $\mathbf{W} \in \mathcal{W}(\mathscr{M})$ if and only if any of the conditions in Proposition 3.1 holds. Observe that the invariance properties in (f)–(i) concern also the choice of $\mathbf{W} \in \mathcal{W}$; not only its generalized inverse. Actually, we will return to this property in due course.

As references to Proposition 3.1, in addition to [40, Sec. 4], we may mention, e.g., [6, Th. 1], [7, Th. 2], [5, Th. 2], [14, p. 468], and [35, Sec. 12.3].

Notice that the equivalence of (g) and (h) of Proposition 3.1, is the same as that between (14) and (15). Moreover, we can conclude that the statement

$$\mathbf{P}_{\mathbf{X};\mathbf{W}^{-}} \in \{\mathbf{P}_{\boldsymbol{\mu}|\mathscr{M}}\} \quad \text{for any choices of } \mathbf{W}^{-} \text{ and } (\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-} \tag{20}$$

is equivalent to the conditions in Proposition 3.1.

Let's take a quick look at some some developments of the equivalence of the statements of Proposition 3.1; see, in particular, Baksalary & Mathew (1990) [5, Sec. 3]. Consider the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, and the following generalized normal equation:

$$\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{A}\mathbf{y}, \tag{21}$$

where $\mathbf{A}$ is a given $n \times n$ matrix. If $\mathbf{A}$ is nonnegative definite (and symmetric) then (21) has a solution for $\boldsymbol{\beta}$ for every $\mathbf{y}$ and the solution minimizes

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{A}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2_{\mathbf{A}}\,.$$

Rao (1971) [40, p. 372] pointed out that we can consider a more general class of matrices $\mathbf{A}$ by allowing $\mathbf{A}$ to be any matrix for which $\boldsymbol{\beta}$ is solvable from (21). Assuming that the model $\mathscr{M}$ is consistent, i.e., $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$, we observe that (21) is solvable for any $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$ if and only if $\mathbf{X}'\mathbf{A}(\mathbf{X} : \mathbf{V})\mathbf{t} \in \mathscr{C}(\mathbf{X}'\mathbf{A}\mathbf{X})$ for all $\mathbf{t} \in \mathbb{R}^{n+p}$, i.e.,

$$\mathscr{C}(\mathbf{X}'\mathbf{A}\mathbf{V}) \subseteq \mathscr{C}(\mathbf{X}'\mathbf{A}\mathbf{X})\,. \tag{22}$$

Rao [40, Th. 4.2] showed that if (22) holds, then for any solution $\boldsymbol{\beta}_0$ of (21) the estimator $\mathbf{X}\boldsymbol{\beta}_0$ is the BLUE of $\mathbf{X}\boldsymbol{\beta}$ if and only if $\mathbf{A}$ is of the form

$$\mathbf{A} = (\mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}')^{-} + \mathbf{J}\,, \tag{23}$$

and satisfies the equality $\mathrm{r}(\mathbf{X}'\mathbf{A}\mathbf{X}) = \mathrm{r}(\mathbf{X})$, with $\mathbf{T}$ and $\mathbf{J}$ being arbitrary matrices such that

$$\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}') = \mathscr{C}(\mathbf{V} + \mathbf{X}\mathbf{T}'\mathbf{X}')\,, \tag{24}$$

and $\mathbf{X}'\mathbf{J}(\mathbf{X} : \mathbf{V}) = (\mathbf{0} : \mathbf{0})$. Baksalary & Puntanen (1989) [6, Th. 1] proved that the condition (24) may be simplified because

$$\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}') \iff \mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{V} + \mathbf{X}\mathbf{T}'\mathbf{X}')\,,$$

c.f. (12), and that, under (24) the condition $\mathrm{r}(\mathbf{X}'\mathbf{A}\mathbf{X}) = \mathrm{r}(\mathbf{X})$ is redundant since for every $\mathbf{A}$ of the form (23):

$$\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}') \implies \mathrm{r}(\mathbf{X}'\mathbf{A}\mathbf{X}) = \mathrm{r}(\mathbf{X})\,. \tag{25}$$

Baksalary et al. [7, Th. 2], showed that the implication (25) may be reversed, in the sense that if

$$\mathrm{r}[\mathbf{X}'(\mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}')^{-}\mathbf{X}] = \mathrm{r}(\mathbf{X}) \quad \text{for every } (\mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}')^{-} \tag{26}$$

then $\mathscr{C}[\mathbf{X}'(\mathbf{V}+\mathbf{X}\mathbf{T}\mathbf{X}')^{-}\mathbf{X}] = \mathscr{C}(\mathbf{X} : \mathbf{V})$; this confirms the equivalence of (f) and (g) in Proposition 3.1. Moreover, they raised the question whether it is possible to relax the condition (26) by requiring only that

$$\mathrm{r}[\mathbf{X}'(\mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}')^{-}\mathbf{X}] \quad \text{is invariant for every } (\mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}')^{-}. \tag{27}$$

Baksalary & Mathew [5, Th. 2] showed that the answer is positive to this question; thus (g) and (d) in Proposition 3.1 are equivalent.

4 Representations of the BLUE

In this section we present an important matrix decomposition in Proposition 4.1 and some its consequences. Before it, however, a few words about the matrix $\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}$ which we denote as

$$\dot{\mathbf{M}} = \mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\,.$$

The matrix $\dot{\mathbf{M}}$ is not necessarily unique for any $(\mathbf{M}\mathbf{V}\mathbf{M})^{-}$; it is unique if and only if $\mathrm{r}(\mathbf{X} : \mathbf{V}) = n$. However, we always have

$$\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{+}\mathbf{M} = (\mathbf{M}\mathbf{V}\mathbf{M})^{+}\mathbf{M} = \mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{+} = (\mathbf{M}\mathbf{V}\mathbf{M})^{+}.$$

In particular, for a positive definite $\mathbf{V}$ we have, for any $(\mathbf{M}\mathbf{V}\mathbf{M})^{-}$,

$$\begin{aligned}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M} &= \mathbf{V}^{-1/2}\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{M}}\mathbf{V}^{-1/2} \\ &= \mathbf{V}^{-1/2}(\mathbf{I}_n - \mathbf{P}_{(\mathbf{V}^{1/2}\mathbf{M})^{\perp}})\mathbf{V}^{-1/2} \\ &= \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1},\end{aligned}$$

where we have used the obvious fact $\mathscr{C}(\mathbf{V}^{1/2}\mathbf{M})^{\perp} = \mathscr{C}(\mathbf{V}^{-1/2}\mathbf{X})$.

Proposition 4.1. *Consider the linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. Let $\mathbf{T}$ be any $p \times p$ matrix such that the matrix $\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}'$ satisfies the condition $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$, i.e., $\mathbf{W} \in \mathcal{W}(\mathscr{M})$, and denote $\dot{\mathbf{M}} = \mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}$. Then*

(a) $\mathbf{P}_{\mathbf{W}}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{P}_{\mathbf{W}} = \mathbf{W}^{+} - \mathbf{W}^{+}\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{+}$,

(b) $\mathbf{P}_{\mathbf{W}}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{P}_{\mathbf{W}} = (\mathbf{M}\mathbf{V}\mathbf{M})^{+} = \mathbf{P}_{\mathbf{W}}\dot{\mathbf{M}}\mathbf{P}_{\mathbf{W}}$,

(c) $\mathbf{P}_{\mathbf{X};\mathbf{W}^{+}} = \mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{+} = \mathbf{P}_{\mathbf{W}} - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{P}_{\mathbf{W}}$,

(d) $\mathbf{P}_{\mathbf{X};\mathbf{W}^{+}} \in \{\mathbf{P}_{\boldsymbol{\mu}|\mathscr{M}}\}$.

For the proof of (a), see [35, Prop. 15.2] and [22, Cor. 2.2]. Some related considerations (in full rank case) appear also in [28, pp. 415–416] and [26, pp. 323–324].

We observe that in light of (8) we have

$$\mathbf{P}_{\mathbf{W}} = \mathbf{P}_{\mathbf{X}} + \mathbf{P}_{\mathbf{M}\mathbf{V}} = \mathbf{H} + \mathbf{P}_{\mathbf{M}\mathbf{V}\mathbf{M}}\,,$$

which implies (b) of Proposition 4.1. Premultiplying (a) by $\mathbf{W}$ and using $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{W}') = \mathscr{C}(\mathbf{W})$ gives

$$\begin{aligned}\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+ &= \mathbf{P}_\mathbf{W} - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{P}_\mathbf{W} \\ &= (\mathbf{I}_n - \mathbf{V}\dot{\mathbf{M}})\mathbf{P}_\mathbf{W}\,. \end{aligned} \tag{28}$$

From (28) we immediately confirm that $\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+$ is invariant with respect to the choice of $\mathbf{W} \in \mathcal{W}$ as was pointed out in the context of Proposition 3.1.

Premultiplying (28) by $\mathbf{H} = \mathbf{P}_\mathbf{X}$ gives further expressions:

$$\begin{aligned}\mathbf{P}_{\mathbf{X};\mathbf{W}^+} &= \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+ \\ &= \mathbf{P}_\mathbf{W} - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{P}_\mathbf{W} \\ &= \mathbf{P}_\mathbf{W} - \mathbf{V}(\mathbf{M}\mathbf{V}\mathbf{M})^+ \\ &= \mathbf{H} - \mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{P}_\mathbf{W} \\ &= \mathbf{H} - \mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^+\mathbf{M}\,. \end{aligned} \tag{29}$$

It is worth emphasizing that in (28) and (29) we use the Moore–Penrose inverse $\mathbf{A}^+$, wherever it is marked while the notation $\mathbf{A}^-$ means that we can use any generalized inverse.

As we have already in (16) observed we have under $\mathscr{M}$

$$\mathbf{P}_{\mathbf{X};\mathbf{W}^-}\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\mathbf{y} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \tilde{\boldsymbol{\mu}}\,.$$

Notice that $\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'$ is invariant with respect to the choice of generalized inverses involved and

$$\mathbf{P}_{\mathbf{X};\mathbf{W}^+} = \mathbf{X}(\mathbf{X}'\mathbf{W}^+\mathbf{X})^+\mathbf{X}'\mathbf{W}^+ = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+$$

for any choice of $\mathbf{W}^-$ and $(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-$.

From (29) we can conclude that under the consistent model $\mathscr{M}$, i.e., assuming that $\mathbf{y} \in \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$,

$$\tilde{\boldsymbol{\mu}} = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\mathbf{y} = \mathbf{y} - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{y} = \mathbf{H}\mathbf{y} - \mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{y},$$

and

$$\mathbf{y} - \tilde{\boldsymbol{\mu}} = \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{y} = \tilde{\boldsymbol{\varepsilon}} = \mathrm{BLUE}\text{'s residual.} \tag{30}$$

The covariance matrix of the $\tilde{\boldsymbol{\mu}} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$ can be expressed as

$$\begin{aligned}\mathrm{cov}(\tilde{\boldsymbol{\mu}}) &= \mathbf{H}\mathbf{V}\mathbf{H} - \mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{V}\mathbf{H} \\ &= \mathrm{cov}(\mathbf{H}\mathbf{y}) - \mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{V}\mathbf{H}\,, \end{aligned}$$

as well as $\mathrm{cov}(\tilde{\boldsymbol{\mu}}) = \mathbf{V} - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{V}$. Notice that

$$\mathrm{cov}(\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}) = \mathrm{cov}(\hat{\boldsymbol{\mu}}) - \mathrm{cov}(\tilde{\boldsymbol{\mu}}) = \mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{V}\mathbf{H}\,.$$

Postmultiplying (28) by $\mathbf{W}$ yields

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}' = \mathbf{W} - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{V}$$

and thereby the BLUE's covariance matrix has a representation

$$\mathrm{cov}(\tilde{\boldsymbol{\mu}}) = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}' - \mathbf{X}\mathbf{T}\mathbf{X}'. \tag{31}$$

The form (31) was first expressed, using $\mathbf{T} = \alpha^2\mathbf{I}_n$, by Rao (1971) [40, p. 382] and Rao & Mitra (1971) [41, p. 289]. Rao [40, p. 384–385] pointed out the use of $\mathbf{W} \in \mathcal{W}$ with condition $\mathscr{C}(\mathbf{W}') = \mathscr{C}(\mathbf{X} : \mathbf{V})$, which, as stated earlier, is actually not needed. For further references regarding the $\mathrm{cov}(\tilde{\boldsymbol{\mu}})$, see, e.g., [7] and [22, 23].

Remark 1. The referee of our paper interestingly commented as follows:

For singular models where we use the matrix $\mathbf{W}$ to get representations for the BLUE of $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, the covariance matrix is a somewhat lengthy matrix expression, not as elegant as the expression given above for a nonsingular $\mathbf{V}$. For singular models, are there some choices of $\mathbf{W}$ that gives some elegant simplifications for the covariance matrix of the BLUE of $\boldsymbol{\mu}$?

Indeed, for a positive definite $\mathbf{V}$ we can choose $\mathbf{W} = \mathbf{V}$ and thus (31) gives the well-known formula $\mathrm{cov}(\tilde{\boldsymbol{\mu}}) = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'$. Similarly, for a weakly singular linear model, i.e., when $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{V})$, we can again choose $\mathbf{W} = \mathbf{V}$ and obtain

$$\mathrm{cov}(\tilde{\boldsymbol{\mu}}) = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-}\mathbf{X})^{-}\mathbf{X}'.$$

But for the question of further choices of $\mathbf{W}$ yielding elegant simplifications we are afraid that we must raise our hands and postpone it for further research. However, as pointed out by Rao (1978) [36], missing the role of the matrix $\mathbf{T}$ in (31) can yield wrong results. Rao points out that the choice of $\mathbf{T} = \alpha^2\mathbf{I}_p$ has some advantages, like even if $\mathbf{V}$ is singular, the matrix $\mathbf{W} = \mathbf{V} + \alpha^2\mathbf{X}\mathbf{X}'$ may be positive definite.

There is a related curious problem: suppose $\mathbf{W}$ is defined as

$$\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{U}'\mathbf{X}' = (\mathbf{V}^{1/2} : \mathbf{X}\mathbf{U})(\mathbf{V}^{1/2} : \mathbf{X}\mathbf{U})',$$

so that $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$. What is the choice for $\mathbf{U}$ making $\mathbf{W}^{-}$ to be also a generalized inverse of $\mathbf{V}$, i.e.,

$$\mathbf{V}(\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{U}'\mathbf{X}')^{-}\mathbf{V} = \mathbf{V}.$$

Groß [12] showed that one such choice is $\mathbf{U} = \mathbf{X}^{+}(\mathbf{I}_n - \mathbf{P}_{\mathbf{V}})$. For related discussion, see also [32].

The ordinary, unweighted sum of squares of errors SSE is defined as

$$\mathrm{SSE}(\mathbf{I}) = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \mathbf{y}'\mathbf{M}\mathbf{y}\,,$$

while the weighted SSE, when $\mathbf{V}$ is positive definite, is

$$\begin{aligned}\mathrm{SSE}(\mathbf{V}) &= \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2_{\mathbf{V}^{-1}} = \|\mathbf{y} - \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}\mathbf{y}\|^2_{\mathbf{V}^{-1}} \\ &= \mathbf{y}'[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{y} \\ &= \mathbf{y}'\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{y} = \mathbf{y}'\dot{\mathbf{M}}\mathbf{y}\,.\end{aligned}$$

In the general case, the weighted SSE can be defined as

$$\mathrm{SSE}(\mathbf{W}) = (\mathbf{y} - \tilde{\boldsymbol{\mu}})'\mathbf{W}^{-}(\mathbf{y} - \tilde{\boldsymbol{\mu}})\,,$$

where $\mathbf{W} \in \mathcal{W}$. Then, recalling that by (30), the BLUE's residual is $\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \tilde{\boldsymbol{\mu}} = \mathbf{V}\dot{\mathbf{M}}\mathbf{y}$, we observe the following:

$$\begin{aligned}\mathrm{SSE}(\mathbf{W}) &= \tilde{\boldsymbol{\varepsilon}}'\mathbf{W}^{-}\tilde{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{\varepsilon}}'\mathbf{V}^{-}\tilde{\boldsymbol{\varepsilon}} = \mathbf{y}'\dot{\mathbf{M}}\mathbf{y} \\ &= \mathbf{y}'[\mathbf{W}^{-} - \mathbf{W}^{-}\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}]\mathbf{y}.\end{aligned}$$

It can be further shown that under $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}\}$, $\mathrm{SSE}(\mathbf{W})$ provides an unbiased estimator of σ^2:

$$\mathrm{E}(\mathbf{y}'\dot{\mathbf{M}}\mathbf{y}/f) = \sigma^2, \quad \text{where } f = \mathrm{r}(\mathbf{V}\mathbf{M})\,.$$

Remark 2. Baksalary et al. (1990) [7, Th. 3] considered the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ and proved that if $\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}' \in \mathcal{W}(\mathscr{M})$ then the equality

$$\mathbf{W} = \mathbf{V}\mathbf{B}(\mathbf{B}'\mathbf{V}\mathbf{B})^{-}\mathbf{B}'\mathbf{V} + \mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}' \tag{32}$$

holds for a matrix $\mathbf{B}$ if and only if

$$\mathscr{C}(\mathbf{V}\mathbf{W}^{-}\mathbf{X}) \subseteq \mathscr{C}(\mathbf{B})^{\perp} \text{ and } \mathscr{C}(\mathbf{V}\mathbf{M}) \subseteq \mathscr{C}(\mathbf{V}\mathbf{B})\,. \tag{33}$$

It is clear that the choice of $\mathbf{B} = \mathbf{M}$ satisfies (33) and thereby also (32) holds for $\mathbf{B} = \mathbf{M}$. Postmultiplying (32) by $\mathbf{W}^{+}$ in that situation gives (c) of Proposition 4.1.

Remark 3. Wang & Liski (1998) [43, p. 45] introduce an interesting matrix inequality by considering estimator $\mathbf{A}\mathbf{y}$ which is unbiased for $\mathbf{B}\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, i.e., $\mathbf{A}\mathbf{X} = \mathbf{B}\mathbf{X}$. Let $\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{U}'\mathbf{X}' \in \mathcal{W}_{\geq}(\mathscr{M})$. Then $\mathbf{B}\mathbf{P}_{\mathbf{X};\mathbf{W}^{+}}\mathbf{y} = \mathrm{BLUE}(\mathbf{B}\mathbf{X}\boldsymbol{\beta})$ and for any $\mathbf{A}$ and $\mathbf{B}$ satisfying $\mathbf{A}\mathbf{X} = \mathbf{B}\mathbf{X}$,

$$\mathrm{cov}(\mathbf{B}\mathbf{P}_{\mathbf{X};\mathbf{W}^{+}}\mathbf{y}) \leq_{\mathsf{L}} \mathrm{cov}(\mathbf{A}\mathbf{y})\,,$$

i.e.,

$$\mathbf{B}\Big[\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}' - \mathbf{X}\mathbf{U}\mathbf{U}'\mathbf{X}'\Big]\mathbf{B}' \leq_{\mathsf{L}} \mathbf{A}\mathbf{V}\mathbf{A}'\,,$$

which is further equivalent to

$$\mathbf{B}\Big[\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\Big]\mathbf{B}' \leq_{\mathsf{L}} \mathbf{A}\mathbf{W}\mathbf{A}'\,. \tag{34}$$

Equality appears in (34) if and only if $\mathbf{A}\mathbf{y}$ is the $\mathrm{BLUE}(\mathbf{B}\mathbf{X}\boldsymbol{\beta})$ which by Proposition 2.1 happens if and only if $\mathscr{C}(\mathbf{V}\mathbf{A}') \subseteq \mathscr{C}(\mathbf{X})$.

Another interesting application of the $\mathcal{W}_{\geq}$-matrix is given by [27] who considered the upper bound for

$$\delta = \mathrm{trace}[\mathrm{cov}(\hat{\boldsymbol{\mu}} \mid \mathscr{M}) - \mathrm{cov}(\tilde{\boldsymbol{\mu}} \mid \mathscr{M})] = \mathrm{trace}[\mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{V}\mathbf{H}]\,,$$

where $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. Without going into more details we may only mention that they based their proof by noting that

$$\delta = \mathrm{trace}[\mathrm{cov}(\hat{\boldsymbol{\mu}} \mid \mathscr{M}_{\mathbf{W}}) - \mathrm{cov}(\tilde{\boldsymbol{\mu}} \mid \mathscr{M}_{\mathbf{W}})]\,,$$

where $\mathscr{M}_{\mathbf{W}} = \{\mathbf{y},\ \mathbf{X}\boldsymbol{\beta},\ \mathbf{W}\}$, with $\mathbf{W} = \mathbf{V} + \alpha^2\mathbf{X}\mathbf{X}' \in \mathcal{W}_{\geq}(\mathscr{M})$ and could generalize the result of Rao (1985) [39] given for a positive definite $\mathbf{V}$.

5 Partitioned Linear Model

Consider then the estimation of $\boldsymbol{\mu}_1 = \mathbf{X}_1\boldsymbol{\beta}_1$ under the partitioned model $\mathscr{M}_{12} = \{\mathbf{y}, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \mathbf{V}\}$ assuming that $\boldsymbol{\mu}_1$ is estimable which is well known to hold if and only if

$$\mathscr{C}(\mathbf{X}_1) \cap \mathscr{C}(\mathbf{X}_2) = \{\mathbf{0}\}\,, \quad \text{i.e.,} \quad \mathrm{r}(\mathbf{M}_2\mathbf{X}_1) = \mathrm{r}(\mathbf{X}_1)\,.$$

Let us denote the small models as $\mathscr{M}_i = \{\mathbf{y}, \mathbf{X}_i\boldsymbol{\beta}_i, \mathbf{V}\}, i = 1, 2$. Corresponding to (19), $\mathbf{W}_i \in \mathcal{W}_{\geq}(\mathscr{M}_i)$ if there exists a matrix $\mathbf{L}_i$ such that

$$\mathbf{W}_i = \mathbf{V} + \mathbf{X}_i\mathbf{L}_i\mathbf{L}_i'\mathbf{X}_i'\,, \quad \mathscr{C}(\mathbf{W}_i) = \mathscr{C}(\mathbf{X}_i : \mathbf{V})\,, \quad i = 1, 2\,. \tag{35}$$

Premultiplying $\mathscr{M}_{12}$ by $\mathbf{M}_2 = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_2}$ yields the reduced model

$$\mathscr{M}_{12\cdot 2} = \{\mathbf{M}_2\mathbf{y},\ \mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1,\ \mathbf{M}_2\mathbf{V}\mathbf{M}_2\}\,,$$

which is a special case of the transformed model $\mathscr{T} = \{\mathbf{Fy}, \mathbf{FX}\boldsymbol{\beta}, \mathbf{FVF}'\}$, where $\mathbf{F} \in \mathbb{R}^{f\times n}$. In view of the Frisch–Waugh–Lovell theorem, see, e.g., [13, Sec. 6], the BLUEs of $\boldsymbol{\theta}_1 = \mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1$ under $\mathscr{M}_{12}$ and $\mathscr{M}_{12\cdot 2}$ coincide. It is noteworthy that $\boldsymbol{\theta}_1$ is estimable under $\mathscr{M}_{12}$ as well as under $\mathscr{M}_{12\cdot 2}$. An explicit expression for the BLUE of $\boldsymbol{\theta}_1$ under $\mathscr{M}_{12\cdot 2}$ can be obtained from

$$\mathrm{BLUE}(\boldsymbol{\theta}_1 \mid \mathscr{M}_{12\cdot 2}) = \mathrm{BLUE}(\boldsymbol{\theta}_1 \mid \mathscr{M}_{12}) = \mathbf{P}_{\mathbf{M}_2\mathbf{X}_1;\mathbf{W}_{rm}^-}\mathbf{M}_2\mathbf{y},$$

where

$$\mathbf{P}_{\mathbf{M}_2\mathbf{X}_1;\mathbf{W}_{rm}^-} = \mathbf{M}_2\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{W}_{rm}^-\mathbf{M}_2\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{M}_2\mathbf{W}_{rm}^- \in \{\mathbf{P}_{\boldsymbol{\theta}_1|\mathscr{M}_{12\cdot 2}}\},$$

and $\mathbf{W}_{rm}$ is an arbitrary $\mathcal{W}$-matrix in $\mathscr{M}_{12\cdot 2}$, i.e., $\mathbf{W}_{rm} \in \mathcal{W}(\mathscr{M}_{12\cdot 2})$. Notice that

$$\mathbf{P}_{\mathbf{M}_2\mathbf{X}_1;\mathbf{W}_{rm}^-}\mathbf{M}_2 \in \{\mathbf{P}_{\boldsymbol{\theta}_1|\mathscr{M}_{12}}\}.$$

Clearly any matrix of the form $\mathbf{M}_2(\mathbf{V} + \mathbf{X}_1\mathbf{K}_1\mathbf{K}_1'\mathbf{X}_1')\mathbf{M}_2$ satisfying

$$\mathscr{C}[\mathbf{M}_2(\mathbf{V} : \mathbf{X}_1\mathbf{K}_1)] = \mathscr{C}[\mathbf{M}_2(\mathbf{V} : \mathbf{X}_1)] = \mathscr{C}(\mathbf{M}_2\mathbf{W}_1), \tag{36}$$

is a $\mathcal{W}_{\geq}$-matrix in $\mathscr{M}_{12\cdot 2}$. Putting $\mathbf{K}_1 = \mathbf{L}_1$ as in (35) we can choose

$$\mathbf{W}_{rm} = \mathbf{M}_2\mathbf{W}_1\mathbf{M}_2 \in \mathcal{W}_{\geq}(\mathscr{M}_{12\cdot 2}).$$

Thus the BLUE of $\boldsymbol{\theta}_1 = \mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1$ under $\mathscr{M}_{12\cdot 2}$ can be expressed as

$$\mathrm{BLUE}(\boldsymbol{\theta}_1 \mid \mathscr{M}_{12\cdot 2}) = \mathbf{M}_2\mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{X}_1)^-\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{y}, \tag{37}$$

where

$$\dot{\mathbf{M}}_2 = \mathbf{M}_2(\mathbf{M}_2\mathbf{W}_1\mathbf{M}_2)^-\mathbf{M}_2.$$

Notice that by Proposition 4.1 the matrix

$$\mathbf{P}_{\mathbf{M}_2\mathbf{X}_1;\mathbf{W}_{rm}^+} = \mathbf{M}_2\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{W}_{rm}^-\mathbf{M}_2\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{M}_2\mathbf{W}_{rm}^+$$

belonging to $\{\mathbf{P}_{\boldsymbol{\theta}_1|\mathscr{M}_{12\cdot 2}}\}$ is unique with respect to the choice of generalized inverses indicated by "$-$" as well as with the choice of $\mathbf{W}_{rm} \in \mathcal{W}_{\geq}(\mathscr{M}_{12\cdot 2})$. It is easy to confirm that $\mathscr{C}(\mathbf{X}_1'\mathbf{M}_2) = \mathscr{C}(\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{X}_1) = \mathscr{C}(\mathbf{X}_1')$ where the last iquality holds if $\boldsymbol{\mu}_1$ is estimable in $\mathscr{M}_{12}$.

It is clear that for estimable $\boldsymbol{\mu}_1$ we have

$$\mathrm{BLUE}(\boldsymbol{\mu}_1 \mid \mathscr{M}_{12\cdot 2}) = \mathrm{BLUE}(\boldsymbol{\mu}_1 \mid \mathscr{M}_{12}) = \mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{X}_1)^-\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{y}. \tag{38}$$

Notice that by (36) $\mathbf{M}_2\mathbf{V}\mathbf{M}_2 \in \mathcal{W}_{\geq}(\mathscr{M}_{12\cdot 2})$ if and only if $\mathscr{C}(\mathbf{M}_2\mathbf{X}_1) \subseteq \mathscr{C}(\mathbf{M}_2\mathbf{V})$, i.e.,

$$\mathscr{C}(\mathbf{X}_1) \subseteq \mathscr{C}(\mathbf{X}_2 : \mathbf{V}).$$

Thus, for example for a positive definite $\mathbf{V}$ and full-rank $\mathbf{X}$ we have

$$\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) = (\mathbf{X}_1'\dot{\mathbf{M}}_{2V}\mathbf{X}_1)^{-1}\mathbf{X}_1'\dot{\mathbf{M}}_{2V}\mathbf{y}, \quad \mathrm{cov}[\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_{12})] = (\mathbf{X}_1'\dot{\mathbf{M}}_{2V}\mathbf{X}_1)^{-1},$$

where

$$\dot{\mathbf{M}}_{2V} = \mathbf{M}_2(\mathbf{M}_2\mathbf{V}\mathbf{M}_2)^-\mathbf{M}_2 = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{V}^{-1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{V}^{-1}.$$

Consider now the following choice for $\mathbf{W}_\ell \in \mathcal{W}_{\geq}(\mathscr{M}_{12})$ (with obvious partitioning)

$$\mathbf{W}_\ell = \mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{U}'\mathbf{X}' = \mathbf{V} + \mathbf{X}\begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}(\mathbf{U}_1' : \mathbf{U}_2')\mathbf{X}',$$

where $\mathbf{U} = (\mathbf{U}_1' : \mathbf{U}_2')'$ has the property

$$\mathscr{C}(\mathbf{W}_\ell) = \mathscr{C}(\mathbf{V} : \mathbf{X}_1\mathbf{U}_1 + \mathbf{X}_2\mathbf{U}_2) = \mathscr{C}(\mathbf{V} : \mathbf{X}_1 : \mathbf{X}_2)\,. \tag{39}$$

Premultiplying (39) with $\mathbf{M}_2$ gives

$$\mathscr{C}(\mathbf{M}_2\mathbf{W}_\ell) = \mathscr{C}[\mathbf{M}_2(\mathbf{V} : \mathbf{X}_1\mathbf{U}_1)] = \mathscr{C}(\mathbf{M}_2\mathbf{W}_1)\,.$$

The nonnegative definiteness of $\mathbf{W}_\ell$ means that $\mathscr{C}(\mathbf{M}_2\mathbf{W}_\ell) = \mathscr{C}(\mathbf{M}_2\mathbf{W}_\ell\mathbf{M}_2)$ and so we have proved the following:

$$\mathbf{W} \in \mathcal{W}_{\geq}(\mathscr{M}_{12}) \implies \mathbf{M}_2\mathbf{W}\mathbf{M}_2 \in \mathcal{W}_{\geq}(\mathscr{M}_{12\cdot 2})\,. \tag{40}$$

Thus the $\mathrm{BLUE}(\boldsymbol{\theta}_1 \mid \mathscr{M}_{12})$ has a representation $\mathbf{P}_{\mathbf{M}_2\mathbf{X}_1;\mathbf{W}_\ell^-}\mathbf{M}_2\mathbf{y}$, where

$$\mathbf{P}_{\mathbf{M}_2\mathbf{X}_1;\mathbf{W}_\ell^-} = \mathbf{M}_2\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{W}_\ell^-\mathbf{M}_2\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{M}_2\mathbf{W}_\ell^- \in \{\mathbf{P}_{\boldsymbol{\theta}_1|\mathscr{M}_{12\cdot 2}}\}$$

so that

$$\begin{aligned}\mathrm{BLUE}(\boldsymbol{\theta}_1 \mid \mathscr{M}_{12}) &= \mathbf{M}_2\mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1)^-\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{y},\\ \mathrm{BLUE}(\boldsymbol{\mu}_1 \mid \mathscr{M}_{12}) &= \mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1)^-\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{y},\end{aligned}$$

where

$$\dot{\mathbf{M}}_{2W} = \mathbf{M}_2(\mathbf{M}_2\mathbf{W}_\ell\mathbf{M}_2)^-\mathbf{M}_2 \quad \text{and} \quad \mathbf{W}_\ell \in \mathcal{W}_{\geq}(\mathscr{M}_{12})\,.$$

Denoting

$$\mathbf{P}_{\mathbf{M}_2\mathbf{X}_1;\mathbf{W}_\ell^+} = \mathbf{M}_2\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{W}_\ell^-\mathbf{M}_2\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{M}_2\mathbf{W}_\ell^+ \in \{\mathbf{P}_{\boldsymbol{\theta}_1|\mathscr{M}_{12\cdot 2}}\}\,,$$

we observe by Proposition 4.1 that

$$\begin{aligned}\mathbf{P}_{\mathbf{M}_2\mathbf{X}_1;\mathbf{W}_\ell^+} &= \mathbf{P}_{\mathbf{M}_2\mathbf{X}_1;\mathbf{W}_{rm}^+}\,,\\ \mathrm{BLUE}(\boldsymbol{\theta}_1 \mid \mathscr{M}_{12}) &= \mathbf{P}_{\mathbf{M}_2\mathbf{X}_1;\mathbf{W}_\ell^-}\mathbf{M}_2\mathbf{y} = \mathbf{P}_{\mathbf{M}_2\mathbf{X}_1;\mathbf{W}_{rm}^-}\mathbf{M}_2\mathbf{y}\,.\end{aligned}$$

Now we can wonder whether (40) holds for *any* $\mathbf{W} \in \mathcal{W}(\mathscr{M}_{12})$ so that $\mathbf{W}$ is not necessarily symmetric nor nonnegative definite. So, let $\mathbf{W}_t$ be of the form $\mathbf{W}_t = \mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}'$, where

$$\mathscr{C}(\mathbf{W}_t) = \mathscr{C}(\mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}') = \mathscr{C}(\mathbf{X} : \mathbf{V})\,. \tag{41}$$

To have (40) holding for any $\mathbf{W}_t \in \mathcal{W}(\mathscr{M}_{12})$ we should have

$$\mathscr{C}(\mathbf{M}_2\mathbf{W}_t\mathbf{M}_2) = \mathscr{C}[\mathbf{M}_2(\mathbf{X} : \mathbf{V})]\,. \tag{42}$$

By (41) we observe that

$$\mathscr{C}(\mathbf{M}_2\mathbf{W}_t\mathbf{M}_2) \subseteq \mathscr{C}(\mathbf{M}_2\mathbf{W}_t) = \mathscr{C}[\mathbf{M}_2(\mathbf{X} : \mathbf{V})]\,, \tag{43}$$

so that (42) holds if and only if

$$\mathscr{C}(\mathbf{M}_2\mathbf{W}_t\mathbf{M}_2) = \mathscr{C}(\mathbf{M}_2\mathbf{W}_t)\,. \tag{44}$$

In other words, the implication

$$\mathbf{W} \in \mathcal{W}(\mathscr{M}_{12}) \implies \mathbf{M}_2\mathbf{W}\mathbf{M}_2 \in \mathcal{W}(\mathscr{M}_{12\cdot 2}) \tag{45}$$

holds if and only if (44) holds; this happens, e.g., if $\mathbf{W}$ is nonnegative definite.

One alternative expression for the BLUE of $\boldsymbol{\mu}_1$ can be obtained by premultiplying the fundamental BLUE-equation

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{VM}) = (\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{0})\,, \text{ where } \mathbf{W} \in \mathcal{W},$$

by $\mathbf{M}_2$ yielding

$$(\mathbf{M}_2\mathbf{X}_1 : \mathbf{0})(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{VM}) = (\mathbf{M}_2\mathbf{X}_1 : \mathbf{0} : \mathbf{0})\,. \tag{46}$$

Because $\mathrm{r}(\mathbf{M}_2\mathbf{X}_1) = \mathrm{r}(\mathbf{X}_1)$, we can, by the rank cancellation rule of Marsaglia & Styan (1974) [31], cancel $\mathbf{M}_2$ in (46) and thus an alternative expression for (38) is

$$\tilde{\boldsymbol{\mu}}_1(\mathscr{M}_{12}) = (\mathbf{X}_1 : \mathbf{0})(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}\mathbf{y}.$$

Let us figure out what is the covariance matrix of the BLUE of estimable $\boldsymbol{\mu}_1 = \mathbf{X}_1\boldsymbol{\beta}_1$ under $\mathscr{M}_{12}$ when

$$\tilde{\boldsymbol{\mu}}_1(\mathscr{M}_{12}) = \mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1)^{-}\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{y} =: \mathbf{Ay}.$$

Notice that $\mathrm{cov}(\tilde{\boldsymbol{\mu}}_1 \mid \mathscr{M}_{12})$ is obviously unique and hence invariant for the choice of representation of the BLUE of $\boldsymbol{\mu}_1$. Choosing $\mathbf{W} = \mathbf{V} + \mathbf{XUU}'\mathbf{X}' \in \mathcal{W}_{\geq}(\mathscr{M}_{12})$, where $\mathbf{U} = (\mathbf{U}_1' : \mathbf{U}_2')'$, we get (after straightforward calculation)

$$\begin{aligned}\mathrm{cov}(\tilde{\boldsymbol{\mu}}_1 \mid \mathscr{M}_{12}) &= \mathbf{AVA}' = \mathbf{A}(\mathbf{W} - \mathbf{XUU}'\mathbf{X}')\mathbf{A}' \\ &= \mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1)^{-}\mathbf{X}_1' - \mathbf{X}_1\mathbf{U}_1\mathbf{U}_1'\mathbf{X}_1'\,,\end{aligned}$$

where, in light of part (e) of Proposition 6.1 in Section 6, $\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1$ can be written as

$$\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1 = \mathbf{X}_1'\Big[\mathbf{W}^{+} - \mathbf{W}^{+}\mathbf{X}_2(\mathbf{X}_2'\mathbf{W}^{+}\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{W}^{+}\Big]\mathbf{X}_1\,.$$

Remark 4. We can generalise the considerations in (41)–(45) for the transformed model $\mathscr{T} = \{\mathbf{Fy}, \mathbf{FX}\boldsymbol{\beta}, \mathbf{FVF}'\}$, where $\mathbf{F} \in \mathbb{R}^{f\times n}$. Then the set of $\mathcal{W}$-matrices is defined as

$$\mathcal{W}(\mathscr{T}) = \Big\{\mathbf{W} : \mathbf{W} = \mathbf{F}(\mathbf{V} + \mathbf{XNX}')\mathbf{F}',\ \mathscr{C}(\mathbf{W}) = \mathscr{C}[\mathbf{F}(\mathbf{X} : \mathbf{V})]\Big\}\,.$$

Choosing $\mathbf{W}_t = \mathbf{V} + \mathbf{XTX}' \in \mathcal{W}(\mathscr{M})$ we have

$$\mathscr{C}(\mathbf{FW}_t\mathbf{F}') \subseteq \mathscr{C}(\mathbf{FW}_t) = \mathscr{C}[\mathbf{F}(\mathbf{X} : \mathbf{V})]\,. \tag{47}$$

If we want that $\mathbf{FW}_t\mathbf{F}' \in \mathcal{W}(\mathscr{T})$, we need to have the equality in (47), which happens if and only if $\mathrm{r}(\mathbf{FW}_t\mathbf{F}') = \mathrm{r}(\mathbf{FW}_t)$. Thus one representation for the BLUE of $\mathbf{FX}\boldsymbol{\beta}$ under $\mathscr{T}$ is

$$\mathbf{FX}[\mathbf{X}'\mathbf{F}'(\mathbf{FW}_t\mathbf{F}')^{-}\mathbf{FX}]^{-}\mathbf{X}'\mathbf{F}'(\mathbf{FW}_t\mathbf{F}')^{-}\mathbf{Fy}\,,$$

where $\mathbf{W}_t \in \mathcal{W}(\mathscr{M})$ and $\mathrm{r}(\mathbf{FW}_t\mathbf{F}') = \mathrm{r}(\mathbf{FW}_t)$ as pointed out by [24, p. 287].

Remark 5. To simplify the considerations in the partitioned model $\mathscr{M}_{12}$ we could consider the subclass $\mathcal{W}_{\#}(\mathscr{M}_{12})$ of $\mathcal{W}_{\geq}(\mathscr{M}_{12})$ defined so that $\mathbf{W} \in \mathcal{W}_{\#}(\mathscr{M}_{12})$ if $\mathbf{W}_i = \mathbf{V} + \mathbf{X}_i\mathbf{L}_i\mathbf{L}_i'\mathbf{X}_i'$, $\mathscr{C}(\mathbf{W}_i) = \mathscr{C}(\mathbf{X}_i : \mathbf{V})$, $i = 1, 2$, and

$$\mathbf{W} = \mathbf{V} + \mathbf{X}_1\mathbf{L}_1\mathbf{L}_1'\mathbf{X}_1' + \mathbf{X}_2\mathbf{L}_2\mathbf{L}_2'\mathbf{X}_2'\,.$$

The benefit in using $\mathcal{W}_{\#}(\mathscr{M}_{12})$ instead of $\mathcal{W}_{\geq}(\mathscr{M}_{12})$ is that some calculations become simpler. For example, if $\mathbf{W} \in \mathcal{W}_{\#}(\mathscr{M}_{12})$, then

$$\mathbf{M}_2\mathbf{W}\mathbf{M}_2 = \mathbf{M}_2\mathbf{W}_1\mathbf{M}_2 \in \mathcal{W}(\mathscr{M}_{12\cdot 2})\,,$$
$$\dot{\mathbf{M}}_2 = \mathbf{M}_2(\mathbf{M}_2\mathbf{W}_1\mathbf{M}_2)^{\cdot}\mathbf{M}_2 = \mathbf{M}_2(\mathbf{M}_2\mathbf{W}\mathbf{M}_2)^{\cdot}\mathbf{M}_2 = \dot{\mathbf{M}}_{2W}\,,$$

while $\mathbf{W} \in \mathcal{W}_{\geq}(\mathscr{M}_{12})$ implies

$$\mathbf{M}_2\mathbf{W}\mathbf{M}_2 \in \mathcal{W}(\mathscr{M}_{12\cdot 2}) \quad \text{and} \quad \mathbf{M}_2\mathbf{W}_1\mathbf{M}_2 \in \mathcal{W}(\mathscr{M}_{12\cdot 2})$$

but the equality $\mathbf{M}_2\mathbf{W}\mathbf{M}_2 = \mathbf{M}_2\mathbf{W}_1\mathbf{M}_2$ does not necessarily hold.

6 Some Properties of the $\perp$

It is interesting to take a further look at the $\perp$-operation and its usefulness in linear models. Let's begin by citing [35, Sec. 5.13].

Proposition 6.1. *Consider the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\beta, \mathbf{V}\}$ and let $\mathbf{W} \in \mathcal{W}(\mathscr{M})$. Then*

$$\mathscr{C}(\mathbf{V}\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{W}^{-}\mathbf{X} : \mathbf{I}_n - \mathbf{W}^{-}\mathbf{W})^{\perp}, \tag{48}$$

where $\mathbf{W}^{-}$ is an arbitrary (but fixed) generalized inverse of $\mathbf{W}$. The column space $\mathscr{C}(\mathbf{V}\mathbf{X}^{\perp})$ can be expressed also as

$$\mathscr{C}(\mathbf{V}\mathbf{X}^{\perp}) = \mathscr{C}\big[(\mathbf{W}^{-})'\mathbf{X} : \mathbf{I}_n - (\mathbf{W}^{-})'\mathbf{W}'\big]^{\perp}.$$

Moreover, let $\mathbf{V}$ be possibly singular and assume that $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{V})$. Then

$$\mathscr{C}(\mathbf{V}\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{V}^{-}\mathbf{X} : \mathbf{I}_n - \mathbf{V}^{-}\mathbf{V})^{\perp} \subseteq \mathscr{C}(\mathbf{V}^{-}\mathbf{X})^{\perp},$$

where the inclusion becomes equality if and only if $\mathbf{V}$ is positive definite.

It is of particular interest to note that the perp symbol $\perp$ falls down, so to say, very "nicely" when $\mathbf{V}$ is positive definite:

$$\mathscr{C}(\mathbf{V}\mathbf{X}^{\perp})^{\perp} = \mathscr{C}(\mathbf{V}^{-1}\mathbf{X})\,,$$

but when $\mathbf{V}$ is singular we have to use a much more complicated rule (48).

Markiewicz & Puntanen [29] reviewed various features of the perp-operation, and proved, e.g., the following: If $\mathbf{W} \in \mathcal{W}$, then

$$\mathscr{C}(\mathbf{V}\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{W}^{-}\mathbf{X})^{\perp} \iff \mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathbb{R}^n.$$

For the following Proposition 6.2, see [35, Sec. 5.13] and [30, Lemma 4]. In this lemma the notation $\mathbf{A}^{1/2}$ stands for the nonnegative definite square root of a nonnegative definite matrix $\mathbf{A}$. Similarly $\mathbf{A}^{+1/2}$ denotes the Moore–Penrose inverse of $\mathbf{A}^{1/2}$ so that $\mathbf{P}_{\mathbf{A}} = \mathbf{A}^{1/2}\mathbf{A}^{+1/2} = \mathbf{A}^{+1/2}\mathbf{A}^{1/2}$.

Proposition 6.2. *Let $\mathbf{W} \in \mathcal{W}_{\geq}(\mathscr{M}_{12})$ and $\dot{\mathbf{M}}_{2W} = \mathbf{M}_2(\mathbf{M}_2\mathbf{W}\mathbf{M}_2)^{-}\mathbf{M}_2$. Then:*

(a) $\mathscr{C}(\mathbf{V}\mathbf{M})^{\perp} = \mathscr{C}(\mathbf{W}\mathbf{M})^{\perp} = \mathscr{C}(\mathbf{W}^{+}\mathbf{X} : \mathbf{Q}_{\mathbf{W}})$, *where* $\mathbf{Q}_{\mathbf{W}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{W}}$,

(b) $\mathscr{C}(\mathbf{W}^{1/2}\mathbf{M}_2)^{\perp} = \mathscr{C}(\mathbf{W}^{+1/2}\mathbf{X}_2 : \mathbf{Q}_{\mathbf{W}})\,,$

(c) $\mathscr{C}(\mathbf{W}^{1/2}\mathbf{M}_2) = \mathscr{C}(\mathbf{W}^{+1/2}\mathbf{X}_2 : \mathbf{Q}_{\mathbf{W}})^{\perp} = \mathscr{C}(\mathbf{W}^{+1/2}\mathbf{X}_2)^{\perp} \cap \mathscr{C}(\mathbf{W})\,,$

(d) $\mathbf{P}_{\mathbf{W}^{1/2}\mathbf{M}_2} = \mathbf{P}_{\mathbf{W}} - \mathbf{P}_{\mathbf{W}^{+1/2}\mathbf{X}_2} = \mathbf{P}_{\mathscr{C}(\mathbf{W}^{+1/2}\mathbf{X}_2)^{\perp}\cap\mathscr{C}(\mathbf{W})}$,

(e) $\mathbf{P}_{\mathbf{W}}\dot{\mathbf{M}}_{2W}\mathbf{P}_{\mathbf{W}} = \mathbf{W}^{+} - \mathbf{W}^{+}\mathbf{X}_2(\mathbf{X}_2'\mathbf{W}^{+}\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{W}^{+}$,

(f) $\mathbf{W}\dot{\mathbf{M}}_{2W}\mathbf{X}_1 = [\mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{W}^{+}\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{W}^{+}]\mathbf{X}_1$.

Proof. Claim (a) follows from Proposition 6.1. Let us take a look, in more details as [30, Sec. 2], at the other statements of Proposition 6.2. We observe that $(\mathbf{W}^{1/2}\mathbf{M}_2)'(\mathbf{W}^{+1/2}\mathbf{X}_2 : \mathbf{Q}_{\mathbf{W}}) = \mathbf{0}$ so that

$$\mathscr{C}(\mathbf{W}^{+1/2}\mathbf{X}_2 : \mathbf{Q}_{\mathbf{W}}) \subseteq \mathscr{C}(\mathbf{W}^{1/2}\mathbf{M}_2)^{\perp}. \tag{49}$$

We further have

$$\begin{aligned}
\mathrm{r}(\mathbf{W}^{+1/2}\mathbf{X}_2 : \mathbf{Q}_{\mathbf{W}}) &= \mathrm{r}(\mathbf{W}^{+1/2}\mathbf{X}_2) + \mathrm{r}(\mathbf{Q}_{\mathbf{W}}) \\
&= \mathrm{r}(\mathbf{X}_2) + n - \mathrm{r}(\mathbf{W}) \,, \\
\mathrm{r}(\mathbf{W}^{1/2}\mathbf{M}_2)^{\perp} &= n - \mathrm{r}(\mathbf{W}^{1/2}\mathbf{M}_2) \\
&= n - [\mathrm{r}(\mathbf{W}^{1/2}) - \dim \mathscr{C}(\mathbf{W}^{1/2}) \cap \mathscr{C}(\mathbf{X}_2)] \\
&= n - \mathrm{r}(\mathbf{W}) + \mathrm{r}(\mathbf{X}_2) \,,
\end{aligned}$$

which confirms the equality in (49), i.e., claim (b) which is obviously equivalent to (c). Part (d) follows from (c):

$$\begin{aligned}
\mathbf{P}_{\mathbf{W}^{1/2}\mathbf{M}_2} &= \mathbf{I}_n - \mathbf{P}_{(\mathbf{W}^{+1/2}\mathbf{X}_2 : \mathbf{Q}_{\mathbf{W}})} = \mathbf{I}_n - (\mathbf{Q}_{\mathbf{W}} + \mathbf{P}_{\mathbf{W}^{+1/2}\mathbf{X}_2}) \\
&= \mathbf{P}_{\mathbf{W}} - \mathbf{P}_{\mathbf{W}^{+1/2}\mathbf{X}_2} = \mathbf{P}_{\mathscr{C}(\mathbf{W}^{+1/2}\mathbf{X}_2)^{\perp}\cap\mathscr{C}(\mathbf{W})} \,.
\end{aligned}$$

In view of (d) we have

$$\begin{aligned}
\mathbf{P}_{\mathbf{W}}\dot{\mathbf{M}}_{2W}\mathbf{P}_{\mathbf{W}} &= \mathbf{P}_{\mathbf{W}}\mathbf{M}_2(\mathbf{M}_2\mathbf{W}\mathbf{M}_2)^{-}\mathbf{M}_2\mathbf{P}_{\mathbf{W}} \\
&= \mathbf{W}^{+1/2}\mathbf{P}_{\mathbf{W}^{1/2}\mathbf{M}_2}\mathbf{W}^{+1/2} \\
&= \mathbf{W}^{+1/2}(\mathbf{P}_{\mathbf{W}} - \mathbf{P}_{\mathbf{W}^{+1/2}\mathbf{X}_2})\mathbf{W}^{+1/2} \\
&= \mathbf{W}^{+} - \mathbf{W}^{+}\mathbf{X}_2(\mathbf{X}_2'\mathbf{W}^{+}\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{W}^{+},
\end{aligned}$$

and hence

$$\begin{aligned}
\mathbf{W}\dot{\mathbf{M}}_{2W}\mathbf{X}_1 &= \mathbf{W}[\mathbf{W}^{+} - \mathbf{W}^{+}\mathbf{X}_2(\mathbf{X}_2'\mathbf{W}^{+}\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{W}^{+}]\mathbf{X}_1 \\
&= [\mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{W}^{+}\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{W}^{+}]\mathbf{X}_1 \,.
\end{aligned}$$

which completes the proof.

Remark 6. Markiewicz & Puntanen [30, p. 11] mention that in claim (f) of Proposition 6.2 the matrix $\mathbf{W}$ can be replaced with $\mathbf{W}_1$ to obtain

$$\mathbf{W}_1\dot{\mathbf{M}}_2\mathbf{X}_1 = [\mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{W}_1^{+}\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{W}_1^{+}]\mathbf{X}_1 \,, \tag{51}$$

where $\dot{\mathbf{M}}_2 = \mathbf{M}_2(\mathbf{M}_2\mathbf{W}_1\mathbf{M}_2)^{-}\mathbf{M}_2$. However, (51) does not hold in general; it holds, for example, if $\mathscr{C}(\mathbf{X}_2) \subseteq \mathscr{C}(\mathbf{W}_1)$.

For completeness we state the following related result, due to [41, p. 140].

Proposition 6.3. *Consider the linear model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\beta, \mathbf{V}\}$ *and denote* $\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}'$, *where* $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$, *and let* $\mathbf{W}^{-}$ *be an arbitrary generalized inverse of* $\mathbf{W}$. *Then*

$$\begin{aligned}
\mathscr{C}(\mathbf{W}^{-}\mathbf{X}) \oplus \mathscr{C}(\mathbf{X})^{\perp} &= \mathbb{R}^n, \qquad \mathscr{C}(\mathbf{W}^{-}\mathbf{X})^{\perp} \oplus \mathscr{C}(\mathbf{X}) = \mathbb{R}^n, \\
\mathscr{C}[(\mathbf{W}^{-})'\mathbf{X}] \oplus \mathscr{C}(\mathbf{X})^{\perp} &= \mathbb{R}^n, \qquad \mathscr{C}[(\mathbf{W}^{-})'\mathbf{X}]^{\perp} \oplus \mathscr{C}(\mathbf{X}) = \mathbb{R}^n.
\end{aligned}$$

7 Linear Sufficiency

A linear statistic $\mathbf{Fy}$, where $\mathbf{F} \in \mathbb{R}^{f\times n}$, is called linearly sufficient for $\mathbf{X}\boldsymbol{\beta}$ under the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, if there exists a matrix $\mathbf{A} \in \mathbb{R}^{n\times f}$ such that $\mathbf{AFy}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$. Correspondingly, $\mathbf{Fy}$ is linearly sufficient for estimable $\mathbf{K}\boldsymbol{\beta}$, where $\mathbf{K} \in \mathbb{R}^{k\times p}$, if there exists a matrix $\mathbf{A} \in \mathbb{R}^{k\times f}$ such that $\mathbf{AFy}$ is the BLUE for $\mathbf{K}\boldsymbol{\beta}$.

The concept of linear sufficiency was essentially introduced in early 1980s by Baksalary & Kala [4, 3] and by Drygas [10]. [4] talked about "linear transformations preserving best linear unbiased estimators" and Drygas [10] introduced the term "linear sufficiency".

By definition, $\mathbf{Fy}$ is linearly sufficient for estimable $\mathbf{K}\boldsymbol{\beta}$ if and only if the equation

$$\mathbf{AF}(\mathbf{X} : \mathbf{VM}) = (\mathbf{K} : \mathbf{0})$$

has a solution for $\mathbf{A}$, which happens if and only if

$$\mathscr{C}\begin{pmatrix}\mathbf{K}'\\ \mathbf{0}\end{pmatrix} \subseteq \mathscr{C}\begin{pmatrix}\mathbf{X}'\mathbf{F}'\\ \mathbf{MVF}'\end{pmatrix}.$$

Sometimes we may use the notation $\mathbf{Fy} \in \mathcal{S}(\mathbf{K}\boldsymbol{\beta})$ to indicate that $\mathbf{Fy}$ is linearly sufficient for $\mathbf{K}\boldsymbol{\beta}$. Moreover, we can denote, symbolically,

$$\mathcal{S}(\mathbf{K}\boldsymbol{\beta}) = \{\mathbf{Fy} : \mathbf{AF}(\mathbf{X} : \mathbf{VM}) = (\mathbf{K} : \mathbf{0}) \text{ for some } \mathbf{A} \in \mathbb{R}^{k\times f}\}.$$

For the proofs of parts (a) and (b) of Proposition 7.1, see [4], and for (c), [3].

Proposition 7.1. *The statistic $\mathbf{Fy}$ is linearly sufficient for $\mathbf{X}\boldsymbol{\beta}$ under the linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ if and only if any of the following equivalent statements holds:*

(a) $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{WF}')$, *where* $\mathbf{W} \in \mathcal{W}_{\geq}$,

(b) $\mathrm{r}(\mathbf{X} : \mathbf{VF}') = \mathrm{r}(\mathbf{WF}')$, *where* $\mathbf{W} \in \mathcal{W}_{\geq}$.

Moreover, $\mathbf{Fy}$ is linearly sufficient for estimable $\mathbf{K}\boldsymbol{\beta}$ under $\mathscr{M}$ if and only if

(c) $\mathscr{C}[\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{K}'] \subseteq \mathscr{C}(\mathbf{WF}')$, *where* $\mathbf{W} \in \mathcal{W}_{\geq}$.

The crucial connection between the concept of linear sufficiency and the transformed model $\mathscr{T} = \{\mathbf{Fy}, \mathbf{FX}\boldsymbol{\beta}, \mathbf{FVF}'\}$ was was proved by Baksalary & Kala [4, 3]: if $\mathbf{K}\boldsymbol{\beta}$ is estimable under $\mathscr{M}$ and $\mathscr{T}$, then

$$\mathbf{Fy} \in \mathcal{S}(\mathbf{K}\boldsymbol{\beta}) \iff \mathrm{BLUE}(\mathbf{K}\boldsymbol{\beta} \mid \mathscr{M}) = \mathrm{BLUE}(\mathbf{K}\boldsymbol{\beta} \mid \mathscr{T}) \text{ w.p. } 1.$$

Thus we do not lose anything essential if we estimate $\mathbf{K}\boldsymbol{\beta}$ under the transformed model $\mathscr{T}$ instead of $\mathscr{M}$.

The next proposition characterizes when $\mathbf{Fy}$ is linearly sufficient for $\boldsymbol{\mu}_1$.

Proposition 7.2. *Let $\boldsymbol{\mu}_1 = \mathbf{X}_1\boldsymbol{\beta}_1$ be estimable under $\mathscr{M}_{12}$ and let $\mathbf{W} \in \mathcal{W}_{\geq}$. Then $\mathbf{Fy}$ is linearly sufficient for $\boldsymbol{\mu}_1$ under $\mathscr{M}_{12}$ if and only if any of the following equivalent conditions holds:*

(a) $\mathscr{C}(\mathbf{W}\dot{\mathbf{M}}_{2W}\mathbf{X}_1) \subseteq \mathscr{C}(\mathbf{WF}')$, *where* $\dot{\mathbf{M}}_{2W} = \mathbf{M}_2(\mathbf{M}_2\mathbf{WM}_2)^{-}\mathbf{M}_2$.

(b) $\mathscr{C}\big\{[\mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{W}^{+}\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{W}^{+}]\mathbf{X}_1\big\} \subseteq \mathscr{C}(\mathbf{WF}')$.

Let us prove Proposition 7.2 along the lines of [24, Sec. 3]. For a different proof, see [21, Th. 2]. One expression for the BLUE of $\boldsymbol{\mu}_1 = \mathbf{X}_1\boldsymbol{\beta}_1$, obtainable from $\mathscr{M}_{12\cdot 2}$, is

$$\mathbf{Ay} := \mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1)^-\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{y},$$

where $\dot{\mathbf{M}}_{2W} = \mathbf{M}_2(\mathbf{M}_2\mathbf{W}\mathbf{M}_2)^-\mathbf{M}_2$ and $\mathbf{W} \in \mathcal{W}_{\geq}(\mathscr{M}_{12})$. On the other hand, the BLUE of $\boldsymbol{\mu}_1$ can be written also as

$$\mathbf{By} := (\mathbf{X}_1 : \mathbf{0})(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\mathbf{y} = \mathbf{K}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\mathbf{y},$$

where $\mathbf{K} = (\mathbf{X}_1 : \mathbf{0}) \in \mathbb{R}^{n\times p}$ and $\mathbf{W} \in \mathcal{W}_{\geq}(\mathscr{M}_{12})$. By the consistency of the model $\mathscr{M}_{12}$ we have $\mathbf{Ay} = \mathbf{By}$ with probability 1, i.e., $\mathbf{AW} = \mathbf{BW}$, which can be transposed to give

$$\mathbf{W}\dot{\mathbf{M}}_{2W}\mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1)^-\mathbf{X}_1' = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{K}'.$$

In light of part (c) of Proposition 7.1, the claim (a) is confirmed by showing that

$$\mathscr{C}[\mathbf{W}\dot{\mathbf{M}}_2\mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1)^-\mathbf{X}_1'] = \mathscr{C}(\mathbf{W}\dot{\mathbf{M}}_{2W}\mathbf{X}_1)\,,$$

i.e., $\mathrm{r}\Big[\mathbf{W}\dot{\mathbf{M}}_{2W}\mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1)^-\mathbf{X}_1'\Big] = \mathrm{r}(\mathbf{W}\dot{\mathbf{M}}_{2W}\mathbf{X}_1)$, which follows from

$$\begin{aligned}\mathrm{r}(\mathbf{W}\dot{\mathbf{M}}_{2W}\mathbf{X}_1) &\geq \mathrm{r}\Big[\mathbf{W}\dot{\mathbf{M}}_{2W}\mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1)^-\mathbf{X}_1'\Big]\\ &\geq \mathrm{r}\Big[\mathbf{W}\dot{\mathbf{M}}_{2W}\mathbf{X}_1(\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{X}_1)^-\mathbf{X}_1'\dot{\mathbf{M}}_{2W}\mathbf{X}_1\Big]\\ &= \mathrm{r}(\mathbf{W}\dot{\mathbf{M}}_{2W}\mathbf{X}_1)\,.\end{aligned}$$

The claim (b) follows from part (f) of Proposition 6.2.

Remark 7. It is of interest to consider some particular properties related to the linear sufficiency condition (c) of Proposition 7.1:

$$\mathbf{Fy} \in \mathcal{S}(\mathbf{X}\boldsymbol{\beta}) \iff \mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{WF}')\,, \quad \text{where } \mathbf{W} \in \mathcal{W}_{\geq}(\mathscr{M}). \tag{52}$$

The matrix $\mathbf{W}$ in (52) belongs to the set $\mathcal{W}$ of (symmetric) nonnegative definite matrices. One could wonder whether the column space $\mathscr{C}(\mathbf{WF}')$ is unique once $\mathbf{F}$ is given, i.e., does it remain invariant for any choice of $\mathbf{W} \in \mathcal{W}_{\geq}$? Kala et al. [24, Ex. 1] provide a counterexample showing that this is not the case. Kala et al. [24, Sec. 4] also studied whether the column space $\mathscr{C}(\mathbf{WF}')$ is invariant for any choice of $\mathbf{W} \in \mathcal{W}(\mathscr{M})$ if $\mathbf{Fy} \in \mathcal{S}(\mathbf{X}\boldsymbol{\beta})$. The answer is positive, and moreover,

$$\mathscr{C}(\mathbf{WF}') = \mathscr{C}(\mathbf{X}) \oplus \mathscr{C}(\mathbf{MVF}') = \mathscr{C}(\mathbf{W}'\mathbf{F}')\,.$$

Kala et al. [24, Th. 4] were also wondering whether in (52) the set $\mathcal{W}_{\geq}$ can be replaced with the more general set $\mathcal{W}$. Interestingly, the answer is positive. As far as we know, in all linear sufficiency considerations appearing in literature, it is assumed that $\mathbf{W}$ is nonnegative definite. However, this is not necessary, and $\mathbf{W}$ can also be nonsymmetric. It may be mentioned, in passing, that the proof is parallel to that of [4, p. 914] who utilize the fact that $\mathbf{By}$ is a BLUE of $\mathbf{X}\boldsymbol{\beta}$ if and only if

$$\mathbf{BW} = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}', \quad \text{where } \mathbf{W} \in \mathcal{W}_{\geq}\,. \tag{53}$$

It is easy to confirm that in (53) the set $\mathcal{W}_{\geq}$ can be replaced with $\mathcal{W}$. Namely we know that $\mathbf{B} \in \{\mathbf{P}_{\boldsymbol{\mu}|\mathscr{M}}\}$ if and only if

$$\mathbf{B} = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^- + \mathbf{EQ}_{\mathbf{W}} \tag{54}$$

for some $\mathbf{E} \in \mathbb{R}^{n\times n}$. Postmultiplying (54) by $\mathbf{W}$ and using $\mathbf{X}'\mathbf{W}^-\mathbf{W} = \mathbf{X}'$ gives (53). On the other hand, if $\mathbf{B}$ satisfies (53) then $\mathbf{B}$ is necessarily of the form (54) for some $\mathbf{E}$ and thereby $\mathbf{B} \in \{\mathbf{P}_{\boldsymbol{\mu}|\mathscr{M}}\}$.

Remark 8. Baksalary & Kala (1981) [4, p. 914] write the following (in our notation):

(a) "If the condition $\mathscr{C}(\mathbf{X}) \subseteq \mathscr{C}(\mathbf{WF}')$, where $\mathbf{W} \in \mathcal{W}_{\geq}(\mathscr{M})$, is satisfied, then each BLUE of $\mathbf{X}\boldsymbol{\beta}$ in the transformed model $\mathscr{T}$ is also a BLUE of $\mathbf{X}\boldsymbol{\beta}$ in the original model $\mathscr{M}$, and vice versa."

It is the phrase *vice versa* that may cause some confusion as stated by [25, Sec. 4]. Let us discuss the meaning of the *vice versa* part along the lines of [16, Sec. 11.6].

Suppose that $\boldsymbol{\eta} = \mathbf{K}\boldsymbol{\beta}$ is estimable under the transformed model $\mathscr{T}$ (and thereby also under $\mathscr{M}$). Then $\mathbf{CFy}$ is the BLUE for $\mathbf{K}\boldsymbol{\beta}$ under $\mathscr{T}$ if and only if $\mathbf{C}$ belongs to the set $\{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{T}}\}$ which is defined as

$$\mathbf{C} \in \{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{T}}\} \iff \mathbf{C}(\mathbf{FX} : \mathbf{FVF}'\mathbf{Q}_{\mathbf{FX}}) = (\mathbf{K} : \mathbf{0})\,.$$

where $\mathbf{Q}_{\mathbf{FX}} = \mathbf{I}_f - \mathbf{P}_{\mathbf{FX}}$. The set of products $\mathbf{CF}$, where $\mathbf{C} \in \{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{T}}\}$, will be denoted as $\{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{T}}\mathbf{F}\}$. It means that each matrix $\mathbf{D} \in \{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{T}}\mathbf{F}\}$ applied to $\mathbf{y}$ provides the BLUE for $\mathbf{K}\boldsymbol{\beta}$ under the transformed model $\mathscr{T}$, i.e.,

$$\mathbf{D} \in \{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{T}}\mathbf{F}\} \iff \mathbf{D} = \mathbf{CF},\ \text{where } \mathbf{C} \in \{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{T}}\}.$$

Consider the multipliers of the response vector $\mathbf{y}$ when playing with the BLUEs under $\mathscr{M}$ and under $\mathscr{T}$; these sets are $\{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{M}}\}$ and $\{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{T}}\mathbf{F}\}$, respectively. Assume further that $\mathbf{Fy}$ is linearly sufficient for $\boldsymbol{\eta}$. Then the inclusion $\{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{T}}\mathbf{F}\} \subseteq \{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{M}}\}$ is straightforward but corresponding equality is more problematic. The following solution was given by [16, Prop. 11.17].

Proposition 7.3. *Let $\boldsymbol{\eta} = \mathbf{K}\boldsymbol{\beta}$ be estimable under $\mathscr{T}$, $\mathbf{W} \in \mathcal{W}_{\geq}(\mathscr{M})$ and assume that $\mathbf{Fy} \in \mathcal{S}(\boldsymbol{\eta})$. Then $\{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{M}}\} = \{\mathbf{P}_{\boldsymbol{\eta}|\mathscr{T}}\mathbf{F}\}$ holds if and only if*

$$\mathbf{Q}_{\mathbf{W}} = \mathbf{Q}_{\mathbf{W}}\mathbf{P}_{\mathbf{F}'}\,, \quad \textit{i.e.,} \quad \mathscr{C}(\mathbf{W})^{\perp} \subseteq \mathscr{C}(\mathbf{F}')\,. \tag{55}$$

In other words, under the linear sufficiency and condition (55)*, each representation of the* BLUE *of* $\mathbf{K}\boldsymbol{\beta}$ *under* $\mathscr{M}$ *is a representation of the* BLUE *under* $\mathscr{T}$ *and vice versa.*

8 Equality of the BLUEs Under Two Models

Let us consider two linear models, $\mathscr{A} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_a\}$ and $\mathscr{B} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_b\}$, having different covariance matrices. Let $\mathbf{W}_a \in \mathcal{W}_{\geq}(\mathscr{A})$ so that for some $\mathbf{U}$

$$\mathbf{W}_a = \mathbf{V}_a + \mathbf{XUU}'\mathbf{X}', \quad \text{where} \quad \mathscr{C}(\mathbf{W}_a) = \mathscr{C}(\mathbf{X} : \mathbf{V}_a)\,.$$

Then one representation for the BLUE of $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{A}$ is

$$\mathbf{P}_{\mathbf{X};\mathbf{W}_a^+}\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{W}_a^+\mathbf{X})^-\mathbf{X}'\mathbf{W}_a^+\mathbf{y}.$$

We can now ask whether $\mathbf{P}_{\mathbf{X};\mathbf{W}_a^+}\mathbf{y}$ continues to be BLUE under $\mathscr{B}$. This happens if and only if

$$\mathbf{X}(\mathbf{X}'\mathbf{W}_a^+\mathbf{X})^-\mathbf{X}'\mathbf{W}_a^+(\mathbf{X} : \mathbf{V}_b\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,,$$

which is obviously equivalent to $\mathbf{X}'\mathbf{W}_a^+\mathbf{V}_b\mathbf{M} = \mathbf{0}$. Further equivalent conditions are given in Proposition 8.1 below which appears in Mitra & Moore (1973) [33, Th. 2.1, Th. 2.2, Note 1]. Some parts they did not prove in details, giving only hints. For a complete proof, see [18, Th. 1].

Proposition 8.1. *Using the earlier notation,* $\mathbf{P}_{\mathbf{X};\mathbf{W}_a^+}\mathbf{y}$ *is the* BLUE *for* $\mathbf{X}\boldsymbol{\beta}$ *also under* $\mathscr{B}$ *if and only if any of the following equivalent conditions holds:*

(a) $\mathbf{X}'\mathbf{W}_a^+\mathbf{V}_b\mathbf{M} = \mathbf{0}$, (b) $\mathscr{C}(\mathbf{V}_b\mathbf{M}) \subseteq \mathscr{C}(\mathbf{W}_a^+\mathbf{X})^{\perp} =: \mathscr{C}(\mathbf{Z})$,

(c) $\mathbf{V}_b = \mathbf{X}\mathbf{R}\mathbf{X}' + \mathbf{Z}\mathbf{S}\mathbf{Z}'$ *for some* $\mathbf{R}$ *and* $\mathbf{S}$*, and* $\mathbf{Z} \in \{(\mathbf{W}_a^+\mathbf{X})^\perp\}$,

(d) $\mathbf{P}_{\mathbf{X};\mathbf{W}_a^+}\mathbf{V}_b$ *is symmetric,*

(e) $\mathscr{C}(\mathbf{W}_a^+\mathbf{X})$ *is spanned by a set of* r *proper eigenvectors of* $\mathbf{V}_b$ *with respect to* $\mathbf{W}_a$*;* $r = \mathrm{r}(\mathbf{X})$,

(f) $\mathscr{C}(\mathbf{X})$ *is spanned by a set of* r *eigenvectors of* $\mathbf{V}_b\mathbf{W}_a^+$.

In Proposition 8.1 we utilize the concept of *proper eigenvectors* following Rao & Mitra (1971) [41, Sec. 6.3]; see also [34], and [42]. To have a brief look at these concepts, let $\mathbf{A}$ and $\mathbf{B}$ be two symmetric $n \times n$ matrices of which $\mathbf{B}$ is nonnegative definite. Let $\lambda \in \mathbb{R}$ be a scalar and $\mathbf{u}$ a vector such that

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{B}\mathbf{u}, \quad \mathbf{B}\mathbf{u} \neq \mathbf{0}\,.$$

Rao & Mitra [41, Sec. 6.3] call λ a proper eigenvalue and $\mathbf{u}$ a proper eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$, or shortly, $(\lambda, \mathbf{u})$ is a proper eigenpair for $(\mathbf{A}, \mathbf{B})$. If $\mathbf{B}$ is singular, there may exist a vector $\mathbf{u} \neq \mathbf{0}$ such that $\mathbf{A}\mathbf{u} = \mathbf{B}\mathbf{u} = \mathbf{0}$, in which case

$$(\mathbf{A} - \lambda\mathbf{B})\mathbf{u} = \mathbf{0}$$

is satisfied with arbitrary λ. Such a vector $\mathbf{u} \in \mathbb{R}^n$ is called an improper eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$. The space of improper eigenvectors is precisely $\mathscr{N}(\mathbf{A}) \cap \mathscr{N}(\mathbf{B}) = \mathscr{C}(\mathbf{A} : \mathbf{B})^\perp$.

What about if we request that *every* representation of BLUE of $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ under $\mathscr{A}$ continues to be BLUE under $\mathscr{B}$, or shortly

$$\{\mathrm{BLUE}(\boldsymbol{\mu} \mid \mathscr{A})\} \subseteq \{\mathrm{BLUE}(\boldsymbol{\mu} \mid \mathscr{B})\}\,, \quad \text{i.e.,} \quad \{\mathbf{P}_{\boldsymbol{\mu}|\mathscr{A}}\} \subseteq \{\mathbf{P}_{\boldsymbol{\mu}|\mathscr{B}}\}\,. \tag{56}$$

As an arbitrary member of $\{\mathbf{P}_{\boldsymbol{\mu}|\mathscr{A}}\}$ can be expressed as

$$\mathbf{X}(\mathbf{X}'\mathbf{W}_a^+\mathbf{X})^-\mathbf{X}'\mathbf{W}_a^+ + \mathbf{E}\mathbf{Q}_{\mathbf{W}_a}\,, \quad \text{where } \mathbf{E} \in \mathbb{R}^{n\times n} \text{ is free to vary,}$$

we conclude that (56) holds if and only if

$$[\mathbf{X}(\mathbf{X}'\mathbf{W}_a^+\mathbf{X})^-\mathbf{X}'\mathbf{W}_a^+ + \mathbf{E}\mathbf{Q}_{\mathbf{W}_a}](\mathbf{X} : \mathbf{V}_b\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,. \tag{57}$$

It is straightforward to conclude that (57) holds for any $\mathbf{E}$ if and only if

$$\mathscr{C}(\mathbf{V}_b\mathbf{M}) \subseteq \mathscr{C}(\mathbf{V}_a\mathbf{M})\,.$$

For the conditions like (56) see, e.g., [33] and [15].

9 Further Remarks

In this paper we have reviewed the properties of matrix $\mathbf{W}$ belonging to the class

$$\mathcal{W}(\mathscr{M}) = \left\{\mathbf{W} \in \mathbb{R}^{n\times n} : \mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{T}\mathbf{X}',\ \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})\right\},$$

where $\mathbf{T}$ can be any $p \times p$ matrix as long as $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$ is satisfied and $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. Corresponding considerations can be done in other models, like linear model with *new observations*, which we will denote as $\mathscr{M}_*$. The *mixed* linear model is a special case of of the model with new observations. In this article we skip the mixed model but will briefly go through the linear model with new observations.

We can extend the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ by considering a $q \times 1$ random vector $\mathbf{y}_*$, which is an unobservable random vector containing new future observations. These new observations are assumed to be generated from

$$\mathbf{y}_* = \mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_* = \boldsymbol{\mu}_* + \boldsymbol{\varepsilon}_*\,,$$

where $\mathbf{X}_*$ is a known $q \times p$ matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the same vector of fixed but unknown parameters as in $\mathscr{M}$, and $\boldsymbol{\varepsilon}_*$ is a q-dimensional random error vector with $\mathrm{E}(\boldsymbol{\varepsilon}_*) = \mathbf{0}$. The covariance matrix of $\mathbf{y}_*$ and the cross-covariance matrix between $\mathbf{y}$ and $\mathbf{y}_*$ are assumed to be known and thus we have

$$\mathrm{E}\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_*\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}_*\end{pmatrix} = \begin{pmatrix}\mathbf{X}\\ \mathbf{X}_*\end{pmatrix}\boldsymbol{\beta}\,, \quad \mathrm{cov}\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_*\end{pmatrix} = \begin{pmatrix}\mathbf{V} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}.$$

This setup can be denoted shortly as

$$\mathscr{M}_* = \left\{\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_*\end{pmatrix}, \begin{pmatrix}\mathbf{X}\\ \mathbf{X}_*\end{pmatrix}\boldsymbol{\beta}, \begin{pmatrix}\mathbf{V} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}\right\}.$$

Our aim is to predict the unobservable $\mathbf{y}_*$ on the basis of the observable $\mathbf{y}$.

The random vector $\mathbf{A}\mathbf{y}$ is a linear unbiased predictor (LUP) of $\mathbf{y}_*$ if $\mathrm{E}(\mathbf{y}_* - \mathbf{A}\mathbf{y}) = \mathbf{0}$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Such a matrix $\mathbf{A} \in \mathbb{R}^{q\times n}$ exists if and only if $\mathscr{C}(\mathbf{X}_*') \subseteq \mathscr{C}(\mathbf{X}')$, i.e., $\mathbf{X}_*\boldsymbol{\beta}$ is estimable under $\mathscr{M}$ and then we say that $\mathbf{y}_*$ is predictable under $\mathscr{M}_*$. Now a LUP $\mathbf{A}\mathbf{y}$ is the best linear unbiased predictor, BLUP, for $\mathbf{y}_*$, if the covariance matrix of the prediction error, subject to the unbiasedness of the prediction, is minimized:

$$\mathrm{cov}(\mathbf{y}_* - \mathbf{A}\mathbf{y}) \le_{\mathsf{L}} \mathrm{cov}(\mathbf{y}_* - \mathbf{A}_\#\mathbf{y}) \quad \text{for all } \mathbf{A}_\# : \mathbf{A}_\#\mathbf{X} = \mathbf{X}_*\,.$$

It appears that the linear predictor $\mathbf{A}\mathbf{y}$ is the BLUP for $\mathbf{y}_*$ if and only if $\mathbf{A} \in \mathbb{R}^{q\times n}$ satisfies the the so-called fundamental BLUP equation

$$\mathbf{A}(\mathbf{X} : \mathbf{V}\mathbf{X}^\perp) = (\mathbf{X}_* : \mathbf{V}_{21}\mathbf{X}^\perp)\,. \tag{58}$$

For (58), see, e.g., [9, p. 294], and [20, p. 1015]. Corresponding to (58), $\mathbf{B}\mathbf{y}$ is the $\mathrm{BLUP}(\boldsymbol{\varepsilon}_*)$ whenever

$$\mathbf{B}(\mathbf{X} : \mathbf{V}\mathbf{X}^\perp) = (\mathbf{0} : \mathbf{V}_{21}\mathbf{X}^\perp)\,.$$

Now the $\mathrm{BLUP}(\mathbf{y}_*)$ under $\mathscr{M}_*$, see, e.g., [17, Sec. 2] and [19, Sec. 4], can be written for example as

$$\begin{aligned}\mathrm{BLUP}(\mathbf{y}_*) &= \mathrm{BLUE}(\boldsymbol{\mu}_*) + \mathrm{BLUP}(\boldsymbol{\varepsilon}_*)\\ &= \mathbf{X}_*\mathbf{B}\mathbf{y} + \mathbf{V}_{21}\mathbf{V}^-(\mathbf{I}_n - \mathbf{G})\mathbf{y}\\ &= \mathbf{X}_*\mathbf{B}\mathbf{y} + \mathbf{V}_{21}\mathbf{W}^-(\mathbf{I}_n - \mathbf{G})\mathbf{y}\\ &= \mathbf{X}_*\mathbf{B}\mathbf{y} + \mathbf{V}_{21}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{y},\end{aligned}$$

where $\mathbf{B} = (\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-$ and $\mathbf{G} = \mathbf{X}\mathbf{B} = \mathbf{P}_{\mathbf{X};\mathbf{W}^-}$ and $\mathbf{W} \in \mathcal{W}(\mathscr{M})$. In particular, if $\mathbf{V}$ is positive definite and $\mathrm{r}(\mathbf{X}) = p$, we obtain

$$\begin{aligned}\mathrm{BLUP}(\mathbf{y}_*) &= \mathbf{X}_*\tilde{\boldsymbol{\beta}} + \mathbf{V}_{21}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\\ &= \mathbf{X}_*\tilde{\boldsymbol{\beta}} + \mathbf{V}_{21}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{y},\end{aligned}$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.

Acknowledgements

The authors are grateful to Professor Bikas Kumar Sinha for the invitation to prepare this article for possible publication in the *Gujarat Journal of Statistical and Data Sciences*. Thanks go also to the anonymous referee for constructive comments.

References

[1] A C Aitken. "On least squares and linear combination of observations." In: *Proc. Roy. Soc. Edinburgh Sect. A* 55 (1935), pp. 42–49. DOI: 10.1017/S0370164600014346.

[2] J K Baksalary. "An elementary development of the equation characterizing best linear unbiased estimators." In: *Linear Algebra Appl.* 388 (2004), pp. 3–6. DOI: 10.1016/S0024-3795(03)00396-3.

[3] J K Baksalary and R Kala. "Linear sufficiency with respect to a given vector of parametric functions." In: *J. Stat. Plann. Inference* 14 (1986), pp. 331–338. DOI: 10.1016/0378-3758(86)90171-0.

[4] J K Baksalary and R Kala. "Linear transformations preserving best linear unbiased estimators in a general Gauss–Markoff model." In: *Ann. Stat.* 9 (1981), pp. 913–916. DOI: 10.1214/aos/1176345533.

[5] J K Baksalary and T Mathew. "Rank invariance criterion and its application to the unified theory of least squares." In: *Linear Algebra Appl.* 127 (1990), pp. 393–401. DOI: 10.1016/0024-3795(90)90352-D.

[6] J K Baksalary and S Puntanen. "Weighted-least-squares estimation in the general Gauss–Markov model." In: *Statistical Data Analysis and Inference*. Ed. by Y Dodge. Amsterdam: North-Holland, 1989, pp. 355–368. DOI: 10.1016/B978-0-444-88029-1.50038-1.

[7] J K Baksalary, S Puntanen, and G P H Styan. "A property of the dispersion matrix of the best linear unbiased estimator in the general Gauss–Markov model." In: *Sankhyā Ser. A* 52 (1990), pp. 279–296. URL: http://www.jstor.org/stable/25050797.

[8] J K Baksalary, C R Rao, and A Markiewicz. "A study of the influence of the 'natural restrictions' on estimation problems in the singular Gauss–Markov model." In: *J. Stat. Plann. Inference* 31 (1992), pp. 335–351. DOI: 10.1016/0378-3758(92)90141-E.

[9] R Christensen. *Plane Answers to Complex Questions: the Theory of Linear Models*. 4th ed. New York: Springer, 2011. DOI: 10.1007/978-1-4419-9816-3.

[10] H Drygas. "Sufficiency and completeness in the general Gauss–Markov model." In: *Sankhyā Ser. A* 45 (1983), pp. 88–98. URL: http://www.jstor.org/stable/25050416.

[11] H Drygas. *The Coordinate-Free Approach to Gauss–Markov Estimation*. Berlin: Springer, 1970. DOI: 10.1007/978-3-642-65148-9.

[12] J Groß. "Special generalized inverse matrices connected with the theory of unified least squares." In: *Linear Algebra Appl.* 264 (1997), pp. 325–327. DOI: 10.1016/S0024-3795(95)07778-2.

[13] J. Groß and S Puntanen. "Estimation under a general partitioned linear model." In: *Linear Algebra Appl.* 321 (2000), pp. 131–144. DOI: 10.1016/S0024-3795(00)00028-8.

[14] D A Harville. *Matrix Algebra From a Statistician's Perspective*. New York: Springer, 1997. DOI: 10.1007/b98818.

[15] S J Haslett and S Puntanen. "A review of conditions under which BLUEs and/or BLUPs in one linear mixed model are also BLUEs and/or BLUPs in another." In: *Calcutta Stat. Assoc. Bull.* 65 (2013), pp. 25–41. DOI: 10.1177/0008068320130103.

[16] S J Haslett et al. "A review of the linear sufficiency and linear prediction sufficiency in the linear model with new observations." In: *Multivariate, Multilinear and Mixed Linear Models*. Ed. by D von Rosen K Filipiak A Markiewicz. Cham: Springer, 2021, pp. 265–318. DOI: 10.1007/978-3-030-75494-5_11.

[17] S J Haslett et al. "Equalities between OLSE, BLUE and BLUP in the linear model." In: *Stat. Pap.* 55 (2014), pp. 543–56. DOI: 10.1007/s00362-013-0500-7.

[18] J Hauke, A Markiewicz, and S Puntanen. "Revisiting the BLUE in a linear model via proper eigenvectors." In: *Combinatorial Matrix Theory and Generalized Inverses of Matrices*. Ed. by R B Bapat et al. India: Springer, 2013, pp. 73–83. DOI: 10.1007/978-81-322-1053-5_7.

[19] J Isotalo, A Markiewicz, and S Puntanen. "Some properties of linear prediction sufficiency in the linear model." In: *Trends and Perspectives in Linear Statistical Inference: LinStat, Istanbul, 2016*. Ed. by D von Rosen M Tez. Cham: Springer, 2018, pp. 111–129. DOI: 10.1007/978-3-319-73241-1_8.

[20] J Isotalo and S Puntanen. "Linear prediction sufficiency for new observations in the general Gauss–Markov model." In: *Commun. Stat. Theory Methods* 35 (2006), pp. 1011–1023. DOI: 10.1080/03610920600672146.

[21] J Isotalo and S Puntanen. "Linear sufficiency and completeness in the partitioned linear model." In: *Acta Comment Univ. Tartu Math.* 10 (2006), pp. 53–67. DOI: 10.12697/ACUTM.2006.10.05.

[22] J Isotalo, S Puntanen, and G P H Styan. "A useful matrix decomposition and its statistical applications in linear regression." In: *Commun. Stat. Theory Methods* 37 (2008), pp. 1436–1457. DOI: 10.1080/03610920701666328.

[23] J Isotalo, S Puntanen, and G P H Styan. "The BLUE's covariance matrix revisited: a review." In: *J. Stat. Plann. Inference* 138 (2008), pp. 2722–2737. DOI: 10.1016/j.jspi.2008.03.029.

[24] R Kala, A Markiewicz, and S Puntanen. "Some further remarks on the linear sufficiency in the linear model." In: *Applied and Computational Matrix Analysis: MatTriad, Coimbra, Portugal, 2015*. Ed. by N Bebiano. Cham: Springer, 2017, pp. 275–294. DOI: 10.1007/978-3-319-49984-0_19.

[25] R Kala, S Puntanen, and Y Tian. "Some notes on linear sufficiency." In: *Stat. Pap* 58 (2017), pp. 1–17. DOI: 10.1007/s00362-015-0682-2.

[26] L R LaMotte. "A direct derivation of the REML likelihood function." In: *Stat. Pap.* 48 (2007), pp. 321–327. DOI: 10.1007/s00362-006-0335-6.

[27] E P Liski, S Puntanen, and S G Wang. "Bounds for the trace of the difference of the covariance matrices of the OLSE and BLUE." In: *Linear Algebra Appl.* 176 (1992), pp. 121–130. DOI: 10.1016/0024-3795(92)90214-U.

[28] A Markiewicz. "On dependence structures preserving optimality." In: *Statist. Probab. Lett.* 53 (2001), pp. 415–419. DOI: 10.1016/S0167-7152(01)00064-5.

[29] A Markiewicz and S Puntanen. "All about the $\perp$ with its applications in the linear statistical models." In: *Open Mathematics* 13 (2015), pp. 33–50. DOI: 10.1515/math-2015-0005.

[30] A Markiewicz and S Puntanen. "Further properties of the linear sufficiency in the partitioned linear model." In: *Matrices, Statistics and Big Data: Selected Contributions from IWMS 2016*. Ed. by S Puntanen S E Ahmed F Carvalho. Cham: Springer, 2019, pp. 1–22. DOI: 10.1007/978-3-030-17519-1_1.

[31] G Marsaglia and G P H Styan. "Equalities and inequalities for ranks of matrices." In: *Linear Multilinear Algebra* 2 (1974), pp. 269–292. DOI: 10.1080/03081087408817070.

[32] S K Mitra. "Unified least squares approach to linear estimation in a general Gauss–Markov model." In: *SIAM J. Appl. Math.* 25 (1973), pp. 671–680. DOI: 10.1137/0125065.

[33] S K Mitra and B J Moore. "Gauss–Markov estimation with an incorrect dispersion matrix." In: *Sankhyā Ser. A* 35 (1973), pp. 139–152. URL: https://www.jstor.org/stable/25049864.

[34] S K Mitra and C R Rao. "Some results in estimation and tests of linear hypotheses under the Gauss–Markoff model." In: *Sankhyā Ser. A* 30 (1968), pp. 281–290. URL: https://www.jstor.org/stable/25041359.

[35] S Puntanen, G P H Styan, and J Isotalo. *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*. Heidelberg: Springer, 2011. DOI: 10.1007/978-3-642-10473-2.

[36] C R Rao. "Least squares theory for possibly singular models." In: *Canad. J. Statist.* 6 (1978), pp. 19–23. DOI: 10.2307/3314821.

[37] C R Rao. "Projectors, generalized inverses and the BLUE's." In: *J. Roy. Statist. Soc. Ser. B* 36 (1974), pp. 442–448. DOI: 10.1111/j.2517-6161.1974.tb01019.x.

[38] C R Rao. "Representations of best linear estimators in the Gauss–Markoff model with a singular dispersion matrix." In: *J. Multivariate Anal.* 3 (1973), pp. 276–292. DOI: 10.1016/0047-259X(73)90042-0.

[39] C R Rao. "The inefficiency of least squares: extensions of the Kantorovich inequality." In: *Linear Algebra Appl.* 70 (1985), pp. 249–255. DOI: 10.1016/0024-3795(85)90056-4.

[40] C R Rao. "Unified theory of linear estimation." In: *Sankhyā Ser. A* 33 (1971), pp. 371–394. URL: https://www.jstor.org/stable/25049750.

[41] C R Rao and S K Mitra. *Generalized Inverse of Matrices and Its Applications*. New York: Wiley, 1971.

[42] A J Scott and G P H Styan. "On a separation theorem for generalized eigenvalues and a problem in the analysis of sample surveys." In: *Linear Algebra Appl.* 70 (1985), pp. 209–224. DOI: 10.1016/0024-3795(85)90053-9.

[43] S G Wang and E P Liski. "Some matrix versions of Cauchy–Schwartz inequality: a statistical approach." In: *Mathematical Statistics and Applied Probability* 13 (1998), pp. 43–50.

[44] G Zyskind and F B Martin. "On best linear estimation and general Gauss–Markov theorem in linear models with arbitrary nonnegative covariance structure." In: *SIAM J. Appl. Math.* 17 (1969), pp. 1190–1202. DOI: 10.1137/0117110.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Agreement Between Raters for Hierarchical Designs with Repeated Measurements

Hairong Shi[1], Soumya Sahu[2], Domenic J. Reda[3], Bikas K. Sinha[4] and Dulal K. Bhaumik[5]

1. *CSPCC, Hines VA Hospital, Illinois, United States*
2. *Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Illinois, United States*
3. *CSPCC, Hines VA Hospital, Illinois, United States Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Illinois, United States*
4. *Indian Statistical Institute, Kolkata, India*
5. *Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Illinois, United States, CSPCC, Hines VA Hospital, Illinois, United States, Department of Psychiatry, University of Illinois at Chicago, Illinois, United States*

Received: 03 Jan 2022 / Revised: 12 Jun 2022 / Accepted: 08 Jul 2022

Abstract

Recent research towards measuring agreement among different methods or raters has received wide attention. The concordance correlation coefficient (CCC) has been used to assess agreement between raters or methods of measurements on continuous scales. However, CCC does not consider repeated measurement structures for multiple raters which may arise, e.g. longitudinal studies in clinical trials or bioassay data with sub-samples. For those designs, random variables underlying the measurements are not independent nor identically distributed. In this article we generalize the concept of CCC for repeated correlated measurements for multiple raters. For comparing agreement between raters, we have performed analyses based on two-level and three-level models and found that a three-level model fits better for data we collected in the Glucosamine/Chondroitin Arthritis Intervention Trial. We introduce the concept of intra-rater agreement in the context of CCC when raters are asked to rate each subject multiple times. Our approach represents an attempt in evaluating the CCC for complex data structures involving multiple levels.

Keywords: Concordance Correlation Coefficient (CCC), Three-level Longitudinal Linear Mixed-effect Model, Agreement

1 Introduction

Medical research involving taking measurements from images requires precision and the ability to replicate those measurements. Inter- and intra-observer variability are common and can introduce both measurement error and bias. Thus, a common quality control measure is necessary when multiple raters take measurements independently. In addition, measurements which historically were taken by humans now can often be obtained using computer programs and in many instances computer measurements have replaced corresponding human measurements. Statistical methods to measure an agreement among multiple raters have been in common use for decades. Thus, it is possible to assess both inter and intra-rater reliability, or agreement, for measurements taken from a set of independent measurements, such as radiographs of knee joints in individuals participating in a clinical trial. Thus, an agreement statistic for multiple raters, (e.g. human and computer algorithms) is needed to compare their performances.

Cohen's Kappa statistic [22] and weighted Kappa statistic[23] are the most popular indices for measuring agreement when responses are nominal. The weighted Kappa statistic has been proposed by Landis and Koch[37] and it is appropriate for assessing agreement when the categories of response

© 2022 Author(s). (https://www.thegsa.in/).

are ordinal. Several authors have provided guidelines for the interpretation of the Kappa statistic, vide, for example, Landis and Koch (1977), Bland and Altman (1986)[11] and Kraemer et al. (2002)[18]. A comprehensive review paper is also worth reporting here (Banerjee et al. (1999)[2]). Recently, some studies have been undertaken to critically examine certain aspects of Cohen's Kappa. These relate to its attaining the negatively extreme value and its standardization [Sinha et al (2006)[65]]. Extensions have also been made to allow for more than two raters, more than two possible ratings, ordinal data and continuous data. In addition, many other applications of the Kappa statistic in a variety of different contexts can be found in the literature and also in a book by Eye and Mun (2010)[69]. Another book dealing with both categorical and continuous measurements for multiple raters and multiple ratings is by Shoukri (2004)[64].

Despite its wide use in psychological, behavioral and social science studies, Cohen's Kappa has severe draw backs and limitations. It works perfectly well when all subjects are rated by two raters with 100 percent 'matching' between two categories. However, it performs poorly in inter-rater agreement at the presence of marginal heterogeneity. On the other hand, for 100% 'dismatching' i.e., 100% matching along the 'off-diagonals' Kappa does not always give the value -1, even though it should. Sinha et. al. (2006)[65] fixed the formula of Kappa in this scenario. Another limitation of using Cohen's Kappa is that it cannot be generalized for multilevel hierarchical designs with continuous outcomes.

Lin (1989)[38] introduced CCC for measuring agreement when outcome measures are continuous. A weighted CCC was proposed by Chinchilli et al.(1996)[17] for repeated measurement designs and a generalized CCC for continuous and categorical data was introduced by King and Chinchilli (2001)[32]. Lin (2000)[47] also introduced the concept of total deviation index (TDI) for measuring individual agreement with applications in laboratory performance and bio-equivalence. Furthermore, Lin et. al. (2002)[45] proposed methods for checking the agreement in terms of coverage probability (CP) when the two measurements are quantitative in nature. The approach proposed and studied in Lin et. al. (2002) has been extended by Hedayat et. al. (2009)[27] for the case of multiple raters.

However, these methods are generally applied to low dimensional data structures. In clinical trials, repeated measurements are taken on individuals over time. It is possible to measure agreement at a single time point, but there is a lack of statistical methods that can take advantage of the longitudinal structure of clinical trials. Measurements following complex data structures typically suffer from measurement error providing reduced statistical power. A method that considers the longitudinal nature of a clinical trial has the potential to provide additional insights into the data structure and answers to questions such as: (i) does inter-rater reliability change over time? (ii) is there an evidence for biased measurements for one rater compared with other raters? (iii) is the bias or reliability influenced by the magnitude of the measurements or characteristics of the individual from whom the measurements are obtained.

In this paper, our goal is to generalize the CCC measure between two raters proposed by Lin (1989) for repeated measurements over time. In doing so, we would like to maintain the definition proposed by Lin (1989). We propose a linear mixed-effect model to address the correlation due to repeated measurements over time which is just an extension of the bivariate normal set up used in Lin (1989). While comparing the CCC measure of our model with the existing works on generalized CCC, we found that the one proposed by Carrasco et al. (2009)[14] lacks the 'mean difference' term in the denominator, thus, it is different from Lin's (1989)[38] CCC. Apart from that, we also noticed that Tsai and Lin (2018)[68] have proposed a CCC measure which is an extension of the CCC measure introduced by King et al. (2007)[33], conceptually different from Lin's CCC (1989)[38]. In that way, we can claim the novelty of our approach, neither Carrasco et al. (2009)[14] nor King et al.(2007)[33]

can be considered as a generalization of Lin's (1989)[38] work.
The fundamental contribution of this article is in generalizing the concept of CCC for multilevel hierarchical designs when subjects are rated repeatedly over time, and the same rater is asked to rate subjects at multiple times at each measuring time point. We address the complexity of intra- and inter-rater agreement for correlated continuous outcomes when multiple raters are involved. Our purpose in this research study is to propose an 'agreement index' in situations dealing with multiple raters and multiple measurements obtained from longitudinal data. In Section 2, we provide our motivational example and describe the complexity of the underlying trial design involving multiple raters in a longitudinal design where repeated measurements are obtained at every measuring time point. In Section 3, we develop the mathematical foundation for computing agreement between raters for a hierarchical design with repeated measures, and discuss estimation of the underlying model parameters by several methods. In Section 4, we provide a limited simulation study to evaluate the performance of our proposed methodology. In Section 5, we analyze the data of the motivational example and provide relevant interesting results. Finally, in Section 5 we discuss flexibility and some limitations of our model, and briefly say how the inferential issues of our study will be addressed in future.

2 Motivational Example

Osteoarthritis (OA) is the most common form of all arthritides, afflicting approximately 27 million persons in the United States. Based on 2010-2012 data from the National Health Interview Survey (NHIS), an estimated 52.5 million adults (22.7%) have self-reported doctor-diagnosed arthritis. Weight bearing large joints, especially the knees, are often involved in OA. People with osteoarthritis often report pain limiting daily activites. In addition, radiographs of the joints generally show changes in the joint such as loss of cartilage, which can be observed as a narrowing of the joint space which increases over time.
The Glucosamine/Chondroitin Arthritis Intervention Trial, (GAIT) N01-AR-9-2236, is an NIH funded, double-blind, five-arm randomized clinical trial, which was designed to determine whether glucosamine, chondroitin sulfate and/or the combination of glucosamine and chondroitin sulfate are more effective than placebo and whether the combination is more effective than glucosamine or chondroitin sulfate alone in the treatment of knee pain associated with osteoarthritis (OA) of the knee[21, 57].
A substudy was also conducted to further examine the effects of these treatments on changes in joint space width over time. Measurements from each radiograph were obtained by two trained manual readers and a computer measurement system. Each radiograph of Joint Space Width (JSW) measurements was evaluated for meeting image quality inclusion criteria for the structural study. The images were blinded and coded at the Hines Cooperative Studies Program (CSP) Coordinating Center as to patient name, participating clinic, treatment group and date the X-ray was taken. The X-rays were read by physician investigators without knowledge of the patient name, participating clinic, treatment, and date of X-ray. These blinded images were read by physician investigators in matched pre-post pairs, but read in a randomly assigned order. Each X-ray was interpreted independently by two investigators: a rheumatologist with extensive experience in clinical investigation and previous experience in radiographic interpretation of clinical trials, and a musculoskeletal radiologist. Each reader reviewed plain hard copy radiographs and measured the JSW with calipers. An additional non-technical rater used the computer program Mdisplay by Buckland-Wright (1994)[12, 40, 54, 24] to measure JSW on digitized images of each film. Mdisplay is a semi-automated program requiring a user to mark the endpoints of the medial tibial and femoralcondyles and then an edge-finding algorithm determines the joint space borders. To determine whether manual measurement of JSW on plain

radiographs is equally as reproducible as computer generated measurements of digitalized radiographs, it is necessary to perform statistical evaluation of the agreement. Ideally, new measurements would be compared with the true value, known as the gold standard; however, in the case of measuring JSW - as is often the case - the gold standard is unknown.

3 Hierarchical Mixed-Effects Model and Estimation of Concordance Correlation Coefficient

Statistical measures for an agreement remain inadequate when hierarchical designs are used to evaluate subjects by multiple raters. To lay down the foundation, consider a study, where multiple raters are rating (e.g. reading x-ray plates) every subject over time. After a time gap, raters are asked again to rate the same x-ray plate that he/she rated before. Thus each rater is rating every subject multiple times at each measuring time point. The idea of this design is displayed in Figure 1, where replicates (level 1) are nested within x-ray plates (level 2), and plates are nested within subjects (level 3).

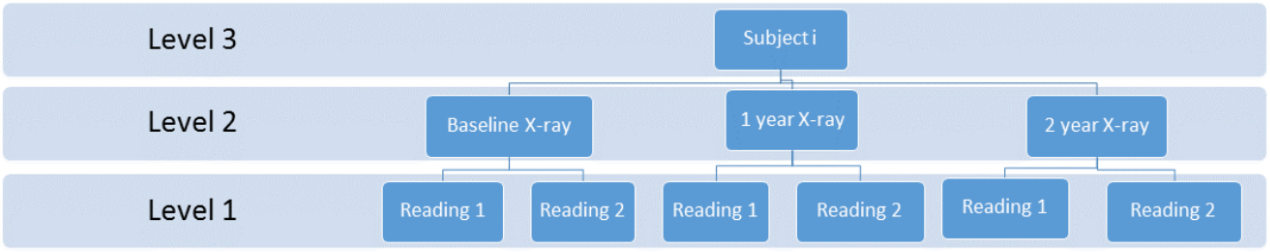


Figure 1: This figure displays a three level hierarchical design with multiple readings at each time for a subject by a rater

A design of this type provides us an opportunity to measure both within and between raters agreement. Denote two raters by X and W. The kth measure (or replicate) obtained from rater X for subject i, at jth time point is denoted by x_{ijk}, a similar measure from rater W is denoted by w_{ijk}, where $i = 1, 2, \cdots, N$, i.e. there are N study participants, $j(i) = 1, 2, \cdots, T_i$, means that the ith subject is measured in a total of T_i time points, and $k_{ij} = 1, 2, \cdots, K_{ij}$ means that the ith subject at the jth time point is measured k_{ij} times. For this study, we propose the following three level mixed-effects models for X and W respectively [30].

$$X_{ijk} = P_{X1i}\beta_X + P_{X2i}V_{Xij} + P_{X3i}U_{Xi} + E_{X_{ijk}}, \tag{1}$$

$$W_{ijk} = P_{W1i}\beta_W + P_{W2i}V_{Wij} + P_{W3i}U_{Wi} + E_{W_{ijk}}. \tag{2}$$

Here, $P_{.1i}$ is the design matrix for fixed effects of dimension $n_i \times p$, p is the number of fixed covarites (e.g. age, sex) in the model, $P_{.2i}$ and $P_{.3i}$ are design matrices for level-2, level-1 effects respectively. β is an $(p \times 1)$ vector of fixed regression coefficients, and $E_{.ijk}, V_{.ij}, U_{.i}$, denote respectively level-1, level-2, and level-3 random effects. $U_{.i}$ is used for random subject effects, and it explains the between-subject variability. $V_{.ij}$ addresses the within-subject variability. Our assumptions on within and between subject variabilities are $V_{.ij} \sim N(0, Q); U_{.i} \sim N(0, G)$, and on errors $E_{ijk} \sim N(0, \Sigma)$. We further assume that U_{Xi}, V_{Xij} and E_{Xi} are independent, and also U_{Wi}, V_{Wij} and E_{Wi} are independent $i \neq i'$. U_{Xi} and $U_{Xi'}$ are independent, and also, U_{Wi} and $U_{Wi'}$ are independent, $i \neq i'$. V_{Xij} and $V_{Xi'j}$ are independent, $i \neq i'$, similarly V_{Wij} and $V_{Wi'j}$ are independent, $i \neq i'$. $Cov(V_{Xij}, V_{Xij'}) \neq 0$, $j \neq j'$,

and $Cov(V_{Wij}, V_{Wij'}) \neq 0, j \neq j'$. $Cov(U_{Xi}, U_{Wi}) \neq 0$, for two raters evaluating the same subject, and $Cov(V_{Xij}, V_{Wij}) \neq 0$, for two raters evaluating the same subject at the same time point. We further assume that all components in $E_{X_{ijk}}$, and also those in $E_{W_{ijk}}$ are independent. Vectors $E_{X_{ijk}}$ and $E_{W_{ijk}}$ are independent.
Note that in the absence of multiple ratings at each time point (i.e. without replications of ratings), our model reduces to the models discussed by Chinchilli et. al. (1996[17]; 2001[32]; 2007[33, 34]), Choudhary, et. al. (2005)[20, 19], Barnhart, et. al. (2005 [5]).

Let $Y_{ijk} = (X'_{ijk}, W'_{ijk})'$, $\mathbf{U}_{\mathrm{i}} = (\mathbf{U}'_{\mathrm{Xi}}, \mathbf{U}'_{\mathrm{Wi}})'$, $\mathbf{V}_{\mathrm{ij}} = (\mathbf{V}'_{\mathrm{Xij}}, \mathbf{V}_{\mathrm{Wij}}{}'))'$.

$$\mathbf{P}_{1i} = \begin{pmatrix} P_{X1i} & 0 \\ 0 & P_{W1i} \end{pmatrix}, \mathbf{P}_{2i} = \begin{pmatrix} P_{X2i} & 0 \\ 0 & P_{W2i} \end{pmatrix}, \mathbf{P}_{3i} = \begin{pmatrix} P_{X3i} & 0 \\ 0 & P_{W3i} \end{pmatrix}.$$

The mixed-effects model for the aforementioned design is:

$$Y_{ijk} = \mathbf{P}_{1i}\boldsymbol{\beta} + \mathbf{P}_{3i}\mathbf{U}_{\mathrm{i}} + \mathbf{P}_{2i}\mathbf{V}_{\mathrm{ij}} + \mathbf{E}_{\mathrm{ijk}}. \tag{3}$$

In the above model, the trend at the population level is measured by the fixed parameter vector $\boldsymbol{\beta}$. Random vectors are used to measure deviations of that trend at different levels. For example, if a subject is being rated by two raters longitudinally at four different time points, and each rater repeats it four times at every measuring time point, then $T_i = 4$, $k = 4$. Figure 2 shows the raw data plotting of a subject for two raters. For each rater, the bold line represents the average trend of the subject, whereas each line represents the trend of ratings over time for each replicate. Thus, for four replicates, we have four lines plus the average trend line for each rater.

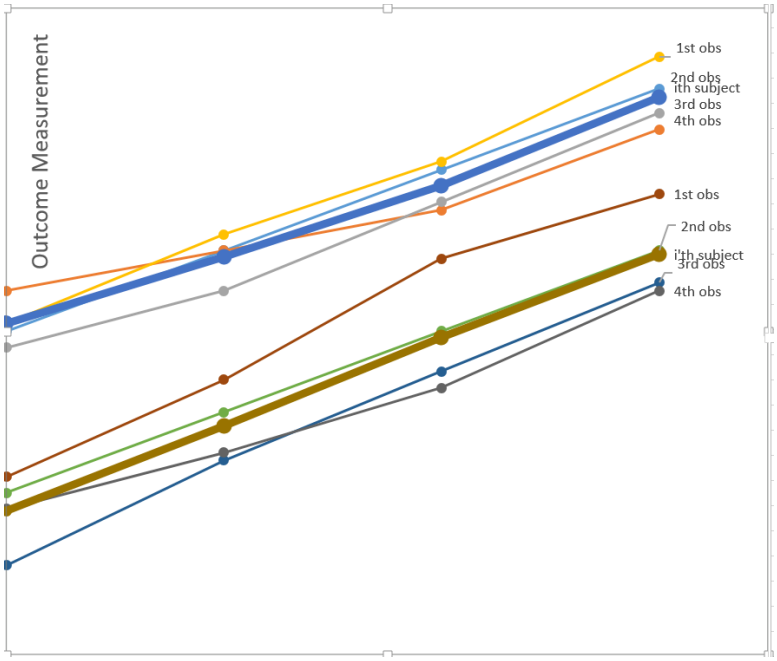


Figure 2: This figure displays trend lines of raw data of two subjects with multiple readings by a rater

Concordance Correlation Coefficient (CCC) for two raters

In the spirit of the general structure described in Model 3, we write a specific model for the ith subject, where x_{ijk} denotes the kth measurement at time j rated by the rater X.

$$x_{ijk} = \beta_0 + \beta_1 j + \theta_{0i} + \theta_{1i} j + \nu_{0ij} + \epsilon_{ijk}, \tag{4}$$

where β_0 is the baseline average of x_{ijk}, β_1 is the slope of x_{ijk} across time, θ_{0i} and θ_{1i} are random intercept and random slope respectively for subject i, ν_{0ij} is random intercept at time j of the i^{th} subject. ν_{0ij} and $\nu_{0ij'}$ are assumed to be uncorrelated. ϵ_{ijk}s are random errors and assumed to be autocorrelated. Similarly, the model for the W rater is:

$$w_{ijk} = \beta_0' + \beta_1' t + \theta_{0i}' + \theta_{1i}' t + \nu_{0ij}' + \epsilon_{ijk}', \tag{5}$$

with all variables similarly defined as in x_{ijk}.
Define $y_{ijkl} = \begin{pmatrix} x_{ijk}|(l=1) & w_{ijk}|(l=2) \end{pmatrix}'$. Combining these two models together, the augmented model for two raters is:

$$\begin{aligned} y_{ijkl} = {} & \beta_0\delta_{(l=1)} + \beta_1\delta_{(l=1)}t + \theta_{0i}\delta_{(l=1)} + \theta_{1i}t\delta_{(l=1)} + \nu_{0ij}\delta_{(l=1)} \\ & + \beta_0'\delta_{(l=2)} + \beta_1'\delta_{(l=2)}t + \theta_{0i}'\delta_{(l=2)} + \theta_{1i}'t\delta_{(l=2)} + \nu_{0ij}'\delta_{(l=2)} + \epsilon_{ijkl}. \end{aligned} \tag{6}$$

Following Model (6), under the distributional assumptions of random effects stated above, the variance covariance matrix of K replicates at time point t by rater X, is denoted by $\mathbf{\Sigma}_{x_{tt}}$, where

$$\Sigma_{x_{tt}} = (\sigma_{\theta_0}^2 + \sigma_{\theta_1}^2 t^2)\mathbf{J}_{K\times K} + \sigma_{\nu_0}^2\mathbf{J}_{K\times K} + \sigma_e^2\mathbf{I}_{K\times K}. \tag{7}$$

Covariance matrix between two different time points (t and t') of the same subject is :

$$\Sigma_{x_{tt'}} = Cov(x_{1t}, x_{1t'}) = (\sigma_{\theta_0}^2 + \sigma_{\theta_1}^2 tt')\mathbf{J}_{K\times K} + \sigma_{\nu_0}^2\mathbf{I}_{K\times K} + \sigma_e^2\mathbf{\Omega}_{K\times K}^{tt'}, \tag{8}$$

Covariance structure for X_i:

$$\Sigma_{X_i} = \begin{pmatrix} \Sigma_{x_{11}} & \Sigma_{x_{12}} & \dots & \Sigma_{x_{TT}} \\ \vdots & \vdots & \vdots & \vdots \\ \Sigma_{x_{T1}} & \Sigma_{x_{T2}} & \dots & \Sigma_{x_{TT}} \end{pmatrix}_{KT\times KT},$$

where $\mathbf{\Omega}_{TK\times TK}$ defines an auto-correlated variance covariance structure, and the tt' th element of it is denoted by $\mathbf{\Omega}_{K\times K}^{tt'}$. The covariance structure for W_i is similar to X_i, and the covariance structure for X_i and $W_{i'}$ is 0, where $i \neq i'$. Let $\sigma_{xw}^2 = Cov(\nu_{0ij}, \nu_{0ij}') + Cov(\theta_{0i}, \theta_{0i}')$, for which between-raters variability is zero for two different subjects. Under these assumptions, the covariance matrix of all T measures of a subject, and that of all subjects rated by raters X and W are respectively

$$\begin{aligned} Cov(X_i, W_i) &= \mathbf{I}_T \otimes (\sigma_{xw}^2\mathbf{J}_{K\times K}), \\ \Sigma_{XW} = Cov(X, W) &= \mathbf{I}_N \otimes [\mathbf{I}_T \otimes (\sigma_{xw}^2\mathbf{J}_{K\times K})], \\ \Sigma_{XX} &= \mathbf{I}_N \otimes \Sigma_{X_i}, \\ \Sigma_{WW} &= \mathbf{I}_N \otimes \Sigma_{W_i}, \end{aligned} \tag{9}$$

Using these notations we define the concordance correlation coefficient [38, 41, 13, 45, 9, 51, 43, 19] between raters X and W and denote it by $\boldsymbol{CCC}_{XW}$, where

$$\boldsymbol{CCC_{xw}} = \frac{tr(\mathbf{\Sigma}_{X+W} - \mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{WW})}{tr(\mathbf{\Sigma}_{XX}) + tr(\mathbf{\Sigma}_{WW}) + (\mu_x - \mu_W)'(\mu_X - \mu_W)}. \tag{10}$$

$$\boldsymbol{CCC_{xw}} = \frac{tr(\mathbf{\Sigma}_{XX} + \mathbf{\Sigma}_{WW} - \mathbf{\Sigma}_{X-W})}{tr(\mathbf{\Sigma}_{XX}) + tr(\mathbf{\Sigma}_{WW}) + (\mu_x - \mu_W)'(\mu_X - \mu_W)} \tag{11}$$

Note that
(i) if $\mu_X = \mu_W$ and $tr(\mathbf{\Sigma}_{XW}) = 0$, then $\boldsymbol{CCC}_{XW} = 0$;

(ii) if $\mu_X = \mu_W$ and $tr(\mathbf{\Sigma}_{XW}) = tr(\mathbf{\Sigma}_{XX}) = tr(\mathbf{\Sigma}_{WW})$, then $\boldsymbol{CCC}_{XW} = 1$;

(iii) if $\mu_X = \mu_W$ and $tr(-\mathbf{\Sigma}_{XW}) = tr(\mathbf{\Sigma}_{XX}) = tr(\mathbf{\Sigma}_{WW})$, then $\boldsymbol{CCC}_{XW} = -1$.

In general, whatever be the parameters $\mu_X, \mu_Y, \mathbf{\Sigma}_{XX}, \mathbf{\Sigma}_{WW}$ and $\mathbf{\Sigma}_{XW}$ and whatever be their interrelations, it follows that (i) $(\mu_x - \mu_W)'(\mu_X - \mu_W) \geq 0$; (ii) $2|tr(\mathbf{\Sigma}_{XW})| \leq tr(\mathbf{\Sigma}_{XX}) + tr(\mathbf{\Sigma}_{WW})$. Consequently, $-1 \leq \boldsymbol{CCC}_{XW} \leq 1$.

An extension of Concordance Correlation Coefficient (CCC) for multiple raters

For the scenario of two raters, we have used two models for raters X and W ((4) & (5)) respectively, and those models are connected by a bivariate covariance structure of the random effects. The idea can be generalized by using a multivariate covariance structure. The combined model (6) can be extended for L raters as follows:

$$y_{ijkl} = \beta_{0l} + \beta_{1l} j + \theta_{0il} + \theta_{1il} j + \nu_{0ijl} + \epsilon_{ijkl}, \tag{12}$$

where, y_{ijkl} denotes kth measurement at time j of ith subject rated by rater l. This model uses the same notations used in (6) except the fact that an extra subscript l has been used in place of the indicators to denote rater specific fixed and random effects.

We assume an unstructured covariance matrix for each of $\begin{pmatrix}\theta_{0i1} & \dots & \theta_{0iL}\end{pmatrix}'$, $\begin{pmatrix}\theta_{1i1} & \dots & \theta_{1iL}\end{pmatrix}'$, $\begin{pmatrix}\nu_{0ij1} & \dots & \nu_{0ijL}\end{pmatrix}'$. These random effects are independent and identically distributed for time-points $j = 1, 2, ..., T$ nested in each subject $i = 1, 2, ..., N$. Let, $var(\theta_{0il}) = \sigma^2_{\theta_{0l}}$, $var(\theta_{1il}) = \sigma^2_{\theta_{1l}}$, $var(\nu_{0ijl}) = \sigma^2_{\nu_{0l}}$ for $l = 1, 2, ..., L$.

Let l and l' be notations for two raters. We denote the vector of all measurements by rater l for ith subject at time point t by $l_{it} = \begin{pmatrix}y_{it1l} & \dots & y_{itKl}\end{pmatrix}'$ and accordingly $l_i = \begin{pmatrix}l'_{i1} & \dots & l'_{iT}\end{pmatrix}'$. For rater l, (7) and (8) can be rewritten as,

$$\Sigma_{l_{tt}} = (\sigma^2_{\theta_{0l}} + \sigma^2_{\theta_{1l}} t^2)\mathbf{J}_{K\times K} + \sigma^2_{\nu_{0l}}\mathbf{J}_{K\times K} + \sigma^2_e \mathbf{I}_{K\times K}. \tag{13}$$

$$\Sigma_{l_{tt'}} = Cov(l_{1t}, l_{1t'}) = (\sigma^2_{\theta_{0l}} + \sigma^2_{\theta_{1l}} tt')\mathbf{J}_{K\times K} + \sigma^2_{\nu_{0l}}\mathbf{I}_{K\times K} + \sigma^2_e \mathbf{\Omega}^{tt'}_{K\times K}, \tag{14}$$

The covariance structure for l_i is $\Sigma_{l_i} = \begin{pmatrix}\Sigma_{l_{11}} & \Sigma_{l_{12}} & \dots & \Sigma_{l_{1T}} \\ \vdots & \vdots & \vdots & \vdots \\ \Sigma_{l_{T1}} & \Sigma_{l_{T2}} & \dots & \Sigma_{l_{TT}}\end{pmatrix}_{KT\times KT}$.

Let, $\sigma^2_{ll'} = Cov(y_{ijkl}, y_{ijkl'}) = Cov(\nu_{0ijl)}, \nu_{0ijl'}) + Cov(\theta_{0il}, \theta_{0il'})$. So, the covariance structure between two raters, l and l' for subject i can be written as $Cov(l_i, l'_i) = \mathbf{I}_T \otimes (\sigma^2_{ll'}\mathbf{J}_{K\times K})$. Using the above notations, (9) can be rewritten as,

$$\begin{aligned}\Sigma_{ll'} &= \mathbf{I}_N \otimes [\mathbf{I}_T \otimes (\sigma^2_{xw}\mathbf{J}_{K\times K})], \\ \Sigma_{ll} &= \mathbf{I}_N \otimes \Sigma_{l_i},\end{aligned} \tag{15}$$

We define, $\mu_{l_i} = E(l_i) = \beta_{0l}\mathbf{1}_{KT\times KT} + \begin{pmatrix}1 & 2 & \dots & T\end{pmatrix}' \otimes [\beta_{1l}\mathbf{1}_{K\times K}]$, for $i = 1, ..., N$, accordingly, $\mu_l = \mathbf{1}_{N\times N} \otimes \mu_{l_1}$.

Generalized $\boldsymbol{CCC}$ for L raters is defined as,

$$\boldsymbol{CCC} = \frac{2\sum_{l<l'} tr(\Sigma_{ll'})}{(L-1)\sum_{l=1}^{L} tr(\Sigma_{ll}) + \sum_{l<l'}(\mu_l - \mu_{l'})'(\mu_l - \mu_{l'})} \tag{16}$$

Note that,

(i) if $tr(\boldsymbol{\Sigma}_{ll'}) = 0$ for all $l \neq l'$, then $\boldsymbol{CCC} = 0$;

(ii) if $\mu_l = \mu_{l'}$ and $tr(\boldsymbol{\Sigma}_{ll'}) = tr(\boldsymbol{\Sigma}_{ll}) = tr(\boldsymbol{\Sigma}_{l'l'})$ for all $l \neq l'$, then $\boldsymbol{CCC} = 1$;

(iii) if $\mu_l = \mu_{l'}$ and $tr(-\boldsymbol{\Sigma}_{ll'}) = tr(\boldsymbol{\Sigma}_{ll}) = tr(\boldsymbol{\Sigma}_{l'l'})$ for all $l \neq l'$, then $\boldsymbol{CCC} = -1$.

In general, (i) $(\mu_l - \mu_{l'})'(\mu_l - \mu_{l'}) \geq 0$; (ii) $2|tr(\boldsymbol{\Sigma}_{ll'})| \leq tr(\boldsymbol{\Sigma}_{ll}) + tr(\boldsymbol{\Sigma}_{l'l'})$ for all $l \neq l'$.. Consequently, $-1 \leq \boldsymbol{CCC} \leq 1$.

Intra-rater Concordance Correlation Coefficient (Intra-CCC)

In the previous subsection we have discussed CCC as a measure of agreement between two or more raters, which can be rephrased as the inter-CCC. In the beginning of this section, we have discussed the hierarchical design (Figure 1), which involves multiple measurements by a particular rater for the same subject at the same time point. This motivates us to develop a measurement of agreement within that rater. Note that when a subject and time point are fixed, the additional variability of the measurement is due to the variability of the unexplained error ϵ. We define intra-CCC for rater l as following,

$$\boldsymbol{CCC}_l^{Intra} = \frac{2\sum_{i,j,k} Cov(y_{ijkl}, y_{ijk'l})}{\sum_{i,j,k}(Var(y_{ijkl}) + Var(y_{ijk'l}))} = \frac{\sigma^2_{\theta_{0l}} + \sigma^2_{\nu_{0l}} + \sum_j j^2\sigma^2_{\theta_{1l}}}{\sigma^2_{\theta_{0l}} + \sigma^2_{\nu_{0l}} + \sum_j j^2\sigma^2_{\theta_{1l}} + \sigma^2_e} \tag{17}$$

Conceptually, the intra-CCC cannot be negative, which is supported by (17), as $\sigma^2_e \geq 0$, and hence $0 \leq \boldsymbol{CCC}_l^{Intra} \leq 1$.

Estimation of Model Parameters

The expectation-maximization (EM) algorithm is used to estimate parameters of the proposed model. In the E-step, with the initial values of the other parameters, we compute the "expected *a posteriori*" or empirical Bayes (EB) estimates of the random effects as well as the conditional variances of the random effects, given the data. In the M-step, given the current values of the random effects, we obtain the maximum marginal likelihood (MML) (or restricted maximum marginal likelihood (REML)) estimates of the regression coefficients, error variances, and the variances of the random effects. The algorithm iterates between the EB and MML or REML estimates until convergence is achieved. In addition, we followed the maximum marginal likelihood method to estimate parameters (Hedeker and Gibbons, 2006[29]). Gauss quadrature is used for numerical integration, and the Newton-Raphson method is used for the optimal solution.

4 Simulation Study

One simulation study is performed for both two and three raters scenario to evaluate the accuracy and reliability of the parameter estimation algorithms. For each of the following scenarios, 10,000 data sets were generated with 30 subjects having X-rays taken at three time-points, and each X-ray subsequently reviewed twice by two/three different raters. The following Models were used to generate data.

Two raters scenario

12 (6 for each rater) observations for each subject were generated according to the following model:

$$\begin{aligned} y_{ijkl} = {} & \beta_0\delta_{(l=1)} + \beta_1\delta_{(l=1)}j + \theta_{0i}\delta_{(l=1)} + \nu_{0ij)}\delta_{(l=1)} \\ & + \beta_0'\delta_{(l=2)} + \beta_1'\delta_{(l=2)}j + \theta_{0i}'\delta_{(l=2)} + \nu_{0ij)}'\delta_{(l=2)} + \epsilon_{ijkl}, \end{aligned} \tag{18}$$

where $l = 1$ denotes the first rater, and $l = 2$ denotes the second rater. Assumptions used to generate data for fixed and random effects parameters are:

(i) Parameters $\beta_0, \beta_0', \beta_1, \beta_1'$ are treated as fixed parameters.

(ii) θ_{0i}, $\nu_{0j(i)}$ and ϵ_{ijkl} are distributed independently.

(iii) $\begin{pmatrix}\theta_{0i}\\ \theta_{0i}'\end{pmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_\theta^2 & \sigma_{\theta\theta'}\\ \sigma_{\theta\theta'} & \sigma_{\theta'}^2\end{bmatrix}\right)$.

(iv) $\begin{pmatrix}\nu_{0ij}\\ \nu_{0ij}'\end{pmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_\nu^2 & \sigma_{\nu\nu'}\\ \sigma_{\nu\nu'} & \sigma_{\nu'}^2\end{bmatrix}\right)$,

(v) $\nu_{0ij'}$ and ν_{0ij}' are independent, when $j' \neq j$.

(vi) $\epsilon \sim N(0, \sigma^2)$, and independent.

For this model, the inter-CCC under (10) or (16) is,

$$CCC = \frac{2 \times 6(\sigma_{\theta\theta'} + \sigma_{\nu\nu'})}{6(\sigma_\theta^2 + \sigma_{\theta'}^2 + \sigma_\nu^2 + \sigma_{\nu'}^2) + ||(\beta_0 - \beta_0')\mathbf{1}_{6\times 6} + \begin{pmatrix}1 & 2 & 3\end{pmatrix}' \otimes [(\beta_1 - \beta_1')\mathbf{1}_{2\times 2}]||_2^2}, \tag{19}$$

where $||.||_2$ denotes L_2 norm. Using (17),the intra-CCC for rater 1 is written as,

$$CCC_1^{Intra} = \frac{\sigma_\theta^2 + \sigma_\nu^2}{\sigma_\theta^2 + \sigma_\nu^2 + \sigma^2} \tag{20}$$

Note that, in this model we have not considered any random slope, so, the variance component of the random slope has been omitted from both numerator and denominator of (17) to produce (20). Intra-CCC for rater 2 can be expressed similarly.

Three raters scenario

18 (6 for each rater) observations were generated for each subject according to the following model:

$$\begin{aligned} y_{ijkl} = {} & (\beta_0 + \beta_1 j + \theta_{0i} + \nu_{0ij})\delta_{(l=1)} \\ & + (\beta_0' + \beta_1' j + \theta_{0i}' + \nu_{0ij}')\delta_{(l=2)} \\ & + (\beta_0'' + \beta_1'' j + \theta_{0i}'' + \nu_{0ij}'')\delta_{(l=3)} + \epsilon_{ijkl}, \end{aligned} \tag{21}$$

where $l = i$ means the ith rater for $i = 1, 2, 3$. Assumptions for generating fixed and random parameters are:

(i) Parameters $\beta_0, \beta_0', \beta_0'', \beta_1, \beta_1', \beta_1''$ are treated as fixed parameters.

(ii) θ_{0i}, ν_{0ij} and ϵ_{ijkl} are distributed independently.

(iii) $\begin{pmatrix}\theta_{0i}\\ \theta'_{0i}\\ \theta''_{0i}\end{pmatrix} \sim N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2_\theta & \sigma_{\theta\theta'} & \sigma_{\theta\theta''}\\ \sigma_{\theta\theta'} & \sigma^2_{\theta'} & \sigma_{\theta'\theta''}\\ \sigma_{\theta\theta''} & \sigma_{\theta'\theta''} & \sigma^2_{\theta''}\end{bmatrix}\right).$

(iv) $\begin{pmatrix}\nu_{0ij}\\ \nu'_{0ij}\\ \nu''_{0ij}\end{pmatrix} \sim N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2_\nu & \sigma_{\nu\nu'} & \sigma_{\nu\nu''}\\ \sigma_{\nu\nu'} & \sigma^2_{\nu'} & \sigma_{\nu'\nu''}\\ \sigma_{\nu\nu''} & \sigma_{\nu'\nu''} & \sigma^2_{\nu''}\end{bmatrix}\right),$

(v) ν_{0ij_1}, ν'_{0ij_2} and ν''_{0ij_3} are mutually independent, when $j_1 \neq j_2 \neq j_3$.

(vi) $\epsilon \sim N(0, \sigma^2)$, and independent.

Using (16), inter-CCC for three raters can be written as,

$$\begin{aligned} CCC &= \frac{A}{B}\\ A &= 2\times 6(\sigma_{\theta\theta'}+\sigma_{\theta\theta''}+\sigma_{\theta'\theta''}+\sigma_{\nu\nu'}+\sigma_{\nu\nu''}+\sigma_{\nu'\nu''})\\ B &= \big(2\times 6(\sigma^2_\theta+\sigma^2_{\theta'}+\sigma^2_{\theta''}+\sigma^2_\nu+\sigma^2_{\nu'}+\sigma^2_{\nu''})\\ &\quad + ||(\beta_0-\beta'_0)\mathbf{1}_{6\times 6}+\begin{pmatrix}1 & 2 & 3\end{pmatrix}'\otimes[(\beta_1-\beta'_1)\mathbf{1}_{2\times 2}]||^2_2\\ &\quad + ||(\beta_0-\beta''_0)\mathbf{1}_{6\times 6}+\begin{pmatrix}1 & 2 & 3\end{pmatrix}'\otimes[(\beta_1-\beta''_1)\mathbf{1}_{2\times 2}]||^2_2\\ &\quad + ||(\beta'_0-\beta''_0)\mathbf{1}_{6\times 6}+\begin{pmatrix}1 & 2 & 3\end{pmatrix}'\otimes[(\beta'_1-\beta''_1)\mathbf{1}_{2\times 2}]||^2_2\big). \end{aligned} \tag{22}$$

Intra-CCC for all raters can be calculated similarly using the expression in (20).
For each simulated data set, fixed effect and random effect parameters are estimated according to the estimation method described in the last section. The performance of this method is evaluated using biases and root mean squared errors (RMSE).
Intra-CCC is used for measuring an agreement among multiple replications measured by a single rater. Inter-CCC is used for measuring the agreement among different raters. In the following tables, the abbreviations are as follows: EST=estimate, STD=Standard Deviation, RMSE=root mean squared error, $Intra_{l=1}$=Intra-CCC for rater 1, $Intra_{l=2}$=Intra-CCC for rater 2, $Intra_{l=3}$=Intra-CCC for rater 3, $Inter2$=Inter-CCC between two raters, $Inter3$=Inter-CCC between three raters.

Table 1 and Table 2 show simulated results when distributional parameters (two raters scenario) are set close to the estimated parameters from the real data. These tables reveal that parameter estimates are close to true values, biases are close to zero with a mixture of positive and negative biases. Although STDs and RMSEs of fixed effects are reasonably satisfactory for smaller sample sizes (e.g. 30), those of variance components are not encouraging. The corresponding CCCs are estimated with high efficiency. Theoretically these estimates are consistent, we observe that STDs decrease for larger sample sizes (e.g. 50). We have performed another simulation study with sample size 50 which shows a significant decrease of STDs for example – STD of σ^2_ν (Table 1) decreases to .17 from .37, STD of $\sigma^2_{\nu'}$ (Table 1) decreases to .13 from .27, STD of $\sigma^2_{\theta''}$ (Table 2) decreases to .07 from .12, STD of $\sigma_{\theta'\theta''}$ (Table 2) decreases to .07 from .11. Taken together, this simulation study demonstrates the accuracy and precision in estimates of the model parameters.

Category	Parameter	True Value	EST(STD)	Bias	RMSE
Fixed	β_0	3.65	3.6385 (0.2238)	-0.0076	0.2240
	β_0'	4.11	4.0998 (0.1932)	-0.0053	0.1933
	β_1	-0.10	-0.0861 (0.0552)	0.0090	0.0560
	β_1'	-0.16	-0.1502 (0.0545)	0.0077	0.0551
Random variance	σ_ν^2	1.36	1.3540 (0.3653)	-0.0083	0.3654
	σ_θ^2	0.11	0.1115 (0.0354)	0.0001	0.0354
	$\sigma_{\nu'}^2$	0.97	0.9669 (0.2663)	-0.0050	0.2663
	$\sigma_{\theta'}^2$	0.10	0.1211 (0.0354)	0.0211	0.0412
	σ^2	0.16	0.1613 (0.0173)	-0.0007	0.0173
	$\sigma_{\nu\nu'}$	1.12	1.1131 (0.3044)	-0.0065	0.3045
	$\sigma_{\theta\theta'}$	0.09	0.0960 (0.0283)	0.0060	0.0289
CCC	$Intra_{l=1}$	0.90	0.8951 (0.0284)	-0.0059	0.0290
	$Intra_{l=2}$	0.93	0.8640 (0.0358)	-0.0664	0.0754
	$Inter$	0.81	0.8014 (0.0434)	-0.0107	0.0447

Table 1: Estimation of model parameters and computation of CCC for two raters scenario.

5 Analysis of Motivational Example Data

In this section, we would like to apply methods discussed in Section 3 to evaluate agreement between different raters. Back to the GAIT trial, subjects enrolled in the GAIT ancillary structural study met the original GAIT inclusion criteria, summarized as being at least 40 years of age with clinical evidence of painful OA in knee(s) for at least the immediate past six months and radiographic evidence of OA as determined by having a Kellgren & Lawrence grade 2- or 3-rated radiograph of the index knee. Of the 1583 original GAIT study participants, 662 were also in the structural study. Patients were asked to continue follow-up even if they stopped taking their assigned treatment. The study treatments were glucosamine hydrochloride (HCl) 500 mg three times daily, sodium chondroitin sulfate 400 mg three times daily, both glucosamine and chondroitin sulfate as above, celecoxib 200 mg once daily or placebo daily.

The metatarsophalangeal films were obtained as previously described, using the method of Buckland-Wright on participants in the GAIT radiographic sub-study at baseline, at one year and two years or until the end of the study.

Baseline analysis

The joint space width (JSW) at baseline of 281 participants was used to evaluate the agreement between raters. The results in this section are based on only one reading by each rater. The mean and standard deviation of the difference of measurements between two raters are presented to provide information regarding the average magnitude of the difference. Inter-rater agreement between readers is assessed statistically by an unconditional intra-class correlation (ICC). ICC is defined by $\rho_u = \sigma^2_{subject}/(\sigma^2_{subject} + \sigma^2_{rater} + \sigma^2_\epsilon)$. The differences in readings between readers were also examined using Bland-Altman plots to graphically investigate systematic differences in disagreement [Bland, 1986, 1993]. Disagreement in a Bland-Altman plot is being indicated by differences that fall within the 95% confidence limits of the mean difference. Table 3 shows that Raters 1 and 2 have higher ICC and lower mean difference on JSW at baseline.

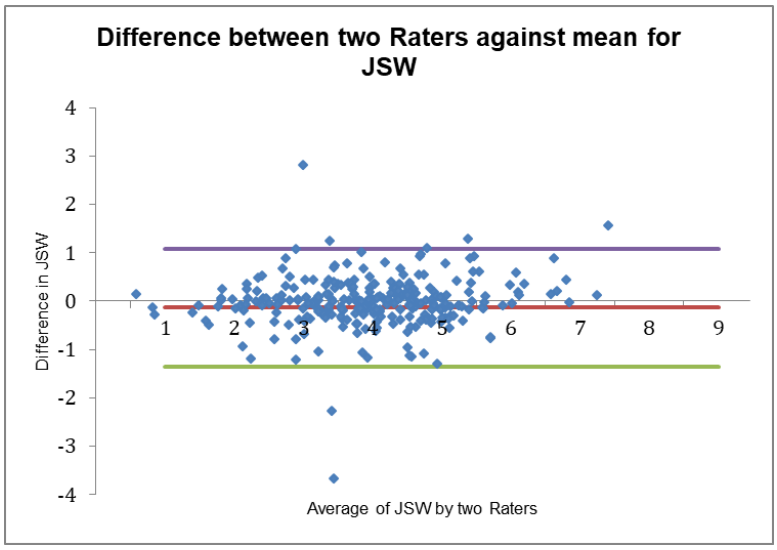


Figure 3: Rater 1 vs Rater 2 Bland-Altman plot n=281 radiographs.

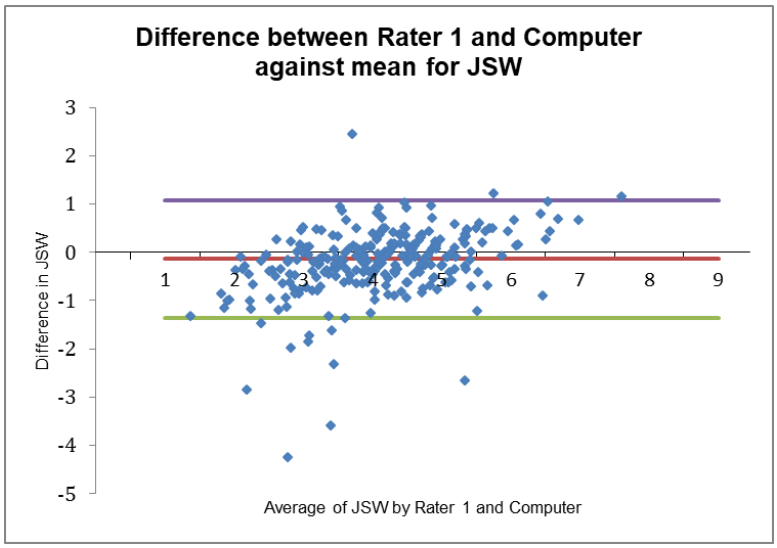


Figure 4: Rater 1 vs Computer Bland-Altman plot n=281 radiographs.

Category	Parameter	True Value	EST(STD)	Bias	RMSE
Fixed	β_0	3.75	3.7481 (0.3333)	-0.0019	0.3333
	β_0'	4.11	4.1109 (0.3203)	0.0009	0.3202
	β_0''	4	3.9996 (0.2969)	-0.0004	0.2969
	β_1	-0.10	-0.1003 (0.1504)	-0.0003	0.1504
	β_1'	-0.16	-0.1612 (0.1451)	-0.0012	0.1451
	β_1''	-0.12	-0.1204 (0.1330)	-0.0004	0.1330
Random variance	σ_ν^2	1.30	1.2518 (0.2272)	-0.0482	0.2323
	σ_θ^2	0.11	0.1459 (0.1309)	0.0359	0.1357
	$\sigma_{\nu'}^2$	1.20	1.1570 (0.2085)	-0.0430	0.2129
	$\sigma_{\theta'}^2$	0.10	0.0958 (0.1222)	-0.0042	0.1223
	$\sigma_{\nu''}^2$	1.00	0.9656 (0.1780)	-0.0344	0.1813
	$\sigma_{\theta''}^2$	0.14	0.1214 (0.1186)	-0.0186	0.1200
	σ^2	0.16	0.1591 (0.0136)	-0.0009	0.0136
	$\sigma_{\nu\nu'}$	1.12	1.0798 (0.2005)	-0.0402	0.2045
	$\sigma_{\nu\nu''}$	1.00	0.9617 (0.1818)	-0.0383	0.1858
	$\sigma_{\nu'\nu''}$	1.00	0.9644 (0.1778)	-0.0356	0.1813
	$\sigma_{\theta\theta'}$	0.08	0.0991 (0.1155)	0.0060	0.1157
	$\sigma_{\theta\theta''}$	0.07	0.1078 (0.1056)	0.0378	0.1122
	$\sigma_{\theta'\theta''}$	0.05	0.0785 (0.1066)	0.0285	0.1103
CCC	$Intra_{l=1}$	0.90	0.8961 (0.0174)	-0.0039	0.0178
	$Intra_{l=2}$	0.89	0.8885 (0.0185)	-0.0015	0.0186
	$Intra_{l=3}$	0.87	0.8746 (0.0207)	0.0046	0.0212
	$Inter3$	0.76	0.7509 (0.0353)	-0.0065	0.0359

Table 2: Estimation of model parameters and computation of CCC for three raters scenario.

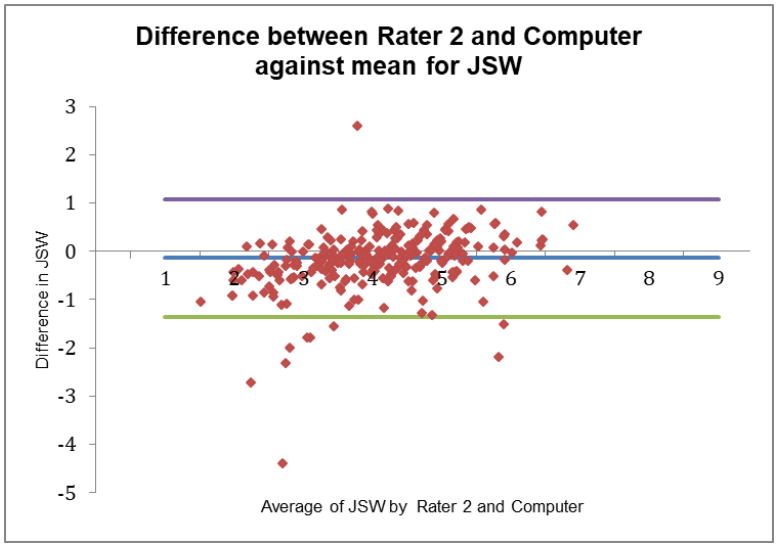


Figure 5: Rater 2 vs Computer Bland-Altman plot n=281 radiographs.

Bland-Altman plots for all evaluated films reveal that Rater 1 was the most reliable as most readings fall within the two standard deviation boundary, whereas readings from Rater 2 are little more scattered, and the Computer is the poorest of all. The plots (Figures 3-5) are based on one reading by each rater at baseline.

	ICC	Mean difference	95% CI
Rater 1 vs Rater 2	0.90	-0.02	-0.08, 0.04
Rater 1 vs Computer	0.80	-0.20	-0.28, -0.12
Rater 2 vs Computer	0.83	-0.18	-0.26, -0.11

Table 3: ICC and its confidence interval for human and computer: Joint Space Width n=281 radiographs.

Analysis for complete longitudinal data

Table 4 shows the Intra-CCC for each rater. Values of Intra-CCC by the mixed model suggest that Rater 1 is the most reliable rater.

Intra-CCC	Rater 1	Rater 2	Rater 3
	0.9190	0.8778	0.8436

Table 4: Intra-CCC estimation for three raters.

Table 5 presents results of CCC between Raters 1 and 2, Raters 1 and 3, and Raters 2 and 3. We use the inter-CCC to show the quantitative agreement level between two raters. CCC between the second rater with the other two raters are relatively small. Also, we estimate CCC by three-level and two-level mixed effects models. The estimates of fixed variables by the three-level model are almost the same as those by the two-level mixed effects model, but the third level variance-covariance reduces the role of the error variance, and increases the covariance between raters slightly. Therefore, we get a higher CCC estimate by three-level models compared to two-level models. In fact, based on the data structure as we have shown before, the three-level model captures different levels of variance and covariance, and the results are closer to the real data. Therefore, estimated CCCs based on three-level models are better than corresponding CCCs from two-level models.

6 Discussion

We have developed statistical methodologies to measure agreement for multiple raters using hierarchical longitudinal designs. These methods use covariance matrices and mean vectors of raters to measure agreement. The methods can be extended to adjust the influence of covariates. In addition, we use intra-CCC for measuring agreement within a rater. Intra-CCC is specifically important when an existing manual system is going to be replaced by artificial intelligence. In this contest, we developed a measure for agreement for all raters jointly.

We did not address the inferential aspect of the problem in this manuscript. Bhaumik et. al. [8] constructed the generalized confidence interval (GCI) for CCC utilizing a bivariate normal distribution of measurements, and also developed a large sample based confidence interval (LSCI). They established satisfactory performance of GCI by providing the desired coverage probability (CP) via simulation. The derivation of the distribution of CCC for our hierarchical design is intractable. Thus

Rater 1 & 2			
Category	Parameters	Three-level model Est(Std)	Two-level model Est(Std)
Fixed	β_0	3.6461 (0.2333)	3.6461 (0.2304)
	$\beta_0^{'}$	3.3876 (0.2342)	3.3876 (0.2339)
	β_1	-0.0952 (0.0723)	-0.0952 (0.0623)
	$\beta_1^{'}$	-0.0632 (0.0634)	-0.0632 (0.0623)
Random variance	σ^2_{θ}	1.3252 (0.3818)	1.3514 (0.3813)
	$\sigma_{\theta\theta^{'}}$	1.0882 (0.3511)	1.1596 (0.3509)
	$\sigma^2_{\theta^{'}}$	1.3964 (0.3940)	1.3991 (0.3940)
	σ^2_{ν}	0.0706 (0.0616)	
	$\sigma_{\nu\nu^{'}}$	0.2142 (0.0374)	
	$\sigma^2_{\nu^{'}}$	0.0000 (.)	
	σ^2	0.4659 (0.0434)	0.4498 (0.0375)
CCC	*Inter*	0.6900	0.6266
Rater 1 & 3			
Fixed	β_0	3.6461 (0.2291)	3.6461 (0.2264)
	$\beta_0^{'}$	4.1051 (0.1955)	4.1051 (0.1935)
	β_1	-0.0952 (0.0576)	-0.0952 (0.0454)
	$\beta_1^{'}$	-0.1579 (0.0532)	-0.1579 (0.0454)
Random variance	σ^2_{θ}	1.3623 (0.3814)	1.3866 (0.3812)
	$\sigma_{\theta\theta^{'}}$	1.1196 (0.3161)	1.1538 (0.3160)
	$\sigma^2_{\theta^{'}}$	0.9719 (0.2746)	0.9868 (0.2744)
	σ^2_{ν}	0.1114 (0.0371)	
	$\sigma_{\nu\nu^{'}}$	0.1025 (0.0272)	
	$\sigma^2_{\nu^{'}}$	0.0831 (0.0319)	
	σ^2	0.1620 (0.0174)	0.2390 (0.0199)
CCC	*Inter*	0.8121	0.7670
Rater 2 & 3			
Fixed	β_0	3.3876 (0.2345)	3.3876 (0.2340)
	$\beta_0^{'}$	4.1051 (0.1989)	4.1051 (0.1983)
	β_1	-0.0632 (0.0645)	-0.0632 (0.0628)
	$\beta_1^{'}$	-0.1579 (0.0645)	-0.1579 (0.0628)
Random variance	σ^2_{θ}	1.3935 (0.3940)	1.3979 (0.3940)
	$\sigma_{\theta\theta^{'}}$	0.9504 (0.3002)	1.0034 (0.3000)
	$\sigma^2_{\theta^{'}}$	0.9461 (0.2745)	0.9504 (0.2744)
	σ^2_{ν}	0.0000 (.)	
	$\sigma_{\nu\nu^{'}}$	0.1589 (0.0282)	
	$\sigma^2_{\nu^{'}}$	0.0000 (.)	
	σ^2	0.4832 (0.0439)	0.4572 (0.0381)
CCC	*Inter*	0.6006	0.5497
Rater 1, 2 & 3			
CCC	*Inter* (Three raters)	0.6593	0.6388

Table 5: Inter-CCC estimation for three-level and two-level models

testing of any hypothesis related to CCC will be based on large sample (Chaudhary 2017[20, 19]). It should be noted that there is no gold standard for CCC in the literature[6, 15, 13, 27, 25], neither in general nor particularly for this problem. It is worth constructing a confidence interval for CCC using bootstrapping or some other approach for small samples. This can be a future direction of research in this area.

Note: Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Acknowledgments

This paper is the result of work supported in part with resources and the use of facilities at the U.S. Department of Veteran Affairs' Hines Cooperative Studies Program Coordinating Center, Hines, Illinois.

References

[1] Sharon Anderson and Walter W Hauck. "Consideration of individual bioequivalence." In: *Journal of Pharmacokinetics and Biopharmaceutics* 18.3 (1990), pp. 259–273.

[2] Mousumi Banerjee et al. "Beyond kappa: A review of interrater agreement measures." In: *Canadian Journal of Statistics* 27.1 (1999), pp. 3–23.

[3] Huiman X Barnhart, Michael Haber, and Jingli Song. "Overall concordance correlation coefficient for evaluating agreement among multiple observers." In: *Biometrics* 58.4 (2002), pp. 1020–1027.

[4] Huiman X Barnhart, Michael J Haber, and Lawrence I Lin. "An overview on assessing agreement with continuous measurements." In: *Journal of Biopharmaceutical Statistics* 17.4 (2007), pp. 529–569.

[5] Huiman X Barnhart, Jingli Song, and Michael J Haber. "Assessing intra, inter and total agreement with replicated readings." In: *Statistics in medicine* 24.9 (2005), pp. 1371–1384.

[6] Huiman X Barnhart et al. "Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement." In: *Journal of Biopharmaceutical Statistics* 17.4 (2007), pp. 721–738.

[7] Ionut Bebu and Thomas Mathew. "Comparing the means and variances of a bivariate lognormal distribution." In: *Statistics in Medicine* 27.14 (2008), pp. 2684–2696.

[8] Dulal K. Bhaumik et al. "Generalized confidence interval for an agreement between raters." In: *Statistics in Medicine* 40.9 (2021), pp. 2230–2238.

[9] Arto Luoma Bikas Kumar Sinha Erkki Liski. "Optimal design for an inverse prediction problem under random coefficient." In: *Indian Soc. Agricultural Statist* 49.1996/97 (1996), pp. 277–288.

[10] J Martin Bland and Douglas G Altman. "Measuring agreement in method comparison studies." In: *Statistical Methods in Medical Research* 8.2 (1999), pp. 135–160.

[11] J Martin Bland and DouglasG Altman. "Statistical methods for assessing agreement between two methods of clinical measurement." In: *The Lancet* 327.8476 (1986), pp. 307–310.

[12] JC Buckland-Wright et al. "Quantitative microfocal radiographic assessment of osteoarthritis of the knee from weight bearing tunnel and semiflexed standing views." In: *The Journal of Rheumatology* 21.9 (1994), pp. 1734–1741.

[13] Josep L Carrasco and Lluis Jover. "Estimating the generalized concordance correlation coefficient through variance components." In: *Biometrics* 59.4 (2003), pp. 849–858.

[14] Josep L Carrasco, Tonya S King, and Vernon M Chinchilli. "The concordance correlation coefficient for repeated measures estimated by variance components." In: *Journal of Biopharmaceutical Statistics* 19.1 (2009), pp. 90–105.

[15] Josep L Carrasco et al. "Distinguishability and agreement with continuous data." In: *Statistics in Medicine* 33.1 (2014), pp. 117–128.

[16] YeaHung Chen and XiaoHua Zhou. "Interval estimates for the ratio and difference of two lognormal means." In: *Statistics in Medicine* 25.23 (2006), pp. 4099–4113.

[17] Vernon M Chinchilli et al. "A weighted concordance correlation coefficient for repeated measurement designs." In: *Biometrics* (1996), pp. 341–353.

[18] Helena Chmura Kraemer, Vyjeyanthi S Periyakoil, and Art Noda. "Kappa coefficients in medical research." In: *Statistics in Medicine* 21.14 (2002), pp. 2109–2129.

[19] Pankaj K Choudhary and Haikady N Nagaraja. *Measuring Agreement: Models, Methods, and Applications*. John Wiley Sons, 2017.

[20] Pankaj K Choudhary and HN Nagaraja. "Assessment of agreement using intersection-union principle." In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 47.5 (2005), pp. 674–681.

[21] Daniel O Clegg et al. "Glucosamine, chondroitin sulfate, and the two in combination for painful knee osteoarthritis." In: *New England Journal of Medicine* 354.8 (2006), pp. 795–808.

[22] J. Cohen. "A Coefficient of Agreement for Nominal Scales." In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46.

[23] J. Cohen. "Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit." In: *Psychological Bulletin* 70.4 (1968), pp. 213–220.

[24] DE Dupuis et al. "Precision and accuracy of joint space width measurements of the medial compartment of the knee using standardized MTP semi-flexed radiographs." In: *Osteoarthritis and Cartilage* 11.10 (2003), pp. 716–724.

[25] Ganesh Dutta and Bikas Kumar Sinha. "Some Further Aspects of Assessment of Agreement involving Bivariate Normal Responses." In: *International Journal of Statistical Sciences* 13 (2013), pp. 1–19.

[26] Alan E Gelfand and Adrian FM Smith. "Sampling-based approaches to calculating marginal densities." In: *Journal of the American statistical association* 85.410 (1990), pp. 398–409.

[27] AS Hedayat, Congrong Lou, and Bikas K Sinha. "A statistical approach to assessment of agreement involving multiple raters." In: *Communications in Statistic Theory and Methods* 38.16-17 (2009), pp. 2899–2922.

[28] Donald Hedeker and Robert D Gibbons. "Application of random-effects pattern-mixture models for missing data in longitudinal studies." In: *Psychological Methods* 2.1 (1997), p. 64.

[29] Donald Hedeker and Robert D Gibbons. *Longitudinal data analysis*. John Wiley & Sons, 2006.

[30] Donald Hedeker, Robert D Gibbons, and Brian R Flay. "Random-effects regression models for clustered data with an example from smoking prevention research." In: *Journal of Consulting and Clinical Psychology* 62.4 (1994), p. 757.

[31] Daniel J Holder and Francis Hsuan. "Moment-based criteria for determining bioequivalence." In: *Biometrika* 80.4 (1993), pp. 835–846.

[32] Tonya S King and Vernon M Chinchilli. "A generalized concordance correlation coefficient for continuous and categorical data." In: *Statistics in Medicine* 20.14 (2001), pp. 2131–2147.

[33] Tonya S King, Vernon M Chinchilli, and Josep L Carrasco. "A repeated measures concordance correlation coefficient." In: *Statistics in Medicine* 26.16 (2007), pp. 3095–3113.

[34] Tonya S King et al. "A class of repeated measures concordance correlation coefficients." In: *Journal of Biopharmaceutical Statistics* 17.4 (2007), pp. 653–672.

[35] K Krishnamoorthy and Thomas Mathew. "Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals." In: *Journal of Statistical Planning and Inference* 115.1 (2003), pp. 103–121.

[36] Nan M Laird, James H Ware, et al. "Random-effects models for longitudinal data." In: *Biometrics* 38.4 (1982), pp. 963–974.

[37] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data." In: *Biometrics* (1977), pp. 159–174.

[38] I Lawrence and Kuei Lin. "A concordance correlation coefficient to evaluate reproducibility." In: *Biometrics* (1989), pp. 255–268.

[39] I Lawrence and Kuei Lin. "Assay validation using the concordance correlation coefficient." In: *Biometrics* (1992), pp. 599–604.

[40] MP Hellio Le Graverand, Eric P Vignon, and Kenneth D Brandt. "Head-to-head comparison of the Lyon Schuss and fixed flexion radiographic techniques. Long-term reproducibility in normal knees and sensitivity to change in osteoarthritic knees." In: *Annals of the Rheumatic Diseases* 67.11 (2008), pp. 1562–1566.

[41] Lawrence Lin. "Overview of agreement statistics for medical devices." In: *Journal of Biopharmaceutical Statistics* 18.1 (2007), pp. 126–144.

[42] Lawrence Lin, AS Hedayat, and Wenting Wu. "A unified approach for assessing agreement for continuous and categorical data." In: *Journal of Biopharmaceutical Statistics* 17.4 (2007), pp. 629–652.

[43] Lawrence Lin, AS Hedayat, and Wenting Wu. *Statistical tools for measuring agreement*. Springer Science & Business Media, 2012.

[44] Lawrence Lin and Lynn D Torbeck. "Coefficient of accuracy and concordance correlation coefficient: new statistics for methods comparison." In: *PDA Journal of Pharmaceutical Science and Technology* 52.2 (1998), pp. 55–59.

[45] Lawrence Lin et al. "Statistical methods in assessing agreement: Models, issues, and tools." In: *Journal of the American Statistical Association* 97.457 (2002), pp. 257–270.

[46] Lawrence Kuei Lin. "A note on the concordance correlation coefficient." In: *Biometrics* 56 (2000), pp. 324–325.

[47] Lawrence Kuei Lin. "Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence." In: *Statistics in Medicine* 19.2 (2000), pp. 255–270.

[48] LIK Lin and V Chinchilli. "Comment on the use of concordance correlation to assess the agreement between two variables-Rejoinder." In: *Biometrics* 53.2 (1997), pp. 777–778.

[49] Mary J Lindstrom and Douglas M Bates. "Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data." In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1014–1022.

[50] Erkki P Liski, Arto Luoma, and Bikas K Sinha. "Optimal designs in random coefficient linear regression models." In: *Calcutta Statistical Association Bulletin* 46.3-4 (1996), pp. 211–230.

[51] Erkki P Liski et al. "Optimal designs for prediction in random coefficient linear regression models." In: *Journal of Combinatorics, Information and System Sciences* 23.1-4 (1998), pp. 1–16.

[52] Roderick JA Little. "Modeling the drop-out mechanism in repeated-measures studies." In: *Journal of the American Statistical Association* 90.431 (1995), pp. 1112–1121.

[53] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[54] Steven A Mazzuca et al. "Pitfalls in the accurate measurement of joint space narrowing in semiflexed, anteroposterior radiographic imaging of the knee." In: *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology* 50.8 (2004), pp. 2508–2515.

[55] Jorge Quiroz and Richard K Burdick. "Assessment of Individual Agreements with Repeated Measurements Based on Generalized Confidence Intervals." In: *Journal of Biopharmaceutical Statistics* 19.2 (2009), pp. 345–359.

[56] Donald B Rubin. "Inference and missing data." In: *Biometrika* 63.3 (1976), pp. 581–592.

[57] Allen D Sawitzke et al. "The effect of glucosamine and/or chondroitin sulfate on the progression of knee osteoarthritis: a report from the glucosamine/chondroitin arthritis intervention trial." In: *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology* 58.10 (2008), pp. 3183–3191.

[58] Robert Schall. "Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar." In: *Biometrics* (1995), pp. 615–626.

[59] Robert Schall and Herman G Luus. "On population and individual bioequivalence." In: *Statistics in Medicine* 12.12 (1993), pp. 1109–1124.

[60] Robert Schall et al. "Towards a practical strategy for assessing individual bioequivalence." In: *Journal of Pharmacokinetics and Biopharmaceutics* 24.1 (1996), pp. 133–149.

[61] William A Scott. "Reliability of content analysis: The case of nominal scale coding." In: *Public Opinion Quarterly* (1955), pp. 321–325.

[62] Shayle R Searle, George Casella, and Charles E McCulloch. *Variance components*. John Wiley & Sons, 2009.

[63] Lewis B Sheiner. "Bioequivalence revisited." In: *Statistics in Medicine* 11.13 (1992), pp. 1777–1788.

[64] Mohamed M Shoukri, MH Asyali, and A Donner. "Sample size requirements for the design of reliability study: review and new results." In: *Statistical Methods in Medical Research* 13.4 (2004), pp. 251–271.

[65] Bikas K Sinha, Pornpis Yimprayoon, and Montip Tiensuwan. "Cohen's kappa statistic: A critical appraisal and some modifications." In: *Calcutta Statistical Association Bulletin* 58.3-4 (2006), pp. 151–170.

[66] Martin A Tanner and Wing Hung Wong. "The calculation of posterior distributions by data augmentation." In: *Journal of the American statistical Association* 82.398 (1987), pp. 528–540.

[67] H Toutenburg. "Rubin, DB: Multiple imputation for nonresponse in surveys." In: *Statistical Papers* 31.1 (1990), pp. 180–180.

[68] Miao-Yu Tsai and Chao-Chun Lin. "Concordance correlation coefficients estimated by variance components for longitudinal normal and Poisson data." In: *Computational Statistics & Data Analysis* 121 (2018), pp. 57–70.

[69] Alexander Von Eye, Patrick Mair, and Eun-Young Mun. *Advances in configural frequency analysis*. Guilford Press, 2010.

[70] E Vonesh and VM Chinchilli. "Linear and Non-Linear Models for the Analysis of Repeated Measurements Marcel Decker." In: *Inc, New York, NY* (1997).

[71] Edward F Vonesh, Vernon M Chinchilli, and Kewei Pu. "Goodness-of-fit in generalized nonlinear mixed-effects models." In: *Biometrics* (1996), pp. 572–587.

[72] Samaradasa Weerahandi. "Generalized Confidence Intervals." In: *Journal of the American Statistical Association* 88.423 (1993), pp. 899–905.

[73] Samaradasa Weerahandi. *Generalized Confidence Intervals*. Springer Series in Statistics, 1995.

[74] P Yimprayoon, M. Tiensuwan, and Bimal K. Sinha. "Some statistical aspects of assessing agreement: theory and applications, (English summary), Festschrift for Tarmo Pukkila on his 60th birthday." In: *Dep. Math. Stat. Philos* (2006), pp. 327–346.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Some Applications of Mixture Regression Modelling Techniques

Tapio Nummi [1], Jyrki Möttönen [2], Janne Salonen [3] and Timothy E. O'Brien [4]

1. *Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland*
2. *Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland*
3. *Finnish Public Sector Pension Provider Keva, Helsinki, Finland*
4. *Loyola University Chicago, Department of Mathematics and Statistics, and Institute of Environmental Sustainability, Chicago, Illinois, USA*

Received: 12 Jun 2022 / Revised: 22 Oct 2022 / Accepted: 21 Oct 2022

Abstract

Many practical data sets are quite heterogeneous in terms of the observed outcome variables, and it may not be possible to model this heterogeneity based on the obtained data. Various approaches, e.g., based on the latent variables, have been proposed for these kind of problems. This paper presents a selection of methods based on the so-called finite mixtures. Of particular interest are the various mixture regression techniques and their extensions with applications to interesting real-world modelling problems.

Keywords: Finite mixtures, Generalized linear model, Mixture normal distribution, Multivariate mixtures, Number of mixture components, Trajectory analysis.

1 Introduction

In this article, we introduce and illustrate basic finite mixture modelling techniques for univariate and multivariate data. In addition to traditional distribution and regression modelling [4, 15, 21], multivariate mixtures [28] and mixture based longitudinal analysis (so-called trajectory analysis (TA) see e.g. [20], [16], [19] and [9]) in one-dimensional and multidimensional case are also considered. For further developments of the basic TA model see [23] and [34] and for the growth curve model [26, 10] see [25]. As theoretical developments and details are provided elsewhere, our focus here is to demonstrate the usefulness of these techniques on real-world applications.

The structure of the article is as follows. General finite mixtures are discussed and illustrated in Section 2, with Section 3 covering and demonstrating finite mixtures in regression modelling. Multivariate mixtures are the focus of Section 4, with multivariate normal mixtures in Section 4 and mixtures of longitudinal data with extensions in Section 4. Our final comments are given in Section 5.

2 Finite Mixtures

In this section, we introduce and illustrate basic finite mixtures of normal distributions. To this end, let $y_1, \ldots, y_n$, be n independent realizations of random variable Y. These observations are assumed to have arisen from K sub-populations. Let $Z_i = k$ if the ith observation belongs to the sub-population k with the probability $\pi_k = P(Z_i = k)$, such that $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$. The conditional density of y_i given $Z_i = k$ is denoted by $f_k(y_i; \boldsymbol{\theta}_k)$, and these densities are typically assumed to arise from the same parametric family with parameters $\boldsymbol{\theta}_k$ for $k = 1, \ldots, K$. The unconditional density for the whole sample

© 2022 Author(s). (https://www.thegsa.in/).

is then

$$f(y_i;\boldsymbol{\xi}) = \sum_{k=1}^{K} \pi_k f_k(y_i;\boldsymbol{\theta}_k), \quad i = 1,\ldots,n,$$

where $\boldsymbol{\xi} = (\pi_1,\ldots,\pi_K,\boldsymbol{\theta}_1',\ldots,\boldsymbol{\theta}_K')'$. In this setting, the sub-populations are considered as unknown latent variables that need to be estimated together with the model parameters $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K$ of respective sub-population density functions.

Mixture of Normal Distributions

Mixture of normal (MN) distributions and their extensions play an important role in the field of finite mixtures, see the discussions in [31], [15] and [21], for example. In [15] it is shown how MN can model different distribution forms, including asymmetric and skewed distributions; see also Example 1 for illustration. It is underscored that almost any distribution form could be approximated arbitrarily close with a suitable choice of a MN. Although many alternative distribution forms could also be used, we highlight that the use of MN can be justified for several reasons. The theory of the normal distribution is very well established and many computer programs have been developed for the analyzes. Further, as noted above, they provide a moderate approximation to many situations that may further be improved by a suitable transformation of the observed data [7, 15].

Parameter estimation is generally achieved using the method of maximum likelihood, and we follow this approach here. Under this setting, the aim is to find an estimate $\hat{\boldsymbol{\xi}}$ which maximizes the log likelihood function

$$l(\boldsymbol{\xi};y_1,\ldots,y_n) = \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \pi_k f_k(y_i;\boldsymbol{\theta}_k)\right\}.$$

Once the model parameters $\boldsymbol{\xi}$ are estimated, each observation is then placed into a suitable sub-population on the basis of posterior probability

$$p_{ik}(\hat{\boldsymbol{\xi}},y_i) = \frac{\hat{\pi}_k f_k(y_i;\hat{\boldsymbol{\theta}}_k)}{\sum_{k=1}^{K}\hat{\pi}_k f_k(y_i;\hat{\boldsymbol{\theta}}_k)} = \frac{\hat{\pi}_k f_k(y_i;\hat{\boldsymbol{\theta}}_k)}{f(y_i;\hat{\boldsymbol{\xi}})}$$

by using the highest $p_{ik}(\hat{\boldsymbol{\xi}},y_i)$ with $k = 1,\ldots,K$. For maximization of the log likelihood function $l(\boldsymbol{\xi};y_1,\ldots,y_n)$ the EM algorithm [3] is often employed, especially when the assumed response variable is chosen from the exponential family of distributions. Implementations of the EM algorithm for mixtures are usually based on the joint distribution of the observed data and cluster memberships information, where cluster memberships are considered as missing data. However, our experience in real data analysis is such that for slightly more complex models, the associated optimization burden is rather unwieldy and the algorithm does not necessarily converge with all the desired number of mixtures K. We also underscore that the runs may need to be repeated with several different initial values to ensure that the algorithm is converged to the maximum point of the likelihood function.

Example 1. *The data used for this example is part of the growth data of 4223 children collected in Finland as described in [32] and [24]. For this example we focus on the 1995 birth cohort. For this cohort, we used all male's 347* $weight$ *and* $height$ *measurements at the age of 15 years, and formed the chosen (derived) response variable Body Mass Index (BMI),*

$$bmi = weight/height^2.$$

Here $weight$ *was measured in kilograms and* $height$ *in meters. This variable is often considered as an important measure when assessing overweight or obesity of adult population. Discussion of the use of BMI for adolescents and children is given in [6].*

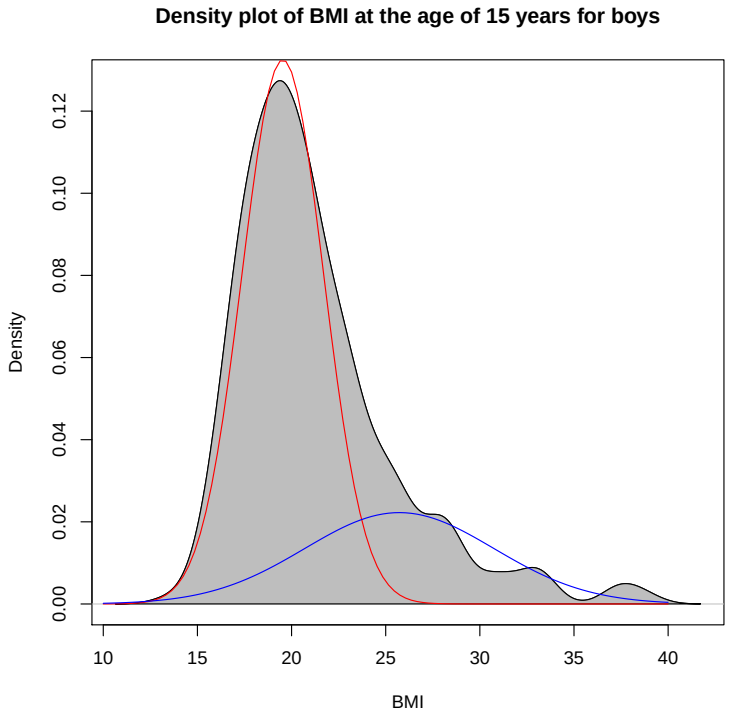


Figure 1: Density plot of BMI at the age of 15 years for boys ($n = 347$) and densities of the fitted sub-populations (red color for population 1 and blue color for population 2).

The density plot of bmi variable is shown in Figure 1. When determining the appropriate number $k = 1, \ldots, 5$ of mixture of normal components, we use the BIC criterion, and obtain the respective values: $BIC(1) = 2003.0$, $BIC(2) = 1905.7$, $BIC(3) = 1915.7$, $BIC(4) = 1931.4$ and $BIC(5) = 1944.3$. On that basis the appropriate number of mixture components is $K = 2$ and the associated parameter estimates are

$$\hat{\pi}_1 = 0.72,\ \hat{\mu}_1 = 19.54,\ \hat{\sigma}_1^2 = 4.69,\ \hat{\pi}_2 = 0.28,\ \hat{\mu}_2 = 25.72 \text{ and } \hat{\sigma}_2^2 = 25.27.$$

Clearly, the high values $\hat{\mu}_2 = 25.72$ and $\hat{\sigma}_2^2 = 25.27$ in the second sub-population indicate that some part of the males in this sub-population have relatively high mean but that this group is quite disperse. We also point out that the right tail of the density function of the first sub-population is very close to the limit $(25\ kg/m^2)$ of overweight or mild obesity for the adult population (Figure 1).

3 Mixture Regression

Whereas the previous section demonstrates mixture modelling of the observed density, we next consider finite mixtures in the context of regression modelling. The generalized mixture regression model can be formulated as

$$f(y_i; \boldsymbol{\xi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{z}_i, \boldsymbol{\alpha}_k) f_k(y_i; \boldsymbol{\theta}_{ki}),$$

where $g(\mu_{ki}) = \boldsymbol{x}_i' \boldsymbol{\beta}_k$, $\mu_{ki} = E(y_i | Z_i = k)$, g is a monotonic and differentiable link function and $\pi_k(\boldsymbol{z}_i, \boldsymbol{\alpha}_k)$ follows a multinomial logit model with (possible) concomitant variables $\boldsymbol{z}_i$ and parameters $\boldsymbol{\alpha}_k$. Then essentially the densities f_k and proportions π_k can have separate models using variables $\boldsymbol{x}_i$ and

concomitant variables $\boldsymbol{z}_i$. The following example considers modelling in a survey data using mixtures of Poisson distribution.

Example 2 (Quality of Work Life: Stress)**.** *As described in [29], the Quality of Working Life Survey is a nationwide survey conducted by Statistics Finland in the years 1977, 1984, 1990, 1997, 2003, 2008, 2013 and 2018. The survey examines the physical, mental and social working environment of employees in addition to collecting information on work, labor market status, working conditions, reconciling work and family life, occupational health and factors at the work organization level.*

In this example our focus is on stress examination in 1984 (n=4362), 1990 (n=4018), 1997 (n=2945), 2003 (n=4097), 2008 (n=4387), 2013 (n=4859) and 2018 (n=4094), where the numbers in parentheses are the corresponding sample sizes. In our study stress is measured by Stresscore which is the sum of 7 individual stress indicator variables each measured on five-point Likert scale. Three variables assess functional somatic symptoms (Headache; Palpitations or Irregular heartbeat; Stomach upset) and four variables (Fatigue, reluctance or inactivity; Difficulty getting to sleep or waking up at night; Tension, nervousness, or irritability; Feeling that everything is beyond power) measure psychological distress. In the quality of work life studies in Finland Stresscore has been used as a measure of stress, the higher the value of the score, the more stressed the person is. Related information on stress issues is given in [12].

The observed distribution of Stresscore is discrete, highly right-skewed and heterogeneous. This makes further investigation of Stresscore using statistical models quite challenging. The approach we take here is to approximate the distribution of Stresscore by a mixture of Poisson distributions. The mixture of Poisson distributions can be written as

$$f(y_i;\boldsymbol{\xi}) = \sum_{k=1}^{K} \pi_k f_k(y_i;\mu_{ki})$$

with

$$f_k(y_i;\mu_{ki}) = \exp(-\mu_{ki})\mu_{ki}^{y_i}/y_i!$$

for $k = 1,\ldots,K$ *and* $i = 1,\ldots,n$*. It is quite easily shown that*

$$\mathrm{E}(y_i) = \sum_{k=1}^{K} \pi_k\mu_{ki} \quad \textit{and} \quad \mathrm{Var}(y_i) = \mathrm{E}(y_i) + \nu_{ij},$$

where

$$\nu_{ij} = \sum_{k=1}^{K} \pi_k\mu_{ki}^2 - \left(\sum_{k=1}^{K} \pi_k\mu_{ki}\right)^2,$$

which equals zero if and only if $\mu_{1i} = \mu_{2i} = \cdots = \mu_{Ki}$*. As such, finite mixtures is one key technique to handle over-dispersion among the observed data.*

Our purpose is to perform a model-based clustering such that more detailed statistical analysis can be successfully performed in sub-populations. Due to changes in working life (e.g. digitalization, globalization or increase in part-time employment), it can be argued that working life stress may have changed (and perhaps increased) over time. Although other variables, such as, gender or employment sector may also be connected with stress measurements, but since their associations may be more nuanced, we have excluded these variables here from our example. For the Stresscores we tested the simple clustering model

$$\log(\mu_{ki}) = \boldsymbol{x}_i'\boldsymbol{\beta}_k,\ i = 1,\ldots,n,$$

where $\boldsymbol{x}_i' = (1, year_i)$*, with* $k = 1,2,3,4$*. The following* $BIC(k)$ *values were obtained:* $BIC(1) = 165547.7$*,* $BIC(2) = 159173.5$*,* $BIC(3) = 159204.2$ *and* $BIC(4) = 159235.0$*, respectively. The*

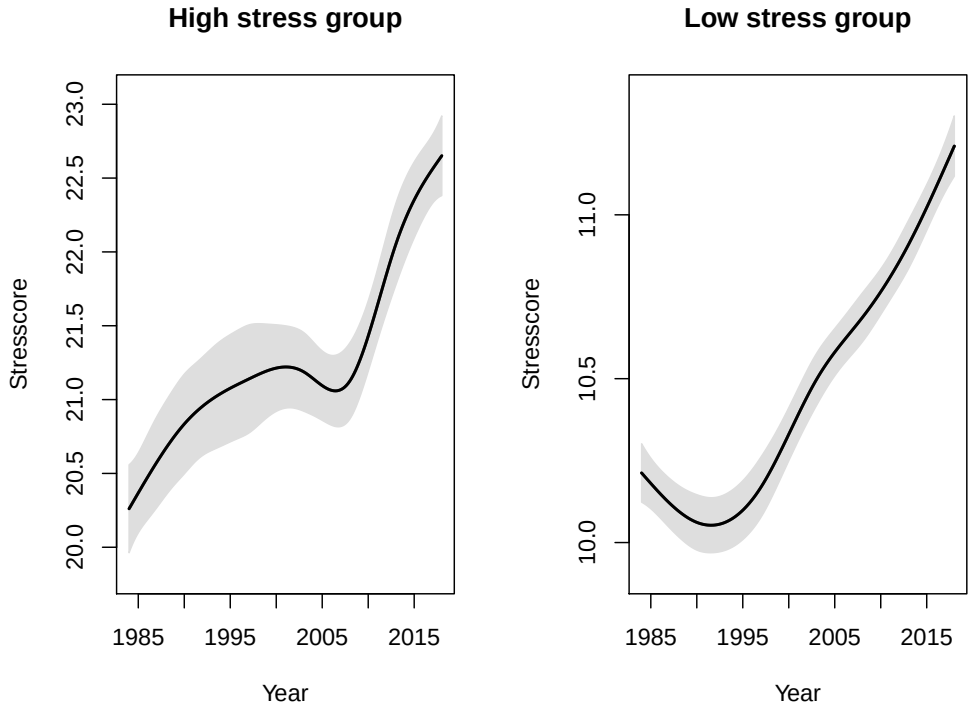


Figure 2: Smoothed mean Stresscore as a function of time in two sub-populations, where estimated proportions are $\hat{\pi}_1 = 0.13$ (High stress group) and $\hat{\pi}_2 = 0.87$ (Low stress group).

minimum $BIC(2) = 159173.5$ *indicates that two different sub-populations are identified here. The estimated mixture proportions are then* $\hat{\pi}_1 = 0.13$ *and* $\hat{\pi}_2 = 0.87$.

From the Figure 2 we note that both the magnitude and time-patterns of stress seem to be quite different for these two groups. Therefore as such, they are not directly comparable with each other with the note that they have been rising since at least 2005. In the figure, the groups have been aptly named as ($k = 1$*, High stress group) and (*$k = 2$*, Low stress group). As highlighted above, the overall stress levels in the left panel are higher than those on the right. In terms of the overall patterns, both panels are generally increasing but with dips in the left panel around 2006 and in the right panel around 1992.*

4 Mixtures of Multivariate Data

Mixture of Multivariate Normal Distributions

In the mixture of multivariate normal (MMN) distribution settings, the density of each p-dimensional $\boldsymbol{y}_i$ can be written as

$$f(\boldsymbol{y}_i; \boldsymbol{\xi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i; \boldsymbol{\theta}_k), \quad i = 1, \ldots, n,$$

where

$$f_k(\boldsymbol{y}_i; \boldsymbol{\theta}_k) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}_k|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_k) \right),$$

$\boldsymbol{\mu}_k$ is the mean vector and $\boldsymbol{\Sigma}_k$ is the variance-covariance matrix for the kth mixture. In the most general case $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the unstructured mean and covariance matrices. However, often some more parsimonious structures or models are imposed either on $\boldsymbol{\mu}_k$ or on $\boldsymbol{\Sigma}_k$ or on both.

Variable	G_1	G_2	G_3	G_4	G_5	G_6
Working life (yrs.)	30.7	10.7	24.3	28.5	5.9	1.4
Age at ret. (yrs.)	55.5	42.2	59.8	58.5	57.2	49.6
Sickness (days)	168	226	173	84	259	165
Wage (Eur.)	2941	1607	1958	2495	389	17
Pension (Eur.)	1569	793	1016	658	216	17

Table 1: Multivariate mixture normal mean estimates for disability retirees.

Example 3 (Predictors of disability retirement)**.** *The comprehensive administrative registers of Finnish statutory earnings-related pension system provides data to study disability pension retirees via multivariate mixture framework. Our study-design consists of the following five continuous variables: the amount of statutory disability pension (in Euros/month), the pre-retirement wage earnings (in Euros/month), the length of working life at the end of 2016 (in years since age 18), the total number of long-term sickness benefit days in 2005–2017, and the age at retirement. The actual data consists of a random sample of* $n = 4160$ *public sector disability pension retirees (in 2017) in cohorts born over period 1954–2000.*

In our analysis $\boldsymbol{\Sigma}_k$ *is modelled using the eigenvalue decomposition as follows (see [28])*

$$\boldsymbol{\Sigma}_k = c_k \boldsymbol{T}_k \boldsymbol{\Lambda}_{*k} \boldsymbol{T}'_k,$$

where $\boldsymbol{T}_k$ *is a matrix of eigenvectors,* $\boldsymbol{\Lambda}_{*k}$ *is a diagonal matrix of scaled eigenvalues and* c_k *is the associated scaling constant. In our larger sample, the interpretation of identified clusters is somewhat cumbersome in practice and, on the other hand, some groups may become so small that they have no practical significance. As such, based on our subjective knowledge and experience, we have chosen six groups (denoted G1-G6 here) with respective mixing proportions:* $\hat{\pi}_1 = 0.1245, \hat{\pi}_2 = 0.2558, \hat{\pi}_3 = 0.2095, \hat{\pi}_4 = 0.2131, \hat{\pi}_5 = 0.1184$ *and* $\hat{\pi}_6 = 0.0775$*. Cluster mean estimates are presented in Table 1.*

The results indicate quite meaningful groups, which are in line with real-life experience. As captured in groups G1, G3 and G4, there is a high share (55%) of disability retirees with long and stable working lives cut short, yet the pension system yields decent level of pension. There is a relatively large group G5 (11.8%) and smallest group G6 (7.7%) with weak attachment to public sector employment, and the final pension accrual remains understandably quite low. In practice these groups consists of people with private sector employment background, and they get pension also from private sector pension providers. The largest group, G2 (25.6%) shows that pension rules together with strong labor market attachment yields moderate pension security, although drawn at quite young age after long sickness periods (226 days). The group G1 with less than 12.5% of the total, underscores the strength of the earnings-related pension system: this group receives highest pension security following long career in the public sector.

Note that early disability retirement, with low wage does not automatically indicate low pension income (cf. G2), because pension system rules compensate for non-completed expected career. Furthermore, the disability pensions indicated by Table 1 are not the final pension, as small earnings-related pension (cf. G2, G4, G5 and G6) is compensated by national pension, which is an universal benefit.

Mixtures of Longitudinal Data

Longitudinal data are typically correlated and also lend themselves to mixture modelling techniques, which we address here - first modelling a single outcome and then several outcome variables.

Trajectory Analysis

Understanding and modelling heterogeneity of longitudinal data has been of lively interest to researchers in various applied fields for several decades. In this case $\boldsymbol{y}_i$ consists of longitudinal sequence of p measurements of the ith individual. Here we focus on the normal general linear model, but also other members of the exponential family are possible. As introduced and coined in [20] and [16], these are so-called *trajectory analyses*; additional details about trajectory modelling approaches are given in [17], [5] and [22]. In the following it is assumed that

$$\boldsymbol{\mu}_k = \boldsymbol{X}_i\boldsymbol{\beta}_k \quad \text{with} \quad \boldsymbol{\Sigma}_k = \sigma_k^2\boldsymbol{I}, \quad k = 1, \ldots, K,$$

where $\boldsymbol{X}_i$ may depend on one or several time-dependent variables. Note that this assumed within-cluster independence does not imply independence at the sample level. In the following two subsections, we apply this basic model and extend it also for the analysis of multivariate longitudinal data.

Let $\mathbf{y}_i = (y_{i_1}, \ldots, y_{i_T})'$, represent a sequence of repeated measurements on an individual i over T, and let $f_i(\mathbf{y}_i; \mathbf{X}_i, \boldsymbol{\xi})$ denote the marginal probability distribution of $\mathbf{y}_i$ with potential time-dependent covariates $\mathbf{X}_i$. It is assumed that $f_i(\mathbf{y}_i; \mathbf{X}_i, \boldsymbol{\xi})$ follows a mixture of K densities

$$f_i(\mathbf{y}_i; \mathbf{X}_i, \boldsymbol{\xi}) = \sum_{k=1}^{K} \pi_k f_{ik}(\mathbf{y}_i; \mathbf{X}_i, \boldsymbol{\theta}_k), \quad \sum_{k=1}^{K} \pi_k = 1 \quad \text{with} \quad \pi_k > 0,$$

where π_k is the probability of belonging to unobserved sub-group k, $f_{ik}(\mathbf{y}_i; \mathbf{X}_i, \boldsymbol{\theta}_k)$ is the density for the kth sub-group and $\boldsymbol{\theta}_k$ is a vector of component-specific parameters for density. The overall likelihood function is then

$$L(\boldsymbol{\xi}) = \prod_{i=1}^{n} f_i(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\xi}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_{ik}(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\theta}_k),$$

where n is the sample size.

Example 4 (Wage trajectories)**.** *Administrative registers of Finnish statutory pension system provide data from pensionable wage earnings of private and public sector employees and of self-employed workers. In addition to earnings, there is also background information including gender, socioeconomic groups and uses of several social security benefits. For this study a random sample ($n = 750$) of individuals was drawn from the nation-wide registers. The sample consists of cohort born in 1960, for which wages and social security benefits are collected annually for the 11-year period from 2010 to 2020.*

The observed distribution of wage earnings (yearly wages) has both excess number of zero or near zero observations and it is highly right-skewed. Our approach is to approximate the wage distribution by a mixture of normal distributions. For the log *wage earnings, we fitted a simple cubic polynomial cluster-specific model*

$$\beta_{0k} + \beta_{1k}t + \beta_{2k}t^2 + \beta_{3k}t^3,$$

where t is the age of a person. After fitting the model for $k = 1, \ldots, 5$ latent groups we obtained the following BIC values: $BIC(2) = 17805.6$, $BIC(3) = 17288.2$, $BIC(4) = 16802.8$ and $BIC(5) = 16460.9$, respectively. The BIC criterion supports the model selection with up to five mixture components k, which was the maximum number of tested components. However, with five components the group sizes would lead to the solution that is not very interesting from a practical point of view. As such, and in the line with our previous comments, we select the three-component solution; this in turn yields reasonable group sizes, over 50 individuals per group. The estimated mixture proportions for the three groups were then $\hat{\pi}_1 = 0.0749$, $\hat{\pi}_2 = 0.8508$ and $\hat{\pi}_3 = 0.0742$, respectively. From Figure 3 we can see the estimated

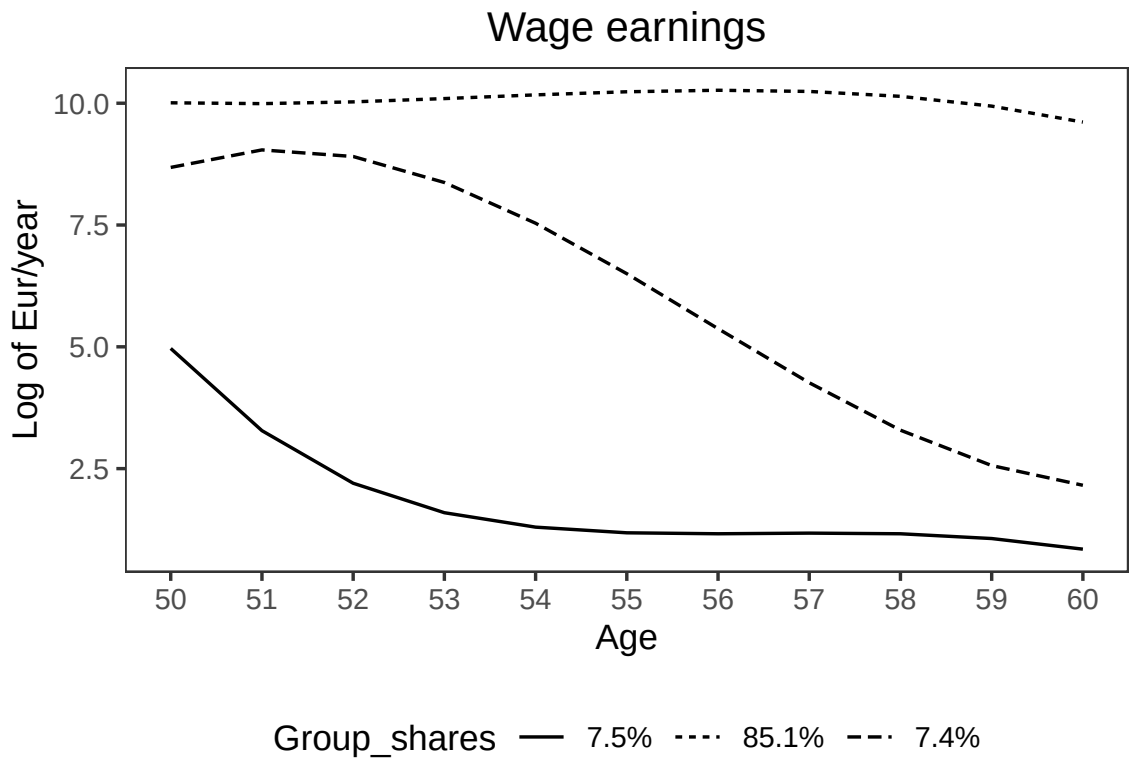


Figure 3: Mean wage earnings trajectories as a function of age in three sub-populations, where estimated proportions are: $\hat{\pi}_1 = 0.0749$ (Low wages), $\hat{\pi}_2 = 0.8508$ (Stable wages) and $\hat{\pi}_3 = 0.0742$ (Falling wages).

wage earnings trajectory models. The groups were given heuristic names Low wages ($k = 1$), Stable wages ($k = 2$) and Falling wages ($k = 3$).

We can conclude that there is a large group (85.1%) with "hump-shaped wages" and solid labor market attachment at mid-life course. Other groups indicate a more fragile labor market attachment and falling wages. Cluster membership information also gives further possibilities for more detailed statistical analysis of explanatory factors (e.g. using multinomial logistic regression) behind the observed cluster composition with socioeconomic factors, for example. Simple statistical analysis with gender showed that Low wages group and Falling wages group consists mainly of women (70%), while the respective share in Stable wages group is 80%.

Multivariate Trajectory Analysis

An interesting extension of the basic TA is the mixture analysis where it is possible to describe the interrelationship of several multiple outcomes followed over time. The multivariate model can be written as follows. Let $\mathbf{y}_i = (\mathbf{y}'_{i1}, \ldots, \mathbf{y}'_{iJ})'$, where $\mathbf{y}_{ij}$ represent a sequence of measurements of the jth independent outcome on an individual i over T, and let $h_{ij}(\mathbf{y}_{ij}|\mathbf{X}_{ij})$ denote the marginal probability distribution of the outcome $\mathbf{y}_{ij}$ with possible time dependent covariates $\mathbf{X}_{ij}$. The marginal mixture distribution of $\mathbf{y}_i$ can now be written as

$$f_i(\mathbf{y}_i|\mathbf{X}_i) = \sum_{k=1}^{K} \pi_k \left(\prod_{j=1}^{J} h_{ijk}(\mathbf{y}_{ij}|\mathbf{X}_{ij}) \right), \sum_{k=1}^{K} \pi_k = 1 \quad \text{with} \quad \pi_k > 0,$$

where π_k is the probability of belonging to sub-group k, $h_{ijk}(\mathbf{y}_{ij}|\mathbf{X}_{ij})$ is the density of jth outcome for the kth sub-group and $\boldsymbol{X}_i = \{\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iJ}\}$. Note that outcomes can now follow different distributions and explanatory variables do not need to be the same for all the outcomes. Applications of multivariate trajectory analysis can be found in [19], [8], [11], [27] and [2], for example.

Example 5 (Wage and social security benefit trajectories)**.** *As shown in Example 4 three mixture components can be identified from longitudinal wage distribution. In this example we extend the analysis by modelling wage earnings and social security benefit days jointly. Social security benefits include the following benefits: long-term unemployment spells, sickness spells, occupational rehabilitation and accident-based spells. For this example, the same small sample ($n = 750$) of individuals born in 1960 is used as in Example 3. The observed distribution of benefit days (yearly days) has both excess number of zeros and is highly left-skewed. Our approach is to approximate the distribution of wage earnings and benefit days by a multivariate mixture of (zero-truncated) normal distributions. The log-transformation is used on the yearly wages as in Example 3. From Table 2 we can see the distribution of the benefit days used in the analysis.*

Gender	Count	50%	75%	95%	99%	Mean	SD
Men	1804	0	3	206	265	29	67
Women	6446	0	21	201	262	31	65
Total	8250	0	16	202	264	31	66

Table 2: Summary of benefit days over years 2010–2020 (Counts, quantiles, means and standard deviations).

For analyzing both log wage earnings and benefit days (yearly total), we tested the third degree clustering model:

$$\beta_{0jk} + \beta_{1jk}t + \beta_{2jk}t^2 + \beta_{3jk}t^3$$

for variable j and cluster k in t (age). Testing the number of clusters with BIC (SAS, proc traj) shows that $BIC(2) = -37388.1$, $BIC(3) = -36720.2$, $BIC(4) = -36218.4$ and $BIC(5) = -35912.4$, respectively. The BIC criterion supports the model with up to five mixture components. In order to compare the results with the Example 4 we fixed the number of clusters as three. The estimated mixture proportions for the three groups are now $\hat{\pi}_1 = 0.0891$, $\hat{\pi}_2 = 0.7862$ and $\hat{\pi}_3 = 0.1245$. From Figure 4 we can see the estimated wage trajectories (left) and benefit trajectories (right) in the sample. The groups were given heuristic names: Weak labor attachment ($k = 1$), Stable labor attachment ($k = 2$) and Falling wages and unemployment ($k = 3$).

It is seen in Figure 4 that with a Stable labor attachment, the number of benefit day is the lowest, as expected. Falling wages and unemployment seems to be associated with highly increasing benefit days with age. In a Weak labor attachment, the number of benefit days would even appear to decrease with age, which may be a little surprising. When compared to clusters in Example 4 a slightly different picture is also drawn of the cluster proportions. Especially the Stable labor attachment appears to be become smaller due to increasing number of benefit days at the end of the career (0.8508 versus 0.7862). The share of Falling wages and unemployment attachment appears to become larger when number of benefit days is also taken into account (0.0749 versus 0.1245).

More detailed analysis of benefit day indicate that 81% of the sample draw sickness benefits. Highest share locates in Falling wages and unemployment group (90%) with average sum of 317 days. In Weak labor attachment group the share is 84% and the average duration 325 days. Unemployment is a persistent problem, but less common as 33% of the sample faces unemployment spells at mid-life course. Highest share locates in Falling wages and unemployment group (83%) with average sum of 757 days. In Weak

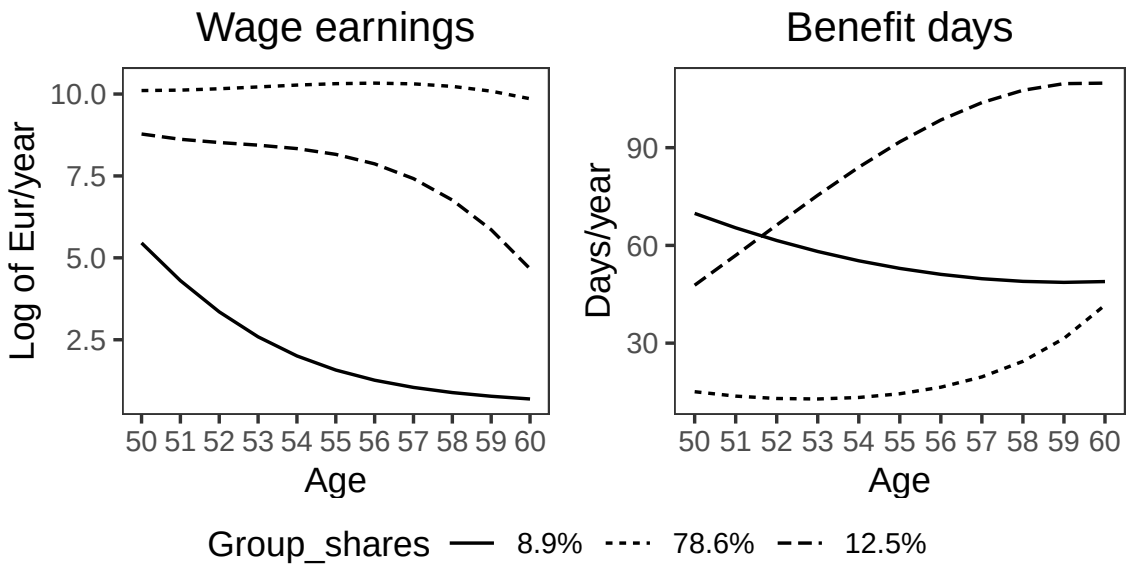


Figure 4: Mean wage earnings and social security benefits trajectories as a function of age in three sub-populations, where estimated proportions are: $\hat{\pi}_1 = 0.0891$ (Weak labor attachment), $\hat{\pi}_2 = 0.7862$ (Stable labor attachment) and $\hat{\pi}_3 = 0.1245$ (Falling wages and unemployment).

labor attachment group the share is 60% and the average duration 672 days. In Stable labor attachment group only 22% face unemployment spells and the duration is shorter (172 days) compared to other groups with a more fragile labor market attachment.

Dual Trajectory Analysis: A Conditional Model

The dual trajectory model is an extension to univariate model, outlined by [18], [16] and [9], which permits us to study the interrelationship between two related longitudinal outcomes in a conditional setup. The approach visualizes the multidimensional associations between the outcomes and it has been applied, for example, in medical sciences [1, 14, 33]. For more discussion on trajectory modelling approaches, see also [2].

The likelihood function of the conditional model stems from the assumptions of univariate trajectory model. The dual modelling includes two conceptual approaches: *general model* and *constrained model* to link two correlated trajectory outcomes. The general model includes the assumption that the two outcomes have the same number of mixture components, whereas the constrained model allows different number of components. Let $\boldsymbol{y}_1 = (y_{11}, y_{12}, y_{13}, \ldots, y_{1T_1})$ and $\boldsymbol{y}_2 = (y_{21}, y_{22}, y_{23}, \ldots, y_{2T_2})$ denote the two longitudinal outcomes to be modelled. As the conditional independence assumption (given group membership) of univariate trajectory model is maintained, the joint conditional distributions of the outcome vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ can be defined as

$$f^k(\boldsymbol{y}_1) = \prod_{t=1}^{T_1} f_t^k(y_{1t})$$

and

$$h^k(\boldsymbol{y}_2) = \prod_{t=1}^{T_2} h_t^k(y_{2t}),$$

where $f_t^k(\cdot)$ and $h_t^k(\cdot)$ are the respective PDFs (probability density functions) or PMFs (probability mass functions). Note that the two outcomes can follow different distributions, thus, for example, $f_t^k(\cdot)$ can be

a a normal PDF and $h_t^k(\cdot)$ can be a Poisson PMF.

Assume that the outcome vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are correlated but they are independently distributed conditional on the memberships in trajectory groups. Then the unconditional likelihood function for the general model is of the form

$$L = \sum_{j=1}^{J} \sum_{k=1}^{K} \pi_{jk} f^j(\boldsymbol{y}_1) h^k(\boldsymbol{y}_2),$$

where π_{jk} is the joint probability of membership in trajectory group j for $\boldsymbol{y}_1$ and trajectory group k for $\boldsymbol{y}_2$. See [16], for example.

The general model assumes J unique mixture components and developmental trajectories which describe the courses of Y_1 and Y_2. Both outcomes should be independently distributed. Therefore the likelihood function for the constrained model weighted by π_j is defined as

$$L = \sum_{j=1}^{J} \pi_j f^j(\boldsymbol{y}_1) h^j(\boldsymbol{y}_2),$$

where π_j is the shared proportion of both Y_1 and Y_2 (See also Section 4.2.2). Using similar data and measures as in Example 5, it is possible to illustrate the practical idea of dual trajectory modelling.

Example 6 (Dual analysis of Wages and Benefits)**.** *The example data includes two distinct, but related outcomes, which can be used to demonstrate the above mentioned constrained model. The analysis proceeds in two stages. In the first step a univariate trajectory analysis assuming zero-truncated normal distribution is done separately on both outcomes: wage earnings and benefit days. The second step includes counting the optimal group-solution in the joint model using the parameters (including group shares) of the univariate analyzes as starting values for the joint analysis.*

For the univariate analysis of both wage earnings and benefit days, we used the simple clustering model:

$$\mathbf{x}_i' \boldsymbol{\beta}_{jk} = \beta_{0jk} + \beta_{1jk} t.$$

A three-group solution for wages and a two-group solution for benefit days yields a tractable model, which can be used in the joint model. Although the group-specific clustering model is simpler than in Examples 4 and 5, nearly identical trajectory groups (cf. Figure 4), with similar substantial implications can be revealed with the joint model. The group shares are different compared to multivariate trajectory model. The mixing probabilities for wage are: $\hat{\pi}_1 = 0.0629$ *(Low wages),* $\hat{\pi}_2 = 0.8596$ *(Stable wages) and* $\hat{\pi}_3 = 0.0774$ *(Falling wages). The corresponding shares for benefit days are:* $\hat{\pi}_1 = 0.7638$ *(Slow increasing) and* $\hat{\pi}_2 = 0.2361$ *(High). The probabilities of group membership for univariate model and dual model can be found from Table 3. The interrelationship across trajectory groups can be presented in several ways by showing the conditional probabilities of the outcomes. Table 4 shows one way to illustrate the joint probabilities across groups.*

The benefit group-specific results indicate that there is a large group with over 95 per cent of Slow increasing group associated with Stable wages trajectory. Nearly 4% of Slow increasing group associated with Low wages trajectory. High use of social security benefits is more evenly associated to the three wage trajectories. About 56% of High benefit group is associated to Stable wage trajectory group. The respective shares for Falling wages trajectory and Low wages trajectory are 29% and 15%.

5 Concluding Remarks

It is demonstrated that methodologies presented here are quite useful in a wide variety of practical analyzes. The presented methodology here is based mainly on the normal distribution with linear modelling. However,

Wage group	1-Low wages	2-Stable wages	3-Falling wages
Univariate model	0.0617	0.8612	0.0769
Dual model	0.0629	0.8596	0.0774
Benefit group	1-Slow increasing	2-High	
Univariate model	0.7410	0.2589	
Dual model	0.7638	0.2361	

Table 3: Probabilities of group membership from univariate model and dual model.

	Benefit group	
Wage group	1-Slow increasing	2-High
1-Low wages	0.036	0.151
2-Stable wages	0.952	0.559
3-Falling wages	0.012	0.290
Total	1.00	1.00

Table 4: Probability of wage group j conditional on benefit group k ($\pi_{j|k}$).

these methods can be generalized in many respects by utilizing, for example, different distribution forms, e.g. [13], or by using non-linear, e.g. [30] or non-parametric regression modelling, e.g. [23], techniques. Modelling and revealing the relative proportions of sub-populations also offers interesting possibilities in different application situations.

6 Acknowledgements

The authors are very grateful to Professor Antti Saloniemi and Emeritus Professor Pekka Virtanen for providing data for the *Quality of Work Life* example and for commenting on the text of the example. The authors are also very grateful for the constructive comments of the anonymous referee.

References

[1] Edward D Barker et al. "Developmental trajectories of male physical violence and theft: relations to neurocognitive performance." en. In: *Archives of General Psychiatry* 64.5 (May 2007), pp. 592–599. URL: https://doi.org/10.1001/archpsyc.64.5.592.

[2] Y. Cheng. "Multivariate Finite Mixture Group-Based Trajectory Modeling With Application To Mental Health Studies." In: *PhD dissertation. University of Saskatchewan, Saskatoon, Canada.* (2021). URL: https://harvest.usask.ca/handle/10388/13665.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), pp. 1–38. ISSN: 00359246. URL: http://www.jstor.org/stable/2984875 (visited on 04/08/2022).

[4] B. Everitt and D. Hand. *Finite Mixture Distributions*. Monographs on Applied Probability and Statistics. London: Chapman and Hall, 1981. URL: https://doi.org/10.1007/978-94-009-5897-5.

[5] Nathalie M. G. Fontaine and Isaac T. Petersen. "Developmental Trajectories of Psychopathology." In: *The Wiley Handbook of Developmental Psychopathology*. Ed. by Luna C. Centifanti and David M. Williams. John Wiley & Sons, Ltd, 2017. Chap. 1, pp. 1–27. ISBN: 9781118554470. URL: https://doi.org/10.1002/9781118554470.ch1.

[6] David S. Freedman and Bettylou Sherry. "The validity of BMI as an indicator of body fatness and risk among children." In: *Pediatrics* 124.Supplement 1 (Sept. 2009), S23–S34. ISSN: 0031-4005. URL: https://doi.org/10.1542/peds.2008-3586E.

[7] Roberto G. Gutierrez et al. "Analysis of Tomato Root Initiation Using a Normal Mixture Distribution." In: *Biometrics* 51.4 (1995), pp. 1461–1468. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2533276 (visited on 04/08/2022).

[8] Hui-Chuan Hsu. "Trajectories of multimorbidity and impacts on successful aging." In: *Experimental Gerontology* 66 (2015), pp. 32–38. ISSN: 0531-5565. URL: https://doi.org/10.1016/j.exger.2015.04.005.

[9] B. Jones and D. Nagin. "Advances in group-based trajectory modeling and an SAS procedure for estimating them." In: *Sociological Methods & Research* 35.4 (2007), pp. 542–571. URL: https://doi.org/10.3390/socsci8090253.

[10] Chinubhai G Khatri. "A note on a MANOVA model applied to problems in growth curve." In: *Annals of the Institute of Statistical Mathematics* 18.1 (1966), pp. 75–86. URL: https://doi.org/10.1007/BF02869518.

[11] K Kuitto, J Salonen, and J. Helmdag. "Gender Inequalities in Early Career Trajectories and Parental Leaves: Evidence from a Nordic Welfare State." In: *Social Sciences* 8.253 (2019), pp. 1–17. URL: https://doi.org/10.3390/socsci8090253.

[12] Richard S Lazarus and Susan Folkman. *Stress, Appraisal, and Coping*. Springer publishing company, New York, NY, 1984.

[13] Sharon X Lee and Geoffrey J McLachlan. "On mixtures of skew normal and skew t-distributions." In: *Advances in Data Analysis and Classification* 7.3 (2013), pp. 241–266.

[14] X. Li et al. "The Trajectories and Correlation between Physical Limitation and Depression in Elderly Residents of Beijing, 1992–2009." In: *PLoS ONE* 7.8 (2012). URL: https://doi.org/10.1371/journal.pone.0042999.

[15] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. USA, NY: John Wiley & Sons, Inc, 2000. URL: http://dx.doi.org/10.1002/0471721182.

[16] D. Nagin. *Group-Based Modeling of Development*. Cambridge, MA: Harvard University Press, 2005.

[17] D. Nagin and C. Odgers. "Group-based trajectory modeling (nearly) two decades later." In: *Journal of Quantitative Criminology* 26.4 (2010), pp. 445–453. URL: https://doi.org/10.1007/s10940-010-9113-7.

[18] D Nagin and R. Tremblay. "Analyzing developmental trajectories of distinct but related behaviors: A group-based method." In: *Psychological Methods* 6.1 (2001), pp. 18–34. URL: https://doi.org/10.1037/1082-989X.6.1.18.

[19] D Nagin et al. "Group-based multi-trajectory modeling." In: *Statistical Methods in Medical Research* 0.0 (2016), pp. 1–9. URL: https://doi.org/10.1177/0962280216673085.

[20] Daniel S Nagin. "Analyzing developmental trajectories: a semiparametric, group-based approach." In: *Psychological methods* 4.2 (1999), pp. 139–177. URL: https://doi.org/10.1037/1082-989X.4.2.139.

[21] Shu Kay Ng, Liming Xiang, and Kelvin Kai Wing Yau. *Mixture Modelling for Medical and Health Sciences*. Chapman and Hall/CRC, 2019. URL: https://doi.org/10.1201/9780429195181.

[22] Hermine Lore Nguena Nguefack et al. "Trajectory Modelling Techniques Useful to Epidemiological Research: A Comparative Narrative Review of Approaches." In: *Clinical Epidemiology* 12 (2020), 1205–1222. URL: https://doi.org/10.2147/CLEP.S265287.

[23] Tapio Nummi et al. "A semiparametric mixture regression model for longitudinal data." In: *Journal of Statistical Theory and Practice* 12.1 (2018), pp. 12–22. URL: https://doi.org/10.1080/15598608.2017.1298062.

[24] Tapio Nummi et al. "A trajectory analysis of body mass index for Finnish children." In: *Journal of Applied Statistics* 41.7 (2014), pp. 1422–1435. URL: https://doi.org/10.1080/02664763.2013.872232.

[25] Yating Pan et al. "Growth curve mixture models with unknown covariance structures." In: *Journal of Multivariate Analysis* 188 (2022). 50th Anniversary Jubilee Edition, p. 104904. ISSN: 0047-259X. URL: https://doi.org/10.1016/j.jmva.2021.104904.

[26] Richard F. Potthoff and S. N. Roy. "A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems." In: *Biometrika* 51.3/4 (1964), pp. 313–326. ISSN: 00063444. URL: http://www.jstor.org/stable/2334137 (visited on 04/08/2022).

[27] Antti Saloniemi et al. "The diversity of transitions during early adulthood in the Finnish labour market." In: *Journal of Youth Studies* 24.7 (2021), pp. 851 –870. ISSN: 13676261. URL: https://doi.org/10.1080/13676261.2020.1776229.

[28] Luca Scrucca et al. "mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models." In: *The R Journal* 8.1 (2016), pp. 289–317. URL: https://doi.org/10.32614/RJ-2016-021.

[29] Statistics Finland. *Quality of work life survey*. https://www.stat.fi/keruu/tolo/index_en.html. Accessed: 30 Dec 2021. 2021.

[30] Tatiana V Tatarinova and Alan Schumitzky. *Nonlinear mixture models: a Bayesian approach*. World Scientific, 2014.

[31] D Michael Titterington, A F M Smith, and U E Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley series in probability and mathematical statistics. John Wiley & Sons Incorporated, 1985.

[32] Nina Vuorela. "Body Mass Index, Overweight and Obesity Among Children in Finland - A Retrospective Epidemilogical Study in Pirkanmaa District Spanning Over Four Decades." PhD thesis. School of Medicine of the University of Tampere, 2011. URL: https://urn.fi/urn:isbn:978-951-44-8446-9.

[33] X. Xiang et al. "Dual Trajectories of Social Isolation and Dementia in Older Adults: A Population-Based Longitudinal Study." In: *Journal of Aging and Health* 33.1-2 (2020), pp. 63–74. URL: https://doi.org/10.1177/0898264320953693.

[34] Jing Yu, Tapio Nummi, and Jianxin Pan. "Mixture regression for longitudinal data based on joint mean–covariance model." In: *Journal of Multivariate Analysis* 190 (2022), p. 104956. ISSN: 0047-259X. URL: https://doi.org/10.1016/j.jmva.2022.104956.

The Panel of the Reviewers in the current issue

Professor Subhash Bagui
Department of Mathematics and Statistics
University of West Florida
Pensacola, FL, USA

Professor Thomas Mathew
Department of Mathematics and Statistics
University of Maryland, Baltimore
USA

Dr. Seema Jaggi
Assistant Director General (HRD)
Agriculture Educational Division
ICAR, Krishi Anusandhan Bhavan II
New Delhi – 110012, INDIA

Professor Aditya Chakraborty
Department of Mathematics and Statistics
University of South Florida
4202 E Fowler Ave., CMC 14
Tampa, FL-33613, USA

Dr. Avinash Dharmadhikari
Pune

Professor K. Jaykumar
Senior Professor
Department of Statistics
University of Calicut
Kerala – 673635, INDIA

Professor Seng-Huat Ong
Department of Actuarial Science and Applied Statistics
UCSI University, 56000 Kuala Lumpur,
Malaysia

Professor. Manisha Pal
Department of Statistics
Calcutta University
35, B. C. Road
Kolkata – 700019, W B, INDIA

Professor. Manas Ranjan Tripathi
Associate Professor
Department of Mathematics
N I T Rourkela
Odisha, INDIA

Dr. P.S.N.V. Prasad Rao
Retired Professor
I S I Kolkata
203, B. T. Road, Kolkata - 700108
INDIA

Dr. Ganesh Dutta
Department of Statistics
B. D. College
Calcutta University,
WB, INDIA

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Memberships:

One can become the Member/Life member of the Gujarat Statistical Association by applying in the prescribed form available at the website of Gujarat Statistical Association (GSA) https://www.thegsa.in/

Subscription rates: Subscription rates for Gujarat Journal of Statistics and Data Science are as follows:

Inland	Rs. 500.00 (Inclusive of postage)
Foreign	U.S. $ 100.00 (Inclusive of postage)

Gujarat Journal of Statistics and Data Science will be available at Gujarat Statistical Association (GSA) site https://www.thegsa.in/. The members of the association, editorial board, and the authors can download the paper(s) free of cost.

Gujarat Journal of Statistics and Data Science (formerly Gujarat Statistical Review) is a peer reviewed research journal in Statistics and published twice in a year in past. Now the journal will be publishing once in a year in the beginning. The journal will publish referred original research papers, reviews, and case studied related to any branch of applied and theoretical Statistics. Book's reviews, letters to the Editor in Chief related to the article/papers of the journal are also welcome for possible publication.

Authors are encouraged to submit manuscripts in electronic form. Author should submit their manuscript as softcopy either directly to the Editor in Chief (ghosh_dkg@rediffmail.com) or through the Managing Editor (bikassinha1946@gmail.com) or Editor (amsseng@gmail.com).

How to prepare manuscript:

Author should prepare and submit their manuscripts electronically using the LaTeX package. Authors are also requested to submit Encapsulated PostScript ("eps") or .png or .jpeg or .pdf format files for figures in both electronic and non - electronic papers. Authors should ensure that laser – printed originals of these figures are of high quality and suitable for scanning. Authors are requested to submit their (1) LaTex file (.tex format) [Refer the LaTex format available at website], (2). Bibliography file (.bib format) [Refer demo.bib file available at website], and (3). pdf file of your article.

Authors are requested to download their manuscripts in the prescribed standard format available at website of Gujarat Statistical Association (GSA) https://www.thegsa.in/

Manuscript Components: Research paper must be written in English. The manuscript must be organized as following:

Title of the paper,
Author(s) name, affiliations and e-mail address,
Abstract (not more than 250 words) without reference,
Key word (not more than six),
AMS subject classification,
Complete articles which include Tables, Figure, graph (if any required),

Acknowledgement,
Conclusion,
Appendices (if required) and finally,
References.

The complete articles will be organized as appropriate number of sections, Subsections, Equations, Figures, Tables, Equations. All Figures and Tables should normally be mentioned explicitly by number and should appear in correct numerical order in the body of the text. Tables should be numbered continuously starting with 1 irrespective of their subsections etc.

The reference list and text citations should agree and be accurate. All references cited in the text must appear in the reference list, similarly, all references listed in the reference list must be cited in the text.

Acronyms and abbreviations should be spelled out the first time they are used unless they are common throughout the discipline. Avoid beginning sentences with a symbol, number, or lowercase letter.

References for research article and book should follow the following:

Research paper:

1. Ghosh, D.K. and Karmakar, P. (1988) Some series of efficiency balanced designs. Aust. Jour. Statist. 30(1), 47-51.
2. Ghosh, D.K. (1989). Construction of confounded designs for mixed factorial experiments, Jour. Statist. Planning and Inference., 23, 253- 261.

Book:

1. Das, M. N. and Giri, N. C. (1973). Design and Analysis of Experiments. Third Edition, New Age International Publishers.

It is the policy of the journal that no submission, or substantially overlapping submission, be published or be under review at another journal or conference at any time during the review process. Submission of a manuscript implies that it has been approved by all authors as well as by the responsible authorities tacitly or explicitly at the institute where the work has been carried out. The publisher and the editors will not be held legally responsible should there be any claims for compensation.

Decision after Review:

After considering the Reviewer's reports, the Editorial board members will make one of the following decisions:

(i) Accept the paper for possible publication
(ii) Request for a minor/major revision
(iii) Reject the publication of the paper

The peer-review process of the journal is a double-blind process. While submitting the revised manuscript, the authors should clearly explain point-by-point as to how the reviewer's comments have been addressed. If the authors disagree with reviewer's comments, the reason(s) should be explained. Decision of the editorial board will be final. Those papers recommended by two reviewers will be accepted for publication. The peer review process may be completed within 4 months.

Gujarat Statistical Association Body

Executive Committee

President (I/C)	Dr N D Shah
Vice President	Dr D K Ghosh
Secretary	Dr Chirag Trivedi
Joint Secretary	Dr R D Patel
Treasurer	Dr (Mrs.) C D Bhavsar

Members

Dr Rakesh Srivastav	Dr Parag Shah
Dr Ashok Shanubhogue	Dr H D Budhbhatti
Dr Aarti Rajguru	Dr L Murlidharan
Dr Sudhir Joshi	Dr A J Patel
Dr Mayuri Pandya	Dr Rajashree Mengore

Dr B B Jani (Editor: Sankhya Vignan)

Co-opted Members

Dr J B Shah	Dr Mrs. Kavita Dave

Invited Members

Dr Arnab Laha	Prof (Smt.) Subha Rani

Gujarat Journal of Statistics and Data Science

1.	Place of Publication:	Gujarat Statistical Association C/0 Department of Statistics Gujarat University, Navrangpura Ahmedabad – 380009, Gujarat, India
2.	Periodicity of Publication:	Annually
3.	Printer's Name:	Gujarat Statistical Association
4.	Address:	Gujarat Statistical Association C/0 Department of Statistics Gujarat University, Navrangpura Ahmedabad – 380009, Gujarat, India
5.	Publisher's Name:	Dilip Kumar Ghosh
6.	Nationality:	Indian
7.	Editor in Chief's Name:	Dilip Kumar Ghosh
8.	Owner's Name:	Gujarat Statistical Association

I, Dilip Kumar Ghosh declare that the particulars provided above are true to the best of my knowledge and belief.

Date: July, 31, 2022

Sd/- Dilip Kumar Ghosh

Signature of Publisher