Austrian Journal of Statistics

AUSTRIAN STATISTICAL SOCIETY

Volume 46, Number 2, 2017

ISSN: 1026597X, Vienna, Austria

Österreichische Zeitschrift für Statistik

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Austrian Journal of Statistics; Information and Instructions

GENERAL NOTES

The Austrian Journal of Statistics is an open-access journal with a long history and is published approximately quarterly by the Austrian Statistical Society. Its general objective is to promote and extend the use of statistical methods in all kind of theoretical and applied disciplines. Special emphasis is on methods and results in official statistics.

Original papers and review articles in English will be published in the Austrian Journal of Statistics if judged consistently with these general aims. All papers will be refereed. Special topics sections will appear from time to time. Each section will have as a theme a specialized area of statistical application, theory, or methodology. Technical notes or problems for considerations under Shorter Communications are also invited. A special section is reserved for book reviews.

All published manuscripts are available at

http://www.ajs.or.at

(old editions can be found at http://www.stat.tugraz.at/AJS/Editions.html)

Members of the Austrian Statistical Society receive a copy of the Journal free of charge. To apply for a membership, see the website of the Society. Articles will also be made available through the web.

PEER REVIEW PROCESS

All contributions will be anonymously refereed which is also for the authors in order to getting positive feedback and constructive suggestions from other qualified people. Editor and referees must trust that the contribution has not been submitted for publication at the same time at another place. It is fair that the submitting author notifies if an earlier version has already been submitted somewhere before. Manuscripts stay with the publisher and referees. The refereeing and publishing in the Austrian Journal of Statistics is free of charge. The publisher, the Austrian Statistical Society requires a grant of copyright from authors in order to effectively publish and distribute this journal worldwide.

OPEN ACCESS POLICY

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge.

ONLINE SUBMISSIONS

Already have a Username/Password for Austrian Journal of Statistics?
Go to http://www.ajs.or.at/index.php/ajs/login

Need a Username/Password?
Go to http://www.ajs.or.at/index.php/ajs/user/register

Registration and login are required to submit items and to check the status of current submissions.

AUTHOR GUIDELINES

The original LaTeX-file guidelinesAJS.zip (available online) should be used as a template for the setting up of a text to be submitted in computer readable form. Other formats are only accepted rarely.

SUBMISSION PREPARATION CHECKLIST

- The submission has not been previously published, nor is it before another journal for consideration (or an explanation has been provided in Comments to the Editor).
- The submission file is preferable in LaTeXfile format provided by the journal.
- All illustrations, figures, and tables are placed within the text at the appropriate points, rather than at the end.
- The text adheres to the stylistic and bibliographic requirements outlined in the Author Guidelines, which is found in About the Journal.

COPYRIGHT NOTICE

Manuscripts should be unpublished and not be under consideration for publication elsewhere. By submitting an article, the author(s) certify that the article is their original work, that they have the right to submit the article for publication, and that they can grant the above license.

Austrian Journal of Statistics

Volume 46, Number 2, 2017

Editor-in-chief: Matthias TEMPL

http://www.ajs.or.at

Published by the AUSTRIAN STATISTICAL SOCIETY

http://www.osg.or.at

Österreichische Zeitschrift für Statistik

Jahrgang 46, Heft 2, 2017

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Impressum

Editor:	Matthias Templ, Zurich University of Applied Sciences
Editorial Board:	Peter Filzmoser, Vienna University of Technology Herwig Friedl, TU Graz Bernd Genser, University of Konstanz Peter Hackl, Vienna University of Economics, Austria Wolfgang Huf, Medical University of Vienna, Center for Medical Physics and Biomedical Engineering Alexander Kowarik, Statistics Austria, Austria Johannes Ledolter, Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien & Management Sciences, University of Iowa Werner Mueller, Johannes Kepler University Linz, Austria Josef Richter, University of Innsbruck Milan Stehlik, Department of Applied Statistics, Johannes Kepler University, Linz, Austria Wolfgang Trutschnig, Department for Mathematics, University of Salzburg Regina Tüchler, Austrian Federal Economic Chamber, Austria Helga Wagner, Johannes Kepler University Walter Zwirner, University of Calgary, Canada
Book Reviews:	Ernst Stadlober, Graz University of Technology
Printed by	Statistics Austria, A-1110 Vienna

Published approximately quarterly by the Austrian Statistical Society, C/o Statistik Austria
Guglgasse 13, A–1110 Wien

Contents

Editorial

This volume include five scientific papers.

The first contribution deals with the problem of estimating two ordered normal means when variances are unknown and unequal. The authors provided R code, which has been made available at http://www.ajs.or.at.

Multidimensional scaling is a very famous topic in statistics and the authors of the second paper shows and solve some problems related to initial configurations and local minima. The corresponding R code have been made available at the AJS website as well.

The third paper introduces new methods to estimate the parameters of a famous distribution in life-time modelling under a censoring scheme.

The authors of the fourth contribution analyses trade flows structures using compositional data analysis methods. The authors provide conviencing examples.

Transport statistics is in the main topic of the last contribution. The authors improve distance estimations in order to provide more reliable estimates of transport volumes in Austria.

Matthias Templ
(Editor-in-Chief)

Zurich University of Applied Sciences
Rosenstrasse 3
CH–8400 Winterthur, Switzerland
E-mail: matthias.templ@gmail.com

Winterthur, 12. Januar 2017

Austrian Journal of Statistics
February 2017, Volume 46, 3–17.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i2.446

Estimation of Order Restricted Normal Means when the Variances Are Unknown and Unequal

Najmeh Pedram
Persian Gulf University

Abouzar Bazyari
Persian Gulf University

Abstract

In the present paper, two normal distributions with parameters μ_i and σ_i^2 where there is an order restriction on the means when the variances are unknown and unequal are considered. Under the squared error loss function, a necessary and sufficient condition for the plug-in estimators to improve upon the unrestricted maximum likelihood estimators uniformly is given. Also under the modified Pitman nearness criterion; a class of estimators is considered that reduce to the estimators of a common mean when the unbiased estimators violate the order restriction. It is shown that the most critical case for uniform improvement with regard to the unbiased estimators is the one when two means are equal. To illustrate the results, two numerical examples are presented.

Keywords: maximum likelihood estimator, order restriction, Pitman nearness, squared error loss function.

1. Introduction

Let X_{ij} be the j th observation of the i th population and be mutually independently distributed as $N(\mu_i, \sigma_i^2)$, $i = 1, 2, \cdots, k$, $j = 1, 2, \cdots, n_i$, where the order restriction on the unknown parameters μ_i, $i = 1, 2, \cdots, k$ is defined as

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k. \tag{1}$$

We consider the following squared error loss function of the estimators of μ_i, $i = 1, 2, \cdots, k$,

$$L(\mu_i, \hat{\mu}_i) = (\hat{\mu}_i - \mu_i)^2. \tag{2}$$

Then the risk is given by

$$R(\mu_i, \hat{\mu}_i) = E[L(\mu_i, \hat{\mu}_i)]. \tag{3}$$

The estimator $\hat{\mu}_i^*$ uniformly improves upon the estimator $\hat{\mu}_i^{**}$, $i = 1, 2, \cdots, k$, under the squared error loss function (2) if and only if

$$R(\mu_i, \hat{\mu}_i^*) \leq R(\mu_i, \hat{\mu}_i^{**}),$$

for all $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$. Note that $\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$ is the unrestricted maximum likelihood estimator of μ_i and is distributed as $N(\mu_i, \sigma_i^2/n_i)$.

Later, many authors, including Brown and Cohen (1974), Khatri and Shah (1974) and Bhattacharya *et al.* (1980) have given a class of improved estimators of the form

$$\hat{\mu}(\gamma) = \gamma \bar{X}_1 + (1-\gamma)\bar{X}_2,$$

where γ is a function of s_1^2 and s_2^2 .
Under the order restriction (1), the maximum likelihood estimator of μ_i is given by

$$\underset{t\geq i}{min}\underset{s\leq i}{max}\frac{\sum_{j=s}^{t} n_j \bar{X}_j/\sigma_j^2}{\sum_{j=s}^{t} n_j/\sigma_j^2}. \tag{4}$$

A possible alternative criterion to evaluate the goodness of estimators, mean squared error (MSE), is Pitmam nearness.
For comparing two estimators $T_i, (i = 1, 2)$ of a single parameter θ, Pitman (1937) proposed the following criterion: T_1 is said to be closer (better) than T_2 if

$$PN_\theta(T_1, T_2) = P\{|T_2 - \theta| > |T_1 - \theta|\} > \frac{1}{2}, \tag{5}$$

for all θ. The probability $PN_\theta(T_1, T_2)$ in (5) is usually called the Pitman nearness of T_1 relative to T_2.
Lee (1981) showed that the estimator (4) uniformly improves upon $\bar{X}_i$. Rao (1980) discussed the similarities and differences of MSE and PMN. Kelly (1989) strengthened Lee (1981)'s result and showed that (4) universally dominates $\bar{X}_i$.
Nayak (1990) defined modified Pitman nearness of an estimator T_1 of θ relative to the other estimator T_2 by

$$MPN_\theta(T_1, T_2) = P\{|T_1 - \theta| < |T_2 - \theta||T_1 \neq T_2\}. \tag{6}$$

If $MPN_\theta(T_1, T_2) \geq 1/2$ for any parameter value, then T_1 is said to be closer to θ than T_2. Gupta and Singh (1992) have applied modified Pitman nearness to the estimation of ordered means of two normal population with common variance and have shown that MLE is closer than the unbiased estimator.
Hwang and Peddada (1994) showed that under arbitary order restriction on μ_i's, (4) universally dominates $\bar{X}_i$ to estimate μ_i if μ_i is a node and proposed estimation procedure also for nonnodal means. (μ_i is said to be a node if, for any j, it is known that either $\mu_j \leq \mu_i$ or $\mu_i \leq \mu_j$).

In this paper, we consider the estimation of two normal means when they are subject to the order restriction

$$\mu_1 \leq \mu_2, \tag{7}$$

and $\sigma_i^2, i = 1, 2$ are unknown and possibly unequal. If σ_i^2's are known, from (7) the restricted maximum likelihood estimators of μ_i's are given by

$$\hat{\mu}_1^* = min\left(\bar{X}_1, \frac{\frac{n_1}{\sigma_1^2}\bar{X}_1 + \frac{n_2}{\sigma_2^2}\bar{X}_2}{\frac{n_1}{\sigma_1^2} + \frac{n_2}{\sigma_2^2}}\right), \tag{8}$$

and

$$\hat{\mu}_2^* = max\left(\bar{X}_2, \frac{\frac{n_1}{\sigma_1^2}\bar{X}_1 + \frac{n_2}{\sigma_2^2}\bar{X}_2}{\frac{n_1}{\sigma_1^2} + \frac{n_2}{\sigma_2^2}}\right). \tag{9}$$

But, if we suppose that σ_i^2's are unknown, so we estimate σ_i^2 by $s_i^2 = \sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/(n_i - 1)$ and replace σ_i^2 with s_i^2 in (8) and (9) and obtain the plug-in estimators as follows

$$\hat{\mu}_1 = min\left(\bar{X}_1, \frac{\frac{n_1}{s_1^2}\bar{X}_1 + \frac{n_2}{s_2^2}\bar{X}_2}{\frac{n_1}{s_1^2} + \frac{n_2}{s_2^2}}\right), \tag{10}$$

and

$$\hat{\mu}_2 = max\left(\bar{X}_2, \frac{\frac{n_1}{s_1^2}\bar{X}_1 + \frac{n_2}{s_2^2}\bar{X}_2}{\frac{n_1}{s_1^2} + \frac{n_2}{s_2^2}}\right). \tag{11}$$

? proposed another type of plug-in estimators $\tilde{\mu}_i'$ obtained by replacing s_i^2 with $\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/n_i$ given in (10) and (11) and proposed results when $\mu_1 = \mu_2$. Chang and Shinozaki (2012) have considered a class of estimators of μ_i, $i = 1, 2$ of the form

$$\hat{\mu}_1(\gamma) = min\{\bar{X}_1, \gamma\bar{X}_1 + (1-\gamma)\bar{X}_2\}, \tag{12}$$

and

$$\hat{\mu}_2(\gamma) = max\{\bar{X}_2, \gamma\bar{X}_1 + (1-\gamma)\bar{X}_2\}. \tag{13}$$

Bazyari (2015) considered the estimators of the monotonic mean vectors for two dimensional normal distributions and compare those with the unrestricted maximum likelihood estimators under two different cases. One case is that covariance matrices are known, the other one is that covariance matrices are completely unknown and unequal.

To illustrate the usefulness of order restriction we have taken the following examples.

Example 1. An experiment was conducted to evaluate the effect of exercise on the age at which a child starts to walk. Let Y denote the age (in months) at which a child starts to walk, the data on Y are given in Tabel 1. (The original experiment consisted of another treatment, however, here we consider only two treatments for simplicity.)

Table 1: The age at which a child first walks.

Treatment (i)	Age (in months)						n_i	$\bar{y}_i$	μ_i
1	9.00	9.50	9.75	10.00	13.00	9.50	6	10.125	μ_1
2	11.00	10.00	10.00	11.75	10.50	15.00	6	11.375	μ_2

The first treatment group received a special walking exercise for 12 minutes per day beginning at age 1 week and lasting 7 weeks. The second group received daily exercises but not the special walking exercises. For treatment i (i=1, 2), let μ_i be the mean age (in months) at which a child starts to walk. However, suppose that the researcher was prepared to assume that the walking exercises would not have negative effect of increasing the mean age at which a child starts to walk, and it was desired that this additional information be incorporated to improve on the statistical analysis. In this case, we have that $\mu_1 \leq \mu_2$.

Example 2. An experiment was done to evaluate the discrimination of men from women. Four psychological test scores, pictorial absurdities, paper form board, tool recognition and vocabulary were given to two different groups of 32 men and 32 women. The data on men and women are for 32 applicants for a professional position requiring 10 or more years of successful schooling (the completion of second-year high school in Ontario, up to a University degree). The 4 tests were each scored according to the number of questions answered successfully. The mean vectors of the two samples are

$$\bar{\boldsymbol{X}}_{\mathbf{1}} = (15.7, 15.91, 27.19, 22.75)', \quad \bar{\boldsymbol{X}}_{\mathbf{2}} = (12.34, 13.91, 16.66, 21.94)'.$$

Let $\boldsymbol{\mu_i} = (\mu_{i1}, \mu_{i2}, \mu_{i3}, \mu_{i4})'$ for $i = 1, 2$, denotes the mean variable for i^{th} group, where μ_{ij}, $j = 1, 2, 3, 4$, denotes the j^{th} element of mean vector $\boldsymbol{\mu_i}$. Suppose that the researcher is prepared to assume that the elements of mean vectors of two populations are subject to the order restriction

$$\mu_{21} < \mu_{11}, \quad \mu_{22} < \mu_{12}, \quad \mu_{23} < \mu_{13}, \quad \mu_{24} < \mu_{14}\ .$$

The rest of this paper is organized as follows. In section 2, we show that the plug-in estimator μ_i uniformly improves upon $\bar{X}_i$ if and only if for all σ_i^2's the risk difference $\bar{X}_i$ and $\hat{\mu}_i$ is nonnegative when $\mu_1 = \mu_2$. In section 3, with respect to modified Pitman nearness, we show that the estimator $\hat{\mu}_i(\gamma)$ improves upon $\bar{X}_i$ uniformly imporoves upon the $\bar{X}_i$ if and

only if $MPN_{\mu_i}(\hat{\mu}_i(\gamma), \bar{X}_i) \geq \frac{1}{2}$ when $\mu_1 = \mu_2$, which is the most critical case for uniform improvement. Further, it is shown that $\hat{\mu}_i(\gamma)$ improves upon $\bar{X}_i$ if and only if $\hat{\mu}(\gamma)$ improves upon $\bar{X}_i$ for the same γ in estimating a common mean. To illustrate the results two numerical examples are presented in section 4. Concluding remarks are given in section 5.

2. Uniformly improved estimator of each of two ordered normal means

We show that the most critical case for $\hat{\mu}_i$ to improve upon $\bar{X}_i$ if and only uniformly is the one when $\mu_1 = \mu_2$.

Theorem 2.1. *The plug-in estimator $\hat{\mu}_1$ uniformly improves upon the unrestricted maximum likelihood estimator $\bar{X}_1$ if and only if for all σ_i^2's the risk of $\hat{\mu}_1$ is not larger than that of $\bar{X}_1$ when $\mu_1 = \mu_2$.*

Proof. Putting $\gamma = \frac{(\frac{n_1}{s_1^2})}{(\frac{n_1}{s_1^2}+\frac{n_2}{s_2^2})}$, $\hat{\mu}_1$ is expressed as

$$\hat{\mu}_1 = min(\bar{X}_1, \gamma\bar{X}_1 + (1-\gamma)\bar{X}_2), \tag{14}$$

and we calculate the risk difference of $\bar{X}_1$ and $\hat{\mu}_1$ as

$$\begin{aligned} &R(\mu_1, \bar{X}_1) - R(\mu_1, \hat{\mu}_1) \\ &= E[(\bar{X}_1 - \mu_1)^2 - \{\gamma(\bar{X}_1 - \mu_1) + (1-\gamma)(\bar{X}_2 - \mu_1)\}^2]I_{\bar{X}_1 \geq \bar{X}_2}, \end{aligned} \tag{15}$$

where I_d denotes the indicator function of the set satisfying the condition d. Making the transformations

$$Z_1 = \bar{X}_1 - \mu_1 \quad , \quad Z_2 = \bar{X}_2 - \mu_1, \tag{16}$$

Z_1 and Z_1 are mutually independently distributed as $N(0, \tau_1^2)$ and $N(\mu, \tau_2^2)$, respectively, where $\mu = \mu_2 - \mu_1 \geq 0, \tau_1^2 = \sigma_1^2/n_1$ and $\tau_2^2 = \sigma_2^2/n_2$. Noting that Z_1, Z_2 and γ are mutually independent, we have from (16)

$$\begin{aligned} R(\mu_1, \bar{X}_1) - R(\mu_1, \hat{\mu}_1) &= E[Z_1^2 - \{\gamma Z_1 + (1-\gamma)Z_2\}^2]I_{Z_1 \geq Z_2} \\ &= 2E[\gamma(1-\gamma)]E[(Z_1 - Z_2)Z_1 I_{Z_1 \geq Z_2}] \\ &+ E[(1-\gamma)^2]E[(Z_1^2 - Z_2^2)I_{Z_1 \geq Z_2}]. \end{aligned} \tag{17}$$

Making the further transformations

$$Y_1 = Z_1 - Z_2 \quad , \quad Y_2 = Z_1 + (\frac{\tau_1^2}{\tau_2^2})Z_2, \tag{18}$$

note that Y_1 and Y_2 are mutually independently distributed as $N(-\mu, \tau_1^2+\tau_2^2)$ and $N\left((\frac{\tau_1^2}{\tau_2^2})\mu, \tau_1^2 + (\frac{\tau_1^4}{\tau_2^2})\right)$, respectively, and

$$Z_1 = \frac{Y_1(\frac{\tau_1^2}{\tau_2^2}) + Y_2}{1 + \frac{\tau_1^2}{\tau_2^2}} \quad , \quad Z_2 = \frac{Y_2 - Y_1}{1 + (\frac{\tau_1^2}{\tau_2^2})}.$$

Then, we have

$$\begin{aligned}
&E[(Z_1 - Z_2)Z_1 I_{Z_1 \geq Z_2}] \\
&= E\left[\frac{\tau_2^2 Y_1\left(Y_1(\frac{\tau_1^2}{\tau_2^2}) + Y_2\right)}{\tau_1^2 + \tau_2^2} I_{Y_1 \geq 0}\right] \\
&= E\left[\frac{\tau_2^2\left(Y_1^2(\frac{\tau_1^2}{\tau_2^2}) + Y_1 E[E[Y_2]]\right)}{\tau_1^2 + \tau_2^2} I_{Y_1 \geq 0}\right] \\
&= E\left[\frac{\tau_1^2(Y_1^2 + \mu Y_1)}{\tau_1^2 + \tau_2^2} I_{Y_1 \geq 0}\right] \\
&\geq \frac{\tau_1^2}{\tau_1^2 + \tau_2^2} E[Y_1^2 I_{Y_1 \geq 0}], \qquad (19)
\end{aligned}$$

and

$$\begin{aligned}
&E[(Z_1^2 - Z_2^2)I_{Z_1 \geq Z_2}] \\
&= E\left[\left(\frac{\tau_2^2}{\tau_1^2 + \tau_2^2}\right)^2 \left[Y_1^2\left(\frac{(\tau_1^2)^2 - (\tau_2^2)^2}{(\tau_2^2)^2}\right) + 2Y_1 Y_2\left(\frac{\tau_1^2 + \tau_2^2}{\tau_2^2}\right)\right] I_{Y_1 \geq 0}\right] \\
&= E\left[\left(\frac{\tau_2^2}{\tau_1^2 + \tau_2^2}\right)^2 \left[Y_1^2\left(\frac{(\tau_1^2)^2 - (\tau_2^2)^2}{(\tau_2^2)^2}\right) + 2Y_1 E[E[Y_2]]\left(\frac{\tau_1^2 + \tau_2^2}{\tau_2^2}\right)\right] I_{Y_1 \geq 0}\right] \\
&= E\left[\frac{(\tau_1^2 - \tau_2^2)Y_1^2 + 2\tau_1^2 \mu Y_1}{\tau_1^2 + \tau_2^2} I_{Y_1 \geq 0}\right] \\
&\geq \frac{\tau_1^2 - \tau_2^2}{\tau_1^2 + \tau_2^2} E[Y_1^2 I_{Y_1 \geq 0}], \qquad (20)
\end{aligned}$$

with equalities for $\mu = 0$ and strict inequalities for $\mu > 0$. Thus we have from (17), (19) and (20)

$$\begin{aligned}
&R(\mu_1, \bar{X}_1) - R(\mu_1, \hat{\mu}_1) \\
&\geq \frac{E[Y_1^2 I_{Y_1 \geq 0}]}{\tau_1^2 + \tau_2^2}\{2\tau_1^2 E[\gamma(1-\gamma)] + (\tau_1^2 - \tau_2^2)E[(1-\gamma)^2]\} \\
&= \frac{E[Y_1^2 I_{Y_1 \geq 0}]}{\tau_1^2 + \tau_2^2}[\tau_1^2 - \{\tau_1^2 E[\gamma^2] + \tau_2^2 E[(1-\gamma)^2]\}] \\
&= \frac{E[Y_1^2 I_{Y_1 \geq 0}]}{E_{\mu_1 = \mu_2}[Y_1^2 I_{Y_1 \geq 0}]}\{R_{\mu_1 = \mu_2}(\mu_1, \bar{X}_1) - R_{\mu_1 = \mu_2}(\mu_1, \hat{\mu}_1)\}, \qquad (21)
\end{aligned}$$

with equality for $\mu = 0$ and strict inequality for $\mu > 0$. Thus, we have shown that $\hat{\mu}_1$ unifromly improves upon $\bar{X}_1$ if and only if for all σ_i^2's the risk difference is not positive when $\mu_1 = \mu_2$, which is the most critical case for uniform improvement. This completes the proof. □

Regarding the improved estimation of μ_2, we have a similar result as follows.

Corollary 2.2. *The plug-in estimator $\hat{\mu}_2$ uniformly improves upon the unrestricted maximum likelihood estimator $\bar{X}_2$ if and only if for all σ_i^2's the risk of $\hat{\mu}_2$ is not larger than that of $\bar{X}_2$ when $\mu_1 = \mu_2$.*

Proof. Since $\mu_1 \leq \mu_2$ can be written as $-\mu_2 \leq -\mu_1$, the result follows directly from theorem (2.1). □

3. Pitman dominates of new plug-in estimators

In this section, we consider estimators of μ_i of the form (12) and (13) and compare them with unbiased estimator $\bar{X}_i$. We first show that for the case when γ is a function of s_1^2 and s_2^2 the most critical case for $\hat{\mu}_i(\gamma)$ to be closer to μ_i than $\bar{X}_i$ is the one when $\mu_1 = \mu_2$. Further, it is shown that $\hat{\mu}_i(\gamma)$ improves upon $\bar{X}_i$ if and only if $\hat{\mu}(\gamma)$ dominates $\bar{X}_i$ in the estimation problem of a common mean.

Theorem 3.1. *Suppose that* $0 \le \gamma \le 1$ *is a function of* s_1^2 *and* s_2^2. *Then*

a)$MPN_{\mu_i}(\hat{\mu}_i(\gamma), \bar{X}_i) \ge \frac{1}{2}$ *for all* $\mu_1 \le \mu_2$ *and for all* σ_1^2 *and* σ_2^2 *if and only if for all* σ_1^2 *and* σ_2^2, $MPN_{\mu_i}(\hat{\mu}_i(\gamma), \bar{X}_i) \ge \frac{1}{2}$ *when* $\mu_1 = \mu_2$.

b)$MPN_{\mu_i}(\hat{\mu}_i(\gamma), \bar{X}_i) \ge \frac{1}{2}$ *for all* $\mu_1 \le \mu_2$ *and for all* σ_1^2 *and* σ_2^2 *if and only if for all* σ_1^2 *and* σ_2^2, $PN_{\mu}(\hat{\mu}_i(\gamma), \bar{X}_i) \ge 1/2$ *to estimate* μ *when* $\mu_1 = \mu_2 = \mu$.

Proof. We need only to give a proof for the case of μ_1.
a) Since $\hat{\mu}_1(\gamma) \ne \bar{X}_1$ if and only if $\bar{X}_2 < \bar{X}_1$ and $\gamma < 1$, we have

$$\begin{aligned}
&MPN_{\mu_1}(\hat{\mu}_1(\gamma), \bar{X}_1) \\
&= P\{|\hat{\mu}_1(\gamma) - \mu_1| < |\bar{X}_1 - \mu_1||\hat{\mu}_1(\gamma) \ne \bar{X}_1\} \\
&= P\{|\gamma\bar{X}_1 + (1-\gamma)\bar{X}_2 - \mu_1| < |\bar{X}_1 - \mu_1||\bar{X}_2 < \bar{X}_1, \gamma < 1\} \\
&= P\{-(\gamma\bar{X}_1 + (1-\gamma)\bar{X}_2 - \mu_1) < (\bar{X}_1 - \mu_1)|\bar{X}_2 < \bar{X}_1, \gamma < 1\} \\
&= P\{\bar{X}_1 - \mu_1 + \gamma\bar{X}_1 + \bar{X}_2 - \gamma\bar{X}_2 - \mu_2 > 0|\bar{X}_2 < \bar{X}_1, \gamma < 1\} \\
&= P\{\bar{X}_1 - \mu_1 + \gamma\bar{X}_1 - \gamma\mu_1 + \bar{X}_2 - \mu_1 - \gamma\bar{X}_2 + \gamma\mu_1 > 0|\bar{X}_2 - \mu_1 < \bar{X}_1 - \mu_1, \gamma < 1\} \\
&= P\{(1+\gamma)Z_1 + (1-\gamma)Z_2 > 0|Z_2 < Z_1, \gamma < 1\}, \qquad (22)
\end{aligned}$$

where $Z_1 = \bar{X}_1 - \mu_1$ and $Z_2 = \bar{X}_2 - \mu_1$ are distributed as $N(0, \tau_1^2)$ and $N(\mu, \tau_2^2)$ respectively, $\mu = \mu_1 - \mu_2$ and $\tau_i^2 = \sigma_i^2/n_i$. Now, we consider the conditional probability

$$P\{0 < (1+\gamma)Z_1 + (1-\gamma)Z_2|Z_2 < Z_1, s_1^2, s_2^2\} \equiv f(\mu),$$

as a function of μ. We need only to show that $f(0) \le f(\mu)$. Putting $d = (1+\gamma)/(1-\gamma))$, we define the sets

$$\begin{aligned}
&A = \{(z_1, z_2)|z_2 \le z_1, -dz_1 \le z_2\}, \quad B = \{(z_1, z_2)|z_2 \le z_1, -dz_1 > z_2\}, \\
&A_1 = \{(z_1, z_2)|z_2 \le z_1, z_2 \ge 0\}, \quad \textit{and} \quad A_2 = \{(z_1, z_2)| - dz_1 \le z_2, z_2 < 0\}.
\end{aligned}$$

Since A_1 and A_2 are disjoint and $A = A_1 \cup A_2$, we have

$$\begin{aligned}
f(\mu) - f(0) &= \frac{P_\mu(A)}{P_\mu(A) + P_\mu(B)} - \frac{P_0(A)}{P_0(A) + P_0(B)} \\
&= \frac{\{P_\mu(A_1)P_0(B) - P_0(A_1)P_\mu(B)\} + \{P_\mu(A_2)P_0(B) - P_0(A_2)P_\mu(B)\}}{\{P_\mu(A) + P_\mu(B)\} \times \{P_0(A) + P_0(B)\}}.
\end{aligned}$$

We first show that $\{P_\mu(A_1)P_0(B) - P_0(A_1)P_\mu(B)\} > 0$ for $\mu > 0$. For that purpose, we note that

$$\begin{aligned}
P_\mu(B) &= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi\tau_2^2}} exp\{-\frac{(z_2-\mu)^2}{2\tau_2^2}\} \int_{z_2}^{-z_2/d} \frac{1}{\tau_1}\phi(z_1/\tau_1)dz_1dz_2 \\
&< exp\{-\frac{\mu^2}{2\tau_2^2}\} \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi\tau_2^2}} exp\{-\frac{z_2^2}{2\tau_2^2}\} \int_{z_2}^{-z_2/d} \frac{1}{\tau_1}\phi(z_1/\tau_1)dz_1dz_2 \\
&= exp\{-\frac{\mu^2}{2\tau_2^2}\}P_0(B). \qquad (23)
\end{aligned}$$

Similarly, we have

$$P_\mu(A_1) = \int_0^\infty \frac{1}{\sqrt{2\pi\tau_2^2}} exp\{-\frac{(z_2-\mu)^2}{2\tau_2^2}\} \int_{z_2}^\infty \frac{1}{\tau_1}\phi(z_1/\tau_1)dz_1dz_2$$
$$> exp\{-\frac{\mu^2}{2\tau_2^2}\}P_0(A_1). \tag{24}$$

From (23) and (24), we see that $\{P_\mu(A_1)P_0(B) - P_0(A_1)P_\mu(B)\} > 0$.
Next, we show that $\{P_\mu(A_2)P_0(B) - P_0(A_2)P_\mu(B)\} > 0$ for $\mu > 0$. We express $P_\mu(A_2)$ as

$$P_\mu(A_2) = \int_{-\infty}^0 \frac{1}{\sqrt{2\pi\tau_2^2}} exp\{-\frac{(z_2-\mu)^2}{2\tau_2^2}\} \int_{-z_2/d}^\infty \frac{1}{\tau_1}\phi(z_1/\tau_1)dz_1dz_2$$
$$= P_\mu\{Z_2 < 0\}E_\mu[g(Z_2)|Z_2 < 0],$$

where $g(z_2) = \int_{-z_2/d}^\infty \phi(z_1/\tau_1)/\tau_1 dz_1$. Since $g(z_2)$ is an increasing function and the conditional distribution of $Z_2 < 0$ is stochastically smallest when $\mu = 0$, we have for $\mu > 0$

$$P_\mu(A_2) > P_\mu\{Z_2 < 0\}E_0[g(Z_2)|Z_2 < 0] = P_0\{A_2\}P_\mu\{Z_2 < 0\}/P_0\{Z_2\}. \tag{25}$$

Similarly, since $h(z_2) = \int_{z_2}^{-z_2/d} \phi(z_1/\tau_1)/\tau_1 dz_1$ is a decreasing function, we have

$$P_\mu(B) = \int_{-\infty}^0 \frac{1}{\sqrt{2\pi\tau_2^2}} exp\{-\frac{(z_2-\mu)^2}{2\tau_2^2}\} \int_{z_2}^{-z_2/d} \frac{1}{\tau_1}\phi(z_1/\tau_1)dz_1dz_2$$
$$< P_\mu\{Z_2 < 0\}E_\mu[h(Z_2)|Z_2 < 0]$$
$$= P_0(B)P_\mu\{Z_2 < 0\}/P_0\{Z_2 < 0\}. \tag{26}$$

From (25) and (26), we have $\{P_\mu(A_2)P_0(B) - P_0(A_2)P_\mu(B)\} > 0$ and we have shown that $f(\mu) > f(0)$ for $\mu > 0$.

b)In the estimation problem of a common mean, as is stated in Kubokawa (1989) and according to the formula (26), $\hat{\mu}(\gamma)$ is closer to μ than $\bar{X}_1$ if and only if

$$P\{(1-\gamma)(U_2-U_1)^2 + 2U_1(U_2-U_1) \le 0\} \ge \frac{1}{2}, \tag{27}$$

where $U_i = \bar{X}_i - \mu$, $i = 1, 2$. Since

$$(1-\gamma)(U_2-U_1)^2 + 2U_1(U_2-U_1) = (U_2-U_1)\{(1+\gamma)U_1 + (1-\gamma)U_2\}, \tag{28}$$

the left-hand side of (27) is expressed as

$$P\{(1-\gamma)(U_2-U_1)^2 + 2U_1(U_2-U_1) \le 0\}$$
$$= P\{U_2 \ge U_1\}P\{(1+\gamma)U_1 + (1-\gamma)U_2 < 0|U_2 \ge U_1\}$$
$$+ P\{U_2 < U_1\}P\{(1+\gamma)U_1 + (1-\gamma)U_2 > 0|U_2 < U_1\}.$$

We notice that

$$P\{(1+\gamma)U_1 + (1-\gamma)U_2 < 0|U_2 \ge U_1\}$$
$$= P\{(1+\gamma)U_1 + (1-\gamma)U_2 > 0|U_2 < U_1\}.$$

Since U_1 and U_2 are symmetrically distributed about the origin, thus

$$P\{U_2 \ge U_1\} = P\{U_2 < U_1\} = \frac{1}{2}. \tag{29}$$

We see that the left-hand side of (26) is equal to

$$P\{(1+\gamma)U_1 + (1-\gamma)U_2 > 0|U_2 < U_1\},$$

which is $MPN_{\mu_1}(\hat{\mu}_1(\gamma), \bar{X}_1)$ given by (22) for the case $\mu_1 = \mu_2$. Therefore, we see from (a) that $MPN_{\mu_1}(\hat{\mu}_1(\gamma), \bar{X}_1)) \ge \frac{1}{2}$ for all $\mu_1 \le \mu_2$ and for all σ_i^2, $i = 1, 2$ if and only if $PN_\mu(\hat{\mu}(\gamma), \bar{X}_i) \ge \frac{1}{2}$ for all μ and for all σ_i^2, $i = 1, 2$. We complete the proof. □

Remark 3.2. In the estimation problem of a common mean, Kubokawa (1989) has given a sufficient condition on sample sizes n_1 and n_2 for $\hat{\mu}(\gamma)$ to be closer to μ than $\bar{X}_i$ for some specified class of γ .

Remark 3.3. We should mention about the general case when γ is a function of $s_i^2, i = 1, 2$ and $(\bar{X}_1 - \bar{X}_2)^2$. We first consider the case when we estimate μ_1 and suppose that $\hat{\mu_1}(\gamma_0)$ is closer to μ_1 than $\bar{X}_1$, where γ_0 is a function of s_i^2 and possibly $(\bar{X}_1 - \bar{X}_2)^2$. For any γ satisfying $\gamma_0 \leq \gamma < 1$ if $\gamma_0 < 1$, $\hat{\mu}_1(\gamma)$ is closer to μ_1 than $\bar{X}_1$. This is seen from (22), since (22) is true even when γ depends on $(\bar{X}_1 - \bar{X}_2)^2$ and (22) is an increasing function of γ.

4. Examples

In this section, to illustrate the results the following numerical examples are presented.

Example 3. Consider two univariate normal distributions, when they are subject to the order restriction $\mu_1 \leq \mu_2$. Six different cases are considered here. We simulate the values of random samples $X_{11}, X_{12}, \cdots, X_{1n_1}$, from the univariate distributions $N(\mu_{1r}, s_{1r})$ with means μ_{1r}, $r = a, b, c$, and known variances s_{1r} respectively. Also the values of random samples $X_{21}, X_{22}, \cdots, X_{2n_2}$, from the univariate normal distributions $N(\mu_{2r}, s_{2r})$ with means μ_{2r}, $r = a, b, c$, and known variances s_{2r}, respectively. In each simulation, the process of computation is repeated 10000 times to get an estimate of sample means $\bar{X}_1$ and $\bar{X}_2$, isotonic estimators of means, i.e. $\hat{\mu}_1$ and $\hat{\mu}_1$ by (12) and (13), and the risk difference $RD_{\bar{X}_1,\hat{\mu}_1} = R(\mu_1, \bar{X}_1) - R(\mu_1, \hat{\mu}_1)$ and $RD_{\bar{X}_2,\hat{\mu}_2} = R(\mu_2, \bar{X}_2) - R(\mu_2, \hat{\mu}_2)$. For differente values of sample sizes and $r = a, b, c$ the results are given in Table 2. From the Table 2, it is completely clear that $\mu_{1a} \leq \mu_{2a}$, $\mu_{1b} \leq \mu_{2b}$ and $\mu_{1c} \leq \mu_{2c}$ and in case 2 (r=b)$[n_1 = 10, n_2 = 15, \mu_1 = 4\mu_2 = 4, s_1 = 2, s_2 = 3]$ and in case 1 (r=a) $[n_1 = 20, n_2 = 25, \mu_1 = 4\mu_2 = 4, s_1 = 5, s_2 = 6]$, the isotonic regression $\hat{\mu}_1$ uniformly has the smaller risk than the unrestricted maximum likelihood estimator, $\bar{X}_1$ and the isotonic regression $\hat{\mu}_2$ uniformly has the smaller risk than the unrestricted maximum likelihood estimator, $\bar{X}_2$, respectively. But in other cases the isotonic regression estimator $\hat{\mu}_1$ uniformly has not the smaller risk than the unrestricted maximum likelihood estimator, $\bar{X}_1$ and the isotonic regression estimator $\hat{\mu}_2$ uniformly has not the smaller risk than the unrestricted maximum likelihood estimator, $\bar{X}_2$, since $RD_{\bar{X}_1,\hat{\mu}_1} < 0$ and $RD_{\bar{X}_2,\hat{\mu}_2} < 0$, respectively. Figure 1 shows the risk deifference $RD_{\bar{X}_1,\hat{\mu}_1} = R(\mu_1, \bar{X}_1) - R(\mu_1, \hat{\mu}_1)$ as a function of $\mu_1 = \mu_2$, where $\mu = \mu_{2r} - \mu_{1r}$, for different values of r. Also, figure 2 shows the risk deifference $RD_{\bar{X}_2,\hat{\mu}_2} = R(\mu_2, \bar{X}_2) - R(\mu_2, \hat{\mu}_2)$ as a function of $\mu_1 = \mu_2$, where $\mu = \mu_{2r} - \mu_{1r}$, for different values of r.

Table 2: Simulation from two univariate normal distributions: the values of risks difference $\hat{\mu}_1$ and $\hat{\mu}_2$.

	Sample sizes	$N(\mu_{1r}, s_{1r})$	$N(\mu_{2r}, s_{2r})$	$RD_{\bar{X}_1,\hat{\mu}_1}$	$RD_{\bar{X}_2,\hat{\mu}_2}$
$case1(r = a)$	$n_1 = 10$	$\mu_{1a} = 3$	$\mu_{2a} = 4$	1.179	-0.235
	$n_2 = 15$	$s_{1a} = 4$	$s_{2a} = 5$		
$case2(r = b)$	$n_1 = 10$	$\mu_{1b} = 4$	$\mu_{2b} = 4$	0.011	0.127
	$n_2 = 15$	$s_{1b} = 2$	$s_{2b} = 3$		
$case3(r = c)$	$n_1 = 10$	$\mu_{1c} = 3$	$\mu_{2c} = 3$	-0.139	-0.110
	$n_2 = 10$	$s_{1c} = 5$	$s_{2c} = 6$		
$case1(r = a)$	$n_1 = 20$	$\mu_{1a} = 4$	$\mu_{2a} = 4$	0.048	0.013
	$n_2 = 25$	$s_{1a} = 5$	$s_{2a} = 6$		
$case2(r = b)$	$n_1 = 20$	$\mu_{1b} = 5$	$\mu_{2b} = 5$	-0.019	-0.067
	$n_2 = 25$	$s_{1b} = 4$	$s_{2b} = 6$		
$case3(r = c)$	$n_1 = 20$	$\mu_{1c} = 7$	$\mu_{2c} = 7$	-0.037	-0.068
	$n_2 = 20$	$s_{1c} = 6$	$s_{2c} = 7$		

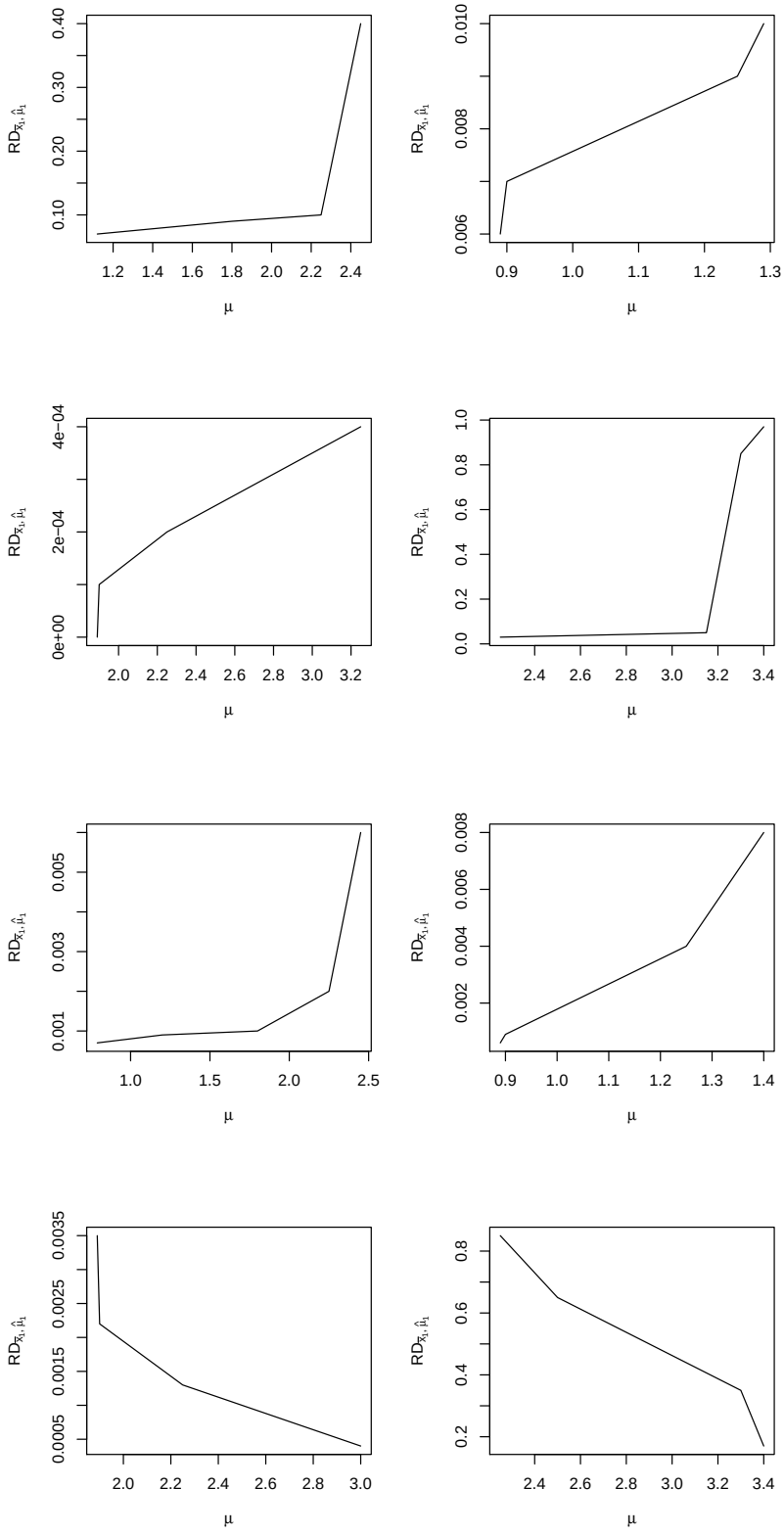


Figure 1: Risk difference $RD_{\bar{X}_1,\hat{\mu}_1} = R(\mu_1, \bar{X}_1) - R(\mu_1, \hat{\mu}_1)$.

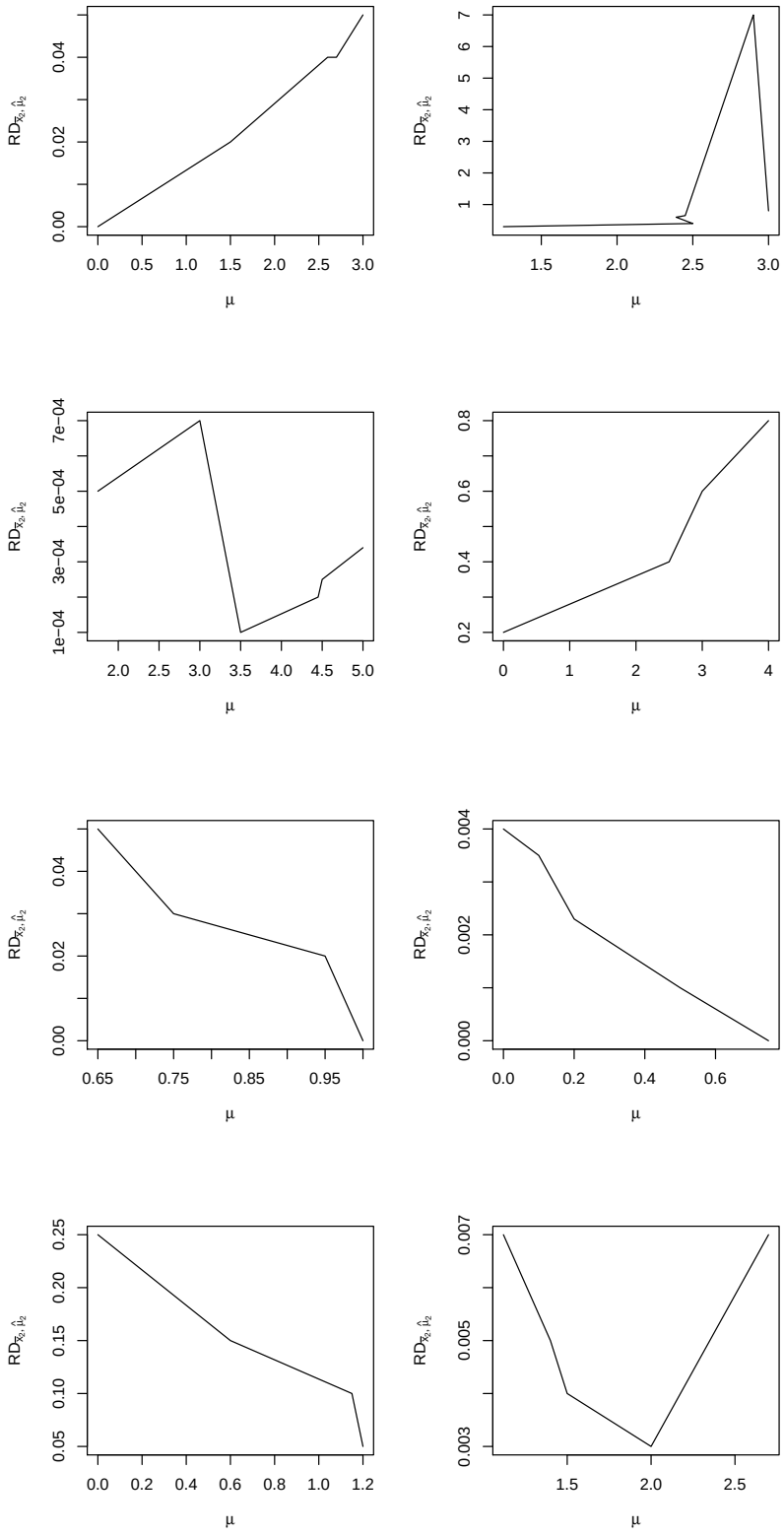


Figure 2: Risk difference $RD_{\bar{X}_2,\hat{\mu}_2} = R(\mu_2, \bar{X}_2) - R(\mu_2, \hat{\mu}_2)$.

Example 4. Consider two univariate normal distributions, when they are subject to the order restriction $\mu_1 \leq \mu_2$. Six different cases are considered here. We simulate the values of random samples $X_{11}, X_{12}, \cdots, X_{1n_1}$, from the univariate distributions $N(\mu_{1r}, s_{1r})$ with means μ_{1r}, $r = a, b, c$, and known variances s_{1r} respectively. Also the values of random samples $X_{21}, X_{22}, \cdots, X_{2n_2}$, from the univariate normal distributions $N(\mu_{2r}, s_{2r})$ with means μ_{2r}, $r = a, b, c$, and known variances s_{2r}, respectively. In each simulation, the process of computation is repeated 10000 times to get an estimate of sample means $\bar{X}_1$ and $\bar{X}_2$, isotonic estimators of means, i.e. $\hat{\mu}_1$ and $\hat{\mu}_1$ by (12) and (13), and $MPN_{\mu_i}(\hat{\mu}_i(\gamma), \bar{X}_i) \geq \frac{1}{2}$. For differente values of sample sizes and $r = a, b, c$ the results are given in Table 3. From the Table 3, it is completely clear that $\mu_{1a} \leq \mu_{2a}$, $\mu_{1b} \leq \mu_{2b}$ and $\mu_{1c} \leq \mu_{2c}$, and modified Pitman nearness of $(\bar{X}_1, \mu_1)$ is greater than $\frac{1}{2}$ for all of cases. Also, the modified Pitman nearness of $(\bar{X}_2, \mu_2)$ is greater than $\frac{1}{2}$ in cases 1,2,3 and 5. But in cases 4 and 6, the modified Pitman nearness of $(\bar{X}_2, \mu_2)$ is not greater than $\frac{1}{2}$. Figure 3 shows $MPN_{\mu_1} = MPN_{\mu_1}(\hat{\mu}_1(\gamma), \bar{X}_1)$ as a function of $\mu_1 = \mu_2$, where $\mu = \mu_{2r} - \mu_{1r}$, for different values of r. Also, figure 4 shows $MPN_{\mu_2} = MPN_{\mu_2}(\hat{\mu}_2(\gamma), \bar{X}_2)$ as a function of $\mu_1 = \mu_2$, where $\mu = \mu_{2r} - \mu_{1r}$, for different values of r.

Table 3: Simulation from two univariate normal distributions: the values of $MPN_{\mu_1} = MPN_{\mu_1}(\hat{\mu}_1(\gamma), \bar{X}_1)$ and $MPN_{\mu_2} = MPN_{\mu_2}(\hat{\mu}_2(\gamma), \bar{X}_2)$.

	Sample sizes	$N(\mu_{1r}, s_{1r})$	$N(\mu_{2r}, s_{2r})$	MPN_{μ_1}	MPN_{μ_2}
$case1(r = a)$	$n_1 = 15$	$\mu_{1a} = 6$	$\mu_{2a} = 7$	0.208	0.568
	$n_2 = 15$	$s_{1a} = 4$	$s_{2a} = 5$		
$case2(r = b)$	$n_1 = 10$	$\mu_{1b} = 4$	$\mu_{2b} = 5$	0.252	0.580
	$n_2 = 20$	$s_{1b} = 5$	$s_{2b} = 7$		
$case3(r = c)$	$n_1 = 15$	$\mu_{1c} = 3$	$\mu_{2c} = 3$	0.303	0.618
	$n_2 = 20$	$s_{1c} = 5$	$s_{2c} = 7$		
$case1(r = a)$	$n_1 = 20$	$\mu_{1a} = 9$	$\mu_{2a} = 9$	0.428	0.379
	$n_2 = 25$	$s_{1a} = 7$	$s_{2a} = 4$		
$case2(r = b)$	$n_1 = 20$	$\mu_{1b} = 6$	$\mu_{2b} = 7$	0.188	0.549
	$n_2 = 20$	$s_{1b} = 4$	$s_{2b} = 5$		
$case3(r = c)$	$n_1 = 15$	$\mu_{1c} = 5$	$\mu_{2c} = 7$	0.074	0.372
	$n_2 = 20$	$s_{1c} = 3$	$s_{2c} = 6$		

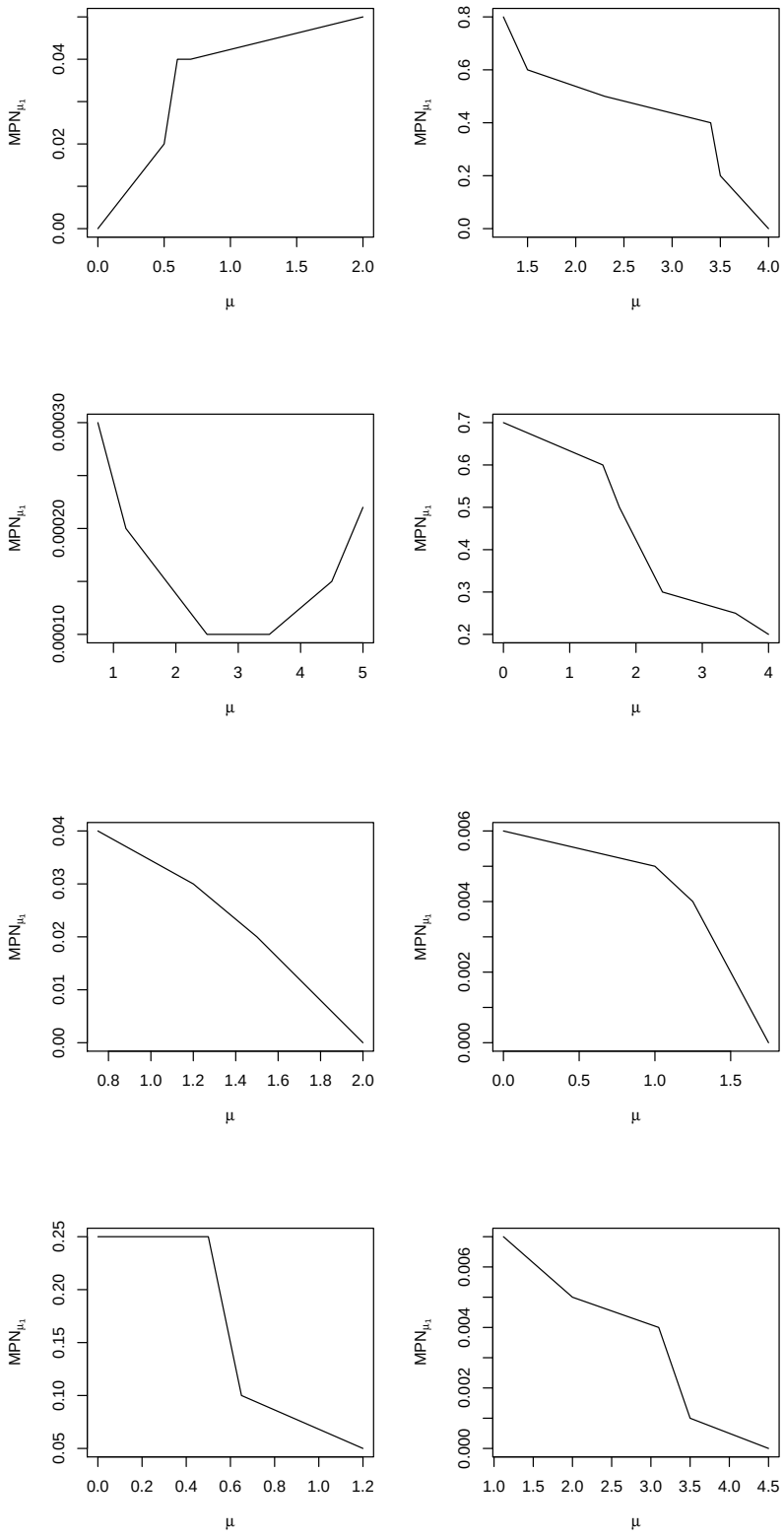


Figure 3: $MPN_{\mu_1} = MPN_{\mu_1}(\hat{\mu}_1(\gamma), \bar{X}_1)$.

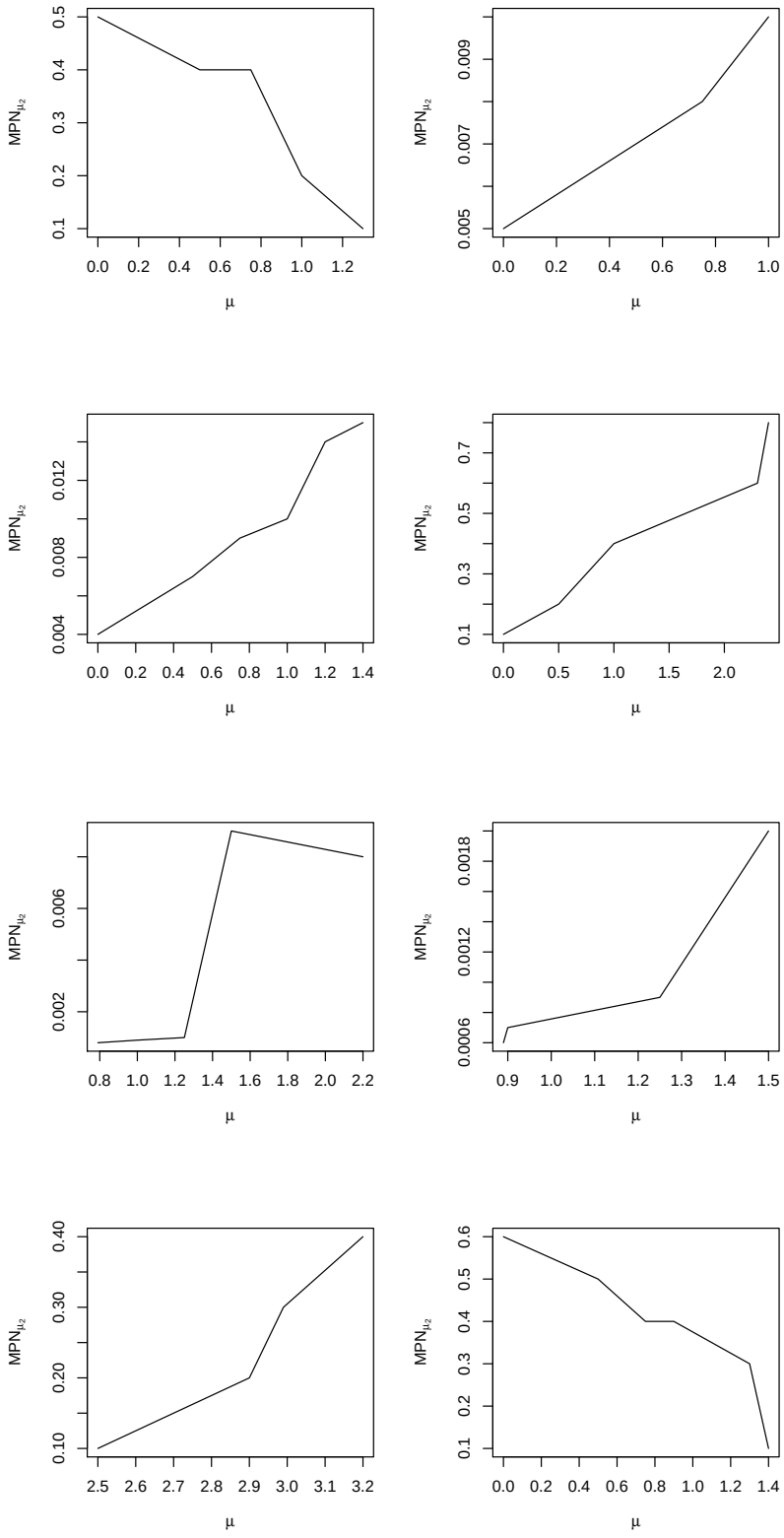


Figure 4: $MPN_{\mu_2} = MPN_{\mu_2}(\hat{\mu}_2(\gamma), \bar{X}_2)$.

5. Conclusion

In this paper, we have deal with the problem of estimating two ordered normal means under the squared error loss function when the variances are unknown and unequal. We showed that the plug-in estimator $\hat{\mu}_1$ uniformly improves upon the unrestricted maximum likelihood estimator $\bar{X}_1$ if and only if for all σ_i^2, the risk of $\hat{\mu}_1$ is not larger than that of $\bar{X}_1$ when $\mu_1 = \mu_2$, and showed that the plug-in estimator $\hat{\mu}_2$ uniformly improves upon the unrestricted maximum likelihood estimator $\bar{X}_2$ if and only if for all σ_i^2, the risk of $\hat{\mu}_2$ is not larger than that of $\bar{X}_2$ when $\mu_1 = \mu_2$. Also, under modified Pitman nearness criterion when the order restriction on variances is not present, it is shown that the most critical case for $\hat{\mu}_i(\gamma)$ to improve upon $\bar{X}_i$ is the one when $\mu_1 = \mu_2$ and that the problem of improving upon $\bar{X}_i$ reduces to the one of a common mean. Also, two numerical examples presented to illustrate the results. In example 1, the data simulated from different bivariate normal distributions. We showed that, in two cases, the isotonic regression estimators uniformly have the smaller risk than the unrestricted maximum likelihood estimator since the risk differences are positive and in the other cases, the isotonic regression estimators uniformly have the smaller risk than the unrestricted maximum likelihood estimator since the risk differences are negative. In example 2, the data simulated from different bivariate normal distributions. We showed that the modified Pitman nearness of $(\bar{X}_1, \mu_1)$ is greater than $\frac{1}{2}$ for all of cases. But, the modified Pitman nearness of $(\bar{X}_2, \mu_2)$ is greater than $\frac{1}{2}$ for some cases.

Acknowledgement

The authors would like to thank the editor and the anonymous referees for their helpful and constructive comments.

References

Bazyari A (2015). "On the Properties of Estimates of Monotonic Mean Vectors for Multivariate Normal Distributions." *Journal of Statistical Theory and Applications*, **14**(1), 89–106.

Bhattacharya CG, *et al.* (1980). "Estimation of a Common Mean and Recovery of Interblock Information." *The Annals of Statistics*, **8**(1), 205–211.

Brown L, Cohen A (1974). "Point and Confidence Estimation of a Common Mean and Recovery of Interblock Information." *The Annals of Statistics*, pp. 963–976.

Chang YT, Shinozaki N (2012). "Estimation of Ordered Means of Two Normal Distributions with Ordered Variances." *Journal of Mathematics and System Science*, **2**(1).

Gupta RD, Singh H (1992). "Pitman Nearness Comparisons of Estimates of Two Ordered Normal Means." *Australian Journal of Statistics*, **34**(3), 407–414.

Hwang JG, Peddada SD (1994). "Confidence Interval Estimation Subject to Order Restrictions." *The Annals of Statistics*, pp. 67–93.

Kelly RE (1989). "Stochastic Reduction of Loss in Estimating Normal Means by Isotonic Regression." *The Annals of Statistics*, pp. 937–940.

Khatri C, Shah K (1974). "Estimation of Location Parameters from Two Linear Models under Normality." *Communications in Statistics-Theory and Methods*, **3**(7), 647–663.

Kubokawa T (1989). "Closer Estimators of a Common Mean in the Sense of Pitman." *Annals of the Institute of Statistical Mathematics*, **41**(3), 477–484.

Lee CIC (1981). "The Quadratic Loss of Isotonic Regression under Normality." *The Annals of Statistics*, pp. 686–688.

Nayak TK (1990). "Estimation of Location and Scale Parameters using Generalized Pitman Nearness Criterion." *Journal of Statistical Planning and Inference*, **24**(2), 259–268.

Pitman EJ (1937). "The "closest" Estimates of Statistical Parameters." **33**(02), 212–222.

Rao CR (1980). *Some Comments on the Minimum Mean Square Error as a Criterion of Estimation.*

Affiliation:

Najmeh Pedram and Abouzar Bazyari
Department of Statistics, Persian Gulf University, Bushehr, Iran
E-mail: ab_bazyari@yahoo.com

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 46 *Submitted:* 2016-03-28
February 2017 *Accepted:* 2016-08-01

Austrian Journal of Statistics
February 2017, Volume 46, 19–32.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i2.561

The Choice of Initial Configurations in Multidimensional Scaling: Local Minima, Fit, and Interpretability

Ingwer Borg
WWU Münster

Patrick Mair
Harvard University

Abstract

Multidimensional scaling (MDS) algorithms can easily end up in local minima, depending on the starting configuration. This is particularly true for 2-dimensional ordinal MDS. A simulation study shows that there can be many local minima that all have an excellent model fit (i.e., small Stress) even if they do not recover a known latent configuration very well, and even if they differ substantially among each other. MDS programs give the user only one supposedly Stress-optimal solution. We here present a procedure for analyzing all MDS solutions resulting from using a variety of different starting configurations. The solutions are compared in terms of fit and configurational similarity. This allows the MDS user to identify different types of solutions with acceptable Stress, if they exist, and then pick the one that is best interpretable.

Keywords: MDS, initial configuration, local minima, procrustes, SMACOF, R.

1. Introduction

Multidimensional scaling (MDS) is a statistical method that optimally maps proximity data on pairs of objects (i.e., data expressing the similarity or the dissimilarity of pairs of objects) into distances among points in a multidimensional space. MDS is used for exploring or testing the structure of proximity data. There are many variants of MDS (see Borg and Groenen 2005; Cox and Cox 2000). In applied research, two-dimensional ordinal MDS is probably the most popular model. Here, the proximity data—converted first to dissimilarity indices δ_{ij} in case the proximities are similarity measures—are optimally mapped into Euclidean distances $d_{ij}(\mathbf{X})$ among points of a two-dimensional Euclidean space (with the coordinate matrix $\mathbf{X}$). The order of the distances corresponds to the order of the data, and ties in the data can be broken in the distances ("primary approach to ties").

The fit of an MDS model to the data is measured by the raw Stress coefficient

$$\text{Stress} = \sum_{i<j} e_{ij}^2 = \sum_{i<j} w_{ij} \left(f(\delta_{ij}) - d_{ij}(\mathbf{X})\right)^2, \tag{1}$$

where w_{ij} are non-negative fixed weights set by the user to weight the importance of error

(e.g., to handle missing data by setting $w_{ij} = 0$ if δ_{ij} is missing and $w_{ij} = 1$ otherwise). The fitted distances are defined by

$$d_{ij}(\mathbf{X}) = \left(\sum_{a=1}^{m} |x_{ia} - x_{ja}|^p \right)^{1/p}, \tag{2}$$

with $p = 2$ in the Euclidean case. The $f(\delta_{ij})$ are *disparities*, that is, dissimilarities optimally re-scaled (within the bounds set by the scale level assigned to the data) so that they approximate the distances as closely as possible. Expressed more technically, disparities are computed by a regression of the dissimilarities onto the distances so that $f(\delta_{ij}) = \widehat{d}_{ij}$.

Since Equation (1) is minimized over both $\mathbf{X}$ and $\widehat{d}_{ij}$, an obvious but trivial solution is choosing $\mathbf{X} = \mathbf{0}$ and all $\widehat{d}_{ij} = 0$. To avoid this solution, Equation (1) needs to be normalized which can be achieved by dividing by the sum of the squared distances. Doing so and taking the square root gives the usual *Stress.1* loss function of MDS:

$$\text{Stress.1} = \sqrt{\sum_{i<j} w_{ij} \left(\widehat{d}_{ij} - d_{ij}(\mathbf{X}) \right)^2 / \sum_{i<j} w_{ij} d_{ij}^2(\mathbf{X})}. \tag{3}$$

This normalization also has the advantage that the Stress value does not depend on the magnitude of configuration $\mathbf{X}$.

An optimal MDS solution is found by using iterative optimization algorithms. They all start with some initial configuration (IC), and then repeatedly move its points in space in small steps until the Stress has converged to a minimum. Modern algorithms guarantee convergence, but the response surface is bumpy and the optimization can easily end up in a local and not the global minimum (Groenen 1993). The choice of the initial configuration can be crucial, because the various local minima sometimes differ radically even if they represent the data almost equally well in terms of overall Stress. For most data, many local optima exist in most MDS models, in particular when using ordinal MDS and when the dimensionality of the MDS solution is low. Suboptimal local minima are particularly likely to occur in case of one-dimensional MDS, where standard MDS programs almost never find the global minimum (Mair and De Leeuw 2015). Conversely, the greater the dimensionality of the MDS space, the smaller the risk for suboptimal local minima (Borg, Groenen, and Mair 2013, p. 61ff.).

The various local minima may differ substantially even if they represent the data almost equally well in terms of overall Stress. Confirmatory MDS (Bentler and Weeks 1978; De Leeuw and Heiser 1980) is designed to check this. It puts additional theory-based restrictions onto the MDS solutions. For instance, allowing the user to specify the MDS dimensions (i.e., the columns of $\mathbf{X}$) with values that must remain fixed up to linear transformations, or requiring that all points of the MDS solution are located on a perfect circle. The MDS algorithm may still find configurations that satisfy these particular models with Stress values that are hardly worse than those produced by substantively blind exploratory MDS. Borg and Lingoes (1980) provide striking examples of cases where minimal-Stress exploratory MDS representations and minimal-Stress theory-compatible MDS representations of the same data are very different, but where both have acceptably low Stress values.

When running MDS in a confirmatory way, the researcher can almost always specify an IC that he or she derives from content theory or from structural relations established in previous empirical research. For example, in psychological research on personal values (such as power, hedonism, and self-direction) where hundreds of studies have led to robust structural expectations, one can predict that the MDS configurations representing correlations among items on the psychological importance of different personal values form a circle. The various personal values are represented as points on this circle in a particular order, and with certain values positioned in opposition to each other (Schwartz and Bilsky 1987; Schwartz, Cieciuch, Vecchione, Davidov, Fischer, Beierlein, Ramos, Verkasalo, Lönnquist, Demirutku, Dirilen-Gumus, and Konty 2014; Borg, Bardi, and Schwartz 2016). Hence, it is easy to formulate

a theoretically sound initial configuration when doing yet another MDS study on personal values.

In exploratory MDS, an initial configuration has to be set up by statistical reasoning. Most MDS programs offer a number of choices for such rational initial configurations. The one that is usually recommended as the best initial configuration—that is, as the one that is most likely to lead to a configuration with minimal Stress—is the "Torgerson" configuration obtained by "classical" MDS (Torgerson 1958; Davison 1978; Borg and Groenen 2005). It is computed by assuming that the dissimilarities are distances, then converting these data to scalar products by squaring and double-centering them, and finally taking the first m eigenvectors of the resulting matrix as the m-dimensional initial configuration.

A second (additional) recommended choice is to run MDS with many different random initial configurations and then pick the configuration with the smallest Stress as the optimal MDS solution. With today's computing power this is a viable alternative, since it only takes seconds to run such an approach, and one can always compare the results of the random and the Torgerson method to check if they lead to the same results.

Unfortunately, applied researchers find themselves in a dilemma when using MDS. They can choose a particular type of initial configuration such as Torgerson. An MDS program will then deliver the supposedly optimal solution based on this choice. Or they can choose the random approach and then the program presents what appears to be the formally best solution based on the analyses of many different initial configurations. In either case, no information is provided on other local-minima solutions even though these solutions may have only marginally higher Stress values but may be theoretically much more meaningful and better interpretable. Yet, even if the solution given by the MDS program is indeed the global minimum solution, an MDS user may simply be interested to see what other local-minima solutions exist, what their Stress values are, and how similar they are. In the following, we study this issue with an artificial data set and describe a systematic approach that can be used by the applied MDS user to answer these questions for his or her data and the particular choice of MDS model.

2. Method

We now describe a procedure for generating local minima solutions in MDS beginning with different initial configurations, and for comparing the configurational similarity of these solutions so that the user can identify those solutions he or she wants to check for their interpretability. The procedure is then illustrated using a simulation study where the true configuration is known. Finally, we look at some applications in real empirical research.

2.1. The procedure

To avoid overlooking local minima solutions in MDS that have an acceptably good fit and that are also substantively interpretable, one must generate many such local minima (if they exist) and then systematically compare them rather than reporting only one Stress-optimal solution. We suggest to achieve this goal in the following way:

1. Run an MDS analysis with a set of different initial configurations (e.g., using many random configurations).

2. Save all resulting MDS solutions and their fit indices (Stress, p-values resulting from permutation tests, etc.).

3. Use generalized Procrustean fitting to eliminate all meaningless differences (i.e., differences not driven by the data) among the MDS solutions.

4. Compute the similarity of each pair of MDS configurations.

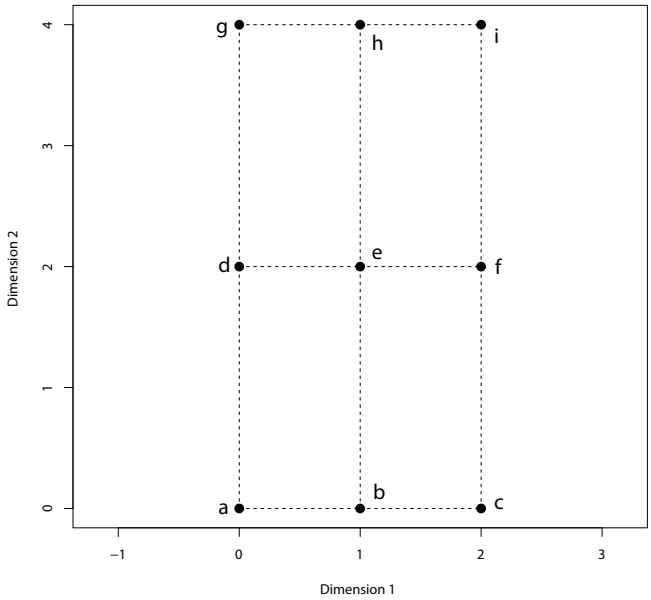


Figure 1: Configuration to be recovered by MDS ("true" configuration)

5. Analyze the similarity structure of the MDS configurations with two-dimensional MDS (to visualize the similarity structure) and cluster analysis (to identify types of MDS configurations).

6. For each type of MDS configuration with a reasonable Stress, plot one prototypical MDS solution and check its interpretability.

7. Pick the MDS solution that is both acceptable in terms of Stress and best interpretable as your MDS solution.

These steps can be easily programmed so that the user has to choose only the particular MDS model he/she wants to fit. Plots and statistics can be produced that allow the user to pick those MDS solutions that he/she wants to study further for their interpretability. A corresponding R (R Core Team 2016) code chunk is given in the supplementary materials. We use the **smacof** package (De Leeuw and Mair 2009) to fit the MDS models.

2.2. Simulation study

We illustrate our procedure by using an artificial case where the true MDS configuration is known. We begin by defining an MDS configuration $\mathbf{X}$ which consists of nine points forming a rectangular grid (see Figure 1). The distances among the nine points are computed. Then, to introduce a somewhat non-linear mapping, their square roots are taken as dissimilarity data for MDS. Thus, there exists a true underlying configuration whose distances are monotonically (but not linearly) related to the dissimilarities.

We then ask if MDS succeeds to recover $\mathbf{X}$ given the above dissimilarities. $\mathbf{X}$, of course, is the global minimum MDS solution with a Stress of zero. We use ordinal MDS because of the non-linear relation of the dissimilarities to the MDS distances in our case. We choose this setup also because ordinal MDS has been a very popular MDS model in applied research. In addition, most MDS programs run ordinal MDS with the additional default specification that ties in the data need not lead to the same distances in the MDS solution (the so called primary approach to ties).

Two types of initial configurations are used in the following, always employing the function `mds()` of the **smacof** package to compute MDS solutions:

1) The first type of IC is the Torgerson configuration. We then add successively more error (randomly sampled from a normal distribution with mean = 0 and $sd = 1, 2, ...$) to the point coordinates of this IC and repeat the MDS analysis for each case. With $sd = 0$ we have the case that is the default IC in most modern MDS programs today, or the IC that is most often recommended as the best single IC based on extensive simulation studies (Borg and Groenen 2005). When adding more noise to a Torgerson IC, we test the case where other solutions with similar Stress exist but would not be found because the algorithm gets stuck in the neighborhood of the Torgerson IC.

2) To safeguard against sub-optimal local minimum solutions, we recommend repeating an MDS analysis with random initial configurations and then pick the resulting MDS solution with the smallest Stress as the best solution. We therefore also run random initial configurations (with coordinate values randomly sampled from a standard uniform distribution), but store each solution and not just the best one. We then compare the various MDS solutions so obtained by inspecting their Stress values and their configurations.

To compare many dozens configurations, a systematic approach is used: We first eliminate all meaningless differences of the various solutions (due to rotations, reflections, dilations, and translations) by Procrustean methods, then measure the overall similarity of each pair of configurations by computing the product-moment correlation of the coordinates of corresponding points, and finally use MDS and cluster analysis to visualize the similarity structure of these correlations to detect possible classes of acceptable MDS solutions.

Specifying the function f in Equation (1) differently, we use the same procedure to study the effect of different ICs for other MDS models, namely ordinal MDS with the so-called secondary approach to ties where tied dissimilarities must be mapped into tied distances ("keep ties"); interval MDS, where the dissimilarities can be linearly transformed; and ratio MDS where the dissimilarities are fixed up to a multiplicative constant.

2.3. Real data applications

Simulations can be illuminating to illustrate a method but they may also be too contrived for the applied researcher, showing applications and solutions for cases that almost never become relevant in empirical research. We therefore use our procedure on some real data sets that have been used before in the applied MDS literature. One such data set is a study by Wish (1971) who asked students about the subjective similarity of 12 different nations. These data are one of the oldest cases of using MDS in attitude research. They were analyzed, for example, by Kruskal and Wish (1978) in their classic introductory textbook on MDS by using ordinal MDS with a Torgerson initial configuration.

A second application uses ratings of 327 psychology students on the importance of personal values as guiding principles in their life (Borg *et al.* 2016). What is scaled here are the inter-correlations of indexes for the so called 10 basic values of the Schwartz theory on values (Schwartz *et al.* 2014). The theory predicts a circular scale, with the points ordered on the circle. This study replicates what was done in literally hundreds of related studies in research on personal values, where MDS has been the cardinal method for testing and refining what is called the Theory of Universals in Values (TUV). Almost always ordinal MDS has been used, but systematic studies of using different initial configurations are not reported.

Finally, a classic MDS application is on data on the subjective similarity of different colors by Ekman (1954). These data were used in the first papers on ordinal MDS by Shepard (1962) and Kruskal (1964) where it was shown that the structure of the observed similarities is an almost perfect color circle with very low Stress.

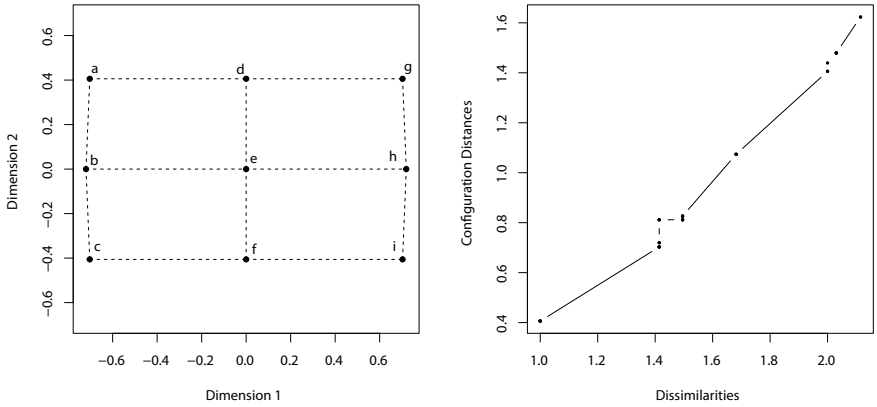


Figure 2: Zero-Stress MDS solution for dissimilarities based on the grid in Figure 1, using Torgerson initial configuration (left panel); Shepard diagram (right panel).

3. Results

3.1. Results: simulation study

A standard ordinal MDS analysis of the dissimilarity data derived from Figure 1 with a Torgerson IC leads to an almost perfect recovery of the underlying configuration, with zero Stress (Figure 2, left panel). The Shepard diagram (right panel of Figure 2) for this configuration shows that the square-root relation of dissimilarities to distances is also closely recovered.

Adding error to the Torgerson IC sometimes leads to MDS solutions with zero Stress but sometimes to radically different solutions that also have zero Stress. Figure 3 gives an example where the points do not form the rectangular grid but are collapsed in one direction to almost a single straight line. The Shepard diagram in the right-hand panel of Figure 3 shows, for example, that **smacof** maps all dissimilarities in the interval $[1.0, 1.4]$ into line segments with essentially zero length.

An even more pronounced step function comes with the solution shown in Figure 4. In this case, MDS maps the dissimilarities into only two types of distances, large ones and small ones. The configuration, therefore, exhibits the shape of an equilateral triangle. Its lines represents

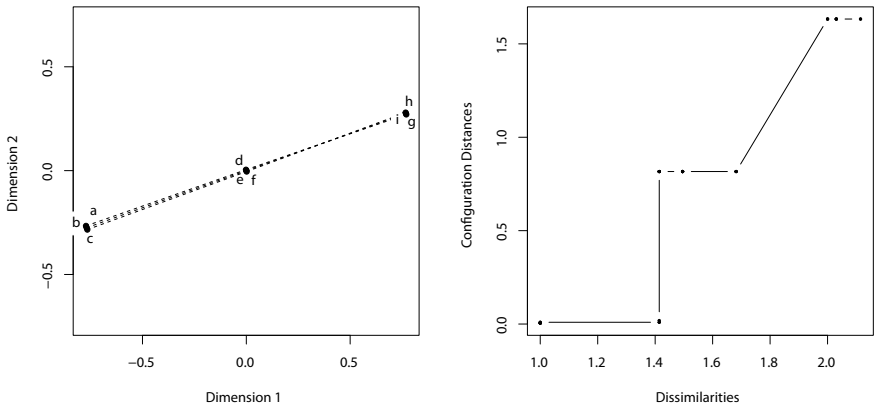


Figure 3: Zero-Stress MDS solution for dissimilarities based on the grid in Figure 1, using Torgerson plus small error initial configuration (left panel); Shepard diagram (right panel).

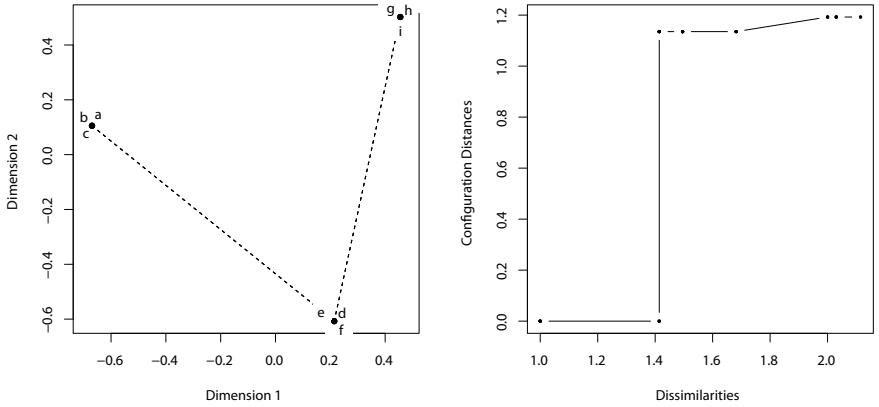


Figure 4: Zero-Stress MDS solution for dissimilarities based on the grid in Figure 1, using Torgerson plus larger error initial configuration (left panel); Shepard diagram (right panel).

the vertical grid lines of the design configuration, collapsed and wired around the triangle's center. Its corner points each summarize the three points of one horizontal grid line of the design configuration.

In a second simulation setting the procedure from Section 2.1 is illustrated on the artificial grid data. 100 different random ICs (drawn from a $U(0, 1)$ distribution) are generated and for each of them an ordinal MDS is fitted. Analyzing the similarities of the resulting MDS solutions shows how bumpy the response surface of ordinal MDS is for this set of data. The structure of the 100 MDS solutions is visualized in Figure 5. This plot is a two-dimensional interval MDS representation of the similarity structure of the 100 ordinal MDS solutions where the

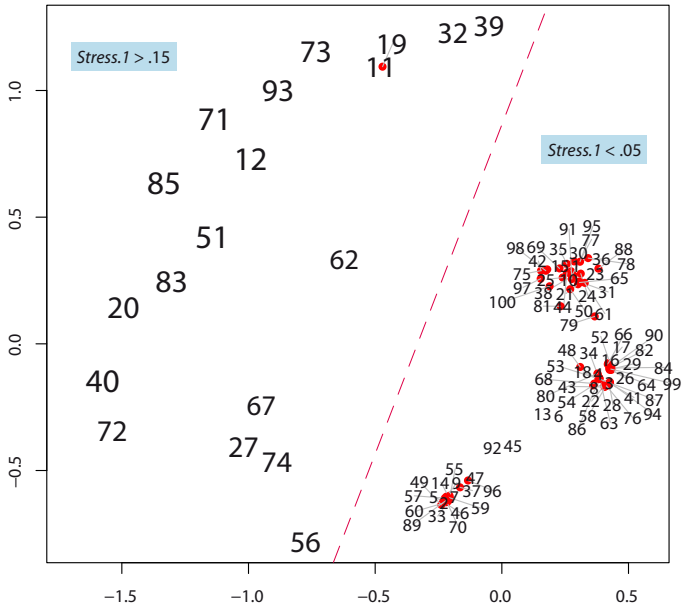


Figure 5: MDS configuration of 100 MDS solutions based on random initial configurations; plotted with **wordcloud** to unclutter point labels; label size corresponds to Stress of respective MDS solution; dashed line partitions plane into high- and low-Stress solutions, resp.

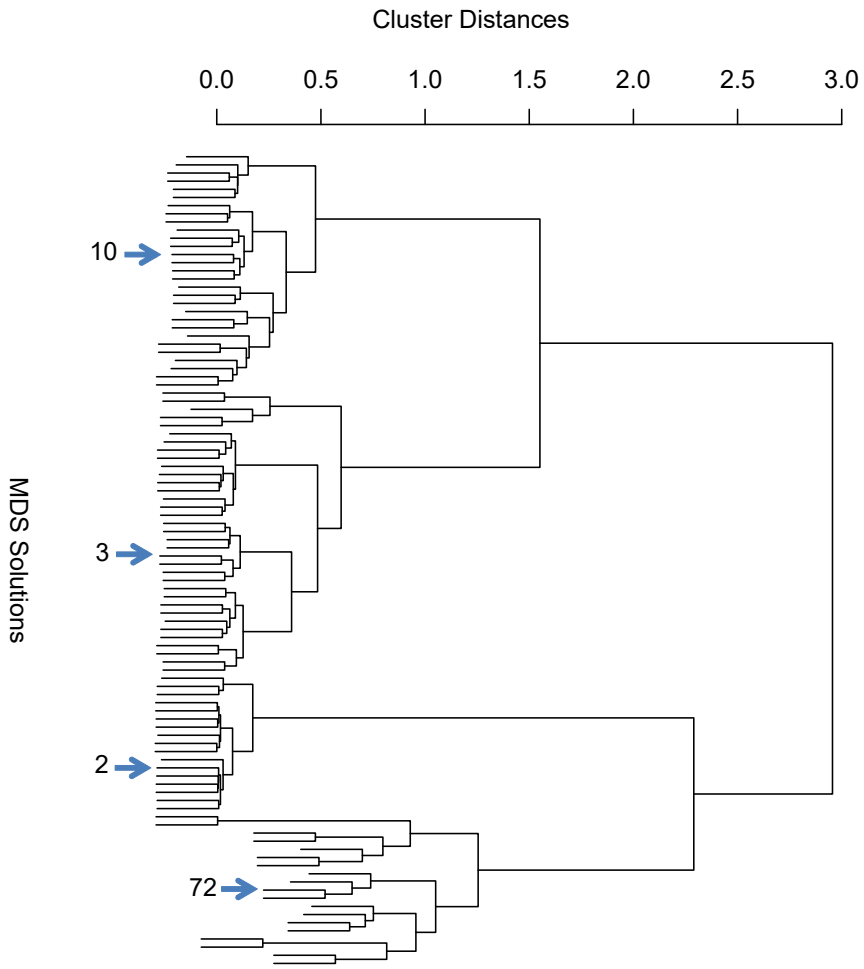


Figure 6: Dendrogram (hierarchical cluster analysis; Ward criterion) of similarity indexes of the 100 MDS solutions based on random initial configurations; arrows point to four prototypes of the major clusters.

similarity of two configurations is measured by intercorrelating (using the Pearson correlation coefficient) for each pair of solutions the coordinates of corresponding points after Procrustean fitting (cf. Borg and Leutner 1985). To unclutter the plot we used the R package **wordcloud** (Fellows 2014). Figure 5 represents each point by its label, except in cases where the point labels would be overlapping. In that case, it plots the points as red dots and connects the point label to the point with a straight line so that the various labels do not overlap (see the point clouds on the right-hand side of the plot).

The figure shows three rather dense clusters of low Stress solutions on the right-hand side, and various widely scattered solutions with high Stress on the left side of the plot (large labels). The plot can be partitioned by a staight line that separates all MDS solutions with poor Stress (left-hand side) from those with acceptably low Stress (right-hand side).

One can also use cluster analysis to study the similarity of the 100 MDS solutions: Figure 6 shows that a hierarchical cluster analysis identifies four major clusters. For each cluster, one configuration is marked by an arrow in the dendrogram as a prototype (the configurations #10, #3, #2, and #72). Figure 7 exhibits these configurations. Only one of them, #72 has a poor fit (*Stress.1* = .236). The other three configurations correspond closely to the configurations in Figure 2 ("grid", #10), Figure 3 ("collapsed horizontals", #3), and Figure 4 ("triangle", #2), respectively. They all have Stress values of almost zero. Thus, formally,

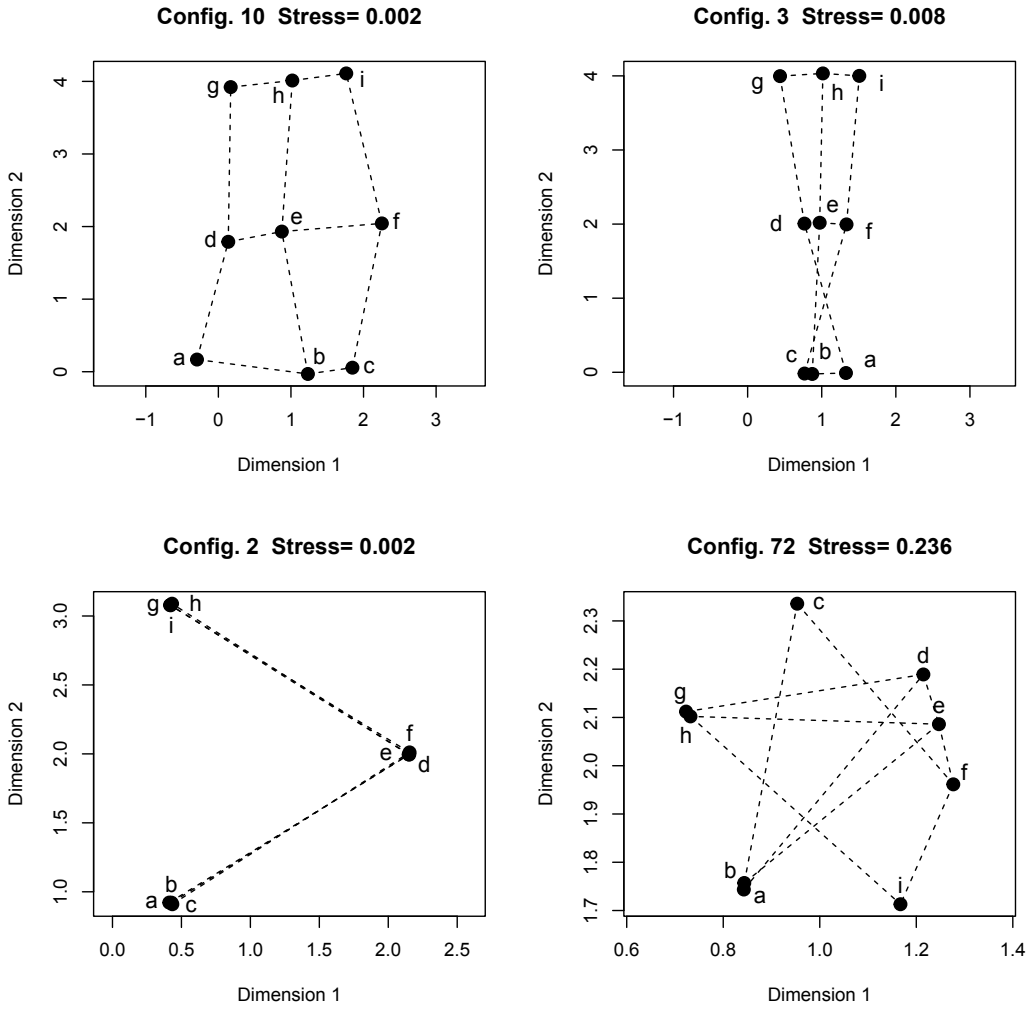


Figure 7: Configurations that represent four different clusters of the dendrogram in Figure 6.

they are all excellent representations of the data in the sense of the ordinal primary approach to ties MDS model, but neither the "collapsed horizontals" nor the "triangle" solutions recover the underlying configuration of Figure 1 very well. Rather, they are examples of degenerate solutions.

Ordinal MDS with the secondary approach to ties ("keep ties") recovers the latent configuration of Figure 1 (and the slightly non-linear relation of dissimilarities and distances) perfectly when using the Torgerson IC or the best random IC. However, using other ICs, one also obtains many undesirable solutions. Some examples are shown in Figure 8. The configurations #5, #11, and #7 swap the points on one or two of the horizontal grid lines of the design configuration. Their Stress values are not zero, but they are all significant by the norms of Spence and Ogilvie (1973). We also find a probability of essentially zero in all cases using the permutation test provided by **smacof**. Hence, if the true configuration is not known, one would likely accept any one of these solutions as "the" MDS solution if there are theoretical reasons that speak for such a choice.

Interval MDS leads to MDS solutions that are either unacceptable local minima with high Stress, or to configurations that closely correspond to the configurations #2, #5, and #7 in Figure 8. They have Stress values of .046, .113, and .113, respectively. All Stress values indicate a highly significant model fit according to **smacof**'s permutation test.

Finally, ratio MDS leads to solutions that are quite similar to those of interval MDS. Of course, the Stress values are higher, even in case of "grid" solutions, because the regression

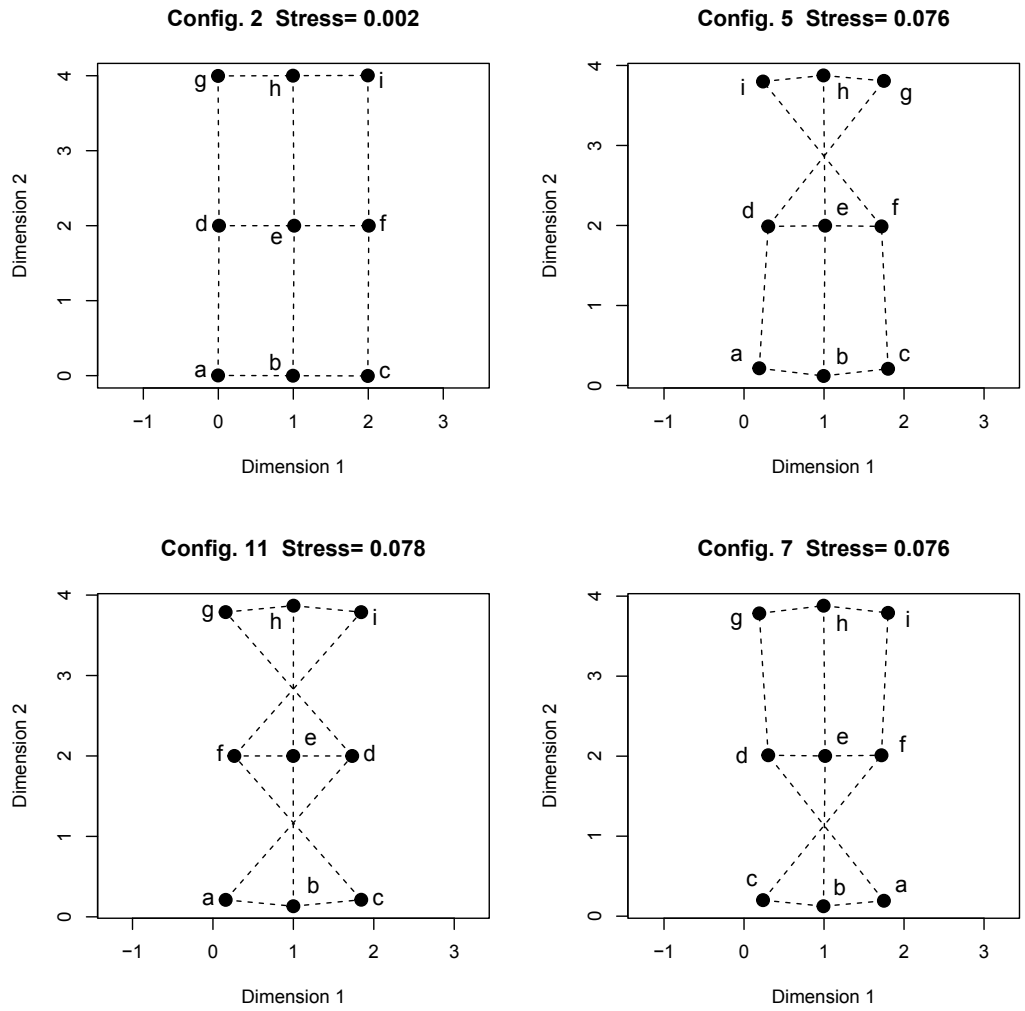


Figure 8: Four local minima solutions using ordinal MDS with the secondary approach to ties.

line in the Shepard diagram must be linear and must run through the origin in ratio MDS.

When adding some jitter to the design configuration, we find similar results in all cases, except when using ordinal MDS with the primary approach to ties. The noise that is added to the design configuration makes all ties in the dissimilarities go away. The dissimilarities are, therefore, all different and this makes it impossible for MDS to generate gross step functions as in Figure 4, for example. Hence, solutions in the form of a triangle are not observed anymore. Rather, the solutions are either grid-like, or they show some swapping of points on the lower and/or on the upper horizontal grid line (as in Figure 7).

3.2. Results: real data applications

We now turns to three classic real data sets. For the data on the subjective similarity of different countries reported by Wish (1971), ordinal MDS starting with random configurations leads to different solutions. Many of them have unacceptably high Stress, but there are two types of solutions with the same minimal Stress of .185. The fit of these solutions is also significant ($p = .04$) using **smacof**'s permutation test. Figure 9 shows these solutions next to each other. They are rather similar but differ in two important details: In configuration #1, the positions of Japan and Israel are swapped in comparison to where they are in configuration #13; moreover, in configuration #1 India is positioned more in the center of the configuration.

The crossed dashed lines in the plots represent two external scales that were optimally fitted

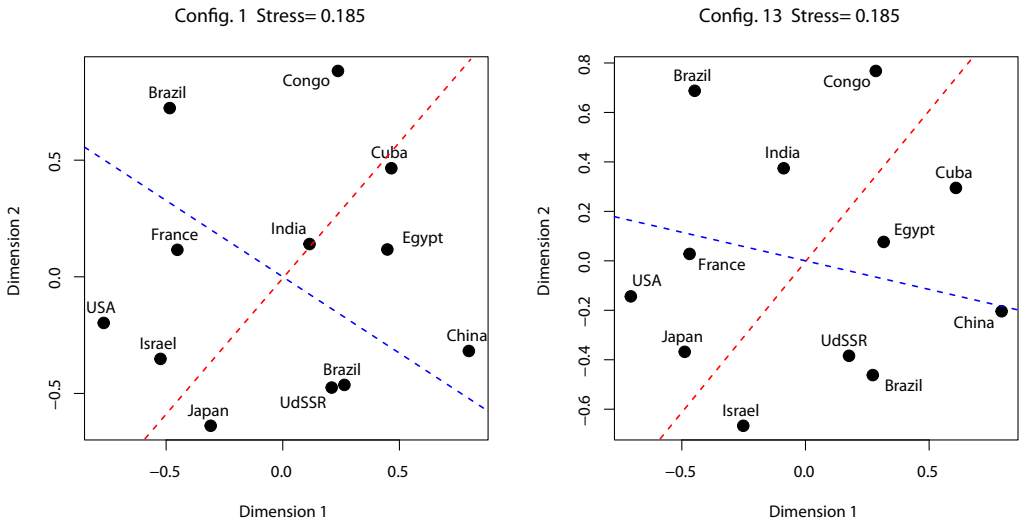


Figure 9: Two MDS solutions for ratings on the subjective similarity of different nations.

into these plots. The external scales show the economic development and the number of inhabitants, respectively, of the 12 countries in 1971 (Borg *et al.* 2013). In case of economic development (red lines), the external scales and the fitted scales correlate with $r = .936$ and .925 in the plots. In case of the number of inhabitants, the correlation is $r = .464$ in configuration #1 but only $r = .303$ in configuration #13. Hence, configuration #1 is the more meaningful MDS solution if one wants to follow Wish (1971) and Wish, Deutsch, and Biener (1972) in interpreting the configuration in terms of these dimensions. However, this solution may not be the one that is reported by the MDS program as the final solution.

A second application uses ratings of students on the importance of personal values as guiding principles in their life (Borg *et al.* 2016). What is scaled here are the inter-correlations of indexes for ten basic values. These values are predicted to form a circular scale, with the points ordered as PO(wer), AC(hievement), HE(donism), ST(imulation), S(elf-)D(irection), UN(iversalism), BE(nevolence), TR(adition), CO(nformity), SE(curity), and back to PO(wer).

Ordinal MDS (using `mds()` with its default settings) leads to the solution shown in the lower left-hand panel of Figure 10 with *Stress.1* = .128. This solution obviously closely corresponds to the predicted value circle even though the circle is somewhat dented, with AC(hievement) moved towards the circle's center. Random initial configurations identify three more local minima. They all have almost the same Stress values. They are also all quite similar, but a closer inspection shows that configuration #2 exhibits a theory-incompatible swapping of CO(nformity) and TR(adition), while configuration #5 moves AC(hievement) somewhat towards the center of the circle (see Figure 10). Configuration #3 on the other hand supports the Schwartz theory perfectly.

As a third example, we study the color similarity data collected by Ekman (1954). For these data, MDS finds a circular configuration with points ordered in terms of the physical wavelengths of the colors that they represent. With both Torgerson or with random initial configurations, the usual ordinal MDS with the primary approach to ties almost always leads to the same solution, the color circle. Configurations that do not exhibit this circle are extremely rare when testing many different initial configurations. Moreover, they all have much higher Stress. This is useful information for the substantive researcher because it shows that these data have an essentially unique MDS representation.

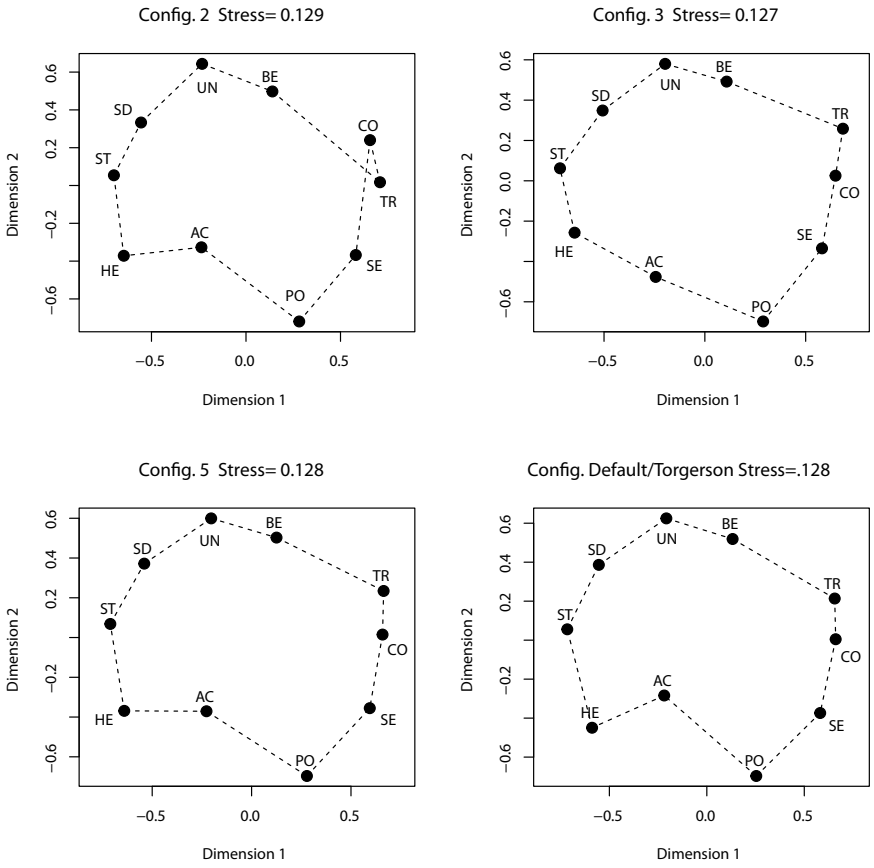


Figure 10: Local minima MDS solutions for intercorrelations of indexes on the importance of personal values.

4. Discussion

The above analyses demonstrate that it can be risky to use MDS without thoroughly studying the effects of different initial starting configurations. Not just in contrived simulations, but also with real data MDS can have many local minima with almost the same Stress but with different configurations, allowing different interpretations. The Torgerson IC proves to be a good rational IC, but (when adding error to the artificial data derived from the design grid in Figure 1), even it does not guarantee to always succeed recovering a known underlying configuration. When using stronger MDS models, particular types of solutions that result from non-smooth step functions (as in Figure 4) cannot occur anymore, but radically different solutions that all have acceptably low Stress values still exist, in particular if the data contain ties and the usual ordinal MDS model with the primary approach to ties is employed.

If MDS is used in an exploratory way to visualize the structure of proximity data and make them accessible to the researcher's eye, there is really nothing to recover. In this case, one should pick the local minimum solution that is best interpretable, ideally even suggesting a content-based law of formation. Of course, this solution should also have an acceptably low Stress value. In the applications discussed above, it turned out that the best-interpretable solutions always had a Stress value that was not worse than the Stress value of other solutions. However, this may not always be true in applied research, and so we recommend to at least take a look at other local minima solutions before accepting the solution offered by the MDS program.

This also holds if one does not use ordinal MDS but stronger MDS models. To make testing the effects of different initial configurations easy in practice, the supplementary materials provide corresponding R code. With this code, users can identify the configurations they want to look

at in more detail. The various plots (together with the fit statistics) are the basis for deciding what to pick as the best MDS representation for the given data. If computing time becomes as issue as in large scale MDS settings, R's facilities for parallelizing the random IC fits can be used (e.g. using the **parallel** package). Since these MDS fits are independent from each other, the job can be distributed easily on machines with multiple cores.

In general, no formal decision rule seems possible that tells the user what solutions satisfy the criterion of having a Stress that is "still acceptable". This decision always requires to consider a set of statistical and content criteria (see Mair, Borg, and Rusch 2016) such as the overall Stress; the composition of the Stress (Stress per point); the assumed error level of the data; the mapping requirements of the chosen MDS model; and the statistical significance, the robustness, the replicability, and the theoretical interpretability of the solution.

References

Bentler PM, Weeks DG (1978). "Restricted Multidimensional Scaling Models." *Journal of Mathematical Psychology*, **17**, 138–151.

Borg I, Bardi A, Schwartz SH (2016). "Does the Value Circle Exist within Persons or Only across Persons?" *Journal of Personality*. Forthcoming.

Borg I, Groenen PJF (2005). *Modern Multidimensional Scaling*. 2nd edition. Springer, New York.

Borg I, Groenen PJF, Mair P (2013). *Applied Multidimensional Scaling*. Springer, New York.

Borg I, Leutner D (1985). "Measuring the Similarity of MDS Configurations." *Multivariate Behavioral Research*, **20**, 325–334.

Borg I, Lingoes JC (1980). "A Model and Algorithm for Multidimensional Scaling with External Constraints on the Distances." *Psychometrika*, **45**, 25–38.

Cox TF, Cox MAA (2000). *Multidimensional Scaling*. 2nd edition. Chapman & Hall, London.

Davison ML (1978). *Multidimensional Scaling*. Wiley, New York.

De Leeuw J, Heiser WJ (1980). "Multidimensional Scaling with Restrictions on the Configuration." In PR Krishnaiah (ed.), *Multivariate Analysis*, volume V, pp. 501–522. North-Holland, Amsterdam.

De Leeuw J, Mair P (2009). "Multidimensional Scaling Using Majorization: SMACOF in R." *Journal of Statistical Software*, **31**(3), 1–30. URL http://www.jstatsoft.org/v31/i03/.

Ekman G (1954). "Dimensions of Color Vision." *Journal of Psychology*, **38**, 467–474.

Fellows I (2014). *wordcloud: Word Clouds*. R package version 2.5, URL https://CRAN.R-project.org/package=wordcloud.

Groenen PJF (1993). *The Majorization Approach to Multidimensional Scaling: Some Problems and Extensions*. Ph.D. thesis, University of Leiden.

Kruskal JB (1964). "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis." *Psychometrika*, **29**, 1–27.

Kruskal JB, Wish M (1978). *Multidimensional Scaling*. Sage, Beverly Hills, CA.

Mair P, Borg I, Rusch T (2016). "Goodness-of-fit Assessment in Multidimensional Scaling and Unfolding." *Multivariate Behavioral Research*. Forthcoming.

Mair P, De Leeuw J (2015). "Unidimensional Scaling." In *Wiley StatsRef: Statistics Reference Online*. Wiley, New York.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schwartz SH, Bilsky W (1987). "Toward a Psychological Structure of Human Values." *Journal of Personality and Social Psychology*, pp. 550–562.

Schwartz SH, Cieciuch J, Vecchione M, Davidov E, Fischer R, Beierlein C, Ramos A, Verkasalo M, Lönnquist JE, Demirutku K, Dirilen-Gumus O, Konty M (2014). "Refining the Theory of Basic Individual Values." *Journal of Personality and Social Psychology*, **103**, 663–688.

Shepard RN (1962). "Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function." *Psychometrika*, **27**, 125–140.

Spence I, Ogilvie JC (1973). "A Table of Expected Stress Values for Random Rankings in Nonmetric Multidimensional Scaling." *Multivariate Behavioral Research*, **8**, 511–517.

Torgerson WS (1958). *Theory and Methods of Scaling*. John Wiley & Sons, New York.

Wish M (1971). "Individual Differences in Perceptions and Preferences among Nations." In CW King, D Tigert (eds.), *Attitude Research Reaches New Heights*, pp. 312–328. American Marketing Association, Chicago, IL.

Wish M, Deutsch M, Biener L (1972). "Differences in Perceived Similarity of Nations." In AK Romney, RN Shepard, SB Nerlove (eds.), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, pp. 289–313. Academic Press, New York.

Affiliation:

Ingwer Borg
Fachrichtung Psychologie
Westfälische-Wilhelms Universität
Fliednerstr. 21
48149 Münster, Germany

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 46 *Submitted:* 2016-04-08
February 2017 *Accepted:* 2016-07-12

Austrian Journal of Statistics
February 2017, Volume 46, 33–47.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i2.566

Estimations of the Parameters of Generalised Exponential Distribution under Progressive Interval Type-I Censoring Scheme with Random Removals

Arun Kaushik, Aakriti Pandey, Sandeep K Maurya, Umesh Singh and Sanjay K Singh
Department of Statistics, Institute of Science,
Banaras Hindu University, India-221 005.

Abstract

The present article aims to point and interval estimation of the parameters of generalised exponential distribution (GED) under progressive interval type-I (PITI) censoring scheme with random removals. The considered censoring scheme is most useful in those cases where continuous examination is not possible. Maximum likelihood, expectation-maximization and Bayesian procedures have been developed for the estimation of parameters of the GED, based on a PITI censored sample. Real datasets have been considered to illustrate the applicability of the proposed work. Further, we have compared the performances of the proposed estimators under PITI censoring to that of the complete sample.

Keywords: statistical computing, Bayesian, maximum likelihood, simulation, progressive interval censoring.

1. Introduction

In Statistical literature several authors have proposed models which are supposed to be competing models (see, Mudholkar and Srivastava (1993), Kondolf and Adhikari (2000), etc.) to Gamma and Weibull distributions. Similarly, Gupta and Kundu (2001a) introduced the generalised exponential distribution (GED) as an alternative to Gamma and Weibull distributions. Nowadays, its has gained popularity in the statistical literature due to its simplicity, and the probability density function (pdf) is very flexible and accommodate wide variety of shapes. The probability density function of the GED is given as,

$$f(x|\alpha,\theta) = \alpha\theta e^{-\theta x}(1-e^{-\theta x})^{\alpha-1}; \qquad x \geq 0, \alpha, \theta > 0, \tag{1}$$

where α is the shape parameter and θ is the scale parameter of the considered model. Its cumulative distribution and survival functions are given by,

$$F(x|\alpha,\theta) = (1-e^{-\theta x})^{\alpha} \tag{2}$$

and

$$S(x|\alpha,\theta) = 1-(1-e^{-\theta x})^{\alpha}; \qquad x \geq 0, \alpha, \theta > 0, \tag{3}$$

respectively. It has been extensively studied by Raquab and Madi (2005); Singh, Singh, Singh, and Prakash (2008) and many others. Gupta and Kundu (2001a); Jaheen (2004); Sarhan (2007); Zheng (2002) discuss its importance over gamma and Weibull distribution which are two most popular distribution used in survival analysis. Gupta and Kundu (2001a) noted that, in many situations, the two-parameter generalised exponential distribution provides a better fit than the two-parameter Weibull distribution. It may be noted here that the GED is a special case of a distribution that was used by Gompertz (1825). Gupta and Kundu (2001b) studied different methods of point estimation for GED parameters which include maximum likelihood estimation, method of moment estimation and probability plot method of estimation based on complete samples. Singh, Singh, and Kumar (2011) discussed the parameter estimation and reliability characteristic of GED under Bayesian paradigm. It is worthwhile to mention here that little attention has been paid to inferences based on censored samples from GED under the Bayesian paradigm, although censoring is quite common in various clinical and life testing experiments.

Situations do arise when the units under study are lost or removed from the experiments while they are still alive i.e., we get censored data in such cases. If the point at which the experiment terminate is time dependent, it is called Type-I censoring. On the other hand, if it is unit dependent, it is called Type-II censoring. Depending on the need and practical considerations, various modified forms of censoring schemes have been discussed in the literature. Aggarwala (2001) proposed a combination of interval Type-I censoring and progressive censoring called progressive Type-I (PTI) interval censoring which naturally arises in most clinical experiments. To have a clear visualization of this censoring scheme, let us consider an experiment with n bladder cancer patients for whom remission times are to be recorded. The patients are called for regular check-ups at scheduled times, and those who turn up are checked. At the first visit, scheduled at time T_1, only $n - R_1$ patients out of the total n patients report, i.e. R_1 patients leave the experiment during the time interval $(0, T_1]$. The experimenter examines these $n - R_1$ patients and finds that cancer has reoccurred in D_1 patients. It may be noted here that the exact time of recurrence for these D_1 patients is not known to the experimenter; he only has the information about the number of recurrences during the time period between the start of the experiment and first visit. At the second visit, scheduled at time T_2, $n - R_1 - D_1 - R_2$ out of the remaining $n - R_1 - D_1$ patients report, i.e. R_2 patients leave the experiment at this stage (during the time interval $(T_1, T_2]$). The experimenter examines these patients and finds that cancer has reoccurred in D_2 patients out of remaining $n - R_1 - D_1 - R_2$ patients, and in this way the experiment continues till the m^{th} visit. At this stage (m^{th} visit) all the remaining $R_m = n - D_1 - D_2 \cdots - D_m - R_1 - R_2 \cdots R_{m-1}$ units are removed, i.e. the experiment is terminated at this stage. Recently Chen and Lio (2010) proposed a methodology to estimate parameters involve in GED under PTI interval censoring under the assumption that the proportions (p_i) of the patients leaving the experiment during $(T_{i-1}, T_i]$ is known in advance, i.e. they prefixed the proportions $p_1, p_2, \ldots, p_m$ and considered that at i^{th} stage, $\lfloor n_i * p_i \rfloor$ patients shall leave the experiment. Here, $\lfloor n_i * p_i \rfloor$ denotes the largest integer less than or equal to $n_i * p_i$. The author's claim that exactly $\lfloor n_i * p_i \rfloor$ patients out of $\lfloor n_i \rfloor$ will drop out of the experiment at the i^{th} stage (visit), seems unrealistic and hypothetical. In fact, the number of patients dropping out from the clinical trial at any stage is beyond the control of the experimenter and cannot be predetermined. It seems more logical and natural to consider these p_i as random variables for the risk of dropping at the i^{th} stage. Perhaps, keeping a similar thought in mind, Yuen and Tse (1996) and Tse, Yang, and Yuen (2000) discussed progressive censoring scheme with binomial removal. Ashour and Afify (2007) have used PITI censoring scheme with binomial removals assuming that the exact value of the lifetimes of the units are observable. In their studies, they have assumed that the number of removals $R_i's$ at the i^{th} stage $(i = 1, 2, \ldots, m)$ is random and follows the binomial distribution with probability p_i. Thus, $R_1 \sim Binomial\,(n,\ p_1)$ and $R_2 \sim Binomial\,(n - D_1 - R_1,\ p_2)$. In general, the number of units dropping at the i^{th} stage, R_i follows the binomial distribution

with parameters $(n - \sum_{l=1}^{i} D_l + R_l,\ p_i)$ for $i = 1, 2, 3, \cdots, m-1$. In this paper, we will consider PITI censored data with binomial removals and develop estimators for the shape and scale parameter under the situation that the exact value of the lifetimes of the units not observable, only the number of units lying in the specified interval of times are known. For the parameter estimation problem, we have considered the most popular loss function, namely the squared error loss function (SELF) which can be easily justified on the grounds of minimum variance unbiased estimation (see Berger 2013, Ch.2). We will compare the performance of the proposed estimators of the parameters obtained under the above stated censoring scheme with the estimates under the complete sample case.

The rest of the paper is organized in the following sections. In Section 2, Classical and Bayes procedures for the estimation of the model parameters based on PITI with binomial removal samples have been developed. Two real datasets has been considered, the first one is related to the survival time of patients with plasma cell myeloma and the second one regarding the number of revolutions in million before failure of groove ball bearings, have been considered for the illustration of the proposed methodology in Section 3. Comparison of the estimators based on simulation study has been provided in Section 4. Finally, conclusions have been summarized in Section 5.

2. Parameter estimation

2.1. Maximum likelihood estimation

In this section, we provide the MLEs of α and θ, the parameters of the lifetime distribution given in equation (1). Let us consider that n units are put on test initially at time $T_0 = 0$, and we record the number of droppings and number of failures during pre-specified time intervals $(T_{i-1}, T_i]$ $(i = 1, 2, \ldots, m)$ amongst the available units; i.e. we get the data consisting of the number of failures $D = (d_1, d_2, \ldots, d_m)$ and number of droppings $R = (r_1, r_2, \ldots, r_m)$ during the time intervals $(0, T_1]$, $(T_1, T_2]$, ..., $(T_{m-1}, T_m]$ through the censoring scheme described in the previous section. It may be noted here that the individual units dropping from the test at

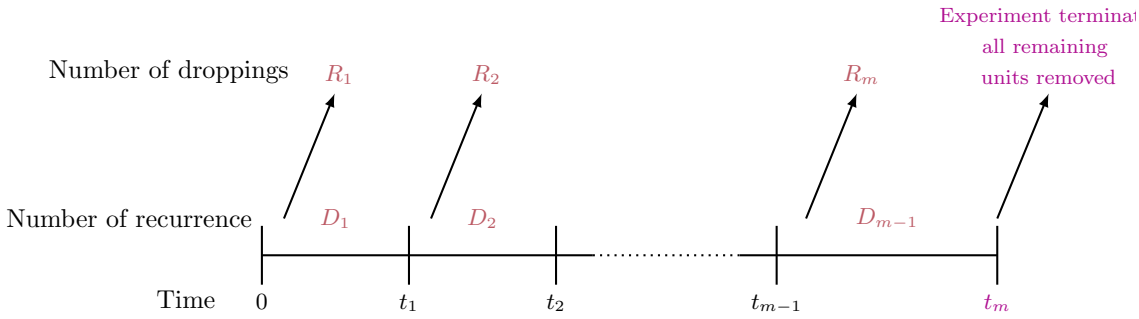


Figure 1: Progressive interval type-I censoring scheme

the i^{th} stage (during the time interval $(T_{i-1}, T_i]$), $i = 1, 2, \cdots m$ are random and independent of each other with certain probability, say p_i, for $i = 1, 2, \cdots m$. Therefore, the number r_1 of units dropping at the 1^{st} stage follows a binomial distribution with parameters (n, p_1) and $r_i;\ i = 2, 3, \cdots m$ follows a binomial distribution with parameters $\left(n - \sum_{l=1}^{i-1}(d_l + r_l), p_i\right)$; i.e.

$$
\begin{aligned}
P\left(r_1|p_1\right) &= \binom{n}{r_1} p_1^{r_1}(1-p_1)^{n-r_1} \\
P\left(r_2|r_1, d_1, p_2\right) &= \binom{n-d_1-r_1}{r_2} p_2^{r_2}(1-p_2)^{n-d_1-r_1-r_2}
\end{aligned}
$$

and in general

$$P\left(r_i|r_1,\ldots,r_{i-1},d_1,\ldots,d_{i-1},p_i\right) = \binom{n-\sum_{j=1}^{i-1}(d_j+r_j)}{r_i} p_i^{r_i}(1-p_i)^{n-\sum_{j=1}^{i-1}d_j-\sum_{j=1}^{i}r_j},$$

for $i = 2,3,\cdots m$ and $r_{m+1} = n - \sum_{j=1}^{m}(d_j - r_j)$. Now the complete likelihood for the observed data can easily be written as

$$\begin{aligned} L(\alpha,\theta|R,D,T) &\propto \prod_{i=1}^{m}[F(T_i)-F(T_{i-1})]^{d_i}\times[1-F(T_{i-1})]^{r_i} \\ &\quad \times P\left(r_i|r_1,\ldots,r_{i-1},d_1,\ldots,d_i,p_{i-1}\right)[1-F(T_m)]^{r_{m+1}} \\ &= \prod_{i=1}^{m}\left[\left(1-e^{-\theta T_i}\right)^{\alpha}-\left(1-e^{-\theta T_{i-1}}\right)^{\alpha}\right]^{d_i}\left[1-\left(1-e^{-\theta T_{i-1}}\right)^{\alpha}\right]^{r_i} \\ &\quad \times\left[1-\left(1-e^{-\theta T_m}\right)^{\alpha}\right]^{r_{m+1}} \quad \binom{n-\sum_{j=1}^{i-1}(d_j+r_j)}{r_i} \\ &\quad \times {p_i}^{r_i}(1-p_i)^{n-\sum_{j=1}^{i-1}d_j-\sum_{j=1}^{i}r_j}. \end{aligned} \tag{4}$$

Above expression bifurcates as

$$L(\alpha,\theta|R,D,T) \propto L_1(\alpha,\theta|R,D,T)L_2(P|R,D,T), \tag{5}$$

where

$$L_1(\alpha,\theta|R,D,T) = \prod_{i=1}^{m}\left[\left(1-e^{-\theta T_i}\right)^{\alpha}-\left(1-e^{-\theta T_{i-1}}\right)^{\alpha}\right]^{d_i}\prod_{i=1}^{m+1}\left[1-\left(1-e^{-\theta T_{i-1}}\right)^{\alpha}\right]^{r_i} \tag{6}$$

Note that $L_2(\cdot)$ is free from α and θ. Thus, to compute ML estimate of α and θ, we require only $L_1(\cdot)$. The corresponding log likelihood function can be written as

$$\begin{aligned} logL_1(T,\alpha,\theta) &= \sum_{i=1}^{m} d_i \ln\left[\left(1-e^{-\theta T_i}\right)^{\alpha}-\left(1-e^{-\theta T_{i-1}}\right)^{\alpha}\right] \\ &\quad + \sum_{i=1}^{m+1} r_i \ln\left[1-\left(1-e^{-\theta T_{i-1}}\right)^{\alpha}\right]. \end{aligned} \tag{7}$$

Hence, the likelihood equations can be obtained as;

$$\begin{aligned} \frac{dlogL}{d\alpha} &= \sum_{i=1}^{m} d_i \frac{\left[\left(1-e^{-\theta T_i}\right)^{\alpha} ln(1-e^{-\theta T_i}) - \left(1-e^{-\theta T_{i-1}}\right)^{\alpha} ln(1-e^{-\theta T_{i-1}})\right]}{\left[(1-e^{-\theta T_i})^{\alpha}-(1-e^{-\theta T_{i-1}})^{\alpha}\right]} \\ &\quad - \sum_{i=1}^{m+1} r_i \frac{\left(1-e^{-\theta T_{i-1}}\right)^{\alpha} ln(e^{-\theta T_{i-1}})}{\left[1-(1-e^{-\theta T_{i-1}})^{\alpha}\right]} = 0 \end{aligned} \tag{8}$$

$$\begin{aligned} \frac{dlogL}{d\theta} &= \sum_{i=1}^{m} d_i \frac{\left[\left(1-e^{-\theta T_i}\right)^{\alpha-1} e^{-\theta T_i}\alpha T_i - \left(1-e^{-\theta T_{i-1}}\right)^{\alpha} e^{-\theta T_{i-1}}\alpha T_{i-1}\right]}{\left[(1-e^{-\theta T_i})^{\alpha}-(1-e^{-\theta T_{i-1}})^{\alpha}\right]} \\ &\quad - \sum_{i=1}^{m+1} r_i \frac{\left(1-e^{-\theta T_{i-1}}\right)^{\alpha-1} e^{-\theta T_{i-1}}\alpha T_{i-1}}{\left[1-(1-e^{-\theta T_{i-1}})^{\alpha}\right]} = 0 \end{aligned} \tag{9}$$

The MLEs of α and θ can be obtained by solving (8) and (9) simultaneously. But it may be noted here that explicit solutions cannot be obtained from the above equations. Thus, we propose the use of a suitable numerical technique to solve these two non-linear equations. One

may use Newton-Raphson or simulated annealing of their variants to solve these equations. This can be routinely done using R or other packages. We have also obtained the observed information matrix,

$$I(\alpha,\theta|data) = \begin{bmatrix} -L_{\alpha\alpha} & -L_{\alpha\theta} \\ -L_{\theta\alpha} & -L_{\theta\theta} \end{bmatrix}, \tag{10}$$

where, all the second partial derivatives of the log-likelihood function $L_{\alpha\alpha}$, $L_{\alpha\theta}$, $L_{\theta\alpha}$ and $L_{\theta\theta}$ are provided in the Appendix-A. Based on it, the asymptotic confidence (AC) interval and standard errors of the parameter estimates can be obtained in the usual way. While using the Newton-Raphson algorithm (the details are provided in the simulation section) to compute the MLEs for the parameters, it is observed that the iterations converge approximately $85\% - 90\%$ of the time.

2.2. Bayesian estimation

In this section, we provide the Bayesian inferences for α and θ, when we have the progressive interval type-I censored data as explained in Figure 1. We have also obtained the highest posterior density (HPD) intervals for both the parameters. Before proceeding further, we make selections for the prior distributions of the parameters. Following Berger and Sun (1993); Raquab and Madi (2005); Singh, Singh, and Kumar (2014), it is assumed that both α and θ are independent gamma variates, having pdfs

$$g_1(\alpha) = \frac{\lambda_1^{\nu_1}}{\Gamma(\nu_1)} e^{-(\lambda_1\alpha)} \alpha^{(\nu_1-1)} \quad ; \qquad 0 < \alpha < \infty,\ \lambda_1 > 0, \nu_1 > 0 \tag{11}$$

and

$$g_2(\theta) = \frac{\lambda_2^{\nu_2}}{\Gamma(\nu_2)} e^{-(\lambda_2\theta)} \theta^{(\nu_2-1)} \quad ; \qquad 0 < \theta < \infty,\ \lambda_2 > 0, \nu_2 > 0, \tag{12}$$

Here, all the hyperparameters λ_1, ν_1, λ_2 and ν_2 are assumed to be known and can be evaluated following the method suggested by Singh, Singh, and Kumar (2013). We compute the Bayes estimate of the unknown parameters under the squared error loss function. Using the priors given in (11) and (12) and the likelihood function (4), the joint posterior density of α and θ for the given data can be written as

$$\begin{aligned} \pi(\alpha,\theta|R,D,T) &= \frac{L(\alpha,\theta|R,D,T)\, g_1(\alpha)\, g_2(\theta)}{\int_0^\infty \int_0^\infty L(\alpha,\theta|R,D,T)\, g_1(\alpha)\, g_2(\theta)\, d\alpha\, d\theta} \\ &= \frac{J}{\int\int_0^\infty J d\alpha d\theta}, \end{aligned} \tag{13}$$

where

$$\begin{aligned} J = J(\alpha,\theta) &= e^{-(\lambda_1\alpha+\lambda_2\theta)} \alpha^{(\nu_1-1)} \theta^{(\nu_2-1)} \prod_{i=1}^{k} \left[(1-e^{-\theta T_i})^\alpha - (1-e^{-\theta T_{i-1}})^\alpha\right]^{d_i} \\ &\quad \times \left[1-(1-e^{-\theta T_i})^\alpha\right]^{r_i}. \end{aligned}$$

Let $h(\cdot)$ be a function of α and θ. Then, the Bayes estimator of $h(\cdot)$ under the squared error loss function is given by

$$\begin{aligned} \hat{h}_B(\alpha,\theta) &= E_\pi(h(\alpha,\theta)) \\ &= \frac{\int\int_0^\infty h(\alpha,\theta)\, J\, d\alpha\, d\theta}{\int\int_0^\infty J\, d\alpha\, d\theta}. \end{aligned} \tag{14}$$

It is clear from the expression (13) that there is no closed form for the estimators, so we suggest using an MCMC procedure to compute the Bayes estimates. After getting MCMC

samples from the posterior distribution, we can find the Bayes estimate for the parameters in the following way

$$[E(\Theta|data)] = \left[\frac{1}{N-N_0}\sum_{i=N_0+1}^{N}\Theta_i\right],$$

where N_0 is burn-in period of the Markov chain and $\Theta_i = [\alpha_i, \theta_i]'$. For computation of the highest posterior density (HPD) interval of Θ, order the MCMC sample of Θ as $\Theta_{(1)}$, $\Theta_{(2)}$, $\Theta_{(3)}$, $\cdots$, $\Theta_{(N)}$. Then construct all the 100(1-γ)% credible intervals of Θ say $(\Theta_{(1)}$, $\Theta_{(N\lfloor 1-\gamma\rfloor+1)})$, $(\Theta_{(2)}, \Theta_{(N\lfloor 1-\gamma\rfloor+2)})$ $\cdots$,$(\Theta_{(\lfloor N\gamma\rfloor)}$, $\Theta_{(N)})$. Finally, the HPD credible interval of α and β is that interval which has the shortest length.

In order to obtain the MCMC samples from the joint posterior density of α and θ, we use the Metropolis-Hastings (M-H) algorithm. We consider a bivariate normal distribution as the proposal density i.e. $N_2(\mu, \Sigma)$ where Σ is the variance-covariance matrix. It may be noted here that if we generate observations from the bivariate normal distribution, we may get negative values also, which are not possible as the parameters under consideration are positive valued. Therefore, we take the absolute value of the generated observations. Following this, the Metropolis-Hastings algorithm associated with the target density $\pi(\cdot)$ and the proposal density $N_2(\mu, \Sigma)$ produces a Markov chain Θ^i through the following steps.

① Set initial values $\Theta_0 = [\alpha_0, \theta_0]'$.

② Generate new candidate parameter values $\Theta_* = [\alpha_*, \theta_*]'$ from $N_2(\mu, \Sigma)$.

③ Calculate the ratio

$$\rho\left(\Theta_*, \Theta_{i-1}\right) = \min\left\{\frac{\pi(\Theta_*)}{\pi(\Theta_{i-1})}, 1\right\}.$$

④ Draw u from uniform(0,1);

$$\begin{cases}\text{Accept } \Theta_* \text{ as } \Theta_i \text{ if } u < \rho\left(\Theta_*, \Theta_{i-1}\right), \\ \text{If } \Theta_* \text{ is not accepted, then } \Theta_i = \Theta_{i-1}.\end{cases}$$

In using the above algorithm, the problem arises as to how to choose the initial guess. Here, we propose the use of the MLEs of (α, θ), obtained by using the method described in Section 2.1, as initial values for the MCMC process. The choice of covariance matrix Σ is also an important issue; see Natzoufras (2009) for details. One choice for Σ would be the asymptotic variance-covariance matrix $I^{-1}(\hat{\alpha}, \hat{\theta})$. While generating M-H samples by taking $\Sigma = I^{-1}(\hat{\alpha}, \hat{\theta})$, we noted that the acceptance rate for such a choice of Σ is about 15%. By acceptance rate, we mean the proportion of times a new set of values is generated at the iteration stages. It is well known that if the acceptance rate is low, a good strategy is to run a small pilot run using a diagonal Σ as a rough estimate of the correlation structure for the target posterior distribution and then re-run the algorithm using the corresponding estimated variance-covariance matrix; for more details see Gelmen, Carlin, Stern, and Rubin (1995, pp. 334-335). Therefore, we have also used the latter described strategy for the calculations in the following sections.

3. Real data application

In this section, we illustrate our proposed methodology with the real examples. The first dataset considered by us represents the survival times for patients with plasma cell myeloma, already reported in Carbone, Kellerhouse, and Gehan (1967). The data contains the response time to therapy of 112 patients with plasma cell myeloma (a tumour of the bone marrow composed of cells normally found in bone marrow) treated at the National Cancer Institute, Bethesda, Maryland. Figure 2 represents the contour plot of negative log-likelihood for the

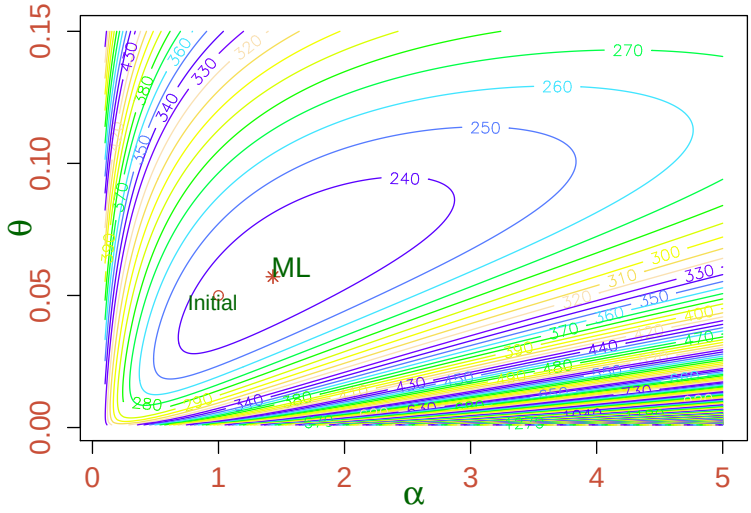


Figure 2: Contour plot for plasma cell myeloma data

considered dataset. The ellipses are obtained by joining those points which have equal values of the negative log-likelihood. Every inner ellipse has a smaller value than that of the outer ellipse. Thus, the innermost ellipse has the minimum value. In other words, the minimum of the minus log-likelihood (maximum of the likelihood) will correspond to the innermost ellipse. We used an arbitrary point (1,0.05) from this innermost ellipse as an initial guess. The MLEs for the dataset are then calculated, using the procedure explained in Section 2.1. Finally, these are obtained as $\hat{\alpha}_{ML} = 1.4325$, $\hat{\theta}_{ML} = 0.0571$. Similarly, a 95% asymptotic confidence intervals for α is obtained as (0.9706, 1.8944) and for β as (0.0420, 0.0727).

To compute the Bayes estimates for the considered dataset, we used the MCMC technique discussed in Section 2.2. Following Robert (2015), we ran three MCMC chains with initial values selected as MLE, MLE - (asymptotic standard deviation) and MLE + (asymptotic standard deviation), respectively. Figure 3 shows the iterations and density plot of samples generated from the posterior distribution using the MCMC technique. From this figure, we see that all the three chains have converged and are well mixed. It is further noted that the posterior of α is approximately symmetric, but the posterior of θ is left skewed. Utilizing these MCMC samples, we computed the Bayes estimates, following the method discussed in Section 2.2, and got $\hat{\alpha}_B = 1.4301$, $\hat{\theta}_B = 0.0581$ under non-informative independent priors. The 95% highest posterior density (HPD) interval estimate for α is obtained as (1.0001,1.6109) and for θ as (0.0424, 0.0719).

The second dataset, considered here, arose in the tests on the endurance of deep groove ball bearings. This data contains the number of million revolutions before failure for each of the 23 ball bearings in the life test and has been reported by (Lawless 2002, pp.228). The data points are exact observations. For the illustration of our methodology, we have generated censored data for a prefixed number of inspections by specifying the inspection times and dropping probabilities.

We fixed the experimentation time as 140 units of time and decided to have 7 inspections during this period. We have considered four different inspection plans. The first plan consists of equally spaced inspection times i.e. at 20, 40, ..., 140 units of time. The next inspection plan is designed under the motivation that if the probability of failure is high during some time interval, an early inspection should be scheduled. Thus, the second inspection plan is based on such a notion. The third inspection plan is designed on the basis of estimated cdf; although such a plan is not feasible in practice we have included it for theoretical interest. First, we calculate $u = F(140, \alpha_{ML}, \theta_{ML})$, then inspection times are obtained as $T_1 = F^{-1}(u/7, \alpha_{ML}, \theta_{ML})$, $T_2 = F^{-1}(2u/7, \alpha_{ML}, \theta_{ML})$, ..., $T_6 = F^{-1}(6u/7, \alpha_{ML}, \theta_{ML})$ and $T_7 = 140$. The fourth inspection plan is chosen so as to

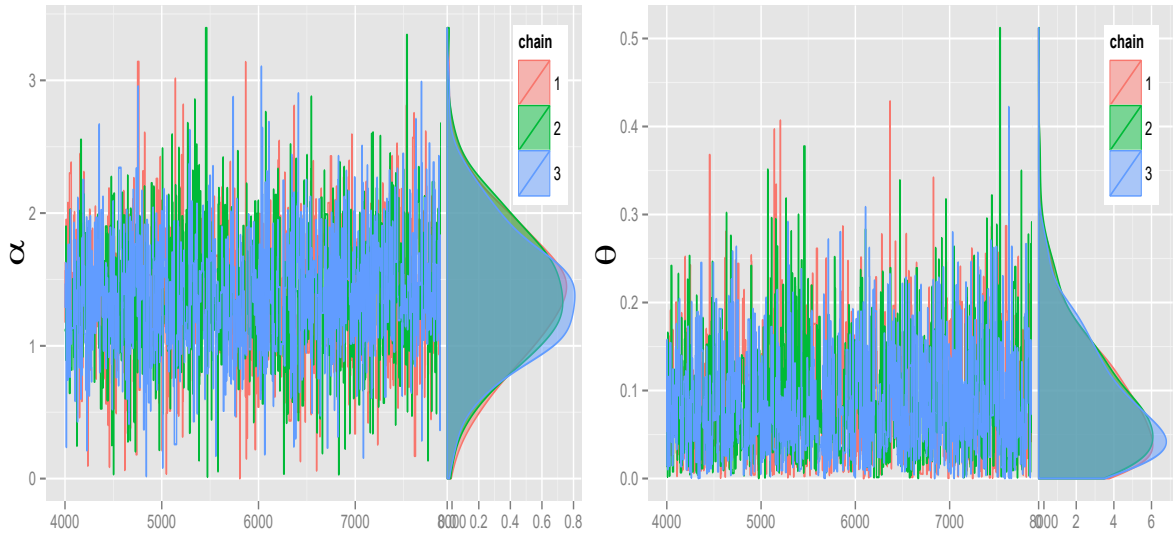


Figure 3: Iteration trace and density plot of MCMC samples for plasma cell myeloma data

have approximately equal probability of failure in each interval of inspection and are approximated to the nearest multiple of 10. The dropping schemes, are selected in the following manner: the first scheme considers the risk of dropping at all the intermediate stages to be zero i.e. $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = 0, p_8 = 1$. In the second scheme, the risk at all stages is equal but not to zero i.e. $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = 0.2, p_8 = 1$. The third scheme is constructed so that the risk of dropping is low in the earlier stages and high in later stages. Contrary to it, in the fourth scheme, the risk of dropping is high in earlier stages and low in the later stages. Lastly, we consider the case when the risk is high at the first stage, but there is no risk at all other stages. These inspection schemes and dropping schemes are summarized in Table 1b and Table 1a, respectively. Under dropping scheme 1 and inspection scheme A, we obtained the number of failures at seven stages as 1, 2, 8, 4, 3, 2, 2 respectively and one dropping at last stage. Following the same procedure, as followed in the previous example, we calculated the ML estimates and Bayes estimates with corresponding interval estimates for the dataset as mentioned above. This result is summarized in the first row of Table 2. The last row of the table provides the ML and Bayes estimates with corresponding interval estimates for complete dataset.

It may be worthwhile to mention here that the number of droppings are random and we are generating the progressive interval type-I censored data from the complete sample data, therefore we can study the average performance of the estimators. For this purpose, we generated 2000 censored datasets of r_i's for given p_i's and accordingly the d_i's from the considered complete dataset. Table 2 provides the average ML and Bayes estimates, along with the AC and HPD interval estimates of the parameters based on the generated censored datasets. It may be seen from this table that the width of the interval estimates under dropping scheme 1, when risk of dropping at all stages is zero, is least of all the estimators under other schemes. It may further be seen that width of the interval estimates under dropping scheme 2 is more than the others. Further, under the 4^{th} scheme the interval width is lesser than those under the 3^{rd} scheme. While studying the effect of inspection time on the performance of the estimators, we noted that the average estimate under inspection scheme A and dropping scheme 1 is close to the estimate obtained for the complete sample case. For other inspection and dropping schemes the average estimates are larger than that obtained for the complete sample case. Similarly, the average width of the interval estimates under scheme A is least among all considered inspections schemes. The width of the interval estimates under scheme B is more than those under scheme A but less than those under scheme C. The width

of the interval estimates under scheme D is largest. It is also noted that as the proportion of droppings increases, the width of the interval estimates increase.

Table 1: Censoring scheme

(a) Dropping Scheme

Name	Dropping Probabilities
1	0, 0, 0, 0, 0, 0, 0, 1
2	0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 1
3	0.2, 0.2, 0.2, 0.1, 0.1, 0.1, 0.1, 1
4	0.1, 0.1, 0.1, 0.1, 0.2, 0.2, 0.2, 1
5	0.25, 0, 0, 0, 0, 0, 0, 1

(b) Inspection Scheme

Name	Inspection times
A	20, 40, 60, 80, 100, 120, 140
C	33.00, 45.60, 51.96, 67.80, 68.88, 98.64, 140
B	36.96, 46.85, 55.88, 65.55, 77.31, 94.54, 140
D	40 50, 65, 70, 100, 120, 140

Table 2: Average ML and Bayes estimates of α and θ along with their AC and HPD intervals for different censoring scheme related to the ball bearing dataset

Inspection Scheme	Dropping scheme	$\hat{\alpha}_{ML}$	AC Interval of α		$\hat{\alpha}_B$	HPD Interval of α		$\hat{\theta}_{ML}$	AC Interval of θ		$\hat{\theta}_B$	HPD Interval of θ	
			Lower	Upper		Lower	Upper		Lower	Upper		Lower	Upper
A	1	5.7260	1.0579	9.3029	5.7261	1.2590	8.2029	0.0198	0.0449	0.0331	0.0326	0.0246	0.0411
A	2	5.6573	0.0000	10.1799	5.6579	0.3648	9.0797	0.0178	0.0454	0.0325	0.0329	0.0241	0.0413
A	3	5.6133	0.0000	9.5056	5.6139	0.0673	8.4048	0.0184	0.0456	0.0322	0.0320	0.0240	0.0417
A	4	5.7297	0.0000	9.4267	5.7302	0.0791	8.3263	0.0184	0.0457	0.0322	0.0323	0.0240	0.0412
A	5	5.6271	0.2748	9.3500	5.6275	0.4651	8.2488	0.0194	0.0452	0.0324	0.0318	0.0245	0.0410
B	1	5.6181	0.0000	9.1913	5.6189	0.0000	8.0901	0.0185	0.0458	0.0331	0.0330	0.0252	0.0417
B	2	7.8164	0.0000	33.0905	7.8160	0.0000	31.9902	0.0170	0.0480	0.0329	0.0332	0.0230	0.0429
B	3	7.9880	0.0000	28.4871	7.9892	0.0000	27.3877	0.0165	0.0472	0.0329	0.0323	0.0236	0.0427
B	4	7.4395	0.0000	17.7741	7.4402	0.0000	16.6733	0.0178	0.0466	0.0328	0.0325	0.0240	0.0423
B	5	6.6683	0.0000	11.3200	6.6683	0.0000	9.0184	0.0192	0.0461	0.0337	0.0324	0.0246	0.0424
C	1	5.8681	0.0000	9.4332	5.8698	0.0000	8.3330	0.0188	0.0456	0.0332	0.0328	0.0250	0.0420
C	2	8.6784	0.0000	46.1006	8.6785	0.0000	44.9996	0.0167	0.0486	0.0340	0.0338	0.0231	0.0440
C	3	7.4340	0.0000	38.6957	7.4343	0.0000	37.5957	0.0164	0.0486	0.0347	0.0333	0.0232	0.0437
C	4	7.0231	0.0000	22.1019	7.0241	0.0000	21.0007	0.0174	0.0476	0.0349	0.0337	0.0240	0.0426
C	5	6.4037	0.0000	12.6525	6.4052	0.0000	9.5506	0.0185	0.0468	0.0347	0.0339	0.0242	0.0424
D	1	7.0398	0.0000	21.8711	7.0412	0.0000	20.7705	0.0178	0.0469	0.0330	0.0327	0.0246	0.0424
D	2	10.794	0.0000	64.2911	10.7940	0.0000	63.1905	0.0161	0.0509	0.0347	0.0336	0.0227	0.0441
D	3	13.624	0.0000	47.2659	13.6252	0.0000	46.1648	0.0162	0.0506	0.0330	0.0325	0.0229	0.0436
D	4	9.4267	0.0000	23.0812	9.4280	0.0000	19.9801	0.0169	0.0487	0.0324	0.0331	0.0236	0.0432
D	5	8.4054	0.0000	13.7967	8.4067	0.0000	12.6970	0.0182	0.0477	0.0328	0.0330	0.0245	0.0427
Complete		5.2525	1.2716	9.2933	—	1.572	9.3819	0.0322	0.0449	—	0.0319	0.0256	0.0427

4. Simulation study

In this section, we have compared the performances of the various estimators on the basis of their bias and mean square error (MSE). It may be mentioned here that the exact expressions for the bias and mean square errors cannot be obtained, because the estimators are not in closed form. Therefore, biases and MSEs are estimated on the basis of a Monte-Carlo simulation study of 2000 samples. For this purpose we generated a specified number of observations from the distribution given in equation (1) for arbitrarily fixed values of the parameters under the specified censoring schemes and calculated different estimates of α and θ following the procedure described in the previous sections. This process was repeated 2000 times to obtain the simulated biases and MSEs. We have computed the MLEs by using the Newton-Raphson algorithm. The estimates of (α, θ) obtained through the Newton-Raphson algorithm are denoted as $(\alpha_{ML}, \theta_{ML})$, respectively. It is noted that Newton-Raphson algorithm has a convergence rate of 85%-90%. We have reported the results omitting the cases where the algorithms do not converge. To simulate a progressive interval type-I censored sample from the considered distribution, we have used the algorithm given by Balakrishnan and Cramer (2014, pp.200) after modifying step ④ as : Determine the number of droppings at the j^{th} stage by generating r_j from $Bin(n - \sum_{j=1}^{i-1}(d_j + r_j), p_j)$.

It may be noted here that the MSE and bias of these estimators will depend on the sample size n, values of α, θ and hyperparameters λ_1, λ_2, ν_1 and ν_2. We considered a number of values

for the sample size n; namely n = 20, 30, 40 and 50. For the choice of the hyper-parameters of the prior distribution, we have considered one set of values as $\lambda_1 = \lambda_2 = \nu_1 = \nu_2 = 0$ which reduces the prior to a non-informative prior. For an informative prior, the hyperparameters are chosen on the basis of the information possessed by the experimenter. In most cases, the experimenter can have a notion of what are the expected value of the parameter and can always associate a degree of belief to this value. In other words, the experimenter can specify the prior mean and prior variance for the parameters. The prior mean reflects the experimenter's belief about the parameter in the form of its expected value, and the prior variance reflects his confidence in this expected value. Keeping this point in mind, we have chosen the hyper-parameters in such a way that the prior mean is equal to the true value of the parameter, and the belief in the prior mean is either strong or weak, i.e. the prior variance is small or large, respectively; for details see Singh *et al.* (2011). The bias of the estimates of parameters, reliability and hazard rate with corresponding MSEs have been calculated, and the results are summarized in Table 3, 4 and 5.

Table 3 provides the absolute bias and MSE of estimates of the parameters along with the reliability and hazard rate at time $t = 1$ for $\alpha = 2.5$, $\theta = 2$ and inspection times 0.2, 0.4 ,0.6, 0.8, 1.0, 1.2, 1.4, 1.6. It can be seen from the table that in general the bias and MSEs decrease as n increases in all the considered cases. It can also be seen that the MSE of the MLE is more than that of the corresponding Bayes estimate in all cases, but the difference between the MSEs of the Bayes and ML estimates decreases for increases in the value of n. It is noted here that bias of the estimates and MSEs under censoring scheme 1 are approximately equal to that of complete sample case (denoted as scheme 0) and smaller than those under other schemes. In most cases it is observed that the bias and MSE under dropping scheme 1 are smallest followed by scheme 5, 4, 3 and 2 sequentially. Bias and MSE of the reliability estimate show a similar trend as observed for the parameter estimates.

Table 4 provides the absolute bias and MSE of the various estimators for different choices of model parameters. Above we noted that as the sample size increases the Bias and MSE decrease, therefore we have reported the results for n=30 only. Similarly, we noted above that under dropping scheme 1 the performance of the estimates are as good as the complete sample case and better than all other schemes. Therefore, we have reported the results for the complete sample case and scheme 1 and scheme 4 only. It may be seen from the table that the bias and MSE of all the considered estimates of α, θ, reliability $S_{ML}(t = 1)$ and hazard rate $H_{ML}(t = 1)$ increase as α increases and/or as θ increases. It is interesting to note that the bias and MSE of all the estimates are smaller when the proportion of droppings are smaller. All the estimates under scheme 1 have, more or less, a similar bias and MSE as that obtained for the complete case; but the bias and MSE of the estimates under scheme 4 are a little larger. The bias and MSE of the Bayes estimates obtained using various priors are presented in Table 5, and we see that as prior confidence in the guessed value increases the MSE decreases.

5. Conclusions

In the present piece of work, we have considered both Classical and Bayesian analysis for the progressive interval type-I censored data when the lifetime of the items follows generalised exponential distribution. The ML estimates do not have explicit forms. Therefore, the Newton-Raphson algorithm has been proposed to compute the MLEs. The Bayes estimates under the squared error loss function also do not exist in explicit form, but, Bayes estimates can be routinely obtained through the use of MCMC technique considering the shape and scale parameters having independent gamma priors. On the basis of this study, we may conclude that the proposed estimation procedures under progressive interval type-I censoring with specific choices of the scheme can be easily implemented. It is also seen that the inspection scheme and dropping schemes have an effect on the performance of the estimators. Thus, if it is possible, it is better to choose a scheme resulting in a fewer number of droppings.

Table 3: Simulated bias (MSE) of estimates of parameters, reliability and hazard rate for fixed $\alpha = 2.5$, $\theta = 2$ and inspection time 0.2(0.2)1.6.

n	Dropping Scheme	α_{ML}	θ_{ML}	α_B	θ_B	$S_{ML}(t=1)$ [a]	$H_{ML}(t=1)$ [b]
20	0[c]	0.5718(1.1600)	0.2025(0.5122)	0.4856(0.7593)	0.0691(0.2791)	0.0056(0.0070)	0.3079(0.1870)
	1	0.5932(1.3333)	0.1842(0.5699)	0.5863(0.9271)	0.0385(0.3443)	0.0057(0.0069)	0.3085(0.1757)
	2	2.0746(4.2198)	0.3975(0.8522)	2.0922(3.7516)	0.2707(0.6738)	0.0065(0.0089)	0.4007(0.3911)
	3	1.7536(3.1505)	0.3920(0.8764)	1.6912(2.6920)	0.2597(0.5979)	0.0059(0.0077)	0.3093(0.3838)
	4	1.7491(3.1034)	0.3861(0.8202)	1.4365(2.6664)	0.1939(0.5585)	0.0058(0.0073)	0.3091(0.3769)
	5	1.0191(1.8213)	0.2974(0.7259)	0.8832(1.4055)	0.1025(0.4666)	0.0057(0.0072)	0.3094(0.2854)
30	0	0.3437(0.8028)	0.1147(0.4058)	0.0998(0.5197)	0.0152(0.2521)	0.0048(0.0047)	0.3079(0.1020)
	1	0.3427(0.8692)	0.1003(0.4130)	0.1557(0.6105)	0.0091(0.2262)	0.0048(0.0046)	0.3082(0.1022)
	2	0.5693(1.3154)	0.2069(0.6577)	0.3680(1.0635)	0.0814(0.4828)	0.0059(0.0059)	0.4008(0.1036)
	3	0.5596(1.2533)	0.1696(0.6296)	0.3012(1.0403)	0.0552(0.4668)	0.0054(0.0054)	0.3088(0.1031)
	4	0.5194(1.2118)	0.1619(0.5703)	0.3989(0.9591)	0.0301(0.4670)	0.0052(0.0053)	0.3091(0.1027)
	5	0.4322(1.0940)	0.1379(0.5208)	0.2671(0.8896)	0.0015(0.3743)	0.0050(0.0050)	0.3084(0.1026)
40	0	0.2522(0.7110)	0.0957(0.3491)	0.0910(0.5352)	0.0047(0.2438)	0.0043(0.0034)	0.3071(0.0960)
	1	0.2536(0.8036)	0.0977(0.3891)	0.0882(0.6733)	0.0198(0.2541)	0.0044(0.0035)	0.3074(0.0958)
	2	0.4758(1.0459)	0.1165(0.7106)	0.2164(0.9158)	0.0761(0.6127)	0.0056(0.0047)	0.3996(0.0975)
	3	0.4140(1.0447)	0.1552(0.7056)	0.3396(0.8940)	0.0386(0.3859)	0.0053(0.0040)	0.3085(0.0965)
	4	0.3450(0.9457)	0.1797(0.5425)	0.1757(0.8136)	0.0327(0.3185)	0.0049(0.0039)	0.3083(0.0969)
	5	0.3318(0.8901)	0.1106(0.4865)	0.1575(0.6956)	0.0316(0.5886)	0.0049(0.0037)	0.3080(0.0961)
50	0	0.2477(0.7118)	0.0949(0.3488)	0.0332(0.6452)	0.0339(0.1837)	0.0040(0.0029)	0.3065(0.0853)
	1	0.2537(0.8043)	0.0971(0.3895)	0.0078(0.7000)	0.0161(0.2413)	0.0041(0.0030)	0.3068(0.0862)
	2	0.3322(0.9056)	0.1096(0.6144)	0.0955(0.7610)	0.0465(0.4660)	0.0048(0.0043)	0.3069(0.0866)
	3	0.3148(0.8990)	0.1198(0.4954)	0.0349(0.8323)	0.0123(0.4802)	0.0046(0.0037)	0.3066(0.0859)
	4	0.3011(0.8619)	0.1015(0.4742)	0.0879(0.7333)	0.0022(0.4111)	0.0043(0.0034)	0.3066(0.0860)
	5	0.2499(0.8065)	0.0885(0.4326)	0.0247(0.7547)	0.0109(0.3040)	0.0041(0.0030)	0.3068(0.0860)

[a] Here, true value of reliability at time 1 is $S(1) = 0.3048$

[b] True value of hazard rate at time 1 is $H(1) = 1.7851$

[c] 0 means complete case, when no dropping and data points collected continuously

Table 4: Simulated bias (MSE) of estimates of parameters, reliability and hazard rate for various choice of parameters and fixed $n = 30$

α	θ	Dropping Scheme	α_{ML}	θ_{ML}	α_B	θ_B	$S_{ML}(t=1)$	$H_{ML}(t=1)$
	0.5	0	0.0329(0.0145)	0.0324(0.0105)	0.0275(0.0130)	0.0210(0.0101)	0.0003(0.0007)	0.0487(0.0219)
	0.5	1	0.0366(0.0143)	0.0330(0.0125)	0.0313(0.0141)	0.0245(0.0111)	0.0060(0.0008)	0.0501(0.0215)
	0.5	4	0.0452(0.0152)	0.0336(0.0132)	0.0399(0.0149)	0.0310(0.0131)	0.0063(0.0009)	0.0541(0.0224)
	1.5	0	0.0345(0.0143)	0.0903(0.0964)	0.0318(0.0130)	0.0743(0.0904)	0.0029(0.0024)	0.1577(0.2474)
0.5	1.5	1	0.0413(0.0155)	0.0913(0.0964)	0.0403(0.0142)	0.0755(0.0948)	0.0082(0.0026)	0.1627(0.2476)
	1.5	4	0.0471(0.0148)	0.0976(0.0968)	0.0448(0.0152)	0.0776(0.0958)	0.0095(0.0027)	0.1574(0.2482)
	2.5	0	0.0449(0.0149)	0.1452(0.2642)	0.0429(0.0152)	0.1301(0.2527)	0.0036(0.0045)	0.2797(0.7334)
	2.5	1	0.0455(0.0156)	0.1451(0.2652)	0.0427(0.0156)	0.1330(0.2558)	0.0042(0.0046)	0.2765(0.7330)
	2.5	4	0.0481(0.0166)	0.1454(0.2666)	0.0454(0.0154)	0.1337(0.2565)	0.0077(0.0048)	0.2728(0.7333)
	0.5	0	0.1641(0.2270)	0.0360(0.0133)	0.1302(0.2242)	0.0312(0.0081)	0.0024(0.0021)	0.0083(0.0074)
	0.5	1	0.1712(0.2273)	0.0435(0.0140)	0.1343(0.2253)	0.0324(0.0092)	0.0038(0.0024)	0.0114(0.0076)
	0.5	4	0.1808(0.2311)	0.0440(0.0141)	0.1385(0.2282)	0.0410(0.0122)	0.0074(0.0026)	0.0042(0.0075)
	1.5	0	0.1878(0.2309)	0.1037(0.1193)	0.1416(0.2243)	0.0866(0.1121)	0.0020(0.0043)	0.0869(0.1132)
1.5	1.5	1	0.1912(0.2324)	0.1068(0.1197)	0.1455(0.2241)	0.0873(0.1135)	0.0167(0.0044)	0.0851(0.1141)
	1.5	4	0.1956(0.2388)	0.1097(0.1202)	0.1516(0.2318)	0.0883(0.1194)	0.0182(0.0045)	0.0891(0.1139)
	2.5	0	0.1892(0.2349)	0.1875(0.3345)	0.1453(0.2340)	0.1543(0.3241)	0.0013(0.0048)	0.1850(0.3636)
	2.5	1	0.1923(0.2362)	0.1918(0.3346)	0.1470(0.2358)	0.1582(0.3241)	0.0026(0.0050)	0.1833(0.3637)
	2.5	4	0.1994(0.2382)	0.1986(0.3352)	0.1485(0.2375)	0.1584(0.3248)	0.0039(0.0052)	0.1904(0.3635)
	0.5	0	0.3222(0.8562)	0.0642(0.0285)	0.2822(0.8474)	0.0368(0.0270)	0.0049(0.0017)	0.0090(0.0040)
	0.5	1	0.3285(0.8576)	0.0719(0.0287)	0.2823(0.8474)	0.0383(0.0268)	0.0203(0.0018)	0.0094(0.0036)
	0.5	4	0.3351(0.8579)	0.0786(0.0297)	0.2863(0.8473)	0.0406(0.0277)	0.0224(0.0021)	0.0117(0.0042)
	1.5	0	0.3476(0.8634)	0.1682(0.2517)	0.2916(0.8413)	0.1263(0.2469)	0.0003(0.0038)	0.0687(0.0739)
2.5	1.5	1	0.3499(0.8644)	0.1698(0.2525)	0.3017(0.8455)	0.1318(0.2482)	0.0019(0.0039)	0.0502(0.0744)
	1.5	4	0.3527(0.8645)	0.1706(0.2532)	0.3082(0.8453)	0.1354(0.2496)	0.0026(0.0043)	0.0612(0.0737)
	2.5	0	0.3909(0.9115)	0.2839(0.6999)	0.3150(0.9071)	0.2353(0.6930)	0.0045(0.0051)	0.1421(0.3133)
	2.5	1	0.3919(0.9125)	0.2826(0.7000)	0.3226(0.9110)	0.2393(0.6969)	0.0062(0.0053)	0.1357(0.3136)
	2.5	4	0.3995(0.9204)	0.2854(0.7008)	0.3310(0.9149)	0.2420(0.6970)	0.0114(0.0054)	0.1262(0.3146)
	3.0	0	0.5388(0.2176)	0.5116(0.1683)	0.4087(0.2304)	0.3261(0.1829)	0.0396(0.0202)	0.7741(0.3364)
	3.0	1	0.5844(0.2234)	0.5090(0.2120)	0.4815(0.2212)	0.3829(0.2117)	0.1274(0.0478)	0.8046(0.3540)
	3.0	4	0.6898(0.2692)	0.5463(0.2120)	0.6183(0.2369)	0.4740(0.2294)	0.1364(0.0286)	0.8580(0.3517)
	4.0	0	0.5530(0.2303)	1.3648(1.4693)	0.4819(0.2352)	1.1533(1.3866)	0.0560(0.0566)	2.3702(3.7346)
3.0	4.0	1	0.6703(0.2267)	1.4112(1.4771)	0.6596(0.2532)	1.1435(1.4505)	0.1490(0.0812)	2.4594(3.7510)
	4.0	4	0.7226(0.2649)	1.4730(1.4784)	0.7047(0.2760)	1.1856(1.4579)	0.1750(0.0492)	2.3688(3.7659)
	5.0	0	0.7038(0.2371)	2.1954(4.0110)	0.6798(0.2532)	2.0040(3.8336)	0.0736(0.0736)	4.2306(11.0138)
	5.0	1	0.6992(0.2392)	2.1816(3.9871)	0.6577(0.2589)	1.9941(3.8587)	0.0733(0.1028)	4.1922(11.0185)
	5.0	4	0.7512(0.2888)	2.2243(4.0108)	0.7009(0.2402)	2.0491(3.8652)	0.1312(0.0979)	4.1031(11.0093)
	3.0	0	2.4743(3.4426)	0.5779(0.2462)	1.9738(3.3973)	0.4890(0.1667)	0.0522(0.0520)	0.1438(0.1548)
	3.0	1	2.5733(3.4655)	0.6817(0.2338)	2.0302(3.3962)	0.4975(0.1861)	0.0636(0.0411)	0.2144(0.1460)
	3.0	4	2.7376(3.4884)	0.6648(0.2113)	2.1179(3.4287)	0.6501(0.2158)	0.1137(0.0763)	0.1135(0.1639)
	4.0	0	2.8501(3.4821)	1.5624(1.8098)	2.1362(3.3767)	1.3086(1.7272)	0.0418(0.0946)	1.3455(1.7215)
4.0	4.0	1	2.8971(3.4904)	1.6341(1.8392)	2.2153(3.3794)	1.3131(1.7141)	0.2563(0.0930)	1.3069(1.7546)
	4.0	4	2.9686(3.5936)	1.6931(1.8472)	2.2918(3.4908)	1.3620(1.7989)	0.2984(0.0876)	1.3602(1.7181)
	5.0	0	2.8733(3.5557)	2.8329(5.0180)	2.2141(3.5167)	2.3504(4.8665)	0.0381(0.1063)	2.7912(5.4545)
	5.0	1	2.9381(3.5388)	2.8915(5.0673)	2.2219(3.5526)	2.3802(4.8844)	0.0744(0.1164)	2.7914(5.4672)
	5.0	4	3.0215(3.5968)	3.0140(5.0419)	2.2480(3.5922)	2.4060(4.9194)	0.0857(0.0869)	2.8584(5.4640)
	3.0	0	4.8491(12.8567)	0.9865(0.4531)	4.2426(12.7171)	0.5767(0.4308)	0.0875(0.0392)	0.1381(0.0866)
	3.0	1	4.9642(12.8929)	1.1139(0.4607)	4.2362(12.7365)	0.5975(0.4296)	0.3245(0.0346)	0.1655(0.0980)
	3.0	4	5.0702(12.8907)	1.2088(0.4734)	4.3373(12.7514)	0.6192(0.4602)	0.3430(0.0472)	0.1805(0.0961)
	4.0	0	5.2182(12.9806)	2.5490(3.7884)	4.4249(12.6350)	1.9183(3.7509)	0.0387(0.1010)	1.0399(1.1152)
5.0	4.0	1	5.2887(13.064)	2.5670(3.8373)	4.5390(12.7071)	1.9886(3.7638)	0.0765(0.0882)	0.7663(1.1307)
	4.0	4	5.2959(12.9793)	2.5738(3.8091)	4.6375(12.7009)	2.0625(3.7921)	0.0876(0.0708)	0.9403(1.1580)
	5.0	0	5.8731(13.7266)	4.2612(10.5378)	4.7567(13.6204)	3.5345(10.4273)	0.0885(0.0966)	2.1410(4.7334)
	5.0	1	5.9029(13.6974)	4.2759(10.5085)	4.8633(13.6841)	3.6289(10.4968)	0.1249(0.1087)	2.0771(4.7301)
	5.0	4	6.0106(13.8543)	4.3315(10.5531)	4.9905(13.7422)	3.6310(10.4785)	0.2096(0.0827)	1.9228(4.7278)

Table 5: Average Bayes estimates (MSE) and 95% HPD intervals based on data simulated under dropping scheme 1 for different choices of prior parameters

$g_1(\alpha)$	$g_2(\theta)$	$\hat{\alpha}_B$	HPD Interval	$\hat{\theta}_B$	HPD Interval
G(4,2)[1]	G(4,2)	0.0849(0.8064)	(3.6262, 2.1042)	0.0111(0.2759)	(1.3450, 2.6745)
G(4,2)	G(1,0.5)	0.0971(0.8342)	(2.0054, 3.7263)	0.0275(0.3157)	(1.3070, 2.7325)
G(4,2)	G(0.4,0.2)	0.1139(0.9135)	(1.8692, 3.7963)	0.0412(0.3564)	(1.0127, 2.9745)
G(1,0.5)	G(4,2)	0.1135(0.9075)	(1.9507, 3.7702)	0.0198(0.3022)	(1.2016, 2.8618)
G(1,0.5)	G(1,0.5)	0.1330(0.9275)	(1.8068, 3.9263)	0.0390(0.3440)	(1.1129, 2.9044)
G(1,0.5)	G(0.4,0.2)	0.1459(0.9324)	(1.7864, 3.9292)	0.0501(0.3674)	(0.9977, 3.0199)
G(0.4,0.2)	G(0.4,0.2)	0.1460(0.9454)	(1.7001, 4.5137)	0.0566(0.3624)	(0.9901, 3.1198)

However, in most practical situations the dropping scheme is not controllable. Therefore, in such situations, the inspection plan should be designed as to result in the least number of droppings. However, under any scheme, the proposed method can be used to obtain the estimates.

We have not considered any covariates in this paper, but in practice often the covariates may be present. It will be interesting to develop statistical procedures for the estimation of the unknown parameters in the presence of covariates. Further, we have considered dropping probabilities at each stage to be fixed, but in real life, these may be random, and a suitable model to capture this randomness can be developed. The work in this direction is under progress.

Acknowledgements

The authors gratefully acknowledge the *Council of Scientific and Industrial Research (CSIR); New Delhi, India* for providing financial assistance to this work. The authors also thankful to journal editor and reviewers for their valuable and constructive comments that help improve the manuscript.

References

Aggarwala R (2001). "Progressive Interval Censoring: Some Mathematical Results with Applications to Inference." *Communications in Statistics - Theory and Methods*, **30**(8-9), 1921–1935.

Ashour SK, Afify WM (2007). "Statistical Analysis of Exponentiated Weibull Family under Type I Progressive Interval Censoring with Random Removals." *Journal of Applied Sciences Research*, **3**(12), 1851–1863.

Balakrishnan N, Cramer E (2014). *The Art of Progressive Censoring*. Springer New York.

Berger JO (2013). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Berger JO, Sun D (1993). "Bayesian Analysis for the Poly-Weibull Distribution." *Journal of the American Statistical Association*, **88**, 1412–1418.

Carbone PP, Kellerhouse LE, Gehan EA (1967). "Plasmacytic Myeloma: A Study of the Relationship of Survival to Various Clinical Manifestations and Anomalous Protein Type in 112 Patients." *The American Journal of Medicine*, **42**(6), 937 – 948. ISSN 0002-9343.

Chen DG, Lio YL (2010). "Parameter Estimations for Generalized Exponential Distribution under Progressive Type-I Interval Censoring." *Computational Statistics and Data Analysis*, **54**(6), 1581–1591.

Gelmen A, Carlin J, Stern H, Rubin D (1995). *Bayesian Data Analysis*. Text in Statistical Science, Chapman & Hall.

Gompertz B (1825). "On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies." *Philosophical Transactions of the Royal Society London*, **115**, 513–585.

Gupta RD, Kundu D (2001a). "Exponentiated Exponential Family : An Alternative to Gamma and Weibull Distributions." *Biometrical journal*, **43**, 117 – 130.

Gupta RD, Kundu D (2001b). "Generalized Exponential Distribution : Different Method of Estimations." *Journal of Statistical Computations and Simulations*, **69**, 315–337.

Jaheen ZF (2004). "Empirical Bayes Inference for Generalized Exponential Distribution Based on Records." *Communication in Statistics: Theory and Methods*, **33(8)**, 1851 – 1861.

Kondolf G, Adhikari A (2000). "Weibull vs. Lognormal Distributions for Fluvial Gravels." *Journal of Sedimentary Research*, **70**(3).

Lawless JF (2002). *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Mudholkar GS, Srivastava DK (1993). "Exponentiated Weibull Family for Analyzing Bath Tub Failure Data." *IEEE Transition in Reliability*, **42**, 299 – 302.

Natzoufras I (2009). *Bayesian Modeling Using WinBugs*. Series B, Wiley.

Raquab MZ, Madi MT (2005). "Bayesian Inference for the Generalized Exponential Distribution." *Journal of Statistical Computation and Simulation*, **75(10)**, 841 – 852.

Robert CP (2015). "The Metropolis-Hastings Algorithm." *arXiv:1504.01896v1 [stat.CO]*.

Sarhan AM (2007). "Analysis of Incomplete Censored Data in Competing Risks Models with Generalized Exponential Distribution." *IEEE Trans Reliability*, **56**, 132 – 138.

Singh R, Singh SK, Singh U, Prakash G (2008). "Bayes Estimator of Generalized Exponential Parameters under LINEX Loss Function using Lindley's Approximation." *Data Science Journal*, **7**, 65–75.

Singh SK, Singh U, Kumar D (2011). "Estimation of Parameters and Reliability Function of Exponentiated Exponential Distribution: Bayesian Approach Under General Entropy Loss Function." *Pakistan Journal of Statistics and Operational Research*, **VII**(2), 199–216.

Singh SK, Singh U, Kumar M (2013). "Estimation of Parameters of Exponentiated Pareto Model for Progressive Type-II Censored Data with Binomial Removals Using Markov Chain Monte Carlo Method." *International Journal of Mathematics and Computation*, **21**(4), 88–102.

Singh SK, Singh U, Kumar M (2014). "Bayesian Estimation for Poission-exponential Model under Progressive Type-II Censoring Data with Binomial Removal and Its Application to Ovarian Cancer Data." *Communications in Statistics-Simulation and Computation*, (just-accepted), 00–00.

Tse SK, Yang C, Yuen HK (2000). "Statistical Analysis of Weibull Distributed Lifetime Data under Type II Progressive Censoring with Binomial Removals." *Journal of Applied Statistics*, **27**(8), 1033–1043. ISSN 0266-4763.

Yuen HK, Tse SK (1996). "Parameters Estimation for Weibull Distributed Lifetimes under Progressive Censoring with Random Removals." *Journal of Statistical Computation and Simulation*, **55**(1-2), 57–71.

Zheng G (2002). "Fisher Information Matrix in Type II Censored Data from Exponentiated Exponential Family." *Biometrical Journal*, **44**, 353 – 357.

Appendix

$$
\begin{aligned}
L_{\alpha\alpha} &= \sum_{i=1}^{m} d_i \frac{\left({\phi_i}^{\alpha}(\ln\phi_i)^2 - \phi_{i-1}^{\alpha}(\ln\phi_{i-1})^2\right)}{\left(\phi_i^{\alpha} - \phi_{i-1}^{\alpha}\right)} \\
&\quad - d_i \frac{\left({\phi_i}^{\alpha}\ln\phi_i - \phi_{i-1}^{\alpha}\ln\phi_{i-1}\right)^2}{\left(\phi_i^{\alpha} - \phi_{i-1}^{\alpha}\right)^2} \\
&\quad - \sum_{i=1}^{m+1} r_i \frac{{\phi_{i-1}}^{\alpha}\left(\ln\phi_{i-1}\right)^2}{1 - {\phi_{i-1}}^{\alpha}} - r_i \frac{{\phi_{i-1}}^{2\alpha}\left(\ln\phi_{i-1}\right)^2}{\left(1 - {\phi_{i-1}}^{\alpha}\right)^2} \\
L_{\theta\theta} &= \sum_{i=1}^{m} -d_i \frac{\left(\phi_i^{\alpha-1}\alpha\,\xi_i - \phi_{i-1}^{\alpha-1}\alpha\,\xi_{i-1}\right)^2}{\left(\phi_i^{\alpha} - \phi_{i-1}^{\alpha}\right)^2} \\
&\quad + d_i\alpha \frac{\phi_i^{\alpha-2}\xi_i(\alpha\xi_i - \phi_i T_i - \xi_i) - \phi_{i-1}^{\alpha-2}\xi_{i-1}(\alpha\xi_{i-1} - \phi_{i-1}T_{i-1} - \xi_{i-1})}{\left(\phi_i^{\alpha} - \phi_{i-1}^{\alpha}\right)} \\
&\quad - \sum_{i=1}^{m+1} r_i \frac{\phi_{i-1}^{\alpha-2}\alpha\xi_{i-1}(\alpha\xi_{i-1} - \phi_{i-1}T_{i-1} - \xi_{i-1})}{1 - \phi_{i-1}^{\alpha}} - r_i \frac{\phi_{i-1}^{2\alpha}\alpha^2{\xi_{i-1}}^2}{\phi_{i-1}^2\left(1 - \phi_{i-1}^{\alpha}\right)^2} \\
L_{\theta\alpha} &= L_{\alpha\theta} = \sum_{i=1}^{m} d_i \frac{\left(\psi_i\phi_i^{\alpha}\ln\phi_i - \psi_{i-1}\phi_{i-1}^{\alpha}\ln\phi_{i-1}\right)}{\left(\phi_i^{\alpha} - \phi_{i-1}^{\alpha}\right)} \\
&\quad - \frac{(\psi_i - \psi_{i-1})(\phi_i^{\alpha}\ln\phi_i - \phi_{i-1}^{\alpha}\ln\phi_{i-1})}{\left(\phi_i^{\alpha} - \phi_{i-1}^{\alpha}\right)^2} \\
&\quad - \sum_{i=1}^{m+1} r_i \frac{\phi_{i-1}^{\alpha}\ln(1-\phi_{i-1})(1-\phi_{i-1})^{\alpha} + \psi_{i-1}^2}{\left(1 - \phi_{i-1}^{\alpha}\right)^2}
\end{aligned}
$$

where, $\phi_i = \phi_i(\theta, T_i) = 1 - e^{-\theta T_i}$ and $\psi_i = \frac{d}{d\theta}\phi_i = T_i e^{-\theta T_i}$

Affiliation:

Arun Kaushik
Department of Statistics, Institute of Science
Banaras Hindu University
Varanasi, India-221005
E-mail: arundevkauhsik@gmail.com
URL: arun-kaushik.github.io

Austrian Journal of Statistics
published by the Austrian Society of Statistics
http://www.ajs.or.at/
http://www.osg.or.at/

Volume 46
February 2017

Submitted: 2016-05-04
Accepted: 2016-09-07

Austrian Journal of Statistics
February 2017, Volume 46, 49–63.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i2.569

Compositional Analysis of Trade Flows Structure

Klára Hrůzová
Palacký University,
Czech Republic

Miroslav Rypka
Palacký University,
Czech Republic

Karel Hron
Palacký University,
Czech Republic

Abstract

Statistical analysis of trade flows structure can significantly help to reveal or to confirm important macroeconomic phenomena. Because of relative character of these multivariate observations, application of standard multivariate methods directly to raw data can lead to meaningless results, affected by trade sizes of different countries. As a way out, it is proposed to employ the logratio methodology that is able to capture interesting features through logratios between compositional parts. Particularly, the perturbation operation together with clr coefficients for coordinate representation of compositions seem to be easy to handle and to interpret for the purpose. Popular exploratory tools, principal component analysis and PARAFAC modeling of three-way data, resulting from a long-term study of the export/import structure, are applied in the compositional context for data from OECD and WIOD databases. The results show that the logratio methodology enables to reveal interesting features of world trade flows and thus provides a preferable alternative to existing exploratory tools.

Keywords: compositional data, perturbation, principal component analysis, PARAFAC, export and import.

1. Introduction

In today's globalised world, export and import play an important role in the country's economic situation. Globalisation causes growth of international trade in goods and services and two structural changes in trade patterns: the increasing importance of emerging economies and rapid growth of trade in intermediate goods as a result of vertical specialisation, meaning that each country is specialised in one or more innovation and production processes and thus it is common for the value chain of a particular final product to span several countries. Trade in intermediate goods currently represents about 56 % of total global trade in goods (Miroudot, Lanz, and Ragoussis 2009) and therefore we intend to explore trade flows broken down by end-use categories to better monitor international trade patterns.

As emphasized in Rodrik (2006) and Hausmann, Hwang, and Rodrik (2007), it is no longer important how much a country exports, but what it exports. Moreover, even manufacturing processes are fragmented, which means that tasks requiring low-skilled labour (e.g. assembling, control) are off-shored to developing countries (or countries with lower labour costs). This contributes significantly to the amount of exports while the value added to the product

in developing countries may be small. Consequently, much more interest in the part is devoted to relative structure of export rather than to its amount in absolute numbers.

The fragmentation of the manufacturing process can be analyzed using input-output tables (see Stehrer, Foster, and de Vries 2010; Timmer, Erumban, Gouma, Los, Temurshoev, de Vries, and Arto 2012; Timmer, Los, Stehrer, and de Vries 2013). The value added may be splitted up by production factors. For the purposes of this article, we distinguish capital, low-skilled, medium-skilled and high-skilled work. We will compare (relative) shares of these factors in value added exports (i.e. domestic value added embodied in final expenditures abroad). Of course, export structure is closely linked to import shares, so they cannot be analyzed separately in order to obtain concise and predicative results.

The aim of the article is to introduce appropriate statistical techniques for analysis and visualization of structure of trade flows in goods. Since we focus on the structure of trade flows, the absolute values of exports and imports are no longer relevant for the analysis. Thus we consider the data as compositional, i.e. carrying only relative information, which leads to a new perspective to the data processing. Although this perspective is recently intensively discussed in many applied fields from geochemistry and chemometrics to social sciences (Pawlowsky-Glahn and Buccianti 2011), just a few papers were published with purely an economic motivation (see Fry 2011, and references therein). On the contrary, even when the authors are aware of relative nature of the underlying economic data, this feature is mostly not (or just sloppily) taken into account for the statistical analysis (Blejer and Fernandez 1980; Devarajan, Swaroop, and Zou 1996).

In the next section, the basics of logratio methodology to compositional data analysis, essential for the purposes of this article, will be recalled together with two speficic methods, applied in the following — principal component analysis and PARAFAC. A particular focus will be devoted to the operation of perturbation, linked to the geometrical structure of compositional data, that enables easily to link the export and import structure of countries. Accordingly, in Section 3, the theoretical contributions are applied to the real-world data of exports and imports, where their structure is explored with respect to end-use categories and factors in value added. In the last section the results are briefly discussed.

2. Logratio methodology to compositional data analysis

To motivate the concept of compositional data, the basic idea will be explained with an example. Let household expenditures on housing, foodstuffs, other goods (including clothing, footwear and durable goods) and services in various countries are of interest. Obviously, their absolute values is hardly comparable due to different price level in each country. On the other hand, the relative structure of expenditures (that can be expressed, e.g., in proportions or percentages) can be quite similar. Consequently, ratios between components as a source of the relevant information, which remains unaltered with any scaling performed, can much better reflect specific situation in various countries than by processing the raw input data. Therefore, when the relative information is of main interest, the sum of components (leading to expression in the local currency, proportions, etc.) should not affect the result of statistical processing. We refer to *scale invariance* of compositional data which is completely violated when the whole analysis is based on the fixed representation of such data.

Technically, compositional data (Aitchison 1986) are strictly positive multivariate observations that carry only relative information. Accordingly, the only relevant information is contained in ratios between parts of a composition. The sample space of representations of compositional data with a prescribed constant sum constraint, the simplex, $\mathcal{S}^D$, consists of D-part compositions $\mathbf{x} = (x_1, \ldots, x_D)'$, where $\sum_{i=1}^{D} x_i = \kappa$ (which equals 100 for the case of percentages and 1 for proportions).

The specific nature of compositional data induces its own geometrical structure, called the Aitchison geometry, which has Euclidean vector space structure. Basic operations of the

Aitchison geometry (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015) are perturbation and powering, defined for compositional vectors $\mathbf{x} \in \mathcal{S}^D$ and $\mathbf{y} \in \mathcal{S}^D$ and a real number α as follows,

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, \ldots, x_D y_D]'; \ \alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, \ldots, x_D^\alpha]',$$

where

$$\mathcal{C}(\mathbf{x}) = \left[\frac{\kappa \cdot x_1}{\sum_{i=1}^D x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^D x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^D x_i}\right]'$$

stands for an arbitrarily chosen representation of the resulting composition (the closure operation). In the standard Euclidean geometry in real space, these two operations correspond to summation of vectors and multiplication of a vector by a scalar, respectively. The operation of perturbation can be also interpreted as shifting with respect to the Aitchison geometry, i.e. as a measure of difference appropriate to compositional change (Aitchison and Ng 2005). The perturbation-subtraction of $\mathbf{x}$ and $\mathbf{y}$,

$$\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus [(-1) \odot \mathbf{y}] = \mathcal{C}[x_1/y_1, \ldots, x_D/y_D]',$$

then represents the relative difference between both compositions. In other words, how the compositions differ in terms of ratios between the corresponding components. Obviously, if all the parts in the resulting composition are the same(neutral elements), the relative contributions conveyed by both compositional vectors coincide.

The Aitchison inner product, norm and distance, defined for two compositions $\mathbf{x}$ and $\mathbf{y}$ as

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}; \ \ \|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \left(\ln \frac{x_i}{x_j}\right)^2};$$

$$d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j}\right)^2}, \tag{1}$$

respectively, complete the Euclidean vector space structure of the Aitchison geometry.

Although the Aitchison geometry closely follows the relative nature of compositional data, most of standard statistical methods cannot be used there as they are designed for the Euclidean geometry in real space (Pawlowsky-Glahn *et al.* 2015). Moreover, in order to apply them to compositions, any such method would need to fulfil three principles, resulting from specific character of compositional data. The first principle is the mentioned scale invariance which means that output of the processing must remain the same irrespective to the change of measurement units. The second one is called subcompositional coherence and is closely related to the previous principle. In particular, when dealing with a subcomposition, which consists only of a selected components of the original composition, results of any analysis should not be in conflict with those of processing the whole composition. The third principle is the permutation invariance, i.e. invariance with respect to change of order of parts in a composition.

Instead of developing specific methods directly in the Aitchison geometry, it is much easier to express compositions in the real space and proceed with standard statistical tools. For this purpose, the so called logratio coordinates, formed with respect to the Aitchison geometry, are utilized (Pawlowsky-Glahn and Buccianti 2011). It depends on the aim of the analysis, which coordinates are the most appropriate.

It turned out that for the purpose of dimension reduction methods, that will be further employed in this study, the clr coefficients (Aitchison 1986), defined as

$$\mathrm{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right]', \tag{2}$$

where $g(\mathbf{x})$ is the geometric mean of all parts of a composition $\mathbf{x}$, form the reasonable choice. The clr coefficients are symmetric in components, each of them expresses (through the corresponding logratio) dominance of a component with respect to average behaviour of the other parts, aggregated by their geometric mean; i.e., the relative contribution of each part to the other components in average is captured. On the other hand, the sum of clr coefficients is zero as they correspond to a generating system with respect to the Aitchison geometry (Pawlowsky-Glahn *et al.* 2015). The reason is that dimension of a D-part composition is just $D-1$. This reflects the fact that it can be represented in a $(D-1)$-dimensional subspace (the simplex of proportions, percentages) without loss of information. It also means that the corresponding covariance matrix of clr coordinates is singular. Although the clr coefficients are thus not coordinates with respect to a basis on the simplex, which would reflect the usual practice, they still possess important properties. The crucial one is an isometry between the Aitchison geometry and the Euclidean space. Concretely, for compositions $\mathbf{x} \in \mathcal{S}^D$ and $\mathbf{y} \in \mathcal{S}^D$ and real numbers $\alpha,\ \beta$ it holds that

$$\mathrm{clr}(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha \cdot \mathrm{clr}(\mathbf{x}) + \beta \cdot \mathrm{clr}(\mathbf{y});$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}) \rangle\,;$$

$$\|\mathbf{x}\|_a = \|\mathrm{clr}(\mathbf{x})\|\,;\ d_a(\mathbf{x}, \mathbf{y}) = d(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y})),$$

Hence, when a composition is expressed in clr coordinates, standard statistical tools (that are able to cope with singularity of the covariance matrix) can be employed.

As pointed out in the previous section, the aim of this article is to analyse the structure of export and import in the end-use categories. The question is how to compare export and import of different countries. In the standard case, one would compute simply differences between components. However, each country has different area, different size of population, different GDP and different structure of the economy. This means that if we would just subtract import from export values, the results could be completely misleading. The problem can be solved using the mentioned perturbation-subtraction, i.e. by taking the ratios of export and import for every end-use category, and further statistical processing in clr coordinates.

2.1. Principal component analysis

Principal component analysis (PCA) is one of the most popular statistical techniques when analysing the multivariate structure of a dataset. The aim of this method is to reduce the data dimension in order to preserve most of the variability which is captured by small number of new variables - principal components (PCs).

Principal components for a mean-centered data matrix $\mathbf{X}_{(n\times D)}$ are obtained through linear transformation $\mathbf{U} = \mathbf{XB}$, where $\mathbf{U}_{(n\times D)}$ is the score matrix, whose columns $(\mathbf{u}_1, \ldots, \mathbf{u}_D)$ are the mentioned principal components, and $\mathbf{B}_{(D\times D)}$ stands for the loading matrix (Härdle and Simar 2012). The first PC is defined to have the largest possible variance, the second PC has to be orthogonal to the previous one and again posses the largest possible variance. Other PCs are defined in the same way.

In order to get principal components, the definiton of the matrix $\mathbf{B}$ is required.The loading matrix can be obtained via eigenvalue decomposition of the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{X}$. Accordingly, $\mathbf{\Sigma} = \mathbf{B\Lambda B}'$, where $\mathbf{\Lambda} = \mathrm{Diag}\{\lambda_1, \ldots, \lambda_D\}$ denotes the diagonal matrix of eigenvalues in decreasing order. In other words, the data matrix $\mathbf{X}$ can be interpreted as a product of the score and loading matrices,

$$\mathbf{X} = \mathbf{UB}' \text{ with } \mathbf{U}'\mathbf{U} = \mathbf{\Lambda}^2 \text{ and } \mathbf{B}'\mathbf{B} = \mathbf{I},$$

where $\mathbf{I}_D$ is the identity matrix. Consequently, bilinear decomposition is obtained.

For representation of the results of PCA, loadings and scores, the graph called biplot (Gabriel 1971; Gower and Hand 1996) is often applied. In the biplot scores (as points) and loading

vectors (as rays) of the first two principal components are displayed. In case of standard multivariate data, the length of the rays approximates the standard deviations of the original variables and the cosine of the angle between two rays displays correlation coefficients between the corresponding variables.

The differences for the compositional biplot (Aitchison and Greenacre 2002; Kynčlová, Filzmoser, and Hron 2016) consist in applying PCA on clr coordinates of $\mathbf{X}$ defined in (2). This implies different interpretation: rays now represent variability of relative dominance of compositional parts with respect to the rest of components, conveyed by the clr variables. Instead of correlation between two clr coefficients (that might be misleading due to singularity of the corresponding covariance matrix) rather variance of the pairwise logratio, approximated by the length of a link between two vertices, is considered. In particular, when the rays (vertices) coincide, the variance $\mathrm{var}(\ln \frac{x_i}{x_j})$ is approximately equal to zero which means that compositional parts x_i and x_j are interchangeable.

2.2. Parallel factor analysis

When in addition to the first two modes (samples, variables) also the third one, corresponding to conditions (like time or several measurement techniques, applied to the same samples), the bilinear PCA is no longer appropriate. One particular case is, when the same samples (countries) are observed for the same variables (end-use categories) in a long-term study, like for several years (occasions). Although it would be possible to analyse the data separately using PCA for each year, or even to apply PCA for the whole unfolded data set, by doing so the three-way structure could not be recognized. To analyse the complex structure of data simultaneously, the method called parallel factor analysis (PARAFAC) or canonical decomposition (CANDECOMP) needs to be applied (Harshman 1970; Carroll and Chang 1970). The data are decomposed into trilinear components where each component consists of one score vector and unlike PCA two loading vectors (though it is also usual to refer to three loading vectors). A PARAFAC model of three-way array (Carroll and Chang 1997) is thus given by three loading matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ with elements a_{if}, b_{jf} and c_{kf} that minimize the sum of squares of the residuals e_{ijk} coming from expression

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{3}$$

for $i = 1, \ldots, n,\ j = 1, \ldots, D$ and $k = 1, \ldots, K$.

The solution of the PARAFAC model (estimation of the loading matrices for a given number of factors F) can be found using alternating least squares (ALS) by assuming the loading vectors of two modes known and then estimating the unknown set of parameters of the last mode using the least squares regression (Carroll and Chang 1997; Kroonenberg 1983). The algorithm works in an iterative manner and under mild conditions it converges to a unique solution (Harshman and Lundy 1984; Stegeman 2006). From the compositional perspective, the rotational invariance of the ALS algorithm (Kruskal 1989) is of particular importance, because it enables to employ any logratio coordinates with the isometry property (like clr coefficients) for the estimation purposes (Di Palma, Gallo, Filzmoser, and Hron 2016). Although PARAFAC or, more generally, statistical modeling of three-way data was recently successfully employed for economic applications (Dell'Anno and Amendola 2015; Veldscholte, Kroonenberg, and Antonides 1998) and its specifics for compositional data were developed (Gallo 2013; Gardlo, Smilde, Hron, Hrdá, Karlíková, Friedecký, and Adam 2016), combination of both aspects (as far as it is known to the authors) is not available in the literature.

Similarly as of PCA, it is popular to display PARAFAC results graphically. Concretely, loading values of the first two components are displayed in terms of three scatterplots, one for each of modes. Subsequently, the obtained information can be merged together in order to get a concise view on the three-way structure. There are not specific features in case of

compositional data here, except to the fact that interpretation of clr variables needs to be taken into account.

3. Applications to trade flows structure

Theoretical considerations, introduced in the previous section, were applied on the real-world data which include the values of exported and imported goods of EU countries and 13 other largest economies of the world (regarding available data of WIOD database). These countries represented more than 85% of the world GDP in 2008. The first data set, trade flows broken down by end-use categories, is available online (http://stats.oecd.org/index.aspx?queryid=32186), the second database - shares of value added broken down by factors can be obtained from the WIOD database (www.wiod.org).

All the computations and graphical outputs were performed using the packages robCompositions (Templ, Hron, and Filzmoser 2011) and ThreeWay (Giordani, Kiers, and Del Ferraro 2014) of statistical software R (R Core Team 2016). Accordingly, the optimal number of components in the PARAFAC model was derived using the NumConvHull procedure (Ceulemans and Kiers 2006).

3.1. Trade flows in end-use categories

Breaking down trade in goods according to their end-use (OECD Directorate for Science, for Economic Analysis, and Statistics 2014) adds a new dimension to the traditional commodity-based trade statistics and provides a link to National Accounts Input-Output Tables, in which flows of goods and services are reported according to end-users. Using the basic domestic end-use categories from the System of National Accounts and the detailed classification systems of trade in goods, bilateral flows of exports and imports can be classified into intermediate goods, household consumption goods and capital goods. However, some kinds of products can be either for intermediate demand and household consumption, or for capital goods in industry and household consumption. Thus it was introduced mixed end-use category which contains personal computers, passenger cars, personal phones, packed medicines and precious goods. The last category, miscellaneous, includes commodities that don't belong to any other categories. To keep the presented study simple, we will not consider this category for further calculations. In Table 1 a small part of the data set is shown for illustration purposes.

The dataset used in this section is called The OECD STAN Bilateral Trade by Industry and End-use (Zhu, Yamano, and Cimper 2011). It firstly released in 2011 to highlight the increasing influence of export and import of intermediate goods. The values of import and export of goods are broken down by industrial sectors and, simultaneously, by end-use categories. Estimates are expressed in nominal terms, in current US dollars, and are collected from more than a hundred reporters and partners, including all 34 members of OECD and a wide range of non-members. Note that for the purpose of standard statistical analysis, without considering the relative nature of data, we would have to convert the current US dollars into constant US dollars in order to employ time. However, we are dealing with compositional data which means that just ratios between categories form the source of relevant information and thus multiplication by any constant does not affect results of the analysis. Following this idea, it is not necessary to convert the currency prior to further statistical processing using the logratio methodology.

As stated above, patterns in the relative structure of export and import of goods cannot be revealed by applying standard multivariate techniques to the raw data as the relevant information is contained exclusively in ratios between the respective components. Nevertheless, for the sake of comparison, principal component analysis was applied both to the original data and to clr coordinates for the year 2012, the most recent complete one in the database.

Obviously, when dealing with economies of different size of trade (with different population, share of trade in economy), straightforward application of PCA (see Figures 1-3) becomes

Table 1: Small part of the OECD data.

EXPORT	Intermediate	Household consumption	Capital	Mixed end-use
AUS	117409481	41592639	47616893	37627503
AUT	91332690	20919015	28060985	12265110.5
BEL	262169002	67877294	33929953	77809975.3
⋮				
TUR	85476254	45804409	13700828	6770951
USA	880162112	159919833	234236609	132854776
TWN	230018351	16278914	41361666	11368675
IMPORT				
AUS	216321786	19537887	6669449	6676966.4
AUT	103594607	30203689	19508560	15607428
BEL	269679952	58898420	33473255	70981894
⋮				
TUR	137206055	14363175	30415227	14167392
USA	1209223479	406593569	298476700	351426871
TWN	212736197	15469315	29972519	10065989

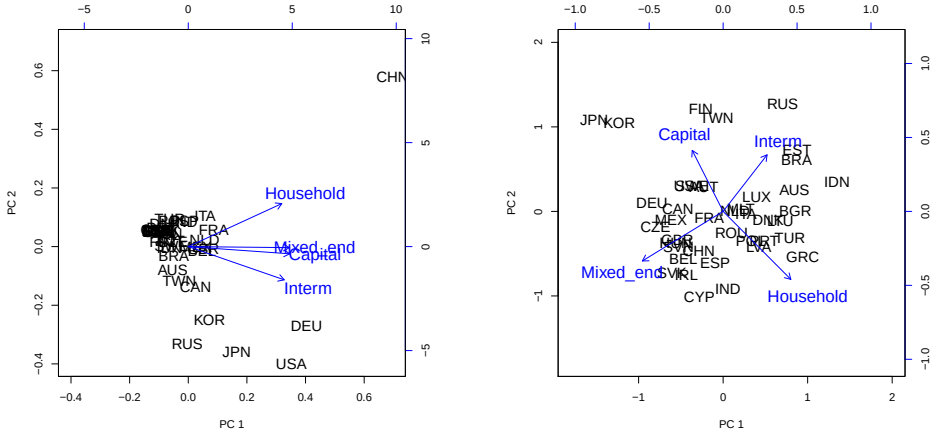


Figure 1: Biplots of export applied to the original data (on the left) and to clr coordinates (on the right).

useless. From the biplots on the left side, it is hard to recognize any structure in the dataset: it either seems that all variables are highly correlated (Figure 1 and 2), or the respective interpretation is doubtful (Figure 3).

In contrast, when relative contributions of the components, conveyed by clr coordinates (here applied to end-use categories), are considered instead, PCA and biplot diagrams are much easier to interpret (see the Figures 1 and 2 on the right). In Figure 1 (on the right), the countries exporting relatively more intermediate goods (Russia, Australia, Brazil), household (Greece, Turkey, India), mixed end-use (middle Europe countries), capital goods (Japan, Korea, Finland) can be well distinguished, no matter of their size.

Similarly, in Figure 2 on the right, the compositional biplot of import is displayed. It is evident that for Asian countries such as Korea, Taiwan, India and China dominance of intermediate and capital goods in relative structure of import can be observed. On the other hand, mixed end-use goods are imported into large countries, namely Russia, Australia, USA and Canada. Middle Europe countries are spread around the origin and Cyprus imports mostly the household consumption goods. This corresponds well to the general perspective of international trade structure of that year (UN 2012).

The perturbation operation can be now used to capture relative differences between export

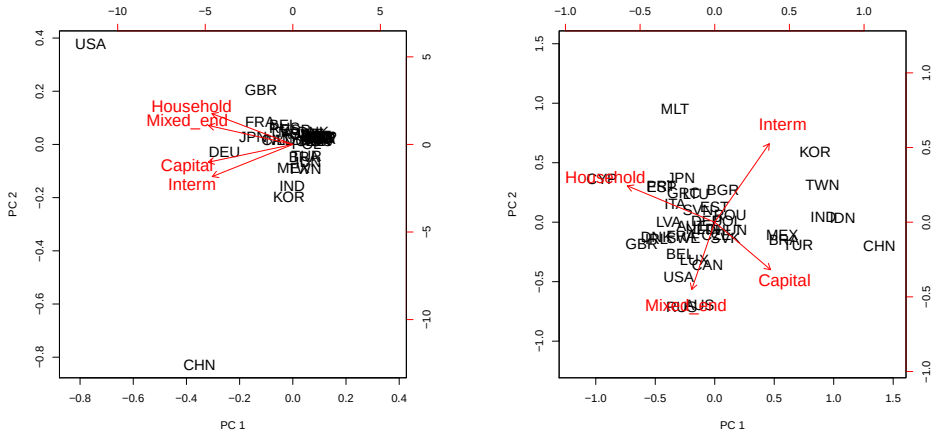


Figure 2: Biplots of import applied to the original data (on the left) and to clr coordinates (on the right).

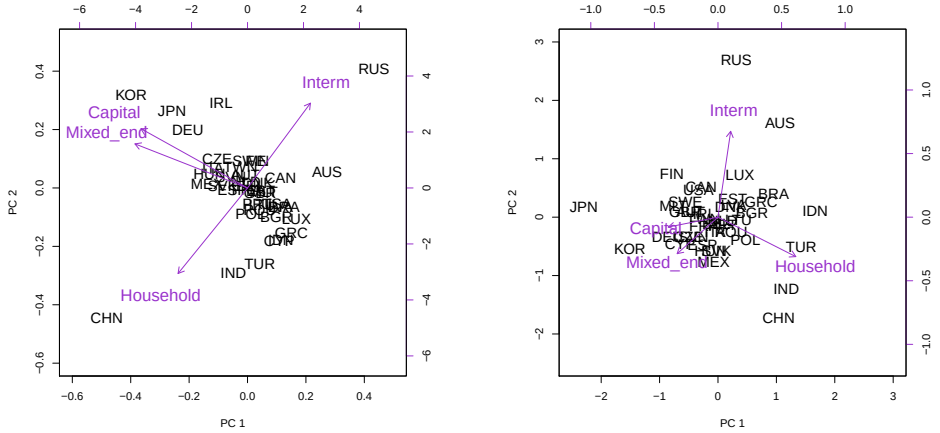


Figure 3: Biplots of differences between export and import, applied to original data (on the left) and to clr coordinates (on the right).

and import structure through ratios between the respective components. Consequently, large values of the (log-)ratios will indicate discrepancy between both international flow aspects. From the respective link in Figure 3 (right) it is visible that the variance of pairwise logratio between export/import ratios of Capital goods and Mixed end-use goods, respectively, is very small. Thus the ratios between exports and imports of these end-use categories are relatively very similar. The cluster of China, India, Indonesia and Turkey lies near the Household goods variable (in terms of its relative dominance with respect to the other categories as conveyed by the respective clr coordinate), thus these countries have the relatively largest surplus of export in this category. Russia and Australia have largest surplus in intermediate goods, while Korea and Japan in capital goods. Although these effects could be even better observed from biplot of the original data, previous results of sole export and import indicate that high variability of mixed end-use and capital goods categories is not relevant by considering relative structure of observations.

In order to include also time variable and to get a complete picture about the development in a larger time scale, also PARAFAC modeling was applied to the perturbed data, i.e. to the ratio of export and import components (after expressing them in clr coordinates) for years 2003–2012. Similar results as for the previous figures were obtained that confirm a certain stability of the export/import structure comparing to the single year 2012, considered above.

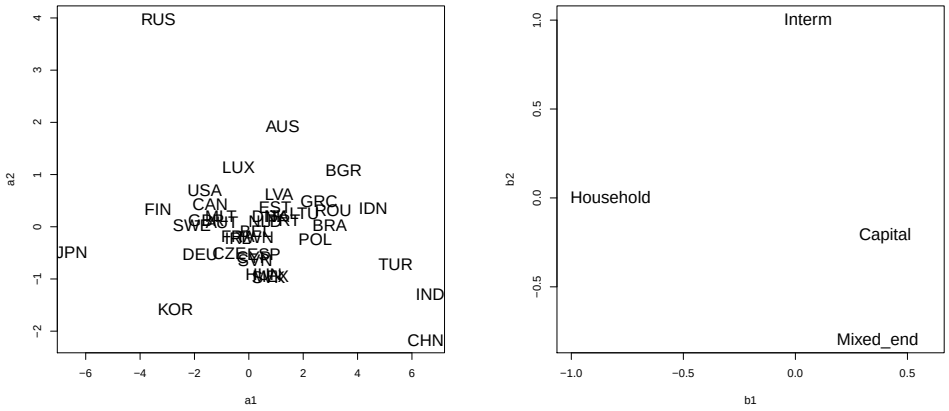


Figure 4: Results of the PARAFAC method for differences between exports and imports, mode A (on the left) and mode B (on the right), using clr coordinates.

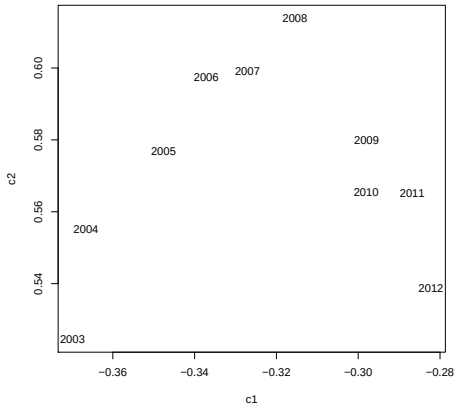


Figure 5: Results of the PARAFAC method for differences between exports and imports, mode C, using clr coordinates.

In the mode A (Figure 4 on the left), corresponding to samples, cluster of China, India, Turkey and Indonesia can be seen, as well as cluster of Japan and Korea. In the middle of the plot there is a group of middle European countries and it also seems that Russia differs significantly from the other countries. Mode B (Figure 4, right plot) confirms the result that components Capital and Mixed end-use goods are relatively very similar, when considering ratios of export and import for the years 2003–2012. And finally, mode C displayed on the Figure 5 shows the development in time, where a clear time pattern with a change point in 2008 is observed, interpretable in terms of global integration. Accordingly, this loading plot well reflects the global crisis in 2008–2009 that has temporarily brought the long-run trend of rising global integration through trade to a halt.

3.2. Trade flows of value added

Since an intensive integration process recently, the flows of value added across countries have become more relevant than the flows of goods. It is caused by the growing effect of the vertical specialization, which can be explained in a way that firms offshore activities to other countries to exploit cost advantages in particular stages of production (for more see Stehrer, Foster, and de Vries 2012; Hummels, Ishii, and Yi 2001). As discussed above, the share of intermediates in trade is significant. In order to distinguish real contribution (represented by value added)

Table 2: Small part of VA data.

VA EXPORT	Capital	High-skilled	Medium-skilled	Low-skilled
AUS	236672.49	108796.630	122716.748	109261.377
AUT	68726.10	49733.400	86969.011	12356.746
BEL	100983.53	59336.529	102739.534	27714.804
⋮				
TUR	204194.6	42477.77	31392.93	39164.41
TWN	79596.1	71722.10	34759.83	22437.47
USA	3632767.3	3073707.79	2353237.78	196386.87
VA IMPORT				
AUS	22149.071	6364.3767	9851.844	5749.7064
AUT	15143.087	6587.8375	11316.180	3869.4331
BEL	29874.281	12481.4846	19397.737	9089.2901
⋮				
TUR	8765.471	4035.702	6835.904	3442.306
TWN	29002.087	6870.004	12055.759	8311.096
USA	261998.337	46909.534	84104.518	48165.174

of each country in its exports and other countries in its imports, the composition of value added export and import was explored.

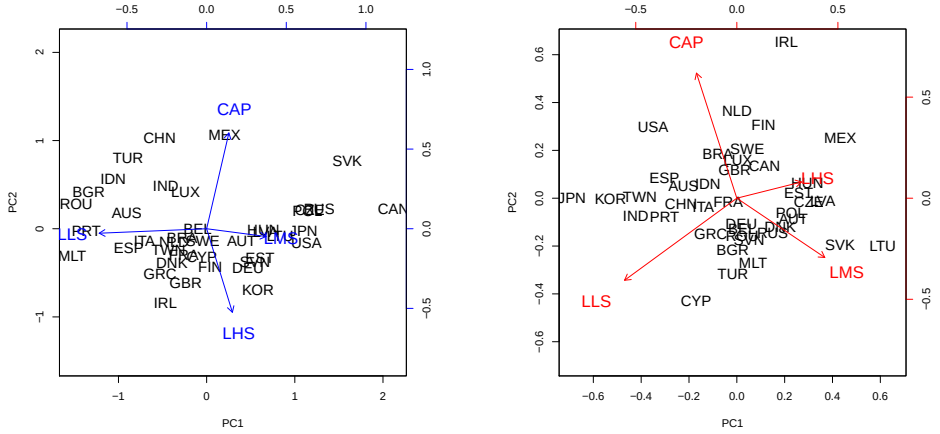


Figure 6: Compositional biplots of value added export (on the left) and import (on the right) of factors for clr coordinates.

The WIOD database (Timmer *et al.* 2012) allows to break value added of final products into factors, namely capital (CAP) and labour (low skilled (LLS), medium skilled (LMS) and high skilled (LHS)). The database comprises gross output and value added by industry for each country and the flow of products across industries and countries in a global input-output matrix. The WIOD database provides a time series of world input-output tables (WIOTs) from 1995 to 2009. The shares of factors in each industry for all considered countries may be found in the Socio Economic Accounts table (may be downloaded from `http://www.wiod.org/new_site/database/seas.htm`).

Our second data set (see Table 2) is obtained from WIOTs and Socio Economic Accounts table in the following way. From WIOTs, we can calculate value added export (VAX) (for detailed treatment see Johnson and Noguera (2012) and Timmer, Dietzenbacher, Los, Stehrer, and Vries (2015)) for each country and each industry. Employing Socio Economic Accounts table we obtain share of each factor in the calculated value added in each industry. Summing by industry we get shares of each factor in VAX for each country. Similarly we can split value added by other countries in imports to each country.

It is well known (Stehrer *et al.* 2010; Timmer *et al.* 2013) that developed countries export relatively more high skilled labour and import more capital. In contrast, developing countries are abundant with low skilled labour and import high skilled labour. This is illustrated by Figure 6 for the year 2009, for which the database provides complete data. Indeed, China, Turkey and Indonesia export relatively more low-skilled labour and capital, southern part of EU low-skilled labour (in sense of their relative contributions with respect to the other components, reflected by clr coordinates). The new countries of the EU have significant abundance in medium-skilled labour as well as United States or Japan. The opposite tendency can be seen in Figure 6 on the right, where compositional biplot of import of factors is displayed.

To see the development in time, the PARAFAC model was applied to data for years 2000–2009. In Figure 7 the results are displayed. By considering modes A and B of export (left panel) together, countries can be divided into two parts. In the left part, the countries exporting relatively more low-skilled labour are clustered (e.g., southern European countries, Turkey, India, Indonesia or China). However, in the right part clusters of countries that export relatively more capital, high- and medium skilled labour can be seen - Canada, USA, Japan, Korea and middle European countries. From the mode B it can be concluded that export of LMS and LHS is quite strongly proportional. Mode C of the left panel reflects the change in year 2004, when an intensive integration process for many European countries as new members of the European Union started.

Similarly as for the case of export, from mode B of the right panel it can be observed that import of LHS and LMS is proportional (though not so closely as for the case of export). Moreover, clusters of countries from mode A are similar to those from the biplot in Figure 6 (right). Accordingly, 1) Ireland, Finland, Sweden, Netherlands and USA, 2) Malta, Cyprus, Portugal, Turkey and Bulgaria, and 3) Japan, India, Taiwan and Korea have similar relative structure of import in terms of value added. In Mode C, the development is not so clear as for the case of export, however it still reflects the exceptional role of the year 2004.

4. Discussion

With development of detailed publicly available databases, it is possible to analyse systematically also the international trade structure. Nevertheless, it is of particular importance to consider carefully the natural properties of the observations at hand prior to their further statistical processing. The case of export and import structure shows that problems with different trade sizes can be overcome by employing the logratio methodology of compositional data. Although PCA (biplot) and PARAFAC are standard tools for analysis and visualization of multivariate data, their application in the compositional and economic contexts simultaneously form the main novelty of the paper. Results of analysing the international trade structure reflects well the general knowledge, as provided regularly by the United Nations (UN) and other institutions.

Apparently, the interpretation provided in the previous section is just illustrative capturing the main features and there is still space for its further extension. For example, differences in factors related to the export can be seen also from much broader perspective. In case of the European Union, one can distinguish "core" EU countries, its southern countries and new countries. Accordingly, the difference in technological structure of export, which is related to the level of skills, is often accounted for problems of Euro (see, e.g., Wierts, Van Kerkhoff, and De Haan 1998). We leave these issues as inspiration for those, who would employ the logratio methodology for more detailed macroeconomic analyses in the future.

Acknowledgments

Authors gratefully acknowledge the support of the Operational Program Education for Com-

petitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic), the grant COST Action CRoNoS IC1408 and the grant IGA_PrF_2016_025 Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc.

References

Aitchison J (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall, London.

Aitchison J, Greenacre M (2002). "Biplots of Compositional Data." *Applied Statistics*, **51**, 375–392.

Aitchison J, Ng KW (2005). "The Role of Perturbation in Compositional Data Analysis." *Statistical Modelling*, **5**, 173–185.

Blejer MI, Fernandez RB (1980). "Effects of Unanticipated Money Growth on Prices and on Output and Its Composition in a Fixed-exchange-rate Open Economy." *Canadian Journal of Economy*, **13**, 82–95.

Carroll JD, Chang J (1970). "Analysis of Individual Differences in Multidimensional Scaling via an n-way Generalization of "eckartyoung" Decomposition." *Psychometrika*, **35**, 283–319.

Carroll JD, Chang J (1997). "PARAFAC. Tutorial and Applications." *Chemometrics and Intelligent Laboratory Systems*, **38**, 149–171.

Ceulemans E, Kiers HAL (2006). "Selecting among Three-mode Principal Component Models of Different Types and Complexities: A Numerical Convex Hull Based Method." *British Journal of Mathematical and Statistical Psychology*, **59**(1), 133–150.

Dell'Anno R, Amendola A (2015). "Social Exclusion and Economic Growth: An Empirical Investigation in European Economies." *Review of Income and Wealth*, **61**(2), 274–301.

Devarajan S, Swaroop V, Zou H (1996). "The Composition of Public Expenditure and Economic Growth." *Journal of Monetary Economics*, **37**, 313–344.

Di Palma MA, Gallo M, Filzmoser P, Hron K (2016). "A Robust Parafac Model for Compositional Data." Under review.

Fry T (2011). "Applications in Economics." In V Pawlowsky-Glahn, A Buccianti (eds.), *Compositional Data Analysis: Theory and Applications*, pp. 318–326. Wiley, Chichester.

Gabriel KR (1971). "The Biplot Graphic Display of Matrices with Application to Principal Component Analysis." *Biometrika*, **58**, 453–467.

Gallo M (2013). "Log-ratio and Parallel Factor Analysis: An Approach to Analyze Three-way Compositional Data." In AN Proto, M Squillante, J Kacprzyk (eds.), *Advanced Dynamic Modeling of Economic and Social Systems*, pp. 209–221. Springer, Heidelberg.

Gardlo A, Smilde AK, Hron K, Hrdá M, Karlíková R, Friedecký D, Adam T (2016). "Normalization Techniques for PARAFAC Modeling of Urine Metabolomic Data." *Metabolomics*, **12**, 117.

Giordani P, Kiers HAL, Del Ferraro MA (2014). "Three-Way Component Analysis Using the R Package ThreeWay." *Journal of Statistical Software*, **57**(7).

Gower JC, Hand DJ (1996). *Biplots.* Chapman & Hall, London.

Härdle WK, Simar L (2012). *Applied Multivariate Statistical Analysis*. Springer, Heidelberg.

Harshman RA (1970). "Foundations of the Parafac Procedure: Models and Conditions for an "Explanatory" Multi-modal Factor Analysis." *Report 10085*, University of California, Los Angeles.

Harshman RA, Lundy ME (1984). "The PARAFAC Model for Three-way Factor Analysis and Multidimensional Scaling." In HG Law (ed.), *Research Methods for Multimode Data Analysis*, pp. 122–215. Praeger, New York.

Hausmann R, Hwang J, Rodrik D (2007). "What You Export Matters." *Journal of Economic Growth*, **12**(1), 1–25.

Hummels D, Ishii J, Yi KM (2001). "The Nature and Growth of Vertical Specialization in World Trade." *Journal of International Economics*, **54**, 75–96.

Johnson RC, Noguera G (2012). "Accounting for Intermediates: Production Sharing and Trade in Value Added." *Journal of international Economics*, **86**(2), 224–236.

Kroonenberg PM (1983). *Three-mode Principal Component Analysis. Theory and Applications*. DSWO Press, Leiden.

Kruskal JB (1989). "Rank, Decomposition, and Uniqueness for 3-way and N-way Arrays." *Multiway Data Analysis*, **33**, 7–18.

Kynčlová P, Filzmoser P, Hron K (2016). "Compositional Biplots Including External Non-compositional Variables." *Statistics*, pp. 453–467. doi:10.1080/02331888.2015.1135155.

Miroudot S, Lanz R, Ragoussis A (2009). "Trade in Intermediate Goods and Services." *OECD Trade Policy Working Paper*, **93**.

OECD Directorate for Science T, for Economic Analysis ID, Statistics (2014). *OECD Bilateral Trade Database by Industry and End-use Category*. OECD Publishing.

Pawlowsky-Glahn V, Buccianti A (eds.) (2011). *Compositional Data Analysis: Theory and Applications*. Wiley, Chichester.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (eds.) (2015). *Modeling and Analysis of Compositional Data*. Wiley, Chichester.

R Core Team (2016). "R: A Language and Environment for Statistical Computing."

Rodrik D (2006). "What's So Special About China's Exports?" *China & World Economy*, **14**(5), 1–19.

Stegeman A (2006). "Degeneracy in Candecomp/Parafac Explained for $p \times p \times 2$ Arrays of Rank $p + 1$ or Higher." *Psychometrika*, **71**(3), 483–501.

Stehrer R, Foster N, de Vries G (2010). "Value Added and Factors in Trade: A Comprehensive Approach." *Dynamics*, **67**.

Stehrer R, Foster N, de Vries G (2012). "Value Added and Factors in Trade: A Comprehensive Approach." *wiiw Working paper*, **80**, 1–22.

Templ M, Hron K, Filzmoser P (2011). "robCompositions: an R-package for Robust Statistical Analysis of Compositional Data."

Timmer M, Erumban AA, Gouma R, Los B, Temurshoev U, de Vries GJ, Arto I (2012). "The World Input-output Database (WIOD): Contents, Sources and Methods." *WIOD Background document*, **40**. URL www.wiod.org.

Timmer MP, Dietzenbacher E, Los B, Stehrer R, Vries GJ (2015). "An Illustrated User Guide to the World Input–output Database: The Case of Global Automotive Production." *Review of International Economics*, **23**(3), 575–605.

Timmer MP, Los B, Stehrer R, de Vries GJ (2013). "Fragmentation, Incomes and Jobs: an Analysis of European Competitiveness." *Economic Policy*, **28**(76), 613–661.

UN (2012). *World Economic Situation and Prospects*. United Nations, New York.

Veldscholte CM, Kroonenberg PM, Antonides G (1998). "Three-mode Analysis of Perceptions of Economic Activities in Eastern and Western Europe." *Journal of Economic Psychology*, **19**(3), 321–351.

Wierts P, Van Kerkhoff H, De Haan J (1998). "Composition of Exports and Export Performance of Eurozone Countries." *JCMS: Journal of Common Market Studies*, **52**(4), 928–941.

Zhu S, Yamano N, Cimper A (2011). "Compilation of Bilateral Trade Database by Industry and End-use Category." In *OECD Science, Technology and Industry Working Papers*. OECD Publishing, Los Angeles.

Affiliation:

Klára Hrůzová
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science
Palacký University
771 46 Olomouc, Czech Republic
E-mail: klara.hruzova@gmail.com

Miroslav Rypka
Department of Geoinformatics, Faculty of Science
Palacký University
771 46 Olomouc, Czech Republic
E-mail: rypkamir@seznam.cz

Karel Hron
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science
Palacký University
771 46 Olomouc, Czech Republic
E-mail: hronk@seznam.cz

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 46 *Submitted:* 2012-06-04
February 2017 *Accepted:* 2016-05-26

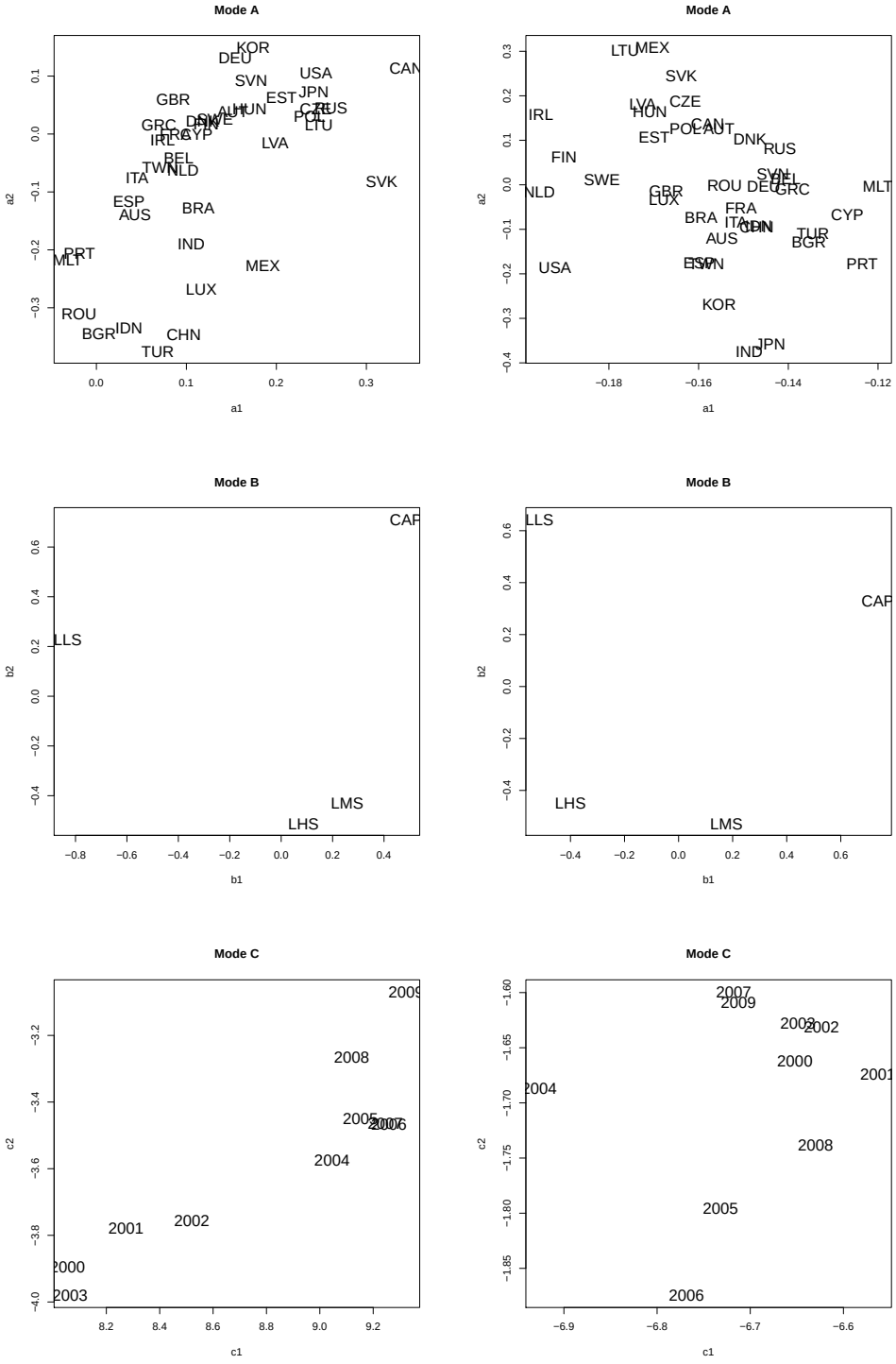


Figure 7: Results of the PARAFAC method for the export (left panel) and import (right panel) of value added using clr coordinates.

Austrian Journal of Statistics
February 2017, Volume 46, 65–80.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i2.576

Improving Road Freight Transport Statistics by Using a Distance Matrix

T. Karner Statistics Austria
B. Weninger Statistics Austria
S. Schuster Statistics Austria
S. Fleck Statistics Austria
I. Kaminger Statistics Austria

Abstract

Distances driven by road freight vehicles are an essential parameter for the calculation of transport volume. In the Austrian road freight survey, places of loading and unloading are recorded on a postal code basis. To derive the actual distances driven from this data, Statistics Austria uses a distance matrix that was first created in the 1980s. While the first version of this matrix was based on manual measurements, it has recently been recreated and updated using modern routing software.

This article describes the methodology on which the current Austrian distance matrix is based. The main points discussed are: how to determine representative centroids for postal code areas; how to deal with journeys within one postal code area; and how to calculate the actual distances using routing software.

The last part of the article compares the distance matrix to odometer readings from the Austrian road freight survey of the reference year 2015. This comparison showed a high positive correlation which indicates the good quality of the developed distance matrix and emphasises its usefulness in road freight transport statistics.

Keywords: road freight transport statistics, distance matrix, transport volume, calculation of distances.

1. Introduction

In the framework of the European Statistical System (ESS) and in context with principle 9 of the European Statistical Code of Practise (Eurostat 2011a) official European Statistics should be produced without excessive burden on respondents. This article presents a method for the estimation of driven distances in kilometres based on a distance matrix. This method could be easily implemented and used in the European road freight survey to simplify the collection of kilometres driven. The survey is based on EU Regulation No 70/2012 (Council of the European Union 2012) and is obligatory for all Member States (MS) except Malta.

The first part of this article describes the theory and the base data for the distance matrix. Additionally, several practical examples are introduced. The second part includes a comparison of the kilometres received from the distance matrix with the kilometres driven based on odometer information of the reference period.

2. The road freight survey

In general, transport statistics provide information on the transport volume and the transport performance of the different modes of transport (road, rail, inland waterways, sea, air and pipelines). Transport volume is the weight of transported goods in tonnes; transport performance is the product of transport volume and the distance in kilometres.

In contrast to other surveys in transport statistics, the nationality principle is applied to the road freight survey instead of the territoriality principle. Furthermore, the road freight survey is performed as sample survey in place of a complete survey.

2.1. The nationality principle

Compared to the territoriality principle, where all movements of a vehicle on a defined territory are observed, the nationality principle is based on collecting data of vehicles registered in the respective country. Hence on the basis of EU Regulation No 70/2012 (Council of the European Union 2012) each member state surveys the journeys of road transport vehicles - with at least a load capacity of 3.5 tonnes or maximum possible weight of 6 tonnes in case of single motor vehicles – performed on public roads within the territory of the member state as well as abroad. Agricultural vehicles, military vehicles and vehicles belonging to central or public administration[1] are not included in the survey.

In the Austrian road freight survey information on all journeys of lorries registered in Austria are collected. Due to the nationality principle there are five types of transport:

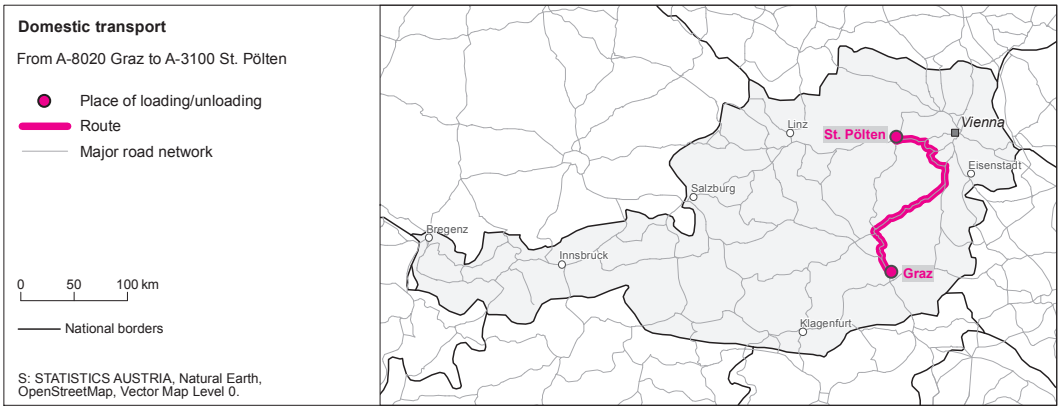
- *Domestic transport*: Place of loading and unloading are both located in Austria. This definition includes cabotage as a special case of domestic transport, as the main focus in this article lies on the territory where the journeys take place and not the nationality of the vehicles (see figure 1a).
- *International dispatch*: Place of loading is in Austria and place of unloading in a different country (see figure 1b).
- *International receipt*: Place of loading is in a different country and place of unloading is in Austria (see figure 1c).
- *Transit*: Place of loading and place of unloading are not in Austria, but the journey leads through Austrian territory (see figure 1d).
- *Other transport abroad*: This kind of transport involves journeys of Austrian road goods vehicles, which do not take place on Austrian territory (see figure 1e).

As a consequence of the nationality principle, the road freight surveys in the member states do not include all transportation on the national territory. Instead, they contain information on transport of all vehicles registered in each member state, irrespective where it was performed.

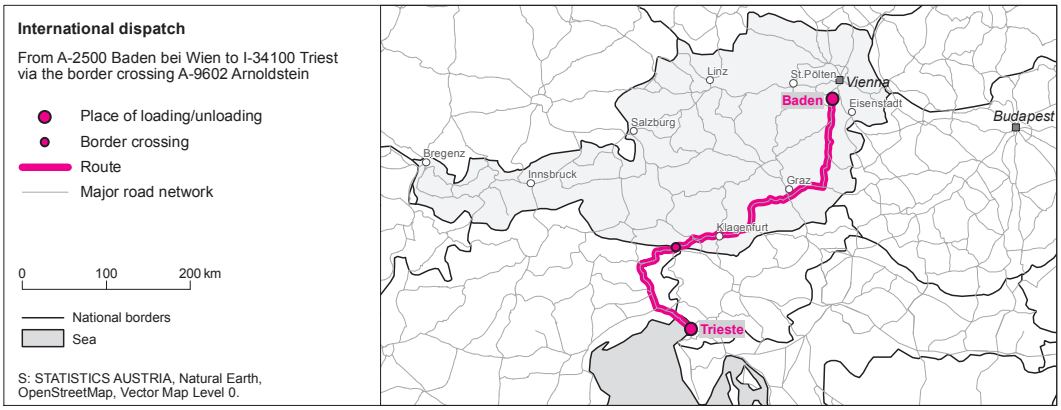
Eurostat receives data sets from all member states and – after several plausibility checks – consolidates them to one comprehensive data set. Based on this comprehensive data set several tables can be generated and with regard to the European Commission Regulation No 6/2003 (European Commission 2003a) are distributed to the national authorities which are responsible for community transport statistics in the particular member states. These authorities[2] have the possibility to complete the statistical coverage of road transport at national level with the provided information.

Obviously, the road freight survey is a cross-national European Statistics. Hence it is of extreme importance, that the quality of the survey in each member state is high level and the concepts within the different surveys are similar and coordinated as well as possible.

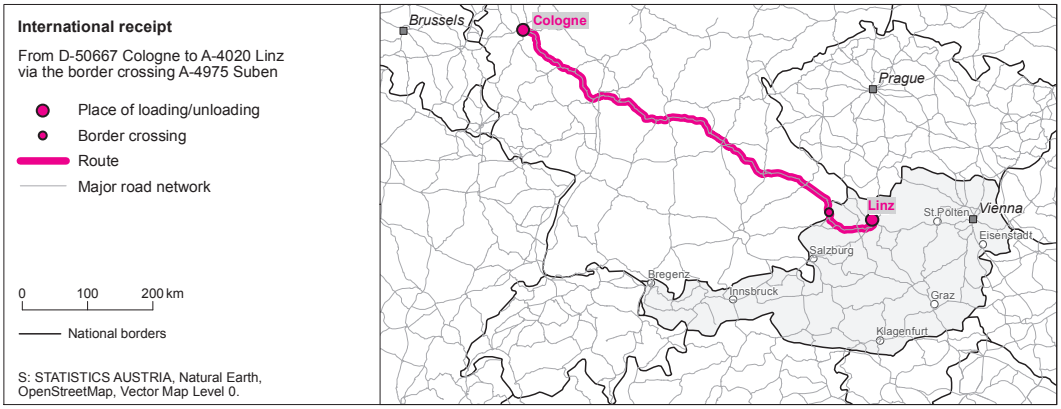
[1] with the exception of goods road vehicles belonging to public undertakings.
[2] In Austria the so called D-tables are transmitted to Statistics Austria.



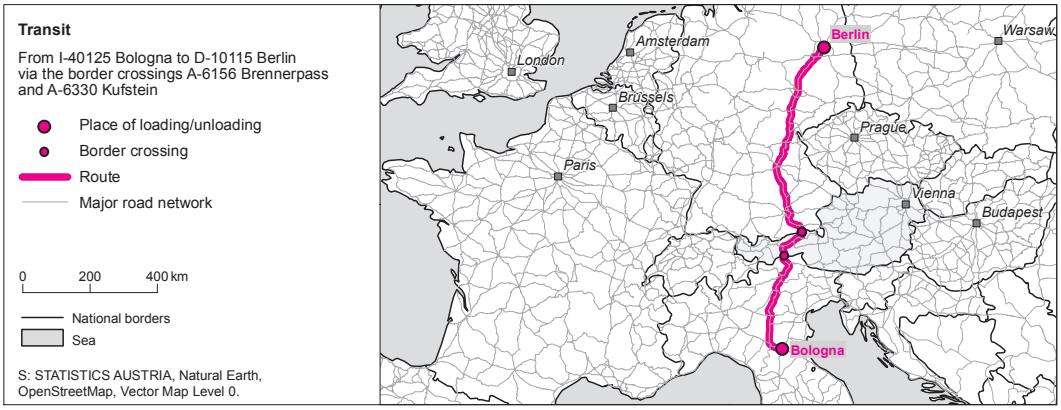
(a) Domestic transport



(b) International dispatch

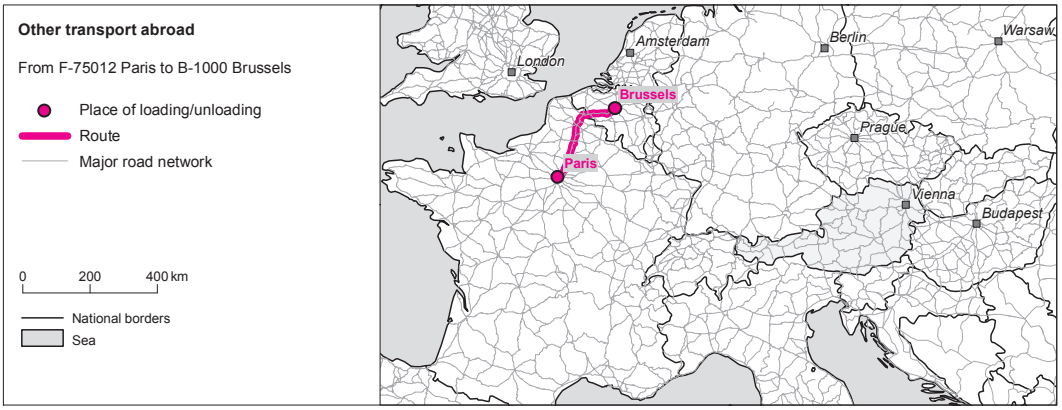


(c) International receipt



(d) Transit

Figure 1: Types of transport



(e) Other transport abroad

Figure 1: Types of transport (continued)

2.2. A sample survey

Due to the number of registered vehicles in the member states and the high amount of journeys, the road freight survey is performed as sample survey. Based on the principles of the European Statistical System, which implicate the reduction of burden on respondents, cost effectiveness and the development of advanced statistics using modern methods, it is not deemed maintainable to implement a complete survey.

The population for the sampling procedure consists of the road freight vehicles registered in each member state. The manual "Road Freight Transport Methodology" (Eurostat 2011b) provides several recommendations for the design of the random sample. These recommendations refer to time periods (normally operations during one reference week), sampling strategies (e.g. considering different sizes of vehicles, separate strata for road tractors) and tips to avoid systematic errors (e.g. refusals, response errors, not adequate coverage of the population).

Regarding the sample size, the thresholds for the percentage standard errors are defined in the European Commission Regulation No 642/2004 (European Commission 2003b). The percentage standard errors of the annual estimates for the main variables tonnes transported, tonne-kilometres performed and total kilometres travelled loaded shall not exceed $\pm$ 5 % (95 % confidence) respectively $\pm$ 7 % if the total stock of road motor vehicles relevant to the survey in a Member State is less than 25 000 or the total stock of vehicles engaged in international transport is less than 3 000.

In Austria, the population of the survey consists of around 66 000 road freight vehicles with a load capacity of at least 2 tonnes or road tractors. Once a year (usually in December) a stratified sampling procedure (load capacity of the local unit, vehicle capacity, region, type of transport) is done for the whole reporting year. As a benefit of this yearly procedure, large companies are informed in advance about the dates of their reference weeks. To avoid a possible bias due to inactive local units or deregistered vehicles during the year, a refreshment sample is performed quarterly. On the whole, a total of 26 000 reporting weeks are collected annually.

2.3. The Questionnaire

Due to the high complexity of reported journeys (e.g. combination of laden or empty journeys, delivery or collection journeys) the design of the questionnaire for the road freight survey is a huge challenge for statisticians. Thus the questionnaire resembles more a log book than a questionnaire typically used in official statistics.

Four main tasks on the development of a questionnaire have to be taken into account in order to collect all relevant data (e.g. place of loading and unloading, type of goods or distances driven):

- The questionnaire should be easy to understand and fill in.
- The respondents burden should be minimised.
- The collected information should be detailed and accurate.
- Several kinds of questionnaires (paper-based, computer-based) should be offered.

Regarding the last point it has to be mentioned, that the target group for the questionnaire is very heterogeneous. In large companies the questionnaires are usually filled in by the staff of the accounting departments, whereas in smaller companies mostly the driver is in charge of it. A study conducted by SYSTRA for the Department for Transport in the UK (Systra 2015), showed that the information sources to complete the road freight survey are quite varying. The companies use for instance run records, drivers reports/day sheets/worksheets, tachograph software, GTS or GPS systems, google maps for distance calculation, fuel cards, vehicle inspection sheets, odometer readings, company diaries, log books or smart phone applications. One result of the study was that companies typically need three different sources to complete the survey. On the one hand, all data was stored electronically and on the other hand there was a mixture of computer based information and hard copy data sources. Therefore, it is useful to design the questionnaire to be easily applicable on different kinds of media (e.g. electronic questionnaire, excel sheets, mobile phone applications or paper questionnaire), to enable each respondent to choose the appropriate kind of questionnaire.

Nevertheless, the questionnaire is complex and dynamic because its length depends on the number and type of journeys during the reference week. Hence, the effort for every respondent might be different. To support the Member States in the development of the questionnaire, Eurostat provides several suggestions through the reference manual.

Collecting information about the distances of journeys

The information of the distance for each journey is one of the essential variables of the survey as it is essential for the calculation of the transport performance. Referring to the Eurostat manual for the road freight survey, the respondents should provide this information for each journey. In practise this information is frequently not available. In the SYSTRA study it became evident that in such cases respondents use the driver's worksheet for the variables place of loading and unloading. Additionally, Google Maps or similar in-house systems are used to calculate the kilometres.

Obviously, collecting data about place of loading, place of unloading and additionally, the kilometres driven between these places raises the burden on respondents and is redundant.

Statistics Austria was aware of these difficulties already in the 1970-ies. For this reason a method was developed, which imputes the kilometres driven between the place of loading and unloading on basis of the postal codes of these places. The fundament of the imputation is a distance matrix which includes all distances between every possible postal code combination in kilometres within Austria and - with limitations – abroad.

The first version of a distance matrix was developed by using meilographs for measuring the lengths of the roads between two postal codes manually or by using algorithms based on air-line distance. It is obvious, that the development of the distance matrix was complex and labour-intensive then and it was also impossible to update it continuously.

Due to the development of modern IT-technology, powerful route planning software and GIS-applications in general, nowadays the automatic generation of distance matrices is a straight forward process. The following part of the article describes methods to improve the road freight transport statistics by using a distance matrix.

3. Methods for calculating a distance matrix

3.1. General considerations

As vehicles registered in Austria could operate anywhere in Europe as well as outside, a European wide matrix would be necessary. Due to the computing time caused by calculating combinations of all European postal codes and the fact that some postal codes may change over time, an ongoing maintenance of the European postal codes would cause far too much effort. Furthermore updates of the distance matrix should be possible at regular intervals (e.g. every five years) without major changeovers of the underlying internal processes.

It is advisable to analyse the most frequently used places of loading and unloading of previous journeys. In Austria more than 90 % of the journeys of vehicles registered in Austria are performed within the national territory. Based on this information it became necessary to subdivide the methods used into journeys on national territory on the one hand and journeys abroad on the other hand. Regarding the high percentage of journeys performed on national territory it was primarily important to develop a particularly accurate matrix referring to these distances.

Concerning the distances abroad it was recommended to find a more common and especially more practicable approach. Therefore it was advisable to find a way to aggregate the postal codes abroad. One possibility would be to use NUTS 3 regions due to the fact that Eurostat already offers correspondences between NUTS 3 regions and postal codes in the tercet-database[3]. Using these already existing correspondences allows the creation of a distance matrix for all NUTS 3 regions within the European Union without considerable effort in the development.

For Austria it was more effective to keep the historical access of using so called postal code regions. These postal code regions have been evaluated in the 1980's based on regional subdivisions which summarize a respective number of territorial neighbouring postal codes. They are located below the NUTS 3 regions and hence they will provide more precise distances. In Austria more than 80 % of journeys abroad accounted to Germany, Italy and Switzerland. Taking this into account, it was required to find a method to calculate the distances for these countries and additionally another access for countries with fewer journeys.

After finding the appropriate regions (NUTS 3 or any other defined region) as a basis for the distance matrix, the next step is to decide on a centroid (geo-coordinate) representing each region. Then the calculation of the distances between all combinations of these centroids (geo-coordinates) can be performed as Origin-Destination Matrix using the appropriate GIS-software (including routing options).

The following part of the article describes the development of a distance matrix for Austria. The description should serve as guide for other countries which are interested in developing a similar system.

3.2. Distances within the country

Finding a representative geo-coordinate

To calculate the distances between postal code areas, a specific geo- coordinate that could be used as centroid for routing tasks had to be determined for each postal code. Initially a purely population weighted centroid based on the population numbers from the Austrian population census of 2011 and the coordinates from the register of buildings and dwellings (AGWR) was chosen. The AGWR contains address details of parcels, buildings and dwellings (including x,y-coordinates) as well as structural data for buildings, dwellings and other usage units. It

[3] http://ec.europa.eu/eurostat/tercet/locality.do

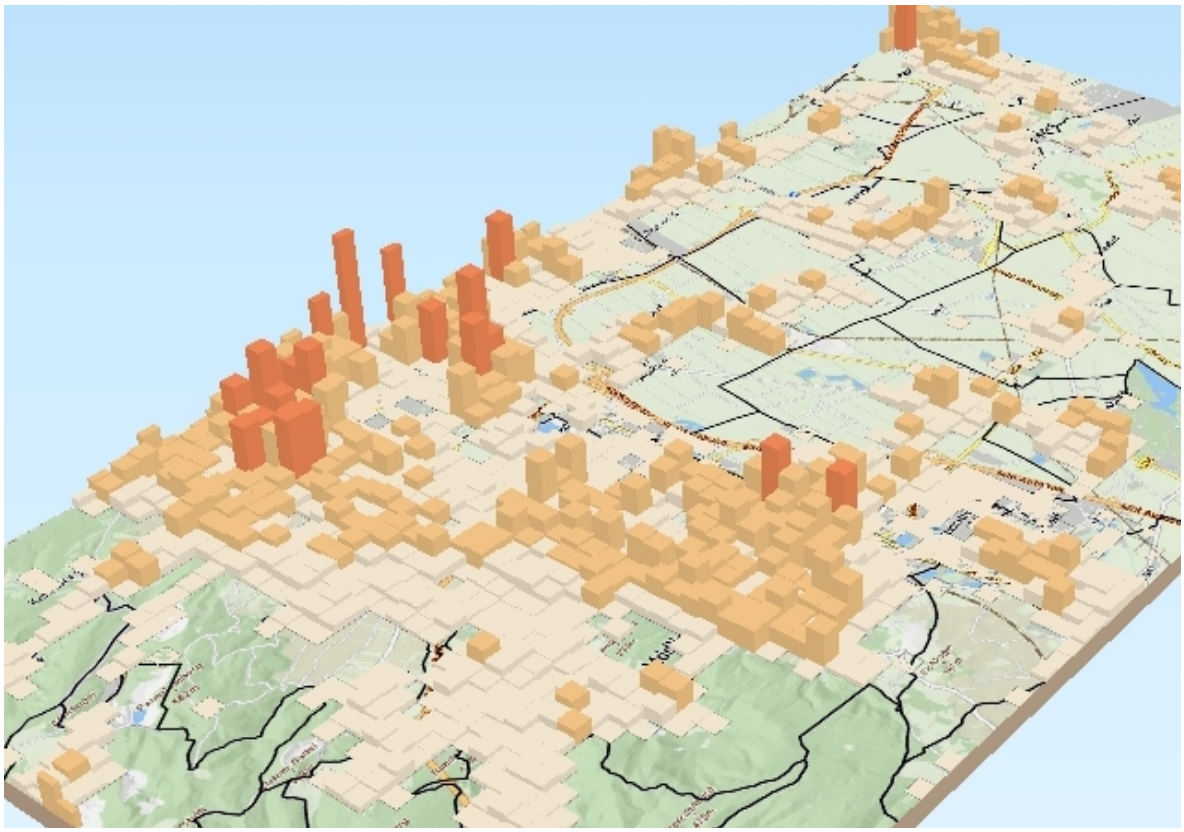


Figure 2: Residential population (district of Mödling and surroundings)

is linked with the Austrian population register, and thus contains the number of people living at any given address.

Nevertheless, this method had some weaknesses as it only considered the residential population. Moreover, industrial and commercial areas were severely underrepresented. Therefore, a new method based on both the residential and the "daytime population" was implemented. For the daytime population, the population is not counted based on the place of residency of an individual, but rather where it is likely to be during the day, e.g. on its work and school place respectively. Figure 2 and figure 3 clarify the large differences between daytime and residential population. Figure 4 depicts centroids dervied from these population measures.

In this new method for determining the central points a combination of daytime and residential population was used. It can be described as follows:

- Determine the weighted centroid of a postal code area based on the sum of the residential population and the daytime population.
- Move this point to the closest building with a residential or daytime population >0 that lies inside an area of permanent settlement.
- Move this point again to the closest street or crossing, considering the rank of the street.

Distances between postal code areas

The calculation of the actual distances between the central points was based on the TomTom routing network and was implemented in ESRI ArcGIS 10.1 with the Network Analyst extension. The routing system allows an accurate distance determination based on several features:

- The subdivision according to road sections, which include the distance from one crossing to the next, whereas for each of these sections the maximum speed is stored.

Figure 3: Daytime population (district of Mödling and surroundings)

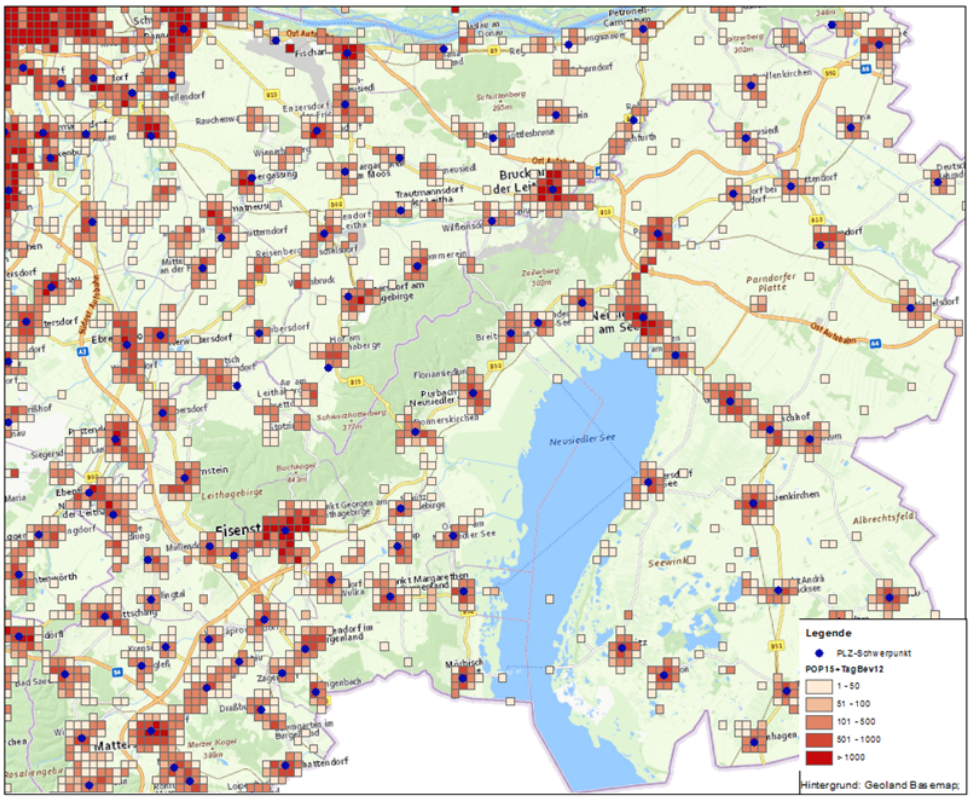


Figure 4: Weighted centroids of postal areas in the northern Burgenland (federal state of Austria)

- Information about restricted road access, one way streets, toll roads and also on the overriding road network
- Information, if a special road section is located in a built-up area.
- Based on the maximum speed and the information on street sections lying in built-up areas and major cities resp. Statistics Austria developed a speed model as bases for the calculations.

The route chosen was the fastest route between two postal code centroids, based on the STAT speed model (Kaminger and Vojtech 2016). Certainly, the fastest connection is not necessarily the shortest one, but experience has shown that mostly the high-level road network is used.

Distances within one postal code area

Journeys within one postal code area have both loading place and unloading place within the same postal code area. Therefore these journeys should be treated separately. These journeys are often "delivery or collection journeys" like e.g. grade supplies for retail stores, beverages deliveries or waste collections.

As the method described above could not be applied for these special cases, a different access based on the geographic extent of a postal code area was developed. Initially, only the centre points of postal code areas were available at Statistics Austria. Information about the geographic borders was not at hand. These points were used to generate a Voronoi-diagram based partitioning of Austrian national territory. The calculation of the transport distance (kmDis) was then based on the diagonal of the minimum bounding rectangle (bounding box) of the respective polygons associated with each postal code.

A straightforward approach would be to define the requested distances as half the diagonal of these bounding boxes and use it for the calculation of the kilometrage:

$$km_{Dis} = \frac{km_{Diagonal}}{2}$$

Regarding the landscape of Austria it is clear that the approach described above does not fit for each area as there are many alpine regions and woods to take into account. Consequently, it was required to choose a refined approach. Thus, the share of the settlement area – the available area for agriculture, settlement and industry - was also considered. First analysis showed that taking the share of the settlement area as factor as it was it resulted in kilometre distances too low for areas with a very low share of the settlement area. Based on this experience it was decided to set the factor to 25 % at least.

$$km_{Dis} = \left(\frac{km_{Diagonal}}{2}\right) \times \max(25\,\%, \text{share of settlement area in per cent})$$

In order to explain the access more practically two examples are presented in the following:

Vienna – Down Town

The bounding box for Vienna's central district has an area of 2.89 square kilometres and a 100 % share of settlement area. Half the diagonal of the bounding box is 1.5 kilometres. Therefore, the resulting distance for Vienna – Down Town is 1.5 kilometres.

Sölden

A totally different example is an alpine region like Sölden im Ötztal. The area of the bounding box is 160.7 square kilometres with a share of 3.47 % settlement area. Half the diagonal of the bounding box is 13.3 kilometres which is longer than the major road within Sölden and therefore an unrealistic high value. The weighting - based on the fact that the share of the settlement area is lower than 25 % - is done with 0.25, resulting in a calculated distance of 3.32 kilometres. This value seems to be plausible due to the fact that the total length of the only major road in this postal code is about 7 kilometres.

3.3. Calculation of the matrix abroad

As with national data, a central point (place of loading/unloading) had to be defined for each region. Since data necessary to calculate the population weighted centroids was not available for all of Europe, a different method had to be developed. If more than 10 trips to/from a postal code region were available, the weighted centre point was defined as the geographic mean of those origins/destinations. Usually that was the geographic centre of the postal code region with the most journeys. As mentioned before, more than 80 % of all journeys concerning foreign countries performed with vehicles registered in Austria have affected Germany, Italy or Switzerland throughout the last years. To be as valid as possible, all journeys of the last eleven years concerning Germany and Italy have been regarded based on the postal code combinations.

For Switzerland or if there were less than 10 trips to/from a postal code region in Germany or Italy available, the central point was defined manually based on local geographic and urban features such as industrial areas or important ports.

For other countries, it was not deemed necessary to pre-calculate any distances. Those are calculated on a case-by-case basis and inserted into the matrix as required.

Alternatively, Karner, Scharl, and Weninger (2014) describe a methodology for determining central points of NUTS 3 regions based on the Urban Clusters (European Commission, 2006) and CORINE land cover (European Environment Agency) datasets. This method can easily be implemented by anyone, as all the necessary data is free of cost available from the respective agencies.

Once a coordinate has been defined for each postal code- or NUTS 3-region, the distances between the regions can be calculated either in a dedicated GIS database or using external routing services such as google maps or open street map. For the distance matrix outside of Austria, the commercial routing software Microsoft Map Point 2011 was used. This was necessary as the routing network used for calculating Austrian domestic transport distances was only available for Austria.

Even on this aggregated level, calculating all possible routes would have been too inefficient. As the methodology presented in this paper is flexible, it is easy to update an existing matrix of pre-calculated distances on demand, if new origin-destination combinations are required.

4. Odometer information as benchmark for the road freight survey

To verify the developed distance matrix as well as the quality of the survey, the distances from the distance matrix were compared with the odometer information received from the questionnaire. This was done with data of the Austrian road freight survey from the reference year 2015.

As previously mentioned, the respondent has to fill in the place of loading and the place of unloading for each journey during the reference week, which is then used to obtain the kilometres driven from the distance matrix. Additionally, the number of kilometres according to the odometer at the beginning and at the end of the week has to be provided. The difference of these data represents the kilometres driven during the reference week.

For a comparison of these two data sources it has to be considered that not every journey is reportable. Journeys on private roads (such as forest roads, roads within a factory, hospital grounds or construction sites) are excluded from the survey, as are winter services (snow removal, gritting) and road maintenance. Therefore, the reported journeys are only a subset of all journeys driven during the reference week.

Another important point is that some odometer readings might be incorrect. As highlighted before and also mentioned in the SYSTRA study, the questionnaire is often filled in by not directly related departments (e.g. accounting) instead of the actual driver. As they might fill in the questionnaire after the vehicle was driven, the provided information (like the actual

Table 1: Comparison of odometer- and distance matrix in million km driven for the year 2015

Variable	Odometer	Distance matrix
Total annual km	12.41	11.85
Own account	5.09	4.79
Hire or reward	6.94	6.71
C10 - C32 Manufacturing	1.08	1.03
E38.1 Waste collection	0.30	0.26
F41 - F43 Construction	0.73	0.70

odometer reading at the start or the end of the week) might be incomplete.

The analysis is based on 19 583 reported reference weeks in 2015. Out of these, 4 617 weeks were eliminated because the received difference of the odometer readings was zero. Assuming that a driven distance of more than 3 000 km per week might be too high and consequently incorrect, 1 372 cases were also excluded. After these plausibility checks 13 594 weeks were used for further analysis.

The results of this comparison are recorded in table 1. This table contains aggregate statistics for 2015 such as total annual kilometres and annual kilometres grouped by NACE (Council of the European Union 2006) and transport type. It can be seen that the differences between the reported kilometres and the kilometres estimated from the distance matrix are very small. As assumed the estimated kilometres are slightly lower than the reported kilometres. This indicates that for this level of detail, the approach described in this paper works quite well.

Figure 5 provides a more detailed look on total annual kilometres driven in 2015. It illustrates the dependency of the odometer reading (horizontal axis) on the reported kilometres taken from the developed distance matrix (vertical axis). The Pearson correlation coefficient of the two variables is 0.91 and shows a high positive linear correlation between the two variables.

Despite the high correlation, outliers were detected. Generally, the comparison of the accumulated kilometres from the distance matrix with the odometer reading is used as plausibility check for the data of the road freight survey in Austria. If the data differs by more than 30 %, the employees of the statistical office contact the respondents to clarify the discrepancies. As the main focus of this analysis was the improvement of the survey with regard to underestimation, vehicles, whose kilometres from the distance matrix were higher than the odometer reading, were accepted and no further enquiry to the respondents was realised. In case of outliers in the opposite way, where these kilometres were lower than the odometer readings, the respondents were called to identify the reasons of the underestimations. The general feedback was that the odometer reading had been incorrect or there had been no reportable journeys during the reference week. This indicates a good quality of the developed distance matrix and emphasises the reasonable use of a distance matrix with regard to the reduction of respondents' burden. Furthermore, as this comparison is an additional possibility for plausibility checks it improves the data quality of the survey.

Another analysis included the classification of the journeys in "Hire or reward" (NACE 49.4 Freight transport by road and removal services) and "Own account" (other NACE positions). Both showed a correlation coefficient of 0.9 which was only marginal lower than in the whole sample (see figure 6). As illustrated, vehicles of "Hire or reward" drive longer distances per week than those belonging to the classification "Own account".

Furthermore, the reference weeks of the companies were analysed by the NACE activities "Manufacturing" (C10-C32), "Waste collection" (C38.1) and "Construction" (F41-F43). It can be seen that both activities "Manufacturing" and "Construction" have a higher correlation coefficient than "Waste collection" (see figure 7). On the one hand, this is due to the fact that

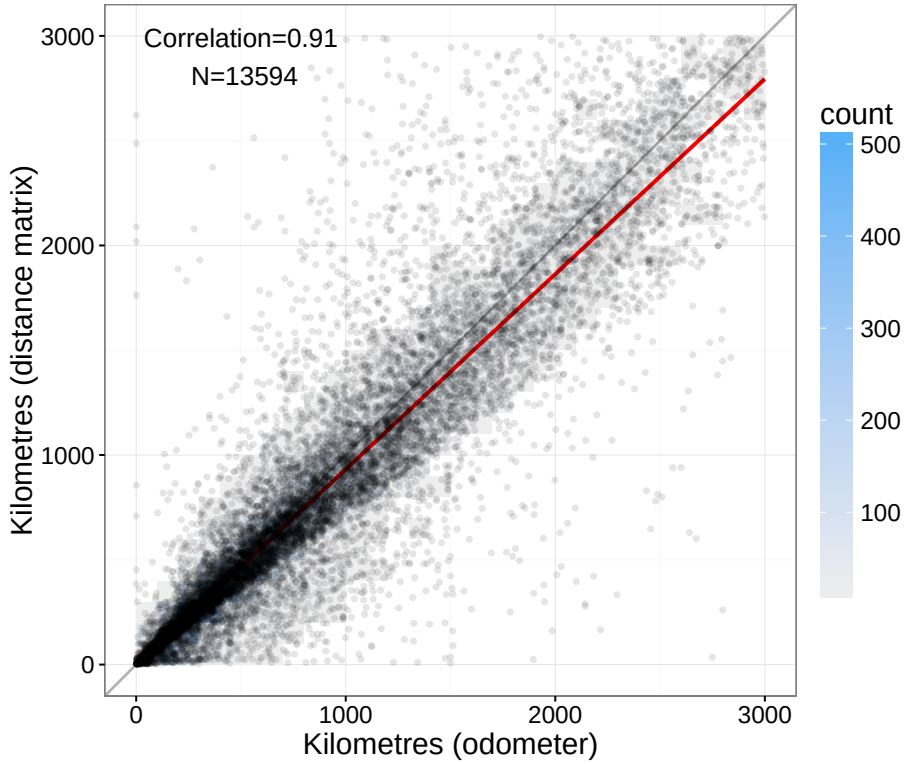


Figure 5: Comparison of odometer information with kilometres from the distance matrix for all vehicles in 2015

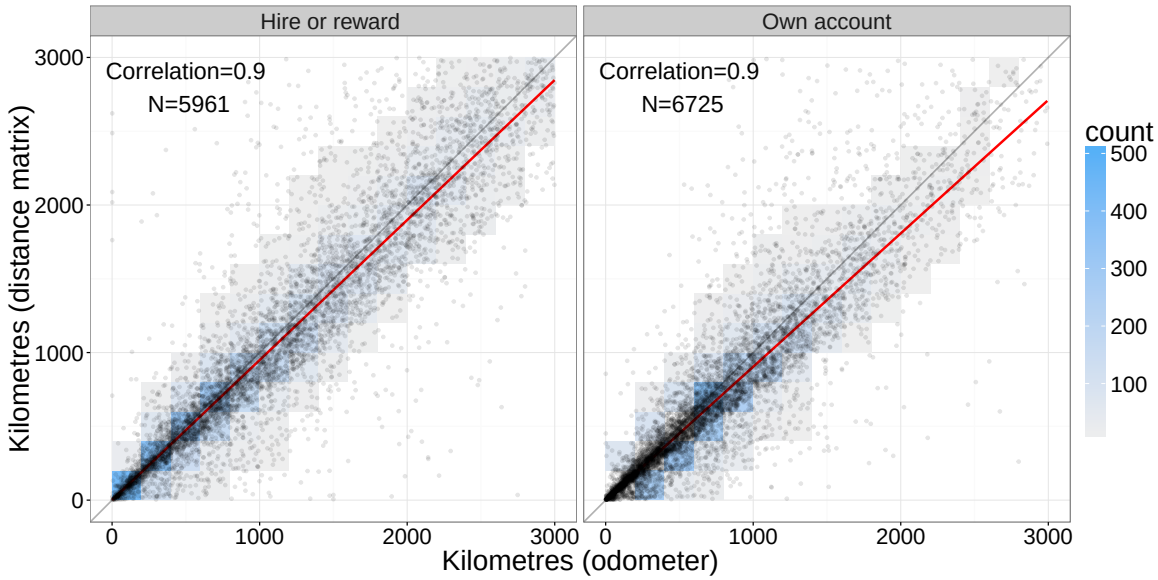


Figure 6: Comparison of odometer information with kilometres from the distance matrix for the classifications "Hire or reward" and "Own account".

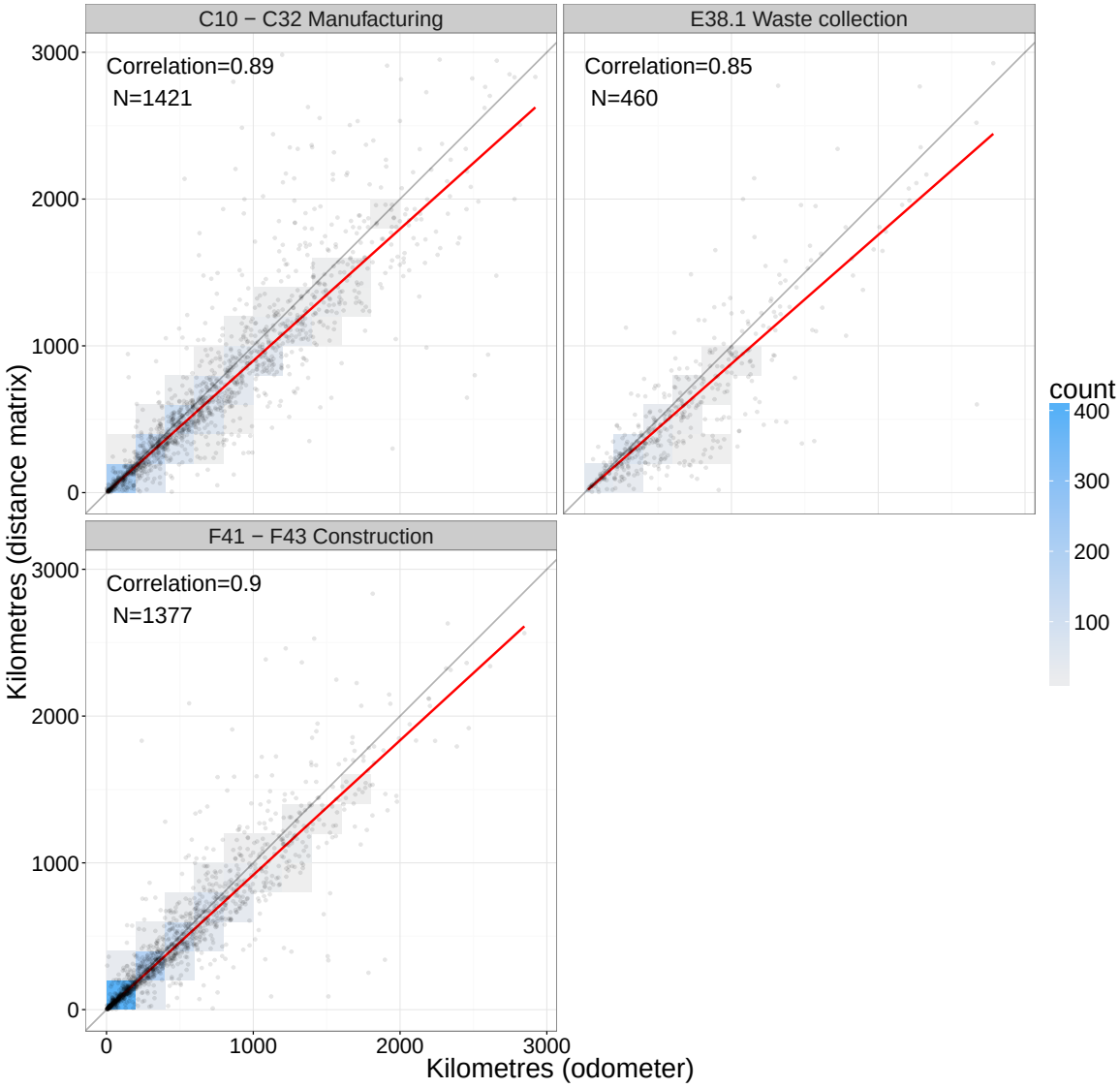


Figure 7: Comparison of odometer information with kilometres from the distance matrix for the classifications “Manufacturing”, “Waste collection” and “Construction”.

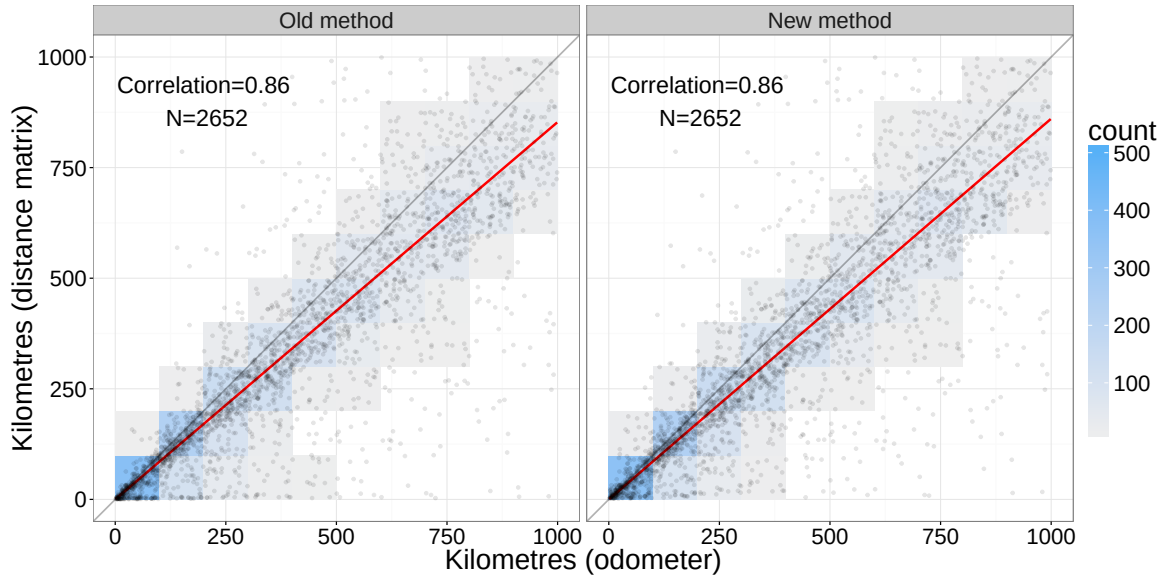


Figure 8: Effect of the updated distance matrix on domestic journeys within one postal code area.

journeys for waste collection are quite derived. The trucks often drive around several streets within one area, for which reason the calculated kilometres are below the actual weekly driven kilometres reported by the odometer information. On the other hand, journeys for collecting waste are short and therefore the already mentioned problem with journeys within one postal code influences the discrepancy. For the activities "Manufacturing" and "Construction" it has to be kept in mind that journeys on the construction site as well as the factory site do not have to be reported. As a consequence, the calculated kilometres naturally have to be lower than the ones of the odometer information.

Regarding the analysis of kilometres within one postal code area it was investigated how the new version described before would affect the discrepancy of the two variables (see figure 8). As there were no data available for the whole reference year, the comparison was limited to the first quarter of 2015. The graph on the left site shows the old approach of the distance matrix (for every journey within one postal code 1 km was taken), whereas on the right site the approximated distance on the basis of the postal code size was used. Both using only journeys within Austria. There was no increase of the correlation coefficient, which remained at 0.86. Nevertheless, when comparing the total transport distances measured by odometer (1.10 million km) with the distances estimated with the old method (0.92 million km) and the new method (0.94 million km), it becomes clear that the new method produces slightly better results.

5. Conclusion

The aim of this article was to point out the advantages of a distance matrix and to present a general guidance to create a distance matrix for a country to facilitate the survey for road freight transport statistics. The distance matrix is a reasonable instrument to decrease the burden of the respondents.

In order to eliminate the obligation to calculate the kilometres driven or to record all odometer readings for the different journeys the distance can be calculated automatically by the statistical office through the place of loading and the place of unloading. It is indispensable to renew and update the distance matrix regularly as infrastructure and population focus change over time.

Together with the odometer information for a specific period it can also be used as an addi-

tional plausibility check to increase the quality of the data, although it is sometimes difficult to compare the sum of all calculated kilometres with the odometer information of the reference week due to journeys which are not reportable.

The comparison of the odometer information with the distances estimated with the distance matrix showed a high positive correlation ($r = 0.91$) for the Austrian data of the reference year 2015. This indicates a good quality of the data and combined with the achieved reduction of respondents burden it strengthens also the use of a distance matrix in road freight transport statistics.

Certainly, there is still some work to be done on the improvement of the presented distance matrix. As seen in the last chapter, distances for driving within one postal code or for specific kinds of transport (e.g. delivery and collection journeys as waste collection) have to be analysed further as there are still discrepancies between the calculated kilometres by the distance matrix and the actual driven kilometres.

For the future there are several approaches possible:

- The estimation of weights for journeys within one postal code region, grouped by parameters such as the length of the high-level road network or the extent of industrial or residential areas.
- Special questionnaires adapted for delivery or collection journeys (e.g. no type of goods) with additional information on kilometres from the respondents.
- The use of mobile apps as a new kind of questionnaire for the road freight transport survey. The main benefit would be the monitoring of accurate kilometres driven based on GPS-technology.

References

Council of the European Union (2006). "Regulation (EC) No 1893/2006 of the European Parliament and of the Council of 20 December 2006 Establishing the Statistical Classification of Economic Activities Nace Revision 2 and Amending Council Regulation (EEC) No 3037/90 As Well As Certain EC Regulations on Specific Statistical Domains." URL http://data.europa.eu/eli/reg/2006/1893/oj.

Council of the European Union (2012). "Regulation (EU) No 70/2012 of the European Parliament and of the Council of 18 January 2012 on Statistical Returns in Respect of the Carriage of Goods by Road." URL http://data.europa.eu/eli/reg/2012/70/oj.

European Commission (2003a). "Commission Regulation (EC) No 6/2003 of 30 December 2002 Concerning the Dissemination of Statistics on the Carriage of Goods by Road." URL http://data.europa.eu/eli/reg/2003/6/oj.

European Commission (2003b). "Commission Regulation (EC) No 642/2004 of 6 April 2004 on Precision Requirements for Data Collected in Accordance With Council Regulation (EC) No 1172/98 on Statistical Returns in Respect of the Carriage of Goods by Road." URL http://data.europa.eu/eli/reg/2004/642/oj.

Eurostat (2011a). "European Statistics Code of Practice for the National and Community Statistical Authorities." *Technical report*, European Commission. URL http://bookshop.europa.eu/is-bin/INTERSHOP.enfinity/WFS/EU-Bookshop-Site/en_GB/-/EUR/ViewPublication-Start?PublicationKey=KS3211955.

Eurostat (2011b). "Road Freight Transport Methodology - Volume 1: Reference Manual for the Implementation of Council Regulation No 1172/98/EC on Statistics on the Carriage of Goods by Road." *Technical report*, European Commission. doi:10.2785/18474. URL http://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-BI-05-001.

Kaminger I, Vojtech N (2016). "Census 2011 — Enriching Commuter Statistics." To be published by Eurostat in 2016.

Karner T, Scharl S, Weninger B (2014). "Estimation of the Domestic Transport Performance From the Consolidated European Road Freight Transport Data." *Austrian Journal of Statistics*, **43**(1), 49. doi:10.17713/ajs.v43i1.8. URL http://dx.doi.org/10.17713/ajs.v43i1.8.

Systra (2015). "Experience of Completing the Continuing Survey of Road Goods Transport." *Technical report*, Department for Transport. URL https://www.gov.uk/government/publications/experience-of-completing-the-continuing-survey-of-road-goods-transport.

Affiliation:

Thomas Karner
Directorate Business Statistics - Transport
STATISTICS AUSTRIA
A-1110 Vienna, Austria
E-mail: thomas.karner@statistik.gv.at

Ingrid Kaminger
Cartography and Geographic Information Systems
STATISTICS AUSTRIA
A-1110 Vienna, Austria
E-mail: ingrid.kaminger@statistik.gv.at

ISSN 1026–597X

Contents