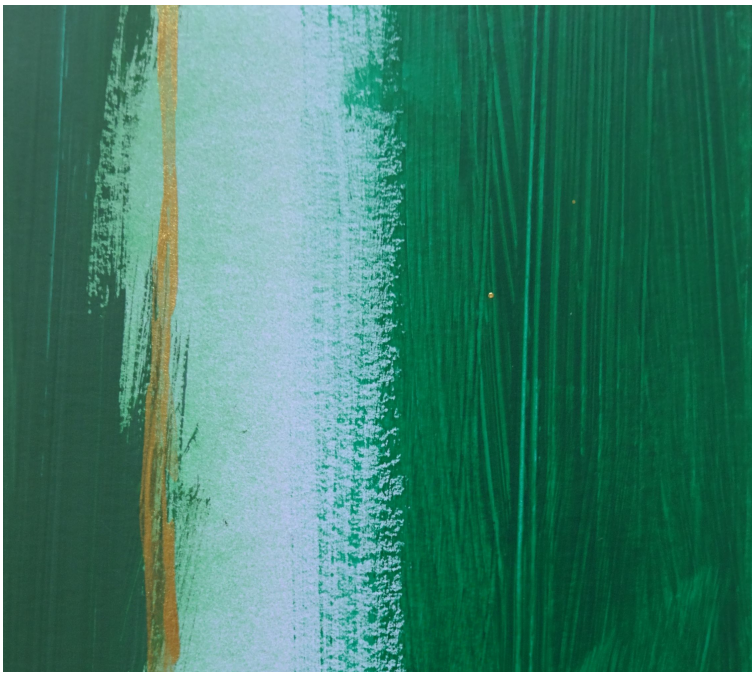
Austrian Journal of Statistics

AUSTRIAN STATISTICAL SOCIETY

Volume 45, Number 2, 2016

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 45, Number 2, 2016

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 45, Heft 2, 2016

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Impressum

Editor:	Matthias Templ, Statistics Austria & Vienna University of Technology
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 Müller, 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
Editorial Help:	Klára Hrůzová, Palacký University, Olomouc, Czech Republic
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 includes four scientific papers and one interview with the following content.

Issues on quality in register-based statistics is a sensitive topic and hard to determine in numbers. While in Germany they established a survey and use small area estimation methods to evaluate the quality of the census register (and to correct the registers information), the authors of the first contribution define a framework of deterministic rules for registers without the help of a survey, and they use the Dempster-Shafer approach to aggregate quality indicators. This is presented in the first paper.

The second paper is dedicated to biostatistics. New models for background correction of affymetrix data are proposed. I would like to thank Herwig Friedl for substantial improvements and the editing of this paper.

The third paper discusses Riesz and Beta-Riesz distributions, important distributions in a multivariate setting.

The fourth contribution shows very interesting aspects and proposals on life expectancy loss due to a certain disease in comparison with a general reference population. Official statistics get married with survival analysis.

I'm proud to announce the interview with Ernst Stadlober. Herwig Friedl and myself conducted the interview in December in Graz in a nice charming atmosphere. This familial and productive atmosphere seems to always be present at the Institute of Statistics at the TU Graz. Some reasons for this can be extracted from this interview.

Matthias Templ
(Editor-in-Chief)

Statistics Austria & Vienna University of Technology
Wiedner Hauptstr. 8–10
A–1040 Vienna, Austria
E-mail: matthias.templ@gmail.com

Vienna, March 2016

Austrian Journal of Statistics
March 2016, Volume 45, 3–14.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i2.97

Quality Assessment for Register-based Statistics - Results for the Austrian Census 2011

Asamer Stat. Aus. | **Astleithner** Univ. Vienna | **Ćetković** WU | **Humer** WU | **Lenk** Stat. Aus. | **Moser** WU | **Rechta** Stat. Aus.

Abstract

In 2011, Statistics Austria carried out its first register-based census. Advantages of using administrative data for statistical purposes are, among others, a reduced burden for respondents and lower cost for the National Statistical Institutes (NSI). However, new challenges, like need for a new approach to the quality assessment of this kind of data arise. Therefore, Statistics Austria developed a comprehensive standardized framework to evaluate data quality for register-based statistics. In this paper, we present the basic concept of this quality framework and provide detailed results from the quality evaluation of the Austrian census of 2011. More specifically, we derive a quality measure for each census attribute from four complementary hyperdimensions. The first three of these hyperdimensions address the documentation of data, the usability of records and an external data validation. The fourth hyperdimension focuses on the quality of data imputations. The proposed framework combines these different quality-related information sources for each attribute to form an overall quality indicator. This procedure allows to track changes in quality during data processing and to compare the quality of different census generations.

Keywords: administrative data, register–based census, quality assessment, Austrian Census 2011.

1. Introduction

The importance of administrative data as an input for statistical purposes has increased steadily in the last decades. Following the example of Scandinavian countries, about one third of the United Nations Economic Commission for Europe (UNECE) members now base their census at least partially on administrative data (UNECE 2014). In Austria, the last survey-based census was conducted in 2001 and was replaced with a register-based census in 2011. The advantages of the new approach are manifold and comprise inter alia a reduced burden for respondents and lower costs. However, quality assessment for administrative data has only received little attention in the statistical literature. Hereafter, we present a standardized quality framework for the assessment of administrative data, which was developed by *Statistics Austria* in the course of the Austrian register-based census of 2011. The procedure tries to aggregate all available (meta-)information to generate a single quality indicator for each attribute for each statistical unit.

This paper is structured as follows. Section 2 introduces the data sources for the register-based

census. In section 3, the quality framework is introduced using the example of the quality assessment for the variable *Legal Marital Status* (LMS). Section 4 provides summary results for the overall quality assessment of the Austrian census and finally section 5 concludes.

2. Sources for the register-based census

A decisive quality–related topic for register–based statistics is the selection of appropriate data sources that supply certain required information. To give an overview for the data sources used in the Austrian census, Figure 1 illustrates the connections between the actual data sources (administrative registers), the topics of the census (data cubes) and the final census data, so-called statistical registers, which are derived from the data cubes.

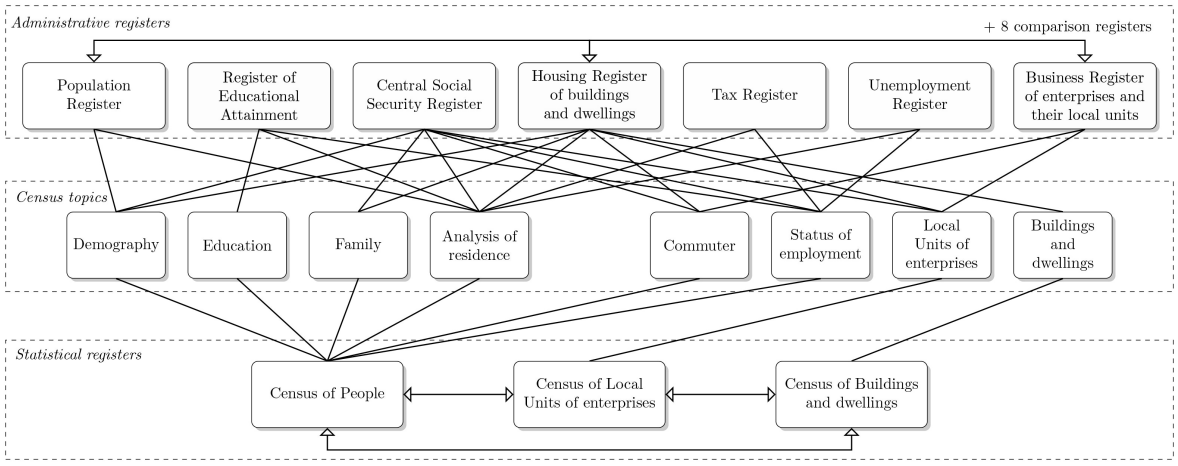


Figure 1: Data sources for the register–based census
The administrative registers provide information for the register based census (first row). This administrative information is gathered in data cubes (second row). The actual census data are "statistical registers", i.e. they are derived from the data cubes (third row).

For the purpose of the census, *Statistics Austria* distinguishes between 7 base registers and 8 comparison registers. While the base registers are needed to provide all attributes of interest for the register-based census, the subset of grey shaded registers form the backbones of the census. Especially, they determine the population count, the number of buildings and dwellings and the quantity of enterprises as well as their local units. The base registers are maintained by *Statistics Austria*. To further improve the quality of the census, the base registers are supported by eight comparison registers, which gather additional information for cross-checks from more than 50 external data holders.[1] For cases where more than one autonomous data source is available for a specific attribute, the different registers are used to mutually validate information. Accordingly, we apply a *principle of redundancy* to improve quality of data (Lenk 2008, p. 3).

3. The quality assessment of administrative data

National Statistical Institutes (NSI) often rely on external data sources for which they are not directly responsible. Such third party data sources may constitute a large part of the required information for e.g. population censuses, as it is the case for Austria. Hence, the relevance of a comprehensive quality assessment in the process of using register–based statistics has to be emphasized. The following approach for the evaluation of administrative data extends previous work by other NSIs in this field (Daas, Ossen, Vis-Visschers, and Arends-Tóth 2009;

[1] If data is not or only partly available in the base registers, information is derived from the comparison registers as well (Berka, Humer, Lenk, Moser, Rechta, and Schwerer 2010, p. 300).

Daas and Fonville 2007) and especially relies on four quality-related hyperdimensions (Berka *et al.* 2010; Berka, Humer, Lenk, Moser, Rechta, and Schwerer 2012).

Besides these quality dimensions, the actual data processing for the Austrian census is conducted in three stages that have to be considered in the quality assessment: the raw data (i.e. the administrative registers i), the combined dataset of these raw data (*Central Data Base*, CDB) and the final dataset, which includes imputations (*Final Data Pool*, FDP). The two latter are referred to as "statistical registers". Figure 2 illustrates the data processing, beginning with the delivery of raw data from the various administrative data holders. The four hyperdimensions, named HD^D, HD^P, HD^E and HD^I, aim to assess the quality for different types of attributes at all stages of the data processing. This yields quality measures which are standardized between zero and one, where a higher value indicates better data quality.

In this process, HD^D describes quality-relevant processes at the register authority, HD^P deals with formal errors in the data and HD^E measures the data quality in comparison to an external source. Finally, HD^I computes the quality of the imputations. $q^n_{i,j}$ denotes the combined quality measure of HD^D, HD^P and HD^E for an attribute j in a register i at the observation level n (e.g. $q_{1,A}$ describes the quality of attribute A in register 1).

The data from Austrian administrative registers can be linked through unique personal identifiers (so-called *branch-specific personal identification number*, bPIN) and are collected in data cubes, namely the CDB. We distinguish three types of attributes in the CDB, which have to be treated differently from a quality perspective: *Unique attributes* (see attribute C in Figure 2) exist only in a single administrative register. *Multiple attributes* are available from more than one administrative register (see attribute A). *Derived attributes* are not covered by any register in the required form. Therefore, such attributes have to be derived from related information (see attribute F and G).

We denote the combination of the raw data quality measures on the CDB level as $q^n_{\odot,j}$. For multiple and derived attributes, the CDB is then again validated using the hyperdimension *External Source* (HD^E), to capture quality-related effects of the register aggregation process. The final quality indicator at this stage is then given by a combination of $q^n_{\odot,j}$ and the HD^E quality, which we refer to as $q^n_{\Psi,j}$, the effective CDB quality indicator.

Finally, the quality of imputed values has to be taken into account. Before the imputation process, the quality of missing values on raw and CDB level is zero by definition. The hyperdimension HD^I assesses the quality of the imputations and replaces the zero default-quality for missings with a new value, which depends on the imputation method. Again, every statistical unit n obtains a quality indicator for every attribute j, which we denote $q^n_{\Omega,j}$, the final FDP quality measure.[2]

Based on this framework, we will in the following illustrate the calculation of the quality indicators in detail using the example of the attribute *Legal Marital Status* (LMS). This process includes the assessment of the quality on the raw data level for each observation in each register as well as the evaluation of the statistical registers CDB and FDP.

3.1. The raw data level

The quality assessment starts by collecting quality information at the raw data level (left boxes in Figure 2), which is obtained via the three hyperdimensions Documentation (HD^D), Pre-processing (HD^P) and External Source (HD^E). A detailed explanation for each hyperdimension is given below.

Hyperdimension Documentation (HD^D)

HD^D describes quality-related processes as well as the data documentation (metadata) at

[2]While all indicators are available on an individual basis, we drop the n superscript in subsequent notation for reasons of readability.

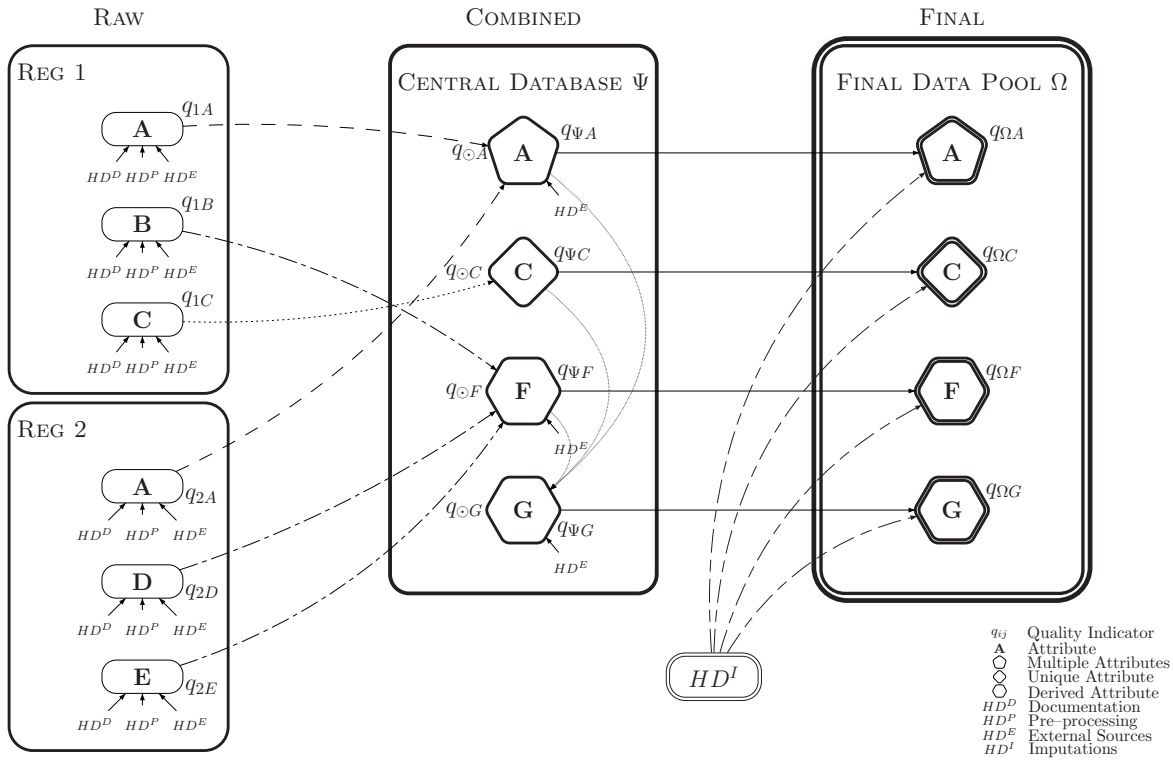


Figure 2: Schematic overview of the quality framework for register–based censuses.
On the raw level, every attribute (A-E) in the administrative registers REG1 and REG2 is evaluated through hyperdimensions HD^D, HD^P, HD^E. Combining all quality-related information yields the CDB quality measure $q_{\Psi,i}$, measured on the level of statistical units. In the final step, the quality of imputations is assessed for the FDP using HD^I. At the end of the process every statistical unit has obtained a quality value between zero and one.

the administrative authorities. Data holders are assigned a degree of confidence and reliability which are monitored using a questionnaire on quality-relevant procedures, that contains several open and scored questions. Administrative authorities are requested to answer the questionnaire for every attribute seperately, to control for different treatment of single variables. The scored questions are used for the quality-assessment and can be divided into four subgroups, covering data historiography, definitions, administrative purpose and data treatment. Through these scored questions, every attribute j in each administrative register i obtains a quality measure $HD^D_{i,j}$ at this stage, as described by equation 1.

$$HD^D_{i,j} = \frac{\text{obtained score}}{\text{achievable score}} \tag{1}$$

The set of open questions serves as complementary information but is not considered in the quality assessment. Table 1 gives an overview over the current set of questions.

For the exemplary case of the LMS, data is obtained from eleven source registers which have to be assessed individually. The calculation of the hyperdimension documentation (HD^D) for each source register is illustrated in Table 2.

The data holders answer quality related questions on a dichotomous (yes/no) or an ordinal scale. For each question, a higher value indicates better quality-related performance of the register. Furthermore, each question is weighted by its relative importance, which was determined by experts in the field of register-based statistics at Statistics Austria. The metadata quality for each register is summarized as the weighted average of these scored questions. For example, a value of 1 for the question "Definitions" in the central population register (CPR) indicates, that the definition of the Legal Marital Status is consistent between the CPR and the register-based census.

In practice, data for a single comparison register may be delivered from up to 20 different

Table 1: Scored Questions — HD Documentation

Data Historiography	Level of measurement
Can we detect data changes over time? [Detect Changes]	dichotomous
Is the information available for the cut–off date?[1] [Cut-off date]	dichotomous
Definitions	
Are the data definitions for the attribute compatible to those of Statistics Austria? [Definitions]	dichotomous
Administrative Purpose	
Is the attribute relevant for the data source keeper? [Relevance]	dichotomous
Does a legal basis for the attribute exist? [Legal basis]	dichotomous
Data Treatment	
How fast are changes edited in the register? [Timeliness]	ordinal
Are the data verified on entry? [Administrative Control]	dichotomous
Are technical input checks applied [Technical Control]?	dichotomous
How good is the data management, i.e. ex post consistency checks? [Data management]	ordinal

[1] This refers to the availability of data for a particular reference date.

data authorities (e.g. regional offices). For such cases, the hyperdimension Documentation is applied separately for each delivery and then processed, so that these sources are aggregated to one comprehensive comparison register. To assess the HD^D quality for such complex cases, the relative contribution to the comparison register (in terms of observations) is used to compute a weighted average for the questionnaire of each (regional) delivery. One example of such a register is the Social Welfare Register (SWR). Consider for example the "Cut-off date" in Table 2. The indicator yields an average value of 0.47 for all data deliveries. Equivalently, for only 47 per cent of the records in the SWR data copies for a specific reference date are available.

The weighted average of the scores on the different questions forms the quality indicator on the register level (HD^D_{LMS}). Results are shown in the last row in Table 2. A lower value, as presented for the Asylum Seekers Register (ASR), indicates bad performance of the attribute in this administrative register with respect to the Hyperdimension HD^D. On the contrary, a high value, as found in the Tax Regsiter (TR), indicates a good performance.

Table 2: Calculation of hyperdimension documentation HD^D for Legal Marital Status (*LMS*)

HD-Subdimension	Weight	ASR	UR	RPS	CAR	CFR	CSSR	CHR	HPSR	SWR	CPR	TR
Detect Changes	1	0.00	1.00	0.87	1.00	0.00	1.00	0.67	0.35	0.51	1.00	1.00
Cut-off date	2	0.00	1.00	0.87	1.00	0.00	1.00	0.67	0.35	0.47	1.00	1.00
Definitions	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Relevance	4	0.00	0.00	0.62	1.00	0.00	1.00	0.67	0.7	0.83	0.00	1.00
Legal basis	1	0.00	1.00	1.00	1.00	0.00	1.00	0.67	0.35	1.00	1.00	1.00
Timeliness	3	1.00	1.00	0.80	1.00	1.00	1.00	1.00	0.81	0.85	1.00	1.00
Administrative Contr	2	0.33	0.67	0.73	1.00	0.33	1.00	0.67	0.73	0.81	1.00	1.00
Technical Contr	2	0.67	0.00	0.70	1.00	0.67	1.00	0.78	0.49	0.77	1.00	1.00
Data management	4	0.33	1.00	0.63	0.67	0.33	0.67	0.78	0.59	0.64	1.00	1.00
HD^D_{LMS}		0.39	0.68	0.86	0.93	0.39	0.93	0.77	0.70	0.74	0.81	1.00

Notes: *ASR*: Asylum Seekers Register, *UR*: Unemployment Register, *RPS*: Register of Public Servants of the Federal State and the Länder, *CAR*: Child Allowance Register, *CFR*: Central Foreigner Register, *CSSR*: Central Social Security Register, *CHR*: Chambers Register, *HPSR*: Hospital for Public Servants Register, *SWR*: Register of Social Welfare Recipients, *CPR*: Central Population Register, *TR*: Tax Register.

Hyperdimension Pre-Processing HD^P

The hyperdimension pre-processing (HD^P) is again based on the actual raw data received by the NSI and computes the share of useless records (missing identification keys, missing

values, values out of range, see Table 3) for each attribute in each register.

Table 3: HD Pre–processing

	Number of observations [Observations]
—	Records without unique bPIN [Missing bPIN]
—	Records with item non–response (but including unique bPIN) [Non resp.]
—	Records with wrong values or values out of range [Out of range]
=	Usable records

The final result of this hyperdimension is given by the ratio of usable records to the total number of records (see equation 2).

$$HD^P_{i,j} = \frac{\text{usable records}}{\text{total number of records}} \tag{2}$$

HD^P results for the LMS in the source registers are shown in Table 4. For this case, most data sources provided formally correct information, resulting in indicators close to 1. An exception are data from the Asylum Seekers Register (ASR) and the Social Welfare Register (SWR) where a significant amount of missing unique personal identifiers (56.1 per cent and 14.4 per cent respectively) can be found. Accordingly, this procedure yields lower quality indicators for these registers.

Table 4: Calculation of the hyperdimension HD^P for the Legal Marital Status (*LMS*)

Register	Observations	Missing bPIN %	Non resp. & Out of range %	HD^P_{LMS}
ASR	66,411	56.12	3.73	0.402
UR	327,702	1.30	7.74	0.910
RPS	640,155	1.66	2.85	0.955
CAR	3,658,263	2.72	0.01	0.973
CFR	747,688	7.67	2.58	0.898
CSSR	8,811,838	6.30	48.30	0.454
CHR	23,904	3.40	41.51	0.551
HPSR	87,954	6.23	38.60	0.552
SWR	263,134	14.44	7.24	0.783
CPR	9,605,679	0.0	33.04	0.670
TR	9,359,027	6.28	9.31	0.844

Hyperdimension External Source HD^E

The hyperdimension (HD^E_{LMS}) is again conducted on the raw data level and assesses the data-quality of the source registers in comparison to an external source, which in our case is given by the Austrian microcensus. The quality indicator is simply calculated as the number of consistent values between each register and the microcensus, divided by the number of all records that could be linked to the microcensus (see equation 3).

$$HD^E_{i,j} = \frac{\text{number of consistent values}}{\text{total number of linked records}} \tag{3}$$

In Table 5, we present results for the comparison of LMS between the raw registers and the microcensus. For example, 1,239 individuals from the Unemployment Register (UR) could be linked to the microcensus. Out of these observations, 1.9 per cent were classified as inconsistent. This yields a $HD^E_{\text{UR,LMS}}$ value of 0.981 for the LMS in the UR.

Final quality on the raw-data level

Given these three quality measures, an overall quality indicator for each attribute on the register level can be derived as the weighted average described in equation 4. While these

Table 5: Calculation of the hyperdimension HD^E for the Legal Marital Status (*LMS*)

Register	Linked observations	Conflicting observations %	HD^E_{LMS}
ASR	10	50.0	0.500
UR	1,239	1.9	0.981
RPS	2,993	4.1	0.959
CAR	13,905	3.0	0.970
CFR	2,235	11.5	0.885
CSSR	20,346	5.8	0.942
CHR	71	11.3	0.887
HPSR	194	2.6	0.974
SWR	576	5.2	0.948
CPR	27,959	2.9	0.971
TR	24,332	8.9	0.910

indicators do not vary per observation (but rather register and attribute), these are still attached to each observation, so that the quality of a single record can be traced throughout the process. In our framework, each hyperdimension has the same weight ($v^D = v^P = v^E = 1/3$), reflecting our assumption on an equal impact of each dimension on the quality measure. The resulting value summarizes the existing quality-related information for each attribute j in each register i.

$$q_{i,j} = v^D \cdot HD^D_{i,j} + v^P \cdot HD^P_{i,j} + v^E \cdot HD^E_{i,j} \tag{4}$$

Table 6 summarizes the combined information for the attribute LMS for each register. Accordingly, we obtain eleven quality indicators for the LMS. The Asylum Seekers Register (ASR) has the lowest quality-measure, while the Child Allowance Register (CAR) delivers the best quality for the LMS. The differences in quality result partly from the different population subgroups covered by the individual registers (families with young children vs. foreign persons). An additional reason for the variation is the varying importance of LMS for different register authorities. LMS is highly relevant for the CAR but less important in the ASR. This fact influences quality outcomes as well.

These quality indicators on register level serve as the main input to the next step of the framework, the evaluation of the LMS in the CDB.

Table 6: Calculation of the quality indicator for the (*LMS*) for the registers

Register	HD^D	HD^P	HD^E	q
ASR	0.397	0.402	0.500	0.433
UR	0.683	0.910	0.981	0.858
RPS	0.864	0.955	0.959	0.926
CAR	0.936	0.973	0.970	0.960
CFR	0.397	0.898	0.885	0.726
CSSR	0.936	0.454	0.942	0.777
CHR	0.778	0.551	0.887	0.739
HPSR	0.706	0.552	0.974	0.744
SWR	0.746	0.783	0.948	0.826
CPR	0.810	0.670	0.971	0.817
TR	1.000	0.844	0.910	0.918

3.2. The Central Data Base (CDB)

In a next step, the data from the raw registers are combined in the Central Data Base (CDB), which covers all attributes of interest for the register–based census. At this level, a quality indicator $q_{\odot,j}$ for each attribute j and statistical unit n is computed. Concerning the evaluation of quality for the CDB we need to distinguish three types of attributes, unique,

multiple and derived as described earlier.[3]

LMS is a multiple attribute, and accordingly shows up in several registers. Since there are multiple data sources which provide *LMS*, a predefined ruleset, based on experience of Statistics Austria, picks the most appropriate value from the underlying registers according to the constellation in the source registers. To assess the validity of this chosen value, all the available information for *LMS* from all registers is taken into account. To aggregate this information, the *Dempster-Shafer Theory* (DST, see Shafer 1992) for the combination of evidence is applied to derive a quality measure for this attribute for each statistical unit. Berka *et al.* (2012) give a detailed explanation of the possibilities to apply DST to the assessment of quality.

To combine the different quality measures for *LMS* on the raw data level, they are interpreted as beliefs in the degree of correctness of each data source. In such a setting, DST combines the existing evidence and takes all available information from the registers into account to form one quality-indicator on the CDB level, denoted $q_{\odot,j}$ for each statistical unit n.

In a further step, the values in the CDB are compared to an external source (reapplying HD^E), to address possible quality issues in the process of combining the raw data to the CDB. The process of picking the values for the CDB has to be independent from the data generation. Otherwise the results of the quality assessment would be skewed. For this reason, the belief values can't be the starting point for picking values from the source registers. The evaluation yields a final quality indicator in the CDB $q_{\Psi,j}$. Table 7 shows the final quality measure on CDB level for LMS ($q_{\Psi,\text{LMS}}$), which is the weighted average of $q_{\odot,\text{LMS}}$ and $HD^E_{\text{CDB,LMS}}$ on CDB level. In this special example $q_{\Psi,\text{LMS}}$ is 0.728. Hence, $HD^E_{\text{CDB,LMS}}$ slightly increases the quality indicator compared to the sole combination of raw quality indicators.

Table 7: The quality for the *LMS* on CDB level

	$q_{\odot,LMS}$	$HD^E_{\text{CDB,LMS}}$	$q_{\Psi,LMS}$
LMS	0.721	0.973	0.728

3.3. The Final Data Pool (FDP)

In the last step of the data generation, values, which are missing in the CDB, are imputed. For the assessment of the data quality of these imputations, the Hyperdimension HD^I is applied. The distinction of imputation methods is a crucial factor for this task (Kausl 2012). As an example, the Austrian census uses different methods, such as deterministic editing, hot-deck techniques and logistic regressions. However, the principle for the evaluation of the imputations is the same for all methods. It is based on both the quality of the inputs and the quality of the imputation model. The quality of the inputs is assessed as a weighted average of the quality of the input variables k, which have already been calculated in the CDB stage. The accuracy of the imputation models m is consistently assessed by using classification rates (Φ), as shown in equation 5, where $q_{\Omega,k}$ is the quality of a specific input variable k and Φ^m_j is the classification rate for a certain imputation method m applied to attribute j.

$$HD^{I,m}_j = \frac{1}{K}\sum_{k=1}^{K} q_{\Omega,k} \cdot \Phi^m_j \tag{5}$$

The classification rate resembles the number of correct imputed values if the model is applied to existing data.[4] Using both the input variables for the imputation process and the classi-

[3] A detailed description of the quality assessment for the three types of attributes in the CDB is given by Berka *et al.* (2010, 2012).

[4] For ordinal variables the distance between the true value and the estimated value is taken into account. For numerical variables, the accuracy of the model is simply the correlation coefficient between the true and the imputed values.

fication rate, the quality of the imputations is derived as the observation-weighted average of both indicators. For a detailed explanation of the quality evaluation of different imputation techniques in this framework, see Schnetzer, Astleithner, Ćetković, Humer, Lenk, Moser, Schwerer, and Rechta (2015).

Table 8 shows the change in the average quality level from the CDB to the FDP stage. The average quality in the CDB is 0.728 ($q_{\Psi,\text{LMS}}$). The missing records on CDB level, which were assumed to have a quality of zero, are imputed on the FDP level and obtain a quality measure as described above. The average of the imputation quality HD^I for the *LMS* is 0.956.[5] Since missing values now have a quality indicator larger than zero, the average quality of the *LMS* has increased in the FDP ($q_{\Omega,\text{LMS}} = 0.949$) vis-a-vis the CDB ($q_{\Psi,\text{LMS}} = 0.728$).

Table 8: Quality monitoring for *LMS* from CDB to FDP level

	$q_{\Psi,\text{LMS}}$	HD^I_{LMS}	$q_{\Omega,\text{LMS}}$
LMS	0.728	0.956	0.949

4. Results of the quality assessment for the Austrian Census 2011

The quality framework, which we outlined above, is used for the quality assessment of the whole Austrian register-based census in 2011. To provide an overall picture for the quality assessment, we present the quality measures for selected attributes on CDB and FDP levels.

Table 9 shows, that while all attributes listed here have received a combined quality indicator on the CDB level (column 2), only multiple attributes and those derived on the CDB level have received an additional HD^E check (column 3). For unique attributes, such as Educational Attainment, this is not necessary, since the CDB values are based directly on the ones available in the single source register. The final CDB indicators in column 4 then either represent the combined CDB value from column 2 (for unique attributes) or the weighted average between 2 and 3 (for multiple attributes and attributes which are derived on CDB-level).

The next column (5) gives an overview of the average quality of imputations. However, not all of the attributes had to be imputed. Only those attributes with imputations received a quality indicator HDI at this stage. Column 7 lists the number of imputations, which can be interpreted as the contribution of HD^I to the final FDP quality indicator in column 8. Attributes derived from imputed data sets (only available in the FDP) are consistently compared to an external source at this stage (column 6).

In the following we exemplarily illustrate some of the main findings for different kinds of attributes. For example, the quality of the multiple attribute *Age*, has been compared to an external source HD^E on CDB level, which in general confirmed the combined raw quality indicator (0.997 versus 0.999). Additionally records have been imputed on FDP level. The quality of these imputations is notably lower than the available data from registers (0.731 vs 0.998), but still lead to a marginal improvement, since the few missing values were considered with the default quality of 0 until now.

The highest data quality, according to our framework, can be found for *Sex*, which is available in a large number of high-quality registers. Additionally, there are no imputations necessary, so that the data quality on FDP level is the same as for the CDB. An example of a unique attribute is *Educational Attainment*. By definition, the quality for such attributes in the source registers is equal to that in the CDB. However, a substantial number of values (293,698) need to be imputed on the FDP level. While the quality of imputed values (0.595) is lower than that of the CDB (0.791), the imputations still lead to an increase in the final quality measure (0.815) because of the zero valuation of missings until this stage.

[5] For further information see the documentation of data quality for the Austrian census and Schnetzer *et al.* (2015)

An extremely simple case is the unique attribute *Field of Educational Attainment*, which has required no imputations, resulting in an equal quality for the raw register level, the CDB and the FDP of 0.819. To further highlight a special example, the *Family Status* is a derived attribute, for which the derivations have been conducted on FDP level. This reflects the fact, that it is constructed from already imputed datasets, only available in the FDP. Therefore, the comparison to an external source is consistently carried out on the FDP-level, as opposed to e.g. *Current Activity Status*, for which this process can be carried out in the CDB.

Table 9: Results for the Austrian census on FDP-level

Attribute	$q_{\odot}$	HD^E	q_{Ψ}	HD^I	HD^E	No. Imp.[1]	q_{Ω}
Age	0.999	0.997	0.998	0.731	.	415	0.998
Sex	1.000	0.999	0.999	.	.	.	0.999
Country of birth	0.987	0.982	0.986	.	.	.	0.986
Country of citizenship	0.985	0.995	0.988	.	.	.	0.988
Legal marital status	0.721	0.973	0.728	0.956	.	1,929,346	0.949
Educational attainment	0.791	.	0.791	0.595	.	293,698	0.815
Field of educational attainment	0.819	.	0.819	.	.	.	0.819
Family status	0.820	.	0.820	0.799	0.999	923,910	0.948
Current activity status	0.909	0.923	0.913	.	.	.	0.913
Status in employment	0.930	0.955	0.930	.	.	.	0.930
Occupation	0.535	0.492	0.535	0.645	.	1,036,480	0.699
Full/part time employment	0.681	.	0.681	0.788	.	133,421	0.707

[1] Number of Imputations. Imputation methods include deterministic editing, derived from imputed FDP attributes as well.

5. Conclusion and outlook

This contribution applies a comprehensive quality-framework that allows for the assessment of quality of administrative data to the Austrian census of 2011. The main aim of the framework is to implement an objective procedure for the evaluation of different kinds of administrative data.

The framework is based on a modular design. This means that it can be used for a variety of purposes. We generate quality indicators for raw data, their combination as well as the combined and imputed data sets, so that the evaluation tracks the whole data generation process. The single modules are connected through user-defined weights (e.g. weighting of hyperdimensions), that allow for a flexible application in accordance with the special needs of a certain application.

As a result, we are able to compare data quality between different processing stages, data revisions, registers and even single attributes. Furthermore, all these indicators are calculated on the level of individual observations at every stage, so that quality-related information can also be tracked for data subsets.

A major asset of this approach is, that the calculated values can also be used to assess the uncertainty of calculations that make use of administrative data. This possibility is currently ongoing research and should be addressed in more detail in future discussions.

From the perspective of NSIs, the framework is especially valuable to monitor data quality over time and assess how it changes as a reaction to e.g. the introduction of new methods or processing tools. For the special application to a register-based census, which has been highlighted in this contribution, the framework allows the monitoring of different data revisions and subsequent censuses to provide detailed quality information to data users and the NSI itself.

The final application of the framework to the Austrian census of 2011 highlights how the quality of single attributes depends on the data authority, the number of available comparison registers and the processes at the NSI (e.g. imputation strategies).

Future research opportunities should be concerned with the applicability of the framework to other areas of interest and the enhancement of the framework to cover even more aspects where possible.

References

Berka C, Humer S, Lenk M, Moser M, Rechta H, Schwerer E (2010). "A Quality Framework for Statistics based on Administrative Data Sources using the Example of the Austrian Census 2011." *Austrian Journal of Statistics*, **Volume 39, Number 4**, 299–308.

Berka C, Humer S, Lenk M, Moser M, Rechta H, Schwerer E (2012). "Combination of Evidence from Multiple Administrative Data Sources: Quality Assessment of the Austrian Register-Based Census 2011." *Statistica Neerlandica*, **Volume 66, Issue 1**, 18–33.

Daas P, Fonville T (2007). "Quality Control of Dutch Administrative Registers: An Inventory of Quality Aspects." *Technical report*, Statistics Netherlands.

Daas P, Ossen S, Vis-Visschers R, Arends-Tóth J (2009). "Checklist for the Quality Evaluation of Administrative Data Sources." *Statistics Netherlands Discussion Paper*, **09042**.

Kausl A (2012). "The Data Imputation Process of the Austrian Register-Based Census." In *Conference Contribution at UNECE Work Session on Statistical Data Editing*. Oslo, Norway.

Lenk M (2008). "Methods of the Register-based Census in Austria." *Technical report*, Statistik Austria, Wien.

Schnetzer M, Astleithner F, Ćetković P, Humer S, Lenk M, Moser M, Schwerer E, Rechta H (2015). "Quality Measurement in Administrative Statistics and the Assessment of Imputations." *Journal of Official Statistics*, **31**(2), 1–17. doi:10.1515/JOS-2015-0015.

Shafer G (1992). "Dempster-Shafer Theory." In SC Shapiro (ed.), *Encyclopedia of Artificial Intelligence*, pp. 330–331. Wiley.

UNECE (2014). "Measuring Population and Housing – Practices of UNECE Countries in the 2010 Round of Censuses." *Technical report*, United Nations Economic Commission for Europe.

Affiliation:

Eva-Maria Asamer
Unit Register–based Census
Statistics Austria
Guglgasse 13
A-1110 Vienna, Austria
E-mail: eva-maria.asamer@statistik.gv.at
URL: http://www.statistik.at

Franz Astleithner
University of Vienna
Institute for Sociology
Rooseveltplatz 2
A-1090 Vienna, Austria
E-mail: franz.astleithner@univie.ac.at
URL: http://www.univie.ac.at

Mathias Moser
WU Vienna
Institute for Macroeconomics
Welthandelsplatz 1
A-1020 Vienna, Austria
E-mail: matmoser@wu.ac.at
URL: http://www.wu.ac.at

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 45
March 2016

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2015-02-10
Accepted: 2015-08-24

Austrian Journal of Statistics
March 2016, Volume 45, 15–33.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i2.92

A Study of Convolution Models for Background Correction of BeadArrays

Rohmatul Fajriyah
Graz University of Technology and Universitas Islam Indonesia

Abstract

The robust multi-array average (RMA), since its introduction in Irizarry, Bolstad, Collin, Cope, Hobbs, and Speed (2003a); Irizarry, Hobbs, Collin, Beazer-Barclay, Antonellis, Scherf, and Speed (2003b); Irizarry, Wu, and Jaffee (2006), has gained popularity among bioinformaticians. It has evolved from the exponential-normal convolution to the gamma-normal convolution, from single to two channels and from the Affymetrix to the Illumina platform.

The Illumina design provides two probe types: the regular and the control probes. This design is very suitable for studying the probability distribution of both and one can apply a convolution model to compute the true intensity estimator.

In this paper, we study the existing convolution models for background correction of Illumina BeadArrays in the literature and give a new estimator for the true intensity, assuming that the intensity value is exponentially or gamma distributed and the noise has lognormal distribution.

Our study shows that one of our proposed models, the gamma-lognormal with the method of moments for parameters estimation, is the optimal one for the benchmarking data set with benchmarking criteria, while the gamma-normal model has the best performance for the benchmarking data set with simulation criteria.

For the publicly available data sets, the gamma-normal and the exponential-gamma models with maximum likelihood estimation method can not be used and our proposed models exponential-lognormal and gamma-lognormal have the best performance, showing a moderate error in background correction and in the parametrization.

Keywords: convolution, background correction, BeadArrays.

1. Introduction

There are various processes in producing data from microarray experiments and each process contributes noise to the data. The noise can be of two types, biological and non-biological. Non-biological noise should be avoided or at least minimized.

Sources for the non-biological noise are, for example, the chip itself, the scanner, or fluctuations in the electric network. Therefore, the data needs to be adjusted. The pre-processing will adjust the intensity value (Huber, Irizarry, and Gentleman 2005a; Huber, von Heydebreck, and Vingron 2005b) and provides an estimate of the true intensity.

To estimate the intensity value, researchers proposed additive and multiplicative models and also *additive-muliplicative error models*, see e.g. Huber, von Heydebreck, and Vingron (2004). In case of additive models, the underlying distribution is generally chosen as normal (log-normal), exponential, or a Gamma-t mixture in the parametric approach (Allen, Chen, and Xie (2009), Bolstad, Irizarry, Astrand, and Speed (2003), Chen, Xie, and Story (2011), Hochreiter, Djork-Arné, and Obermayer (2006), Irizarry *et al.* (2003a,b, 2006), and Plancade, Rozenholc, and Lund (2011, 2012)).

Irizarry *et al.* (2003a,b, 2006) and Bolstad *et al.* (2003), on the Affymetrix platform, have estimated the true intensity values based on a convolution model in the background correction step of their robust multi-array average (RMA) pre-processing method. They assumed that the true intensity is exponentially distributed and the background noise is normally distributed.

Plancade *et al.* (2011, 2012) showed that the RMA model (in Bolstad *et al.* (2003) and Irizarry *et al.* (2003a,b, 2006)) does not fit Illumina BeadArrays: using the exponential-normal convolution leads to a large distance between the observed and the modeled intensities. They proposed, instead, the implementation of a gamma distribution for the intensity value and normal distribution for the noise.

The simulation study of Plancade *et al.* (2011, 2012) showed that the gamma-normal model performs better than the existing exponential-normal convolution model, giving a more accurate and correct fit for the observed intensities in Illumina BeadArray.

Using a Gamma distribution for the intensity values in Illumina BeadArrays has been first suggested by Xie, Wang, and Story (2009).

The studies of Baek, Son, and MacLachlan (2007) (on the background correction of the image processing) and Chen *et al.* (2011) show that the noise distribution is usually skewed in different degrees. In their studies, based on simulated and real data sets, Baek *et al.* (2007) conclude that the gamma distribution is well suited for the noise. It accounts for the intensities with a positive lower bound and is very flexible in its shape, including asymmetric exponential type and symmetric normal type.

The proposed convolution of exponential-gamma distribution by Chen *et al.* (2011) improves the intensity estimation and the detection of differentially expressed genes in the case when the intensity to noise ratio is large and the noise has a skewed distribution.

In view of the remarks above, it is natural to model both the true intensity and the background noise in Illumina BeadArrays as gamma distributed. In an earlier version of this paper we have developed an estimator for the true intensity based on the gamma-gamma convolution model of RMA. However, this model does not fit very well the Illumina benchmarking data set. Independently, Triche, Weisenberger, Berg, Laird, and Siegmund (2013) proposed and applied the gamma-gamma model to pre-process Illumina methylation arrays.

In this paper we introduce a new model for background correction in Illumina BeadArrays where the true intensity value is exponentially or gamma distributed and the noise has a lognormal distribution. As we will see, this model avoids the difficulties with the gamma-gamma model and has an overall satisfactory performance.

We note that a new method reducing the bias of the maximum likelihood estimator of the shape parameter of the gamma distribution was proposed by Zhang (2013). But since our samples are very large, bias is not a problem in our studies.

We compare the performance of the models on the Illumina spike-in data set, based on various criteria: root and mean square error (RMSE), L_1 error, Kullback-Leibler (K-L) coefficient, and some adapted criteria from Affycomp (Cope, Irizarry, Jaffee, Wu, and Speed 2004). These criteria are measuring the reproducibility, accuracy, precision, specificity, and sensitivity of the expression measure of each model. We then provide a simulation study to measure the consistency of the error of background correction and the parametrization. The description and some details on these criteria and simulation can be found at `http://rfajriyah.staff.`

uii.ac.id/category/supplementary/

Our paper is organized as follows. In Section 2 we review previous work related to the background correction for Illumina BeadArrays. Our proposed models are described in Sections 3.1 and 3.2. Section 3.3 explains the benchmarking and the simulation studies on the benchmarking data set and Section 3.4 compares the performance of all models in the public data sets. Finally, Section 4 states the conclusions and indications of future work.

2. Previous work

Affymetrix is the pioneer and most widely used platform for microarray gene expression experiments. The tools and algorithms to handle the data are numerous, both free and commercial. Some methods for pre-processing are available. Examples for the background correction step are: MAS5.0 by Affymetrix, multiplicative model based expression index (MMBE) by Li and Wong (2001), RMA in Irizarry *et al.* (2003a,b, 2006) and Bolstad *et al.* (2003), GC-RMA by Wu, Irizarry, Gentleman, Martinez-Murillo, and Spencer (2004) and maximum likelihood estimation based on the normal-exponential convolution model by Silver, Ritchie, and Smyth (2009).

Illumina is one of the alternative platforms and is increasingly popular. A few statistical methods have been developed for BeadArray data and there is no consensus yet for the pre-processing steps (Shi, Oshlack, and Smyth 2010). Xie *et al.* (2009) mention that for the background correction step, Illumina bead studio gives two options (no background correction and background substraction) and the packages for BeadArrays in R provide three options (no background correction, background substraction and RMA background correction).

Ding, Xie, Park, Xiao, and Story (2008) extended the RMA model by proposing the model-based background correction method (MBCB) and showed that their model leads to a more precise determination of the gene expression and a better biological interpretation of Illumina BeadArray data.

The studies of Chen *et al.* (2011) and Plancade *et al.* (2011, 2012) show that their background correction models are made by adapting the RMA Affymetrix model. As Forcheh, Verbeke, Kasim, Lin, Shkedy, Talloen, Gohlmann, and Clement (2012), pointed out, most preprocessing methods for Illumina BeadArrays are taken from the Affymetrix microarray platform.

In general, the background correction is applied toward each array, where in each array there are probes, probesets and genes (terminology for the Affymetrix platform) or bead and bead-type level probes (terminology for the Illumina platform).

At the Illumina platform, each gene is only targeted by one bead-type, which has been represented by about 30 time replications. If we can have a raw benchmarking data set, then it is possible to have all bead-type level probes of the raw data intensities.

The current publicly available benchmarking data set for the Illumina platform is the raw data from the bead studio, which is the average of the bead-type level probes, not background corrected and of unnormalized intensity. Therefore, the background correction in this paper is applied to the gene intensity in each array.

Suppose we have J arrays and for each array there are I regular genes. Our convolution model is as follows:

$$P_{ij} = S_{ij} + B_{ij}\,, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J\,, \tag{1}$$

where P_{ij}, S_{ij}, and B_{ij} are the observed signal, true signal, and noise intensity, respectively. The S_{ij} are i.i.d. and so are the B_{ij}; the S_{ij} are independent of the B_{ij}. The P_{ij} are observable. Our task is to recover the unknown signals S_{ij} from the P_{ij}. To do this, for each array j we also have M observable noise intensities B_{0mj}, $m = 1, \ldots, M$; the B_{0mj} are i.i.d. with the same distribution as the B_{ij} and are independent of all of the S_{ij}, B_{ij}. The S_{ij} and B_{ij} have a known type of distribution (exponential, gamma, normal, lognormal) with unknown parameters.

2.1. Background correction by RMA

The RMA method was developed for the Affymetrix platform, where the design of arrays is different from the Illumina one. In this platform, one has perfect match and mismatch probe design. The RMA uses only the perfect match probe, which is the targeted probe of the intended gene. For those not familiar with the Affymetrix platform, refer to Bolstad (2004), Bolstad *et al.* (2003), and Irizarry *et al.* (2003a,b, 2006) for further information.

In modeling the intensity values, the RMA model (Bolstad *et al.* (2003) and Irizarry *et al.* (2003a,b, 2006)) assumes that the intensity values are affected by the noise of the chip. By referring to Equation (1), in the RMA model P_{ij} is the observed bead-type level probe intensity, S_{ij} is the true signal with $S_{ij} \sim f_1(s_{ij};\theta_j) = \text{Exp}(\theta_j)$, $\theta_j > 0$, and B_{ij} is the background noise of the chip with $B_{ij} \sim f_2(b_{ij};\mu_j,\sigma_j^2) = \mathcal{N}(\mu_j,\sigma_j^2)$, $\mu_j \in \mathbb{R}$, $\sigma_j^2 > 0$. To avoid negative intensity values, we truncate B_{ij} at 0 from below, i.e. we replace B_{ij} by $\max\{B_{ij}, 0\}$; this will not change its density function $f_2(b_{ij};\mu_j,\sigma_j^2)$ for $b_{ij} > 0$.

Assuming independence, the joint density of the two-dimensional random variable (S_{ij}, B_{ij}) is

$$f_{S_{ij},B_{ij}}(s_{ij},b_{ij};\mu_j,\sigma_j^2,\theta_j) = \theta_j \exp\Big\{-\theta_j s_{ij}\Big\} f_2(b_{ij};\mu_j,\sigma_j^2), \qquad b_{ij}, s_{ij} > 0\,.$$

Furthermore, the transformation formula for two-dimensional densities gives that joint density of S and P is

$$f_{S_{ij},P_{ij}}(s_{ij},p_{ij};\mu_j,\sigma_j^2,\theta_j) = \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_{ij}-\mu_j)\right\} f_2(s_{ij};\mu_{S.P,j},\sigma_j^2), \qquad 0 < s_{ij} < p_{ij}, \quad (2)$$

where $\mu_{S.P,j} = p_{ij} - \mu_j - \theta_j\sigma_j^2$. From (2) we get the marginal density of P_{ij} and the conditional density of S_{ij} given P_{ij} as

$$f_{P_{ij}}(p_{ij};\theta_j,\mu_j,\sigma_j^2) = \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_{ij}-\mu_j)\right\}\left(\Phi\left(\frac{\mu_{S.P,j}}{\sigma_j}\right) + \Phi\left(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j}\right) - 1\right)$$

$$f_{S_{ij}|P_{ij}}(s_{ij}|p_{ij};\mu_{S.P,j},\sigma_j^2) = \frac{f_2(s_{ij};\mu_{S.P,j},\sigma_j^2)}{\Phi\left(\frac{\mu_{S.P,j}}{\sigma_j}\right) + \Phi\left(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j}\right) - 1}.$$

The background adjusted intensity is computed as the conditional expectation of the true signal given the observed intensity, i.e.

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \frac{1}{\Phi\left(\frac{\mu_{S.P,j}}{\sigma_j}\right) + \Phi\left(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j}\right) - 1}\int_0^{p_{ij}} s_{ij} f_2(s_{ij};\mu_{S.P,j},\sigma_j^2)ds_{ij}.$$

The substitution $s_{ij} = \mu_{S.P,j} + \sigma_j t$ yields

$$\int_0^{p_{ij}} s_{ij} f_2(s_{ij};\mu_{S.P,j},\sigma_j^2)ds_{ij} = \mu_{S.P,j}\left(\Phi\left(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j}\right) + \Phi\left(\frac{\mu_{S.P,j}}{\sigma_j}\right) - 1\right) + \sigma_j\left(\phi\left(\frac{\mu_{S.P,j}}{\sigma_j}\right) - \phi\left(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j}\right)\right)$$

and thus (see Bolstad *et al.* (2003))

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \mu_{S.P,j} + \sigma_j\frac{\phi(\frac{\mu_{S.P,j}}{\sigma_j}) - \phi(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j})}{\Phi(\frac{\mu_{S.P,j}}{\sigma_j}) + \Phi(\frac{p-\mu_{S.P,j}}{\sigma_j}) - 1}. \quad (3)$$

Note that modelling the noise as a truncated normal variable has the consequence that the noise equals 0 with a positive probability p_0, a rather unpleasant feature of the model. As

pointed out in Xie *et al.* (2009), however, in practical cases p_0 is rather small, so this problem can be disregarded. To avoid this difficulty, one can model the noise as the absolute value of a $\mathcal{N}(\mu, \sigma^2)$ variable, which changes the calculations above. However, since in this paper we will provide a background correction model fitting the reality considerably better, we do not give the details here.

2.2. Exponential-normal MBCB

Xie *et al.* (2009) use the same underlying distributions in (1) for background correction. The difference with the RMA (Bolstad *et al.* (2003) and Irizarry *et al.* (2003a,b, 2006)) are

1. Xie *et al.* (2009) use $+\infty$ as the upper bound of the integral to compute the marginal density function and the conditional expectation of the true intensity value. On the other hand, RMA uses p as the upper bound of the integration.

   The background corrected intensity value of Xie *et al.* (2009) is

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \mu_{S.P,j} + \sigma_j \frac{\phi(\frac{\mu_{S.P,j}}{\sigma_j})}{\Phi(\frac{\mu_{S.P,j}}{\sigma_j})}.$$

2. Under the convolution model (1), where the true intensity value is assumed exponentially distributed and the noise is normally distributed, we need to estimate the parameters θ_j, μ_j, and σ_j^2. Xie *et al.* (2009) offer three estimation methods: the method of moments, maximum likelihood estimation, and a Bayesian approach. On the other hand, RMA applies the *ad-hoc* method.

Ding *et al.* (2008) use the exponential-normal convolution model to correct the background of the Illumina platform by using Markov chain Monte Carlo simulation.

2.3. Gamma-normal convolution

Plancade *et al.* (2011, 2012) introduced gamma-normal convolution to model the background correction of Illumina BeadArray. The model is based on the RMA background correction of Affymetrix GeneChip. Plancade *et al.* (2011, 2012) assume that the intensity value is gamma distributed and the noise is normally distributed.

Under model (1), $f_{P_{ij}}$ is the convolution of $f_{S_{ij}}$ and $f_{B_{ij}}$. The background corrected intensity is computed as the conditional expectation of S_{ij} given $P_{ij} = p_{ij}$, i.e.

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \frac{\int s_{ij} f^{\text{gam}}_{\alpha_j,\beta_j}(s_{ij}) f^{\text{norm}}_{\mu_j,\sigma_j}(p_{ij} - s_{ij}) ds_{ij}}{\int f^{\text{gam}}_{\alpha_j,\beta_j}(s_{ij}) f^{\text{norm}}_{\mu_j,\sigma_j}(p_{ij} - s_{ij}) ds_{ij}}, \tag{4}$$

where

$$f^{\text{gam}}_{\alpha_j,\beta_j}(s) = \frac{\beta_j^{\alpha_j} s^{\alpha_j - 1} \exp(-\beta_j s)}{\Gamma(\alpha_j)}, \qquad \alpha_j, \beta_j, s > 0$$

is the gamma density. When S_{ij} is gamma distributed and B_{ij} is normally distributed, then (4) has no analytic expression like (3). Plancade *et al.* (2011, 2012) implemented the Fast Fourier Transform to estimate the parameters and to correct the background. For the background correction with Fast Fourier Transform, the corrected intensity (4) is rewritten as

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \frac{\alpha_j \beta_j \int f^{\text{gam}}_{\alpha_j+1,\beta_j}(s_{ij}) f^{\text{norm}}_{\mu_j,\sigma_j}(p_{ij} - s_{ij}) ds_{ij}}{\int f^{\text{gam}}_{\alpha_j,\beta_j}(s_{ij}) f^{\text{norm}}_{\mu_j,\sigma_j}(p_{ij} - s_{ij}) ds_{ij}},$$

since $s_{ij} f^{\text{gam}}_{\alpha_j,\beta_j}(s_{ij}) = \alpha_j \beta_j f^{\text{gam}}_{\alpha_j+1,\beta_j}(s_{ij})$ is valid for every $s_{ij} > 0$.

2.4. Exponential-gamma convolution

Under model (1), Chen *et al.* (2011) proposed for the distribution of the true intensity and its noise the exponential and gamma distribution, respectively. Therefore, $S_{ij} \sim f_1(s_{ij};\theta_j) = \text{Exp}(\theta_j)$ and $B_{ij} \sim f_2(b_{ij};\alpha_j,\beta_j) = \text{Gamma}(\alpha_j,\beta_j)$, where $s_{ij}, b_{ij}, \theta_j, \alpha_j, \beta_j > 0$.

The corrected background intensity of Chen *et al.* (2011) is

$$\mathsf{E}(S_{ij}|P_{ij}=p_{ij}) = p_{ij} - \frac{\int_0^{p_{ij}} b_{ij}^{\alpha_j} \exp(-(\beta_j-\theta_j)b_{ij})db_{ij}}{\int_0^{p_{ij}} b_{ij}^{\alpha_j-1} \exp(-(\beta_j-\theta_j)b_{ij})db_{ij}}.$$

3. Results

Now we present the results of the two proposed convolution models: the exponential-lognormal convolution in Section 3.1 and the gamma-lognormal convolution in Section 3.2. In each section, the formula for the background corrected intensity value is derived and methods to estimate the parameters are explained. Section 3.3 present the benchmarking results, i.e. a performance comparison of all models at the Illumina spike-in data set. Section 3.4 present the performance comparison of all models for the public data sets. The simulation study results are presented in Sections 3.3 and 3.4.

For further details on the benchmarking criteria, supplemental plots and simulations see Supplementary_Materials at http://rfajriyah.staff.uii.ac.id/category/supplementary/.

3.1. Exponential-lognormal convolution

Background correction Consider model (1) when the true intensity S_{ij} is exponentially distributed, i.e.

$$S_{ij} \sim f_1(s_{ij};\theta_j) = \theta_j \exp(-\theta_j s_{ij}), \qquad \theta_j, s_{ij} > 0,$$

and the background noise B_{ij} is lognormally distributed, i.e.

$$B_{ij} \sim f_2(b_{ij};\mu_j,\sigma_j^2) = \frac{1}{b_{ij}\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log b_{ij}-\mu_j)^2}{2\sigma_j^2}\right), \qquad \mu_j \in \mathbb{R},\ \sigma_j^2, b_{ij} > 0.$$

Then the joint density function of S_{ij} and B_{ij} equals

$$f_{S_{ij},B_{ij}}(s_{ij},b_{ij}) = \frac{\theta_j \exp(-\theta_j s_{ij})}{b_{ij}\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log b_{ij}-\mu_j)^2}{2\sigma_j^2}\right),$$

and thus the joint density function of S_{ij} and P_{ij} is

$$f_{S_{ij},P_{ij}}(s_{ij},p_{ij}) = \frac{\theta_j \exp(-\theta_j s_{ij})}{(p_{ij}-s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij}-s_{ij})-\mu_j)^2}{2\sigma_j^2}\right), \qquad s_{ij} < p_{ij}.$$

Consequently, the marginal density function of P_{ij} equals

$$f_{P_{ij}}(p_{ij}) = \int_0^{p_{ij}} \frac{\theta_j \exp(-\theta_j s_{ij})}{(p_{ij}-s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij}-s_{ij})-\mu_j)^2}{2\sigma_j^2}\right) ds_{ij}.$$

Using the substitution $\log(p_{ij} - s_{ij}) = z_{ij}$ we get

$$\begin{aligned}
f_{P_{ij}}(p_{ij}) &= \int_{-\infty}^{\log p_{ij}} \frac{\theta_j \exp(-\theta_j(p_{ij} - e^{z_{ij}}))}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij}-\mu_j)^2}{2\sigma_j^2}\right) dz_{ij} \\
&= \frac{\theta_j \exp(-\theta_j p_{ij})}{\sigma_j\sqrt{2\pi}} \int_{-\infty}^{\log p_{ij}} \exp\left(-\frac{(z_{ij}-\mu_j)^2}{2\sigma_j^2}\right) \sum_{k=0}^{\infty} \frac{\theta_j^k e^{kz_{ij}}}{k!} dz_{ij} \\
&= \frac{\theta_j \exp(-\theta_j p_{ij})}{\sigma_j\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \int_{-\infty}^{\log p_{ij}} \exp\left(-\frac{(z_{ij}-\mu_j)^2}{2\sigma_j^2} + kz_{ij}\right) dz_{ij} \\
&= \frac{\theta_j \exp(-\theta_j p_{ij})}{\sigma_j\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \exp\left(k\left(\mu_j + \frac{k}{2}\sigma_j^2\right)\right) \int_{-\infty}^{\log p_{ij}} \exp\left(-\frac{(z_{ij}-(\mu_j + k\sigma_j^2))^2}{2\sigma_j^2}\right) dz_{ij} \\
&= \theta_j \exp(-\theta_j p_{ij}) C_{a,j},
\end{aligned}$$

where

$$C_{a,j} = \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \exp\left(k\left(\mu_j + \frac{k}{2}\sigma_j^2\right)\right) \Phi\left(\frac{\log p_{ij} - (\mu_j + k\sigma_j^2)}{\sigma_j}\right).$$

The conditional density function of S_{ij} given $P_{ij} = p_{ij}$ is now obtained as

$$f_{S_{ij}|P_{ij}}(s_{ij}|p_{ij}) = \frac{\exp(\theta_j(p_{ij} - s_{ij}))}{(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}C_{a,j}} \exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right)$$

with conditional mean

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \frac{\exp(\theta_j p_{ij})}{C_{a,j}} \int_0^{p_{ij}} \frac{s_{ij}\exp(-\theta_j s_{ij})}{(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right) ds_{ij}.$$

Using the substitution $\log(p_{ij} - s_{ij}) = z_{ij}$, this equals

$$\begin{aligned}
\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) &= \frac{p_{ij}}{C_{a,j}} \int_{-\infty}^{\log p_{ij}} \frac{\left(1 - \frac{e^{z_{ij}}}{p_{ij}}\right)\exp(\theta_j e^{z_{ij}})}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij}-\mu_j)^2}{2\sigma_j^2}\right) dz_{ij} \\
&= \frac{p_{ij}}{C_{a,j}} \left[C_{a,j} - \int_{-\infty}^{\log p_{ij}} \frac{\frac{e^{z_{ij}}}{p_{ij}}\exp(\theta_j e^{z_{ij}})}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij}-\mu_j)^2}{2\sigma_j^2}\right) dz_{ij}\right] \\
&= p_{ij} - \frac{\exp\left(\mu_j + \frac{\sigma_j^2}{2}\right)}{C_{a,j}} \int_{-\infty}^{\log p_{ij}} \frac{\exp(\theta_j e^{z_{ij}})}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij}-(\mu_j+\sigma_j^2))^2}{2\sigma_j^2}\right) dz_{ij} \\
&= p_{ij} - \frac{\exp\left(\mu_j + \frac{\sigma_j^2}{2}\right)}{C_{a,j}} \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \exp\left(k\left(\mu_j + \frac{k+2}{2}\sigma_j^2\right)\right) \\
&\qquad \times \int_{-\infty}^{\log p_{ij}} \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij}-(\mu_j+(k+1)\sigma_j^2))^2}{2\sigma_j^2}\right) dz_{ij} \\
&= p_{ij} - \frac{C_{b,j}}{C_{a,j}} \exp\left(\mu_j + \frac{\sigma_j^2}{2}\right),
\end{aligned}$$

where

$$C_{b,j} = \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \exp\left(k\left(\mu_j + \frac{k+2}{2}\sigma_j^2\right)\right) \Phi\left(\frac{\log p_{ij} - (\mu_j + (k+1)\sigma_j^2)}{\sigma_j}\right).$$

Parameter estimation To estimate the parameters θ_j, μ_j, and σ_j^2, $j = 1, \ldots, J$ in the exponential-lognormal model we can use various methods.

1. Maximum likelihood estimation (MLE):
   This is implemented by applying the *optim* function in R to maximize the log-likelihood function of the jth array

$$\sum_{i=1}^{I}(\log\theta_j - \theta_j p_{ij} + \log C_{a,j;K}) + \sum_{m=1}^{M}\left(-\frac{(\log b_{0mj} - \mu_j)^2}{2\sigma_j^2} - \frac{1}{2}\log(2\pi\sigma_j^2 b_{0mj}^2)\right),$$

   where p_{ij} and b_{0mj} are the observed values of P_{ij} and B_{0mj}. Note that $C_{a,j}$ in the log-likelihood function is defined by the infinite series at the end of the previous section. However, the terms of this infinite series decrease very rapidly and thus we can cut off the series at a proper index K giving $C_{a,j;K}$, making it suitable for its computation in R. The index K is chosen as the smallest integer for which $|(C_{a,j;K} - C_{a,j;K-1})/C_{a,j;K-1}| < 0.001$ holds.

2. Method of moments:
   Note that the method of moments estimator of the parameter θ in an exponential distribution is the reciprocal of the sample mean. Since the S_{ij} are not observable, but the B_{0mj} are, we consider Equation (1) and estimate $1/\theta_j$ by the difference $\text{mean}(p_{ij}) - \text{mean}(b_{0mj})$. Further, the parameters μ_j and σ_j^2 in the lognormal part are estimated by the sample mean and variance of the observed $\log b_{0mj}$ values.

3. Plug-in estimator:

   (a) We calculate the MLE of the parameter in the model of S_{ij} by utilizing the observed sample p_{ij} instead of s_{ij}. This is justified by the fact that, in some sense, the distribution of S_{ij} is similar to the distribution of P_{ij}.
   (b) We estimate μ_j and σ_j^2 through MLE based on B_{0mj} as described above.

3.2. Gamma-lognormal convolution

Background correction Consider now model (1) and assume that the true intensity S_{ij} is gamma distributed, i.e.

$$S_{ij} \sim f_1(s_{ij}; \alpha_j, \beta_j) = \frac{\beta_j^{\alpha_j} s_{ij}^{\alpha_j - 1}\exp(-s_{ij}\beta_j)}{\Gamma(\alpha_j)}, \qquad \alpha_j, \beta_j, s_{ij} > 0,$$

and that the background noise B_{ij} is lognormally distributed, i.e.

$$B_{ij} \sim f_2(b_{ij}; \mu_j, \sigma_j^2) = \frac{1}{b_{ij}\sigma_j\sqrt{2\pi}}\exp\left(-\frac{(\log b_{ij} - \mu_j)^2}{2\sigma_j^2}\right), \qquad \mu_j \in \mathbb{R},\ \sigma_j^2 > 0.$$

The joint density function of S_{ij} and B_{ij} is

$$f_{S_{ij},B_{ij}}(s_{ij}, b_{ij}) = \frac{\beta_j^{\alpha_j} s_{ij}^{\alpha_j - 1}\exp(-s_{ij}\beta_j)}{\Gamma(\alpha_j)} \frac{1}{b_{ij}\sigma_j\sqrt{2\pi}}\exp\left(-\frac{(\log b_{ij} - \mu_j)^2}{2\sigma_j^2}\right)$$

and thus the joint density function of S_{ij} and P_{ij} is

$$f_{S_{ij},P_{ij}}(s_{ij}, p_{ij}) = \frac{\beta_j^{\alpha_j} s_{ij}^{\alpha_j - 1}\exp(-s_{ij}\beta_j)}{\Gamma(\alpha_j)} \frac{1}{(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}}\exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right).$$

Hence, the marginal density function of P_{ij} is obtained as

$$f_{P_{ij}}(p_{ij}) = \int_0^{p_{ij}} \frac{\beta_j^{\alpha_j} s_{ij}^{\alpha_j - 1} \exp(-s_{ij}\beta_j)}{\Gamma(\alpha_j)} \frac{1}{(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right) ds_{ij}.$$

Using the substitution $\log(p_{ij} - s_{ij}) = z_{ij}$ we get

$$\begin{aligned} f_{P_{ij}}(p_{ij}) &= \int_{-\infty}^{\log p_{ij}} \frac{\beta_j^{\alpha_j} p_{ij}^{\alpha_j - 1} \left(1 - \frac{e^{z_{ij}}}{p_{ij}}\right)^{\alpha_j - 1} \exp(-p_{ij}\beta_j + e^{z_{ij}\beta_j})}{\Gamma(\alpha_j)\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij} - \mu_j)^2}{2\sigma_j^2}\right) dz_{ij} \\ &= \frac{\beta_j^{\alpha_j} p_{ij}^{\alpha_j - 1} \exp(-p_{ij}\beta_j)}{\Gamma(\alpha_j)\sigma_j\sqrt{2\pi}} \int_{-\infty}^{\log p_{ij}} \left(1 - \frac{e^{z_{ij}}}{p_{ij}}\right)^{\alpha_j - 1} \exp(e^{z_{ij}\beta_j}) \exp\left(-\frac{(z_{ij} - \mu_j)^2}{2\sigma_j^2}\right) dz_{ij} \\ &= \frac{\beta_j^{\alpha_j} p_{ij}^{\alpha_j - 1} \exp(-p_{ij}\beta_j)}{\Gamma(\alpha_j)} C_{c,j}, \end{aligned}$$

where

$$C_{c,j} = \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} \frac{(-1)^k \binom{\alpha_j - 1}{k}}{p_{ij}^k n!} \beta_j^n \exp\left((k+n)\left(\mu_j + (k+n)\frac{\sigma_j^2}{2}\right)\right) \Phi\left(\frac{\log p_{ij} - (\mu_j + (k+n)\sigma_j^2)}{\sigma_j}\right).$$

The conditional density function is now obtained as

$$f_{S_{ij}|P_{ij}}(s_{ij}|p_{ij}) = \frac{\exp((p_{ij} - s_{ij})\beta_j) s_{ij}^{\alpha_j - 1}}{C_{c,j} p_{ij}^{\alpha_j - 1}(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right)$$

with respective conditional mean

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \frac{e^{p_{ij}\beta_j}}{C_{c,j} p_{ij}^{\alpha_j - 1}} \int_0^{p_{ij}} \frac{s_{ij}^{\alpha_j} e^{-s_{ij}\beta_j}}{(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right) ds_{ij}.$$

Substituting $\log(p_{ij} - s_{ij}) = z_{ij}$ this becomes

$$\begin{aligned} \mathsf{E}(S_{ij}|P_{ij} = p_{ij}) &= \frac{p_{ij}}{C_{c,j}} \int_{-\infty}^{\log p_{ij}} \frac{\left(1 - \frac{e^{z_{ij}}}{p_{ij}}\right)^{\alpha_j} \exp\left(e^{z_{ij}\beta_j}\right)}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij} - \mu_j)^2}{2\sigma_j^2}\right) dz_{ij} \\ &= \frac{p_{ij} C_{d,j}}{C_{c,j}}, \end{aligned} \tag{5}$$

where

$$C_{d,j} = \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} \frac{(-1)^k \binom{\alpha_j}{k}}{p_{ij}^k n!} \beta_j^n \exp\left((k+n)\left(\mu_j + (k+n)\frac{\sigma_j^2}{2}\right)\right) \Phi\left(\frac{\log p_{ij} - (\mu_j + (k+n)\sigma_j^2)}{\sigma_j}\right).$$

Parameter estimation To estimate the parameters α_j, β_j, μ_j, and σ_j^2 in (5), we can use either of the following methods.

1. Maximum likelihood estimation:
   This is implemented by applying the *optim* function in R to maximize the log-likelihood function of the jth array

$$\begin{aligned} &\sum_{i=1}^{I} \left(\log(C_{c,j;K}) + (\alpha_j - 1)\log(p_{ij}) - p_{ij}\beta_j + \alpha_j \log(\beta_j) - \log(\Gamma(\alpha_j))\right) \\ &+ \sum_{m=1}^{M} \left(-\frac{(\log b_{0mj} - \mu_j)^2}{2\sigma_j^2} - \frac{1}{2}\log(2\pi\sigma_j^2 b_{0mj}^2)\right). \end{aligned}$$

   Similarly to the exponential-lognormal model, in the computation of $C_{c,j}$ the cutoff index K is chosen according to the criteria $|(C_{c,j;K} - C_{c,j;K-1})/C_{c,j;K-1}| < 0.002$.

2. Method of moments:
The implementation of this method for the jth array is done by recalling that in case of a gamma distribution with parameters α_j and β_j, the method of moments estimator for α_j/β_j and α_j/β_j^2 is the sample mean and the sample variance, respectively. Thus, considering Equation (1) we get

$$\hat{\beta}_j = \frac{\text{mean}(s_{ij})}{\text{var}(s_{ij})} = \frac{\text{mean}(p_{ij}) - \text{mean}(b_{0mj})}{\text{var}(p_{ij}) - \text{var}(b_{0mj})},$$

as also

$$\hat{\alpha}_j = \hat{\beta}_j \text{ mean}(s_{ij}) = \frac{(\text{mean}(s_{ij}))^2}{\text{var}(s_{ij})} = \frac{(\text{mean}(p_{ij}) - \text{mean}(b_{0mj}))^2}{\text{var}(p_{ij}) - \text{var}(b_{0mj})}.$$

Furthermore, μ_j and σ_j^2 are estimated by the mean and variance of $\log b_{0mj}$.

3. Plug-in estimator:
In equation (1), P_{ij} and B_{0mj} are observable intensities. Therefore, the plug-in estimator is implemented by

   (a) estimating α_j and β_j through MLE based on the p_{ij} values, and
   (b) estimating μ_j and σ_j^2 through MLE based on the b_{0mj} values.

3.3. Benchmarking

Benchmarking data set Illumina platform has provided a benchmarking data set, the Illumina spike-in Dunning, Barbosa-Morais, Lynch, Tavaré, and Ritchie (2008). These spike-in probes are targeting bacterial and viral genes absent from the mouse genome. These were added at specific concentrations on each sample. Therefore the change in expression level of a particular spike between samples is known a priori. The expression levels of the non-spikes should not change between samples.

There are twelve different concentrations of spike: 1000 picomolar (pM), 300, 100, 30, 10, 3, 1, 0.3, 0.1, 0.03, 0.01, and 0.00 pM. It was replicated four times. Therefore, there are 48 samples and each sample has regular and control bead-type level probes.

There are approximately about 48000 bead-type level probes for each sample and in addition the 33 spike-in bead-type level probes are added into it. For the control, there are 1616 bead-type level probes. These control experiments are the benchmarking data sets of Illumina and are used to compare low-level analysis methods such as in Affymetrix platform.

Performance studies We compare all convolution models: Irizarry *et al.* (2003a,b, 2006) and Bolstad *et al.* (2003): RMA (Exponential-Normal), Plancade *et al.* (2011, 2012): Gamma-Normal, Chen *et al.* (2011): Exponential-Gamma, Xie *et al.* (2009): Exponential-Normal adjusted for Illumina BeadArrays with MLE for the parameters, Bayesian approach and the method of moments, and the proposed models: exponential-lognormal and gamma-lognormal.

We will call the methods above, respectively, as follows: ENr, GN, EG, ENm, ENm, ENn, ELNn, ELNm, ELNp, GLNn, GLNm, and GLNp. We use the MBCB package (Allen *et al.* (2009) and Xie *et al.* (2009)) to adjust the intensity values of these existing models ENr, ENm, ENmc and ENn. Except that, the GN uses the NormalGamma package (Plancade *et al.* (2011)).

Table 1 shows that the GLNn reproduces the Illumina concentration better than others. The ENr shows the closest performance toward the GLNn. Note that the computation of the Kullback-Leibler coefficient is implemented in each array j based on the nominal concentrations O in Table 1 and the observed intensities P in Table 2, and the value in each

Table 1: Reproducibility of each method toward the Illumina spike-in concentration

Model	RMSE	K-L
ENr	1.346	51310
ENn	1.407	41010
ENm	1.483	23170
ENmc	1.483	23170
EG	1.470	20660
GN	1.521	58480
ELNn	1.411	41200
ELNm	1.489	21280
ELNp	1.423	37800
GLNn	**1.323**	**4333**
GLNm	1.510	29630
GLNp	10.700	−115400

Table 2: Reproducibility of each method toward the Illumina spike-in based on the experiment data

Model	RMSE	K-L
ENr	7.251	1141000
ENn	7.127	1062000
ENm	6.927	926500
ENmc	6.927	926200
EG	6.919	907900
GN	7.100	1183000
ELNn	7.124	1062000
ELNm	6.904	911600
ELNp	7.092	1035000
GLNn	**6.825**	**793400**
GLNm	6.937	968400

table is the median Kullback-Leibler coefficient based on $J = 42$ arrays. The Kullback-Leibler coefficient for two positive sequences $(X_{1j}, \ldots, X_{Ij}), (S_{1j}, \ldots, S_{Ij})$ is computed as $\text{K-L}_j = \sum_{i=1}^{I} X_{ij} \log(X_{ij}/S_{ij})$, which can also be negative if $S_{ij} > X_{ij}$ for all or for most i. This is a sign that the S are overestimating X, where X could be O (Table 1) or P (Table 2).

Therefore we exclude the GLNp model from further comparisons. The behavior of GLNp which is different from other models, also shown at the supplemental plots.

Table 2 shows how each method reproduces the data from the experiment. We see that GLNn can be considered to reproduce it better than others, based on the RMSE, and the Kullback-Leibler coefficient. Tables 1 and 2 provide insight about how the performance comparison among the models would be conducted further.

In the first part, we compute the adopted Affycomp benchmarking criteria, based on the data after background correction and their log transformation. In the second part, in the simulation, the MSE_{bc} and the L_1 error will be computed based on the log transformation of the experiment and the nominal concentration data.

The log transformations that we use here, respectively, for the benchmarking and the FFPE data sets are as follows

$$y = \log_2\left(x + \sqrt{x^2+1}\right) \qquad \text{and} \qquad y = \log_2\left(x + 1 + \sqrt{x^2+1}\right),$$

Table 3: Median SD, IQR, and 99.9% percentiles of log fold change for non spike-in between replicates for each model.

Model	Median SD	IQR	99.9%
ENr	**0.027**	**0.062**	0.415
ENn	0.043	0.089	0.441
ENm	0.069	0.139	0.486
ENmc	0.069	0.139	0.486
EG	0.065	0.134	0.477
GN	0.051	0.098	0.520
ELNn	0.045	0.093	0.442
ELNm	0.071	0.145	0.489
ELNp	0.049	0.100	0.449
GLNn	0.038	0.075	**0.398**
GLNm	0.076	0.080	0.507

Table 4: The signal detect R^2 by regressing the nominal and observed value for each model for the Illumina spike-in.

Model	R^2	Low.R^2	Med.R^2	High.R^2
ENr	0.959	0.618	**0.698**	**0.559**
ENn	0.958	0.622	0.695	0.557
ENm	0.957	0.635	0.695	0.558
ENmc	0.957	0.635	0.695	0.558
EG	0.957	0.633	0.695	0.558
GN	0.956	**0.650**	0.697	0.555
ELNn	0.958	0.624	0.695	0.557
ELNm	0.957	0.636	0.694	0.558
ELNp	0.958	0.627	0.695	0.557
GLNn	**0.960**	0.609	0.696	0.558
GLNm	0.956	0.637	0.694	0.558

where x is the nominal concentration O or the observed intensity value P.

First part In Table 3 it is shown that the ENr method provides the smallest variation and IQR and the GLNn model provides the smallest 99.9% percentiles of log fold change for the non spike-in between replicates. The largest variation, IQR, and 99.9% percentiles, respectively, are observed for the GLNm, the ELNm, and the GN method.

In Table 4 it is shown that, in general, all methods perform similar to each other. The GLNn models have the highest signal detect R^2. The GN model has the highest R^2 at low concentration but has the lowest R^2 at high concentration. This means that the GN model works better at low concentration. On the other hand the ENr shows that it works better at medium and high concentrations, which is followed closely by GLNn model.

If we divide the concentrations into two categories, where high concentration means that the nominal concentration is at least 3 pM and low concentration means that the nominal concentration is at most 1 pM, the GLNn model has the highest R^2 (the data is not shown here). It means, in general and at high concentrations, the GLNn offers a better fit than other models. As in Table 4, Table 5 shows that all models have similar performance, although the GLNn model has the highest R^2 of nominal concentration against observed log-fold-change. Table 6 provides the results from the computation of the AUC value. The table shows that all models have a better accuracy at medium concentrations than at low and high concentrations. The ENr performs very poor at the low concentrations, but the GLNm performs best. At high

Table 5: The R^2 observed log-fold-change against nominal log-fold-changes for the spike-in genes.

Model	Obs-intended-fc.R^2	Obs-(low)-int-fc.R^2
ENr	0.976	0.989
ENn	0.974	0.990
ENm	0.972	0.985
ENmc	0.972	0.985
EG	0.972	0.986
GN	0.970	0.987
ELNn	0.974	0.990
ELNm	0.972	0.985
ELNp	0.973	0.990
GLNn	**0.978**	**0.991**
GLNm	0.971	0.984

Table 6: AUC value for each model.

Model	Low concentration AUC	Medium concentration AUC	High concentration AUC	Average AUC	All
ENr	0.450	0.987	**0.785**	0.585	0.886
ENn	0.518	0.987	0.764	0.631	0.899
ENm	0.573	0.987	0.741	0.667	0.911
ENmc	0.573	0.987	0.741	0.667	0.911
EG	0.567	0.987	0.746	0.664	0.910
GN	0.552	0.987	0.723	0.651	0.904
ELNn	0.524	0.987	0.763	0.635	0.900
ELNm	0.574	0.987	0.741	0.668	0.912
ELNp	0.534	0.987	0.761	0.642	0.902
GLNn	0.498	**0.987**	0.784	0.619	0.896
GLNm	**0.579**	0.987	0.730	**0.671**	**0.913**

concentrations, the ENr performs the best and it is followed by the GLNn. But in general, the highest AUC is achieved by all the model with the MLE parameter estimation methods: the GLNm, ELNm, and ENm.

The computation, which is based on all arrays, provides the results where all models have the AUC greater than 0.9. According to Zhu, Zeng, and Wang (2010), the AUC between 0.9 and 1.0 is classified as excellent in measuring the accuracy. Therefore, based on Table 6, we can identify that there are some models excellency accurate in predicting the gene expression.

Second part We do $N = 100$ simulations to assess the performance of each model. The bias of the background correction is assessed by the MSE_{bc}, and the bias of the parameterization is assessed by the L_1 error.

From the simulation results in Table 7 we can see that results for the EG model are not available, because the MBCB package did not work at the log transformation that we have chosen. The GN model performs best, by providing the smallest bias for the background correction and the parameters. A close performance is achieved by the ELN, particularly by the ELNn. The GLNn does not have an optimal performance on the MSE_{bc}, but we still can consider its performance good, concerning that the bias of the parameters are similar to other proposed models and GN.

One of the proposed models, GLNm has the highest bias on the MSE_{bc} and the parameter α. In our view this happens because we use an approximation in estimating the true intensity

Table 7: The simulation results on the spike-in data set.

Model	MSE_{bc}	L_1 error			
		α	β	μ	σ
ENr	0.045		0.664	46.58	11.44
ENn	0.049		0.625	41.61	2.806
ENm	0.038		0.610	58.92	2.040
ENmc	0.036		0.610	62.77	2.039
GN	**0.030**	**0.0003**	**0.007**	0.013	**0.015**
ELNn	0.048		0.009	0.0004	0.018
ELNm	0.039		0.840	0.0004	0.018
ELNp	0.061		0.472	0.0004	0.018
GLNn	0.216	0.052	0.055	0.0004	0.018
GLNm	84.37	38.86	0.851	**0.0003**	0.017

value. The EN models (ENr, ENm, ENn and ENmc) have considerably better performance at MSE_{bc}, but are not good at the parametrization. The bias on the parametrization of the noise is higher than in other models.

3.4. The public data sets

Based on the results from Section 3.3, we compare the performance of these models on some public data sets. We would like to know how good these models are in real data samples. Here, we choose to use the formalin-fixed, paraffin-embedded (FFPE) data sets from Waldron, Ogino, Hoshida, Shima, Reed, Simpson, Baba, Nosho, Segata, Vargas, Cummings, Lakhani, Kirkner, Giovannucci, Quackenbush, Golub, Fuchs, Parmigiani, and Huttenhower (2012): the FFPE of tumors from colorectal cancer patients (GSE32651, 1003 samples), breast cancer metastases of the lymph node and autopsy tissues (GSE32490: GSE32489, 120 samples). Each sample has 24526 bead-type level probes.

The links for the data set are `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32651` and `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32490`.

Currently the FFPE archival samples are widely available in million and it is a great source of information in medical studies about some diseases, for example cancer. This data type is suffering from the RNA degradation, which leads to poor performance in array-based studies. However, the Illumina's DASL assays could provide high-quality data from this degraded RNA samples.

Comparing the performance of these background correction models certainly would help researchers to choose the appropriate background correction for their data, particularly if their data is the FFPE type.

The background correction for the FFPE data set is implemented in three steps:

step 1 Do the quality control (QC) to the raw FFPE data. In this paper, we used the *ffpe* package in R (Waldron (2013)).

step 2 Do the data transformation $\log_2(P_{ij} + 1 + \sqrt{P_{ij}^2 + 1}$ to the raw FFPE data after QC and estimate the background correction parameters based on it. The estimators of true intensity value and the background correction are based on the regular and negative control probe intensity data, respectively.

step 3 Compute the true intensity value (the adjusted intensity estimator) based on the BC parameters at step 2.

Table 8: Simulation results on the GSE32651 data set.

Model	MSE_{bc}	L_1 error			
		α	β	μ	σ
ENr	0.058		0.657	297.67	22.61
ENn	0.281		0.572	1.984	2.627
ENm	0.086		0.591	28.05	1.715
ENmc	**0.036**		0.610	62.77	2.039
ELNn	0.275		**0.025**	0.001	0.018
ELNm	0.059		0.826	**0.0004**	0.018
ELNp	0.672		0.487	0.001	0.018
GLNn	0.838	**0.340**	0.527	0.001	0.018
GLNm	84.15	71.87	0.887	0.0005	**0.018**

Table 9: Simulation results on the GSE32489 data set.

Model	MSE_{bc}	L_1 error			
		α	β	μ	σ
ENr	**0.093**		0.665	67.17	9.511
ENn	0.863		0.509	0.936	2.039
ENm	0.182		0.558	14.20	1.712
ENmc	0.184		0.556	14.18	1.049
ELNn	1.055		0.857	0.002	0.018
ELNm	0.116		0.781	0.001	0.018
ELNp	1.247		0.461	0.002	0.018
GLNn	1.348	**0.332**	**0.497**	0.002	**0.018**
GLNm	164.98	22.24	0.805	**0.0004**	**0.018**

The results of our computation are in Tables 8 and 9. From these tables we can see that there are no EG and GN models. Neither of these models can work on these data sets. For some samples in the data set, both models fail to compute the parameters which has the consequence that the true intensity value cannot be provided.

We decided to remove the EG and GN models from further comparisons in both FFPE data sets. Here we provide the results of the rest of the models only.

Tables 8 and 9 consistently show that the bias of the parameters of noise in the EN models are higher than the proposed models. For the parameter β, the ELNn has the smallest bias and it is followed by the ELNp and the GLNn. With regard to the bias of the background correction, the EN models show the smallest bias in both of FFPE data sets.

4. Conclusions and indication of future work

We have studied additive models of background correction for BeadArrays and proposed some new models where the true intensity is assumed to have exponential or gamma distribution and the noise is lognormally distributed. We have derived the estimator of the true intensity value of the proposed models.

Further, we compared the performance of all models, based on the benchmarking and public data sets. In the benchmarking data set we adopted the criteria from the Affycomp (Cope *et al.* 2004) and for the simulation study we used the criteria which have been used in Xie *et al.* (2009), Chen *et al.* (2011) and Plancade *et al.* (2011, 2012). For the public data sets, we only used the criteria for the simulation study.

We have seen in Sections 3.3 and 3.3 that EN, EG, GN and GLN perform rather similar. However, the GLNn model has provided the highest reproducibility in comparison to other

models. From the Affycomp criteria we can provide the following points:

1. The ENr and GLNn provide the lowest variation between replicates and all models using the MLE estimation method have a higher variation than others.

2. The GLNn model has the highest signal detect R^2 in general and in high concentration. This means the GLNn model is the best fitted for the gene expression.

3. The GLNn model, based on the MvA plot, produces the least number of genes which should not be expressed but are nevertheless expressed. On the other hand, the GN model provides the largest number of such genes.

4. All models with the MLE estimation method have a higher average AUC value, which means that they provide a better accuracy in predicting the gene expression.

5. The ENr and GLNn have the lowest IQR of log fold-change between replicates.

6. Points 1 and 2 show that the GLNn and ENr are more accurate and precise in modelling the gene expression and points 3 and 5 show that the specificity and sensitivity of the GLNn and ENr model are better than others.

In the simulation study, the best performance in estimating the signal by measuring its background correction and parametrization errors is achieved by the GN model. It is followed by our proposed ELN models. It has been shown that the GLNn does not perform optimally at the MSE_{bc} criterion, but for the parametrization this model still can be considered good.

In the FFPE public data set, the GN and EG models cannot be implemented. This is in strong contrast with the fact that in the simulation study of benchmarking data set, the GN model has the best performance.

The EN models show the highest bias in the parametrization in both public data sets and the lowest bias in the background correction. Our proposed models, except the GLNm, show the lowest bias in the parametrization in both data sets and a moderate bias in the background correction.

Based on the results from the benchmarking data and the public data sets, we would suggest researchers the following:

1. if the GN model works properly at the data set at hand (i.e. the estimated signals in all arrays can be computed by this model and the simulation criteria for this data at this model are low) then use the GN model to correct the background.

2. if the GN model fails, then use our proposed models, particularly the GLNn model. The reason for not choosing the ELN models is that the value of the parameter α from the benchmarking data set is less than 1, around 0.2. Therefore, the gamma model is more appropriate to model the true intensity distribution than the exponential one. We believe that the right approximative computation of the GLN models will lead to a better performance than the current approximation.

   The ELN models perform better than the original EN models, due to the fact that not only the regular probes, but also the control probes are skew-distributed (Chen *et al.* 2011). Therefore, these models could be the second choice after the GLN, when the GN model does not work.

3. With regard to the computation time, at the benchmarking data set the EN models are working faster than the others. They are followed by the ELNp, ELNn, and EG. The GLNn and the ENmc are the third fastest, then come the GN and the ELNm, which are followed by the GLNm as the slowest one.

One of the purposes of using microarray technology is finding the genes which are expressed differentially due to some disease or condition. Therefore, it is important to investigate the effect of bias of the background correction and the parametrization toward the differentially expressed genes. This will be our future work.

Acknowledgements

This paper is part of the author's Ph.D dissertation written under the direction of Prof. István Berkes. We would like to thank Herwig Friedl, Ernst Stadlober, Levi Waldron, and an anonymous referee for their comments leading to a substantial improvement of the paper. We thank Florian Endel for discussion and suggestion how to make the R programming more efficient. We thank Yevgeniy Grigoryev and Ken Laing for their generous permission to use their pictures in the Supplementary material at http://rfajriyah.staff.uii.ac.id/category/supplementary/. Financial support from the Austrian Science Fund (FWF), Project P24302-N18 is gratefully acknowledged.

References

Allen JD, Chen M, Xie Y (2009). "Model-Based Background Correction (MBCB): R Methods and GUI for Illumina Bead-array Data." *Journal of Cancer Science and Therapy*, **1**(1), 25–27.

Baek J, Son YS, MacLachlan GJ (2007). "Segmentation and Intensity Estimation of Microarray Images Using a Gamma-t Mixture Mode." *Bioinformatics*, **23**(4), 458–465.

Bolstad BM (2004). *Low Level Analysis of High-Density Oligonucleotide Array Data: Background, Normalization and Summarization*. Ph.D. thesis, University of California, California, Berkeley.

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003). "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance." *Bioinformatics*, **19**(2), 185–193.

Chen M, Xie Y, Story MD (2011). "An Exponential-Gamma Convolution Model for Background Correction of Illumina BeadArray Data." *Communication in Statistics: theory and methods*, **40**(17), 3055–3069.

Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP (2004). "A Benchmark for Affymetrix GeneChip Expression Measures." *Bioinformatics*, **20**, 323–331.

Ding LH, Xie Y, Park S, Xiao G, Story MD (2008). "Enhanced Identification and Biological Validation of Differential Gene Expression via Illumina Whole-Genome Expression Arrays through the Use of the Model-Based Background Correction Methodology." *Nucleic Acids Research*, **36**(10: e58).

Dunning MJ, Barbosa-Morais NL, Lynch AG, Tavaré S, Ritchie ME (2008). "Statistical Issues in the Analysis of Illumina Data." *BMC Bioinformatics*, **9**(85).

Forcheh AC, Verbeke G, Kasim A, Lin D, Shkedy Z, Talloen W, Gohlmann HW, Clement L (2012). "Gene Filtering in the Analysis of Illumina Microarrays Experiments." *Statistical Applications in Genetics and Molecular Biology*, **11**(2).

Hochreiter S, Djork-Arné, Obermayer K (2006). "A New Summarization Method for Affymetrix Probe Level Data." *Bioinformatics*, **22**(8), 943–949.

Huber W, Irizarry RA, Gentleman R (2005a). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter Preprocessing Overview. Springer.

Huber W, von Heydebreck A, Vingron M (2004). "Error Models for microarray Intensities." *Technical Report Paper 6*, Bioconductor Project Working Papers.

Huber W, von Heydebreck A, Vingron M (2005b). "An Introduction to Low-Level Analysis Methods of DNA Microarray Data." *Technical Report Paper 9*, Bioconductor Project Working Papers.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003a). "Summaries of Afymetrix GeneChip Probe Level Data." *Nucleic Acids Research*, **31**(4).

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003b). "Exploration, Normalization and Summaries of High Density Oligonucleotide Array Probe Level Data." *Biostatistics*, **4**(2), 249–264.

Irizarry RA, Wu Z, Jaffee HA (2006). "Comparison of Affymetrix geneChip Expression Measures." *Bioinformatics*, **22**(7), 789–794.

Li C, Wong WH (2001). "Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection." *Proceeding national Academy of Sciences*, **98**(1), 31–36.

Plancade S, Rozenholc Y, Lund E (2011). "Improving Background Correction for Illumina BeadArrays: the Normal-Gamma Model."

Plancade S, Rozenholc Y, Lund E (2012). "Generalization of the Normal-Exponential Model: Exploration of a More Accurate Parameterisation for the Signal Distribution on Illumina BeadArrays." *BMC Bioinformatics*, **13**(329).

Shi W, Oshlack A, Smyth GK (2010). "Optimizing the Noise versus Bias Trade-off for Illumina Whole Enome Expression Beadchips." *Nucleic Acids Research*, **38**(22: e204).

Silver JD, Ritchie ME, Smyth GK (2009). "Microarray Background Correction: Maximum Likelihood Estimation for the Normal-Exponential Convolution Model." *Biostatistics*, **10**, 352–363.

Triche TJ, Weisenberger DJ, Berg DVD, Laird PW, Siegmund KD (2013). "Low-level Processing of Illumina Infinium DNA Methylation BeadArrays." *Nucleic Acids Research*, pp. 1–11.

Waldron L (2013). *ffpe: Quality Assessment and Control for FFPR Microarray Expression Data*, r package version 1.4.0 edition.

Waldron L, Ogino S, Hoshida Y, Shima K, Reed AEM, Simpson PT, Baba Y, Nosho K, Segata N, Vargas AC, Cummings M, Lakhani SR, Kirkner GJ, Giovannucci E, Quackenbush J, Golub TR, Fuchs CS, Parmigiani G, Huttenhower C (2012). "Expression Profiling of Archival Tumors for Long-term Health Studies." *Clinical Cancer Research*.

Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004). "A Model-Based Background Adjustment for Oligonucleotide Expression Arrays." *Journal of the American Statistical Association*, **99**(468), 909 – 917.

Xie Y, Wang X, Story MD (2009). "Statistical Methods of Background Correction for Illumina BeadArray Data." *Bioinformatics*, **25**(6), 751–757.

Zhang J (2013). "Reducing the Bias of the Maximum Likelihood Estimator of the Shape Parameter for the Gamma Distribution." *Computational Statistics*, **28**, 1715 – 1724.

Zhu W, Zeng N, Wang N (2010). "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations." Http://www.nesug.org/Proceedings/nesug10/hl/hl07.pdf.

Affiliation:

Rohmatul Fajriyah
Institute of Statistics
Graz University of Technology
8010 Graz, Austria
E-mail: fajriyah@student.tugraz.at

Department of Statistics
Universitas Islam Indonesia
Jl. Kaliurang Km. 14.4
55584 Jogjakarta, Indonesia
E-mail: rfajriyah@fmipa.uii.ac.id
URL: http://rfajriyah.staff.uii.ac.id/

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 45 *Submitted:* 2014-11-13
March 2016 *Accepted:* 2014-11-13

Austrian Journal of Statistics
March 2016, Volume 45, 35–51.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i2.55

Riesz and Beta-Riesz Distributions

José A. Díaz-García
Universidad Autónoma Agraria Antonio Narro

Abstract

This article derives several properties of the Riesz distributions, such as their corresponding Bartlett decompositions, the inverse Riesz distributions, the distribution of the generalised variance and the density of their eigenvalues for real normed division algebras. In addition, introduce a kind of generalised beta distribution termed beta-Riesz distribution for real normed division algebras. Two versions of this distributions are proposed and some properties are studied.

Keywords: Beta-Riesz distribution, Riesz distribution, generalised beta function and distribution, real, complex quaternion and octonion random matrices.

1. Introduction

It is imminent the important role played by Wishart and beta distributions type I and II in the context of multivariate statistics. In particular, the relationship between these two distributions to obtain the beta distribution in terms of the distribution of two Wishart matrices.

In the last three decades, the family of elliptical contoured distributions has emerged as a robust alternative for dealing with non-normal samples. A number of well known distributions belong to this class, such is the case of normal, t, Bessel, Kotz type, logistic, Pearson type II and IV, among many others; but also infinitely many new distributions can be constructed by choosing a suitable kernel corresponding to a measurable function. Elliptical distributions are characterized by several properties, but for the context of sampling, their large variability of kurtosis and weight of tails, can assure a best explanation of the sample, rather than the usual forced fit to a normal model in presence of non explicable extremal points. Another important property of this set resides in the normal invariant statistics; i.e., assume that certain random matrix $\mathbf{X}$ follows a matrix multivariate elliptical distribution, then, many matrices of the form $\mathbf{Y} = f(\mathbf{X})$, for special functions $f(\cdot)$, are invariant under the complete family of elliptical distribution, in the sense that the distribution of $\mathbf{Y}$ is invariant, independently of the particular distribution of $\mathbf{X}$, in fact the distribution of $\mathbf{Y}$ coincides with the case where $\mathbf{X}$ has a matrix multivariate normal distribution, see Fang and Zhang (1990) and Gupta and Varga (1993).

However, we must note that the addressed invariance only holds under a probabilistic de-

pendence assumption; i.e., if the matrix $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$ follows an elliptical distribution, then $\mathbf{X}_1$ and $\mathbf{X}_2$ must be probabilistically dependent (recall that $\mathbf{X}_1$ and $\mathbf{X}_2$ are probabilistically independent if $\mathbf{X}$ has a matrix multivariate normal distribution, Gupta and Varga (1993)). Now, let $\mathbf{X}^T$ denotes the transpose of $\mathbf{X}$ and consider a non negative definite matrix $\mathbf{B}$, where $\mathbf{B}^{1/2}$ denotes its non negative squared root, see Fang and Zhang (1990), then, the matrix $\mathbf{F} = (\mathbf{X}_2^T\mathbf{X}_2)^{-1/2}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_2^T\mathbf{X}_2)^{-1/2}$ is said to have a matrix multivariate Beta type II distribution, moreover, the same distribution is obtained when $\mathbf{X}$ is assumed to follow matrix multivariate distribution, see Fang and Zhang (1990) and Gupta and Varga (1993).

Now, if the random matrix $\mathbf{X}$ follows a matrix multivariate elliptical distribution and the random matrix $\mathbf{W} = \mathbf{X}^T\mathbf{X}$ is defined, for every particular elliptical distribution, $\mathbf{W}$ follows also a different distribution. The distribution of $\mathbf{W}$ is usually known as the generalised Wishart distribution, and it heritages several properties of elliptical contoured distributions.

By another hand some recent advances in probability has reached interesting general distributions, such as the case of Riesz distribution, due to Hassairi and Lajmi (2001), and named under the denomination of Riesz natural exponential family (Riesz NEF); a distribution based in an special case of the well known Riesz measure given by Faraut and Korányi (1994, p. 137). In fact, Riesz distribution includes Wishart and Gamma matrix multivariate distributions as particular cases. Now, integrating theories have also appeared in other contexts. For example, in matrix multivariate distribution theory, some extensions from real to complex or quaternion or octonion fields appeared separately with great theoretical effort in many cases, the results in each field arose unconnected each other for years; only recent a new approach in the context of real normed division algebras could integrate, with an unified theory, all the addressed dispersed results in such sets of numbers. Following this tendency, recently, Díaz-García (2015a) proposes two versions of the Riesz distribution for real normed division algebras. Alternatively, Díaz-García (2015b) shows that the two versions of Riesz distributions correspond to two generalised Wishart distributions, both derived from certain matrix multivariate elliptical distributions, which are termed matrix multivariate Kotz-Riesz distributions.

Faraut and Korányi (1994), and subsequently Hassairi, Lajmi, and Zine (2005), propose a beta-Riesz distribution, which contains as a special case to the beta distribution obtained in terms of the distribution Wishart, which shall be named beta-Wishart, all this subjects in the context of simple Euclidean Jordan algebras. Such beta-Riesz distribution is obtained analogously to the beta-Wishart distribution, but starting with a Riesz distribution.

Based in these last two versions of the Riesz distributions, it is possible to obtain two versions of the beta-Riesz distributions, which by analogy with the beta-Wishart distributions are termed beta-Riesz type I. As in classical beta-Wishart distribution, in addition it is feasible to propose two version for the beta-Riesz distribution type II. Each of the two versions for each beta-Riesz distributions of type I and II, (both versions for each) contain as particular cases to beta-Wishart distribution type I and beta-Wishart distribution type II, respectively.

Thus, there is no doubt about the theoretical and applied potential of Riesz distribution into the setting in the setting of the integrative modern multivariate analysis. In general, every problem possible ruled by a Wishart process with considerable constraints, potentially can be studied under a more robust Riesz distribution; namely, some opportunities for consideration involve estimation of covariance matrices and principal component estimation, under any of the classical or bayesian approaches. Beta-Riesz distribution, for example, arises in a natural way in bayesian inference, when the two parameter a priori distributions are probabilistically independent. To prove this fact, recall the example about matrix $\mathbf{X}$, referred in the third paragraph of this section, and note that several situations consider are governed by probabilistically independent matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, both of them following a Kotz-Riesz elliptical distribution. Under these assumptions, the distribution of the corresponding random matrix $\mathbf{F}$ does not follow a beta-Wishart, but the beta-Riesz distribution. Then, if the two parameters, $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ are distributed as $(\mathbf{X}_1^T\mathbf{X}_1)$ and $(\mathbf{X}_2^T\mathbf{X}_2)$, respectively; i.e. $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ follows

independent Riesz distributions, then the parameter matrix defined as $\mathbf{F} = \boldsymbol{\delta}_2^{-1/2}\boldsymbol{\delta}_1\boldsymbol{\delta}_2^{-1/2}$, has a beta-Riesz distribution.

Although during the 90's and 2000's were obtained important results in theory of random matrices distributions, the past 30 years have reached a substantial development. Essentially, these advances have been archived through two approaches based on the *theory of Jordan algebras* and the *theory of real normed division algebras*. A basic source of the mathematical tools of theory of random matrices distributions under Jordan algebras can be found in Faraut and Korányi (1994); and specifically, some works in the context of theory of random matrices distributions based on Jordan algebras are provided in Massam (1994), Casalis and Letac (1996), Hassairi and Lajmi (2001), and Hassairi *et al.* (2005), and the references therein. Parallel results on theory of random matrices distributions based on real normed division algebras have been also developed in random matrix theory and statistics, see for example Gross and Richards (1987), Dumitriu (2002), Forrester (2005), Díaz-García and Gutiérrez-Jáimez (2011, 2013), among others. In addition, from mathematical point of view, several basic properties of the matrix multivariate Riesz distribution under *the structure theory of normal j-algebras* and under *theory of Vinberg algebras* in place of Jordan algebras have been studied, see Ishi (2000) and Boutouria and Hassairi (2009), respectively.

From a applied point of view, the relevance of *the octonions* remains unclear. An excellent review of the history, construction and many other properties of octonions is given in Baez (2002), where it is stated that... **However, there is still no *proof* that the octonions are useful for understanding the real world**. We can only hope that eventually this question will be settled one way or another."

For the sake of completeness, in this article the case of octonions is considered, but the veracity of the results obtained for this case can only be conjectured. Nonetheless, Forrester (2005, Section 1.4.5, pp. 22-24) it is proved that the bi-dimensional density function of the eigenvalue, for a Gaussian ensemble of a 2×2 octonionic matrix, is obtained from the general joint density function of the eigenvalues for the Gaussian ensemble, assuming $m = 2$ and $\beta = 8$, see Section 2. Moreover, as is established in Faraut and Korányi (1994) and Sawyer (1997) the result obtained in this article are valid for the *algebra of Albert*, that is when hermitian matrices ($\mathbf{S}$) or hermitian product of matrices ($\mathbf{X}^*\mathbf{X}$) are 3×3 octonionic matrices.

This article studies two versions for beta-Riesz distributions type I and II for real normed division algebras. Section 2 reviews some definitions and notation on real normed division algebras. And also, introduces other mathematical tools as three Jacobians with respect to Lebesgue measure and some integral results for real normed division algebras. Section 3 proposes diverse properties of two versions of the Riesz distributions as their Bartlett decompositions, inverse Riesz distributions, the distribution of the generalized variance and the distribution of their eigenvalues. Section 4 introduces two generalised beta functions and, in terms of these, two beta-Riesz distributions of type I and II are obtained for real normed division algebras. Also, the relationship between the Riesz distributions and the beta-Riesz distributions are studied. This section concludes studying the eigenvalues distributions of beta-Riesz distributions type I and II in their two versions for real normed division algebras.

2. Preliminary results

A detailed discussion of real normed division algebras may be found in Baez (2002) and Neukirch, Prestel, and Remmert (1990). For convenience, we shall introduce some notation, although in general we adhere to standard notation forms.

For our purposes: Let $\mathbb{F}$ be a field. An *algebra* $\mathfrak{A}$ over $\mathbb{F}$ is a pair $(\mathfrak{A}; m)$, where $\mathfrak{A}$ is a *finite-dimensional vector space* over $\mathbb{F}$ and *multiplication* $m : \mathfrak{A} \times \mathfrak{A} \to A$ is an $\mathbb{F}$-bilinear map; that

is, for all $\lambda \in \mathbb{F}$, $x, y, z \in \mathfrak{A}$,

$$\begin{aligned} m(x, \lambda y + z) &= \lambda m(x; y) + m(x; z) \\ m(\lambda x + y; z) &= \lambda m(x; z) + m(y; z). \end{aligned}$$

Two algebras $(\mathfrak{A}; m)$ and $(\mathfrak{E}; n)$ over $\mathbb{F}$ are said to be *isomorphic* if there is an invertible map $\phi : \mathfrak{A} \to \mathfrak{E}$ such that for all $x, y \in \mathfrak{A}$,

$$\phi(m(x, y)) = n(\phi(x), \phi(y)).$$

By simplicity, we write $m(x; y) = xy$ for all $x, y \in \mathfrak{A}$.

Let $\mathfrak{A}$ be an algebra over $\mathbb{F}$. Then $\mathfrak{A}$ is said to be

1. *alternative* if $x(xy) = (xx)y$ and $x(yy) = (xy)y$ for all $x, y \in \mathfrak{A}$,
2. *associative* if $x(yz) = (xy)z$ for all $x, y, z \in \mathfrak{A}$,
3. *commutative* if $xy = yx$ for all $x, y \in \mathfrak{A}$, and
4. *unital* if there is a $1 \in \mathfrak{A}$ such that $x1 = x = 1x$ for all $x \in \mathfrak{A}$.

If $\mathfrak{A}$ is unital, then the identity 1 is uniquely determined.

An algebra $\mathfrak{A}$ over $\mathbb{F}$ is said to be a *division algebra* if $\mathfrak{A}$ is nonzero and $xy = 0_{\mathfrak{A}} \Rightarrow x = 0_{\mathfrak{A}}$ or $y = 0_{\mathfrak{A}}$ for all $x, y \in \mathfrak{A}$.

The term "division algebra", comes from the following proposition, which shows that, in such an algebra, left and right division can be unambiguously performed.

Let $\mathfrak{A}$ be an algebra over $\mathbb{F}$. Then $\mathfrak{A}$ is a division algebra if, and only if, $\mathfrak{A}$ is nonzero and for all $a, b \in \mathfrak{A}$, with $b \neq 0_{\mathfrak{A}}$, the equations $bx = a$ and $yb = a$ have unique solutions $x, y \in \mathfrak{A}$.

In the sequel we assume $\mathbb{F} = \Re$ and consider classes of division algebras over $\Re$ or "*real division algebras*" for short.

We introduce the algebras of *real numbers* $\Re$, *complex numbers* $\mathfrak{C}$, *quaternions* $\mathfrak{H}$ and *octonions* $\mathfrak{O}$. Then, if $\mathfrak{A}$ is an alternative real division algebra, then $\mathfrak{A}$ is isomorphic to $\Re$, $\mathfrak{C}$, $\mathfrak{H}$ or $\mathfrak{O}$.

Let $\mathfrak{A}$ be a real division algebra with identity 1. Then $\mathfrak{A}$ is said to be *normed* if there is an inner product $(\cdot, \cdot)$ on $\mathfrak{A}$ such that

$$(xy, xy) = (x, x)(y, y) \qquad \text{for all } x, y \in \mathfrak{A}.$$

If $\mathfrak{A}$ is a *real normed division algebra*, then $\mathfrak{A}$ is isomorphic $\Re$, $\mathfrak{C}$, $\mathfrak{H}$ or $\mathfrak{O}$.

There are exactly four normed division algebras: real numbers ($\Re$), complex numbers ($\mathfrak{C}$), quaternions ($\mathfrak{H}$) and octonions ($\mathfrak{O}$), see Baez (2002).

Let $\mathfrak{A}$ be a division algebra over the real numbers. Then $\mathfrak{A}$ has dimension either 1, 2, 4 or 8. In other branches of mathematics, the parameters $\alpha = 2/\beta$ and $t = \beta/4$ are used, see Edelman and Rao (2005) and Kabe (1984), respectively.

Let $\mathcal{L}_{m,n}^{\beta}$ be the set of all $m \times n$ matrices of rank $m \leq n$ over $\mathfrak{A}$ with m distinct positive singular values, where $\mathfrak{A}$ denotes a *real finite-dimensional normed division algebra*. Let $\mathfrak{A}^{m \times n}$ be the set of all $m \times n$ matrices over $\mathfrak{A}$. The dimension of $\mathfrak{A}^{m \times n}$ over $\Re$ is βmn. Let $\mathbf{A} \in \mathfrak{A}^{m \times n}$, then $\mathbf{A}^* = \bar{\mathbf{A}}^T$ denotes the usual conjugate transpose.

We denote by $\mathfrak{S}_m^{\beta}$ the real vector space of all $\mathbf{S} \in \mathfrak{A}^{m \times m}$ such that $\mathbf{S} = \mathbf{S}^*$. In addition, let $\mathfrak{P}_m^{\beta}$ be the *cone of positive definite matrices* $\mathbf{S} \in \mathfrak{A}^{m \times m}$. Thus, $\mathfrak{P}_m^{\beta}$ consist of all matrices $\mathbf{S} = \mathbf{X}^*\mathbf{X}$, with $\mathbf{X} \in \mathcal{L}_{m,n}^{\beta}$; then $\mathfrak{P}_m^{\beta}$ is an open subset of $\mathfrak{S}_m^{\beta}$.

Let $\mathfrak{D}_m^{\beta}$ consisting of all $\mathbf{D} \in \mathfrak{A}^{m \times m}$, $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_m)$. Let $\mathfrak{T}_U^{\beta}(m)$ be the subgroup of all *upper triangular* matrices $\mathbf{T} \in \mathfrak{A}^{m \times m}$ such that $t_{ij} = 0$ for $1 < i < j \leq m$.

For any matrix $\mathbf{X} \in \mathfrak{A}^{n\times m}$, $d\mathbf{X}$ denotes the *matrix of differentials* (dx_{ij}). Finally, we define the *measure* or volume element $(d\mathbf{X})$ when $\mathbf{X} \in \mathfrak{A}^{m\times n}, \mathfrak{S}_m^\beta$, and $\mathfrak{D}_m^\beta$, see Dumitriu (2002) and Díaz-García and Gutiérrez-Jáimez (2011).

If $\mathbf{X} \in \mathfrak{A}^{m\times n}$ then $(d\mathbf{X})$ (the Lebesgue measure in $\mathfrak{A}^{m\times n}$) denotes the exterior product of the βmn functionally independent variables

$$(d\mathbf{X}) = \bigwedge_{i=1}^{m}\bigwedge_{j=1}^{n} dx_{ij} \quad \text{where} \quad dx_{ij} = \bigwedge_{k=1}^{\beta} dx_{ij}^{(k)}.$$

If $\mathbf{S} \in \mathfrak{S}_m^\beta$ (or $\mathbf{S} \in \mathfrak{T}_U^\beta(m)$ with $t_{ii} > 0$, $i = 1, \ldots, m$) then $(d\mathbf{S})$ (the Lebesgue measure in $\mathfrak{S}_m^\beta$ or in $\mathfrak{T}_U^\beta(m)$) denotes the exterior product of the exterior product of the $m(m-1)\beta/2 + m$ functionally independent variables,

$$(d\mathbf{S}) = \bigwedge_{i=1}^{m} ds_{ii} \bigwedge_{i<j}^{m}\bigwedge_{k=1}^{\beta} ds_{ij}^{(k)}.$$

Observe, that for the Lebesgue measure $(d\mathbf{S})$ defined thus, it is required that $\mathbf{S} \in \mathfrak{P}_m^\beta$, that is, $\mathbf{S}$ must be a non singular Hermitian matrix (Hermitian definite positive matrix).

If $\mathbf{\Lambda} \in \mathfrak{D}_m^\beta$ then $(d\mathbf{\Lambda})$ (the Legesgue measure in $\mathfrak{D}_m^\beta$) denotes the exterior product of the βm functionally independent variables

$$(d\mathbf{\Lambda}) = \bigwedge_{i=1}^{n}\bigwedge_{k=1}^{\beta} d\lambda_i^{(k)}.$$

Now, we show three Jacobians in terms of the β parameter, which are proposed as extensions of real, complex or quaternion cases, see Díaz-García and Gutiérrez-Jáimez (2011).

Lemma 2.1 *Let* $\mathbf{X}$ *and* $\mathbf{Y} \in \mathfrak{P}_m^\beta$ *matrices of functionally independent variables, and let* $\mathbf{Y} = \mathbf{A}\mathbf{X}\mathbf{A}^* + \mathbf{C}$, *where* $\mathbf{A} \in \mathcal{L}_{m,m}^\beta$ *and* $\mathbf{C} \in \mathfrak{P}_m^\beta$ *are matrices of constants. Then*

$$(d\mathbf{Y}) = |\mathbf{A}^*\mathbf{A}|^{(m-1)\beta/2+1}(d\mathbf{X}), \tag{1}$$

where $|\mathbf{B}|$ *denotes the determinant of* $\mathbf{B}$.

Lemma 2.2 (Cholesky's decomposition) *Let* $\mathbf{S} \in \mathfrak{P}_m^\beta$ *and* $\mathbf{T} \in \mathfrak{T}_U^\beta(m)$ *with* $t_{ii} > 0$, $i = 1, 2, \ldots, m$. *Then*

$$(d\mathbf{S}) = \begin{cases} 2^m \prod_{i=1}^{m} t_{ii}^{(m-i)\beta+1}(d\mathbf{T}) & \text{if } \mathbf{S} = \mathbf{T}^*\mathbf{T}; \\ 2^m \prod_{i=1}^{m} t_{ii}^{(i-1)\beta+1}(d\mathbf{T}) & \text{if } \mathbf{S} = \mathbf{T}\mathbf{T}^*. \end{cases} \tag{2}$$

Lemma 2.3 *Let* $\mathbf{S} \in \mathfrak{P}_m^\beta$. *Then, ignoring the sign, if* $\mathbf{Y} = \mathbf{S}^{-1} + \mathbf{C}$, $\mathbf{C} \in \mathfrak{P}_m^\beta$ *is a matrix of constants,*

$$(d\mathbf{Y}) = |\mathbf{S}|^{-\beta(m-1)-2}(d\mathbf{S}). \tag{3}$$

Next is stated a general result, that is useful in a variety of situations, which enable us to transform the density function of a matrix $\mathbf{X} \in \mathfrak{P}_m^\beta$ to the density function of its eigenvalues, see Díaz-García (2014).

Lemma 2.4 *Let $\mathbf{X} \in \mathfrak{P}_m^\beta$ be a random matrix with a density function $f_{\mathbf{X}}(\mathbf{X})$. Then the joint density function of the eigenvalues $\lambda_1, \ldots, \lambda_m$ of $\mathbf{X}$ is*

$$\frac{\pi^{m^2\beta/2+\varrho}}{\Gamma_m^\beta[m\beta/2]} \prod_{i<j}^{m} (\lambda_i - \lambda_j)^\beta \int_{\mathbf{H}\in\mathfrak{U}^\beta(m)} f(\mathbf{H}\mathbf{L}\mathbf{H}^*)(d\mathbf{H}) \tag{4}$$

where $\mathbf{L} = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$, $\lambda_1 > \cdots > \lambda_m > 0$, $(d\mathbf{H})$ is the normalised Haar measure, $\Gamma_m^\beta[a]$ denotes the Gamma function for the space $\mathfrak{S}_m^\beta$ (Gross and Richards 1987) and

$$\varrho = \begin{cases} 0, & \beta = 1; \\ -m, & \beta = 2; \\ -2m, & \beta = 4; \\ -4m, & \beta = 8. \end{cases}$$

Finally, let's recall the multidimensional convolution theorem in terms of the Laplace transform. For this purpose, let's use the complexification $\mathfrak{S}_m^{\beta,\mathfrak{C}} = \mathfrak{S}_m^\beta + i\mathfrak{S}_m^\beta$ of $\mathfrak{S}_m^\beta$. That is, $\mathfrak{S}_m^{\beta,\mathfrak{C}}$ consist of all matrices $\mathbf{Z} \in (\mathfrak{F}^{\mathfrak{C}})^{m\times m}$ of the form $\mathbf{Z} = \mathbf{X} + i\mathbf{Y}$, with $\mathbf{X}, \mathbf{Y} \in \mathfrak{S}_m^\beta$. It comes to $\mathbf{X} = \operatorname{Re}(\mathbf{Z})$ and $\mathbf{Y} = \operatorname{Im}(\mathbf{Z})$ as the *real and imaginary parts* of $\mathbf{Z}$, respectively. The *generalised right half-plane* $\mathbf{\Phi}_m^\beta = \mathfrak{P}_m^\beta + i\mathfrak{S}_m^\beta$ in $\mathfrak{S}_m^{\beta,\mathfrak{C}}$ consists of all $\mathbf{Z} \in \mathfrak{S}_m^{\beta,\mathfrak{C}}$ such that $\operatorname{Re}(\mathbf{Z}) \in \mathfrak{P}_m^\beta$, see (Gross and Richards 1987, p. 801).

Definition 2.1 *If $f(\mathbf{X})$ is a function of $\mathbf{X} \in \mathfrak{P}_m^\beta$, the Laplace transform of $f(\mathbf{X})$ is defined to be*

$$g(\mathbf{T}) = \int_{\mathbf{X}\in\mathfrak{P}_m^\beta} \operatorname{etr}\{-\mathbf{X}\mathbf{Z}\} f(\mathbf{X})(d\mathbf{X}). \tag{5}$$

where $\mathbf{Z} \in \mathbf{\Phi}_m^\beta$ and $\operatorname{etr}(\cdot) = \exp(\operatorname{tr}(\cdot))$.

Lemma 2.5 *If $g_1(\mathbf{Z})$ and $g_2(\mathbf{Z})$ are the respective Laplace transforms of the densities $f_{\mathbf{X}}(\mathbf{X})$ and $g_{\mathbf{Y}}(\mathbf{Y})$ then the product $g_1(\mathbf{Z})g_2(\mathbf{Z})$ is the Laplace transform of the convolution $f_{\mathbf{X}}(\mathbf{X}) * g_{\mathbf{Y}}(\mathbf{Y})$, where*

$$h_{\mathbf{\Xi}}(\mathbf{\Xi}) = f_{\mathbf{X}}(\mathbf{X}) * g_{\mathbf{Y}}(\mathbf{Y}) = \int_{\mathbf{0}<\mathbf{\Upsilon}<\mathbf{\Xi}} f_{\mathbf{X}}(\mathbf{\Upsilon}) g_{\mathbf{Y}}(\mathbf{\Xi} - \mathbf{\Upsilon})(d\mathbf{\Upsilon}), \tag{6}$$

with $\mathbf{\Xi} = \mathbf{X} + \mathbf{Y}$ and $\mathbf{\Upsilon} = \mathbf{X}$.

3. Riesz distributions

This section shows two versions of the Riesz distributions (Díaz-García 2015a) and the study of their Bartlett decompositions. Also, inverse Riesz distributions and the joint density of its eigenvalues are obtained.

Definition 3.1 *Let $\mathbf{\Sigma} \in \mathbf{\Phi}_m^\beta$ and $\kappa = (k_1, k_2, \ldots, k_m) \in \Re^m$.*

1. *Then it is said that $\mathbf{X}$ has a Riesz distribution of type I if its density function is*

$$\frac{\beta^{am+\sum_{i=1}^m k_i}}{\Gamma_m^\beta[a,\kappa]|\mathbf{\Sigma}|^a q_\kappa(\mathbf{\Sigma})} \operatorname{etr}\{-\beta\mathbf{\Sigma}^{-1}\mathbf{X}\}|\mathbf{X}|^{a-(m-1)\beta/2-1} q_\kappa(\mathbf{X})(d\mathbf{X}) \tag{7}$$

for $\mathbf{X} \in \mathfrak{P}_m^\beta$ and $\operatorname{Re}(a) \geq (m-1)\beta/2 - k_m$; where $\Gamma_m^\beta[a,\kappa]$ is the generalised gamma function of weight κ and $q_\kappa(\mathbf{A})$ is the highest wight vector or generalised power of $\mathbf{A}$ (see Gross and Richards 1987); denoting this fact as $\mathbf{X} \sim \Re_m^{\beta,I}(a,\kappa,\mathbf{\Sigma})$.

2. *Then it is said that* $\mathbf{X}$ *has a Riesz distribution of type II if its density function is*

$$\frac{\beta^{am-\sum_{i=1}^{m}k_i}}{\Gamma_m^\beta[a,-\kappa]|\mathbf{\Sigma}|^a q_\kappa(\mathbf{\Sigma}^{-1})}\operatorname{etr}\{-\beta\mathbf{\Sigma}^{-1}\mathbf{X}\}|\mathbf{X}|^{a-(m-1)\beta/2-1}q_\kappa(\mathbf{X}^{-1})(d\mathbf{X}) \tag{8}$$

for $\mathbf{X}\in\mathfrak{P}_m^\beta$ *and* $\operatorname{Re}(a)>(m-1)\beta/2+k_1$*; where* $\Gamma_m^\beta[a,-\kappa]$ *is the generalised gamma function of weight* κ *proposed by Khatri (1966); denoting this fact as* $\mathbf{X}\sim\mathfrak{R}_m^{\beta,II}(a,\kappa,\mathbf{\Sigma})$.

Theorem 3.1 *Let* $\mathbf{\Sigma}\in\mathbf{\Phi}_m^\beta$ *and* $\kappa=(k_1,k_2,\ldots,k_m)\in\Re^m$*. And let* $\mathbf{Y}=\mathbf{X}^{-1}$.

1. *Then if* $\mathbf{X}$ *has a Riesz distribution of type I the density of* $\mathbf{Y}$ *is*

$$\frac{\beta^{am+\sum_{i=1}^{m}k_i}}{\Gamma_m^\beta[a,\kappa]|\mathbf{\Sigma}|^a q_\kappa(\mathbf{\Sigma})}\operatorname{etr}\{-\beta\mathbf{\Sigma}^{-1}\mathbf{Y}^{-1}\}|\mathbf{Y}|^{-(a+(m-1)\beta/2+1)}q_\kappa(\mathbf{Y}^{-1})(d\mathbf{Y}) \tag{9}$$

for $\operatorname{Re}(a)\geq(m-1)\beta/2-k_m$ *and is termed as* inverse Riesz distribution of type I.

2. *Then if* $\mathbf{X}$ *has a Riesz distribution of type II the density of* $\mathbf{Y}$ *is*

$$\frac{\beta^{am-\sum_{i=1}^{m}k_i}}{\Gamma_m^\beta[a,-\kappa]|\mathbf{\Sigma}|^a q_\kappa(\mathbf{\Sigma}^{-1})}\operatorname{etr}\{-\beta\mathbf{\Sigma}^{-1}\mathbf{Y}^{-1}\}|\mathbf{Y}|^{-(a+(m-1)\beta/2+1)}q_\kappa(\mathbf{Y})(d\mathbf{Y}) \tag{10}$$

for $\operatorname{Re}(a)>(m-1)\beta/2+k_1$ *and it said that* $\mathbf{Y}$ *has a* inverse Riesz distribution of type II.

Proof. It is immediately noted that $(d\mathbf{X})=|\mathbf{Y}|^{-\beta(m-1)-2}(d\mathbf{Y})$ and from (7) and (8). □

Note that, the density function (9) was studied previously by Tounsi and Zine (2012) in real case.

Observe that, if $\kappa=(0,0,\ldots,0)$ in two densities in Definition 3.1 and Theorem 3.1 the matrix multivariate gamma and inverse gamma distributions are obtained. As consequence, in this last case if $\mathbf{\Sigma}=2\mathbf{\Sigma}$ and $a=\beta n/2$, the Wishart and inverse Wishart distributions are obtained, too.

Theorem 3.2 *Let* $\mathbf{T}\in\mathfrak{T}_U^\beta(m)$ *with* $t_{ii}>0$*,* $i=1,2,\ldots,m$ *and define* $\mathbf{X}=\mathbf{T}^*\mathbf{T}$.

1. *If* $\mathbf{X}$ *has a Riesz distribution of type I, (7), with* $\mathbf{\Sigma}=\mathbf{I}_m$*, then the elements* t_{ij} $(1\leq i\leq j\leq m)$ *of* $\mathbf{T}$ *are all independent. Furthermore,* $t_{ii}^2\sim\mathcal{G}^\beta(a+k_i-(i-1)\beta/2,1)$ *and* $\sqrt{2}t_{ij}\sim\mathcal{N}_1^\beta(0,1)$ $(1\leq i<j\leq m)$.

2. *If* $\mathbf{X}$ *has a Riesz distribution of type II, (8), with* $\mathbf{\Sigma}=\mathbf{I}_m$*, then the elements* t_{ij} $(1\leq i\leq j\leq m)$ *of* $\mathbf{T}$ *are all independent. Moreover,* $t_{ii}^2\sim\mathcal{G}^\beta(a-k_i-(i-1)\beta/2,1)$ *and* $\sqrt{2}t_{ij}\sim\mathcal{N}_1^\beta(0,1)$ $(1\leq i<j\leq m)$.

Where $x\sim\mathcal{G}^\beta(a,\alpha)$ *denotes a gamma distribution with parameters* a *and* α *and* $y\sim\mathcal{N}_1^\beta(0,1)$ *denotes a random variable with standard normal distribution for real normed division algebras. Moreover, their respective densities are*

$$\mathcal{G}^\beta(x:a,\alpha)=\frac{1}{(\alpha/\beta)^a\Gamma[a]}\exp\{-\beta x/\alpha\}x^{a-1}(dx),$$

and

$$\mathcal{N}_1^\beta(y:0,1)=\frac{1}{(2\pi/\beta)^{\beta/2}}\exp\{-\beta y^2/2\}(dy)$$

where $x\in\mathfrak{L}_{1,1}^\beta$*,* $y\in\mathfrak{P}_1^\beta$*,* $Re(a)>0$ *and* $\alpha\in\mathbf{\Phi}_1^\beta$*, see Díaz-García and Gutiérrez-Jáimez (2011).*

Proof. This is given for the case of Riesz distribution type I. The proof for Riesz distribution type II is the same thing. The density of $\mathbf{X}$ is

$$\frac{\beta^{am+\sum_{i=1}^m k_i}}{\Gamma_m^\beta[a,\kappa]} \operatorname{etr}\{-\beta\mathbf{X}\}|\mathbf{X}|^{a-(m-1)\beta/2-1} q_\kappa(\mathbf{X})(d\mathbf{X}). \tag{11}$$

Since $\mathbf{X} = \mathbf{T}^*\mathbf{T}$ we have

$$\begin{aligned}
\operatorname{tr}\mathbf{X} &= \operatorname{tr}\mathbf{T}^*\mathbf{T} = \sum_{i\le j}^m t_{ij}^2, \\
|\mathbf{X}| &= |\mathbf{T}^*\mathbf{T}| = |\mathbf{T}|^2 = \prod_{i=1}^m t_{ii}^2, \\
q_\kappa(\mathbf{X}) &= q_\kappa(\mathbf{T}^*\mathbf{T}) = |\mathbf{T}^*\mathbf{T}|^{k_m} \prod_{i=1}^{m-1} |\mathbf{T}_i^*\mathbf{T}_i|^{k_i-k_{i+1}} = \prod_{i=1}^m t_{ii}^{2k_i},
\end{aligned}$$

and by Theorem 2.2 noting that $dt_{ii}^2 = 2t_{ii}dt_{ii}$, then

$$\begin{aligned}
(d\mathbf{X}) &= 2^m \prod_{i=1}^m t_{ii}^{\beta(m-i)+1} \left(\bigwedge_{i\le j} dt_{ij}\right), \\
&= \prod_{i=1}^m \left(t_{ii}^2\right)^{\beta(m-i)/2} \left(\bigwedge_{i=1} dt_{ii}^2\right) \wedge \left(\bigwedge_{i<j} dt_{ij}\right).
\end{aligned}$$

Substituting this expression in (11) we find that the joint density of the t_{ij} $(1 \le i \le j \le m)$ can be written as

$$\prod_{i=1}^m \frac{\beta^{a+k_i-(i-1)\beta/2}}{\Gamma[a+k_i-(i-1)\beta/2]} \exp\{-\beta t_{ii}^2\} \left(t_{ii}^2\right)^{a+k_i-(i-1)\beta/2-1} \left(dt_{ii}^2\right)$$

$$\times \prod_{i<j}^m \frac{1}{(\pi/\beta)^{\beta/2}} \exp\{-\beta t_{ij}^2\}(dt_{ij}),$$

only observe that

$$\begin{aligned}
\frac{\beta^{am+\sum_{i=1}^m k_i}}{\Gamma_m^\beta[a,\kappa]} &= \frac{\beta^{am+\sum_{i=1}^m k_i - m(m-1)\beta/4}}{\beta^{-m(m-1)\beta/4}\pi^{m(m-1)\beta/4}\prod_{i=1}^m \Gamma[a+k_i-(i-1)\beta/2]} \\
&= \prod_{i=1}^m \frac{\beta^{a+k_i-(i-1)\beta/2}}{\Gamma[a+k_i-(i-1)\beta/2]} \prod_{i<j}^m \frac{1}{(\pi/\beta)^{\beta/2}}.
\end{aligned}$$

□

In analogy to generalised variance for Wishart case, the following result gives the distribution of $|\mathbf{X}|$ when $\mathbf{X}$ has a Riesz distribution type I or type II.

Theorem 3.3 *Let $v = |\mathbf{X}|/|\mathbf{\Sigma}|$. Then*

1. *if $\mathbf{X}$ has a Riesz distribution of type I, (7), the density of v is*

$$\prod_{i=1}^m \mathcal{G}^\beta\left(t_{ii}^2 : a+k_i-(i-1)\beta/2, 1\right).$$

2. *if $\mathbf{X}$ has a Riesz distribution of type II, (8), the density of v is*

$$\prod_{i=1}^m \mathcal{G}^\beta\left(t_{ii}^2 : a-k_i-(i-1)\beta/2, 1\right).$$

where t_{ii}^2, $i = 1, \ldots, m$, are independent random variables.

Proof. This is immediately from Theorem 3.2, noting that if

$$\mathbf{B} = \mathbf{\Sigma}^{-1/2}\mathbf{X}\mathbf{\Sigma}^{-1/2} = \mathbf{T}^*\mathbf{T},$$

with $\mathbf{T} \in \mathfrak{T}_U^\beta(m)$ and $t_{ii} > 0$, $i = 1, 2, \ldots, m$, then

$$|\mathbf{B}| = \prod_{i=1}^m t_{ii}^2 = |\mathbf{X}|/|\mathbf{\Sigma}| = v.$$

□

Theorem 3.4 *Let $\mathbf{\Sigma} = \mathbf{I}_m$ and $\kappa = (k_1, k_2, \ldots, k_m)$, $k_1 \geq k_2 \geq \cdots \geq k_m \geq 0$, $k_1, k_2, \ldots, k_m$ are nonnegative integers.*

1. *Let $\lambda_1, \ldots, \lambda_m$, $\lambda_1 > \cdots > \lambda_m > 0$ be the eigenvalues of $\mathbf{X}$. Then if $\mathbf{X}$ has a Riesz distribution of type I, the joint density of $\lambda_1, \ldots, \lambda_m$ is*

$$\frac{\beta^{am+\sum_{i=1}^m k_i}\ \pi^{m^2\beta/2+\varrho}}{\Gamma_m^\beta[m\beta/2]\Gamma_m^\beta[a,\kappa]} \prod_{i<j}^m (\lambda_i - \lambda_j)^\beta \exp\left\{-\beta\sum_{i=1}^m \lambda_i\right\}$$
$$\times \prod_{i=1}^m \lambda_i^{a-(m-1)\beta/2-1} \frac{C_\kappa^\beta(\mathbf{L})}{C_\kappa^\beta(\mathbf{I}_m)}. \tag{12}$$

   where $\mathbf{L} = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ and $\operatorname{Re}(a) \geq (m-1)\beta/2 - k_m$.

2. *Let $\delta_1, \ldots, \delta_m$, $\delta_1 > \cdots > \delta_m > 0$ be the eigenvalues of $\mathbf{X}$. Then if $\mathbf{X}$ has a Riesz distribution of type II, the joint density of their eigenvalues is*

$$\frac{\beta^{am-\sum_{i=1}^m k_i}\ \pi^{m^2\beta/2+\varrho}}{\Gamma_m^\beta[m\beta/2]\Gamma_m^\beta[a,\kappa]} \prod_{i<j}^m (\delta_i - \delta_j)^\beta \exp\left\{-\beta\sum_{i=1}^m \delta_i\right\}$$
$$\times \prod_{i=1}^m \delta_i^{a-(m-1)\beta/2-1} \frac{C_\kappa^\beta(\mathbf{D}^{-1})}{C_\kappa^\beta(\mathbf{I}_m)}). \tag{13}$$

   where $\mathbf{D} = \operatorname{diag}(\delta_1, \ldots, \delta_m)$, $\operatorname{Re}(a) > (m-1)\beta/2 + k_1$.

Where ϱ is defined in Lemma 3 and $C_\kappa^\beta(\cdot)$ denotes the zonal spherical functions or spherical polynomials, see Gross and Richards (1987) and Faraut and Korányi (1994, Chapter XI, Section 3).

Proof. 1. From Lemma 3

$$\frac{\beta^{am+\sum_{i=1}^m k_i}\ \pi^{m^2\beta/2+\varrho}}{\Gamma_m^\beta[m\beta/2]\Gamma_m^\beta[a,\kappa]|\mathbf{\Sigma}|^a q_\kappa(\mathbf{\Sigma})} \prod_{i<j}^m (\lambda_i - \lambda_j)^\beta$$
$$\int_{\mathbf{H}\in\mathfrak{U}^\beta(m)} \operatorname{etr}\{-\beta\mathbf{H}\mathbf{L}\mathbf{H}^*\}|\mathbf{H}\mathbf{L}\mathbf{H}^*|^{a-(m-1)\beta/2-1} q_\kappa(\mathbf{H}\mathbf{L}\mathbf{H}^*)(d\mathbf{H}).$$

Therefore,

$$\frac{\beta^{am+\sum_{i=1}^m k_i}\ \pi^{m^2\beta/2+\varrho}}{\Gamma_m^\beta[m\beta/2]\Gamma_m^\beta[a,\kappa]|\mathbf{\Sigma}|^a q_\kappa(\mathbf{\Sigma})} \prod_{i<j}^m (\lambda_i - \lambda_j)^\beta \prod_{i=1}^m \lambda_i^{a-(m-1)\beta/2-1} \exp\left\{-\beta\sum_{i=1}^m \lambda_i\right\}$$

$$\int_{\mathbf{H}\in\mathfrak{U}^{\beta}(m)} q_{\kappa}(\mathbf{HLH}^{*})(d\mathbf{H}),$$

the result is follow from (Gross and Richards 1987, Equation 4.8(2) and Definition 5.3) and Faraut and Korányi (1994, Chapter XI, Section 3).

2. Is proved similarly. □

4. Generalised beta distributions: Beta-Riesz distributions

This section defines several versions for the beta functions and their relation with the gamma functions type I and II. In these terms, the beta-Riesz distributions type I and II are defined. Finally, diverse properties are studied.

4.1. Generalised c-beta function

A generalised of *multivariate beta function* for the cone $\mathfrak{P}_m^{\beta}$, denoted as $\mathcal{B}_m^{\beta}[a,\kappa;b,\tau]$, can be defined as

$$\int_{\mathbf{0}<\mathbf{S}<\mathbf{I}_m} |\mathbf{S}|^{a-(m-1)\beta/2-1} q_{\kappa}(\mathbf{S})|\mathbf{I}_m-\mathbf{S}|^{b-(m-1)\beta/2-1} q_{\tau}(\mathbf{I}_m-\mathbf{S})(d\mathbf{S}) \tag{14}$$

where $\kappa=(k_1,k_2,\ldots,k_m)\in\Re^m$, $\tau=(t_1,t_2,\ldots,t_m)\in\Re^m$, $\mathrm{Re}(a)>(m-1)\beta/2-k_m$ and $\mathrm{Re}(b)>(m-1)\beta/2-t_m$. This is defined by Faraut and Korányi (1994, p. 130) for Euclidean simple Jordan algebras. In the context of multivariate analysis, this generalised beta function can be termed *generalised c-beta function type I*, as analogy to the correspondence case of matrix multivariate beta distribution, and using the term c-beta as abbreviation of classical-beta. In the next theorem we introduce the *generalised c-beta function type II* and its relation with the generalised gamma function.

Theorem 4.1 *The generalised c-beta function type I can be expressed as*

$$\int_{\mathbf{R}\in\mathfrak{P}_m^{\beta}} |\mathbf{R}|^{a-(m-1)\beta/2-1} q_{\kappa}(\mathbf{R})|\mathbf{I}_m+\mathbf{R}|^{-(a+b)} q_{-(\kappa+\tau)}(\mathbf{I}_m+\mathbf{R})(d\mathbf{R})$$

$$= \frac{\Gamma_m^{\beta}[a,\kappa]\Gamma_m^{\beta}[b,\tau]}{\Gamma_m^{\beta}[a+b,\kappa+\tau]},$$

where $\kappa=(k_1,k_2,\ldots,k_m)\in\Re^m$, $\tau=(t_1,t_2,\ldots,t_m)\in\Re^m$, $Re(a)>(m-1)\beta/2-k_m$ *and* $Re(b)>(m-1)\beta/2-t_m$. *The integral expression is termed* generalised c-beta function type II.

Proof. Let $\mathcal{U}(\mathbf{I}_m-\mathbf{S})^{*}\mathcal{U}(\mathbf{I}_m-\mathbf{S})=(\mathbf{I}_m-\mathbf{S})$ the Cholesky decomposition of $(\mathbf{I}_m-\mathbf{S})$ where $\mathcal{U}(\mathbf{I}_m-\mathbf{S})\in\mathfrak{T}_U^{\beta}(m)$ and define $\mathbf{R}=\mathcal{U}(\mathbf{I}_m-\mathbf{S})^{*-1}\mathbf{S}\,\mathcal{U}(\mathbf{I}_m-\mathbf{S})^{-1}$ then

$$\begin{aligned}\mathbf{R} &= \mathcal{U}(\mathbf{I}_m-\mathbf{S})^{*-1}(\mathbf{I}_m-(\mathbf{I}_m-\mathbf{S}))\,\mathcal{U}(\mathbf{I}_m-\mathbf{S})^{-1}\\ &= \mathcal{U}(\mathbf{I}_m-\mathbf{S})^{*-1}\mathcal{U}(\mathbf{I}_m-\mathbf{S})^{-1}-\mathbf{I}_m\end{aligned}$$

Thus $(\mathbf{I}_m+\mathbf{R})=\mathcal{U}(\mathbf{I}_m-\mathbf{S})^{*-1}\mathcal{U}(\mathbf{I}_m-\mathbf{S})^{-1}$. By Lemma 2.3

$$\begin{aligned}(d\mathbf{R}) &= |\mathcal{U}(\mathbf{I}_m-\mathbf{S})^{*-1}\mathcal{U}(\mathbf{I}_m-\mathbf{S})^{-1}|^{(m-1)\beta+2}(d\mathbf{S})\\ &= |\mathbf{I}_m+\mathbf{R}|^{(m-1)\beta+2}(d\mathbf{S}),\end{aligned}$$

therefore $(d\mathbf{S})=(\mathbf{I}+\mathbf{R})^{-(m-1)\beta-2}(d\mathbf{R})$. Now remember that $q_{\kappa}(\mathbf{T}^{*-1}\mathbf{A}\mathbf{T}^{-1})=q_{\kappa}(\mathbf{A})q_{-\kappa}(\mathbf{B})=q_{\kappa}(\mathbf{A})q_{\kappa}^{-1}(\mathbf{B})$ for $\mathbf{B}=\mathbf{T}^{*}\mathbf{T}$, we have

$$|\mathbf{R}|^{a-(m-1)\beta/2-1} q_{\kappa}(\mathbf{R}) = |\mathbf{I}_m-\mathbf{S}|^{-a+(m-1)\beta/2+1}|\mathbf{S}|^{a-(m-1)\beta/2-1} q_{\kappa}(\mathbf{S})q_{-\kappa}(\mathbf{I}_m-\mathbf{S})$$

and

$$|\mathbf{I}_m+\mathbf{R}|^{b-(m-1)\beta/2-1}q_\tau(\mathbf{I}+\mathbf{R})=|\mathbf{I}_m-\mathbf{S}|^{-b+(m-1)\beta/2+1}q_{-\kappa}(\mathbf{I}_m-\mathbf{S}),$$

then

$$|\mathbf{I}_m+\mathbf{R}|^{-b+(m-1)\beta/2+1}q_{-\tau}(\mathbf{I}+\mathbf{R})=|\mathbf{I}_m-\mathbf{S}|^{b-(m-1)\beta/2-1}q_{\kappa}(\mathbf{I}_m-\mathbf{S}).$$

from where the desired result is obtained.

For the expression in terms of generalised gamma function, let $\mathbf{B}=\mathcal{U}(\boldsymbol{\Xi})^*\mathbf{S}\mathcal{U}(\boldsymbol{\Xi})$ in (14), such that $\boldsymbol{\Xi}=\mathcal{U}(\boldsymbol{\Xi})^*\mathcal{U}(\boldsymbol{\Xi})$. Then $(d\mathbf{S})=|\boldsymbol{\Xi}|^{-(m-1)\beta/2-1}(d\mathbf{B})$, and

$$\begin{aligned}&\mathcal{B}_m^\beta[a,\kappa;b,\tau]|\boldsymbol{\Xi}|^{a+b-(m-1)\beta/2-1}q_{\kappa+\tau}(\boldsymbol{\Xi})\\&\qquad=\int_{\mathbf{0}}^{\boldsymbol{\Xi}}|\mathbf{B}|^{a-(m-1)\beta/2-1}q_\kappa(\mathbf{B})|\boldsymbol{\Xi}-\mathbf{B}|^{b-(m-1)\beta/2-1}q_\tau(\boldsymbol{\Xi}-\mathbf{B})(d\mathbf{B}).\end{aligned}$$

Taking Laplace transform of both size, by (7), the left size is

$$\begin{aligned}&\int_{\boldsymbol{\Xi}\in\mathfrak{P}_m^\beta}\mathcal{B}_m^\beta[a,\kappa;b,\tau]\,\mathrm{etr}\{-\boldsymbol{\Xi}\mathbf{Z}\}|\boldsymbol{\Xi}|^{a+b-(m-1)\beta/2-1}q_{\kappa+\tau}(\boldsymbol{\Xi})(d\boldsymbol{\Xi})\\&\qquad=\mathcal{B}_m^\beta[a,\kappa;b,\tau]\Gamma_m^\beta[a+b;\kappa+\tau]|\mathbf{Z}|^{-(a+b)}q_{\kappa+\tau}(\mathbf{Z}^{-1}),\end{aligned}$$

and applying Lemma 2.5, $g_1(\mathbf{Z})$ is

$$\int_{\boldsymbol{\Xi}\in\mathfrak{P}_m^\beta}\mathrm{etr}\{-\boldsymbol{\Xi}\mathbf{Z}\}|\boldsymbol{\Xi}|^{a-(m-1)\beta/2-1}q_\kappa(\boldsymbol{\Xi})(d\boldsymbol{\Xi})=\Gamma_m^\beta[a;\kappa]|\mathbf{Z}|^{-a}q_\kappa(\mathbf{Z}^{-1}),$$

and $g_2(\mathbf{Z})$ is given by

$$\int_{\mathbf{B}\in\mathfrak{P}_m^\beta}\mathrm{etr}\{-\mathbf{B}\mathbf{Z}\}|\mathbf{B}|^{b-(m-1)\beta/2-1}q_\kappa(\mathbf{B})(d\mathbf{B})=\Gamma_m^\beta[b;\tau]|\mathbf{Z}|^{-b}q_\tau(\mathbf{Z}^{-1}).$$

Thus, equally

$$\mathcal{B}_m^\beta[a,\kappa;b,\tau]=\frac{\Gamma_m^\beta[a,\kappa]\Gamma_m^\beta[b,\tau]}{\Gamma_m^\beta[a+b,\kappa+\tau]}.$$

□

4.2. Generalised k-beta function

Alternatively, a generalised of *multivariate beta function* for the cone $\mathfrak{P}_m^\beta$, can be defined and denoted as

$$\mathcal{B}_m^\beta[a,-\kappa;b,-\tau]=\int_{\mathbf{0}<\mathbf{S}<\mathbf{I}_m}|\mathbf{S}|^{a-(m-1)\beta/2-1}q_\kappa\left(\mathbf{S}^{-1}\right)|\mathbf{I}_m-\mathbf{S}|^{b-(m-1)\beta/2-1}q_\tau\left((\mathbf{I}_m-\mathbf{S})^{-1}\right)(d\mathbf{S}) \tag{15}$$

where $\kappa=(k_1,k_2,\ldots,k_m)\in\Re^m$, $\tau=(t_1,t_2,\ldots,t_m)\in\Re^m$, $\mathrm{Re}(a)>(m-1)\beta/2+k_1$ and $\mathrm{Re}(b)>(m-1)\beta/2+t_1$. Again, in the context of multivariate analysis, this generalised k-beta function can be termed *generalised k-beta function type I*, as an analogy to the corresponding case of matrix multivariate beta distribution and using the term k-beta as abbreviation of Khatri-beta. Next theorem introduces the *generalised k-beta function type II* and its relation with the generalised gamma function proposed by Khatri (1966).

Theorem 4.2 *The generalised k-beta function type II can be expressed as*

$$\int_{\mathbf{R}\in\mathfrak{P}_m^\beta}|\mathbf{R}|^{a-(m-1)\beta/2-1}q_\kappa(\mathbf{R}^{-1})|\mathbf{I}_m+\mathbf{R}|^{-(a+b)}q_{-(\kappa+\tau)}\left((\mathbf{I}_m+\mathbf{R})^{-1}\right)(d\mathbf{R})$$

$$= \frac{\Gamma_m^\beta[a,-\kappa]\Gamma_m^\beta[b,-\tau]}{\Gamma_m^\beta[a+b,-\kappa-\tau]},$$

where $\kappa = (k_1, k_2, \ldots, k_m) \in \Re^m$, $\tau = (t_1, t_2, \ldots, t_m) \in \Re^m$, $Re(a) > (m-1)\beta/2 + k_1$ *and* $Re(b) > (m-1)\beta/2 + t_1$. *The integral expression is termed* generalised k-beta function type II.

Proof. The proof is analogous to the given for Theorem 4.1. □

Observe that if $\kappa = (0, \ldots, 0) \in \Re^m$ and $\tau = (0, \ldots, 0) \in \Re^m$ in (14), Theorem 4.1, (15) and Theorem 4.2 the classical beta function is obtained, see Herz (1955).

4.3. c-beta-Riesz and k-beta-Riesz distributions

As an immediate consequence of the results of the previous section, next the c-beta-Riesz and k-beta-Riesz distributions types I and II are defined.

Definition 4.1 *Let* $\kappa = (k_1, k_2, \ldots, k_m) \in \Re^m$ *and* $\tau = (t_1, t_2, \ldots, t_m) \in \Re^m$.

1. *Then it said that* $\mathbf{S}$ *has a* c-beta-Riesz distribution of type I *if its density function is*

$$\frac{1}{\mathcal{B}_m^\beta[a,\kappa;b,\tau]}|\mathbf{S}|^{a-(m-1)\beta/2-1}q_\kappa(\mathbf{S})|\mathbf{I}_m-\mathbf{S}|^{b-(m-1)\beta/2-1}q_\tau(\mathbf{I}_m-\mathbf{S})(d\mathbf{S}), \tag{16}$$

where $\mathbf{0} < \mathbf{S} < \mathbf{I}_m$ *and* $Re(a) > (m-1)\beta/2 - k_m$ *and* $Re(b) > (m-1)\beta/2 - t_m$.

2. *Then it said that* $\mathbf{R}$ *has a* c-beta-Riesz distribution of type II *if its density function is*

$$\frac{1}{\mathcal{B}_m^\beta[a,\kappa;b,\tau]}|\mathbf{R}|^{a-(m-1)\beta/2-1}q_\kappa(\mathbf{R})|\mathbf{I}_m+\mathbf{R}|^{-(a+b)}q_{-(\kappa+\tau)}(\mathbf{I}_m+\mathbf{R})(d\mathbf{R}), \tag{17}$$

where $\mathbf{R} \in \mathfrak{P}_m^\beta$ *and* $Re(a) > (m-1)\beta/2 - k_m$ *and* $Re(b) > (m-1)\beta/2 - t_m$.

Similarly we have

Definition 4.2 *Let* $\kappa = (k_1, k_2, \ldots, k_m) \in \Re^m$ *and* $\tau = (t_1, t_2, \ldots, t_m) \in \Re^m$.

1. *Then it said that* $\mathbf{S}$ *has a* k-beta-Riesz distribution of type I *if its density function is*

$$\frac{1}{\mathcal{B}_m^\beta[a,-\kappa;b,-\tau]}|\mathbf{S}|^{a-(m-1)\beta/2-1}q_\kappa(\mathbf{S}^{-1})|\mathbf{I}_m-\mathbf{S}|^{b-(m-1)\beta/2-1}q_\tau\left((\mathbf{I}_m-\mathbf{S})^{-1}\right)(d\mathbf{S}), \tag{18}$$

where $\mathbf{0} < \mathbf{S} < \mathbf{I}_m$ *and* $Re(a) > (m-1)\beta/2 + k_1$ *and* $Re(b) > (m-1)\beta/2 + t_1$.

2. *Then it said that* $\mathbf{R}$ *has a* k-beta-Riesz distribution of type II *if its density function is*

$$\frac{1}{\mathcal{B}_m^\beta[a,-\kappa;b,-\tau]}|\mathbf{R}|^{a-(m-1)\beta/2-1}q_\kappa(\mathbf{R}^{-1})|\mathbf{I}_m+\mathbf{R}|^{-(a+b)}q_{-(\kappa+\tau)}\left((\mathbf{I}_m+\mathbf{R})^{-1}\right)(d\mathbf{R}), \tag{19}$$

where $\mathbf{R} \in \mathfrak{P}_m^\beta$ *and* $Re(a) > (m-1)\beta/2 + k_1$ *and* $Re(b) > (m-1)\beta/2 + t_1$.

Observe that the relationship between the densities (16) and (17), and between the densities (18) and (19) are easily obtained from the theorems 4.1 and 4.2, respectively.

The following result state the relation between the Riesz and beta-Riesz distributions.

Theorem 4.3 *Let* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *be independently distributed as Riesz distribution type I, such that* $\mathbf{X}_1 \sim \mathfrak{R}_m^{\beta,I}(a,\kappa,\mathbf{\Sigma})$ *and* $\mathbf{X}_2 \sim \mathfrak{R}_m^{\beta,I}(b,\tau,\mathbf{\Sigma})$, $Re(a) > (m-1)\beta/2 + k_1$ *and* $Re(b) > (m-1)\beta/2 + t_1$. *Let*

$$\mathbf{S} = \mathcal{U}(\mathbf{X}_1+\mathbf{X}_2)^{*-1}\mathbf{X}_1\mathcal{U}(\mathbf{X}_1+\mathbf{X}_2)^{-1},$$

where $\mathcal{U}(\mathbf{X}_1+\mathbf{X}_2) \in \mathfrak{T}_U^{\beta}(m)$ *is such that* $(\mathbf{X}_1+\mathbf{X}_2) = \mathcal{U}(\mathbf{X}_1+\mathbf{X}_2)^*\mathcal{U}(\mathbf{X}_1+\mathbf{X}_2)$. *Then* $\mathbf{S}$ *and* $(\mathbf{X}_1+\mathbf{X}_2)$ *are independent,* $\mathbf{S}$ *has a c-beta-Riesz distribution type I and* $(\mathbf{X}_1+\mathbf{X}_2) \sim \mathfrak{R}_m^{\beta,I}(a+b,\kappa+\tau,\mathbf{I}_m)$.

Proof. The joint density of $\mathbf{X}_1$ and $\mathbf{X}_2$ is given by

$$\frac{\beta^{(a+b)m+\sum_{i=1}^m(k_i+t_i)}}{\Gamma_m^{\beta}[a,\kappa]\Gamma_m^{\beta}[b,\tau]|\mathbf{\Sigma}|^{a+b}q_{\kappa+\tau}(\mathbf{\Sigma})}\operatorname{etr}\{-\beta\mathbf{\Sigma}^{-1}(\mathbf{X}_1+\mathbf{X}_2)\}|\mathbf{X}_1|^{a-(m-1)\beta/2-1}q_{\kappa}(\mathbf{X}_1)$$

$$\times|\mathbf{X}_2|^{b-(m-1)\beta/2-1}q_{\tau}(\mathbf{X}_2)(d\mathbf{X}_1)\wedge(d\mathbf{X}_2).$$

Let $\mathbf{Y} = \mathbf{X}_1+\mathbf{X}_2$ and $\mathbf{Z} = \mathbf{X}_1$, then, $(d\mathbf{X}_1)\wedge(d\mathbf{X}_2) = (d\mathbf{Y})\wedge(d\mathbf{Z})$. Then the joint density of $\mathbf{Y}$ and $\mathbf{Z}$ is given by

$$\frac{\beta^{(a+b)m+\sum_{i=1}^m(k_i+t_i)}}{\Gamma_m^{\beta}[a,\kappa]\Gamma_m^{\beta}[b,\tau]|\mathbf{\Sigma}|^{a+b}q_{\kappa+\tau}(\mathbf{\Sigma})}\operatorname{etr}\{-\beta\mathbf{\Sigma}^{-1}\mathbf{Y}\}|\mathbf{Z}|^{a-(m-1)\beta/2-1}q_{\kappa}(\mathbf{Z})$$

$$\times|\mathbf{Y}-\mathbf{Z}|^{b-(m-1)\beta/2-1}q_{\tau}(\mathbf{Y}-\mathbf{Z})(d\mathbf{Y})\wedge(d\mathbf{Z}).$$

Let $\mathbf{W} = \mathcal{U}(\mathbf{Y})^*\mathcal{U}(\mathbf{Y})$, with $\mathcal{U}(\mathbf{Y}) \in \mathfrak{T}_U^{\beta}(m)$ and $\mathbf{Z} = \mathcal{U}(\mathbf{Y})^*\mathbf{S}\mathcal{U}(\mathbf{Y})$. Observing that $\mathcal{U}(\mathbf{Y})$ is a function of $\mathbf{W}$

$$(d\mathbf{Y})\wedge(d\mathbf{Z}) = |\mathcal{U}(\mathbf{Y})^*\mathcal{U}(\mathbf{Y})|^{\beta(m-1)/2+1}(d\mathcal{U}(\mathbf{Y})^*\mathcal{U}(\mathbf{Y}))\wedge(d\mathbf{S})$$

Hence the joint density of $\mathbf{S}$ and $\mathbf{W} = \mathcal{U}(\mathbf{Y})^*\mathcal{U}(\mathbf{Y})$ is

$$\frac{\beta^{(a+b)m+\sum_{i=1}^m(k_i+t_i)}}{\Gamma_m^{\beta}[a+b,\kappa+\tau]|\mathbf{\Sigma}|^{a+b}q_{\kappa+\tau}(\mathbf{\Sigma})}\operatorname{etr}\{-\beta\mathbf{\Sigma}^{-1}\mathcal{U}(\mathbf{Y})^*\mathcal{U}(\mathbf{Y})\}|\mathcal{U}(\mathbf{Y})^*\mathcal{U}(\mathbf{Y})|^{a+b-\beta(m-1)/2-1}$$

$$q_{\kappa+\tau}(\mathcal{U}(\mathbf{Y})^*\mathcal{U}(\mathbf{Y}))(d\mathcal{U}(\mathbf{Y})^*\mathcal{U}(\mathbf{Y}))$$

$$\times\frac{\Gamma_m^{\beta}[a+b,\kappa+\tau]}{\Gamma_m^{\beta}[a,\kappa]\Gamma_m^{\beta}[b,\tau]}|\mathbf{S}|^{a-(m-1)\beta/2-1}q_{\kappa}(\mathbf{S})|\mathbf{I}-\mathbf{S}|^{b-(m-1)\beta/2-1}q_{\tau}(\mathbf{I}-\mathbf{S})(d\mathbf{S}).$$

which shows that $\mathbf{W} = \mathcal{U}(\mathbf{Y})^*\mathcal{U}(\mathbf{Y}) = \mathbf{X}_1+\mathbf{X}_2 \sim \mathfrak{R}_m^{\beta,I}(a+b,\kappa+\tau,\mathbf{\Sigma})$ independently of $\mathbf{S}$ with a *c*-beta-Riesz distribution type I. □

Theorem 4.4 *Let* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *be independently distributed as Riesz distribution type I, such that* $\mathbf{X}_1 \sim \mathfrak{R}_m^{\beta,I}(a,\kappa,\mathbf{\Sigma})$ *and* $\mathbf{X}_2 \sim \mathfrak{R}_m^{\beta,I}(b,\tau,\mathbf{\Sigma})$, $Re(a) > (m-1)\beta/2 + k_1$ *and* $Re(b) > (m-1)\beta/2 + t_1$. *Let*

$$\mathbf{R} = \mathcal{U}(\mathbf{X}_2)^{*-1}\mathbf{X}_1\mathcal{U}(\mathbf{X}_2)^{-1},$$

where $\mathcal{U}(\mathbf{X}_2) \in \mathfrak{T}_U^{\beta}(m)$ *is such that* $\mathbf{X}_1 = \mathcal{U}(\mathbf{X}_2)^*\mathcal{U}(\mathbf{X}_2)$. *Then* $\mathbf{S}$ *has a c-beta-Riesz distribution type II.*

Proof. From Theorem 4.1 we know that if $\mathbf{S}$ has a *c*-beta-Riesz distribution type I then $\mathbf{R} = \mathcal{U}(\mathbf{I}_m-\mathbf{S})^{*-1}\mathbf{S}\mathcal{U}(\mathbf{I}_m-\mathbf{S})^{-1}$ has a *c*-beta-Riesz distribution type II. In addition the theorem establish that if $\mathbf{X}_1$ and $\mathbf{X}_2$ be independently distributed as Riesz distribution type I, such that $\mathbf{X}_1 \sim \mathfrak{R}_m^{\beta,I}(a,\kappa,\mathbf{\Sigma})$ and $\mathbf{X}_2 \sim \mathfrak{R}_m^{\beta,I}(b,\tau,\mathbf{\Sigma})$ then $\mathbf{R} = \mathcal{U}(\mathbf{X}_1)^{*-1}\mathbf{X}_1\mathcal{U}(\mathbf{X}_2)^{-1}$. Thus, the desired result is follow if we proof that

$$\mathbf{R} = \mathcal{U}(\mathbf{I}_m-\mathbf{S})^{*-1}\mathbf{S}\mathcal{U}(\mathbf{I}_m-\mathbf{S})^{-1} = \mathcal{U}(\mathbf{X}_2)^{*-1}\mathbf{X}_1\mathcal{U}(\mathbf{X}_2)^{-1}.$$

With this aim in mind, let $\mathcal{U}(\mathbf{Y}) \in \mathfrak{T}_U^{\beta}(m)$, such that $\mathbf{X}_1 + \mathbf{X}_2 = \mathcal{U}(\mathbf{Y})^* \mathcal{U}(\mathbf{Y})$, then $\mathbf{S} = \mathcal{U}(\mathbf{Y})^{*-1}\mathbf{X}_1\mathcal{U}(\mathbf{Y})^{-1}$. Now, if $\mathbf{X}_2 = \mathcal{U}(\mathbf{X}_2)^*\mathcal{U}(\mathbf{X}_2)$, with $\mathcal{U}(\mathbf{X}_2) \in \mathfrak{T}_U^{\beta}(m)$. We have

$$\begin{aligned}
\mathbf{I}_m - \mathbf{S} &= \mathbf{I}_m - \mathcal{U}(\mathbf{Y})^{*-1}\mathbf{X}_1\mathcal{U}(\mathbf{Y})^{-1} \\
&= \mathcal{U}(\mathbf{Y})^{*-1}(\mathcal{U}(\mathbf{Y})^*\mathcal{U}(\mathbf{Y}) - \mathbf{X}_1)\mathcal{U}(\mathbf{Y})^{-1} \\
&= \mathcal{U}(\mathbf{Y})^{*-1}\mathbf{X}_2\mathcal{U}(\mathbf{Y})^{-1} \\
&= \mathcal{U}(\mathbf{Y})^{*-1}\mathcal{U}(\mathbf{X}_2)^*\mathcal{U}(\mathbf{X}_2)\mathcal{U}(\mathbf{Y})^{-1} \\
&= \left(\mathcal{U}(\mathbf{X}_2)\mathcal{U}(\mathbf{Y})^{-1}\right)^* \left(\mathcal{U}(\mathbf{X}_2)\mathcal{U}(\mathbf{Y})^{-1}\right) \\
&= \mathcal{U}(\mathbf{I}_m - \mathbf{S})^*\mathcal{U}(\mathbf{I}_m - \mathbf{S}).
\end{aligned}$$

This least equally is obtained observing that $\left(\mathcal{U}(\mathbf{X}_2)\mathcal{U}(\mathbf{Y})^{-1}\right) \in \mathfrak{T}_U^{\beta}(m)$, then $\left(\mathcal{U}(\mathbf{X}_2)\mathcal{U}(\mathbf{Y})^{-1}\right) = \mathcal{U}(\mathbf{I}_m - \mathbf{S})$. Therefore

$$\begin{aligned}
\mathcal{U}(\mathbf{I}_m - \mathbf{S})^{*-1}\mathbf{S}\mathcal{U}(\mathbf{I}_m - \mathbf{S})^{-1} &= \left(\mathcal{U}(\mathbf{X}_2)\mathcal{U}(\mathbf{Y})^{-1}\right)^{*-1}\mathbf{S}\left(\mathcal{U}(\mathbf{X}_2)\mathcal{U}(\mathbf{Y})^{-1}\right)^{-1} \\
&= \mathcal{U}(\mathbf{X}_2)^{*-1}\mathcal{U}(\mathbf{Y})^*\mathbf{S}\mathcal{U}(\mathbf{Y})\mathcal{U}(\mathbf{X}_2)^{-1} \\
&= \mathcal{U}(\mathbf{X}_2)^{*-1}\mathbf{X}_1\mathcal{U}(\mathbf{X}_2)^{-1}.
\end{aligned}$$

From where the desired result is obtained. □

The following theorems 4.5 and 4.6 contain versions for k-beta-Riesz distributions of theorems 4.3 and 4.4, whose proofs are similar.

Theorem 4.5 *Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be independently distributed as Riesz distribution type II, such that $\mathbf{X}_1 \sim \mathfrak{R}_m^{\beta,II}(a, \kappa, \mathbf{\Sigma})$ and $\mathbf{X}_2 \sim \mathfrak{R}_m^{\beta,II}(b, \tau, \mathbf{\Sigma})$, $Re(a) > (m-1)\beta/2 + k_1$ and $Re(b) > (m-1)\beta/2 + t_1$. Let*

$$\mathbf{S} = \mathcal{U}(\mathbf{X}_1 + \mathbf{X}_2)^{*-1}\mathbf{X}_1\mathcal{U}(\mathbf{X}_1 + \mathbf{X}_2)^{-1},$$

where $\mathcal{U}(\mathbf{X}_1 + \mathbf{X}_2) \in \mathfrak{T}_U^{\beta}(m)$ is such that $(\mathbf{X}_1 + \mathbf{X}_2) = \mathcal{U}(\mathbf{X}_1 + \mathbf{X}_2)^*\mathcal{U}(\mathbf{X}_1 + \mathbf{X}_2)$. Then $\mathbf{S}$ has a k-beta-Riesz distribution type I.

Theorem 4.6 *Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be independently distributed as Riesz distribution type I, such that $\mathbf{X}_1 \sim \mathfrak{R}_m^{\beta,II}(a, \kappa, \mathbf{\Sigma})$ and $\mathbf{X}_2 \sim \mathfrak{R}_m^{\beta,II}(b, \tau, \mathbf{\Sigma})$, $Re(a) > (m-1)\beta/2 + k_1$ and $Re(b) > (m-1)\beta/2 + t_1$. Let*

$$\mathbf{R} = \mathcal{U}(\mathbf{X}_2)^{*-1}\mathbf{X}_1\mathcal{U}(\mathbf{X}_2)^{-1},$$

where $\mathcal{U}(\mathbf{X}_2) \in \mathfrak{T}_U^{\beta}(m)$ is such that $\mathbf{X}_2 = \mathcal{U}(\mathbf{X}_1)^*\mathcal{U}(\mathbf{X}_1)$. Then $\mathbf{S}$ has a k-beta-Riesz distribution type II.

4.4. Some properties of the c-beta-Riesz and k-beta-Riesz distributions

This section derives the distributions of eigenvalues for c-beta-Riesz and k-beta-Riesz distributions type I and II. First consider the following integrals:

$$Q(\kappa, \tau, \mathbf{A}, \mathbf{B}) = \int_{\mathfrak{U}^{\beta}(m)} q_{\kappa}(\mathbf{HAH}^*) q_{\tau}\left(\mathbf{HBH}^*\right)(d\mathbf{H})$$

and

$$Q_1(\kappa, \tau, \mathbf{A}, \mathbf{B}) = \int_{\mathfrak{U}^{\beta}(m)} q_{\kappa}(\mathbf{HAH}^*) q_{-(\kappa+\tau)}\left(\mathbf{HBH}^*\right)(d\mathbf{H})$$

Theorem 4.7 *Let $\mathbf{\Sigma} \in \mathbf{\Phi}_m^{\beta}$, $\kappa = (k_1, k_2, \ldots, k_m)$, $k_1 \geq k_2 \geq \cdots \geq k_m \geq 0$, $k_1, k_2, \ldots, k_m$ are nonnegative integers and $\tau = (t_1, t_2, \ldots, t_m)$, $t_1 \geq t_2 \geq \cdots \geq t_m \geq 0$, $t_1, t_2, \ldots, t_m$ are nonnegative integers.*

1. *Let* $\mathbf{L} = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$, $\lambda_1 > \cdots > \lambda_m > 0$ *be the eigenvalues of* $\mathbf{S}$. *Then if* $\mathbf{S}$ *has a c-beta-Riesz distribution of type I, the joint density of* $\lambda_1, \ldots, \lambda_m$ *is*

$$\frac{\pi^{m^2\beta/2+\varrho}}{\Gamma_m^\beta[m\beta/2]\mathcal{B}_m^\beta[a,\kappa;b,\tau]} \prod_{i<j}^m (\lambda_i-\lambda_j)^\beta \prod_{i=1}^m \lambda_i^{a-(m-1)\beta/2-1} \prod_{i=1}^m (1-\lambda_i)^{b-(m-1)\beta/2-1}$$

$$Q(\kappa,\tau,\mathbf{L},\mathbf{I}_m-\mathbf{L}) \left(\bigwedge_{i=1}^m d\lambda_i\right),$$

where $0 < \lambda_i < 1$, $i = 1, \ldots, m$ *and* $Re(a) > (m-1)\beta/2 - k_m$ *and* $Re(b) > (m-1)\beta/2 - t_m$.

2. *Let* $\Delta = \operatorname{diag}(\delta_1, \ldots, \delta_m)$, $\delta_1 > \cdots > \delta_m > 0$ *be the eigenvalues of* $\mathbf{R}$. *Then if* $\mathbf{R}$ *has a c-beta-Riesz distribution of type II, the joint density of their eigenvalues is*

$$\frac{\pi^{m^2\beta/2+\varrho}}{\Gamma_m^\beta[m\beta/2]\mathcal{B}_m^\beta[a,\kappa;b,\tau]} \prod_{i<j}^m (\delta_i-\delta_j)^\beta \prod_{i=1}^m \delta_i^{a-(m-1)\beta/2-1} \prod_{i=1}^m (1-\delta_i)^{-(a+b)}$$

$$Q_1(\kappa,\tau,\mathbf{\Delta},(\mathbf{I}_m+\mathbf{\Delta})) \left(\bigwedge_{i=1}^m d\delta_i\right),$$

where $\delta_i > 0$, $i = 1, \ldots, m$ *and* $Re(a) > (m-1)\beta/2 - k_m$ *and* $Re(b) > (m-1)\beta/2 - t_m$.

ϱ *is defined in Lemma 2.4.*

Proof. This is due to applying the Lemma 2.4 in (16) and (17). □

This section conclude establishing the Theorem 4.7 for the case of the k-beta-Riesz distributions.

Theorem 4.8 *Let* $\mathbf{\Sigma} \in \mathbf{\Phi}_m^\beta$, $\kappa = (k_1, k_2, \ldots, k_m)$, $k_1 \geq k_2 \geq \cdots \geq k_m \geq 0$, $k_1, k_2, \ldots, k_m$ *are nonnegative integers and* $\tau = (t_1, t_2, \ldots, t_m)$, $t_1 \geq t_2 \geq \cdots \geq t_m \geq 0$, $t_1, t_2, \ldots, t_m$ *are nonnegative integers.*

1. *Let* $\mathbf{L} = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$, $\lambda_1 > \cdots > \lambda_m > 0$ *be the eigenvalues of* $\mathbf{S}$. *Then if* $\mathbf{S}$ *has a k-beta-Riesz distribution of type I, the joint density of* $\lambda_1, \ldots, \lambda_m$ *is*

$$\frac{\pi^{m^2\beta/2+\varrho}}{\Gamma_m^\beta[m\beta/2]\mathcal{B}_m^\beta[a,-\kappa;b,-\tau]} \prod_{i<j}^m (\lambda_i-\lambda_j)^\beta \prod_{i=1}^m \lambda_i^{a-(m-1)\beta/2-1} \prod_{i=1}^m (1-\lambda_i)^{b-(m-1)\beta/2-1}$$

$$Q\left(\kappa,\tau,\mathbf{L}^{-1},(\mathbf{I}_m-\mathbf{L})^{-1}\right) \left(\bigwedge_{i=1}^m d\lambda_i\right),$$

where $0 < \lambda_i < 1$, $i = 1, \ldots, m$ *and* $Re(a) > (m-1)\beta/2 + k_1$ *and* $Re(b) > (m-1)\beta/2 + t_1$.

2. *Let* $\Delta = \operatorname{diag}(\delta_1, \ldots, \delta_m)$, $\delta_1 > \cdots > \delta_m > 0$ *be the eigenvalues of* $\mathbf{R}$. *Then if* $\mathbf{R}$ *has a k-beta-Riesz distribution of type II, the joint density of their eigenvalues is*

$$\frac{\pi^{m^2\beta/2+\varrho}}{\Gamma_m^\beta[m\beta/2]\mathcal{B}_m^\beta[a,-\kappa;b,-\tau]} \prod_{i<j}^m (\delta_i-\delta_j)^\beta \prod_{i=1}^m \delta_i^{a-(m-1)\beta/2-1} \prod_{i=1}^m (1-\delta_i)^{-(a+b)}$$

$$Q_1\left(\kappa,\tau,\mathbf{\Delta}^{-1},(\mathbf{I}_m+\mathbf{\Delta})^{-1}\right) \left(\bigwedge_{i=1}^m d\delta_i\right),$$

where $\delta_i > 0$, $i = 1, \ldots, m$ *and* $Re(a) > (m-1)\beta/2 + k_1$ *and* $Re(b) > (m-1)\beta/2 + t_1$.

ϱ is defined in Lemma 2.4.

Finally observe that if in all result of this section are taking $\kappa = (0, 0, \ldots, 0) \in \Re^m$ and $\tau = (0, 0, \ldots, 0) \in \Re^m$ the obtained results are the corresponding to matrix multivariate beta distributions of type I and II.

Conclusions

Finally, note that the real dimension of real normed division algebras can be expressed as powers of 2, $\beta = 2^n$ for $n = 0, 1, 2, 3$. On the other hand, as observed from Kabe (1984), the results obtained in this work can be extended to hypercomplex cases; that is, for complex, bicomplex, biquaternion and bioctonion (or sedenionic) algebras, which of course are not division algebras (except the complex algebra). Also note, that hypercomplex algebras are obtained by replacing the real numbers with complex numbers in the construction of real normed division algebras. Thus, the results for hypercomplex algebras are obtained by simply replacing β with 2β in our results. Alternatively, following Kabe (1984), it can be concluded that, results are true for '2^n-ions', $n = 0, 1, 2, 3, 4, 5$, emphasising that only for $n = 0, 1, 2, 3$ are the result algebras, in fact, real normed division algebras.

Acknowledgements

The author wish to thank the Editor and the anonymous reviewers for their constructive comments on the preliminary version of this paper. This article was written under the existing research agreement between the first author and the Universidad Autónoma Agraria Antonio Narro, Saltillo, México.

References

Baez JC (2002). "The Octonions." *Bull. Amer. Math. Soc.*, **39**, 145 – 205.

Boutouria I, Hassairi A (2009). "Riesz Exponential Families on Homogeneous cones." *http://arxiv.org/abs/0906.1892*.

Casalis M, Letac G (1996). "The Lukacs-Olkin-Rubin Characterization of Wishart Distributions on Symmetric Cones." *Ann. Statist.*, **24**, 768–786.

Díaz-García JA (2014). "Integral Properties of Zonal Spherical Functions, Hypergeometric Functions and Invariant Polynomials." *J. Iranian Statist. Soc.*, **13(1)**, 83 – 124.

Díaz-García JA (2015a). "Distributions on Symmetric Cones I: Riesz Distribution." *http://arxiv.org/abs/1211.1746v2*.

Díaz-García JA (2015b). "A Generalised Kotz Type Distribution and Riesz Distribution." *http://arxiv.org/abs/1304.5292v2*.

Díaz-García JA, Gutiérrez-Jáimez R (2011). "On Wishart Distribution: Some Extensions." *Linear Algebra Appl.*, **435**, 1296 – 1310.

Díaz-García JA, Gutiérrez-Jáimez R (2013). "Spherical Ensembles." *Linear Algebra Appl.*, **438**, 3174 – 3201.

Dumitriu I (2002). *Eigenvalue Statistics for Beta-Ensembles*. Massachusetts Institute of Technology, Cambridge, MA.

Edelman A, Rao RR (2005). "Random Matrix Theory." *Acta Numer.*, **14**, 233 – 297.

Fang KT, Zhang YT (1990). *Generalized Multivariate Analysis*. Science Press, Beijing, Springer-Verlang.

Faraut J, Korányi A (1994). *Analysis on Symmetric Cones*. Clarendon Press, Oxford.

Forrester PJ (2005). *Log-Gases and Random Matrices*. http://www.ms.unimelb.edu.au/~matpjf/matpjf.html.

Gross KI, Richards DSP (1987). "Special Functions of Matrix Argument I: Algebraic Induction Zonal Polynomials and Hypergeometric Functions." *Trans. Amer. Math. Soc.*, **301**, 478 – 501.

Gupta AK, Varga T (1993). *Elliptically Contoured Models in Statistics*. Kluwer Academic Publishers, Dordrecht.

Hassairi A, Lajmi S (2001). "Riesz Exponential Families on Symmetric Cones." *J. Theoret. Probab.*, **14**, 927 – 948.

Hassairi A, Lajmi S, Zine R (2005). "Beta-Riesz Distributions on Symmetric Cones." *J. Satatist. Plan. Inference*, **133**, 387 – 404.

Herz CS (1955). "Bessel Functions of Matrix Argument." *Ann. of Math.*, **61(3)**, 474 – 523.

Ishi H (2000). "Positive Riesz Distributions on Homogeneous Cones." *J. Math. Soc. Japan*, **52(1)**, 161 – 186.

Kabe DG (1984). "Classical Statistical Analysis Based on a Certain Hypercomplex Multivariate Normal Distribution." *Metrika*, **31**, 63 – 76.

Khatri CG (1966). "On Certain Distribution Problems Based on Positive Definite Quadratic Functions in Normal Vector." *Ann. Math. Statist.*, **37**, 468 – 479.

Massam H (1994). "An Exact Decomposition Theorem and Unified View of Some Related Distributions for a Class of Exponential Transformation Models on Symmetric Cones." *Ann. Statist.*, **22(1)**, 369–394.

Neukirch J, Prestel A, Remmert R (1990). *Numbers*. Springer, New York.

Sawyer P (1997). "Spherical Functions on Symmetric Cones." *Trans. Amer. Math. Soc.*, **349**, 3569 – 3584.

Tounsi M, Zine R (2012). "The Inverse Riesz Probability Distribution on Symmetric Matrices." *J. Multivar. Anal.*, **111**, 174 – 182.

Affiliation:

José A. Díaz-García
Universidad Autónoma Agraria Antonio Narro
Calzada Antonio Narro 1923
Col. Buenavista
25315 Saltillo, Coahuila, México
E-mail: jadiaz@uaaan.mx

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 45 *Submitted:* 2014-08-16
March 2016 *Accepted:* 2015-11-01

Austrian Journal of Statistics
March 2016, Volume 45, 53–67.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i2.121

Life Expectancy Comparison between a Study Cohort and a Reference Population

Georg Zimmermann
Department of Neurology,Christian Doppler Klinik, Salzburg
Paracelsus Medical University, Salzburg
Paris Lodron University of Salzburg

Abstract

Questions of life expectancy are highly relevant in clinical practice, in particular if one is interested in the life expectancy loss due to a certain disease in comparison with the general population. Therefore, we propose a methodology which enables the researcher to compare more easily life expectancies between a study cohort and a reference population. Firstly, we take the formula commonly used in official statistics and adapt it to a survival analytic setting, thus establishing a link between official statistics and survival analysis. Further, we discuss two commonly encountered sources of potentially severe bias, and how to remedy them. We hope that the proposed approach facilitates the use of life tables, particularly in clinical studies with mortality as primary endpoint.

Keywords: life expectancy, life table, life expectancy comparison, survival analysis, expected remaining life, Larynx dataset.

1. Introduction

In medical studies, the survival experience of a patient cohort is often expressed in terms of some risk measure, for example, the risk of dying. However, sometimes, this type of quantities is not completely satisfying: A patient suffering from a certain disease is probably more interested in his or her loss of remaining lifetime than the mere information that he or she has a risk of dying elevated by some factor compared to healthy persons, because the first quantity is much easier and more naturally interpretable than the latter one. Consequently, the doctors should also be able to provide their patients with information about life expectancies. But interestingly, it seems as if the number of studies merely examining some risk measures such as standardized mortality ratios or hazard ratios is much higher than the number of life expectancy analyses. In particular, life expectancy comparisons between a patient group and some reference population are hardly available, although this kind of question is a very natural and highly relevant one.

There are basically two measures of life expectancy (for another concept of "life expectancy", see Bradshaw, Stobie, Knuiman, Briffa, and Hobbs 2015): The so-called *mean residual life* is used in survival analysis, a certain branch of mathematical statistics (Kalbfleisch and Pren-

tice 2002), whereas the *average expectation of life* or *life expectancy at exact age* x is popular in the context of life table calculations in offical statistics (Chiang 1979). The latter measure can also be used in analyses of study cohorts, which is a quite popular method in life expectancy analyses (Nusselder, Slockers, Krol, Slockers, Looman, and van Beeck 2013; DuGoff, Canudas-Romo, Buttorff, Leff, and Anderson 2014; Guaraldi, Cossarizza, Franceschi, Roverato, Vaccher, Tambussi, Garlassi, Menozzi, Mussini, and D'Arminio Monforte 2014; Nagai, Kuriyama, Kakizaki, Ohmori-Matsuda, Sone, Hozawa, Kawado, Hashimoto, and Tsuji 2011). On the other hand, survival analytic quantities such as Kaplan meier estimators are sometimes used, too (Chang, Lu, Lee, Hwang, Cheng, and Wang 2015; Luangasanatip, Hongsuwan, Lubell, Limmathurotsakul, Teparrukkul, Chaowarat, Day, Graves, and Cooper 2013). Moreover, there are also "mixtures" between these two concepts insofar as, at first, a survival analytic model is fitted to the data and, then, life expectancies are calculated based on the life table methodology (Li, Hüsing, and Kaaks 2014; Strauss, DeVivo, Paculdo, and Shavelle 2006; Gaitatzis, Johnson, Chadwick, Shorvon, and Sander 2004). However, in some of these studies (Li *et al.* 2014; Nagai *et al.* 2011; Strauss *et al.* 2006), life expectancy comparisons are made only between different subgroups of the study cohort, but not between the patients and some reference population. Even in those cases where comparisons are carried out (Luangasanatip *et al.* 2013; Gaitatzis *et al.* 2004), some mathematical and conceptual issues are not completely clear. Therefore, clarification is needed as to how a study cohort and some reference population can be compared with regard to their life expectancy. Such clarification should rest on both statistically and conceptually rigorous arguments. These theoretical clarifications will hopefully lead to a methodology which can facilitate understanding by statistics practitioners, too. Furthermore, we will establish a link between official statistics and survival analysis, which is, in our opinion, a desirable goal not only for scientific, but also for practical reasons (see the closing part of this paper).

2. Basic quantities measuring life expectancy

To set the stage, let's assume that we want to define the life expectancy for a person of exact age x years. As it would hardly make any sense to take an unrestricted range for x, we assume $x \in \{0, 1, ..., x_{max}\}$, where x_{max} is chosen appropriately, for example, $x_{max} = 95$ or $x_{max} = 100$. Furthermore, let X denote a nonnegative continuous random variable measuring the time from birth to death of this person. For sake of simplicity, we do not take gender into account here.
Then, in the context of life table calculations, the following definitions are popular (Hanika and Trimmel 2005):

(i) The *death probability in the age interval* $[x, x+1)$ is defined as

$$q_x := P(x \leq X < x+1 | X \geq x).$$

(ii) Let $l_0 \in \mathbb{N}$. The *number of survivors at age* x is defined as

$$l_x := \begin{cases} l_0 & x = 0 \\ l_{x-1} \cdot (1 - q_{x-1}) & x \geq 1 \end{cases}$$

(iii) The *number of years lived in the age interval* $[x, x+1)$ is defined as

$$L_x := \frac{1}{2}\left(l_x + l_{x+1}\right).$$

Based on these quantities, we define the *life table life expectancy (LTLE) at age* x as

$$E_x := \frac{1}{l_x} \sum_{y=x}^{x_{max}} L_y. \tag{1}$$

Note that this quantity is usually referred to as *life expectancy at exact age* x. However, we choose this terminology to stress the contrast to the life expectancies calculated for the study cohort (see below). Next, we turn to the survival analytic setting, which requires slightly different assumptions. We consider a situation typically encountered in clinical studies, namely that we have a study cohort which is followed over a certain time period. Let T be a non-negative continuous random variable measuring the time from the starting point of the study. Then, the *mean residual life at time* $t \geq 0$ is defined as

$$r(t) := E[T - t | T \geq t].$$

If $E[T] < \infty$, it can be shown that for all $t \geq 0$, we have

$$r(t) = \frac{1}{S(t)} \int_t^\infty S(u) \mathrm{d}u,$$

where $S(t) = P(T > t), t \geq 0$, is the so-called *survivor function*. The survivor function can be estimated either by some step function (e.g., the Kaplan-Meier estimator) or by specifying a certain parametric model such as the Weibull model.
Since this model will be used for illustrative purposes in the theoretical considerations below, we shall state now how it is specified. A Weibull model is characterized by setting

$$S(t) = exp(-\lambda t^p),$$

where $\lambda, p > 0$. According to the impact on the survival function, λ is usually called *scale parameter*, whereas p is referred to as *shape parameter*. As in other parametric models used in survival analysis, covariates can be easily incorporated by an additional term of the form $exp(\beta' \mathbf{x})$: The survivor function is then given as

$$S(t, \mathbf{x}) = exp(-\lambda t^p exp(\beta' \mathbf{x})).$$

The parameters of such a model can be estimated by means of maximum likelihood estimation, which eventually yields an estimator $\hat{S}$ of the survivor function S. For details concerning these pieces of survival analytic theory, we refer to Kalbfleisch and Prentice (2002).

3. Comparing a study cohort to some reference population

Now, the question arises if a link between these two life expectancy measures can be established in order to compare the study cohort with a reference population. In general, we can't just calculate the differences between the mean residual lifetimes (originating from the cohort data) and the corresponding LTLEs: For example, in the context of life tables, the deaths are assumed to be uniformly distributed in each age interval $[x, x + 1)$, whereas this is not the case if we, for example, fit a Weibull model to the study cohort data. However, if we take a closer look at the LTLE formula, it turns out that there is a possibility for the survivor function S to enter the stage. We will see in the following sections that for subsequent years after start of follow-up, we are thus able to calculate life expectancies for the study cohort.

3.1. Mathematical justification

To begin with, note that for all $x \in \{0, 1, ..., x_{max}\}$, we have

$$E_x = \frac{1}{2} + \frac{\tilde{S}(x_{max} + 1)}{2\tilde{S}(x)} + \frac{1}{\tilde{S}(x)} \sum_{k=x+1}^{x_{max}} \tilde{S}(k), \tag{2}$$

where $\tilde{S}(x) = P(X > x)$.
For a proof of this equality, see Appendix A.

Now, based on the LTLE formula (2) just derived, we want to calculate the life expectancy at time $t \geq 0$ after start of follow-up for a person from the study cohort. The only thing left is to examine the relation between $S(t) = P(T > t)$ and $\tilde{S}(x) = P(X > x)$. Recall that X measures the time from birth to death, whereas T takes start of follow-up as the time origin. Therefore, if we let a_E denote the exact age of that person at start of follow-up, we have $T = X - a_E$. Thus, we get that for all $x \in \{a_E, a_{E+1}, ..., x_{max}\}$ (note that $x < a_E$ wouldn't make any sense since we are only interested in time points after start of follow-up!), the following equations hold:

$$\tilde{S}(x) = P(X > x) = P(X - a_E > x - a_E) = P(T > x - a_E) = S(x - a_E).$$

So, we can write (2) as

$$E_x = \frac{1}{2} + \frac{S(x_{max} - a_E + 1)}{2S(x - a_E)} + \frac{1}{S(x - a_E)} \sum_{k=x+1}^{x_{max}} S(k - a_E),$$

for $x \in \{a_E, a_{E+1}, ..., x_{max}\}$. If we set $t := x - a_E$ and do some re-indexing, we can define the *study cohort life expectancy* as

$$SCLE(t) := \frac{1}{2} + \frac{S(x_{max} - a_E + 1)}{2S(t)} + \frac{1}{S(t)} \sum_{k=t+1}^{x_{max} - a_E} S(k), \tag{3}$$

where $t \in \{0, 1, ..., x_{max} - a_E\}$.

Now, we are ready to compare life expectancies between the study cohort and a reference population: At first, we have to pick a certain estimator $\hat{S}$ for the survivor function S of the study cohort (e.g., the survivor function of a regression model fitted to the data). Then, we calculate $SCLE(t)$ for some values $t \in \{0, 1, ..., x_{max} - a_E\}$ of interest. Due to the derivation carried out above, the corresponding value from the life tables which should be taken for comparison is E_{t+a_E}.

3.2. Conceptual considerations

So far, we have established a method where we use the LTLE formula not only for calculations in the context of life tables, but also for quantities based on a survival analytic model. However, although the structure of the formula is the same now, we may need further refinements in certain situations to make sure that we control several sources of bias.

Bias due to "infinite life" models

To begin with, the follow-up time in clinical studies is usually restricted to one or several years. As a consequence, we expect that in general, some of the subjects will be still alive at the end of the study period. For these persons, we don't have the exact survival times. This phenomenon itself, known as right-censoring, which is a key issue in survival analytic methodology, doesn't cause any serious problems. But, especially when we want to estimate life expectancies, we have to be aware of the fact that we often can't avoid making extrapolations to some extent. For example, to calculate the life expectancy of a subject aged 20 at study entry using formula (3), we need values $S_{20}(j)$ up to $j = x_{max} - 19$. By the subscript 20, we indicate that we take the age at entry as a covariate into account here (e.g., by specifying some regression model as mentioned above). Most likely, $x_{max} - 19$ will be greater than the follow-up time, which means that we have to extrapolate beyond the range of our data. In other words, it may not be possible to judge whether, for example, the estimate of $S_{20}(50)$ is reliable or not because most likely, the follow-up time won't be 50 years. Observe that although we may well have data of patients aged 70 at study entry, we must not assume that $S_{70}(0) = S_{20}(50)$

holds in general. Thus, we see that looking at the values of the survivor functions of the patients who were fairly old at time of entry doesn't help solving the problems concerning extrapolation. Therefore, especially for the patients who were quite young at study entry, the estimation of at least some values of the survivor function can be crucial. This point is very important to keep in mind, especially when using models such as the Weibull model which yields values $S(t)$ for all $t \geq 0$: Although it seems as if we could immediately get all we need for the calculation of life expectancies from such a model, we must not use these values without examining their appropriatness. For example, if we consider a Weibull model, it can be shown that the so-called hazard function, which measures the "instantaneous risk of dying", is given as

$$h(t) = \lambda p t^{p-1},$$

which is either decreasing ($p < 1$), increasing ($p > 1$) or constant ($p = 1$) (Kalbfleisch and Prentice 2002). So, if our data yields a maximum likelihood estimate $\hat{p}$ of p which is smaller than 1, this suggests that the risk of dying decreases over time. Whereas this could be reasonable to some extent (for example, think of a risk reduction due to protective effects of clinical monitoring etc.), this is most likely wrong for large values of t: At some time point, the "force of death" will be stronger than any protective effects. If we don't take this issue into account, we will get life expectancy estimators which may be extremely biased since the behaviour of the hazard discussed above translates to a biased survivor function.

To deal with this problem, it may help to pick a model which allows for more flexibility (e.g., more complicated shapes of the hazard function). Although such questions of model appropriatness are important and should not be neglected, it can well be that in some cases, the values of $S(t)$ remain unrealistic: For example, if we have a cohort of relatively young subjects and the follow-up time is, say, 5 years, we can't see a substantial increase in the risk of dying at age 50 simply because we don't observe persons at this age!

Consequently, we suggest defining some kind of "breakpoint" value t_B from which on we don't "trust" the model any more, or at least put some restrictions on the use of model death probabilities (i.e., the q_j's in formula (6) in the appendix, based on the survivor function of the model), which are defined as

$$q(t) := 1 - \frac{S(t+1)}{S(t)}, t \in \{0, 1, ..., x_{max} - a_E\}.$$

For example, it would make sense to take the maximum of the model and the life table death probability q_{t+a_E} for time points t greater than t_B. From a medical point of view, this means that from a certain time point after diagnosis onwards, you assume that the patients are at a risk of dying which is at least as high as for the reference population. To illustrate this idea, assume that we have decided to set $t_B = 5y$. This means that for the first 5 years following study entry, we use $q(0), q(1), q(2), ..., q(5)$ as annual death probabilities for our calculations. But, for all time points $t = 6, 7, ..., x_{max} - a_E$, we take $max\{q(t), q_{t+a_E}\}$ as annual death probability. Of course, if $x_{max} - a_E < 6$, that is, the patient is fairly old at time of diagnosis, this modification doesn't have any effect on the life expectancies because we only need annual death probabilities up to, for example, 3 years after study entry.

Besides, putting a restriction like the one above on the death probabilities for the entire time range would not make sense (unless you have good reasons for such a decision) since we would thus wipe out any protective effects due to, for example, regular visits at the hospital. Note that at least in some studies, it's indeed reasonable to account for such a sort of effect: For example, if we consider cancer patients, we can rightfully assume that they are scheduled for regular examinations, which may imply that some other diseases are detected earlier than in healthy subjects. Note that this does not mean the results are biased: Of course, if we, for example, examined the effect of a certain therapy on the survival of patients suffering from brain tumors, it's obvious that the effect measure would indeed be biased when the control group contains far more subjects with cardiovascular disease than the treatment group. But, note that in this case, the occurence of brain tumors is not supposed to be

related to cardiovascular disease. By contrast, in the example with the regular medical checks from above, we don't have confounding because as we stated before, the main point is that these examinations are caused by the fact that the patient suffers from cancer. Therefore, the possibly beneficial effect actually turns out to be a part of the disease effect.
Moreover, protective disease-related effects may also be caused by the fact that a patient is sometimes forced to change his or her lifestyle: For example, if you have epilepsy, you are usually not allowed to drive a car any more. Likewise, you won't do any dangerous activities such as parachuting or climbing. Thus, you most likely reduce your risk of dying substantially compared to the overall population.
Summing up, we want to point out that although at first sight, it seems as if protective effects introduce bias to the results, we see that in fact, the contrary is true - if we wiped out these effects in our analysis, we would actually eliminate a part of the disease effect! Of course, it's theoretically possible to eliminate the impact of the protective effects mentioned above in order to arrive at some, let's say, "cleaned" effect estimate which, loosely speaking, measures the effect of the disease "itself" (i.e., the loss of life due to medical reasons only). However, what's the practical relevance of such an estimate? For example, is there any use in calculating the effect of epilepsy "itself" (i.e., the impact of the fact that the patients don't drive cars is somehow eliminated) if there isn't anyone who has epilepsy and is allowed to drive a car? To cut it short, we would then have calculated an effect for a person which actually doesn't exist!
To turn back to the breakpoint issue, we must admit that the choice of such a breakpoint value is a crucial task. We suggest taking a look at carefully collected and thoroughly analysed data such as the most recent life tables for the population the patients come from. For example, the Austrian life tables from 2000/02 show a monotonic increase of the death probabilities from age 30 onwards (Hanika and Trimmel 2005). Alternatively, medical experts may also provide useful information concerning this issue. Moreover, an appropriate breakpoint value may be suggested by the design of the study, for example, if at some time point, the members of the study cohort are not scheduled for regular visits at the hospital and thus don't take advantage of protective effects any more. However, keep in mind that for example, there can be protective effects which may influence the survival experience for the entire time range (e.g., if the patients aren't allowed to drive cars any more).
To sum things up, it may be quite difficult to find an appropriate breakpoint value. We suggest taking not only one of the approaches outlined above, but various different considerations into account. For example, medical experts may have quite a good idea of the amount of risk decrease due to protective effects. On the other hand, a look at some life tables may indicate at which age the "force of death" is most likely stronger than any protective effects. Of course, we must admit that this way of solving the problem might not be satisfying at first sight. But, on the other hand, this means of correction is quite easy to carry out and helps to avoid an amount of bias that can be quite large. To cut it short, having a breakpoint which is roughly appropriate is definitely better than introducing no breakpoint value at all: Note that if we have an extremely unrealistic assumption such as a decreasing risk of dying even for large values of t, we sum up this bias when calculating life expectancies and thus get results which are actually useless. So, with our method, we can substantially reduce this kind of bias, with the only drawback that we don't have a formally justified method at hand to determine the breakpoint value t_B.
By the way, recall that the extrapolation issue was the main point we started from. If we are indeed interested in calculating life expectancies, we suggest following the "breakpoint" idea discussed in detail above. But, instead of calculating this quantity (i.e., the expected number of years the subject has left to live), we may also think of restricting ourselves to looking only at some time span such as, for example, 5 years after study entry. All we have to do is to choose x_{max} appropriately. To stay with our example, we just set $x_{max} = a_E + 5$. Note that due to the concept underlying the life table calculations (see section 2), we then get the expected number of years a subject aged a_E has left to live during the following 5 years. Especially in studies with short follow-up time, extrapolation is quite a serious

concern. To circumvent this problem, the idea outlined above may be a good alternative to life expectancy calculations. Likewise, when analyzing data from patients who are at a very high risk of dying in a moreorless short time interval after study entry (e.g., if some surgical procedure is defined as the start of follow-up), there may be no need to calculate life expectancies - instead, something like a "average 5-year lifetime" as proposed above would perhaps be a more useful measure.

Bias due to changes in life expectancy over time

Now, let's turn to a second important issue, namely the information underlying the data from the study cohort and the life table, respectively. What we haven't mentioned so far is the fact that in a clinical study, the subjects often enter the study at different time points, for example, when the diagnosis of a certain disease is defined as the starting point. As far as the calculation of estimates within the study cohort is concerned, this does not cause problems at all: If the follow-up period is quite short, the difference in entry times is negligible, and even if the subjects are followed over several years, we can for instance fit a regression model which includes a covariate "year of entry" or something like that. To judge if one should account for the time point of entry, we suggest using variable selection methods. In addition to that, a look at the life tables for the population of the area the subjects (mainly) belong to can be also useful to get an idea of the amount of change during the study period. However, if we turn to comparisons with life table quantities, we have to be more careful. Again, if the recruitment as well as the follow-up period are quite short, say, for example, one year, it suffices to take one single reference life table for comparisons. This is also appropriate if hardly any changes in the survival experience of the general population can be observed over the study period. But what if the recruitments and/or follow-ups stretch over several years **and** we observe substantial changes in the general population? At first sight, the solution to this problem is easy: If we want to compare the life expectancy of a patient aged a_E diagnosed in, say, 2000, at time of diagnosis (i.e., in the notation from above, $t = 0$) to the corresponding value of the reference population, we just take the number $LTLE(a_E)$ from the life table of 2000 and calculate the difference to $SCLE(0)$. Analogously, if a comparison of life expectancies 5 years after diagnosis is desired, we calculate $SCLE(5)$ and compare this quantity with $LTLE(a_E + 5)$ from the life table of 2005.
However, note that a period life table is always based on the survival experience of a population in a very certain year (or a quite short time period, say, for example, three years, see Hanika and Trimmel 2005). For the calculation of life expectancies, the annual death probabilities for future years are estimated by the corresponding values of the reporting period, which means that the status quo is carried forward (Hanika and Trimmel 2005). Although this basically makes sense, we get into troubles when using the LTLEs for comparison purposes: To turn back to the example from above, let's see what happens if we use $LTLE(a_E)$ from the life table of 2000 as comparison value for $SCLE(0)$. We assume that the follow-up time is 10 years, and, as stated above, a substantial change in life expectancies can be seen in the life tables published during this time span. Then, the key point is that the subject on study has been followed over these years and, thus, the changes in survival experience are somehow implicitely accounted for, whereas the comparison value from the life table of 2000 is based on information of that certain year only and therefore "ignores" the changes observed from 2000 to 2010. To bridge this information discrepancy, we propose the following approach: Suppose a patient enters the study at age a_E in the year y, where $y \in \{y_S, y_S + 1, ..., y_E\}$. By y_S and y_E, we denote the year of study start and end, respectively. For sake of simplicity, we don't assume separate recruiting and follow-up periods, although the method outlined below can be carried out analogously for this case. Although a_E can basically take values from 0 to x_{max}, for the following algorithm, we exclude x_{max} because for a person aged x_{max}, we only need the values $l_{x_{max}}$ and $l_{x_{max}+1}$ to calculate his or her life expectancy, so the quantities we have in the table of the year y are already the ones we need. Secondly, our study ends in y_E, so the life table of this year doesn't require an update since if we assumed any knowledge

concerning survival experience in $y_E + 1, y_E + 2$ and so forth, we would again have different degrees of information in the study cohort and the population.
Recall that for LTLE calculations (i.e., for life expectancy calculations based on formula (1) in section 2), we need the numbers of survivors for all ages from a_E to $x_{max} + 1$, which are denoted by $l_{a_E}, \ldots, l_{x_{max}+1}$. The algorithm used for this purpose can be described as follows:

$$l_{a_E} := l_{a_E,y}, \quad l_{a_E+1} := l_{a_E+1,y}$$
$$l_{a_E+k} := \begin{cases} l_{a_E+k-1} \cdot (1 - q_{a_E+k-1,y+k-1}) & k \in \{2, 3, ..., \min(y_E - y + 1, x_{max} - a_E + 1)\} \\ l_{a_E+k-1} \cdot (1 - q_{a_E+k-1,y_E}) & k \in \{y_E - y + 2, y_E - y + 3, ..., x_{max} - a_E + 1\} \end{cases}$$

The first line means that for a_E and $a_E + 1$, we simply take the values from the life table of the year of entry y. Then, the numbers of survivors are calculated according to the following scheme: To get l_{a_E+2}, we multiply l_{a_E+1} with the annual survival probability for the age $a_E + 1$ from the life table of the year $y + 1$. In this way, we proceed until we reach either $x_{max} - a_E + 1$ or $y_E - y + 1$. Now, if $x_{max} - a_E \leq y_E - y$, we are done since this means that we already have all numbers of survivors up to $l_{x_{max}+1}$. Otherwise, the remaining l_{a_E+k}'s are calculated by applying the "standard" method of life table construction, using the annual survival probabilities of the life table y_E.
So, to sum things up, we basically follow the usual formula for the number of survivors, but with the important modification that for the entire study period, we take all available information on the population's survival experience into account because our calculations are based on year-wise survival probabilities. In other words, we thus get "dynamic" life expectancies for the general population which are, in terms of the information used, indeed comparable to the SCLEs based on the study cohort's data.

To close this section, we briefly discuss an issue of practical importance. As already mentioned at the beginning of this paper, life expectancy is a measure which can be quite easily understood not only by researchers, but also by people not involved in statistics or medicine. Therefore, it's desirable to use the results from a life expectancy analysis as a nice interpretable piece of information the doctors can communicate to their patients. However, we should be aware of the fact that possible changes in life expectancy are crucial for answering the question if we can indeed take the values we've calculated as estimates for the losses or gains in lifetime of future patients. For example, if the study was terminated at the end of 2014, it's most likely fine if a doctor communicates these values to patients who are currently at his clinic. But, if the follow-up ended in 2000, you should be more cautious of course: Although the "dynamic" population life expectancies can quite easily be re-calculated by additionally taking the life tables from 2001 to 2015 into account, the life expectancies for the patient cohort can't be updated since the study was terminated in 2000. However, the patients' life expectancies may also substantially change over time, partly due to the changes in the population, but also as a consequence of therapeutic advances, etc. Of course, if we fitted a regression model with year of entry to the data, we could theoretically plug in the current year as covariate value and finally get a life expectancy estimate. But, obviously, the corresponding regression coefficient was estimated using data from patients who entered the study before 2000, so we would, again, extrapolate beyond the range of our data.
So, to sum things up, before carrying over the results of studies with limited follow-up time to future patients, one should carefully think about changes in life expectancies which may have occurred since the end of the study. For this purpose, information on medical as well as demographic developments and trends are needed. Besides, it should be mentioned that sometimes, the follow-up time can be considered as unlimited: For example, if you actually have the original data (i.e., date of birth, etc.) of the patients at hand and link it with a death registry (e.g., with some probabilistic record linkage method, see Oberaigner and Stühlinger (2005)), the survival experience of the patient cohort can indeed by updated at any time. Based on this data, the survival analytic model is fitted again, and life expectancies are

calculated afterwards. Thus, in this case, the estimates are reliable even for a person being at the doctor's office right now.

4. Data example

To illustrate the methods outlined above, let's take a look at a dataset which was reported by Kardaun (1983). The cohort consists of 90 males diagnosed with laryngeal cancer between 1970 and 1978 at a Dutch hospital. The dataset contains their survival times (in years), that is, the time from entry to either death or Jan 1, 1983, as well as the year of diagnosis, the age at study entry and the stage of cancer on a scale from 1 to 4. Some descriptives of this data are contained in Table 1. For details concerning this dataset, we refer to Kardaun (1983) and Klein and Moeschberger (2003). This dataset is available in the survival package (Therneau 2014) in R (R Core Team 2015).

Table 1: Some descriptives for the Larynx dataset.

number of patients	90
person years of follow-up	377.8
number of deceased (%)	50 (55.6)
median age (range)	65 (41-86)
median diagnosis year (range)	1974 (1970-1978)
stage 1 (%)	33 (36.7)
stage 2 (%)	17 (18.9)
stage 3 (%)	27 (30.0)
stage 4 (%)	13 (14.4)

Following the considerations of the previous chapters, we now carry out the following steps using R:

1. We fit a Weibull regression model with stage, age at diagnosis and year of diagnosis to the data by using the packages survival and SurvRegCensCov (Hubeaux and Rufibach 2014). Note that the 4 stages are coded as dummy variables in the following way: We introduce 3 dummy variables *dummyA, dummyB* and *dummyC* such that stage 1 corresponds to $dummyA = dummyB = dummyC = 0$. The other stages are coded as $100, 010$ and 001, respectively. The results of the model fit are displayed in the table below.

Table 2: Results of Weibull model fit for the Larynx dataset.

	estimate	standard error
scale (λ)	0.094	0.525
shape (p)	1.121	0.141
dummyA	0.181	0.463
dummyB	0.665	0.356
dummyC	1.780	0.431
age	0.019	0.014
year	-0.022	0.073

Recall that the survival function of a Weibull regression model is given as $S(t, \mathbf{x}) = \exp(-\lambda \mathbf{t}^{\mathbf{p}} \exp(\boldsymbol{\beta}' \mathbf{x}))$. As indicated by the estimate of the shape parameter, the probability of surviving is decreasing moderately fast. To turn to the impact of the different stages on survival, the estimates of the dummy variables show what we expect for medical reasons: The higher the stage, the lower the value of the survival function is. Likewise, it's small suprise that the probability of surviving decreases with age. Interestingly, patients who were diagnosed later seem to have a better chance of surviving. However, all the interpretations I've mentioned have to be taken with caution since the estimated standard errors are quite large in most cases.

2. To deal with the "infinite life" problem discussed in section 3.2., we take a look at the age-specific death probabilities of the annual Dutch life tables from 1970 to 1982 which are available upon request at the website of Statistics Netherlands (Statistics Netherlands 2015). Obviously, there is a remarkable increase from age 40 onwards. As the patients' ages at diagnosis range from 41 to 86 years, it's reasonable to assume that the whole study cohort is exposed to a relatively high, increasing risk of dying. Therefore, the breakpoint should be set to a fairly small value (of course, the choice of the breakpoint can also be based on medical considerations, as mentioned above). We decided to take 0, 3 and 5 years as breakpoints and compare the results. This means that from 1, 4 and 6 years after diagnosis onwards, respectively, we take the maximum of the model and the life table death probabilities instead of merely using the model death probabilities, as described in section 3.2. In step 3, these calculations will be illustrated by an example.

3. For sake of simplicity, we only take one certain covariate combination, namely a person aged 65 which enters the study in 1974 with the diagnosis of stage 2 cancer. We are interested in this person's life expectancy at time of entry (i.e., t = 0) and 2 years after entry (t = 2), respectively. To calculate the life expectancies based on the model, we at first take $\hat{S}$, the estimate of the Weibull regression model CDF from step 1, and plug in the given covariate values for stage, age and year as well as $t = 0$ (or $t = 2$). Then, we calculate the annual death probabilities $1 - \hat{S}(1)/\hat{S}(0), 1 - \hat{S}(2)/\hat{S}(1)$ and so forth, as stated in section 3.2. and in the proof of (2) in the appendix. As the Dutch life tables contain values up to an age of 99 years, we need annual death probabilities up to $1 - \hat{S}(t_{max} + 1)/\hat{S}(t_{max})$, where $t_{max} = 99 - 65 = 34$. To explain the latter formula, recall that we want to calculate the life expectancy for a person aged 65 at time of diagnosis. To make sure that there is a correspondence between the highest age x_{max} in the life table and the greatest time point t_{max} after diagnosis, we choose the latter value such that if we "follow" the patient aged 65 at time of diagnosis for t_{max} years, we arrive exactly at the highest life table age x_{max}. Once we have the annual death probabilities, we immediately get all the other life table quantities, especially the life expectancies, by using the definitions given in section 2. Keep in mind that the value of the breakpoint is crucial for the actual usage of the probabilities just calculated: For example, if we set $t_B = 0$, we at first use $1-\hat{S}(1)/\hat{S}(0)$. But then, we take the maximum of $1 - \hat{S}(2)/\hat{S}(1)$ and the annual death probability for a 66-year old from the life table of 1975, the maximum of $1 - \hat{S}(3)/\hat{S}(2)$ and the annual death probability for a 67-year old from the life table of 1976, and so on, as described in section 3.2.
   It should be mentioned that for the life expectancy calculations, we can take formula (3) as well, which is more straightforward than the method outlined above. However, it's sometimes good to have the annual death probabilities at hand, too. Therefore, we used the life table formulas in our calculations.

4. To assess the variability of the estimates calculated in the previous step, we calculate 95% confidence intervals using the bias-corrected and accelerated (BCa) bootstrap method (Carpenter and Bithell 2000). More to the point, we draw a sample of size $n = 90$ with replacement from the dataset, carry out the steps 1-3 2000 times and save the results (i.e., the life expectancy estimates) in an array. Finally, we use the bootBCa function in the rms package (Harrell 2015) to get the desired confidence intervals.

5. To turn to the population life tables, we observe life expectancy changes up to 1.6 years in the period from 1970 to 1982. Especially for ages between 0 and 50, the life expectancy increase exceeds 1 year. Therefore, we should use "dynamic" life expectancies: Based on the Dutch life tables, we create a new array containing the "dynamic" life expectancies by implementing the algorithm proposed in section 3.2.

6. Finally, we take the "dynamic" life expectancies for a 65-year old in 1974 and a 67-year old in 1976 and subtract them from SCLE(0) and SCLE(2) calculated in step

3, respectively. In the same manner, we get confidence intervals for the resulting life expectancy differences: As we consider the population life expectancies as fixed values, we can simply take the confidence intervals from step 4 and subtract the population values from the interval endpoints.

The results of the procedure outlined above are displayed in Table 2 and Table 3. Recall that a negative value indicates a decreased life expectancy of the patient, whereas a positive number corresponds to an increased life expectancy of the cohort member compared to a healthy person. So, obviously, a laryngeal cancer patient with the characteristics mentioned above tends to have decreased life expectancies compared to the population. However, the variability of the estimates is fairly high (recall that the Weibull parameter estimates come with a relatively high standard error, see Table 2). Consequently, it is hard to tell if the life expectancy loss is substantial or not. Likewise, when taking only the point estimates for 0 and 2 years after the start, there seems to be a slight time trend. However, the corresponding confidence intervals look quite similar, so actually, the data does not allow a clear statement about any time trends in the life expectancy differences.

Table 3: Life expectancies for a patient aged 65 at time of entry in 1974 and corresponding values for the reference population. SCLE = study cohort life expectancy (i.e., life expectancy of the patient with the covariate values mentioned above), LTLEd = dynamic life table life expectancy. SCLEs are given with 95% confidence intervals in brackets.

	$SCLE(0)$	$SCLE(2)$	$LTLEd(65,74)$	$LTLEd(67,76)$
$t_B = 0$	8.86 (5.36, 14.17)	8.36 (4.79, 12.94)	14.16	12.94
$t_B = 3$	8.86 (5.27, 14.04)	8.36 (4.80, 12.97)	14.16	12.94
$t_B = 5$	8.86 (5.18, 14.01)	8.36 (4.62, 12.85)	14.16	12.94

Table 4: Life expectancy differences with 95% confidence intervals for a patient aged 65 at time of entry in 1974. SCLE = study cohort life expectancy (i.e., life expectancy of the patient with the covariate values mentioned above), LTLEd = dynamic life table life expectancy

	$SCLE(0) - LTLEd(65,74)$	$SCLE(2) - LTLEd(67,76)$
$t_B = 0$	-5.29 (-8.80,0.01)	-4.58 (-8.14, 0.00)
$t_B = 3$	-5.29 (-8.89, -0.11)	-4.58 (-8.14, 0.03)
$t_B = 5$	-5.29 (-8.98, -0.14)	-4.58 (-8.32, -0.08)

Moreover, the point estimates of the life expectancy differences seem to be a bit surprising at first sight because they are -5.29 years for $t = 0$ and -4.58 years for $t = 2$, no matter which of the three breakpoint values we choose! When having a look at the annual death probabilities based on the model, we see that right from t = 0 onwards, these values are much higher than the corresponding quantities for the population, except for time points which are quite far away from the time of entry (in the example above, the annual survival probabilities of the patients exceed the corresponding values for the population from 20 years after start onwards). Consequently, it's small suprise that it does not matter if we set t_B to 0, 3 or 5 - the model probabilities will be higher anyway.
However, the choice of t_B may be crucial if we consider other covariate combinations, for example, if the value of age at entry is large. Moreover, the impact of the breakpoint heavily depends on the parameters of the Weibull CDF, especially on the scale λ and the shape p. To take a hypothetical example, we set $\lambda = 0.01$, instead of $\lambda = 0.09$ as in our model fit above (actually, this choice might not be that hypothetical because the standard error of the maximum likelihood estimate of λ is quite large!). Then, the values of $SCLE(0)$ are 0.26, 1.01 and 1.71 for $t_B = 0, 3, 5$, respectively. Thus, we see that different choices of t_B can indeed affect the results! In other words, one should be aware of the fact that introducing a breakpoint is not a question of how pedantic you are: There may be situations where this idea helps avoiding a substantial amount of bias. As discussed in section 3.2., the survival probabilities of the patients may be at least partially unrealistic for time points not covered

by the follow-up time (i.e., time points which are "far away" from the start). To turn back to the hypothetical example from above, the change in the life expectancy differences indicates that the survival probabilities of the cohort are higher than the corresponding values for the population soon after the start. However, it is sometimes crucial to decide which breakpoint is the most appropriate one. For several possibilities of solving this problem, we refer to the discussion in section 3.2.

5. Summary and closing remarks

In this paper, we have described an approach to life expectancy comparisons which is based on a re-formulation of the formula used in the context of life table calculations. Thus, we have made sure that the study cohort and the population quantities are structurally the same. Then, we have discussed two common major sources of bias and how to remedy them. We'd like to emphasize once more that this is not a question of only peripheral interest: If those sources of bias aren't controlled, the results will be unreliable or, in fact, even useless.
It should be mentioned that the methodology proposed above can be applied to quite a broad variety of settings: If we want to set up a more complicated model for the study cohort and take several covariates such as type of disease, health status, etc. into account, all we have to do is to fit an appropriate survival analytic model to the study cohort data. Then, we can proceed in the way outlined above and make life expectancy comparisons with the age- and gender-matched "dynamic" values from the life tables. Furthermore, although our work is motivated by questions arising in a medical context (see the introductory section), the proposed method can be used in other scientific branches (e.g., the Social Sciences), too. In general, we hope that our method is reasonably well to understand not only for statisticians, but also for researches who want to use it for their data analyses.
Finally, we'd like to stress the importance of thinking about links between applied mathematical statistics and official statistics. Apart from the scientific progress gained, there is also a practical reason for such considerations: As to the life expectancy comparisons, we could theoretically consider a case-control design and set up a matched control group recruited from the population as well. However, this would require substantial (monetary) efforts especially in long-term studies or in studies where quite large sample sizes are desired. So, it is a better idea to take high-quality and easy-to-access data such as the period population life tables for comparison purposes.

Acknowledgements

The author thanks Yvonne Höller, Claudia A. Granbichler and Eugen Trinka (Department of Neurology, Christian Doppler Klinik, Salzburg) as well as Arne Bathke (Department of Mathematics, Paris Lodron University, Salzburg) for many constructive discussions about this paper's content. Moreover, the author is grateful to the editor and the reviewer for their valuable suggestions, which led to an improvement of several parts of this paper.

Appendix

Recall that in section 3.1., we stated that for all $x \in \{0, 1, ..., x_{max}\}$, we have

$$E_x = \frac{1}{2} + \frac{\tilde{S}(x_{max}+1)}{2\tilde{S}(x)} + \frac{1}{\tilde{S}(x)} \sum_{k=x+1}^{x_{max}} \tilde{S}(k), \tag{4}$$

where $\tilde{S}(x) = P(X > x)$.

To prove this, we first expand E_x: According to the definitions given above, we have

$$\begin{aligned} E_x &= \frac{1}{l_x}\sum_{k=x}^{x_{max}} L_k = \frac{1}{2l_x}\sum_{k=x}^{x_{max}}(l_k + l_{k+1}) \\ &= \frac{1}{2l_x}\left(\sum_{k=x}^{x_{max}} l_k + \sum_{k=x+1}^{x_{max}+1} l_k\right) \\ &= \frac{1}{2l_x}\left(l_x + l_{x_{max}+1} + 2\sum_{k=x+1}^{x_{max}} l_k\right). \end{aligned} \tag{5}$$

Now, we use the recursive definition of l_k for $k \geq 1$ to obtain

$$l_k = l_x \prod_{j=x}^{k-1}(1-q_j), k \in \{x+1, x+2, ..., x_{max}+1\}.$$

When applying this to (5), l_x cancels out, and thus we get

$$E_x = \frac{1}{2} + \frac{1}{2}\prod_{j=x}^{x_{max}}(1-q_j) + \sum_{k=x+1}^{x_{max}}\prod_{j=x}^{k-1}(1-q_j). \tag{6}$$

Now, according to the definition of q_j, we have

$$P(X > j+1 | X \geq j) = 1 - P(j \leq X < j+1 | X \geq j) = 1 - q_j$$

for $j = 0, 1, ..., x_{max}$. Applying some basic probability theory leads to the equation

$$1 - q_j = P(X > j+1 | X \geq j) = \frac{P(X > j+1)}{P(X \geq j)} = \frac{\tilde{S}(j+1)}{\tilde{S}(j)}, j = 0, 1, ..., x_{max}.$$

Combining this result with (6) yields

$$E_x = \frac{1}{2} + \frac{1}{2}\prod_{j=x}^{x_{max}}\frac{\tilde{S}(j+1)}{\tilde{S}(j)} + \sum_{k=x+1}^{x_{max}}\prod_{j=x}^{k-1}\frac{\tilde{S}(j+1)}{\tilde{S}(j)} = \frac{1}{2} + \frac{\tilde{S}(x_{max}+1)}{2\tilde{S}(x)} + \frac{1}{\tilde{S}(x)}\sum_{k=x+1}^{x_{max}}\tilde{S}(k),$$

which completes the proof.

References

Bradshaw P, Stobie P, Knuiman M, Briffa T, Hobbs M (2015). "Life Expectancy After Implantation of a First Cardiac Permanent Pacemaker (1995–2008): A Population-Based Study." *Int J Cardiol.*, **190**, 42–46.

Carpenter J, Bithell J (2000). "Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians." *Statist. Med.*, **19**, 1141–1164.

Chang K, Lu T, Lee K, Hwang J, Cheng C, Wang J (2015). "Estimation of Life Expectancy and the Expected Years of Life Lost Among Heroin Users in the Era of Opioid Substitution Treatment (OST) in Taiwan." *Drug Alcohol Depend.*

Chiang C (1979). *Life Table and Mortality Analysis*. World Health Organization. URL http://apps.who.int/iris/bitstream/10665/62916/1/15736_eng.pdf?ua=1.

DuGoff E, Canudas-Romo V, Buttorff C, Leff B, Anderson G (2014). "Multiple Chronic Conditions and Life Expectancy: a Life Table Analysis." *Med Care*, **52**(8), 688–694.

Gaitatzis A, Johnson A, Chadwick D, Shorvon S, Sander J (2004). "Life Expectancy in People with Newly Diagnosed Epilepsy." *Brain*, **127**, 2427–2432.

Guaraldi G, Cossarizza A, Franceschi C, Roverato A, Vaccher E, Tambussi G, Garlassi E, Menozzi M, Mussini C, D'Arminio Monforte A (2014). "Life Expectancy in the Immune Recovery Era: the Evolving Scenario of the HIV Epidemic in Northern Italy." *J Acquir Immune Defic Syndr.*, **65**(2), 175–181.

Hanika A, Trimmel H (2005). "Sterbetafel 2000/02 für Österreich." *Stat Nachr Osterr Stat Zent Amt*, **2**, 121–131. URL http://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bevoelkerung/sterbetafeln/index.html.

Harrell F (2015). *rms: Regression Modeling Strategies.* R package version 4.3-1., URL http://CRAN.R-project.org/package=rms.

Hubeaux S, Rufibach K (2014). *SurvRegCensCov: Weibull Regression for a Right-Censored Endpoint with Interval-Censored Covariate.* R package version 1.3, URL http://CRAN.R-project.org/package=SurvRegCensCov.

Kalbfleisch J, Prentice R (2002). *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, Inc. ISBN 9780471363576.

Kardaun O (1983). "Statistical Analysis of Male Larynx-Cancer Patients–A Case Study." *Statistical Nederlandica*, **37**, 103–126.

Klein J, Moeschberger M (2003). *Survival Analysis. Techniques for Censored and Truncated Data.* Springer.

Li K, Hüsing A, Kaaks R (2014). "Lifestyle Risk Factors and Residual Life Expectancy at Age 40: a German Cohort Study." *BMC Med.*, **12**(59). URL http://www.biomedcentral.com/1741-7015/12/59.

Luangasanatip N, Hongsuwan M, Lubell Y, Limmathurotsakul D, Teparrukkul P, Chaowarat S, Day N, Graves N, Cooper B (2013). "Long-Term Survival After Intensive Care Unit Discharge in Thailand: a Retrospective Study." *Crit Care*, **17**(5). URL http://www.ccforum.com/content/17/5/R219.

Nagai M, Kuriyama S, Kakizaki M, Ohmori-Matsuda K, Sone T, Hozawa A, Kawado M, Hashimoto S, Tsuji I (2011). "Impact of Walking on Life Expectancy and Lifetime Medical Expenditure: the Ohsaki Cohort Study." *BMC Open*, **1**(2). URL http://bmjopen.bmj.com/content/1/2/bmjopen-2011-000240.full.

Nusselder W, Slockers M, Krol L, Slockers C, Looman C, van Beeck E (2013). "Mortality and Life Expectancy in Homeless Men and Women in Rotterdam: 2001âĂŞ2010." *PLoS One*, **8**(10). URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073979.

Oberaigner W, Stühlinger W (2005). "Record Linkage in the Cancer Registry of Tyrol, Austria." *Methods Inf Med*, **44**, 1–5.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Statistics Netherlands (2015). Accessed: 2015-07-31, URL http://www.cbs.nl/en-GB/menu/home/default.htm.

Strauss D, DeVivo M, Paculdo D, Shavelle R (2006). "Trends in Life Expectancy after Spinal Cord Injury." *Arch Phys Med Rehabil.*, **87**(8), 1079–1085.

Therneau T (2014). *A Package for Survival Analysis in S*. R package version 2.37-7, URL http://CRAN.R-project.org/package=survival.

Affiliation:

Georg Zimmermann
Department of Neurology, Christian Doppler Klinik
A-5020 Salzburg, Austria
Spinal Cord Injury and Tissue Regeneration Center, Paracelsus Medical University,
A-5020 Salzburg, Austria
Department of Mathematics, Paris Lodron University
A-5020 Salzburg, Austria
E-mail: georg.zimmermann@pmu.ac.at

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 45
March 2016

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2015-08-04
Accepted: 2015-11-26

Austrian Journal of Statistics
March 2016, Volume 45, 69–80.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i2.303

Die Theorie lebt in der Praxis. Ein Interview mit Ernst Stadlober

Ernst Stadlober
TU Graz

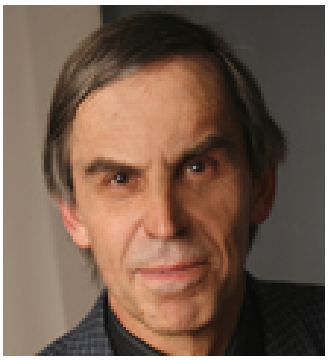
Herwig Friedl
TU Graz

Matthias Templ
TU Wien & Statistics Austria

Abstract

Das Interview mit Ernst Stadlober wurde von Herwig Friedl und Matthias Templ am 18.12.2015 geführt. Es zeichnet ein Bild des beruflichen Werdeganges von Ernst Stadlober, von seinen Anfängen wo er mit fix gesetztem *seed* auf deterministischem Wege über die *Random Number Generation* zu seiner sehr breiten Ausrichtung der Statistik fand. Viele erfolgreich angewandte Forschungsprojekte mit Partnern aus Verwaltung, Industrie und Wirtschaft bezeugen ebenso seine Erfolgsgeschichte als auch die beispiellose intensive Betreuung von Studenten an der TU Graz. Man kann zurecht behaupten, dass Ernst Stadlober ein breites Methodenspektrum aus dem Gebiet der Statistik beherrscht und es trotzdem schaffte in viele Spezialgebiete auch tiefer vorzudringen.

Ernst Stadlobers berufliche Heimat war und ist das Statistikinstitut der TU Graz, das er seit 1998 auch leitet. Dazwischen war er auf Forschungsaufenthalten an der Stanford University/USA und der TH Darmstadt und hatte eine Lehrstuhlvertretung an der Universität Kiel. Bis heute hat er 12 Dissertationen und mehr als 90 Diplom-/Masterarbeiten betreut. Zum Repertoire seiner Lehre zählt die (Angewandte) Statistik, Zeitreihenanalyse, Stochastische Modellierung und Simulation, Versuchsplanung und einiges mehr. Zusätzlich blickt er heute auf über eine Reihe von 100 Vorträgen sowie auf etwa 80 Publikationen aus dem Bereich der Biostatistik, Computerstatistik und Angewandten Statistik zurück.



Keywords: Interview, Statistik, Statistikstudium, Projekte.

Matthias Templ: *Danke für deine Zusage uns ein Interview zu geben. Nach einer Reihe von sehr interessanten Interviews im Austrian Journal of Statistics, setzen wir die Interviewreihe mit dir fort. Vielleicht sollten wir von deinen Anfängen in der Statistik zuerst sprechen. Am Statistikinstitut der TU Graz hast du bei Ulrich Dieter dissertiert. Meines Wissens war er eine große Persönlichkeit.*

Ernst Stadlober: Ja, das war ein sehr umtriebiger, vielseitiger Mathematiker. Der kommt an sich aus der Zahlentheorie und hat eben damals, wie es so üblich war, in der Mathematik an der TU Graz diese Stelle für Mathematische Statistik angenommen. Das

war das Normale, dass ein Mathematiker so einen Bereich übernommen hat. Prof. Dieter hat dann auch die Vorlesungen Wahrscheinlichkeitstheorie und Statistik gehalten und die natürlich eher aus mathematischer Sicht durchgeführt. Er selbst hat nie selber Statistik betrieben in dem Sinn. Ich bin durch ihn in den Forschungsbereich der Zufallszahlenerzeugung reingekommen. In der Lehre hat er damals am Anfang auch die Optimierungstheorie abgedeckt, weil zu der Zeit haben wir an der TU Graz noch keine Professur dafür gehabt. Mein Einstieg war nicht die Statistik sondern die Optimierung. Da hab' ich dann auch meine Diplomarbeit geschrieben, im Bereich der Optimierung. Das war eine Untersuchung über das so genannte quadratische Zuordnungsproblem, ein Vergleich von Computerheuristiken. Da war schon der Ansatz, wo man sagt: man hat die Kombination zwischen der theoretischen Analyse von Methoden und den praktischen Vergleichen von Methoden – auch am Computer.

Matthias Templ: *Und das bereits in den frühen* 80*er Jahren, mit Veröffentlichungen von dir mit computergestütztem Inhalt und computerorientierten Methoden.*

Ernst Stadlober: Ja, wie gesagt, der Einstieg war dann stark computerorientiert. Ich habe die Assistentenstelle 1976 bekommen, knapp nach dem Abschluss meines Diploms. Dann ist zu dieser Zeit auch Rudi Dutter nach Graz gekommen, als Assistent und der hat einen starken Computerbezug mitgebracht. Er hat damals die ersten Vorlesungen in Computerstatistik bei uns angeboten. Also mit SPSS und BMDP zum Beispiel hat er das durchgeführt. Das war dann für mich quasi der Einstieg in die ersten Anwendungen mit realen Datenbeispielen. Dadurch ist man langsam reingekommen in diese Datenlandschaft: wie kann man die Methoden, die man theoretisch gelernt hat, umsetzen, wo sind sozusagen auch die Hemmschwellen. Das war also ein ganz guter Anfang in dieser Problematik. In der Lehre habe ich damals schon relativ viel übernehmen müssen, Übungen und auch eigene Vorlesungen. Obwohl ich das Doktorat noch gar nicht hatte, habe ich zum Beispiel Vorlesungen wie Maßtheorie, Wahrscheinlichkeitstheorie, Warteschlangentheorie oder Stochastische Prozesse abgehalten, neben der Übungsbetreuung und Betreuung von Projekten. Wie gesagt, der Einstieg in die Rechnergeschichte war so, dass wir im Studium Fortran gelernt haben und ich habe in weiterer Folge sehr viel in Fortran programmiert.

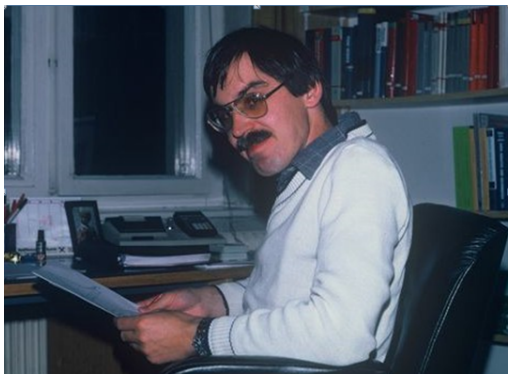


Foto 1981: Assistent an TU Graz

Herwig Friedl: *Ich kann mich sogar daran erinnern, als Rudi Dutter mit APL gekommen ist und wir dann gemeinsame Seminare gemacht haben, das war dann vielleicht ein paar Jahre später.*

Ernst Stadlober: Ja das APL war damals sehr interessant und spannend. Das Problem in APL war halt die Tatsache, dass die Tastatur anders zu benutzen war. Wenn du „mal" drücken wolltest beim APL, dann war das Malzeichen ganz woanders zu finden als in Fortran. Und ich kann mich erinnern, ich habe ein Jahr parallel gearbeitet, einmal mit Fortran, einmal mit APL. Dann habe ich APL schließlich aufgegeben, weil das hin und her mit unterschiedlichen Tastaturen war nicht mehr zu packen. Das war das große Hindernis für APL, obwohl es von der Sprache her sehr gelungen war, es war schon vektororientiert und matrixorientiert. Die Matrixinversion war ein Befehl. In Fortran hast du das alles explizit ausprogrammieren müssen mit den Schleifen und so weiter. [*lacht*]

Herwig Friedl: *Hast du eigentlich deine Algorithmen in Fortran implementiert, oder auch noch in anderen Sprachen?*

Ernst Stadlober: In Fortran, ja. In APL auch – und in Assembler. Wir haben ja damals auch diese Zufallszahlen untersucht – Prof. Dieter hat sich mit gleichverteilten beschäftigt und Prof. Ahrens, sein Koautor, mit den nicht-gleichverteilten. Da waren die zwei

Foto 1993: mit Ahrens und Dieter

ja Gurus auf dem Gebiet: Ahrens-Dieter war so eine Trademark in der Scientific Community. Ich habe dann versucht Algorithmen zu finden für bestimmte Verteilungen, und die haben wir dann in Fortran programmiert, aber den Generator für die gleichverteilten Zufallszahlen zum Beispiel, den haben wir immer in Assembler programmiert, weil der war die Basis von allen anderen Algorithmen. Der hat dann möglichst schnell und effizient sein müssen, damals auf einem UNIVAC Großrechner. Wir haben zur damaligen Zeit natürlich alle Tricks ausgenutzt in Richtung effizienter Multiplikation und Speicherung, um möglichst rechenzeitsparend und auch speichersparend zu arbeiten. Also das war ja eine ganz andere Zeit mit wenig Speicherplatz und langsamen Rechnern. Die Optimierung war immer in der Richtung, dass du Methoden findest, die diese beiden Aspekte gut genug abdecken konnten.

Herwig Friedl: *Aber seitdem ich dich kenne hast du immer mehr oder weniger relativ angewandt gearbeitet und bei der Anwendung natürlich auch relativ computergestützt.*

Ernst Stadlober: Ja, das auf alle Fälle.

Herwig Friedl: *So irgendwie die neue Schiene "Rudi Dutter/Ernst Stadlober" die Dieter überhaupt nicht vertreten hat. Dieter war reiner Theoretiker. Der hat mit einem Blatt Papier und einem Kugelschreiber alles gemacht und bei dir hat man dann gelernt, wie die Sache in der Praxis tatsächlich angegangen werden könnte.*

Ernst Stadlober: Ja eben, wobei der Dieter jetzt nicht der Freak war am Computer, hat er trotzdem die Offenheit gehabt dafür. Er hat ja mit Jo Ahrens zusammen gearbeitet, Jo Ahrens war der Computermensch, und Uli Dieter hat eher die theoretischen Sachen entwickelt. Den Vorteil dieser beiden Teile hat er durchaus gesehen und er hat uns nie irgendeinen Hemmschuh in den Weg gelegt und gesagt, ok, der macht jetzt zu viel Anwendung oder arbeitet viel mit Computer, sondern im Gegenteil. Er war zum Beispiel der erste, der das TeX nach Graz gebracht hat zum Beispiel, oder das Maple. Wo damals die Leute im Rechenzentrum gesagt haben: „Was ist denn das?", so auf die Art. Kannst du dich noch erinnern [*zu Friedl*], der Johann Theurl war damals der Chef, der war eine recht dominierende Persönlichkeit und hat solche Sachen immer abgeblockt. Wenn der Dieter mit irgendeinem Computerband aus Deutschland gekommen ist, wo interessante neue Programme drauf waren, hat er zunächst abgeblockt. Und wie gesagt, der Dieter hat dann durchaus den Riecher gehabt für neue Entwicklungen und hat uns vieles in dieser Richtung ermöglicht.

Matthias Templ: *Ich denke du hattest nie Angst vor Neuem. Ich habe ein bisschen recherchiert, in welchen Bereichen der Statistik und Themengebieten du aktiv warst. Das überspannt Bereiche wie Chemometrie, Biometrie, Versuchspläne, Zeitreihenanalyse und Prognose, Stochastik und Portfolio-Optimierung. Überdies hast du Diplomarbeiten in den verschiedensten Gebieten betreut, z.B. in Themen wie Mischmodelle, Resamplingverfahren, Anwendungen der Linguistik und eben, wie du schon erwähnt hast, Random Number Generators. Wegen dieser sehr breiten Ausrichtung komme ich zum Schluss, dass du sehr offen für neue Themen sein musst.*

Ernst Stadlober: Ja, das Spannende ist, wenn man jetzt als Mathematiker draufblickt auf diese Dinge. Gott sei Dank kommen wir ja aus der Mathematik und wir haben die Strukturen der Methoden gelernt, von der Mathematik her. Dann habe ich eine Anwendung und sehe, eigentlich ist das immer das gleiche Schema für mich als Mathematiker: ich habe das abstrakte Modell, ich muss nur den Übergang schaffen vom Abstrakten zur konkreten Anwendung. Dann ist es vollkommen egal, ob ich mit einem linguistischen Problem arbeite, mit einem Problem aus der Prozessindustrie, mit einem Problem aus der Medizin, ich habe sozusagen immer mein gleiches abstraktes Konstrukt. Das Problem ist also: „Wie schaffe ich den Link dazwischen?“. Das heißt, wie kann ich mit dem, der mit einem Problem zu mir kommt, erst einmal die gleiche Sprache finden. Ich weiß, am Anfang haben wir ja herausfinden müssen – da war der Herwig ja auch immer wieder eingespannt im medizinischen Bereich – was die Leute überhaupt wollen. Also, dass man überhaupt fragt: „Ok, was willst du jetzt überhaupt?“. Das waren oft sehr diffuse Vorstellungen und sehr diffuse Aussagen. Die haben wir zuerst einmal ein bisschen strukturieren und auf den Punkt bringen müssen, das war der erste Schritt. Und dann verstehen, aha, jetzt haben wir es soweit, jetzt können wir schauen, wie passt das zusammen mit irgendeinem statistischen Modell. Also wie kann ich die Fragestellung übertragen – die Merkmale und so weiter. Wenn man diesen Schritt einmal gelernt hat, und versucht, den Partner zu verstehen, dann wird im Prinzip immer der gleiche Mechanismus angewendet, natürlich mit unterschiedlichen Ausprägungen. Also ich sehe das eher so als einheitliches Konstrukt, dass ich sage, egal mit welchem Problem du heute kommst, ich kann das in mein Konzept allgemeiner einbringen. Und das ist das Spannende an unserer Statistik, an der sogenannten Angewandten Statistik, die wirklich angewandt ist, dass du durch die Verbindung zum Abstrakten irrsinnig viel erreichen und machen kannst.

Foto 1983: Promotion

Matthias Templ: *Dann gibt es aber heutzutage wahrscheinlich zu wenige Angewandte Statistiker, weil sich immer mehr Wissenschaftler auf einen Bereich spezialisieren. Dort sind sie dann natürlich Spezialisten, aber nicht mehr offen für andere Sachen.*

Ernst Stadlober: Wie gesagt, das ist natürlich ein Grundproblem, weil man natürlich jetzt an der Uni primär, wenn man sich habilitieren will, fokussiert sein muss auf so einen engen Bereich. Ich habe mich damals auf meinen Forschungsbereich, die Zufallszahlenerzeugung konzentriert. Da habe ich mich auf Algorithmen gestürzt und neue Verfahren ausgearbeitet und entwickelt. Die Methoden nicht nur entwickelt, sondern auch in entsprechende Packages verpackt, die zugänglich waren. Das Ende war, darauf bin ich heute noch stolz, das Softwarepackage CRAND oder WinRand, das war eine Implementierung von Zufallszahlengeneratoren für über 30 Verteilungen, das in *C* geschrieben worden ist. Das wurde dann als offene Plattform von uns angeboten. Das war vielleicht ein bisschen wie – von der Größe natürlich überhaupt nicht vergleichbar mit *R* – aber so ein ganz kleiner Beitrag, wo ich sage, da ist etwas, was man der Öffentlichkeit zugänglich macht, etwas Aktuelles. Das ist es tatsächlich noch immer, Teile davon sind implementiert worden in anderen Packages. Es ist auch in *R* ein bisschen was von uns drinnen und es gibt immer noch Methoden, die ich vor mehr als 20 Jahren entwickelt habe, die heute nach wie vor am Stand der Technik sind. Also da sieht man, wenn man etwas intensiv betreibt und ernsthaft versucht, es effizient zu machen, dass dies durchaus Bestand haben kann – auch in unserer Zeit, wo es sehr kurzlebig zugeht im Bereich der Softwareentwicklung und so. Aber gewisse Grundprinzipien ändern sich ja trotzdem nicht.

Aber wie gesagt, das was ich jetzt als Vorteil sehe in unserem Bereich ist, dass wir Mathematik studiert haben und ausgehend von der Mathematik haben wir die grundlegende Methodik sauber gelernt und das ist ein Fundament, und wenn Interesse besteht

auch raus zu gehen, das muss ich mich halt trauen als Mathematiker. Dabei verlasse ich den sicheren Boden von „Definition, Satz und Beweis“, und beschäftige mich mit sogenannten schmutzigen Daten, wo nichts mehr exakt stimmt, von der Theorie her, auf das muss ich mich einlassen. Den Mut haben, auch wenn man am Anfang überhaupt nicht weiß, ob irgendetwas funktioniert, einfach reinzugehen und das auszuprobieren. Das haben wir auch erst lernen müssen. Bei den ersten angewandten Projekten habe ich selber irrsinnig viel Bauchweh gehabt, ob das überhaupt funktioniert und das war die Hemmung des Mathematikers. Weil da habe ich gesagt, jetzt verlasse ich einen sicheren Boden und ich weiß überhaupt nicht was mich da vielleicht erwartet. Scheitere ich überhaupt mit meinen Methoden? Und natürlich haben wir dann Schritt für Schritt kleine Erfolge gehabt in der Anwendung und haben gesehen, dass die Modelle ganz gut funktionieren und dass wir das umsetzen können. Und dann wirst du natürlich immer frecher und bekommst immer mehr Selbstbewusstsein, weil du sagst irgendetwas kommt immer heraus im Endeffekt – das ist unser Anspruch, wenn irgendjemand kommt und am Anfang sagt „Ich weiß nicht so recht, ob da was drinnen ist.“, aber ich habe die Erfahrung gemacht „Irgendwas schaffen wir schon, was nicht ganz sinnlos ist.“. Und mit dem arbeite ich heute, quasi mit dem Vertrauen das man durch die jahrelange Erfahrung bekommen hat.

Foto 1988: vor Habilitation

Matthias Templ: *Das hat sich also insofern weiter entwickelt, dass du irre viele Projekte mit der Wirtschaft und anderen Instituten erfolgreich durchgeführt hast.*

Ernst Stadlober: Ja, das ist dadurch entstanden. Am Anfang, wie gesagt, habe ich mich im Bereich der Zufallszahlen habilitiert und habe dann eben, als Dozent die normale Lehrverpflichtung für die Mathematiker gehabt. Dann sind diese angewandten Projekte, so ein paar kleine, gekommen. Das erste war zum Beispiel, da war der Herwig Friedl ja als Student noch dabei, diese Geschichte mit den Lungenfunktionsdaten. Damals hat in der Steiermark Prof. Karl Harnoncourt, ein Bruder von unserem Dirigenten Nikolaus Harnoncourt, das recht groß aufgezogen. AKL hat das geheißen, das Kürzel. Da ist ein Bus etabliert gewesen, der ist in der ganzen Steiermark herumgefahren und hat einfach die Leute zusammen getrommelt, um diesen Test zu machen. Das heißt, das waren natürlich keine selektierten Daten, natürlich nicht so, wie man sich das heute vorstellt bei einer wissenschaftlich sauberen Studie. Das war die sogenannte Normalbevölkerung einfach zusammen gefischt. Die Daten haben wir dann bekommen und das waren, glaube ich, 15000 oder so etwas – erstaunlich viel für die damalige Zeit.

Herwig Friedl: *Das war ein Big-Data-Problem.*

Ernst Stadlober: Ja und wir sind damals an die Grenzen vom UNIVAC Rechner und dem Programmpaket BMDP gekommen. Wir haben wirklich versucht die Daten zu modellieren. Das waren zwar nur multiple Regressionsmodelle, aber immerhin. Für die damalige Zeit ein irrsinniger Rechenaufwand und da kann der Herwig auch gewisse Schmankerl erzählen, was da alles drinnen war, in diesen Daten bei der Selektion – Identifikation der Leute, zum Beispiel.

Oder ein und derselbe Mensch hat einmal „Joseph Maier“ geheißen und einmal „Maier Joseph“ und einmal „Dr. Maier Josef“. Da haben wir dann, kannst du dich noch erinnern, solche Selektionen gehabt von Listen von Leuten wo wir erkannt haben, der war fünf mal so drinnen und fünf mal anders. Weil die Sozialversicherungsnummer vielleicht auch nicht dabei war. Also es war gar nicht so einfach.

Herwig Friedl: *Aber es war eine gewaltige Datensituation jedenfalls. Ich bin mir nicht so sicher, ob du mit* 15000 *da richtig liegst, ich hab so irgendwie sogar* 80000 *im Kopf.*

Ernst Stadlober: Im Endeffekt haben wir daraus Österreichische Normwerte entwickelt, für die Mediziner, eingebaut in ihre Apparaturen. Für die Berechnung haben wir die Daten dann natürlich vorher selektiert. Dafür waren immer noch auswertbare Daten in der Größenordnung von über 10000 Probanden vorhanden. Der Vorteil war, dass wir eine europa- oder sogar weltweit einzigartige Bandbreite bzgl. des Alters gehabt haben: Untersuchungen von sechsjährigen Kindern bis 85-jährigen Senioren. Das heißt, die Leute, die mit 85 Jahren noch zum Bus gehen konnten, die waren noch dabei. Es war in dem Sinn natürlich nicht mehr die Normalbevölkerung dieser Altersgruppe, sondern es waren nur die Gesunden drinnen, die extrem Gesunden. Aber immerhin haben wir dann Modelle entwickeln können für diese große Altersbandbreite, was es zur damaligen Zeit sonst noch gar nicht gegegeben hatte.

Herwig Friedl: *Kannst du dich noch an diesen Hilfeschrei erinnern, den du bekommen hast, als bei einem Patienten das Modell angewandt worden ist und eine negative geschätzte Vitalkapazität resultiert hat.*

Ernst Stadlober: Ja, das Schmankerl muss ich erzählen. Da ruft mich der Arzt aus der Lungenabteilung des LKH an, ganz vorwurfsvoll, dass unser Modell falsch ist: bei einem Probanden kommen negative Werte heraus. Ich habe mir gedacht: „Das gibt es ja gar nicht.“. Weil wir hatten das Modell damals evaluiert bezüglich aller möglichen Bereiche, da waren wir immer sehr gewissenhaft und genau und haben jedes Mal, wenn wir irgendetwas aus der Hand gegeben haben, exakt dokumentiert, wofür das Modell gilt; Alter von-bis, Größe von-bis, Geschlecht – immer genau angegeben. Dann frage ich da weiter und sag': „Was war denn das für ein Patient?“. Dann hat sich herausgestellt, dass das ein Liliputaner war. In unserem Modell ist natürlich die Körpergröße sehr wesentlich eingegangen – das ist klar. Und der Liliputaner war 1.20 m und für den war natürlich alles falsch. Das Modell war für die Größen von 1.60 m weg bei den Männern oder so, zulässig. Der Arzt hat das aber extrapoliert hinunter zur Körpergröße des Liliputaners und ist dann natürlich zu negativen Modellwerten gekommen. Daraufhin habe ich den Arzt gefragt: „Haben Sie überhaupt einmal nachgesehen, wofür das Modell eigentlich gilt? Sehen Sie dort irgendeinen Hinweis, dass das auch für Körpergröße 1.20 m gilt?“. Ja, dann hat er schön langsam eingesehen, dass das nicht unser Fehler war, sondern er in der Anwendung auch aufpassen muss. Ähnliche Geschichten haben wir oft erlebt mit Anwendern, die uns vorgeworfen haben, dass wir etwas falsch gemacht haben. Aber meistens haben wir das dann ausräumen können, da irgendetwas in der Anwendung nicht so passierte, wie das eigentlich vorgesehen war. Also wir haben generell immer versucht, das, bevor wir etwas aus der Hand gegeben haben, immer für uns zu evaluieren und abzuchecken, dass da nichts schief gehen kann – also nichts grob schief gehen kann.

Ja, aber es war damals eine wertvolle Erfahrung, diese Lungenfunktionen, weil da haben wir das erste Mal, wie der Herwig richtig gesagt hat, mit sogenannten Big-Data gearbeitet. Wir haben also die Möglichkeit gehabt uns durch diese Datenflut zu wühlen und zu selektieren. Das heißt, wir haben eigentlich ein wenig selber geübt und trainiert – wir waren ja auch am Anfang. Heute würde ich das natürlich ganz anders machen, das ist ja klar mit der Erfahrung. Aber immerhin, wir haben eine sehr gute Case-Study gehabt aus der Praxis. Und ein anderes Erlebnis ist auch ganz lustig. Wir haben Untersuchungen gehabt von bestimmten Gruppen wie Feuerwehrmänner zum Beispiel. Aichfeld – Murau war damals ein großes Thema in den 80er Jahren, wie die Draken kommen sollten in die Steiermark – die Vorgänger vom Eurofighter. Da hat es in der Steiermark die sogenannte Draken-Studie gegeben, vom Landeshauptmann beauftragt, und da sollten bestimmte Fachleute eben untersuchen, ob die Region Aichfeld vielleicht schon so stark belastet ist, dass zusätzliche Düsenjäger quasi keinen Sinn machen. Da sind wir beauftragt worden, zu vergleichen, ob es Unterschiede gibt in den Lungenfunktionswerten von Personen im Aichfeld und von Personen der Region Murau, aufgrund der Umweltsituation. Da war im Prinzip kein Unterschied nachweisbar. Da hatten wir auch Daten von

Feuerwehrleuten, daraus haben wir für unsere Studenten einen Datensatz ausgewählt für die Lehre, der heißt `aimu`-Datensatz. Dieser geistert eben seit Ende der 80er Jahre herum und ist ein recht einfacher Datensatz mit 79 Datenpunkten, wo aber von jedem Feuerwehrmann bestimmte Merkmale drinnen sind. Das heißt du kannst einfache Modelle erklären, du kannst die Verteilungen der Merkmale untersuchen, da hast du stetige Variablen, du hast Gruppierungsvariablen und der Datensatz bringt für uns selber in der Lehre immer wieder neue Überraschungen, weil man die Daten immer wieder aus neuen Blickwinkeln anschaut. Und das ist mittlerweile für uns ein legendärer Datensatz. Wenn wir `aimu` erwähnen, lachen wir schon und sagen: „Das ist unser `aimu`-Datensatz, den haben wir schon so lieb gewonnen. Den geben wir nicht mehr aus der Hand.“. Und den Studenten ist er auch schon sehr vertraut.

Herwig Friedl: *Gekommen bist du ja, wie du schon gesagt hast, von der Zufallszahlengeneratorebene. Dann hast du dich, daran kann ich mich noch gut erinnern, beworben als Nachfolger um die Stelle von Sepp Gölles. Das war für mich damals ziemlich interessant, da ich mir nicht unbedingt vorstellen konnte, dass jemand von der random numbers Seite kommend, in die Angewandte Statistik so ohne irgendetwas hinein schlittert. Du hast die Stelle bekommen, und dann hast du richtig losgelegt mit Anwendungen und Projekten, hast an das Institut eine gewaltige Anzahl von größeren, kleineren, andauernden Projekten mit Firmen, auch mit dem Land Steiermark oder mit der Stadt Graz, geholt und jetzt mehr oder weniger dieses Institut, zumindest diese Seite vom Institut aufgebaut, hier im Land Steiermark eine extrem gut anerkannte, perfekt positionierte Gruppe von Leuten etabliert, die immer wieder in der Lage sind, derartige Studien durchzuführen.*

Und das was mir jetzt am Schluss sehr gut gefallen hat, war die Zusammenarbeit mit dem Umweltamt, also dein tolles Engagement in Richtung `PM10` oder dergleichen. Am Anfang glaube ich war dies eher eine sehr lokale Problematik wo du da mitgespielt hast – Graz betreffend. Das ist in der Zwischenzeit aber schon viel umfassender geworden. Ihr habt auch zusammengearbeitet mit den Südtirolern, Bozen und so weiter. Kannst du uns da ganz kurz so einen Sketch von dieser Linie erzählen?

Ernst Stadlober: Naja das war so, 1997 habe ich diese Professur, die Nachfolge vom Sepp Gölles, bekommen, und das war natürlich Angewandte Statistik. Da habe ich mir gesagt, ok, es ist klar, jetzt habe ich diese Stelle „Angewandte Statistik“, ich hatte auch als Vorbild Josef Gölles, der ja wirklich einer der Pioniere da war in Graz, der die Statistik nach außen getragen hat in unterschiedliche Bereiche. Er hat auch im Rechenzentrum eine kleine Gruppe gehabt und dann im Joanneum Research, und das war für mich ein Vorbild, dass ich mir sagte, schau, so etwas könnte ich weiter treiben in der Art: ich habe eine gute akademische Basis, habe auch gute Studenten aus der Mathematik zur Verfügung und ich versuche die Studenten schön langsam in diese Richtung zu bringen, jene die interessiert sind für angewandte Projekte. Und wie gesagt, dadurch, dass wir vorher diese Lungenfunktionsgeschichten schon sehr intensiv betrieben haben, bin ich schon selber ein bisschen auf den Geschmack gekommen, dass ich mir dachte, das dürfte auch allgemein interessant werden für andere Themen. Der Anfang war bei uns die medizinische Statistik, von der Anwendung eben, über die Lungenfunktionsschiene. Und diese Umweltproblematik ist dann gekommen, weil ich Kontakte gehabt habe zum Umweltamt und da hatte Graz ein Problem als man angefangen hat diese `PM10`-Werte zu messen, und die EU–Grenzwerte, die neu eingeführt wurden, nicht einhalten konnte. Dann hat man gleich erkannt: „Graz hat ein Problem, was machen wir?“. Daraufhin haben wir uns zusammengesetzt; das Umweltamt und ich mit Partnern aus der TU Graz haben dann gemeinsam ein EU-Projekt beantragt, zusammen mit Klagenfurt und Südtirol, Bozen, wo wir versucht haben, diese `PM10`-Problematik zu bearbeiten. Da waren einige Fachleute dabei aus anderen Bereichen, die technische Untersuchungen und Messungen gemacht haben. Oder wir haben auch eine Firma aus Bayern gehabt, die ein neues Messsystem ausprobiert hat. Unsere Aufgabe war es statistisch zu untersuchen, wie es mit der Ver-

teilung von PM10–Konzentrationen aussieht. Haben wir eine Möglichkeit Prognosen zu rechnen für PM10 Tagesmittelwerte, damit man vielleicht, wenn die Politik auf irgendein System einsteigt, über die Prognosen Hinweise geben kann, wann gewisse Maßnahmen zu setzen sind. Das war eben der Ausgangspunkt, dieses EU-Projekt, wo ich gemerkt habe, dass man als Statistiker mit ganz unterschiedlichen Partnern zusammen arbeiten kann. Als Quintessenz des Projekts haben wir Modelle entwickelt für die Prognose von Tagesmittelwerten, sowohl für Bozen und Klagenfurt, als auch für Graz. Das einzige, was seitdem noch laufend umgesetzt wird sind die Prognosen in der Wintersaison in Graz, weil wir für Graz die Daten online im steirischen Messystem haben und das Grazer Umweltamt so offen ist und uns jedes Jahr damit beauftragt. Für die anderen zwei Städte haben wir zwar die Modelle entwickelt, aber das hat nicht gelebt, weil sich dort niemand mehr darum gekümmert hat. Das ist jetzt so ein Projekt, das läuft mittlerweile seit 2003 glaube ich. Ich habe dann auch Kontakte zu Kollegen in Brünn gehabt und mit denen zusammen, die haben auch ähnliche Modelle für Brünn entwickelt, wurde dann vor kurzem eine Arbeit geschrieben mit einem Modellvergleich, wo wir gezeigt haben, dass wir beides recht gut einbetten können in ein gemeinsames System. Im Prinzip sind das einfache GLMs oder multiple Regressionsmodelle. Das Modell selber ist da nicht so das Entscheidende, sondern das Entscheidende ist: „Was soll ich in das Modell reinbringen?“. Das ist da die Hauptarbeit gewesen. Welche Messgrößen beschreiben mir jetzt zum Beispiel die Meteorologie am besten. Und welche Merkmale bekomme ich auch als Messungen und was habe ich dann sonst noch zur Verfügung.

Das heißt die Hauptarbeit ist da nicht die Methodikarbeit in der Statistik, sondern das Vorfeld, dass ich mir überlegen muss von der Sache her, welche Merkmale habe ich zur Verfügung, welche kann ich zuverlässig messen und welche bringen dann für mein Modell auch Information. Das habe ich auch bei den anderen Projekten immer wieder erkannt: das Wichtigste ist das Vorfeld, dass ich einmal abchecke, was liegt vor und wie kann ich die relevanten Merkmale herausfinden – welche Merkmale soll ich beobachten, damit ich zu einem plausiblen Modell komme mit einer Aussagekraft für die Anwendung.

Foto 1997: Professor

Herwig Friedl: *Gerade die verschiedenen methodischen Aspekte, die du bei deinen Projekten immer angewandt hast, haben als Quintessenz dann auch eine Modifikation der Lehre hier vor Ort gehabt. Ich denke nur, eine Vorlesung wie Stochastische Simulation hat es am Anfang, als ich studiert habe, nicht gegeben. Eine Vorlesung Angewandte Statistik mit den Inhalten, wie du sie jetzt machst, Design of Experiments, Time Series, das hat es alles mehr oder weniger vor 30 Jahren nicht gegeben. Also haben ja auch diese praktischen Aspekte irgendwie dahingehend einen Einfluss gehabt. Wie siehst du das?*

Ernst Stadlober: Das ist wirklich eine sehr gute und spannende Frage, weil ich bin der Ansicht, dass derjenige, der eine Vorlesung durchführt, selbst am meisten davon profitiert. Ich lerne durch eine Vorlesung. Und ich habe immer den Ansatz gehabt, so wie der Herwig gesagt hat, ok, jetzt haben wir bestimmte Themen, wir brauchen gewisse Methoden – wie lernen wir die selber am besten? Ich muss die Methoden und deren Anwendungen ja selber beherrschen. Das Beste ist, wenn ich eine Vorlesung mache auf dem Gebiet, dann lerne ich auch die Details sauber. Ich muss ja tiefer gehen, ich muss den Studenten das beibringen und gleichzeitig lerne ich den Stoff. Sozusagen: ich bin dann der, der Hauptprofiteur ist! Das war typisch bei der Zeitreihenvorlesung zum Beispiel, die hat es vorher nie gegeben, da war niemand dafür da und es war aber der Bedarf da von Studierenden aus anderen Fachrichtungen, Zeitreihen zu lernen. Dann habe ich mir gedacht, mache ich halt die Vorlesung Zeitreihenanalyse, da muss ich mich jetzt einmal hineinknien und das gut verstehen von der Methodik. Ich habe das generell so gemacht mit den Vorlesungen, ich habe immer versucht die mathematische Basis ganz

sauber zu machen, aber sofort auch mit Anwendungen verbunden. Das heißt, ich habe in jeder Vorlesung natürlich den Computeraspekt drinnen – das heißt, es gibt bei mir keine Statistikvorlesung, wo nicht praktische Übungen sind mit *R*, wo ich selbst auch die Beispiele mit dem *R* rechne. Die Studenten haben dann kleine Projekte als Übungsaufgaben. Das heißt diese Kombination ist für mich etwas ganz Wertvolles, auch in der Vorlesung: zuerst die Grundmethodik und gleich, wenn die Leute die Grundmethodik können, sofort: „Wie schaut die Umsetzung aus?". Und wie gesagt, wir haben ja auch nicht nur bei den Vorlesungen diesen Aspekt, dass wir selbst etwas lernen wollen, sondern wir haben bewusst auch unsere Seminare gemeinsam immer so ausgewählt, dass wir uns sagen, jetzt haben wir ein Thema, da gibt es jetzt neue Methoden in dem und dem Bereich, da wissen wir noch zu wenig – machen wir ein Seminar. Das heißt, der Herwig und ich haben jahrelang jedes Semester ein Seminar gehalten, wo wir immer – auch wir als Seminarleiter – gelernt und vorgetragen haben. Wir haben bewusst solche Themen gewählt, da wissen wir noch nicht so gut Bescheid, gehen wir in den Bereich hinein – zusammen mit den Studenten. Das heißt, das war für mich der Sinn von Seminaren, dass man nicht irgendein totes Thema gibt, damit die Studenten beschäftigt sind, sondern wir haben immer versucht auch das mit einem Lernprozess für uns zu verknüpfen. Den Blickwinkel haben wir eben gehabt und deswegen haben wir so viele Anwendungen in breiten Bereichen. Ja gut, das ist jetzt verknüpft damit, dass wir uns selber weiter gebildet haben, gemeinsam, sehr breit in der Statistik. Das war nicht nur ich, sondern der Herwig hat ja auch mitgemacht, das heißt, wir haben beide jetzt wirklich ein sehr vielfältiges, breites Portfolio an Methoden, die wir kennen. Ich habe also Seminarordner und elektronische Unterlagen, wo Beispiele ausgearbeitet sind von jedem Bereich und mir ist es oft schon passiert, dass irgendjemand kommt, der sagt, ich habe ein Problem in dem Bereich, dann schnappe ich den entsprechenden Ordner und sage, da hast du den Ordner, da drinnen sind genau die Grundlagen für das, was du brauchst. Da brauchst du kein Buch, da haben wir ausgearbeitete Beispiele von Studenten, die super präsentiert wurden. Darauf haben wir auch immer Wert gelegt: das Seminar war bei uns einerseits lernen für alle und andererseits, die Studenten sollten ja nicht nur den Inhalt lernen, sondern auch die Präsentation beherrschen. Das heißt, wir haben auch sehr viel Wert darauf gelegt, dass die Studenten die Präsentation schriftlich sauber ausarbeiten und dann vortragen. Das was sie programmiert hatten war offenzulegen. Das alles war für uns ein Anliegen, weil heute sollen die Absolventen ja auch diese Softskills können. Und das haben wir versucht mit Inhalt zu füllen. Es war für mich nie ein Anliegen, Studenten in einen Rhetorik-Kurs zu schicken. Da diskutieren sie irgendein Thema und lernen präsentieren, wie man den Computer einschaltet und wie man das vorführt. Wie gesagt, bei uns muss die Präsentation verknüpft sein mit Inhalt. Dann lernt man das für sein Fach. Wie muss ich jetzt mein Fach rüberbringen in der Präsentation und da bin ich nicht unbescheiden, da braucht man nicht unbescheiden sein, weil das hat wirklich gut funktioniert. Wir sehen immer wieder wie toll die Studierenden dann bei der Präsentation der Masterarbeit oder Dissertation agieren. Wir haben jetzt kaum mehr irgendwelche Präsentationen, wo ich sage, die sind nicht brauchbar. Da haben die Studenten auch bei uns ein sehr hohes Niveau erreicht, eben durch diese Schulung im Seminar.

Herwig Friedl: *Jetzt wirst du uns ja unglücklicherweise in absehbarer Zeit leider abhanden kommen und diese vier Wände mit privaten Wänden zurück tauschen. Ich möchte zumindest von dir jetzt, vielleicht auch aus rein persönlichem Interesse, ein bisschen darüber etwas hören, wie du die Rolle der Statistik, die Ausbildung in der Statistik, die Zukunft der Projekte hier an der TU Graz siehst.*

Ernst Stadlober: Naja, die Zukunft hängt natürlich stark davon ab von welchen Personen das weiter betrieben werden kann. Einer der Mitstreiter ist ja der Herwig, Gott sei Dank, der bleibt ja doch noch einige Jahre tätig und meine Hoffnung ist, dass eben

meine Nachfolge durchaus auch aus dem Holz geschnitzt ist, dass sie nicht nur im stillen Kämmerlein sitzt und die eigenen Forschungsinteressen verfolgt, sondern auch so ein bisschen ein Typ ist, der vielseitig ist und sich auch nicht scheut da reinzugehen in die Anwendung. Was wir liefern können ist ein gut ausgebautes Netzwerk, womit man am Beginn Unterstützung anbieten kann. Da habe ich sicher kein Problem, dass ich in der Übergangszeit dann mithelfe, wenn noch eine gewisse Scheu besteht Kontakte aufzunehmen. Es wäre schade, weil das Netzwerk ist ja schon vorher vom Sepp Gölles gelegt worden, und ich habe das was der Herr Gölles angefangen hat im Wesentlichen weiter betrieben – natürlich in meiner, in unserer, anderen persönlichen Art, aber ich würde schon sagen, ich habe auch schon in gewisser Weise auf ein bestimmtes Netzwerk aufgebaut. Ich habe es dann halt mit meinen Möglichkeiten ausgefüllt und weiter ausgebaut. Ich hoffe, dass meine Nachfolge das auch nutzt. Es gibt gute Voraussetzungen von verschiedenen Seiten, sie kann das mit ihren neuen Ideen füllen und das vielleicht weiter treiben in Richtung einer Verknüpfung zwischen der guten mathematischen Ausbildung und der Anwendung in Projekten. Und wie gesagt, wir haben immer wieder gute Studenten, die Studierendenbasis ist bei uns sehr gut, die sind immer interessiert an Projekten und ich habe da meistens sehr gute Erfahrungen gemacht mit Studierenden, die in Firmen gearbeitet haben. Die Mitarbeiter in Firmen sind dann immer sehr überrascht von der hohen Qualität unserer Studenten. Ich habe schon einige Studentinnen gehabt, die in Firmen hineingegangen sind, wo keine einzige Frau gearbeitet hat, zum Beispiel bei Magna-Steyr in Graz. Diese Diplomandin hat eine Versuchsplanung gemacht für Klebeprozesse, wo sie den eigenen Versuchsaufbau selber zusammengebaut hat auf einem großen Tisch. Die konnte sich problemlos dort behaupten. Die Mitarbeiter dort waren sehr glücklich, dass endlich auch einmal eine Frau dabei ist, die ein anderes Leben in die Bude bringt. Und wie gesagt, die hat das super gemacht.

Zweites Beispiel, vor kurzem bei der AT&S in Hinterberg. Da wurde jetzt auch eine Masterarbeit fertig gestellt. Bei einem Termin mit dem erweiterten Vorstand haben die nur so geschwärmt von dieser Studentin: „Die müssen wir behalten, weil die ist so gut und die passt so gut rein in unser Team.“. Mittlerweile arbeitet sie auch dort, aber wie gesagt, das ist für mich auch so, dass man dadurch auch gewisse Hürden abbaut, indem man die Studentinnen einfach integriert. So wie bei anderen Problematiken, wenn man etwas nicht kennt, lehnt man einmal ab, aber wie gesagt, da hat es bei uns immer eher positive Erfahrungen gegeben. Frauen bewähren sich jetzt im technischen Bereich. Wir haben vor Kurzem so ein Brainstorming gemacht, was haben wir für eine Geschlechterverteilung bei den Studierenden? Unsere 50 abgeschlossenen Master/Diplomarbeiten in den letzten 10 Jahren wurden von 25 Frauen und 25 Männern geschrieben. Wir erreichen also eine exakte Geschlechterparität. Also das ist bei uns ganz normal ohne viel Zusatzaufwand passiert. Wir haben das irgendwie einfach so genommen wie es ist und wir haben viele gute Studenten ausgewählt, und da waren halt auch genug gute Frauen dabei. Ich kann jetzt nicht sagen, dass ich mit den Studentinnen prinzipiell andere Erfahrungen gemacht habe, als mit den Studenten. Das war ziemlich ausgewogen: in jeder Gruppe gab es Gute und weniger Gute.

Matthias Templ: *Ich möchte nur ganz kurz hier anknüpfen und habe eine letzte Frage. Also ich nehme dich auch kurz vor der Emeritierung wahnsinnig aktiv wahr.*

Ernst Stadlober: Pensionierung heißt das bei mir! [*lacht*]

Matthias Templ: *Und ich erlebe auch, dass ihr mit Euren Studenten ein familiäres Verhältnis habt.*

Herwig Friedl: *Ab und zu tanzt er sogar mit ihnen! [lacht]*

Matthias Templ: *Aber wirst du uns in der Statistik auch weiter erhalten bleiben?*

Ernst Stadlober: Also, wie gesagt, meine Pläne in nächster Zeit sind so, dass ich in der Übergangsphase noch im Pflichtlehrprogramm mitmachen werde. Ich arbeite bereits an kleineren Projekten mit Partnern, die noch über meine Pensionierung hinaus laufen. Wir haben ja auch Kontakte zu unseren Kompetenzzentren, zum Beispiel gibt es das ViF (das virtuelle Fahrzeug) die ganz interessante Problematiken haben, da sitzen auch Diplomanden und Absolventen von uns. Da ist es nach wie vor gut, wenn man einfach mit der Expertise, die man hat als erfahrener Statistiker, dabei ist. Mein Rat wird ja auch offensichtlich noch gewünscht. Ich dränge mich nicht auf, aber anscheinend ist es doch noch so, dass meine Expertise nachgefragt wird. Wie gesagt, ich bin natürlich mit Leib und Seele Statistiker und das werde ich bis zu meinem Lebensende bleiben, aber ich habe auch Gott sei Dank eine Familie, in der ich gut aufgehoben bin, mit einer großen Kinderschar und auch Enkel, das heißt, ich versuche halt beides irgendwie unter einen Hut zu bringen – bis jetzt ist das halbwegs gelungen – und wenn die Gesundheit mitspielt, das weiß man ja nie, werde ich sicher noch einige Zeit mich da einbringen, wenn es gewünscht ist. Das ist keine Frage.

Was wir jetzt nicht konkret erwähnt haben ist, dass wir wahnsinnig viele Diplomanden haben. Also, wir haben überproportional viele im Vergleich zur Mathematik. Da haben wir gerade gestern noch mit dem Vizerektor ein Gespräch gehabt, dass wir da auch extrem stark die ganze Studienrichtung tragen, OR und Statistik, und zusätzlich noch die Finanzmathematik. Ich habe ja auch so eine kleine Präsentation zusammen gestellt, eine Übersicht von den Projekten und Diplomarbeiten, wo auch Zahlen drinnen stehen. Was wir auch an Dissertationen und so betreut haben insgesamt, am Institut. Ich bin ja nicht alleine, aber der Herwig und ich haben in der Statistik gemeinsam sehr viel gemacht – beide haben wir jetzt sehr viel betreut, oder betreuen immer noch. Also, wie gesagt, das ist für mich jetzt schon ein Zeichen, dass es schade wäre, wenn durch die Nachfolgegeschichte irgendetwas schief geht. Wir haben auch einen guten Stand innerhalb der TU Graz, weil ich mit Ingenieuren immer viel zusammen gemacht habe, z.B. Dissertationen betreut. Das hat natürlich auch die Wirkung, dass wir innerhalb der TU Graz auch eine gewisse Anerkennung erfahren. Die schätzen unseren Bereich durchaus. Wir waren natürlich auch meistens bereit, als Ansprechpartner da zu sein. Dazu muss man halt auch bereit sein – es ist immer ein Geben und Nehmen.

Matthias Templ: *Dann möchte ich mich herzlich für das Gespräch bedanken.*

Herwig Friedl: *Genug des Guten.*

Ernst Stadlober: Ich danke auch. Genug des Guten.

Affiliation:

Ernst Stadlober
Institute of Statistics
Graz University of Technology
Kopernikusgasse 24/III
A-8010 Graz, Austria
Email: e.stadlober@tugraz.at

Herwig Friedl
Institute of Statistics
Graz University of Technology
Kopernikusgasse 24/III
A-8010 Graz, Austria
Email: HFriedl@TUGraz.at

Matthias Templ
Vienna University of Technology &
Statistik Austria
A-1040 Vienna, Austria
E-mail: matthias.templ@gmail.com

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 45
March 2016

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2015-12-18
Accepted: 2016-01-26

~~NEWS~~ FUN

Are statisticians approximative normal?

Since the answer is no, we can continue with a “true” story related to the fourth paper in this issue.

A statistician’s wife had twins. He was delighted. He rang the minister who was also delighted. “Bring them to church on Sunday and we’ll baptize them,” said the minister. “No,” replied the statistician. “Baptize one. We’ll keep the other as a control.”

A statistician is a mathematician broken down by age and sex.

Statisticians do it normal.
Statisticians do it when it counts.
Statisticians do it with 95% confidence.
Statisticians do it with a large number...
Statisticians do it with only a 5% chance of being rejected.
Statisticians do it with two (samples) at the same time.
Statisticians probably do it.

If you choose an answer to this question at random, what is the chance you will be correct?
A) 25% B) 50% C) 60% D) 25%

ISSN 1026–597X

Contents