DESIGN AND ESTIMATION IN BUSINESS SURVEYS

SELECTED TOPICS

BY PATRICIA DÖRR

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

DR. RER. POL.

TO THE DEPARTMENT IV
AT TRIER UNIVERSITY

ON 09 JUNE 2020

SUPERVISORS:

PROF. DR. RALF MÜNNICH (TRIER UNIVERSITY)
PROF. DR. LEONHARD FRERICK (TRIER UNIVERSITY)

PATRICIA DÖRR
Augustinusstraße 15

54296 Trier
pdoerr@freenet.de

EDUCATION:

Since October 2019:	Master in Mathematics, FernUniversität Hagen, Germany
April 2016 – June 2020:	Doctoral Research in Survey Statistics, Trier University, Germany
October 2015 – March 2019:	Bachelor in Business Mathematics, Trier University, Germany
February 2018 – August 2018:	Erasmus scholarship, Università degli Studi di Firenze, Italy
October 2013 – March 2017:	Master in Economics (Focus on European Political Economy), Trier University
October 2013 – March 2016:	Master in Survey Statistics, Trier University
September 2012 – June 2013:	Exchange phase of the bi-national (Germany - France) Bachelor in Economics and Management, (Spécialité Monnaie, Banques et Finance) Université Paris X, France
April 2010 – June 2013:	Bachelor in Economics and Management, Johannes Gutenberg-University Mainz, Germany
March 2010:	Abitur, Carl-Bosch-Gymnasium Ludwigshafen, Germany

ABSTRACT

Estimation and therefore prediction – both in traditional statistics and machine learning – encounters often problems when done on survey data, i.e. on data gathered from a random subset of a finite population. Additional to the stochastic generation of the data in the finite population (based on a superpopulation model), the subsetting represents a second randomization process, and adds further noise to the estimation. The character and impact of the additional noise on the estimation procedure depends on the specific probability law for subsetting, i.e. the survey design. Especially when the design is complex or the population data is not generated by a Gaussian distribution, established methods must be re-thought. Both phenomena can be found in business surveys, and their combined occurrence poses challenges to the estimation.

This work introduces selected topics linked to relevant use cases of business surveys and discusses the role of survey design therein: First, consider micro-econometrics using business surveys. Regression analysis under the peculiarities of non-normal data and complex survey design is discussed. The focus lies on mixed models, which are able to capture unobserved heterogeneity e.g. between economic sectors, when the dependent variable is not conditionally normally distributed. An algorithm for survey-weighted model estimation in this setting is provided and applied to business data.

Second, in official statistics, the classical sampling randomization and estimators for finite population totals are relevant. The variance estimation of estimators for (finite) population totals plays a major role in this framework in order to decide on the reliability of survey data. When the survey design is complex, and the number of variables is large for which an estimated total is required, generalized variance functions are popular for variance estimation. They allow to circumvent cumbersome theoretical design-based variance formulae or computer-intensive resampling. A synthesis of the superpopulation-based motivation and the survey framework is elaborated. To the author's knowledge, such a synthesis is studied for the first time both theoretically and empirically.

Third, the self-organizing map – an unsupervised machine learning algorithm for data visualization, clustering and even probability estimation – is introduced. A link to Markov random fields is outlined, which to the author's knowledge has not yet been established, and a density estimator is derived. The latter is evaluated in terms of a Monte-Carlo simulation and then applied to real world business data.

ACKNOWLEDGEMENTS

The author would like to acknowledge the fruitful conversations with Ralf Münnich, his trained eye on discrepancies in simulation set-ups, intellectual and moral support and a lot of freedom for trials and errors. In addition, but not limited to this, he pointed me to the topic of Chapter 3.

Thanks go to Jan Pablo Burgard who pointed me to the topic of Chapter 2 and gave many advices to the design of scientific papers and Monte-Carlo simulation studies.

Regarding the mathematical considerations in Chapter 4, the author is grateful for the spontaneous meetings with Leonhard Frerick and his quick familiarization with a new mathematical field. The topic came up during an exchange semester at the University of Florence and meetings with Alessandra Petrucci. Thanks for that great experience.

Finally, the author thanks the German Federal Statistical Institute that has financially supported the position as research assistant.

CONTENTS

I ESTIMATORS FOR REGRESSION, FINITE POPULATION VARIANCE AND PROBABILITY DENSITY 1

1 INTRODUCTION 3
1.1 Motivation . . . 3
1.2 Outline . . . 6
1.3 Introduction to Point Estimation Theory and Survey Sampling . . . 7

2 MULTILEVEL REGRESSION ANALYSIS IN BUSINESS SURVEYS 17
2.1 Introduction to Design-sensitive Regression Analysis . . . 17
2.2 Survey-weighted Mixed Models . . . 21
2.2.1 Model Formulation at the Population Level . . . 21
2.2.2 Likelihood Approach under Survey Sampling . . . 25
2.2.3 Maximization of the HT Joint Log-likelihood . . . 28
2.2.4 Simulation Studies . . . 33
2.3 Application to Business Surveys . . . 40
2.3.1 Theoretical Background . . . 40
2.3.2 Data Description . . . 42
2.3.3 Estimation and Evaluation . . . 42

3 GVFS IN BUSINESS SURVEYS 49
3.1 Introduction to Variance Estimation . . . 49
3.2 Theoretical Background of the GVF . . . 52
3.2.1 A Model-based Motivation of the GVF . . . 52
3.2.2 Error Decomposition of the GVF Prediction . . . 56
3.2.3 GVFs and the Design Effect . . . 60
3.2.4 GVFs in a Design-based Framework . . . 61
3.3 GVFs in Practice . . . 62
3.3.1 Common Shapes of GVFs . . . 63
3.3.2 Quality Measures . . . 65
3.3.3 Estimation of GVFs . . . 68
3.3.4 GVFs in Small Area Estimation . . . 69
3.4 Simulation Study . . . 70
3.4.1 DGP in the Model-design Framework . . . 71
3.4.2 Objectives of the Study . . . 72
3.4.3 Simulation Results . . . 73

4 A MULTIVARIATE PROBABILITY DENSITY ESTIMATOR USING SOMS 79
4.1 Introduction to SOMs . . . 79
4.2 The Basic SOM . . . 80
4.2.1 Algorithmic Description . . . 80

4.2.2 Properties of the SOM 84
4.3 Distance Measures in SOMs 87
4.4 Variations of the SOM 90
4.4.1 Simplex Arrangements 90
4.4.2 Related Energies and Random Field Theory 91
4.4.3 Optimization of an Alternative Energy Function . . 96
4.5 Growing SOMs 100
4.6 Summary of the Final Algorithm and Theory 103
4.7 Machine Learning with Survey Data 106
4.8 Simulation Study 107
4.8.1 Simulation Set-up 107
4.8.2 Simulation Results 110
4.8.3 Some Remarks on the Simulation Study 112
4.9 Application to Business Surveys 113
4.10 Conclusion 116
5 SUMMARY AND OUTLOOK 119

II APPENDIX 123
A PROOFS 125
B ADDITIONAL INFORMATION ON (SIMULATION) DATA 137

BIBLIOGRAPHY 151

LIST OF FIGURES

Figure 2.1 MC-Distribution of $\hat{\beta}$ under $Y_i^{(1)} \sim \Gamma(1/\eta_i, 0.5)$ . . . 37
Figure 2.2 MC-Distribution of $\hat{\rho}$ under $Y_i^{(1)} \sim \Gamma(1/\eta_i, 0.5)$ . . . 38
Figure 2.3 MC-Distribution of $\hat{\lambda}$ under $h(Y_i^{(2)}) \sim N(0, 0.16)$ . . 39
Figure 2.4 MC-Distribution of $\hat{\beta}$ under $h(Y_i^{(2)}) \sim N(0, 0.16)$. 39
Figure 2.5 MC-Distribution of $\hat{\rho}$ under $h(Y_i^{(2)}) \sim N(0, 0.16)$ . . 40
Figure 3.1 MC Relative Error of Variance Estimators 75
Figure 3.2 Average R^2 of the GVF Regression vs. Average Relative Prediction Error 76
Figure 3.3 Average $\bar{\delta}_{0.95}$ of the GVF Regression vs. Average Relative Prediction Error 77
Figure 4.1 Concept of the SOM 81
Figure 4.2 MISE and KL of Estimators for P_1 111
Figure 4.3 MISE and KL of Estimators for P_2 112
Figure 4.4 MISE and KL of Estimators for P_3 113
Figure 4.5 Density Estimation of Total Sales – I 115
Figure 4.6 Density Estimation of Total Sales – Weighted . . . 116
Figure B.1 MC Performance of Direct Variance Estimators . . 147

2.1 Exogeneous Variables for Regression Models 43
2.2 Point Estimates for the Logit Mixed Models 44
2.2 Point Estimates for the Logit Mixed Models 45
2.2 Point Estimates for the Logit Mixed Models 46
B.1 Overview on the Electronic Appendices 137
B.1 Overview on the Electronic Appendices 138
B.1 Overview on the Electronic Appendices 139
B.2 Partitions in the Simulation Study in Chapter 2 140
B.3 Partitions in the Simulation Study in Chapter 3 140
B.4 MC Relative Error of GVF Prediction - Model (3.23a) - under SRS 141
B.5 MC Relative Error of GVF Prediction - Model (3.23a) - under StratRS 142
B.6 MC Relative Error of GVF Prediction - Model (3.23a) - under Stratified TSC 143
B.7 MC Relative Error of GVF Prediction - Model (3.23b) - under SRS 144
B.8 MC Relative Error of GVF Prediction - Model (3.23b) - under StratRS 145
B.9 MC Relative Error of GVF Prediction - Model (3.23b) - under Stratified TSC 146
B.10 MC Relative Error of GVF Prediction - Model (3.23c) - under SRS 146
B.11 MC Relative Error of GVF Prediction - Model (3.23c) - under StratRS 147
B.12 MC Relative Error of GVF Prediction - Model (3.23c) - under Stratified TSC 148
B.13 MC Relative Error of GVF Prediction - Model (3.23d) - under SRS 148
B.14 MC Relative Error of GVF Prediction - Model (3.23d) - under StratRS 149
B.15 MC Relative Error of GVF Prediction - Model (3.23d) - under Stratified TSC 150

LIST OF ALGORITHMS

Algorithm 2.1 MCEM Algorithm with Importance Sampling . . . 34
Algorithm 4.1 On-line SOM 82
Algorithm 4.2 Batch SOM 83
Algorithm 4.3 Extended SOM Mini-Batch Algorithm 105

ACRONYMS

AIC	Akaike Information Criterion
BEEPS	Business Environment and Enterprise Performance Surveys
bmu	Best Matching Unit
BSOM	Batch Self-organizing Map
cdf	Cumulative Distribution Function
DGP	Data Generating Process
CLT	Central Limit Theorem
EM	Expectation Maximization
FH	Fay-Herriot Estimator
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
GMM	Gaussian Mixture Model
GVF	Generalized Variance Function
GREG	Generalized Regression Estimator
HT	Horvitz-Thompson Estimator
iid	identically and independently distributed
LMM	Linear Mixed Model
MC	Monte Carlo
MCEM	Monte Carlo Expectation Maximization
MISE	Mean Integrated Squared Error
MSE	Mean Squared Error
ML	Maximum Likelihood
MRF	Markov Random Field
NSI	National Statistical Institute
OLS	Ordinary Least Squares
pdf	Probability Density Function
pps	Probability Proportional to Size
PSU	Primary Sampling Unit
PUM	Public Use Microdata
SAE	Small Area Estimation

SOM	Self-organizing Map
StratRS	Stratified Random Sampling
SRS	Simple Random Sampling
TSC	Two-Stage Cluster Sampling

Part I

ESTIMATORS FOR REGRESSION, FINITE POPULATION VARIANCE AND PROBABILITY DENSITY

1

INTRODUCTION TO CHALLENGES IN ESTIMATION IN BUSINESS SURVEYS

1.1 MOTIVATION

The inherent difference between business surveys and other surveys is the nature of the sampling unit: For example, social surveys sample persons or households and ecological surveys sample plants or animals. Businesses, however, are organizational entities. Like other entities such as households or non-human sampling units, a deputy respondent that speaks for the surveyed units. Furthermore, the nature of an organizational entity like the size and composition of manpower, the operating sector or returns are much more fluid than in other entities, say households. These properties require sometimes special survey designs to capture enough of the variability between businesses and within businesses over time without overloading the respective deputy respondent, especially in small businesses.

What is a business and when is a survey a business survey? In the following, we shall understand a survey as a random subsetting process of a finite index set where each index is assigned to one business unit, that is an organizational unit that provides goods and services in exchange for monetary funds. This means, that in the following, we exclude unlike Cox and Chinnappa [1995] institutions from the definition of business surveys because the peculiarites of businesses like properties of owners and topmanagers (cf. Section 2.3) or skewness of monetary flows (cf. Section 4.9) are not generally applicable for institutions.The terms 'firm' and 'business' are used interchangeably in the following.

Besides problems that are common to most surveys - such as missing data, frictions between sampling and study population, or the auxiliary use of administrative data - business surveys face very specific problems: Business sizes are highly skewed and whilst large businesses are often sampled with probability close to one [Cox and Chinnappa, 1995], there are often cut-off rules for the smallest firms [Bee et al., 2007]. This leads to extreme designs with very diverging design weights [Burgard et al., 2014] or sample informativity that can affect the quality of estimators [Pfeffermann and Sverchkov, 2007]. This becomes even more problematic when the variable of interest to be estimated is demanded for small geographic areas or business domains [Cox and Chinnappa, 1995, Rao and Choudhry, 1995] and in addition is not (conditionally) normally distributed [Chandra et al., 2009, Fabrizi et al., 2017].

Skewed distributions of the variable of interest lead to problematic estimators: Variances for the common Horvitz-Thompson Estimator (HT) [Horvitz and Thompson, 1952] increase and variance estimators can become instable. Alternative estimators like the Generalized Regression Estimator (GREG) can be applied [Deville and Särndal, 1992] because - using auxiliary information - they have possibly smaller variances than the HT. The possible efficiency gain using the GREG was demonstrated in a simulation study based on business survey data by Lee and Croal [1989]. Though asymptotically unbiased, Hedlin et al. [2001] demonstrate that an appropriate model choice is essential for the reliability of the GREG in practice. Furthermore, the GREG relies heavily on a linear relationship between the auxiliary data and the variable of interest. An inappropriate assisting linear model when there is a nonlinear relation between the variable of interest and the auxiliaries, can counteract the desired efficiency gain.

Regression analysis – both, an interest per se and the foundation of model-assisted (like the GREG)and model-based estimators in finite populations – with business data has challenges, too: When data do not meet normality assumptions and when the survey design is complex, classical estimators based on least squares or Maximum Likelihood (ML) do not rely anymore on correct distributional assumptions and can return (even asymptotically) biased results. The focus in this work shall lie on the violation of (conditional) distributive assumptions on the dependent variable. We consider skewed continuous random variables distributed according to a (modified) power-normal distribution and non-normal elements of the exponential family. Both types of dependent random variables cannot be modelled adequately using a standard linear model, but appplying Generalized Linear Models (GLMs) or linear models under data transformations.

It is possible that the data analyist desires, in addition, a mixed model structure, i.e. a combination between fixed and random effects, to account for unobserved heterogeneity, between, for example, different geographic locations (e.g. relevant in Small Area Estimation (SAE)) or economic sectors or between units in panel-structured data. Then, extreme survey designs are difficult to adjust for. The quality of the estimated model translates directly to the quality of predictions arising therefrom, which is relevant for finite population estimators relying on models [Valliant et al., 2000]. An illustration for estimators based on Gaussian Mixture Models (GMMs) is given in Burgard and Dörr [2018]. Furthermore, nonlinear transformations of the dependent variable prohibit the simple plug-in solution for conditional expectations/ predictions. A computationally feasible alternative has thus to be found when such models are of interest.

SAE can also serve as a link to another problem that is discussed in this work: Whilst the already named difficulties arise mostly in unit-lavel SAE, area-level SAE [Fay and Herriot, 1979] especially in business surveys encounter another problem: They theoretically require variances of the direct estimators, usually the HT, that are not available in practice. Plugging in estimators thereof, however, is problematic because they might be unstable due to small sample sizes and skewness of the variable of interest. Amongst others, Maples et al. [2009] and Kubacki and Jędrzejczak [2012] suggest to use Generalized Variance Function (GVF) predictions in the Fay-Herriot Estimator (FH) estimator to stabilize the point estimation.

Though GVFs have consequently even an application in point estimation, they might be of interest for second order statistics as well: Besides a reduced computational effort in case of complex survey designs, the assumption of a functional relation between an estimator's expectation and variance adds additional information to the estimation and therefore can reduce instabilities in variance estimation. Even when the sample size is adequate for a reliable first order statistics (that is, when SAE is not necessary), it might be insufficient for higher order statistics that are nonetheless important to judge the estimator's quality. Instabilities can result from extreme designs and peculiarities of business-related variables and may thus occur in business surveys. The behaviour of GVFs as an interesting alternative to direct variance estimators should consequently be studied theortically and using simulations.

In the subsequent chapters, the synthesis of distributional assumptions on the characteristics observed and survey design play therefore a major role: Both the estimation of model parameters from a complex survey design and prediction for non-observed units from the finite population must account for the additional randomness implied by the draw of a sub-sample under a probability law. We understand statistics under distributional assumptions on the characteristics attributed to an index in the finite population as model-based statistics. In contrast, statistics on the finite population concerned with the random subsetting process, is referred to as design-based statistics. The underlying probability space has as consequence also different settings for estimators. For the synthesis of model- and design based statistics, a joint probability law is required and will be defined in Section 1.3 together with implications on point estimation theory.

The probability law underlying the characteristics of the finite population, the so called Data Generating Process (DGP) is not always known. Sometimes neither information on a family of distributions characterized by a parameter space, which the DGP belongs to, is available. A non-parametric estimator of the distribution can thus be of interest, too, especially in multivariate statistics in order to get an explorative idea of

the data. A new estimator based on a machine learning algorithm, the Self-organizing Map (SOM) is introduced and serves as a contrast to traditional problems in model- and design-based statistics like parameter (point) and variance estimation for estimators of finite population totals under complex survey designs.

1.2 OUTLINE

The different chapters of this thesis aim at the study of the research questions implied by the challenges on business surveys named above. Though inherently connected through the application to business surveys, the research topics, i.e. mixed models under survey design, theory and simulation on GVFs under a joint mode-design framework and machine-learning based probability estimation can be distinctively splitted into different chapters. These are briefly introduced in the following.

A regression parameter estimator for Generalized Linear Mixed Models (GLMMs) and Linear Mixed Models (LMMs) under dual and Box-Cox transformation, introduced in Burgard and Dörr [2018] and Dörr and Burgard [2019] is outlined in Chapter 2. The properties of the estimator are studied and compared to existing methods in novel simulation studies and the estimator is applied to business data. The simulation set-up is such that it corresponds to the joint design-model probability measure introduced in Section 1.3. The Monte-Carlo simulations under finite populations demonstrate that accounting for complex designs by survey-weighting of the estimators does barely harm efficiency when weighting is not necessary and improves the estimation in case of informativity.

Chapter 3 gives an overview about GVFs: The theoretical motivation of GVFs is discussed, their relation to survey design and an error decomposition of GVFs. As the theoretical motivation, which is model-based, diverges from the GVFs' application, i.e. design-based statistics, again the joint model-design probability measure introduced in Section 1.3 is required to put GVFs into a consistent framework. Comprehensive simulation studies round the analysis of GVFs off. As the true functional relation between an estimator's expectation and variance – when existing – is usually unknown to the analyst, the performance of various common GVFs is compared on different types of variables. Furthermore, GVFs are compared with the direct variance estimator. These problems relate to the model-based aspects of GVFs. However, there are also design-based aspects to be studied: Usually, the impact of survey design on the predictive quality of GVFs are not discussed in the literature and often, simulations are only run on one sample realization. The simulation studies here thus constitute a novelty as the repeated drawing from a finite population under different survey designs allows to separate the effect of

a single survey realization from the general behaviour of GVF. Classical indicators for the adequacy of GVFs are studied as well.

Finally SOMs are discussed in Chapter 4, and how they can serve in multivariate density estimation. This chapter goes beyond most of the typical SOM literature and studies variations of SOMs as an optimization problem. A new link to Markov Random Fields (MRFs) is uncovered that will help to relate the original optimization problem to a computationally easier one. Inherent to the optimization problem, a density estimator can be identified is related to classical density estimators, namely GMMs: This relation helps to get an idea about the convergence properties of the machine learning based estimator. This last chapter may be seen as building a bridge between traditional problems of statistical inference – including point estimation in case of Chapter 2 and variance estimation in Chapter 3 – and a new frontier in statistics with topics in machine learning and big data.

The rest of this chapter is dedicated to a brief introduction into survey sampling – a branch of applied statistics that is omnipresent in this work. Design-based statistics and model-based statistics and point estimation therein are discussed and finally synthesized. In that introduction, the mathematical notation is established that is used throughout this thesis except noted otherwise.

1.3 INTRODUCTION TO POINT ESTIMATION THEORY AND SURVEY SAMPLING

First, we introduce the mathematical notation linked to the main features of a survey that are named in Särndal et al. [1992, Chapter 1.2 & 1.3]. A finite index set $U = \{1, \dots, N\}$ is called the finite population (of size N). In practice, the finite population needs to be defined precisely in order to identify all unit $i = 1, \dots, N$ that belong to U. For example, it must be clarified whether those persons with a permanent (legal) residence (U_1) or with the nation's citizenship (U_2) constitute a country's population as $U_1 \neq U_2$. To each unit i in a well defined population U, there is attributed a vector of characteristics $(\mathbf{y}_i, \mathbf{z}_i) \in \mathcal{Y} \times \mathcal{Z} \subseteq \mathbb{R}^{p+p'}$, where the array of characteristics $\mathbf{y} = (\mathbf{y}_i : i = 1, \dots, N)$ is called the variable of interest and the array $\mathbf{z} = (\mathbf{z}_i : i = 1, \dots, N)$ is an array of auxiliary information. We set $\times_{i=1}^{N}(\mathcal{Y} \times \mathcal{Z}) =: \mathcal{Y}_N \times \mathcal{Z}_N = \Omega$. In the case of regression modelling, we replace $\mathbf{y}_i \in \mathcal{Y} \subseteq \mathbb{R}^p$ by $y_i \in \mathcal{Y} \subset \mathbb{R}$ and use partially $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^p$ instead of $\mathbf{z}_i \in \mathcal{Z} \subseteq \mathbb{R}^{p'}$ like it is convention.

A statistic of interest is the value of function $g_U(\mathbf{y})$ with

$$g_U : \times_{k=1}^{N} \mathcal{Y} \to \mathbb{R}^q, \quad \mathbf{y} \mapsto g_U(\mathbf{y}) \quad . \tag{1.1}$$

Delimiting the term 'statistic' in that way differs slightly from Särndal et al. [1992, p. 33]. However, later an interaction between survey sam-

pling and classical statistical models $\mathcal{M}$ are introduced and assuming that $(\mathbf{y}, \mathbf{z})$ are outcomes of a random process, the original understanding of 'statistic' is restablished. For the joint model-design approach to be introduced later, it is required that g_U be measurable in $\times_{k=1}^{N} \mathcal{Y}$ with respect to a probability measure $P_{\mathcal{M}_\theta}$ to be defined later. In the framework of Rubin-Bleuer and Kratina [2005] this means Borel measurability but we seek to generalize the definitions given there. An example statistic of interest is the population total τ_y or population mean μ_y

$$\boldsymbol{\tau}_y := \sum_{i \in U} \mathbf{y}_i, \qquad \boldsymbol{\mu}_y := \frac{1}{N}\tau_y \quad . \tag{1.2}$$

A survey realization is a subset $s \subseteq U$, where s is a realization of a random variable $S \sim P_D$. By this, we exclude all non-probability surveys. We set $S = \mathrm{id}_{\mathcal{S}}$ and assume therefore S to be P_D-measurable random variable mapping from the subset $\mathcal{S}$ of 2^U with $\sum_{s \in \mathcal{S}} P_D(s) = 1$ and $P_D(s) > 0$ for all $s \in \mathcal{S}$. P_D is called in the following a survey design, and might furthermore depend on the population variables $\mathbf{z}$, $P_D = P_D(\cdot; (\mathbf{y}, \mathbf{z}))$, that are also called design variables in this context. Note that the inclusion of the variable of interest $\mathbf{y}$ in P_D is not usual but will help us define sample informativity later on.

The probability space of S is thus $(\mathcal{S}, 2^{\mathcal{S}}, P_D)$. S can be equivalently represented by the random vector $\mathbf{1}_S := (\mathbb{1}_S(i) : i \in U)$, where $\mathbb{1}_A$ is the indicator variable equal to 1 if the argument is included in A and 0 otherwise. As there are only finitely many state spaces of $\mathbf{1}_S$ (due to the finiteness of $\mathcal{S}$, $|\mathcal{S}| \leqslant 2^N$), all moments of the random variable $\mathbf{1}_S$ exist and P_D is completely described by the moments of $\mathbf{1}_S$. We can therefore state $S \simeq \mathbf{1}_S$ and $(\mathcal{S}, 2^{\mathcal{S}}, P_D) \simeq (\{0,1\}^N, 2^{\{0,1\}^N}, P_D)$.

The expectation of the i^{th} element in $\mathbf{1}_S$ is called the first order inclusion probability for unit $i \in U$ is

$$\pi_i := P_D(i \in S; (\mathbf{y}, \mathbf{z})) = \sum_{s \in \mathcal{S}} P_D(S = s; (\mathbf{y}, \mathbf{z})) \cdot \mathbb{1}_s(i) = E_D[\mathbb{1}_S(i)] \tag{1.3}$$

and the i-j-th second moment is the second order inclusion probability

$$\begin{aligned} \pi_{ij} &:= P_D(i \in S \wedge j \in S; \mathbf{z}) = \sum_{s \in \mathcal{S}} P_D(S = s; (\mathbf{y}, \mathbf{z})) \cdot \mathbb{1}_s(i) \cdot \mathbb{1}_s(j) \\ &= E_D[\mathbb{1}_S(i) \cdot \mathbb{1}_S(j)] \quad . \end{aligned} \tag{1.4}$$

Depending on the design, the second and higher order statistics of $\mathbf{1}_S$ can be difficult to calculate and are often not released. This is in so far critical as they are required for variance estimation of point estimators

under P_D (cf. Section 3.1). If the expected sample size $E_D(|S|)$ is set to n, we thus get the property [Robinson and Särndal, 1983]

$$\begin{aligned} n = E_D[|S|] &= \sum_{s \in \mathcal{S}} P_D\left(S = s; (\mathbf{y}, \mathbf{z})\right) \cdot \sum_{i \in U} \mathbb{1}_s(i) \\ &= \sum_{i \in U} \sum_{s \in \mathcal{S}} P_D\left(S = s; (\mathbf{y}, \mathbf{z})\right) \cdot \mathbb{1}_s(i) = \sum_{i \in U} \pi_i \end{aligned} \tag{1.5}$$

and for the design weights $w_i := \frac{1}{\pi_i}$ it holds that

$$\begin{aligned} E_D\left[\sum_{i \in U} w_i \cdot \mathbb{1}_S(i)\right] &= \sum_{s \in \mathcal{S}} P_D\left(S = s; (\mathbf{y}, \mathbf{z})\right) \sum_{i \in U} w_i \cdot \mathbb{1}_s(i) \\ &= \sum_{i \in U} w_i \sum_{s \in \mathcal{S}} \mathbb{1}_s(i) \cdot P_D\left(S = s; (\mathbf{y}, \mathbf{z})\right) \\ &= \sum_{i \in U} w_i P_D(i \in S) = N \quad , \end{aligned} \tag{1.6}$$

where subscripts on E refer to the underlying probability measure. In general, sampling can take place with and without replacement where the second case is common in practice whereas 'with replacement' is comfortable for theoretical analysis. However, the definition of our probability space is not applicable to sampling with replacement because in that case, mathematical set theory cannot be as the number of draws of a unit matters, i.e. with replacement, the sample $\{i, i, j\}$, $i, j \in U$ is possible and not equal to $\{i, j\}$. To keep the mathematical notation simple, we assume throughout this thesis sampling without replacement.

An estimator in the context of survey sampling is thus a function g_S that aims at statements about $g_U(\mathbf{y}, \mathbf{z})$ [Rubin-Bleuer and Kratina, 2005]. Formally, we have

$$g_S : \mathcal{S} \times \Omega \to \mathbb{R}^q$$

or equivalently

$$g_S : \{0, 1\}^N \times \Omega \to \mathbb{R}^q \quad . \tag{1.7}$$

Note the different input spaces of g_S and g_U: Not only does the estimator g_S take as input a realization of the random variable S, but also auxiliary information. This shall help in the estimation process but is not necessary for the finite population statistic. If the sample size fixed by P_D, i.e. $|S| \equiv n$, we might also take as input space $\times_{k=1}^{n}(\mathcal{Y} \times \mathcal{Z})$. We abbreviate a sub-array with elements $s \subset U$ of $\mathbf{y}$ and $\mathbf{z}$ by $\mathbf{y}_s := (\mathbf{y}_i : i \in s)$ and $\mathbf{z}_s := (\mathbf{z}_i : i \in s)$ respectively. The input space of the sub-arrays $(Y_i, Z_i : i \in s)$ is denoted by $\Omega_s = \mathcal{Y}_s \times \mathcal{Z}_s$.

Such an estimator has to be analysed with respect to the probability law P_D. However, like the population statistic (1.1), it is important that

$g_S(s,\cdot)$ be measurable for any $s \in \mathcal{S}$ in its second argument with respect to a probability measure $P_{\mathcal{M}_\theta}$ to be defined later in order to adjust it later to a joint probability measure. The analytical computation of estimator g_S's k-th moment under P_D is

$$E_D\left[g_S(S,\mathbf{y},\mathbf{z})^k\right] = \sum_{s\in\mathcal{S}} P_D(S=s;(\mathbf{y},\mathbf{z}))\cdot g_S(s,\mathbf{y},\mathbf{z})^k \quad . \tag{1.8}$$

Going back to the population total $\boldsymbol{\tau}_y$ as statistic of interest, it is thanks to the representation of the design through $\mathbf{1}_S$ easy to define the HT [Horvitz and Thompson, 1952] as

$$\hat{\boldsymbol{\tau}}_y(s,\mathbf{y}) \triangleq \sum_{i\in U} \frac{1}{\pi_i}\cdot \mathbb{1}_s(i)\cdot \mathbf{y}_i \quad , \tag{1.9}$$

whose variance and the estimation thereof will play a role in Chapter 3. Another commonly used estimator is the GREG [Deville and Särndal, 1992]. The GREG requires knowledge about the population totals of the auxiliary variables $\mathbf{z}$. Then, we can define

$$\hat{\boldsymbol{\tau}}_y^{\mathrm{GREG}}(s,\mathbf{y},\mathbf{z}) \triangleq \hat{\tau}_y + B_s^T(\hat{\boldsymbol{\tau}}_z - \boldsymbol{\tau}_z) \tag{1.10}$$

where

$$B_s \triangleq \left(\sum_{i\in U}\frac{\mathbb{1}_s(i)}{\pi_i}\mathbf{z}_i\mathbf{z}_i^T\right)^{-1}\sum_{i\in U}\frac{\mathbb{1}_s(i)}{\pi_i}\mathbf{z}_i\mathbf{y}_i^T \quad .$$

The product $B_s^T\hat{\boldsymbol{\tau}}_z$ – and B_s is a nonlinear statistic of S itself – complicates the analysis of moments of $\hat{\boldsymbol{\tau}}_y^{\mathrm{GREG}}$. The unbiasedness (cf. Definition 1) is therefore no longer assured.

Definition 1 (Design unbiasedness)**.** *An estimator* $g_S : \mathcal{S}\times\Omega \to \mathbb{R}^q$ *is* design unbiased *if*

$$E_D[g_S(S,\mathbf{y},\mathbf{z})] = \sum_{s\in\mathcal{S}} g_S(s,\mathbf{y},\mathbf{z})\cdot P_D(S=s;(\mathbf{y},\mathbf{z})) = g_U(\mathbf{y}) \tag{1.11}$$

(cf. Lehmann [1983, p. 5] adapted to survey sampling).

Nonetheless, comfortable asymptotic results for the GREG can be derived, leading to another problem of estimators on $\mathcal{S}$: Common asymptotic properties are not possible to state due to the finiteness of U and therefore the upper bound of $|S|$.

The finiteness of U requires special attention for statements on asymptotic statistical properties such as consistency or central limit theorems. For better distinction in the next paragraph, denote U of size N as U_N. The upper bound of S is possible to grow for nested populations

$\ldots \subset U_N \subset U_{N+1} \subset \ldots$ when the sampling rules in P_D (switch to notation P_{D_N}) are adjusted to the population size N. Boistard et al. [2015] give an overview about the necessary theoretical set-up to establish central limit theorems for the HT [Horvitz and Thompson, 1952]. This set-up is similar to that in Rubin-Bleuer and Kratina [2005], seeks to unify infinite and finite population statistics and shall be briefly introduced and adapted to our requirements in the following, because it is related to the statistical models used in Chapters 2 and 3.

The survey design P_{D_N} on a finite population U_N may depend – as already mentioned – on design variables $(\mathbf{y}, \mathbf{z}) \in \times_{i=1}^{N}(\mathcal{Y} \times \mathcal{Z}) =: \Omega_N \subseteq \mathbb{R}^{N\times(p+p')}$. Boistard et al. [2015] assume the design variables to be non-negative, but this is due to convention: None of the proofs in Boistard et al. [2015] requires the design variables to be non-negative and we will drop the non-negativity assumption. In the following, it is only necessary that P_{D_N} be measurable with respect to a sigma-field $\mathcal{A}_N$ on Ω_N. When this is the case, P_{D_N} defines a Markov kernel.

In the following, the key additional assumption is that the variables $(\mathbf{y}_i, \mathbf{z}_i) \in \mathcal{Y} \times \mathcal{Z}$, $i \in U_N$, are realizations of a random process $(Y_i, Z_i) \sim_{iid} P_{\mathcal{M}_\theta}$, $\otimes_{i=1}^{N} P_{\mathcal{M}_\theta} = P_{\mathcal{M}_\theta}^N$ on the space $(\Omega_N, \mathcal{A}_N)$. To simplify the analysis later on, we assume that $\mathcal{A}_N$ is a product sigma-field of the fields $\mathcal{A}_y$ on $\mathcal{Y}_N$ and $\mathcal{A}_z$ on $\mathcal{Z}_N$. If required, let analogously to the outcomes $\mathbf{y}$ and $\mathbf{z}$ denote $Z = (Z_i : i \in U)$ and $Y = (Y_i : i \in U)$ to be short cuts for projections from Ω_N to $\mathcal{Z}_N$ and $\mathcal{Y}_N$ respectively, $Z = \text{proj}_z \circ (Y, Z)$ and $Y = \text{proj}_y \circ (Y, Z)$. Sub-arrays Y_s and Z_s are the projections $\text{proj}_y^s \circ (Y, Z)$ and $\text{proj}_z^s \circ (Y, Z)$ mapping from Ω_N to $\mathcal{Y}_s = \times_{i\in s}\mathcal{Y}$ and $\mathcal{Z}_s = \times_{i\in s}\mathcal{Z}$ respectively. The projection proj^s maps from Ω_N to $\Omega_s = \times_{i\in s}(\mathcal{Y} \times \mathcal{Z}) = \mathcal{Y}_s \times \mathcal{Z}_s$. As we assume product sigma-fields, let $\mathcal{A}_s = \mathcal{A}_{ys} \otimes \mathcal{A}_{zs}$ be the field on Ω_s, i.e. the field on $\text{proj}^s(\Omega_N)$.

As Y is the variable of interest, it is sometimes useful to define the conditional probability of Y given Z rather than the joint probability: Let $A = A_y \times A_z$ be an arbitrary element of $\mathcal{A}_N$ such that $\text{proj}_z^{-1}(A_z) = \mathcal{Y}_N \times A_z \in \mathcal{A}$. If $P_{\mathcal{M}_\theta}^N\left(\text{proj}_z^{-1}(A_z)\right) > 0$, the conditional probability $P_{\mathcal{M}_\theta}^N(Y \in A_y | Z \in A_z)$ is defined as

$$P_{\mathcal{M}_\theta}^N(Y \in A_y | Z \in A_z) = \frac{P_{\mathcal{M}_\theta}^N(A_y \times A_z)}{P_{\mathcal{M}_\theta}^N\left(\text{proj}_z^{-1}(A_z)\right)} \quad . \tag{1.12}$$

If no $A \in \mathcal{A}_N$ is specified, we refer to the conditional probability as $P_{\mathcal{M}_\theta}^N\left(\text{proj}_y^{-1} \cdot | \text{proj}_z\right)$. Note that $P_{\mathcal{M}_\theta}^N\left(\text{proj}_z^{-1}(A_z)\right)$ is the marginal probability that the event A_z occurs.

The law $P_{\mathcal{M}_\theta}^N$ enables us to define a joint model-design probability measure $P_{\mathcal{M}_\theta, D_N}$ on the product algebra $\mathcal{A}_N$ on Ω_N for design variables and

variables of interest, $((Y_i, Z_i) : i \in U_N) =: (Y, Z)$ and S, and the sample variable S defined on $\left(2^U, 2^{(\in^U)}, P_D\right)$ as

$$P_{\mathcal{M}_\theta, D_N}(s \times A) := \int_A P_{D_N}(S = s; (Y, Z)(\omega))\, P^N_{\mathcal{M}_\theta}(d\,\omega) \tag{1.13}$$

for any $A \in \mathcal{A}_N$ [Rubin-Bleuer and Kratina, 2005, Boistard et al., 2015]. We need 2^U rather than $\mathcal{S}$ for the input space as the survey design is conditioned on Z and it may hold that $P_D(s; (Y, Z)(\omega_1)) = 0$ but $P_D(s; (Y, Z)(\omega_2)) > 0$. This is unproblematic because for given $(Y, Z) = (\mathbf{y}, \mathbf{z})$ any samples in $2^U \setminus \mathcal{S}$ can be assigned probabilities equal to zero.

Note that the definition (1.13) differs slightly from the original one in Rubin-Bleuer and Kratina [2005], due to the fact that we allow for informative designs as mentioned previously. The joint probability spaces $\left(2^U \times \Omega_N, 2^{(\in^U)} \otimes \mathcal{A}_N, P_{\mathcal{M}_\theta, D_N}\right)$ require the introduction of new projections proj_s and proj_{sy} from $2^U \times \Omega_N$ to 2^U and $2^U \times \mathcal{Y}_N$ respectively. Note the difference of proj^s mapping from Ω_N to Ω_s and proj_s.

If no asymptotics are considered, i.e. the limit behavior is not under study, we refer like previously to U_N as U, (1.13) is simply $P_{\mathcal{M}_\theta, D}$, P_{D_N} becomes P_D and $(\Omega_N, \mathcal{A}_N)$ is abbreviated to $(\Omega, \mathcal{A})$.

$P_{\mathcal{M}_\theta, D_N}$ allows to study statistical models that rely on subsets of sample realizations that were generated by a scheme P_D. However, the probability measure (1.13), also allows to assume $N \to \infty$ and for adequate P_{D_N} asymptotic results can be derived. If we assume that $U_N \subset U_{N+1}$ for all $N \in \mathbb{N}$, we can interpret

$$\{(Y_i, Z_i)\}_{i \in \mathbb{N}} \quad \text{and} \quad \left\{P_{\mathcal{M}_\theta, D_N}\right\}_{N \in \mathbb{N}}$$

as stochastic processes. For the analysis in the limits, often $\{\{(Y_i, Z_i) : i \in U_N\}\}_{N \in \mathbb{N}}$ is considered.

P_{D_N} together with the data growth rule $P^N_{\mathcal{M}_\theta}$ allows us thus to define design consistency of estimators $g_{S,N} : \mathcal{S}_N \times \Omega_N \to \mathbb{R}^q$ as N and n go to infinity.

Definition 2 (Design consistency). *A sequence of estimators* $\{g_{S,N}\}_{N \in \mathbb{N}}$ *where* $g_{S,N} : \mathcal{S}_N \times \Omega_N \to \mathbb{R}^q$ *for a statistic* g_{U_N} *is* design consistent *if for every* $\omega \in \Omega_N$ *that is not element of the joint null sets of the elements in* $\left\{P^N_{\mathcal{M}_\theta}\right\}_{N \in \mathbb{N}}$ *and arbitrary* $\varepsilon > 0$

$$\lim_{N \to \infty} P_{D_N}(|g_{S,N}(S, Y(\omega), Z(\omega)) - g_{U_N}(Y(\omega))| > \varepsilon; Z(\omega)) = 0 \quad . \tag{1.14}$$

[Pfeffermann, 1993, Definition 2].

Possibly Definition 2 is too restrictive as we require this property for all elements that are not element of joint null sets, i.e. (Y, Z) be adapted to

a complete filtration of the probability space $\left(\Omega_{\mathbb{N}}, \mathcal{A}_{\mathbb{N}}, P^{\mathbb{N}}_{\mathcal{M}_\theta}\right)$. However, for less restrictive definitions – that are not required in the following – a more detailed study on the limit behavior on null sets of $\mathcal{A}_N$ is required.

Remark 1 (Design consistency of the HT)**.** *The HT converges in* L^2 *to the finite population total and thus converges in probability under the conditions on the design* P_D *and* Y *given in Chauvet [2014]. Convergence in probability means that the HT is design consistent.*

Remark 2 (Design consistency of the GREG)**.** *If the HT is* $\sqrt{n}$*-consistent and the calibration weights converge uniformly (cf. conditions C.2 and C.3 in Kim and Park [2010]), the GREG is design consistent [Kim and Park, 2010, Theorem 1 and Equation 15].*

The definition of $P_{\mathcal{M}_\theta, D}$ also requires an adaption of the Definition of design unbiasedness to the model-design context.

Definition 3 (Design consistency (revised))**.** *An estimator* $g_S : \mathcal{S} \times \Omega \to \mathbb{R}^q$ *for a statistic* g_U *is* design consistent *if for almost every* $\omega \in \Omega$ *under* $S \sim P_D$

$$E_D\left[g_S\left(S, (Y, Z)(\omega)\right)\right] = g_U\left(Y(\omega)\right) \quad . \tag{1.15}$$

Often, findings on the asymptotic behavior of finite population estimators have requirements on the sampling design P_D [Chauvet, 2014, for example]. In that context the design entropy is of special concern.

Definition 4 (Entropy)**.** *The* entropy *of a survey design* P_D *is defined as*

$$H\left(P_D\right) \triangleq -\sum_{s \in \mathcal{S}} P_D(s) \log P_D(s) \tag{1.16}$$

with the convention that $0 \log 0 = 0$ *[Berger, 1998].*

The unique design $\tilde{P}_{D_N}$ with maximum entropy amongst those with fixed sample size $|S| \equiv n$ is rejective sampling. Rejective sampling draws units from U_N with a pre-specified probability and with replacement. If a unit is drawn twice, the complete sample is rejected. The procedure is repeated until n different units are drawn. For central limit theorems on the HT, Berger [1998] shows that such a characteristic of the design is the 'high entropy' property:

Definition 5 (High Entropy Design)**.** *A survey design* P_{D_N} *has high entropy when the Kullback-Leibler divergence between* P_{D_N} *and* $\tilde{P}_{D_N}$

$$D\left(P_{D_N}, \tilde{P}_{D_N}\right) \triangleq \sum_{s \in \mathcal{S}_N} P_{D_N}(s) \log\left(\frac{P_{D_N}(s)}{\tilde{P}_{D_N}(s)}\right)$$

goes to zero.

If the entropy is high, Berger [1998] shows the asymptotic normality of the HT. Amongst the other well studied designs are stratified two-stage sampling designs [Krewski and Rao, 1981] and some unequal probability designs like the rejective sampling $\tilde{P}_{D_N}$ [Berger, 1998].

Another remark is on sample informativity: Both Rubin-Bleuer and Kratina [2005] and Boistard et al. [2015] require for their analysis the sampling design to be independent of Y, i.e. $P_D(\cdot; Y, Z) = P_D(\cdot; Z)$ or equivalently $Y \perp S$. In the next chapter, we will call this *non-informativity* of the design. Though the independence is a necessary condition for their analysis, it is nonetheless thinkable that the condition does not hold. In the following, we introduce sample informativity concepts in the framework of the joint probability law.

Assume now that the estimator $g_{S,N}$ aims to estimate the value $g_{\mathcal{M}_\theta}(\theta)$ of a statistic $g_{\mathcal{M}_\theta} : \Theta \to \mathbb{R}^q$ where $((Y_i, Z_i) : i \in U_N) \sim P^N_{\mathcal{M}_\theta}$ for all $N \in \mathbb{N}$ and $\theta \in \Theta$ is unknown. In that case, we can introduce model-design consistency.

Definition 6 (Model-design consistency)**.** *Assume a sequence of estimators* $\{g_{S,N}\}_{N\in\mathbb{N}}$ *with* $g_{S,N} : \mathcal{S}_N \times \Omega_N \to \tilde{\Theta} \supset \Theta$ *for a statistic* $g_{\mathcal{M}_\theta}$*. The sequence is* model-design consistent *if*

$$\lim_{N\to\infty} P_{\mathcal{M}_\theta, D_N}\left(|g_{S,N}(S, Y, Z) - g_{\mathcal{M}_\theta}(\theta)| > \varepsilon\right) = 0 \tag{1.17}$$

for every $\varepsilon > 0$ *and* $\theta \in \Theta$*.*

If we set $P_D = \mathbb{1}_U$, that is $S \equiv U$ almost surely, we call such an estimator $g_{U,N} := g_{U_N}$ simply *model consistent*. The law $P_{\mathcal{M}_\theta, D_N} = P_{\mathcal{M}_\theta, D}$ allows us also to define model-design unbiasedness analogously to Definition 3.

Definition 7 (Model-design unbiasedness)**.** *An estimator* $g_S : \mathcal{S} \times \Omega \to \tilde{\Theta}$ *is* model-design unbiased *for a statistic* $g_{\mathcal{M}_\theta}$ *if for every* $\theta \in \Theta$ *it holds that*

$$E_{\mathcal{M}_\theta, D}[g_S(S, Y, Z)] = \int_{\mathcal{S}\times\Omega} g_S(s, Y, Z)\, \mathrm{d}\, P_{\mathcal{M}_\theta, D} = g_{\mathcal{M}_\theta}(\theta) \quad . \tag{1.18}$$

Again assuming a sequence of laws $\{P_{\mathcal{M}_\theta, D_N}\}_{N\in\mathbb{N}}$, unbiasedness can also hold asymptotically. Again, if $P_D = \mathbb{1}_U$ and $S \equiv U$ almost surely, that is, a 'degenerate' survey design, the estimator $g_S \triangleq g_U$ is simply called *model unbiased* if Definition 7 holds.

In the context of mixed models or GVFs, we often find the term predictor. This notion is often used without a precise definition. A predictor is usually a special estimator for a finite population statistic g_U under a model $\mathcal{M}_\theta$. In order to provide an exact reference for the following chapters, we give here Definition 8.

Definition 8 (Predictor). *A function* $g_S : \mathcal{S} \times \Omega \to \mathbb{R}^q$ *is a* predictor *for the non-sampled units* $U \setminus S$ *if the estimand is* $Y_{U\setminus S}$ *or a function* $g_U\left((Y_{U\setminus S}, \mathbf{y}_S)\right)$.

In contrast to estimation, the statistical proximity of a predictor g_S with respect to $g_U\left((Y_{U\setminus S}, \mathbf{y}_S)\right)$ must be a proximity criterion of $P^N_{\mathcal{M}_\theta,D}$-measurable, non-constant functions. For an estimator g_S in model-based statistics, the function that is proxied in a statistical sense is a constant, namely $g_{\mathcal{M}_\theta}(\theta)$. The attempt to make statements on g_U as a random variable under a model $\mathcal{M}$ rather than on a constant statistic is the biggest difference between a predictor and an estimator in model-based statistics. This peculiarity of predictors needs an adjustment of model-design unbiasedness and consistency to model-based predictors: Definition 6 is applicable to predictors if the statistic $g_{\mathcal{M}_\theta}(\theta)$ is replaced by $E_{\mathcal{M}_\theta,D}[g_U(Y)] = E_{\mathcal{M}_\theta}[g_U(Y)]$ [Isaki and Fuller, 1982]. Similarly, replace in Definition 7 $g_{\mathcal{M}_\theta}(\theta)$ with $E[g(Y)]$. Though the estimand is $g_U\left(Y_{U\setminus S}, \mathbf{y}_S\right)$ a replacement in the definitions with $E_{\mathcal{M}_\theta,D}\left[g_U\left(Y_{U\setminus S}, \mathbf{y}_S\right)\right]$ makes no sense because S is variable and therefore the indices in the sub-arrays $Y_{U\setminus S}$ and $\mathbf{y}_S$ change; requiring once to condition on $Y_i = \mathbf{y}_i$ for $S = s$ and once to integrate Y_i for $S = s'$. This, however, means that the consistency and unbiasedness of a predictor heavily depends on the correct choice of the model family $\mathcal{M}$ and a good parameter estimator for θ.

2

MULTILEVEL REGRESSION ANALYSIS IN BUSINESS SURVEYS

2.1 INTRODUCTION TO DESIGN-SENSITIVE REGRESSION ANALYSIS

This chapter builds upon the discussion papers Burgard and Dörr [2019] and Dörr and Burgard [2019]. Though the topic of design-sensitive regression modelling is not new in the literature [Fuller, 1975, Isaki and Fuller, 1982, Pfeffermann, 1993, Kott, 2018, e.g.], and even not to the (generalized) LMM literature [Pfeffermann et al., 1998, Rabe-Hesketh and Skrondal, 2006, e.g.], the introduced algorithm in Burgard and Dörr [2019] and extended in Dörr and Burgard [2019] has some peculiarities. But before discussing the special case of random effects models and the stochastic optimization employed in the later discussed algorithm, we first review the role of survey design in regression analysis in general.

As already mentioned in Section 1.3, empirical researchers often use an estimator g_S in regression analysis that is an element of a sequence of estimators $\{g_{U_N}\}_{N\in\mathbb{N}}$ for a statistic $g_{\mathcal{M}_\theta}(\theta)$ with model parameter $\theta \in \Theta$. It is assumed that the population only consists of the sample S, i.e. $P_D = \mathbb{1}_U$ and $S \equiv U$ almost surely, which is rarely the case. The problem then lies often in the assumption that good statistical properties of the sequence $\{g_{U_N}\}_{N\in\mathbb{N}}$ apply on g_S regardless the design P_D. This assumption, however, is only justified when the sample design is *non-informative*.

Definition 9 (Sample Informativity)**.** *A survey design* $P_D(\cdot;(Y,Z))$ *is* non-informative *if it holds for every* $s \in 2^U$ *and* $A \in \mathcal{A}$ *and almost every* $\mathbf{z} \in \text{proj}_z(A)$

$$\begin{aligned} &P_{\mathcal{M}_\theta,D}\left(\left\{U \times \left(\text{proj}_y^s(A) \times \mathcal{Y}_{U\setminus s}\right)\right\} | \text{proj}_z\right) \\ &\quad = P_{\mathcal{M}_\theta}^N\left(\text{proj}_y^s(A) \times \mathcal{Y}_{U\setminus s} | \text{proj}_z\right) \quad , \end{aligned} \tag{2.1}$$

Otherwise, the survey design is informative.

Pfeffermann [1993, Definition 3] gives almost the same definition using the density of $P_{\mathcal{M}_\theta,D}$. However, he has forgotten to account for the variability of S in his condition. That means, that the condition in Pfeffermann [1993, Definition 3] needs to hold for every sample $s \in 2^U$. Furthermore, $Y \perp S|Z$ is equivalent to the design being non-informative and both is shown in Appendix A.

Remark 3. *Definition 9 for a non-informative design is equivalent to the statement that* $Y \perp S$ *given the information* Z*.*

As $U \in 2^U$, it follows from Definition 9 that we can simplify the definition of design non-informativity.

Remark 4. *A survey design $P_D(\cdot;(Y,Z))$ is non-informative if the marginal distribution of Y given Z under $P^N_{\mathcal{M}_\theta}$ is the same as under $P_{\mathcal{M}_\theta,D}$, i.e. for any $A \in \mathcal{A}$ and almost every $\mathbf{z}$,*

$$P^N_{\mathcal{M}_\theta}\left(\text{proj}_y^{-1} \in A | \text{proj}_z\right) = P_{\mathcal{M}_\theta,D}\left(\text{proj}_{sy}^{-1} \in \{U \times A\} \cdot | \text{proj}_z\right)$$

When the inference based on $P^N_{\mathcal{M}_\theta}$ and $P_{\mathcal{M}_\theta,D}$ given the auxiliary information is identical, it is possible to choose for a drawn sample $S = s$ the appropriate element n from a sequence of estimators $\{g_{U_N}\}_{N\in\mathbb{N}}$ where $|U_n| = |s|$, set $g_S = g_{U_n}$ and to rely on the properties of the estimator on $P^N_{\mathcal{M}_\theta}\left(\text{proj}_y^{s\,-1} \in \cdot | \text{proj}_z\right)$. For example, known (asymptotic) behavior of g_{U_k} for the moments can be applied with Remark 5 noting that for non-informative P_D, $E_D\left[g_{U_n}(Y_s)\right] = g_{U_n}(Y_s)$.

Remark 5. *Assume that first and second $P^N_{\mathcal{M}_\theta}$-moments of $g_S(s,Y,Z)$ exist for every $s \in 2^U$. In that case, in Appendix A it is shown that*

$$E_{\mathcal{M}_\theta,D}\left[g_S(S,Y,Z)\right] = E_{\mathcal{M}_\theta}\left[E_D\left[g_S(S,Y,Z)\right]\right] \tag{2.2}$$

and

$$\begin{aligned}\text{Var}_{\mathcal{M}_\theta,D}\left[g_S(S,Y,Z)\right] = {} & \text{Var}_{\mathcal{M}_\theta}\left[E_D\left[g_S(S,Y,Z)\right]\right] \\ & + E_{\mathcal{M}_\theta}\left[\text{Var}_D\left[g_S(S,Y,Z)\right]\right] \quad .\end{aligned} \tag{2.3}$$

That means, a design unbiased estimator for the statistic $g_U(Y)$ which is in turn model-unbiased for the statistic $g_{\mathcal{M}_\theta}(\theta)$ is model-design unbiased. Furthermore, there is a decomposition for the triple $\left(g_S, g_U, g_{\mathcal{M}_\theta}(\theta)\right)$ into the effect of using g_U for the estimation of $g_{\mathcal{M}_\theta}(\theta)$ (inter-group variance – the variance of the expected finite population estimator) and the effect of only having a random subset available, i.e. using g_S instead of g_U (intra-group variance – the expected sample variance). However, note that the added stochasticity by the subsetting process does not bother the expectation if as estimator a design unbiased g_S is used rather than the complete realization (meaning g_U), but adds variability to the estimation procedure.

When the probability laws $P^N_{\mathcal{M}_\theta}$ and P_D are combined like in Equation (1.13), two questions arise consequently: First, what happens to an estimator $g_S \in \{g_{U_N}\}_{N\in\mathbb{N}}$ when the assumption of non-informativity is violated and second, is there an alternative estimator $\tilde{g}_S$ that accounts for informativity?

In the rest of the chapter, focus lies on regression analysis. Like indicated in the introduction, we thus shift the notation from (Y,Z) to

(Y, X) with X_i being a $\mathbb{R}^p$-valued random vector with realization $\mathbf{x} = (\mathbf{x}_i : i \in U)$ and Y_i taking values in $\mathbb{R}$ for each $i \in U$. Kott [2018] gives an overview about design-sensitive (generalized) linear models – and focuses on regression models where design-sensitivity is achieved by the use of survey weights

$$w_i \approx \frac{1}{E_D[\mathbb{1}_S(i)]} \quad . \tag{2.4}$$

Survey weights can differ from design weights $\frac{1}{E_D[\mathbb{1}_S(i)]}$ due to calibration and non-response adjustements. In general, survey and design weights should be close to each other, though. In his analysis, he even accounts for the case of calibrated weights in regression analysis and outlines the consistency of estimated regression parameters under weighting. From Section 2.2 on, we focus on weights that return the HT model-design unbiased and call these also survey weights.

The weighting becomes necessary in Kott's (2018) analysis if there is an $i \in U$ such that we have for the zero-centered and independently distributed error ε_i a conditional expectation $E_{\mathcal{M}_\theta, D}(\varepsilon_i | X_i, w_i) \neq 0$. An alternative derivation for weighted linear regression estimators is as follows with estimand $\boldsymbol{\beta}$ and $\mathrm{Var}_{\mathcal{M}_\theta}[\varepsilon_i] \equiv \sigma^2$:

The classical Ordinary Least Squares (OLS) regression estimator $\hat{\boldsymbol{\beta}}_S$, given the data $\mathbf{x} \in \mathcal{X}^N \subset \mathbb{R}^{N \times p}$ and the sample matrix $\mathrm{diag}(\mathbf{1}_S)$ with diagonal vector $\mathbf{1}_S = (\mathbb{1}_S(i) : i \in U)$ and zero else, is

$$\hat{\boldsymbol{\beta}}_S = \left(\mathbf{x}^T \mathrm{diag}(\mathbf{1}_S)\mathbf{x}\right)^{-1} \mathbf{x}^T \mathrm{diag}(\mathbf{1}_S) Y \qquad Y \triangleq (Y_i : i \in U)$$

Assume that the sample design is non-informative. In that case, we have (given $X = \mathbf{x}$ fixed)

$$\begin{aligned}
&E_{\mathcal{M}_\theta, D}\left[\hat{\boldsymbol{\beta}}_S | X = \mathbf{x}\right] = \int_{\mathcal{S} \times \mathbb{R}} \hat{\boldsymbol{\beta}}_S \, \mathrm{d}\, P_{\mathcal{M}_\theta, D}(\cdot | X = \mathbf{x}) \\
&= \int_{\mathcal{S}} \int_{\mathbb{R}} \hat{\boldsymbol{\beta}}_S \, \mathrm{d}\, P^N_{\mathcal{M}_\theta}(\cdot | X = \mathbf{x}) \, \mathrm{d}\, P_D \\
&= \int_{\mathcal{S}} E_{\mathcal{M}_\theta}\left[\left(\mathbf{x}^T \mathrm{diag}(\mathbf{1}_S)\mathbf{x}\right)^{-1} \mathbf{x}^T \mathrm{diag}(\mathbf{1}_S)(\mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \, | X = \mathbf{x}\right] \mathrm{d} P_D \\
&= \int_{\mathcal{S}} \boldsymbol{\beta} \, \mathrm{d}\, P_D = \boldsymbol{\beta} \quad .
\end{aligned} \tag{2.5}$$

Furthermore, $\hat{\beta}_S$ is model-design consistent for non-informative P_D because for any $\eta > 0$ and almost every $\mathbf{X}$, we have

$$\begin{aligned}
&P^N_{\mathcal{M}_\theta,D}\left(\|\hat{\beta}_S - \beta\| > \eta | X = \mathbf{x}\right) \\
&\quad \leqslant E_{\mathcal{M}_\theta,D}\left[\mathrm{tr}\left(\left(\mathbf{x}^T \mathrm{diag}(\mathbf{1}_S)\mathbf{x}\right)^{-1} \cdot \varepsilon\varepsilon^T\right) | X = \mathbf{x}\right] \frac{1}{\eta^2} \\
&\quad = E_{\mathcal{M}_\theta,D}\left[\mathrm{tr}\left(\left(\mathbf{x}^T \mathrm{diag}(\mathbf{1}_S)\mathbf{x}\right)^{-1}\right) | X = \mathbf{x}\right] \frac{\sigma^2}{\eta^2} \\
&\quad \leqslant \mathrm{tr}\left(E_{\mathcal{M}_\theta,D}\left[\mathbf{x}^T \mathrm{diag}(\mathbf{1}_S)\mathbf{x} | X = \mathbf{x}\right]\right)^{-1} \frac{\sigma^2}{\eta^2} \\
&\quad \leqslant \frac{1}{\sum_{i\in U} \pi_i \mathbf{x}_i^T \mathbf{x}_i} \frac{\sigma^2}{\eta^2} \qquad (2.6)
\end{aligned}$$

and for $n, N \to \infty$ where $X_i \sim_{iid} P^{\mathrm{proj}_x^{-1}}_{\mathcal{M}_\theta}$, this last term goes to zero.

If the sample is informative, on the other hand, the second equality in (2.5) and the third line in (2.6) is not valid any more because $P_D = P_D(\cdot; (Y, X)) \neq P_D(\cdot; X)$ and the OLS estimator is biased and inconsistent. To see the latter, note that the design is informative if there exists a nonlinear, non-negative function f such that $P_D = f(\varepsilon)$. If the design remains informative in the limit, this relation remains as $N \to \infty$. The $P_{\mathcal{M}_\theta,D}$-expectation of the minimization problem

$$\frac{1}{N} \cdot \sum_{i\in U} \mathbb{1}_S(i) \cdot (Y_i - \mathbf{x}_i^T \beta)^2 \qquad (2.7)$$

is then constantly different from the finite population based minimization problem and does not even converge to zero for $N \to \infty$. Then, the solution to the minimization problem (2.7) also deviates systematically from the finite population solution $\hat{\beta}_U$ even in the limit.

With survey weights $\mathbf{w} = (w_1, \ldots, w_N)$, $w_i = \frac{1}{E_D[\mathbb{1}_S(i)]}$ the weighted OLS estimator,

$$\begin{aligned}
\hat{\beta}^w_S &= \left(\mathbf{x}^T W_S \mathbf{x}\right)^{-1} \mathbf{x}^T W_S Y \\
W_S &= \mathrm{diag}(w_1 \cdot \mathbb{1}_S(1), \ldots, w_N \cdot \mathbb{1}_S(N)) \quad ,
\end{aligned}$$

on the other hand, is the outcome of the stochastic minimization problem given $X = \mathbf{x}$

$$\frac{1}{N} \cdot \sum_{i\in U} \mathbb{1}_S(i) \cdot w_i \cdot (Y_i - \mathbf{x}_i^T \beta)^2 \qquad (2.8)$$

and we get as the $P_{\mathcal{M}_\theta,D}$-expectation of the gradient of (2.8)

$$\frac{-2}{N} \cdot \sum_{i\in U} \mathbb{1}_S(i) \cdot w_i \cdot (Y_i - \mathbf{x}_i^T \beta)\mathbf{x}_i$$

the vector

$$\begin{aligned}
&\frac{-2}{N}\sum_{s\in 2^U}\int_{\mathcal{Y}_N} P_D(s;(Y,X))\sum_{i\in U}\mathbb{1}_s(i)\cdot w_i\cdot(Y_i-\mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i\,\mathrm{d}\,P^N_{\mathcal{M}_\theta}(\cdot|\mathbf{x})\\
&=\frac{-2}{N}\int_{\mathcal{Y}_N}\left(\sum_{s\in 2^U}P_D(s;(Y,X))\sum_{i\in U}\mathbb{1}_s(i)\cdot w_i\cdot(Y_i-\mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i\right)\mathrm{d}\,P^N_{\mathcal{M}_\theta}(\cdot|\mathbf{x})\\
&=\frac{-2}{N}\int_{\mathcal{Y}_N}\sum_{i\in U}(Y_i-\mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i\,\mathrm{d}\,P^N_{\mathcal{M}_\theta}(\cdot|X=\mathbf{x})\\
&\overset{ind}{=}\frac{-2}{N}\sum_{i\in U}E_{\mathcal{M}_\theta}[\varepsilon_i|X=\mathbf{x}]\,\mathbf{x}_i=0\quad,
\end{aligned}$$

which is equal to the first order condition for the finite population OLS estimator $\hat{\boldsymbol{\beta}}_U$, namely $\frac{-2}{N}\sum_{i\in U}Y_i-\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_U=0$ which in turn yields a model-unbiased estimator for $\boldsymbol{\beta}$.

With Remark 5, we find thus that the survey-weighted gradient is model-design unbiased for the finite population first order condition. Furthermore, we get as model-design covariance matrix of the gradient the diagonal matrix $\frac{4}{N^2}I_N\cdot\sigma^2$, which converges to the zero matrix. Thus, the survey-weighted gradient is model-design consistent (apply Markov's inequality to show this). Confer Lumley [2010, Chapter 5] for a more detailed derivation of $\hat{\boldsymbol{\beta}}_S^w$.

The unbiasedness and consistency of $\hat{\boldsymbol{\beta}}_S^w$ without the non-informativity of the design. We can thus conclude that even in the simplest case of regression analysis, informative designs can be problematic when estimates for superpopulation parameters are desired. Survey-weighting of parameter estimators, though, gives a certain advantage with respect to desirable statistical properties in contrast to unweighted estimators because the joint model-design distribution is explicitly taken into account. In the next section, survey-weighting is discussed for a more complex regression problem. In the statistical problem formulation presented below, OLS is no more applicable due to the introduction of additional random effects into the model. Instead, (Pseudo-)ML estimators are used. The procedure, however, is similar: One seeks to find a design-unbiased estimator for the objective function that theoretically would be optimized on the complete finite population if the latter had been observed. A discussion for ML estimation in a finite population context is given in Royall [1976].

2.2 SURVEY-WEIGHTED MIXED MODELS

2.2.1 *Model Formulation at the Population Level*

Assume that to each unit $i\in U$, a vector of characteristics $(y_i,\mathbf{x}_i^T,\mathbf{z}_i)^T$ is attributed and $y_i\in\mathcal{Y}\subset\mathbb{R}$, $\mathbf{x}_i\in\mathcal{X}\subset\mathbb{R}^p$ and $\mathbf{z}_i\in\mathcal{Z}\subset\mathbb{R}^q$. Both arrays,

$\mathbf{x}$ and $\mathbf{z}$, are explanatories, but will play different roles in the following. Like in the previous section, the random process of the explanatories $\mathbf{x} := (\mathbf{x}_i : i \in U)$ and $\mathbf{z} := (\mathbf{z}_i : i \in U)$ shall be disregarded – the analysis is conditional on $\mathbf{x}$ and $\mathbf{z}$. In this section, we assume as DGP mixed effect models of the type

$$\eta_i | G \sim_{\text{ind}} N\left(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T G, \varphi^2\right), \qquad \varphi^2 \in [0, \infty) \tag{2.9a}$$

$$G \sim N(0, \Sigma) \tag{2.9b}$$

$$Y_i \sim_{\text{ind}} F(h^{-1}(\eta_i; \lambda), \sigma^2) \quad \text{for all} i \in U, \tag{2.9c}$$

where $N(\mathbf{m}, C)$ represents the (multivariate) normal distribution with expectation (vector) $\mathbf{m}$ and variance (covariance matrix) C. F is a distribution that will be discussed in the following.

We differentiate between two scenarios: First, consider the distribution of $\eta_i | G$ to be degenerate, i.e. $\varphi^2 = 0$. In that case, be h^{-1} an inverse link function and let F belong to the exponential family. In this case, the DGP from (2.9a) to (2.9c) describes a Generalized Linear Mixed Model (GLMM) [McCulloch, 1997, Booth and Hobert, 1999]. For some distributions F from the normal family, another scaling parameter $1 \neq \sigma^2 > 0$ is required.

The second scenario considers $\varphi^2 > 0$ and h^{-1} to be the inverse of the Box-Cox or Dual transformation where in both cases $h : (0, \infty) \to \mathbb{R}$ and

$$h_{BC}(y; \lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log y, & \text{if } \lambda = 0 \end{cases} \tag{2.10}$$

for the family of Box-Cox transformations [Box and Cox, 1964] and

$$h_D(y; \lambda) = \begin{cases} \frac{y^\lambda - y^{-\lambda}}{2\lambda}, & \text{if } \lambda > 0 \\ \log y, & \text{if } \lambda = 0 \end{cases} \tag{2.11}$$

for the family of Dual transformations [Yang, 2006]. When we have general notation with $h \in \{h_{BC}, h_D\}$, we omit the subscript. Then, (2.9a) to (2.9c) describe (approximately) a mixed model under data transformation with Y_i following an element F from the family of power distributions such that $h(Y_i, \lambda) \sim_{\text{ind}} \stackrel{d}{=} \eta_i$. In that case, F does not require an additional scaling parameter σ^2. That means, in the first setting, the scaling is determined by σ^2 and in the second by φ^2. We refer in the following to the scaling parameter simply by σ^2. From the context, it should be clear whether a scaling parameter in (2.9a) or (2.9c) is meant.

The approximation results from the fact that in Equations (2.10) and (2.11), $\lambda > 0$ requires for the inverse h^{-1} that the dependent variable be greater than zero. For normally distributed random variables (such

as the conditional distribution of $\boldsymbol{\eta} = (\eta_i \in U)$), this is not the case but holds with high probability for high expectations and relatively small variances. The precise model formulation for scenario two therefore would require a truncated normal distribution in (2.9a). This second type of mixed models is studied in Gurka et al. [2006] and Rojas-Perilla et al. [2017].

The symmetric variance-covariance matrix Σ is built of at most $\frac{q(q+1)}{2}$ different elements summarized in the vector $\boldsymbol{\rho}$. The model $\mathcal{M}$'s parameter vector is thus $\theta = (\boldsymbol{\beta}^T, \boldsymbol{\rho}^T, \sigma^2, \lambda)^T$, where λ might be omitted depending on the particular model set-up.

Note that in Section 1.3, we required an identically and independently distributed (iid) structure for the random variables (Y_i, Z_i). Usually, G can be splitted into sub-vectors,
$G^T = (G_{11}^T, \ldots, G_{1q_1}^T, G_{21}^T, \ldots, G_{2q_2}^T, \ldots, G_{kq_k}^T)$ such that $G_{l_1m_1} \sim_{iid} G_{l_2m_2}$ if $l_1 = l_2$ and $G_{l_1m_1} \perp G_{l_2m_2}$ for any $l_1 \neq l_2$ and m_1,m_2. For $l_1 = l_2$ and $m_1 \neq m_2$, such sub-vectors are (by the construction of $Z = \mathbf{z}$) added to different units $i \in U$. Define an isomorphic map to the partitions of $\{1, \ldots, q\}$

$$\iota : \{(l, m) : l = 1, \ldots, k,\ m = 1, \ldots, q_m\} \to \{\mathcal{C} : \cup_{C \in \mathcal{C}} C = \{1, \ldots, q\}\}$$

such that the conditions above are fulfilled for the corresponding sub-vectors of G. Stacking together such units $i \in U$ that share the same random sub-vectors of G, i.e. $(\tilde{Y}_{\iota(l,m)} := (Y_i : Z_{i,\iota(l,m)} \not\equiv 0))$ we can restablish the iid assumptions for $(\tilde{Y}, \tilde{X}, \tilde{Z})$. In that case, ML estimators for $\boldsymbol{\rho}$ are consistent if no additional covariance structure is imposed on G_{lm}.Because the model law is not any more a product measure for (Y, X, Z), we refer in the following to the joint distribution of (Y, X, Z) simply as $P_{\mathcal{M}_\theta}$ instead of $P_{\mathcal{M}_\theta}^N$.

Let $\boldsymbol{\gamma}$ be a realization of G. The finite population joint log-likelihood that would be used for parameter estimation under a purely model-based framework ignoring P_D (say g_U in the previous section), is

$$\mathcal{LL}(\mathbf{y}, \boldsymbol{\gamma}; \theta) \triangleq \underbrace{\sum_{i \in U} \nu_i}_{=:A} + B \tag{2.12a}$$

with

$$\nu_i \triangleq \left(\frac{\eta_i y_i - b(\eta_i) + c(y_i)}{a(\lambda)} \right) \tag{2.12b}$$

in the GLMM setting and

$$\nu_i \triangleq \left(-\frac{(\eta_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{z}_i^\top \boldsymbol{\gamma})^2}{2\sigma^2} + \log \frac{\partial h(y_i;\lambda)}{\partial y_i} - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log \sigma^2 \right) \tag{2.12c}$$

in the power transformation setting and

$$B = -\frac{1}{2}\log\det\Sigma - \frac{1}{2}\boldsymbol{\gamma}^\top \Sigma^{-1}\boldsymbol{\gamma} - \frac{q}{2}\log(2\pi) \quad . \tag{2.12d}$$

For GLMMs, the functions a, b and c are known and depend on the distribution F. The fact that the first term in (2.12b) is $\eta_i y_i$ implies that we consider here canonical link functions h. Note that η_i is fixed conditional on $\boldsymbol{\gamma}$ in the GLMM framework whilst it is a normal random variable with variance σ^2 under the power transformation framework. Due to the positivity requirement for power transformations, the population log-likelihood (2.12c) is only approximate, ignoring the truncation.

Both shapes in (2.12) reveal the nice property that – assuming that $\boldsymbol{\gamma}$ could be observed – optimization of $\mathcal{LL}$ with respect to $\boldsymbol{\rho}$ is independent from that of the fixed effects parameters $(\boldsymbol{\beta}^\top, \sigma^2, \lambda)$, which eases the computational burden.

For the GLMM setting, note that the Hessian of (2.12b) is globally concave in the fixed model parameters and consequently, the maximizer

$$\hat{\theta}^{pop} = \arg\max \mathcal{LL}(\mathbf{y}, \boldsymbol{\gamma}; \theta) \tag{2.13}$$

is unique given that $\boldsymbol{\rho}$ is such that Σ is unique. For example, $\hat{\boldsymbol{\rho}}$ is unique when G is block diagonal with repeating blocks and without a structure other than symmetry and positive definiteness imposed on disjoint blocks. A proof can be found in the Appendix A. Like in the GLM setting [Wedderburn, 1976], thus, GLMMs have unique joint maximum likelihood estimators under the canonical link, with unstructured random effects covariance matrix.

For the setting under Box-Cox and Dual transformations, however, the log-likelihood is more problematic: First of all, note that the log-likelihood (2.12c) is for a 'true' $\lambda \neq 0$ only approximative because the support of $h(Y_i;\lambda)$ is bounded for the Box-Cox transformation, which is in contrast to a Gaussian random variable's support [Hernandez and Johnson, 1980b, Yang, 2006].

Second, for a well-defined maximum likelihood estimator, the parameter space of λ must be necessarily restricted to $[0, \infty)$ for the Dual transformation to make models identifiable. This restriction is not always mentioned in the literature [Yang, 2006].

Third, though $\hat{\boldsymbol{\beta}}_U$ and $\hat{\rho}_U$ for the finite population are uniquely determined given $\hat{\lambda}_U$, the concentrated log-likelihood has second order

derivatives with respect to λ that are only concave in an environment around $\hat{\lambda}_U$. This is in so far problematic as numerical optimization algorithms act locally and depend on the choice of starting values. The partial differentiation of (2.12c) with respect to (β^T, λ) yields the score function. For the Box-Cox transformation, the score is equivalent to the estimating equations discussed in Foster et al. [2001]. The score's roots on $[0, \infty)$ for λ are not necessarily unique in expectation, thus, the ML is not $P_{\mathcal{M}_\theta}$-consistent if the compact interval $I \subset [0, \infty$ in which one searches for λ is too large.

Hernandez and Johnson [1980b] give conditions under which (2.12c) yields strongly consistent parameter estimators under Box-Cox transformations for the fixed effects – amongst others, a unique global maximum of $E_{\mathcal{M}_\theta}[\mathcal{LL}]$ must exist. The findings in Hernandez and Johnson [1980b] are easily transferable to the mixed effect model as shown in Appendix A.

Assuming that the consistency conditions hold for θ and its parameter space Θ, we can thus go on to introduce a survey design P_D. Again, the aim is to have a $P_{\mathcal{M}_\theta, D}$-consistent estimator for θ.

2.2.2 *Likelihood Approach under Survey Sampling*

From the finite population U, a random sample $S \subset U$ is drawn with survey design P_D, $S \sim P_D(\cdot; (Y, X, Z))$. Because of a possible design informativity (cf. Section 2.1), it is thus desirable to construct an estimator g_S for the model parameter θ, that is not only consistent for $S \perp Y$ given X and Z (cf. Remark 3) but also in the general case. Let $P_{\mathcal{M}_\theta, D}$ be defined like in (1.13). The iid assumption comes from rearranging to $(\tilde{Y}, \tilde{X}, \tilde{Z})$ like in the previous section. We define further the random subset $\tilde{S} \subseteq \{(l, m) : l = 1, \ldots, k,\ m = 1, \ldots, q_l\}$ where $\tilde{S}(S) := \{(l, m) : \exists i \in S : Z_{i\iota(l,m)} \neq 0\}$. This means that $\tilde{S}$ captures the indices whose corresponding element in G appears also impacts Y_S. G_S is the sub-vector of G such that there are units in S which have added elements from G_S.

The unweighted sample likelihood $\mathcal{LL}_S$ is

$$\mathcal{LL}_S(\mathbf{y}, \boldsymbol{\gamma}; \theta) \triangleq \underbrace{\sum_{i \in U} \mathbb{1}_S(i) \cdot \nu_i}_{=: A_S} + B_S \tag{2.14}$$

where

$$B_S \triangleq -\frac{1}{2} \log\det \Sigma_S - \frac{1}{2} \boldsymbol{\gamma}_S^T \Sigma_S^{-1} \boldsymbol{\gamma}_S - \frac{\dim \Sigma_s}{2} \log(2\pi) \tag{2.15}$$

and

$$\Sigma_S \triangleq \operatorname{diag}\left(\underbrace{\Sigma_1, \ldots, \Sigma_1}_{\sum_{m=1}^{q_1} \mathbb{1}_{\tilde{S}}((1,m))\text{- times}}, \ldots, \underbrace{\Sigma_k, \ldots, \Sigma_k}_{\sum_{m=1}^{q_k} \mathbb{1}_{\tilde{S}}((k,m))\text{- times}} \right). \tag{2.16}$$

In a first step, assume that the survey design is such that the probability equals one that the outcome $\mathbf{z}_S := (\mathbf{z}_i : i \in S)$ has full column rank q. This means that P_D is constructed such that for a grouping $j = 1, \ldots, k$ in G, $G^\top = (G_{11}^\top, \ldots, G_{1q_1}^\top, \ldots, G_{k1}^\top, \ldots, G_{kq_k}^\top)$ with $\sum_{j=1}^{k} |G_{j1}| = q$, and for all $G_{jm} \sim G_{jv}$ and $G_{jm} \perp G_{j'v}$, at least one observation of the finite population U enters the sample, i.e. $\tilde{S} \equiv \{1, \ldots, q\}$. Then we can simplify from B_S to B. For example, this is the case when the survey design is stratified.

S can also be represented by the vector $\mathbf{1}_S$ (cf. Section 1.3); one gets therefore for the first moment in such designs

$$\begin{aligned}
\mathrm{E}_{\mathcal{M}_\theta, D}\left[\mathcal{LL}_S | X, Z\right] &= \int_{\mathcal{S}\times\Omega} \mathcal{LL}_S \, \mathrm{d}\, P_{\mathcal{M}_\theta, D}(\cdot|X, Z) \\
&\overset{\text{Eq.(1.13)}}{=} \sum_{s\in\mathcal{S}} \int_\Omega \mathcal{LL}_s \cdot P_D\left(s; Y, X, Z\right) \, \mathrm{d}\, P_{\mathcal{M}_\theta}(\cdot|X, Z) \\
&= \int_\Omega \sum_{s\in\mathcal{S}} \mathcal{LL}_s \cdot P_D\left(\mathbf{1}_s; Y, X, Z\right) \, \mathrm{d}\, P_{\mathcal{M}_\theta}(\cdot|X, Z) \\
&= \int_\Omega \sum_{i\in U} P_D(i \in S; Y, X, Z) \cdot v_i + B \, \mathrm{d}\, P_{\mathcal{M}_\theta}(\cdot|X, Z) \qquad (2.17) \\
&\neq \int_\Omega \sum_{i\in U} v_i + B \, \mathrm{d}\, P_{\mathcal{M}_\theta} = \mathrm{E}_{\mathcal{M}_\theta}\left[\mathcal{LL}|X, Z\right] \quad . \qquad (2.18)
\end{aligned}$$

For a non-informative design, however, the integrals conditioned on (X, Z) can be interchanged and we get for (2.17) the same maximizer due to the separability of the optimization of ρ and $(\beta, \lambda, \sigma^2)$. Theorem 5.1 from Rubin-Bleuer and Kratina [2005] holds and thus, the ML estimators conditional on (X, Z) defined by $\arg\max \mathcal{LL}_S$ converge in $P_{\mathcal{M}_\theta, D}$-probability to θ under the conditions that are required for convergence in $P_{\mathcal{M}_\theta}$. Note, though, that the maintenance of strong consistency is not proven.

Because Theorem 5.1 from Rubin-Bleuer and Kratina [2005] is not applicable under non-informative design, alternatives to the estimator $\arg\max \mathcal{LL}_S$ for θ are required that account for survey design. Note that $\mathcal{LL}$ has the shape of a population total ($A = \sum_{i\in U} v_i$) plus a constant (B) that can – under a stratified design like outlined above – be estimated by the HT-estimator [Horvitz and Thompson, 1952]

$$\widehat{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}, S; \theta) := \sum_{i\in U} w_i \cdot \mathbb{1}_S(i) \cdot v_i + B \tag{2.19}$$

where B is independent from the sample because of the design assumption on P_D. Clearly,

$$E_D\left[\widehat{\mathcal{LL}}\right] = \mathcal{LL} \quad . \tag{2.20}$$

Furthermore, as shown in Appendix A, the application of Remark 5 yields

$$E_{\mathcal{M}_\theta,D}\left[\widehat{\mathcal{LL}}|X,Z\right] = E_{\mathcal{M}_\theta}\left[\mathcal{LL}|X,Z\right] \quad . \tag{2.21}$$

The finite population estimator $\frac{1}{N}\mathcal{LL}$ is consistent due to the law of large numbers, $\frac{\mathcal{LL}}{N} \xrightarrow[N\to\infty]{P_{\mathcal{M}_\theta}} E_{\mathcal{M}_\theta}[\mathcal{LL}]$, for a population growth where also the dimension of Z grows ($q \to \infty$) because the iid assumption holds for $(\tilde{Y},\tilde{X},\tilde{Z})$. Furthermore, Remark 1 states the consistency of the HT estimator under certain survey designs. Stratified Random Sampling (StratRS) belongs to these designs when the number of strata (i.e. the groupings $j = 1,\dots,k$ of G) increases and the sample size within strata is fixed/ increases slower than the growth rate of stratum size [Chauvet, 2014, Proposition 3.1]. We can thus conclude from the second-order differentiability of $\mathcal{LL}$ the continuity of $\arg\max \mathcal{LL}$ and therefore the model-design consistency of

$$\hat{\theta} = \arg\max_{\tilde{\theta}} \widehat{\mathcal{LL}}(Y,G,S;\tilde{\theta}) \quad . \tag{2.22}$$

In a next step, assume that the survey design P_D is not necessarily stratified. In that case, it is not assured anymore that observations containing information on G_{jv}, $v = 1,\dots,q_j$ enter the sample S ($\tilde{S}$ is in that case not a partition of $\{1,\dots,q\}$). In that case, $B \neq B_S$ does not enter as such the sample likelihood, because only observations Y_S containing information from the random sub-vector γ_S are in the sample S. Consequently, also the sample variance-covariance matrix Σ_S (2.16) is random. If the survey design depends on G, i.e. $P_D = P_D(\cdot;Y,G,X,Z) \neq P_D(\cdot;Y,X,Z)$, the unweighted ML covariance matrix (cf. the derivation thereof in the Appendix A) is not consistent any more.

Though Σ_S resembles in its structure the variance-covariance matrix Σ, it is usually of lower dimension as some of the block matrices Σ_j (or the m-th repetition thereof, cf. Equation (2.16)) are missing. Survey-weighting of the first component in (2.19) thus yields a disproportion between the terms $A_S^w = \sum_{i\in U} w_i \cdot \mathbb{1}_S(i) \cdot v_i$ and B_S in the sense that it does not correspond – in expectation – to the relation between A and B in Equation (2.12), which can distort parameter estimation if the survey design is informative (but not informative on G).

Assume therefore that there is a smaller finite population U_n with $n = E_D[|S|]$ generated by the same statistical model $\mathcal{M}_\theta$. Obviously, the

asymptotic findings hold for the estimators of this population, too, as $U_n \in \{U_N\}_{N\in\mathbb{N}}$. Then, the scaled joint log-likelihood

$$\widetilde{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}, S; \theta) \triangleq \sum_{i\in U} \tilde{w}_i \cdot \mathbb{1}_S(i) v_i + B_S \tag{2.23}$$

$$\tilde{w}_i \triangleq \frac{w_i}{\sum_{i\in U} \mathbb{1}_S(i) \cdot w_i} |S| \tag{2.24}$$

can be interpreted as an estimator of this reduced population joint log-likelihood if $S \perp G|Y, X, Z$. This seems a plausible assumption in practice as the term G is usually referred to as 'unobserved' heterogeneity - it is thus a theoretical concept. If clustering or stratification is part of P_D, this is therefore rather implemented using the observables (Y, X, Z).

An alternative to the rescaling of the HT joint log-likelihood would be a weighting procedure for the terms in B_S, leading to a weighted covariance matrix. This would be complicated, though, as one would require $P_D(\iota(j, m) \in \tilde{S})$ at least for those random sub-vectors that are in $\tilde{S}$. This is however the probability that $P_D\left(\mathcal{C}_{jm} \subset S\right)$ where $\mathcal{C}_{jm} = \left\{i \in U : Z_{i,\iota(jm)} \not\equiv 0\right\}$, requiring to know higher order inclusion probabilities of the units in U.

In the following, we use for generalization $\boldsymbol{\gamma}_S$, remembering that $\boldsymbol{\gamma}_S = \boldsymbol{\gamma}$ and $\Sigma_S = \Sigma$ under Z-stratified designs. Furthermore, we omit the explanatories $\mathbf{x}$ and $\mathbf{z}$, though the analysis is conditioned on the random outcomes.

2.2.3 *Maximization of the HT Joint Log-likelihood*

The difficulty of ML estimation in this context is that the random vector realization $\boldsymbol{\gamma}$ is not observed in practice. The Expectation Maximization (EM)-algorithm [Dempster et al., 1977], though, offers a chance to get the ML estimator given a good starting value θ_0 for the iterative optimization procedure. For GLMMs, the likelihood is globablly concave (cf. Subsection2.2.1) meaning that $\mathcal{LL}$ only has one stationary point equal to its global maximum – EM-maximization of $\widetilde{\mathcal{LL}}$ therefore should yield a model-design consistent estimator. For LMMs under power transformations, on the other hand, global concavity is not given and the choice of θ_0 is more important.

The EM algorithm iterates, within the survey sampling framework, for a given realization s and $\mathbf{y}$ between the calculus of the expectation of

the joint log-likelihood given parameter estimate θ_k, sample s and the observed information $\mathbf{y}_s = (\mathbf{y}_i : i \in s)$

$$\mathrm{E}_{\mathcal{M}_\theta, D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta)|\theta_k\right] \triangleq \int_{\mathbb{R}^{|\tilde{s}|}} \sum_{i \in U} \tilde{w}_i \cdot \mathbb{1}_S(i) \cdot \nu_i + B_S \, \mathrm{d}\, P_{\mathcal{M}_{\theta_k}}(\mathrm{proj}_G^{s\,-1}) \tag{2.25a}$$

and the maximization of this conditional expectation

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} \mathrm{E}_{\mathcal{M}_\theta, D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta)|\theta_k\right] . \tag{2.25b}$$

The applications of the EM algorithm for unobserved heterogeneity in longitudinal studies dates back to Laird and Ware [1982], where the author discussed precisely a LMM framework. Wu [1983] shows that the EM algorithm converges to a stationary point of the joint log-likelihood under conditions that are fulfilled here, namely the continuity of $\mathrm{E}_{\mathcal{M}_\theta, D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta)|\theta_k\right]$ and $\nabla_\theta \mathrm{E}_{\mathcal{M}_\theta, D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta)|\theta_k\right]$ in both θ and θ_k. In the GLMM framework, even convergence to the global maximum is guaranteed due to the unimodality of $\widetilde{\mathcal{LL}}$ [Wu, 1983]. In consequence, at least for an appropriate choice of θ_0 in an environment around θ, convergence to the ML estimators for the rescaled HT joint log-likelihood (2.23) is assured, which implies the model-design consistency of the EM estimators for the number of optimization steps k going to infinity.

The E-step (2.25a) is problematic, though. If $h \neq \mathrm{id}$, the conditional density of γ_s given s and $\mathbf{y}_s$ is not normal anymore. Even for $h = \mathrm{id}$, the (conditional) normality assumption for the unobserved heterogeneity can be violated for a complex design P_D. Thus, contrary to the argument in Laird and Ware [1982] who ignored survey design, analytical integration can become cumbersome. When the dimension of the random effects q is large, also numerical approximation loses accuracy and can become impractical. A possible solution to that problem is Monte Carlo (MC) integration, which has already been applied successfully to GLMMs [McCulloch, 1997, Booth and Hobert, 1999, Zipunnikov and Booth, 2006]. However, note that none of the named studies used Monte Carlo integration when survey weights or transformations on the dependent variable were required. When the E-step of the EM algorithm is approximated by MC integration, this modified algorithm is called Monte Carlo Expectation Maximization (MCEM) algorithm [Wei and Tanner, 1990]. In summary, the idea is to sample in E-step k $b = 1, \ldots, B_k$ times

$$G_s^{(b)} \sim F(G|\mathbf{y}, s; \theta_k) \tag{2.26a}$$

and to use the MC mean in step k as an estimator for the integral:

$$\hat{E}_{\mathcal{M}_\theta,D}\left[\widetilde{\mathcal{LL}}\right] \triangleq \frac{1}{B_k}\sum_{b=1}^{B_k}\widetilde{\mathcal{LL}}\left(\mathbf{y}, G^{(b)}, s; \theta\right) \tag{2.26b}$$

and to maximize this approximation in the subsequent M-step:

$$\theta_{k+1} \triangleq \arg\max_\theta \hat{E}_{\mathcal{M}_\theta,D}\left[\widetilde{\mathcal{LL}}\right] \tag{2.26c}$$

[Neath, 2013]. Equation (2.26b) converges for $B_k \to \infty$ by the strong law of large numbers with probability one, if $G \perp S|X, Y, Z$ to the expectation of (2.23). Remember that the conditional orthogonality is necessary for $\widetilde{\mathcal{LL}}$ to be a consistent estimator for the reduced finite population. The law of large numbers, however, is only applicable for $B_k \to \infty$ within one E-step k. For $k \to \infty$, the behavior is not so obvious. Furthermore, for consistency results on $P_{\mathcal{M}_\theta,D}$, also the number of observations, $|S|$ needs tend to infinity which requires a profounder study of the convergence behaviour of MCEM. Neath [2013] summarizes conditions under which the MCEM algorithm converges for curved exponential families. However, these are not the models considered here and thus, we refer to Sherman et al. [1999, Assumptions A1′ and A2′]. For almost sure convergence of $\{\theta_k\}_{k\in\mathbb{N}}$, it must hold for some $\delta > 0$ and $B_k \equiv B$

$$\lim_{B\to\infty} B^\delta E_{\mathcal{M}_\theta,D}\left[\sup_{\theta_k}\|\arg\max_{\theta'}\hat{E}_{\mathcal{M}_\theta,D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta')|\theta_k\right] - \arg\max_{\theta'} E_{\mathcal{M}_\theta,D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta')|\theta_k\right]\|\right] < \infty \tag{2.27a}$$

and

$$\arg\max_\theta E_{\mathcal{M}_\theta,D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta)|\theta_k\right] \text{ is Lipschitz in } \theta_k. \tag{2.27b}$$

Sherman et al. [1999] add to the requirements another limit which is because they employ a Gibbs sampler for the MC E-step. Depending on the Lipschitz constant, different requirements on δ are imposed to guarantee almost sure convergence of the MCEM. In the next paragraph, we employ independent sampling for $G^{(b)}$, $b = 1, \ldots, B$, so we omit this limit condition here. Condition (2.27a) means that uniform convergence in L^1 is required for the MC integration – the optimizer θ_{k+1} of the MC E-step must be close to that one of the original E-step under all possible predecessors θ_k with a certain speed faster than $\mathcal{O}(B^\delta)$. Condition (2.27b) assures good (and stricter than usual) properties of the original EM algorithm that become necessary to link EM and MCEM: It is equivalent

to requiring that the set of estimators based on $\theta_k \in \Theta$ are equicontinuous sets. If (2.27b) holds with a Lipschitz constant < 1, the sequence $\arg\max \mathrm{E}_{\mathcal{M}_\theta, D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta)\right]$ has a fix point for any $\theta_0 \in \Theta$ implying a set of stationary points. In condition i) of Theorem 1 in Sherman et al. [1999], the rate with whom the number of MC samples B must increase for $k \to \infty$ is given for contractions. If the Lipschitz constant is greater or equal to one, however, Sherman et al. [1999] provide faster growth rates for B. Booth and Hobert [1999] state convergence of MCEM for GLMMs, too, but do not give the necessary growth rates for B. For an overview about additional convergence results for MCEM, see Neath [2013].

Though we do not give a proof for the consistency of MCEM in the situation presented here, we have the following intuition: arg max is implicitly defined by the solution to the first order conditions and is continuous in its argument. For Θ compact and because $\mathrm{P}_{\mathcal{M}_\theta}$ is also continuous in θ, there exists a $\theta_k \in \Theta$ such that

$$\begin{aligned}
\sup_{\tilde{\theta}} \| & \arg\max_{\theta} \hat{\mathrm{E}}_{\mathcal{M}_\theta, D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta) | \tilde{\theta}\right] - \\
& \arg\max_{\theta} \mathrm{E}_{\mathcal{M}_\theta, D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta) | \tilde{\theta}\right] \| \\
= \| & \arg\max_{\theta'} \hat{\mathrm{E}}_{\mathcal{M}_\theta, D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta') | \theta_k\right] - \\
& \arg\max_{\theta'} \mathrm{E}_{\mathcal{M}_\theta, D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta') | \theta_k\right] \|
\end{aligned}$$

As $B \to \infty$, this term converges in probability to zero. If one can now show uniform integrability, it follows the L^1 convergence, i.e. Condition 2.27a. Furthermore, we have argued for the local convergence of the EM algorithm. This implies an environment around θ for which $\arg\max \mathrm{E}_{\mathcal{M}_\theta, D}\left[\widetilde{\mathcal{LL}}(\mathbf{y}, G, s; \theta) | \tilde{\theta}\right]$ is a contraction and no additional requirements on δ are needed.

Note further that convergence also requires the sequence $\{B_k\}_{k \in \mathbb{N}}$ needs to increase: Otherwise, the MC error persists and might for close to optimal $\hat{\theta}$ be so large that with positive probability, one quits the optimal neighborhood [Neath, 2013].

Still, the MCEM composed of the iterative application of Equations (2.26) poses a problem in practice: Sampling from the conditional distribution of G given $S = s$ and the observable random variables, $F(G|\mathbf{y}, s; \theta_k)$, is not straightforward – the distribution F might even be unknown. However, we have the relation

$$F(G|Y, S; \theta_k) = \frac{F(Y, G, S; \theta_k)}{F(Y, S; \theta_k)} \tag{2.28}$$

where the numerator is the joint distribution of Y, G and S and the denominator is the joint distribution of Y and S; G is integrated out.

Consequently, $F(G|Y,S;\theta_k)$ is proportional to $\exp(\widetilde{\mathcal{LL}})$. It is thus possible to generate random draws $\boldsymbol{\gamma}^{(b)}$, $b = 1,\ldots,B$ from a proposal with the same support as F (which is $\mathbb{R}^{\dim \Sigma_s}$) and to reweight the realizations in function of $\exp\left(\widetilde{\mathcal{LL}}(\boldsymbol{\gamma}^{(b)})\right)$ and $b = 1,\ldots,B_k$. This is an importance sampling approach where the normalizing constant $F(Y,S;\theta_k)$ is not necessary, i.e. self-normalized importance sampling [Owen, 2013, Chapter 9.2] with importance weights in step k for realization $\boldsymbol{\gamma}^{(b)}$ of $G^{(b)} \overset{d}{=} G$ equal to

$$\omega_k^{(b)} \triangleq \frac{\exp\left(\widetilde{\mathcal{LL}}(\mathbf{y},G^{(b)},s;\theta_k)\right)/h(\boldsymbol{\gamma}^{(b)}|\mathbf{y},s;\theta_k)}{\sum_{\iota=1}^{B}\exp\left(\widetilde{\mathcal{LL}}(\mathbf{y},G^{(\iota)},s;\theta_k)\right)/h(\boldsymbol{\gamma}^{(\iota)}|\mathbf{y},s;\theta_k)} \quad , \tag{2.29}$$

where h is the proposal density, possibly with parameters based on the available information or current parameter estimate θ_k. If importance sampling is used, Equations (2.26a) and (2.26b) must be replaced by

$$G^{(b)} \sim H(G|\mathbf{y},s;\theta_k) \tag{2.30a}$$

and

$$\hat{E}_{\mathcal{M}_\theta,D}\left[\widetilde{\mathcal{LL}}\right] \triangleq \frac{1}{B}\sum_{b=1}^{B}\omega_k^{(b)}\,\widetilde{\mathcal{LL}}\left(\mathbf{y},G^{(b)},s;\theta\right) \quad . \tag{2.30b}$$

There exist several suggestions for proposal distributions. Booth and Hobert [1999] argue that the joint likelihood – in their case, without (scaled) survey weighting – equals the product of the distribution of $Y|G$ and the distribution of G. Using hence $N(0,\Sigma_k)$ as proposal in step k simplifies the computation of the survey weights as the importance weight (2.29) for realization b in iteration step k reduces to

$$\frac{\exp\left(A_s^{(b)}\right)}{\sum_{\iota=1}^{B_k}\exp\left(A_s^{(\iota)}\right)} \quad .$$

However, for moderately large diagonals in Σ, the 'true', unobserved realizations $\boldsymbol{\gamma}$ may lie so far from the expectation 0 that this proposal becomes very inefficient. Again, the MC can then be too large to get close to optimal estimates. The same holds for the suggested rejection sampling or sequential MC sampling [Booth and Hobert, 1999] which can take a lot of iterations to pass through the sample space of G.

Instead, Pinheiro and Bates [1995] and Booth and Hobert [1999] suggest to center a symmetric distribution around the mode $\boldsymbol{\gamma}_k$ in iteration k of $\widetilde{\mathcal{LL}}(\mathbf{y},\boldsymbol{\gamma},s;\theta_k)$ in MCEM-step k and to use as variance-covariance matrix of the proposal H the inverse of the negative Hessian $-\left(\nabla^2\widetilde{\mathcal{LL}}\right)^{-1}$

at $\boldsymbol{\gamma}_k$. The proposal is motivated by a second order Taylor approximation of $\widetilde{\mathcal{LL}}$ around $\boldsymbol{\gamma}_k$:

$$\widetilde{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}, S; \theta_k) \approx \widetilde{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}_k, S; \theta_k) + \underbrace{\nabla^T \widetilde{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}_k, S; \theta_k)}_{=0} \cdot (\boldsymbol{\gamma} - \boldsymbol{\gamma}_k)$$
$$-\frac{1}{2} (\boldsymbol{\gamma} - \boldsymbol{\gamma}_k)^T \left(- \left(\nabla^2 \widetilde{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}_k, S; \theta_k) \right)^{-1} \right)^{-1} \cdot (\boldsymbol{\gamma} - \boldsymbol{\gamma}_k) \ . \tag{2.31}$$

Taking the exponential, we get a term that is proportional to a multivariate normal distribution with the named parameter values. Whilst Booth and Hobert [1999] recommend the multivariate t-distribution to favor sampling from the tails, Pinheiro and Bates [1995] proposes the multivariate normal distribution. The simulation studies in Burgard and Dörr [2019] and Dörr and Burgard [2019], suggest that the multivariate normal distribution is a good proposal.

From a computational point of view, this proposal is burdensome because an additional optimization must take place in each E-step k to determine the mode $\boldsymbol{\gamma}_k$. Furthermore, as multivariate normals are generated from the standard normal by multiplication with the covariance matrix' square root and addition of the expectation, a matrix C is required such that $CC^T = -\left(\nabla^2 \widetilde{\mathcal{LL}}(\mathbf{y}, \boldsymbol{\gamma}_k, S; \theta_k) \right)^{-1}$. Powell [1987] elaborates a version of the BFGS algorithm that returns such a matrix C as the square root of a working Hessian. Together with the importance sampling and the BFGS-algorithm suggested in Powell [1987], the MCEM algorithm is summarized in Algorithm 2.1.

The importance sampling approach and the suggested proposal distribution have another advantage: Even the simulation of the random effects accounts for the survey design as it is a function of $\widetilde{\mathcal{LL}}$. This is by purpose as integration shall be done with respect to $P_{\mathcal{M}_\theta, D}$. Alternative weighted estimators – for example You and Rao [2002] elaborate a weighted fixed effects estimator $\hat{\boldsymbol{\beta}}$ given the estimated variance components $\hat{\boldsymbol{\rho}}$ and $\hat{\sigma}^2$ that were estimated from an unweighted model – do not account for the design when estimating $\boldsymbol{\rho}$. Other weighted estimators are either restricted to LMMs or a diagonal random effects covariance matrix Σ [Pfeffermann et al., 1998] or at least restricted to a nested random effects structure [Rabe-Hesketh and Skrondal, 2006].

2.2.4 *Simulation Studies*

2.2.4.1 *Simulation Set-up*

DATA GENERATING PROCESS

The importance of survey weighting in regression analysis receives increasing attention. Olson et al. [2003] and Fairlie and Robb [2009] are

Algorithm 2.1 MCEM Algorithm with Importance Sampling

Require: Training data, survey weights, start values θ_0, convergence criterion, $K \in \mathbb{N}$, $B_0 \in \mathbb{N}$

Ensure: Vector of parameter estimators $\hat{\theta}$

for $k = 0, \ldots, K$ **do**

Determine $\gamma_k = \arg\max \widetilde{\mathcal{LL}}(\mathbf{y}, \gamma, S; \theta_k)$ and store C with $CC^T = -\left(\nabla^2 \widetilde{\mathcal{LL}}(\mathbf{y}, \gamma_k, S; \theta_k)\right)^{-1}$ ▷ [Powell, 1987]

Sample $b = 1, \ldots, B_{k-1}$ ▷ (E-Step)

$$U^{(b)} \sim N(0, I_q)$$

and generate

$$G^{(b)} \leftarrow \gamma_k + C \cdot U^{(b)}$$

Calculate importance weights:

$$\omega_k^{(b)} \leftarrow \frac{\exp \widetilde{\mathcal{LL}}\left(\mathbf{y}, G^{(b)}, S, ; \theta_k\right)}{\exp\left(-\frac{1}{2}\left(G^{(b)} - \gamma_k\right)^T \left(CC^T\right)\left(G^{(b)} - \gamma_k\right)\right)}$$

$$\omega_k^{(b)} \leftarrow \frac{\omega_k^{(b)}}{\sum_{\iota=1}^{B_k} \omega_k^{(\iota)}}$$

▷ (The rest of the proposal cancels out due to self-normalization)

Determine

$$\theta_{k+1} \leftarrow \arg\max_{\theta \in \Theta} \frac{1}{B_k} \sum_{b=1}^{B_k} \omega_k^{(b)} \widetilde{\mathcal{LL}}\left(\mathbf{y}, G^{(b)}, S; \theta\right)$$

▷ (M-Step)

if convergence criterion is met **then**

break

end if

Increase $B_{k+1} \geqslant B_k$

end for

Set $\hat{\theta} \leftarrow \theta_k$

examples for the use of weights in regression analysis on business surveys. But for regression in combination with a mixed effects structure and possibly a non-normal dependent variable, though, the impact of the sample design is not yet intensively studied. In addition, the functioning of the presented MCEM algorithm must be validated. Hence, we describe in the following simulation studies similar to Burgard and Dörr [2019] and Dörr and Burgard [2019]. For scenarios with other distributions from the exponential family or the Box-Cox transformation, consult these references.

Examples for informative designs in simulation studies are given in Pfeffermann et al. [1998] and Rabe-Hesketh and Skrondal [2006]. Similar to these, the design informativity in our simulation study will be based on the idiosyncratic error $Y_i - h^{-1}(\eta_i)$ given $G = \gamma$, h^{-1} being either the inverse link function or the inverse of the data transformation. However, both Pfeffermann et al. [1998] and Rabe-Hesketh and Skrondal [2006] also employ design informativity based on G, which is not plausible in practice and for which the previous analysis of our estimation algorithm is not applicable. We ignore design informativity at this level.

The DGP (which will be reproduced in each simulation run $r = 1, \dots, 1000$) is

$$\eta_i = \beta_0 + x_{1,i}\beta_1 + x_{2,i}\beta_2 + G_{1,d} + (1, x_{1,i})^T G_{2,g} \tag{2.32a}$$

$$G_{1,d} \sim_{iid} N\left(0, \sigma_1^2\right), \qquad d = 1, \dots, 10 \tag{2.32b}$$

$$G_{2,g} \sim_{iid} N\left(0, \begin{pmatrix} \sigma_2^2 & \rho \\ \rho & \sigma_3^2 \end{pmatrix}\right), \quad g = 1, \dots, 20 \tag{2.32c}$$

if unit i is assumed to be in $U_d \cap U_g$ where both sets stem from possibly independent partitions of the finite population, $U = \dot{\cup}_{d=1}^{10} U_d$ and $U = \dot{\cup}_{g=1}^{2} 0 U_g$. The population size is $|U| = N = 1000$ and the assignment of unit i to the subsets U_d, $d = 1, \dots, 10$ and U_g, $g = 1, \dots, 20$ are completely random and generated once for all 1000 runs. Furthermore, the subsets have different (random) sizes – which often destabilizes variance estimation. The cross-combination of partitions $\{U_g\}_{g=1,\dots,20}$ and $\{U_d\}_{d=1,\dots,10}$ are given in Table B.2 in Appendix B.

The parameters are set to $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2) = (7, -1, 1)$ and $\boldsymbol{\rho}^T = (\sigma_1^2, \sigma_2^2, \rho, \sigma_3^2) = (0.5625, 0.7, 0.3, 0.5)$. In a first scenario, a generalized linear mixed model with a Gamma distributed dependent variable with scale 0.5 and expectation $E_{\mathcal{M}_\theta}[Y_i | G_1, G_2] = \frac{1}{\eta_i}$ is assumed, which means that the DGP continues with

$$Y_i^{(1)} \sim \Gamma\left(1/\eta_i, 0.5\right) \quad , \tag{2.32d}$$

where negative realizations η_i are disregarded in the simulation; remember that $\mathcal{LL}$ was only an approximation for a truncated normal under power transformations.

In a second scenario, a mixed effects model with a dual transformation, the DGP goes on with

$$h\left(Y_i^{(2)};\lambda\right) \sim N\left(\eta_i, 0.16\right) \tag{2.32e}$$

and $\lambda = 0.3$ and h is defined in Equation (2.11). Both scenarios generate a (positive-valued) skewed variable of interest in the population like they can be found in business surveys, for example variables such as 'investment' or 'return of sales'. Note that the auxiliary variables $X_{1,i} \sim \text{Unif}[-0.5, 0.5]$ and $X_{2,i} \sim \text{Unif}[0, 1]$ are generated once for each $i \in U$ and kept constant across the $R = 1000$ simulation runs.

SAMPLING RANDOMIZATION

From each of the $r = 1, \ldots, 1000$ generated populations, three samples with fixed sample size $n = 200$ are drawn under three different survey designs. The first design is non-informative and is a Probability Proportional to Size (pps) (without replacement) design with

$$P_{D_0}(i \in S; \mathbf{X}) \propto x_{2,i} \quad . \tag{2.33}$$

The other two designs are informative with respect to the DGP. Specifically, they are functions of the errors of the dependent variable $Y^{(j)}$, $j = 1, 2$

$$\varepsilon_i^{(l)} \triangleq Y_i^{(l)} - E_{\mathcal{M}_\theta}\left[Y_i^{(l)} | G_1, G_2, \mathbf{x}_i, \mathbf{z}_i\right] \tag{2.34a}$$

$$e_i^{(l)} \triangleq \begin{cases} 6 \text{ if } \varepsilon_i^{(l)} < F_N^{-1}(0.25) \\ 4 \text{ if } \varepsilon_i^{(l)} \in \left[F_N^{-1}(0.25), F_N^{-1}(0.5)\right) \\ 2 \text{ if } \varepsilon_i^{(l)} \in \left[F_N^{-1}(0.5), F_N^{-1}(0.75)\right) \\ 1 \text{ else} \end{cases} \tag{2.34b}$$

$$P_{D_l}(i \in S; \mathbf{X}, Y^{(j)}) \propto e_i^{(l)} \qquad l = 1, 2. \tag{2.34c}$$

F_N denotes here the empirical Cumulative Distribution Function (cdf) on the finite population. For similar informative designs based on the dependent variable's error, see Rabe-Hesketh and Skrondal [2006] and Burgard and Dörr [2019] and Dörr and Burgard [2019]. For the dual transformation, the expectation $E_{\mathcal{M}_\theta}\left[Y_i^{(j)} | G_1, G_2, \mathbf{x}_i, \mathbf{z}_i\right]$ was calculated via MC-integration. Note that the designs P_{D_1} and P_{D_2} need to be generated in each simulation run. All three survey designs do not reflect the partitioning of the finite population: Neither $\{U_d\}_{d=1,\ldots,10}$ nor $\{U_g\}_{g=1,\ldots,20}$ serve as strata or Primary Sampling Units (PSUs).

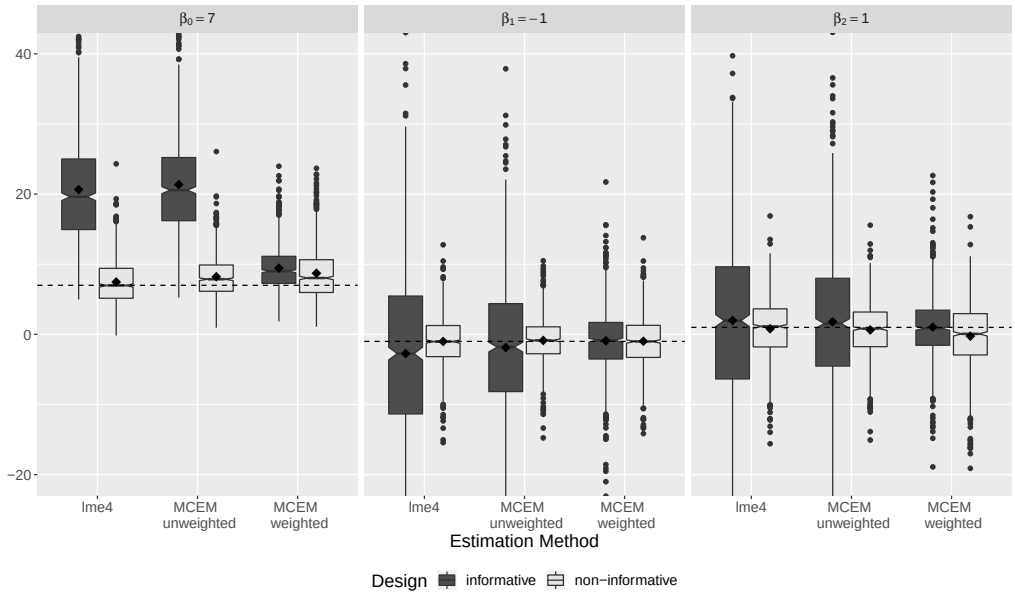
All sampling designs P_{D_j}, $j = 0, 1, 2$, were implemented using the R-package `sampling` [Tillé and Matei, 2016]. As a benchmark, the GLMM is

also estimated with `lme4` using penalized weighted least squares [Bates, 2018]. For the mixed model under the dual transformation, the estimation procedure introduced in Gurka et al. [2006] for Box-Cox transformations and extended by Rojas-Perilla et al. [2017] to dual transformations was programmed by the author. The code is based on `lme4`. An inaccuracy in Gurka et al. [2006] and adopted by Rojas-Perilla et al. [2017] that leads to small estimation biases is discussed in Appendix A. We circumvent this inaccuracy by the optimization via grid-search of the stated, original log-likelihood depending on λ. As start values for Algorithm 2.1, we set the results of `lme4` with the survey weights plugged into the heteroskedasticity weighting procedure.

2.2.4.2 *Simulation Results*

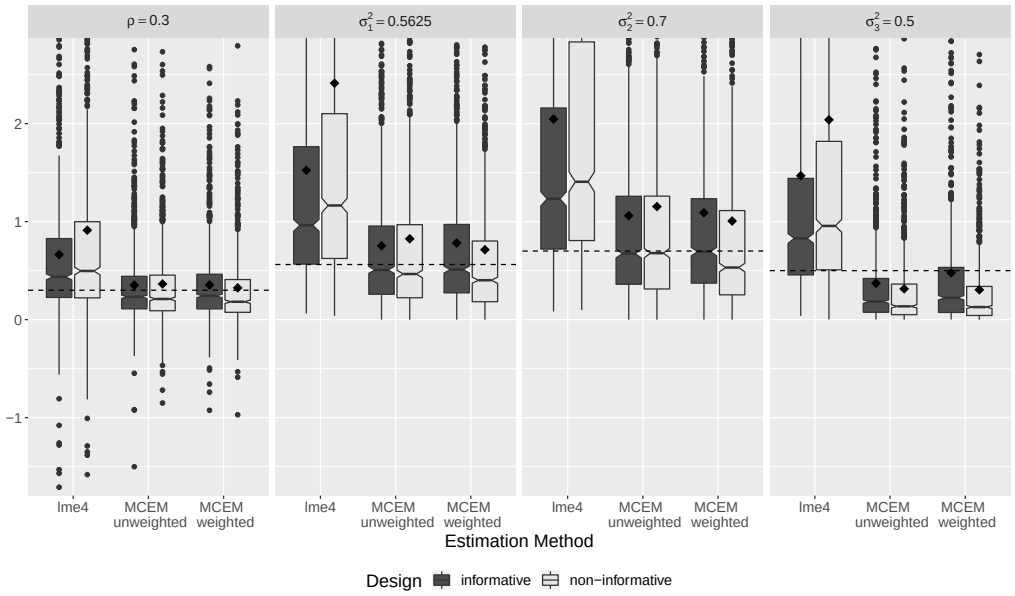
The simulation results are summarized in Figures 2.1 to 2.5. Boxplot 2.1 reveals that both estimation methods (the MCEM and `lme4`) perform well in the fixed parameter estimation for the non-informative desgin when the dependent variable is conditionally Gamma distributed. Furthermore, the inclusion of survey weights does not impact the efficiency of the estimator as can be concluded from the inter-quartile length. Under informative design P_{D_1}, however, both `lme4` and Algorithm 2.1 without weights perform similarly bad in the estimation of the slope parameter, although the unweighted MCEM proves a little stabler in the slope parameter estimation. The inclusion of survey weights in the MCEM reduce on the other hand the bias in the intercept under informative design almost completely. For the random effects components, however, `lme4`

Figure 2.1: MC-Distribution of $\hat{\beta}$ under $Y_i^{(1)} \sim \Gamma(1/\eta_i, 0.5)$



yields even for the non-informative design biased results under the conditional Gamma distribution (Boxplot 2.2). The bias is sometimes even higher than for the informative design, which gives a hint that the variability of the estimator is very high, even under favorable designs. The MCEM is both, weighted and unweighted, closer to the true parameter values. For the random effects components, the weighting does not play a role because the information in P_{D_1} does not account for the random effects. N In the dual case, on the other hand, the MCEM yields similar

Figure 2.2: MC-Distribution of $\hat{\rho}$ under $Y_i^{(1)} \sim \Gamma(1/\eta_i, 0.5)$



results for $\hat{\lambda}$ for non-informative and informative designs P_{D_0} and P_{D_2} as can be seen in Boxplot 2.3. The `lme4`-based estimator slightly overestimates the true parameter $\lambda = 0.3$. For the fixed effects parameters, especially the unweighted method using `lme4` yields a biased intercept (cf. Figure 2.4) – even under the non-informative design. When the transformation parameter λ is estimated biasedly, as happened in Figure 2.3, this translates directly to biased fixed parameter estimators. The bias due to the informative design, though, is hardly observable in the results. The average intercept estimate for the weighted and unweighted MCEM differ only slightly. However, it can be learned that the outliers are fewer and less extrem under weighting. For the random effects components $\boldsymbol{\rho}$, the same that was already observed for the GLMM can also be found in Figure 2.5: Whilst due to the biases in $\hat{\lambda}$ and $\hat{\beta}_0$, `lme4` cannot return reliable estimators for the variance components, the MCEM yields estimators that are close to the true values. Only the variance component σ_3^2 seems to be more biased. Comparing the weighted and the unweighted MCEM, the

Figure 2.3: MC-Distribution of $\hat{\lambda}$ under $h(Y_i^{(2)}) \sim N(0, 0.16)$

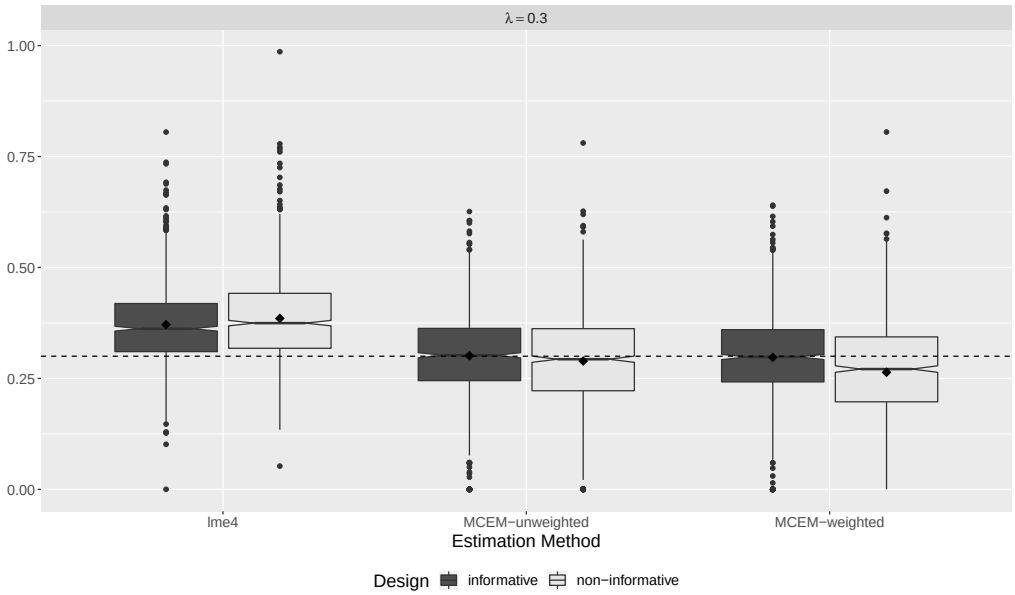
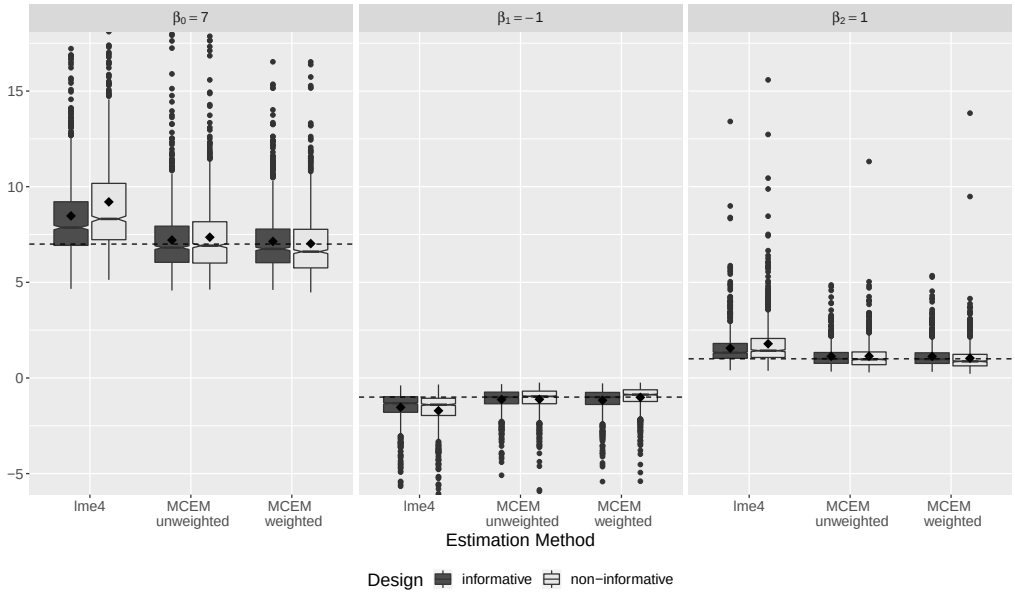
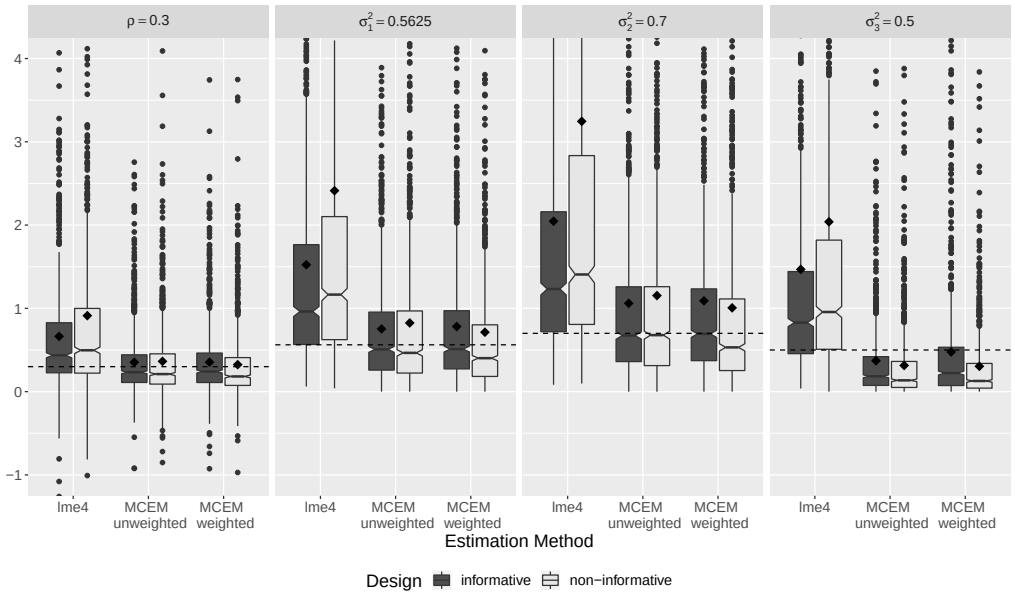


Figure 2.4: MC-Distribution of $\hat{\beta}$ under $h(Y_i^{(2)}) \sim N(0, 0.16)$



use of survey weights especially improves the estimation of σ_3^2 but has no impact to the estimation of ρ due to the reasons discussed above.

Figure 2.5: MC-Distribution of $\hat{\rho}$ under $h(Y_i^{(2)}) \sim N(0, 0.16)$



2.3 APPLICATION TO BUSINESS SURVEYS

The introduced GLMMs are applied to a real business Public Use Microdata (PUM) file and the question whether a firm's access to finance or credit/ loan approval depends on the gender of the top manager/ gender decomposition of the owners. Note that the inference based on the presented results has to be evaluated critically due to the difficulties in standard error estimation for the MCEM [Burgard and Dörr, 2018].

2.3.1 *Theoretical Background*

The role of the leading persons' gender for a firm's access to finance gained a lot of popularity in the 1990s and early 2000s with ambiguous results [Fay and Williams, 1993, Fabowale et al., 1995, Wilson et al., 2007, Carter et al., 2007]: Whilst Fay and Williams [1993] found a moderate impact of gender for less educated business owners on credit approvals, neither Fabowale et al. [1995] nor Wilson et al. [2007] do so. The diverging results may also be due to the different disciplines and respective methodologies involved in this research. Furthermore, psychological and sociological models also motivate qualitative research [Wilson et al., 2007].

When empirical data are used, sample sizes are often small and only hypothesis tests for equal means can be implemented, which is often done in the cited literature. Time period and location vary strongly between the empirical studies: Fabowale et al. [1995] studies Canadian businesses, Fay and Williams [1993] analyzes the situation in New Zealand and the studies of Wilson et al. [2007] and Carter et al. [2007] seem to rely on the same data gathered in the United Kingdom. The data set used here studies firms from Central Asia and Eastern Europe and includes several countries.

As often only means are statistically compared, the economic sector where the businesses act, may also play a major role [Fay and Williams, 1993, Rosa et al., 1996]. These differ across the studies but concentrate mainly on business service [Fay and Williams, 1993, Chell and Baines, 1998]. Here, manufacturing sectors as well as business services are considered and the economic sector is controlled for via a random intercept.

The access of finance may also translate to a firm's performance: Rosa et al. [1996] and Chell and Baines [1998] study the role of the owner's gender on small business performance and though their focus does not lie on 'access to finance' as a determinant, this can clearly be considered a factor on a firm's development. A rather similar performance of firms regardless the owners' gender might be a hint that the access to finance is not biased: Otherwise, women run businesses would possibly perform worse due to inadequate financial equipment.

Also for firm performance, the impact of gender is found to be ambiguous. Chell and Baines [1998] do not find a gender difference whilst Rosa et al. [1996], find gender differences in terms of sales turnover and capital assets – a gender gap in these performance measures that persists even when the firm's age is accounted for, which may also be a major variable [Rosa et al., 1996]. However, note that the model used was a simple linear regression model, ignoring the skewness of these data. Other results may be found using Box-Cox or Dual transformations like discussed previously.

Note that the cited studies focus mostly on small businesses. The data set which is available for our analysis has a survey cut-off level for firms with less than 5 employees and also includes larger firms. In that case, it is often the case that businesses are owned by more than one or two owners. For these reasons and because there exists only an indicator for the existence of female firm owners (and no decomposition of the ownership), we will rely on the gender of the firm's topmanager. Consequently, the data used here differs in terms of the sampling population (more economic sectors, larger firms, other countries and time periods) as well as in terms of available variables (gender of topmanager vs. owner), which might add new aspects to the open research question of the role of gender for a firm's access to the financial market.

2.3.2 *Data Description*

The PUM used is the enterprise survey of the World Bank[1], restricted to central Asian and Eastern European countries, i.e. the Business Environment and Enterprise Performance Surveyss (BEEPSs). These surveys are a joint project of the World Bank, the European Bank for Reconstruction and Development, the European Investment Bank, and the European Commission. Survey years included in the panel are 2002, 2005, 2007 and 2009. The survey design is stratified and ignores firms with less than 5 employees [World Bank, 2009]. The strata are cross-classifications of economic sectors, firm size and geographical location. Depending on the size of the concerned country's economy, some of the strata are collapsed.

After a data editing process that omitted missing observations in the dependent and explanatory variables, there remain only observations from 2009 from 22 different countries in 17 economic sectors. The data editing process is given in the electronic appendix (cf. Table B.1), the overview on model variables is given in Table 2.1.

As learned from the theoretical discussion in Section 2.1 and the preceeding simulation study, the design can be considered to be non-informative when the design variables are adequately considered in the regression analysis [Pfeffermann, 1993]. This is an advantage because the stratification variables are included in the PUM whilst survey weights are not available for every observation.

2.3.3 *Estimation and Evaluation*

We run four different regressions on the introduced data set. All models use the same explanatory variables but partly rely on different subsets of the PUM. The variables used and their abbrevations are listed in Table 2.1. The explanatories were chosen either to account for the design (economic sector and country) or because they were found to be indicators in previous studies (such as the firm's age [Rosa et al., 1996], size of firm [Riding and Swift, 1990], economic sectors) or because they were judged to be indicators for the economic performance (capacity utilization, share of exports shows the international competitivity, investment in research and development and growth of sales indicate the growth/ expansion tendency) or the firm's assessment on the financial market (are collaterals necessary?). Alternatively, if a collateral was asked for the most recent credit, this can give a hint to riskier activities that are less probable to be approved. In addition, Riding and Swift [1990] found that to women, higher collateral requirements are posed. Furthermore, we account for the top-manager's experience; if women have on aver-

1 https://www.enterprisesurveys.org/, 30 March 2020

age less years of experience, that would also be reason to refuse a loan or credit. Furthermore, Fay and Williams [1993] pointed out the role of education for loan granting, especially in interaction with gender. However, in a pre-analysis of the data, the interaction between both the collateral requirement and gender as well as experience and gender did not prove to be significant. An overview on the model variables and their abbreviations is given in Table 2.1. As mentioned previously, sur-

Table 2.1: Exogeneous Variables for Regression Models

Short Name	Description
finance_high_obstacle	Access to finance is considered to be a major or very severe obstacle to the current operations
pessimistic_approval	Firm did not apply for loan or grant because of negative approval was anticipated
loan_rejected	Firm applied for a loan or credit in the last fiscal year that was rejected
female_topmanager	Firm's topmanager is female
collateral_required	Financing required a collateral for the most recent loan or credit
b7	Topmanager's years of experience working in this sector
l1	Permanent, full-time employees end of last fiscal year
age_firm	Number of years since firm began operations
invest_rd	Firm invested last year directly or indirectly in research and development
sales_growth	Average growth rate of total annual sales over last two fiscal years (calculated from LCU)
f1	Firm's capacity utilization in last fiscal year (percent)
d3d	Share of direct and indirect exports to annual sales last fiscal year
a4b	Industry Sector (17 Sectors) (random intercept)
country	Firm's location (random intercept)

vey weights are not available for each observation in the PUM. Therefore we account for the survey design by the inclusion of the stratifiers 'num-

ber of employees', 'country' and 'economic sector'. As no weights are used, an alternative estimation procedure to 2.1 is the R-package `lme4`. However, note that in previous simulation studies Burgard and Dörr [2019] demonstrated shortcomings of the `glmer` command due to integral approximations. Therefore, we include both estimators – based on `lme4` and Algorithm 2.1 – in our analysis. This also allows us to get an idea about statistical significance, as the `lme4` estimators performed better than the expected Fisher information in Burgard and Dörr [2018] for non-informative designs. As Algorithm 2.1 yields a stochastic optimization, we run the algorithm 100 times on the PUM to evaluate the volatility of the point estimates. Here, the average of the MCEM results are presented in Table 2.2.

In general, we find that the estimates stemming from Algorithm 2.1 and `lme4` are similar. Most differences can be found for the intercept and the random effects variances, the latter was also found in the simulation study. Because of the overall agreement of the estimation methods on the fixed effects parameter, those statistically significant ($p < 0.01$:***, $p < 0.05$:**, $p < 0.1$:*) under the standard errors estimated by `lme4`, are marked. The correct estimation of standard errors under the MCEM algorithm is still future work.

Table 2.2: Point Estimates for the Logit Mixed Models

Explanatory	Significance under `lme4`	Avrg. MCEM Estimate	Estimate lme4
Dependent variable: `finance_high_obstacle`			
`(Intercept)`		-0.40	-0.55
`age_firm`		-0.00	0.15
`b7`		-0.00	0.42
`collateral_requiredTRUE`		0.44	-0.00
`d3d`		-0.00	-0.00
`f1`	***	-0.01	1.09
`female_topmanagerTRUE`		0.16	-0.00
`invest_rdTRUE`		0.27	0.25
`l1`		-0.00	0.05
`loan_rejectedTRUE`	***	1.08	-0.01
`sales_growth`		0.05	-0.00
`a4b` (RE, σ^2_{a4b})		0.0031	0.1573
`country` (RE, $\sigma^2_{country}$)		0.0105	0.0000
Dependent variable: `finance_high_obstacle`			

Table 2.2: Point Estimates for the Logit Mixed Models

Explanatory	Significance under lme4	Avrg. MCEM Estimate	Estimate lme4
(Intercept)		0.03	-0.04
age_firm		-0.00	0.05
b7		-0.01	0.49
collateral_requiredTRUE	**	0.54	-0.00
d3d		-0.00	-0.00
f1	***	-0.02	-0.00
female_topmanagerTRUE		0.05	0.20
invest_rdTRUE		0.20	0.03
l1	*	-0.00	-0.02
sales_growth		0.03	-0.00
a4b (RE, σ^2_{a4b})		0.0022	0.1081
country (RE, $\sigma^2_{country}$)		0.0099	0.0000
Dependent variable: pessimistic_approval			
(Intercept)		1.83	2.56
age_firm		0.00	0.27
b7		-0.08	1.24
collateral_requiredTRUE		0.61	-0.11
d3d		-0.02	-0.11
f1		-0.05	0.00
female_topmanagerTRUE		0.64	-0.08
finance_high_obstacleTRUE		-0.81	-0.08
invest_rdTRUE		-0.33	-0.06
l1		-0.09	-0.03
sales_growth		-0.10	-1.18
a4b (RE, σ^2_{a4b})		0.0101	1.8058
country (RE, $\sigma^2_{country}$)		0.0100	0.1052
Dependent variable: loan_rejected			
(Intercept)	***	-1.59	-1.86
age_firm		-0.00	-0.09
b7		-0.02	0.58
collateral_requiredTRUE		0.61	-0.01
d3d		-0.00	-0.00

Table 2.2: Point Estimates for the Logit Mixed Models

Explanatory	Significance under `lme4`	Avrg. MCEM Estimate	Estimate lme4
`f1`		-0.01	-0.00
`female_topmanagerTRUE`		0.03	0.22
`invest_rdTRUE`		0.27	-0.06
`l1`		-0.00	-0.01
`sales_growth`		-0.07	-0.00
`a4b` (RE, σ^2_{a4b})		0.0049	0.3812
`country` (RE, $\sigma^2_{country}$)		0.0126	0.0000

A look on the random effects variance for all four regression models reveals that the role of the economic sector for the firm's situation on the financial market, though, seems to be overstated in contrast to the previous studies – the random effects variance is often very close to zero. This might be due to the fact, that different economic sectors differ mostly in the operating firms' structure, which is already accounted for by the fixed effects for firm size and age. Developing economic sectors – and declining economic sectors as well – may already be summarized in a firm's average growth of total sales or capacity utilization. In the following, we discuss the results for the fixed effects.

To start with, prior to a model for the probability of credit approval, we estimated a model to control whether female business leaders consider access to finance generally more problematic than male. This idea is due to the finding in Fabowale et al. [1995] that women do not seem to be discriminated by loan officers but sometimes feel inappropriately treated by loan officers. This model is estimated twice, first using a subset of firms that applied for a loan in the last fiscal year. There, it is found that an indicator for the economic situation of the firm – capacity utilizatio – is a significant variable for the perception of financial access. Firms that are economically healthier tend to look at access to finance less critically. Also female run businesses perceive access to finance more difficult though the result is – according to `lme4` – not significant. Note however, that the size of the effect and the sign depends on the estimation algorithm.

Furthermore, firms that were not granted the loan perceive access to finance significantly (in terms of `lme4`) more problematic. As this observation is very intuitive but might bias the firms perception, a control model is estimated leaving out this variable and therefor the data set becomes a larger subset of firms, including those that did not apply for a loan or credit. In that control regression, other factors that give hints on

the economic situation of the firm such as number of full-time employees and the posed requirement for collaterals become relevant. When a firm is asked for a collateral in order to receive a loan, the financial institution seems to evaluate the credit as risky due to the firm's economic situation. Though the sign of the parameter for the gender has the expected sign, we can neither in the first nor in the control regression confirm that female top-managers perceive access to finance more difficult than males (note however, that it was not always the top-manager that responded to the survey).

Next, Wilson et al. [2007] pointed out that female business leaders may possibly feel discouraged to even apply for a credit or loan, so the third regression model checks whether women led businesses are less optimistic to apply for loans and therefore did not even enter the PUM subset of the first regression. Again, the slope parameter for the top-managers has the expected sign (in contrast to the `lme4` result) but does not seem to be significant. Neither are other explanatory variables. Nonetheless, the sign of the gender variable changes comparing to `lme4` and even the size of the effect enlarges. This is in accordance with the first and second regression where women led businesses also considered access to finance more often an obstacle, though not significantly. This requires further analysis in future social science research.

Finally, we do not find that the rejection of a loan application is significantly more probable when the topmanager is female (cf. Table 2.2), and it is noteworthy that the MCEM algorithm yields a much smaller value than `lme4` for the parameter which indicates that further inspection in future research may be necessary. Note that the reduced effect using MCEM is consistent with the third regression and indicates that the hypothesis of Wilson et al. [2007] seems plausible – if women do not apply for credits because they are pessimistic, it does not matter that they are finally not discriminated. So the fourth regression in combination with the third one doubts the `lme4` results.

We conclude our real data study thus without finding a clear significant discrimination of women business leader in the financial sector, adding therefore to the evidence of Riding and Swift [1990] and Fabowale et al. [1995] also for larger businesses and in other countries. However, the diverging results for the top-manager variable between `lme4` and MCEM in size and even in the estimate's sign are in accordance with the hypothesis in Wilson et al. [2007] and indicate that still further research, for example in reliable MCEM standard error estimators, is needed for a more profound assessment.

3

GENERALIZED VARIANCE FUNCTIONS FOR VARIANCE ESTIMATION IN BUSINESS SURVEYS

3.1 INTRODUCTION TO VARIANCE ESTIMATION

Holland [1986] differentiates between causal and associative inference in statistics. Whilst the preceding chapter dealt with the relation between variables, where one variable Y_i was considered to be the response of the 'exposure' to variables Z_i (and X_i), this chapter takes a look on associative statistics in the terms of Holland [1986], on a finite population, variables are purely descriptive and randomness enters the framework through random subsetting. In the terminology of this work, we deal here thus with concepts of design-based statistics.

Though correct point estimation like it was discussed in the previous chapter is essential to infer correctly from a complex sample, the estimation of an estimator's second moments is essential, too: Without information on an estimator's expected squared deviation, conclusions about the distance between the statistic of interest and its estimator g_S, for example in terms of confidence intervals or hypotheses testing, are not feasible. In causal inference, this squared deviation is used to assess the relevance of the exposure to control variables. In associative statistics, on the other hand, second moments of g_S capture the insecurity about finite population characteristics when only parts thereof are observed. Second moments, though, require knowledge of the unobserved characteristics to assess the variability of g_S and are therefore also unobservable in practice. Consequently, estimators for second order statistics become necessary.

In large scale surveys, however, the amount of variables or the construction of nonlinear indicators out of linear estimators (meaning that the statistic g_U is not linear in its argument) challenges the estimation of second moments of an estimator g_S for g_U. This is especially true when the survey design P_D is complex. When a survey's outcome must be published [German Federal Statistical Office, 2009, for example] in terms of point and variance estimate of summary statistics, the mass of variables can then lead to a computational burden due to the possible complexity of P_D or g_S.

Sometimes, a combination of these two complexities can be reduced to a problem of the former thanks to linearization methods for g_S. In the

following, we study therefore first linear estimators g_S. Thus, assume a complex design P_D and a linear estimator g_S in $\mathbf{y}$, i.e.

$$g_S : \mathcal{S} \times \Omega \to \mathbb{R}^q, \quad g_S(s, \mathbf{y}, \mathbf{z}) = \sum_{i \in U} w_i \cdot \mathbb{1}_s(i) \cdot \mathbf{a}_i^T \cdot \mathbf{y}_i$$

with $\mathbf{a} := (\mathbf{a}_i : i \in U)$, $\mathbf{a}_i \in \mathbb{R}^{q \times p}$. Without any restriction let $\mathbf{a}_i \equiv I_p$, that is, consider the HT estimator (1.9) for the p variables of interest, $g_S = \hat{\boldsymbol{\tau}}_y$. We use the fact that P_D can also be described by the moments of $\mathbf{1}_S = (\mathbb{1}_S(i) : i \in U)$. We get consequently with $w_i = \pi_i^{-1}$, $\mathbf{w} = (w_i : i \in U)$, for the variance of $\hat{\boldsymbol{\tau}}_y$ [Haziza et al., 2008]

$$\begin{aligned} \mathrm{Var}_D\left[\hat{\boldsymbol{\tau}}_y\right] &= \mathrm{Var}_D\left[\sum_{i \in U} \frac{\mathbb{1}_S(i)}{\pi_i} \mathbf{y}_i\right] \\ &= \mathbf{y}^T W \mathrm{Var}_D\left[\mathbf{1}_S\right] W \mathbf{y}, \quad W := \mathrm{diag}(\mathbf{w}) \\ &= \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \cdot \pi_j\right) \frac{1}{\pi_i} \frac{1}{\pi_j} \mathbf{y}_i \mathbf{y}_j^T \end{aligned} \tag{3.1}$$

And for the second moment of $\mathbf{1}_S$, we have

$$\mathrm{Var}_D\left[\mathbf{1}_S\right] = \mathrm{E}_D\left[\mathbf{1}_S \mathbf{1}_S^T\right] - \mathrm{E}_D\left[\mathbf{1}_S\right] \mathrm{E}_D\left[\mathbf{1}_S\right]^T =: \Pi \tag{3.2a}$$

$$\begin{aligned} [\Pi]_{ij} &= P_D(i \in S \wedge j \in S) - P_D(i \in S) \cdot P_D(j \in S) \\ &= \pi_{ij} - \pi_i \cdot \pi_j \end{aligned} \tag{3.2b}$$

Equation (3.1) encompasses three problems: First, $\mathbf{y}$ is only observed for those units that enter the sample, i.e. only $\mathbf{y}_S$ is known to the researcher. Second, π_{ij}, that is the second order inclusion probability (cf. Section 1.3), can be difficult to compute or difficult to store electronically for $N \gg 0$. Third, even if the inclusion probabilites were known, they would often be published only for $i \in s$ with sample realization $S = s$. This makes the development of variance estimators necessary. Variance estimators that solve the problem of unknown inclusion probabilitys for units in $U \setminus S$ are the HT variance estimator

$$\widehat{\mathrm{Var}}_D\left[\hat{\boldsymbol{\tau}}_y\right] = \sum_{i \in U} \sum_{j \in U} \mathbb{1}_S(i) \cdot \mathbb{1}_S(j) \cdot \left(\frac{\pi_{ij} - \pi_i \cdot \pi_j}{\pi_{ij}}\right) \frac{1}{\pi_i} \frac{1}{\pi_j} \mathbf{y}_i \mathbf{y}_j^T \tag{3.3a}$$

and for $|S| = \mathrm{E}_D[|S|]$ the Sen-Yates-Grundy variance estimator

$$\begin{aligned} \widehat{\mathrm{Var}}_D\left[\hat{\boldsymbol{\tau}}_y\right] = -\frac{1}{2} \sum_{i \in U} \sum_{j \in U} \mathbb{1}_S(i) \cdot \mathbb{1}_S(j) \cdot \left(\frac{\pi_{ij} - \pi_i \cdot \pi_j}{\pi_{ij}}\right) \\ \cdot \left(\frac{1}{\pi_i} \mathbf{y}_i - \frac{1}{\pi_j} \mathbf{y}_j\right) \left(\frac{1}{\pi_i} \mathbf{y}_i - \frac{1}{\pi_j} \mathbf{y}_j\right)^T \end{aligned} \tag{3.3b}$$

[Haziza et al., 2008]. Though the Estimators 3.3 only rely on observed information, both Equations 3.3a and 3.3b still require second order inclusion probabilities, which are not always known. Matei and Tillé [2005] and Haziza et al. [2008] present an overview of existing methods to approximate π_{ij} through π_i and π_j and mention the conditions on P_D under which these approximations hold: Amongst others the design must be pps and

$$\sum_{i \in U_N} \pi_i \cdot (1 - \pi_i) \xrightarrow[N \to \infty]{} \infty \tag{3.4}$$

and P_D must be a 'high entropy design' [Haziza et al., 2008], confer Definitions 4 and 5.

Concerning the nonlinearity of some estimators g_S, amongst others Woodruff [1971] suggests to linearize the estimator using a first order Taylor approximation. For not twice continuously differentiable g_S, generalizations like the Gâteaux-Differential in the influence function approach may be used [Hampel, 1974, Deville, 1999]. However, these techniques require that the variance estimators for linear statistics are easily copmuted under P_D. If this is not the case, other methods must first be used to get (at least approximate) variance estimators for linear linear statistics. Amongst these methods are resampling techniques and GVFs. The most popular resampling techniques are the jackknife method Quenouille [1956] and variations thereof as well as the bootstrap [Efron, 1979].

The basic idea of the jackknife is to measure the impact of each random variable realization $\mathbb{1}_S(i) \cdot \mathbf{y}_i$ entering the estimator g_S by the 'residual' difference $g_S(\mathbf{y}_S) - g_S\left((\mathbf{y}_j : j \in S \wedge j \neq i)\right)$ where the second term is a version of the estimator g_S that uses one sample unit less. The average squared impact is then used as a variance estimator [Efron and Stein, 1981, Shao and Wu, 1989]. Like for the original bootstrap, the original assumption was that P_D is a Simple Random Sampling (SRS) without replacement because both ideas stem from model-based statistics. Defining a probability measure $P_y := \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{\mathbf{y}_i\}}$, though, allows to transfer the model-based origins to a more design-based view where the statistic of interest is considered to be a function of P_y, $g_U(P_y)$ and an estimator g_S is considered to be a function of an estimator $\hat{P}_y$, i.e. $g_S = g_U(\hat{P}_y)$. Though more elaborated extensions to design-based statistics exist, there are also counter examples for which the classical delete-1-jackknife is not consistent under more complex survey designs or for non-smooth statistics g_U [Shao and Wu, 1989].

For the bootstrap, the theoretical principle is that for $i = 1, \ldots, n$ realizations $Y_i \sim_{iid} F$, the distributions of

$$g_S(Y), \quad Y_i \sim_{iid} F$$

and

$$g_S(Y^{(b)}), \quad Y_i^{(b)} \sim_{iid} \hat{F}_n \quad ,$$

where $\hat{F}_n$ is the empirical cdf, should behave similarly. Note that the formulation again dates back to the origins of the bootstrap in model-based statistics. As $\hat{F}$ is known for a sample $\mathbf{y}$, the properties of $g_S(\mathbf{y}^{(b)})$ can either be studied analytically or through MC methods with $b = 1, \ldots, B$ and $B \gg 0$. Again, the SRS assumption is problematic for survey sampling and extensions are necessary [Preston, 2009, for example] because the random variable in survey statistics is S and from one realization s of S, P_D cannot be restablished as the design variables or variables of interest are unknown for $U \setminus S$. Survey bootstrap are therefore an approximation to the survey design P_D. For example, Preston [2009] might be interpreted to mimic the sampling design at least up to the first (and possibly second order) moments by simple sampling rules on the alternative population s and is thus very mechanical.

A third alternative besides linearization and resampling – that can also be combined with the former – are Generalized Variance Functions (GVFs). Assuming a functional relationship between an estimator's expectation and its variance allows to predict variances given only a few tuples of point and variance estimates for the estimation of the functional relation. The theoretical motivation lying behind this assumption, some analytical results concerning the statistical properties of GVFs and extensive simulation studies are presented in the subsequent sections.

3.2 THEORETICAL BACKGROUND OF THE GVF

3.2.1 *A Model-based Motivation of the GVF*

Early works on GVFs date back to [Wolter, 1985, Chapter 5] and one of the first applications was the Current Population Survey in the US [Valliant, 1987]. Basically, the idea of GVFs is to publish predicted squared sampling errors based on a model that relates the sample estimator to its variance. The use of GVFs therefore leads to a shift from estimation to prediction and the prediction is based on a model $\mathcal{M}_\theta$ that is assumed to underly the (survey) estimator's input data. As the input data is a subset of the finite population, this means that $\mathcal{M}_\theta$ is assumed to have generated the finite population. Estimating the model, statements on the unobserved part of the population are therefore possbile (cf. Definition 8). In the context of GVFs, these statements concern the sampling variance.

Under some DGPs, there is a relation between a random variable's expectation and its variance. For example,

$$Y \sim \text{Bin}(n,p) \Rightarrow \text{E}_{\mathcal{M}}[Y] = np \text{ and}$$

$$\text{Var}_{\mathcal{M}}[Y] = \text{E}_{\mathcal{M}}[Y] - \frac{1}{n}\,\text{E}_{\mathcal{M}}[Y]^2 \tag{3.5}$$

$$Y \sim \text{Pois}(\lambda) \Rightarrow \text{E}_{\mathcal{M}}[Y] = \lambda \text{ and } \text{Var}_{\mathcal{M}}[Y] = \text{E}_{\mathcal{M}}[Y] \tag{3.6}$$

$$Y \sim \mathcal{G}(\alpha,\beta) \Rightarrow \text{E}_{\mathcal{M}}[Y] = \alpha\beta \text{ and } \text{Var}_{\mathcal{M}}[Y] = \beta \cdot \text{E}_{\mathcal{M}}[Y] \quad . \tag{3.7}$$

Let thus be all characteristics $\mathbf{y}_i$, $i \in U$, outcomes of such random variables Y_i that are iid under $P_{\mathcal{M}_\theta}$ with parameter vector θ. Define the relation between $\text{E}_{\mathcal{M}_\theta}[Y_i]$ and $\text{Var}_{\mathcal{M}_\theta}[Y_i]$ by $\tilde{f}$:

$$\tilde{f}_y \to \mathbb{R}^{p\times p}; \quad \tilde{f}\left(\text{E}_{\mathcal{M}_\theta}[Y_i]\,;\theta\right) = \text{Var}_{\mathcal{M}_\theta}[Y_i]$$

We define an additional parameter vector $\boldsymbol{\psi}$ that includes θ and possibly parameters of the survey design P_D or the finite population (such as N). We have then for the HT estimator - under a non-informative design and using Remark 5 and the variance of the HT (3.1) -

$$\text{E}_{\mathcal{M}_\theta,D}\left[\sum_{i\in U} w_i \cdot \mathbb{1}_S(i)\cdot Y_i\right] = \text{E}_{\mathcal{M}_\theta}\underbrace{\left[\sum_{i=1}^{N} Y_i\right]}_{=\text{E}_D[\hat{\tau}_y]} = N\text{E}_{\mathcal{M}_\theta}[Y_1] \tag{3.8a}$$

$$\begin{aligned}
\text{Var}_{\mathcal{M}_\theta,D}\left[\sum_{i\in U} w_i \cdot \mathbb{1}_S(i)\cdot Y_i\right] &= \text{Var}_{\mathcal{M}_\theta}\left[\sum_{i=1}^{N} Y_i\right] \\
&\quad + \sum_{i\in U}\sum_{j\in U}\frac{\pi_{ij}-\pi_i\cdot\pi_j}{\pi_i\cdot\pi_j}\text{E}_{\mathcal{M}_\theta}\left[Y_i\cdot Y_j^T\right] \\
&= N\cdot\tilde{f}\left(\text{E}_{\mathcal{M}_\theta}[Y_1];\theta\right) + \sum_{i\in U}\frac{1}{\pi_i}\cdot\tilde{f}\left(\text{E}_{\mathcal{M}_\theta}[Y_1]\,;\theta\right) \\
&\quad + \text{E}_{\mathcal{M}_\theta}[Y_1]\cdot\text{E}_{\mathcal{M}_\theta}\left[Y_1^T\right]\cdot\sum_{i\in U}\sum_{j\in U}\frac{\pi_{ij}-\pi_i\cdot\pi_j}{\pi_i\cdot\pi_j} \\
&=: f\left(\text{E}_{\mathcal{M}_\theta}[Y_1]\,;\boldsymbol{\psi}\right)
\end{aligned} \tag{3.8b}$$

Given that the assumed model is true, it is thus possible to establish a relation between the HT's expectation and its variance when the design is known up to the second moments. Note that the design P_D thus also plays a role for the GVF, which is often ignored in the application.

Example 1. *Let*

$$Y_i \sim_{iid} \text{Pois}(\lambda)$$

and

$$P_D(s) := \begin{cases} \frac{1}{\binom{N}{n}}, & \textit{if } |s| = n \\ 0, & \textit{else} \end{cases}$$

that is a SRS design without replacement. Hence, $\mathrm{E}_{\mathcal{M}_\theta}[Y_1] = \mathrm{Var}_{\mathcal{M}_\theta}[Y_1] = \lambda$ *and we have therefore* $\pi_i \equiv \frac{n}{N}$ *and* $\pi_{ij} \equiv \frac{n^2-n}{N^2-N}$ *for the HT*

$$\mathrm{Var}_{\mathcal{M}_\theta,\mathrm{D}}\left[\hat{\tau}_y\right] = N\lambda + \frac{N^2}{n}\lambda - \frac{N-n}{(N-1)n}\cdot\lambda^2 \quad .$$

Note that the relationship is thus quadratic although the underlying relationship between expectation and variance is linear and the finite population estimator is linear.

From (3.8b) we can nonetheless conclude that if the survey design is such that $\pi_{ij} - \pi_i\pi_j \to_{N\to\infty} 0$ the last term approaches zero and in that case, $\mathrm{Var}_{\mathcal{M}_\theta,\mathrm{D}}\left[\hat{\boldsymbol{\tau}}_y\right]$ approximates a linear relationship to $\tilde{f}\left(\mathrm{E}_{\mathcal{M}_\theta}[Y_1];\theta\right)$.

Remark 6. *Assume that* $g_S^{(j)}$, $j \in \mathcal{Q}$ *are unbiased estimators for the population total and that for* $Y_i \sim_{\mathrm{iid}} P_{\mathcal{M}_\theta}$ *it holds that* $\mathrm{Var}_{\mathcal{M}_\theta}[Y_1] = \tilde{f}\left(\mathrm{E}_{\mathcal{M}_\theta}[Y_1];\theta\right)$. *When the survey design* P_D *is SRS without replacement and fixed sample size, the GVF* f *in* (3.8b) *simplifies to*

$$\begin{aligned}\mathrm{Var}_{\mathcal{M}_\theta,\mathrm{D}}\left[g_S^{(j)}(S,Y)\right] = &\left(N + \frac{N^2}{n}\right)\cdot\tilde{f}\left(\mathrm{E}_{\mathcal{M}_\theta}[Y_1];\theta\right)\\ &- \frac{N-n}{(N-1)n}\mathrm{E}_{\mathcal{M}_\theta}[Y_1]\cdot\mathrm{E}_{\mathcal{M}_\theta}\left[Y_1^T\right] \quad .\end{aligned}$$

Similar results hold for the GREG (1.10): For a slope matrix B, we set

$$\sum_{i\in U}\underbrace{\mathbf{y}_i - \mathbf{z}_i^T B}_{=:\mathbf{e}_i} \quad .$$

Note that B is in this context not the parameter $(\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T\mathbf{y}$ from (1.10) but a model parameter of θ and the finite population version is an estimator thereof. If $\mathbf{e}_i$ is interpreted as the realization of an iid zero centered L^2-integrable random variable ε_i ($\mathrm{Var}_{\mathcal{M}_\theta}[\varepsilon_i] =: \Sigma$ for all $i \in U$), we get (given the realization $\mathbf{z} = (\mathbf{z}_i : i \in U)$ of Z)

$$\begin{aligned}\mathrm{Var}_{\mathcal{M}_\theta,\mathrm{D}}\left[\hat{\boldsymbol{\tau}}_y^{\mathrm{GREG}}\right] &= \mathrm{Var}_{\mathcal{M}_\theta,\mathrm{D}}\left[\sum_{i\in U} w_i\cdot\mathbb{1}_S(i)\cdot\varepsilon_i\right]\\ &= \mathrm{Var}_{\mathcal{M}_\theta}\left[\mathrm{E}_{\mathrm{D}}\left[\sum_{i\in U} w_i\cdot\mathbb{1}_S(i)\cdot\varepsilon_i\right]\right]\\ &\quad + \mathrm{E}_{\mathcal{M}_\theta}\left[\mathrm{Var}_{\mathrm{D}}\left[\sum_{i\in U} w_i\cdot\mathbb{1}_S(i)\cdot\varepsilon_i\right]\right] \qquad (3.9a)\\ &= N\Sigma + \left(\sum_{i\in U}\frac{1}{\pi_i}\right)\Sigma =: f\left(\mathrm{E}_{\mathcal{M}_\theta}[Y_1];\boldsymbol{\psi}\right) \qquad (3.9b)\end{aligned}$$

which is a constant function of $E_{\mathcal{M}_\theta}[Y_1]$ with the constant depending on P_D. Note that using $\hat{B} = (\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T\mathbf{y}$ yields $\sum_{i\in U}\mathbf{e}_i = 0$ regardless whether there is a linear relationship between Y and Z. Therefore, the design unbiasedness would hold using $\hat{B}$ instead of B and the first term in (3.9a) would still equal zero. However, if the model assumption is violated (i.e. $E_{\mathcal{M}_\theta}[\varepsilon_i] \neq 0$, using $\hat{B}$ does not equal anymore $\Sigma\sum_{i\in U}\frac{1}{\pi_i}$ but $E_{\mathcal{M}_\theta}[\varepsilon_1]\cdot E_{\mathcal{M}_\theta}[\varepsilon_1]^T\sum_{i,j\in U}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j}$.

The GREG has therefore one advantage and one inconvenient if the model is true: Then, only first order inclusion probabilities are required; which is handy in application. However, in that case the expectation of Y (and therefore also of $\hat{\tau}_y^{GREG}$) has no predictive power on the variance, which contradicts the idea of GVF usage.

When we consider in the following estimators g_S of the population total that are (at least approximately) unbiased, we can plug $N^{-1}E_{\mathcal{M}_\theta,D}[g_S(S,Y,Z)]$ into (3.8b) and get therefore a (approximate) relation between $\mathrm{Var}_{\mathcal{M}_\theta,D}[g_S(S,Y,Z)]$ and $E_{\mathcal{M}_\theta,D}[g_S(S,Y,Z)]$.

The problem with relation f is therefore first, that it depends on P_D and second, relates to $\tilde{f}$ but is not of the same type as shown in Example 1 for medium size finite populations or unfavorable survey designs. Furthermore, not even $\tilde{f}$ may be known. Therefore, ψ is estimated and a relation f is simply assumed to hold. For estimation, different variables of interest, $Y_i^{(j)}$, $i \in U$ and $j \in \mathcal{Q}$ are required such that for some $j \in \tilde{\mathcal{Q}} \subset \mathcal{Q}$, $E_{\mathcal{M}_\theta}[g_S(S,Y,Z)]$ and the variance $\mathrm{Var}_{\mathcal{M}_\theta}[g_S(S,Y,Z)]$ are known. In practice, these values are unknown neither and estimators are plugged into the estimation method for ψ.

In the following, we shall focus on a general family of estimators $\left\{g_S^{(j)}\right\}_{j\in\mathcal{Q}}$ that are approximately unbiased for the population total in order to justify GVFs like above. Though auxiliary variables Z may be used in $g_S^{(j)}$, $j \in \mathcal{Q}$, we omit the auxiliary variables in the following for better readability. Obviously, variables $\{Y_S^{(j)}\}_{j\in\mathcal{Q}}$ should originate from the same model $\mathcal{M}_\theta$ and underlie the same subsetting process P_D in order to get good estimators for ψ.

The idea of GVFs is then that for a set of similar estimators $g_S^{(j)}$, $j \in \mathcal{Q}$, it is possible to determine the parameter vector ψ, and to avoid direct calculus of $\mathrm{Var}_{\mathcal{M}_\theta,D}\left[g_S^{(\iota)}(S,Y^{(\iota)})\right]$ through the use of the prediction

$$\widetilde{\mathrm{Var}}^{\star}_{\mathcal{M}_\theta,D}\left[g_S^{(\iota)}(S,Y^{(\iota)})\right] = f\left(E_{\mathcal{M}_\theta,D}\left[g_S^{(\iota)}(S,Y^{(\iota)})\right];\psi\right) \tag{3.10}$$

for an index $\iota \in \mathcal{Q}$. If f and ψ were correct and $E_{\mathcal{M}_\theta,D}\left[g_S^{(\iota)}(S,Y^{(\iota)})\right]$ known, we would have $\widetilde{\mathrm{Var}}^{\star}_{\mathcal{M}_\theta,D}\left[g_S^{(\iota)}(S,Y^{(\iota)})\right] = \mathrm{Var}_{\mathcal{M}_\theta,D}\left[g_S^{(\iota)}(S,Y^{(\iota)})\right]$.

This is the optimal case. In summary, Equation (3.10), though, includes four main difficulties: First, the integrals are with respect to $P_{\mathcal{M}_\theta,D}$, implying that not only the estimators $g_S^{(\iota)}$ must be similar (i.e. approximately unbiased for the population total) but also the random variables $Y^{(\iota)}$ must follow the same probability law. Second, the relationship f is not necessarily known to the data analyst. National Statistical Institutes (NSIs) and other applicants of GVFs, though, often recur to very general families $\mathcal{F}$ that encompass many relations f stemming from common distribution functions (for example, confer Valliant et al. [2000, Chapter 10.3] or Di Consiglio et al. [2013]). Third, the equality in (3.8a) and (3.8b) only holds for a given design P_D. Fourth, for all $j \in \mathcal{Q}$, neither $E_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]$ nor $\mathrm{Var}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]$ are usually known which entroubles the estimation of $\boldsymbol{\psi}$. Given these problems, one can formulate conditions on P_D and $P_{\mathcal{M}_\theta}$ under which GVFs that are implementable in practice are consistent. Valliant [1987] gives such conditions for stratified two-stage designs. These conditions include amongst others a certain limit behaviour of P_{D_N} under growing populations U_N (so that the last term in the last but one line of (3.8b) approaches zero), design non-informativity, $P_{\mathcal{M}_\theta,D_N}$-unbiasedness of $g_S^{(j)}$ for all $j \in \mathcal{Q}$ and the correct choice of f. Note that the decomposition in (3.8b) is more general: If it is possible to estimate $\boldsymbol{\psi}$ consistently, which is the case when f is a homeomorphism in θ and $E_{\mathcal{M}_\theta}[Y_1]$ and known, then predictions

$$\widetilde{\mathrm{Var}}_{\mathcal{M}_\theta,D}\left[g_S(S,Y,Z)\right] := f\left(g_S(S,Y,Z); \hat{\boldsymbol{\psi}}\right)$$

should be consistent, too. Given that f is chosen correctly by the data analyst, we conduct in the following an error decomposition of realistic implementations of GVFs similar to Cho et al. [2002].

3.2.2 *Error Decomposition of the GVF Prediction*

First, note that $g_S^{(j)}$ needs to be a model-design unbiased estimator such that the exchange between $\mathrm{NE}_{\mathcal{M}_\theta}\left[Y_1^{(j)}\right]$ and $E_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]$ like in (3.8b) is valid. Assume furthermore that it is a $P_{\mathcal{M}_\theta,D}$-consistent estimator. As both, expectation and variance of $g_S^{(j)}$ are usually not even known for a subset $j \in \tilde{\mathcal{Q}} \subset \mathcal{Q}$, a plug-in estimator for the estimation of $\boldsymbol{\psi}$ is usually used. That is, a variance estimator $\widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]$ and $g_S^{(j)}\left(S, Y^{(j)}\right)$ replace the true variance and expectation and yield the estimator

$$\hat{\boldsymbol{\psi}} \triangleq \underset{\tilde{\boldsymbol{\psi}}\in\Psi}{\arg\min} \sum_{j\in\tilde{\mathcal{Q}}} \ell_j\left(\widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right], f\left(g_S^{(j)}\left(S, Y^{(j)}\right); \tilde{\boldsymbol{\psi}}\right)\right), \tag{3.11}$$

where Ψ is the parameter space of ψ, ℓ a loss function and ξ_j a random variable that accounts for the error stemming from the replacement of the expectation of $g_S^{(j)}\left(S, Y^{(j)}\right)$ by its estimator. In fact, we get

$$\xi_j \triangleq f\left(E_{\mathcal{M}_\theta, D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]; \psi\right) - f\left(g_S^{(j)}\left(S, Y^{(j)}\right); \psi\right) \quad . \tag{3.12}$$

If f is linear, we conclude thus that ξ_j is zero centered and its second moment exists if that of $g_S^{(j)}\left(S, Y^{(j)}\right)$ exists (which is assured if $\mathrm{Var}_{\mathcal{M}_\theta}\left[Y^{(j)}\right]$ exists). If f is nonlinear but continuous, $f\left(g_S^{(j)}\left(S, Y^{(j)}\right); \psi\right)$ is a consistent but not necessarily unbiased estimator. In that case, we can only state that $\xi_j \xrightarrow[N\to\infty]{P_{\mathcal{M}_\theta, D^N}} 0$.

In a next step, it must be analysed to what extent the replacement of the 'true' variance with a variance estimator affects the estimation of ψ and thus finally the prediction. Remark 5 states that in expectation, $\mathrm{Var}_D\left[g_S(S, Y, Z)\right]$ is smaller than $\mathrm{Var}_{\mathcal{M}_\theta, D}\left[g_S^{(j)}(S, Y^{(j)}, Z)\right]$. This means that the Estimator (3.11) yields a bias for ψ with respect to the estimator resulting from

$$\hat{\psi}^\star \triangleq \underset{\tilde{\psi}\in\Psi}{\arg\min} \sum_{j\in\tilde{Q}} \ell_j\left(\mathrm{Var}_{\mathcal{M}_\theta, D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right], f\left(g_S^{(j)}\left(S, Y^{(j)}\right); \tilde{\psi}\right)\right) \tag{3.13}$$

because $\widehat{\mathrm{Var}}_D\left[g_S(j)\left(S, Y^{(j)}\right)\right]$ may be seen as an (usually consistent) estimator for $E_{\mathcal{M}_\theta}\left[\mathrm{Var}_D\left[g_S(j)\left(S, Y^{(j)}\right)\right]\right]$ but does not account for the first term in (2.3). However, note that the first term in (3.8b), which corresponds to the first term in (2.3), becomes negligible if the design is such that $\sum_{i\in U} \frac{1}{\pi_i}$ grows faster than linearly and again $\pi_{ij} - \pi\pi_j \to 0$ and therefore, the estimator $\hat{\psi}$ gets better with increasing N and n.

These studies demonstrate that the realistic GVF predictor

$$\widetilde{\mathrm{Var}}_{\mathcal{M}_\theta, D}\left[g_S^{(j)}(S, Y^{(j)})\right] = f\left(g_S^{(\iota)}(S, Y^{(j)}); \hat{\psi}\right) \tag{3.14}$$

induces several errors into the prediction process. We suggest the following decomposition:

$$\begin{aligned}
&\mathrm{Var}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}(S,Y^{(j)})\right] - \widetilde{\mathrm{Var}}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}(S,Y^{(j)})\right] = \\
&\quad \underbrace{\left(f\left(E_{\mathcal{M}_\theta,D}\left[g_S^{(j)}(S,Y^{(j)})\right];\psi\right) - f\left(g_S^{(j)}(S,Y^{(j)});\psi\right)\right)}_{=:\xi_j} \\
&\quad + \underbrace{\left(f\left(g_S^{(j)}(S,Y^{(j)});\psi\right) - f\left(g_S^{(j)}(S,Y^{(j)});\hat{\psi}^\star\right)\right)}_{=:A_j} \\
&\quad + \underbrace{\left(f\left(g_S^{(j)}(S,Y^{(j)});\hat{\psi}^\star\right) - f\left(g_S^{(j)}(S,Y^{(j)});\hat{\psi}\right)\right)}_{=:B_j} \quad . \qquad (3.15)
\end{aligned}$$

The term A_j is the error that stems from the estimation process of ψ when the correct model-design variance (or an unbiased estimator thereof) was used. Under the conditions under which the loss minimization (3.13) yields a consistent estimator (that is especially an increasing training size $|\tilde{Q}|$), this term should go to zero in probablitiy, too. Hence, the only possible bias – under the correct choice of f – is term B_j, which stems from using $\widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]$ instead of $\mathrm{Var}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]$ due to the underestimation of the model-design variance by the design estimator. Note thus, that we require for consistency of GVF predictions not only the consistency of g_S (for the reduction of ξ_j in probability) and its variance estimator (for the reduction of B_j), but also an increasing number of variables $j \in Q$ in general and in the training set $\tilde{Q}$ (for the reduction of A_j.

This decomposition differs from Cho et al. [2002] in that sense that Cho et al. [2002] contrast the GVF prediction with the direct variance estimator $\widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]$, whilst in (3.15), the design variance only enters the estimator $\hat{\psi}$. As the aim of GVFs is to predict correctly the estimator's variance, the contrast to the 'true' model-design variance seems more plausible, especially when later on efficiency comparisons with estimators $\widehat{\mathrm{Var}}_D\left[g_S^{(j),}\left(S,Y^{(j)}\right)\right]$ are envisaged. Furthermore, the decomposition in Cho et al. [2002] can be misleading because the probability law that is assumed remains unclear: From their efficiency comparisons, it seems that P_D is the law to be studied but from their GVF definition, the estimators' variance is assumed to be a random variable implying that the law applied there is $P_{\mathcal{M}_\theta,D}$. The derivation of GVFs in Valliant et al. [2000, Chapter 10.3] also suggests that the underlying probability law is $P_{\mathcal{M}_\theta,D}$, in accordance with the analysis given here.

The error decomposition in (3.15) allows to derive an approximate formula for the Mean Squared Error (MSE) of GVFs under the assump-

tion that f is twice continuously differentiable. In that case, a first order Taylor approximation yields

$$\begin{aligned}
&\mathrm{MSE}_{\mathcal{M}_\theta,D}\left[\widetilde{\mathrm{Var}}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]\right] \\
&\quad\approx \left(\nabla f\left(\mathrm{E}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right],\psi\right)\right)^T \cdot \mathrm{MSE}\left[\begin{pmatrix} g_S^{(j)}\left(S,Y^{(j)}\right) \\ \hat{\psi}\end{pmatrix}\right] \\
&\quad\cdot \left(\nabla f\left(\mathrm{E}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right],\psi\right)\right) \quad , \qquad (3.16)
\end{aligned}$$

where terms A_j^2 (approximately the variance of $\hat{\psi}$) and B_j^2 (approximately the squared bias of $\hat{\psi}$) are found in the product of $\mathrm{MSE}\left[\hat{\psi}\right]$ with the first derivative and ξ_j^2 appears in the multiplication of $\mathrm{MSE}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]$ with the first derivative. The matrix notation is used to account for a possible covariance between the estimation error of $\hat{\psi}$ and $g_S^{(j)}\left(S,Y^{(j)}\right)$ but for least squares estimation of $\hat{\psi}$, the covariance is usually zero.

Ignoring a possibly non-zero covariance with the estimated parameter vector, the MSE of the GVF predictor is thus a linear function in $\mathrm{Var}_{\mathcal{M}_\theta}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]$. For the model-design MSE of the design-based variance estimator, we have

$$\begin{aligned}
&\mathrm{MSE}_{\mathcal{M}_\theta,D}\left[\widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]\right] \\
&\quad= \mathrm{E}_{\mathcal{M}_\theta,D}\left[\left(\hat{V}_D\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]-\mathrm{E}_{\mathcal{M}_\theta}\left[\mathrm{Var}_D\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]\right]\right)^2\right] \\
&\qquad+\mathrm{Var}_{\mathcal{M}_\theta}\left[\mathrm{E}_D\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]\right]^2 \qquad (3.17)
\end{aligned}$$

If thus the shape of f is favorable, $P_{\mathcal{M}_\theta}$ and the population size are such that the variance of the population total is relatively large, whilst P_D returns a great variability in the sample (cf. the stratified Two-Stage Cluster Sampling (TSC) with $P_{\mathcal{M}_\theta} \in \{\mathrm{Pois}(24), \Gamma(8,3)\}$ in the simulation study), there are noticeable efficiency gains.

Note that the preceding analysis can be relatively easily applied to estimators like the HT or GREG, for which the stated assumptions hold. Using linearization methods, the derivations also become applicable to nonlinear statistics, when the GVF are applied on the linearized variables whose population total then become of interest. Valliant [1992] uses another approach for nonlinear statistics: He first determines the variance of the nonlinear statistic as a function of the statistic itself under a pre-defined probability model $P_{\mathcal{M}_\theta}$ and estimates the model directly. Of course, this approach is also valid, but theoretically more cumbersome than application of the GVF to the linearized estimators, for which functional relations $\tilde{f}$ and f are then assumed.

3.2.3 *GVFs and the Design Effect*

Most of the preceding analysis simlifies a lot under SRS, as examplifies the scenario in Example 1. If there is a relation between variance and expectation of an estimator under SRS,

$$\mathrm{Var}_{\mathcal{M}_\theta,\mathrm{srs}}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right] = f\left(\mathrm{E}_{\mathcal{M}_\theta,\mathrm{srs}}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right],\psi\right) \quad , \tag{3.18}$$

then the use of the design effect [Kish, 1965]

$$\mathrm{deff}_D \triangleq \frac{\mathrm{Var}_D\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]}{\mathrm{Var}_{\mathrm{srs}}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]} \tag{3.19}$$

can help to motivate the use of GVFs under complex designs, too. First assume that the expectation of $g_S^{(j)}$ is the same under both $P_{\mathcal{M}_\theta,\mathrm{srs}}$ and $P_{\mathcal{M}_\theta,D}$, which holds, for example, for the HT or the GREG. Then, it is easy to adjust for the design in (3.18): We get

$$\mathrm{Var}_D\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right] = \frac{f\left(\mathrm{E}_{\mathcal{M}_\theta,\mathrm{srs}}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right],\psi\right)}{\mathrm{deff}_D} \quad . \tag{3.20}$$

Cao et al. [2012] do this for the case of binary $Y^{(j)}$. For a logartihmic model (take the logarithm on both sides of (3.20)), the negative logarithmic design effect becomes a summand and hence, the design effect simply impacts the regression intercept when $\log f$ is a linear regression. If the linear regression intercept is zero under f and SRS (for example, under the binomial, exponential or Poisson distribution), the size of the design effect then can be directly estimated. For example, Cao et al. [2012] relate their GVF for binary data to the design effect.

Note that in this set-up – as can alread be learned from the fact that the left hand side of (3.19), the design effect has to be constant for all variables $j \in \mathcal{Q}$, again underpinning the importance that variable set $\mathcal{Q}$ only contains variables where expectation and variance relate in the same way and with the same parameters.

Accounting for deff_D is important as most of the samples used for GVF application are stratified two stage samples [Johnson and King, 1987, Cao et al., 2012]. If on the other hand, estimates of the design effect are available (the true value is usually unknown), they may also be used in estimating the GVF as is motivated by Equation (3.20).

However, note that the fact that the design effect is estimated adds again an explanatory with measurement error. In the case that the estimator for the design effect is unstable, the inclusion of deff_D into the GVF might be more destablizing than helpful.

3.2.4 *GVFs in a Design-based Framework*

In practice, GVFs are used in a design-based framework, where the variance of an estimator for an unknown finite population statistic has to estimated. The justification of GVFs under such a set-up is

$$\mathrm{Var}_D\left[g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)\right] = g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)^2 - \mathrm{E}_D\left[g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)\right]^2 + \xi_j$$

where

$$\xi_j \triangleq \mathrm{E}_D\left[g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)^2\right] - g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)^2 \quad .$$

Obviously, the error is unbiased, $\mathrm{E}_D\left[\xi_j\right] = 0$. Knowing that for a consistent variance estimator $\widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)\right]$, we have [Cho et al., 2002]

$$\widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)\right] = \mathrm{Var}_D\left[g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)\right] + \zeta_j \tag{3.21}$$

with ζ_j converging to o in probability. Consequently,

$$\begin{aligned}\widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)\right] = {} & g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)^2 - \mathrm{E}_D\left[g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)\right]^2 \\ & + \xi_j + \zeta_j \quad .\end{aligned} \tag{3.22}$$

Note that this relation holds always independently from the DGP of $\mathbf{y}$ and the design P_D, but the distribution of $\xi_j + \zeta_j$ is unknown and as $\mathrm{E}_D[\zeta_j]$ is not necessarily zero because the variance estimator is only asymptotically unbiased, the sum has not necessarily zero expectation neither.

Hence, the application of the design effect, in this context, complicates the analysis as then in

$$\begin{aligned}\widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)\right] = {} & \mathrm{deff}_D \cdot g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)^2 \\ & - \mathrm{deff}_D \cdot \mathrm{E}_{srs}\left[g_S^{(j)}\left(S,\mathbf{y}^{(j)}\right)\right]^2 + \xi_j\end{aligned}$$

the estimator $g_S^{(j)}$ and the error ξ_j would be evaluated under $S \sim P_{srs}$, even when $\mathrm{E}_D\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right] = \mathrm{E}_{srs}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]$, and such a sample realization is usually not available when $S \sim P_D \neq P_{srs}$. Consequently, the interpretation of regression coefficients as average design effect like in Johnson and King [1987] or Cao et al. [2012] is not applicable, as the intercept of the regression needs always to be non-positive (which does not always happen in practice) and is a product of two terms where the design effect is only a component of. Nonetheless, an estimator of the design effect could be derived from the regression result. Therefore, these works can be put into the previously outlined model-design context.

From Equation (3.22), it follows that the functional relation f therefore is always quadratic with restricted slope parameters and for regression estimation, it is furthermore hoped that $\widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S, \mathbf{y}^{(j)}\right)\right] = f\left(g_S^{(j)}\left(S, \mathbf{y}^{(j)}\right); \boldsymbol{\psi}\right) + \varepsilon_j$ has a normally distributed error $\varepsilon_j \sim N(0, \sigma_j^2)$ with $\varepsilon_j = \xi_j + \zeta_j$. A logarithmic transformation as it is discussed in Section 3.3.1 is not theortically justifiable in a purely design-based framework; in contrast to the relative variance regressed on the inverse point estimator.

For the estimation of the parameter vector $\boldsymbol{\psi}$, in addition, it has to be kept in mind that $\{g_S^{(j)}\left(S, Y^{(j)}\right)\}_{j \in \Omega}$ are correlated as they stem from the same sample S, as classical estimation methods assume that the input data are independent. Furthermore, Model (3.22) imposes $E_D\left[g_S^{(j)}\left(S, \mathbf{y}^{(j)}\right)\right] \equiv c$ for all $j \in \Omega$. This is a very restrictive assumption, but is similar to the assumption $Y^{(j)} \sim P_{\mathcal{M}_\theta} \quad \forall j \in \Omega$ in the model-design framework.

From (3.21) it is obvious that the MSE of the design variance is $E_D\left[\zeta_j^2\right]$. For a GVF prediction

$$\widetilde{\mathrm{Var}}_D^{\star}\left[g_S^{(j)}\left(S, \mathbf{y}^{(j)}\right)\right] = g_S^{(j)}\left(S, \mathbf{y}^{(j)}\right)^2 - E_D\left[g_S^{(j)}\left(S, \mathbf{y}^{(j)}\right)\right]$$

(ignoring that the second term has to be estimated) stemming from (3.22), we have on the other hand

$$\mathrm{MSE}_D\left[\widetilde{\mathrm{Var}}_D\left[g_S^{(j)}\left(S, \mathbf{y}^{(j)}\right)\right]\right] = \mathrm{MSE}_D\left[\xi_j + \zeta_j\right]$$

and it is not a priori clear which MSE is smaller: If the two errors are negatively correlated, one could observe an efficiency gain in using the GVF prediction, otherwise, there is an efficiency loss even without accounting for the need to estimate $E_D\left[g_S^{(j)}\left(S, \mathbf{y}^{(j)}\right)\right]^2$ via OLS.

3.3 GVFS IN PRACTICE

In this section, the application of GVFs in practice is discussed. Especially, the most common GVF shapes and their motivation are presented. The problem of the preceding analysis is that GVFs cannot be evaluated in practice because the estimator's 'true' variances, both the design and the model-design variance, are unknown. Therefore, other indicators for a GVF model's predictive power that can be used in practice are necessary. Such quality measures are discussed in the Subsection 3.3.2.

GVFs are most often used in large-scale household surveys like the Current Population Survey [Wolter, 1985, Chapter 5], the American Community Survey [Fuller and Tersine Jr, 2010] and the Consumer Expenditure

Survey [Eltinge and Sukasih, 2001]. However, they were also studied in the Current Employment Statistics [Cho et al., 2014] and in school surveys [Salvucci et al., 1995] – as schools are institutions, the latter would already be interpreted as business survey in Cox and Chinnappa [1995]. Applications that come closest to business surveys like they are considered here (i.e. without institutions) are found in Alegria and Scott [1991] and Filiberti et al. [2007], both looking at the economic sector of forestry and agriculture. Pavone and Russo [1999] apply GVFs to classical business variables such as 'annual turnover', 'value added', 'labor costs' and 'investments'.

3.3.1 *Common Shapes of GVFs*

Many GVF applications are motivated by the binary case (cf. Model (3.5)) where the relative variance, i.e. the variance of random variable divided by its squared expectation, equals

$$\operatorname{relVar}_{\mathcal{M}_\theta}\left[\hat{\theta}\right] = \frac{1-\theta}{N\cdot\theta} \quad ,$$

where $\hat{\theta}$ is the estimator of the success probability θ and N is the number of independent trials. A more general shape that encompasses this relative variance as a special case is then for $j \in \mathbb{Q}$

$$\operatorname{relVar}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right] = \psi_0 + \psi_1\frac{1}{g_S^{(j)}\left(S,Y^{(j)}\right)} + \psi_2\frac{1}{g_S^{(j)}\left(S,Y^{(j)}\right)^2} + \varepsilon_j \quad , \tag{3.23a}$$

where the estimator $g_S^{(j)}$ takes the place of the original parameter. ε_j is here and in the subsequent equations considered to be a zero-centered random error resulting from using $g_S^{(j)}\left(S,Y^{(j)}\right)$ instead of its expectation. It encompasses the errors that were discussed in Section 3.2. Model (3.13) encomapsses mean estimators not only for binomially, but also Poisson (3.6) and Gamma (3.7) distributed variables $Y^{(j)}$. And given the results from Section 3.2.4, this shape is also justifiable with the design-based framework in which GVFs are usually used in. Salvucci et al. [1995] estimates a square root version of (3.23a) – as the square is a nonlinear function, the choice between the relative variance and the coefficient of variation as dependent variable in Model (3.23a) depends on the statistic that shall be estimated unbiasedly under otherwise correct model assumptions. Copeland et al. [2006] discusses Model (3.23a), too.

Note, however, the difference: The shape is motivated using $P_{\mathcal{M}_\theta}$ whilst the discussion in Section 3.2 relied on $P_{\mathcal{M}_\theta,D}$ and as seen in Equation (3.8b), we have only $f = \mathcal{O}\left(N\tilde{f}\right)$ and not $f \propto \tilde{f}$ even under 'good' survey designs when the population size is too small. Theoretical motivation

of GVF models other than (3.23a) are therefore only approximative and suggestive.

Model (3.23a), however, does not prevent predictions from becoming negative. This motivates the log-log regression model

$$\log \mathrm{Var}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right] = \psi_0 + \psi_1 \log g_S^{(j)}\left(S,Y^{(j)}\right) + \varepsilon_j \quad , \tag{3.23b}$$

from which predictions are always positive due to the exponentiation. Pavone and Russo [1999] and Cho et al. [2002] use the log-log regression and also the Australian Bureau of Statistics as well as Statistics Canada [Di Consiglio et al., 2013]. Substracting on both sides $2\log g_S^{(j)}\left(S,Y^{(j)}\right)$ yields

$$\log \mathrm{relVar}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right] = \psi_0 + (\psi_1 - 1)\log\left(\frac{1}{g_S^{(j)}\left(S,Y^{(j)}\right)}\right) + \varepsilon_j \quad ,$$

which is the log-log version of the regression model (3.23a) with $\psi_2 = 0$.

Alegria and Scott [1991] investigate other shapes for GVFs, which are partially nonlinear and require knowledge of some standard errors in advance. Note though that their Model (4) corresponds to Model (3.23b). These models, however, do not seem to be used in other applications of GVFs. Finally, Di Consiglio et al. [2013] mentions that a nonlinear regression of the type

$$\mathrm{relVar}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right] = \frac{\psi_0}{g_S\left(S,Y^{(j)}\right)^{\psi_1}} + \varepsilon_j \tag{3.23c}$$

is used by the Greek Statistical Authority. This model can also be captured by Model 3.23b when the different distributional assumptions on the error ε_j are ignored. The base model, of which the afore discussed models can be seen as a transformation, is the linear regression with quadratic term of $g_S\left(S,Y^{(j)}\right)$,

$$\mathrm{Var}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right] = \psi_0 + \psi_1 \cdot g_S^{(j)}\left(S,Y^{(j)}\right) + \psi_2 \cdot g_S^{(j)}\left(S,Y^{(j)}\right)^2 + \varepsilon_j \quad . \tag{3.23d}$$

Otto and Bell [1995] use this regression model in GVF estimation, which are again used for other estimations.

Note though that the choice to use the coefficient of variation or the relative variance or the basic quadratic model or a logarithmic transformation thereon depends on the distributional assumptions on the error

term ε_j. The respective model, though, is usually chosen such that normality, $\varepsilon_j \sim N\left(0, \sigma_j^2\right)$, holds approximately as most often, the loss function ℓ_j (cf. Equation (3.11)), is the squared error. Of course, the Models (3.23) can all include additional information, for example in terms of an estimated design effect [Pavone and Russo, 1999, for example] or the sample size [Cho et al., 2002].

Another issue when the relative variance or the logarithmic variance is used as dependent variable in the regression, is back-transformation for the GVF prediction. Let $P^{GVF}_{\mathcal{M}_\theta,D}$ denote the probability law for the GVF given the design P_D and the DGP $P_{\mathcal{M}_\theta}$.

If the logarithmic regression for the model-design variance is assumed to be correct (Model (3.23b)), $P^{GVF}_{\mathcal{M}_\theta,D}$ is the log-normal distribution of the random variable

$$\exp(\psi_0) \cdot g_S^{(j)}\left(S, Y^{(j)}\right)^{\psi_1} \quad .$$

Then, the correct prediction is $E_{GVF}\left[\exp(\psi_0) \cdot g_S^{(j)}\left(S, Y^{(j)}\right)^{\psi_1}\right]$ under the model, given Y and S

$$\begin{aligned} & E_{GVF}\left[\exp\left(\psi_0 + \psi_1 \cdot \log g_S^{(j)}\left(S, Y^{(j)}\right) + \varepsilon_j\right) | S, Y^{(j)}\right] \\ & = g_S^{(j)}\left(S, Y^{(j)}\right)^{\psi_1} \cdot \exp\left(\psi_0 + \frac{\sigma_j^2}{2}\right) \\ & \neq g_S^{(j)}\left(S, Y^{(j)}\right)^{\psi_1} \cdot \exp\left(\psi_0\right) \end{aligned} \tag{3.24}$$

(cf. Definition 8). Thus, back-transformation must account for nonlinearities. If, in contrast, Model (3.23a) is assumed to be correct, it follows that

$$\begin{aligned} & E_{GVF}\left[\psi_0 g_S^{(j)}\left(S, Y^{(j)}\right)^2 + \psi_1 g_S^{(j)}\left(S, Y^{(j)}\right) + \psi_2 \right. \\ & \qquad \left. + \varepsilon_j g_S^{(j)}\left(S, Y^{(j)}\right)^2 | S, Y^{(j)}\right] \\ & = \psi_0 \cdot g_S^{(j)}\left(S, Y^{(j)}\right)^2 + \psi_1 \cdot g_S^{(j)}\left(S, Y^{(j)}\right) + \psi_2 \end{aligned}$$

as one model assumption is $\varepsilon_j \perp g_S^{(j)}\left(S, Y^{(j)}\right)$. The variance of the prediction is, though, $\sigma_j^2 \cdot g_S^{(j)}\left(S, Y^{(j)}\right)$. That means that even if ε_j is homoskedastic, this does not hold for the variance prediction.

3.3.2 *Quality Measures*

Like already mentioned, an error analysis like in (3.15) is not feasible in practice because the model-design variance $\mathrm{Var}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]$ is

unknown. Thus, the question arises how a practitioner can assess the predictive quality of her estimated GVF model. As the GVF are usually considered to be a regression problem (cf. Section 3.3.1), it is common practice to use quality measures for regression.

The most frequently used measure is the (adjusted) R^2, that is, the share of variation explained by the model, used for example by Salvucci et al. [1995] and Hinrichs [2003]. If the inclusion of additional auxiliary information is at discussion, the Akaike Information Criterion (AIC) is a measure, for example, for model selection. However, in contrast to the R^2, the AIC only allows to compare GVFs of the same shape with a varying number of explanatories. This is why the right hand side of the regression models in Otto and Bell [1995] differ extremely (partially involving highly nonlinear regression), but the left hand sides of the GVF equal constantly $\widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]$. The AIC is based on the negative log-likelihood of the regression model, it is thus not surprising that the log-likelihood has been used to assess GVFs, too [Maples et al., 2009]. The problem with all these measures is that they give information about the model fit to the training data (here the tuples $\left(g_S\left(S, Y^{(j)}\right), \widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]\right)$ with $j \in \tilde{\mathcal{Q}}$) but are limited in the assessment of predictive quality – the prediction error is downward biased when observations are used both, prediction assessment and training. This underlines that the choice of the training subset $\tilde{\mathcal{Q}} \subset \mathcal{Q}$ whose tuples are used to estimate ψ should be a good choice of variables from $\mathcal{Q}$: In practice, often sub-population counts are of interest, e.g. differentiated by age group and gender. If then only tuples $\left(g_S\left(S, Y^{(j)}\right), \widehat{\mathrm{Var}}_D\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]\right)$ of age×gender categories are used for training, but $\mathcal{Q}$ also includes sub-population counts, say for employment status, the estimated model has poor adjustment to $\mathcal{Q} \setminus \tilde{\mathcal{Q}}$.

Gershunskaya and Dorfman [2013] offer a new quality measure particularly for GVFs, which does not originate from regression analysis. Their approach has the nice property that it allows to re-adjust GVF predictions if these are found to (over-) underestimate structurally the true variances, where the bias is detected based on this new quality measure that we call henceforth $\bar{\delta}_{1-\alpha}$, where $1-\alpha$ is a typical confidence level, here $1-\alpha = 0.95$. Let $g_S^{(j)}\left(S, Y^{(j)}\right)$ be the HT of the variable of interest $Y^{(j)}$. Berger [1998] proves that for high entropy sampling designs P_D, the asymptotic normality of the HT holds,

$$\frac{g_S^{(j)}\left(S, Y^{(j)}\right) - E_{\mathcal{M}_\theta, D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]}{\sqrt{\mathrm{Var}_{\mathcal{M}_\theta, D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]}} \xrightarrow[d]{} N(0,1) \quad .$$

If the distribution of the standardized HT is approximately standard normal for large n and N and all $j \in \mathcal{Q}$, Gershunskaya and Dorfman [2013] argue that this implies that standardization of the HT by the square root of good GVF predictions $\widetilde{\mathrm{Var}}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]$,

$$z_j \triangleq \frac{g_S^{(j)}\left(S,Y^{(j)}\right) - \mathrm{E}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]}{\sqrt{\widetilde{\mathrm{Var}}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]}} \tag{3.25a}$$

or an estimator thereof

$$\hat{z}_j \triangleq \frac{g_S^{(j)}\left(S,Y^{(j)}\right) - \tilde{g}_S^{(j)}\left(S,Y^{(j)}\right)}{\sqrt{\widetilde{\mathrm{Var}}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}\left(S,Y^{(j)}\right)\right]}} \quad , \tag{3.25b}$$

($\tilde{g}_S^{(j)}\left(S,Y^{(j)}\right)$ being another unbiased estimator of $\sum_{i=1}^{N} Y_i$) should yield approximately for $(1-\alpha)\cdot 100$ percent of the variables a value in the $100\cdot(1-\alpha)$ standard normal confidence interval $I_{(1-\alpha)} = (z_{\alpha/2}, z_{1-\alpha/2})$. We define therefore the estimated quality measure

$$\delta \triangleq \frac{1}{|\mathcal{Q}|}\sum_{j\in\mathcal{Q}} \mathbb{1}_{I_{(1-\alpha)}}(\hat{z}_j) \quad . \tag{3.26}$$

For $\tilde{g}_S^{(j)}$ in (3.25b), Gershunskaya and Dorfman [2013] use a balanced half sample replication; arguing that only one replication is needed as it is not intended to do variance estimation, and errors should cancel out amongst the variables $\{Y_{j\in\mathcal{Q}}^{(j)}\}$. Therefore, the computational burden for the quality measure $\bar{\delta}_{1-\alpha}$ is manageable.

If $\bar{\delta}_{1-\alpha}$ does not approximately equal $1-\alpha$, adjustments to the GVF are necessary under the normality assumption because then the predictions are structurally too high ($\bar{\delta}_{1-\alpha} > 1-\alpha$) or too low ($\bar{\delta}_{1-\alpha} < 1-\alpha$). Adjustments can be done either by calibration [Gershunskaya and Dorfman, 2013] or possibly another shape is more adequate. Note however, that the quality of the measure $\bar{\delta}_{1-\alpha}$ relies on asymptotic assumptions for the HT which need not necessarily hold for P_D and U.

Concerning the transferability of $\bar{\delta}_{1-\alpha}$, when the GREG rather than the HT is employed in the large scale survey, Kim and Park [2010] states that $\hat{\tau}_y^{\mathrm{GREG}}$ is asymptotically equivalent to the HT (cf. Remark 2) and therefore also normal, because convergence in probability implies convergence in distribution. So $\bar{\delta}_{1-\alpha}$ is also applicable for the GREG. For other estimators, though, asymptotic normality must be proved to argue the use of $\bar{\delta}_{1-\alpha}$. However, the applicability of GVFs to other estimators with possibly other estimands than the population total must be studied a priori, too. Furthermore, note that $\bar{\delta}_{1-\alpha}$ also allows to check the asymptotic

assumptions: Different values for α could be plugged in and if there are different conclusions for different values in α, this indicates that the asymptotics might not hold.

3.3.3 *Estimation of GVFs*

As the shapes of GVFs introduced in Section 3.3.1, are (non-)linear regressions, standard regression estimators are usually used to estimate the parameter vector ψ. Indeed, OLS is common practice to determine $\hat{\psi}$ [Valliant et al., 2000, chapter 10.3]. However, this requires the additional assumption that not only $\varepsilon_j \sim_{ind} N(0, \sigma_j^2)$ but also $\sigma_j^2 \equiv \sigma^2$ for all $j \in Q$.

A less restrictive assumption is

$$\sigma_j^2 \propto \text{relVar}_{\mathcal{M}_\theta, D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]^2$$

which is, for example, suggested in Wolter [1985, Chapter 5.4] and Valliant et al. [2000, Chapter 10.3]. The assumption is motivated by the observation that the variance is estimated less stably as the point estimator (and thus presumably its expectation) increases. Then, a good estimation method for ψ is weighted OLS, in practice the estimated relative variance is used rather than $\text{relVar}_{\mathcal{M}_\theta, D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]^2$.

Alternatively, noting that the proportionality of σ_j is with the dependent variable, Wolter [1985] suggests to use an iterative estimation algorithm assuming that

$$\sigma_j^2 \propto \widetilde{\text{Var}}_{\mathcal{M}_\theta, D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right] \quad .$$

In words, this assumption means that first and second moment of the GVF prediction are assumed to be in a linear relation.

As a last weighting scheme for weighted OLS, Copeland et al. [2006] uses as weights the logarithmic relative variance, which corresponds to the assumption that

$$\sigma_j^2 \propto \frac{1}{\log \text{relVar}_{\mathcal{M}_\theta, D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]} \quad ,$$

which is hard to justify from a theoretical point of view: Observations $\tilde{Q}$ that have a higher relative variance are weighted higher, contrary to the assumptions in Wolter [1985] and Valliant et al. [2000]. Furthermore, it might happen, that such weights become negative when the relative variance is smaller than 1.

Note that such weighting schemes are in contrast with the necessary assumption that for all $j \in Q$, we have $Y^{(j)} \sim P_{\mathcal{M}_\theta}$. In addition, the fact

that estimators for the correct heteroskedasticity weights are used because the (relative) variances are not available, can counteract the goal to stabilize estimation.

Valliant et al. [2000, Chapter 10.3] further notes that in practice, as we have pointed out in Section 3.2, rather than the model-design variance, an estimator $\widehat{\text{Var}}_D\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right]$ is used. Similarly, not the the expectation of the final population statistic, $E_{\mathcal{M}_\theta}\left[g_U^{(j)}\left(Y^{(j)}\right)\right]$, is the estimand of the HT, but the random finite population outcome $g_U^{(j)}\left(Y^{(j)}\right)$ of but an estimator thereof is used. Hence, one could think of an error-in-variables model to estimate the GVF. However, such a model requires additional assumptions about the kind of error in variables – that adds another source of possible bias when violated.

3.3.4 *GVFs in Small Area Estimation*

Besides their application in large-scale surveys with a big number of variables of interest, a typical application of GVFs is area-level SAE. The basic set-up here consists of the FH [Fay and Herriot, 1979]. The goal of SAE is to estimate/ predict finite population statistics for $m = 1, \dots, M$ pairwise disjoint subsets of the population $U = \cup_{m=1}^{M} U_m$ of size $|U_m| =: N_m$,

$$g_{U_m} : \times_{k=1}^{N_m} \mathbb{R}^p \to \mathbb{R}^q \quad (\mathbf{y}_i : i \in U_d) \mapsto g_{U_m}\left((\mathbf{y}_i : i \in U_m)\right) \tag{3.27}$$

In the following, we assume g_{U_m} to be a sub-population total $\sum_{i \in U_m} y_i$ with $y_i \in \mathbb{R}$. For each index $m = 1, \dots, M$, the Fay-Herriot Estimator (FH) assumes two models to hold for the sub-population totals of interest τ_{y_m} and the HT thereof, namely

$$\tau_{y_m} \sim_{\text{ind}} N\left(\boldsymbol{\tau}_{x_m}^T \boldsymbol{\beta}, \sigma^2\right) \tag{3.28a}$$

$$\hat{\tau}_{y_m} \sim_{\text{ind}} N\left(\tau_{y_m}, \text{Var}_D\left[\hat{\tau}_{y_m}\right]\right) \quad , \tag{3.28b}$$

where $\boldsymbol{\tau}_{x_m}$ is the known sub-population total of the p-dimensional auxiliary variables. Obviously, the models for $m = 1, \dots, M$ all share the parameters $\boldsymbol{\beta}$ and σ^2. This assumption of a common structure in the DGP can yield efficiency gains because then observations from additional domains $l = 1, \dots, M$, $l \neq m$, can included in the estimation process for

τ_{y_m}, thus increasing implicitely the sample size for the domain m. Consequently, the MSE of the FH estimator

$$\hat{\tau}_y^{FH} \triangleq \tilde{\boldsymbol{\tau}}_x \left(\tilde{\boldsymbol{\tau}}_x^T V^{-1} \tilde{\boldsymbol{\tau}}_x\right)^{-1} \tilde{\boldsymbol{\tau}}_x^T V^{-1} (\hat{\tau}_{y_m} : m = 1, \dots, M) \tag{3.28c}$$

$$\tilde{\boldsymbol{\tau}}_x \triangleq \begin{pmatrix} \boldsymbol{\tau}_{x_1} \\ \vdots \\ \boldsymbol{\tau}_{x_M} \end{pmatrix}, \quad V \triangleq \operatorname{diag}\left(\sigma^2 + \operatorname{Var}_D[\hat{\tau}_{y_m}] : m = 1, \dots, M\right)$$

is smaller than $\operatorname{Var}_D[(\hat{\tau}_{y_m} : m = 1, \dots, M)]$.

Note the link between model and design in the FH: The sub-population total is a realization of a normal random variable with expectation $\boldsymbol{\tau}_{x_m}^T \boldsymbol{\beta}$ and variance σ^2. In the notation of Remark 5, $\sigma^2 = \operatorname{Var}_{\mathcal{M}_\theta}[E_D[\hat{\tau}_{y_m}]]$ and given $\mathbf{y}$, $E_{\mathcal{M}_\theta}[\operatorname{Var}_D[\hat{\tau}_{y_m}] | Y = \mathbf{y}] = \operatorname{Var}_D[\hat{\tau}_{y_m}]$.

However, in real world application, $\operatorname{Var}_D[(\hat{\tau}_{y_m} : m = 1, \dots, M)]$ is not available and must be estimated from a random subset $S \cap U_m$ by $\widehat{\operatorname{Var}}_D[\hat{\tau}_{y_m}]$ and then plugged into (3.28c) [Otto and Bell, 1995]. This practice reveals a paradox, though: Whilst the FH is motivated by the HT's inefficiency (i.e. high variance) for small sample sizes $U_m \cap S$, an estimated second order statistic of $\hat{\tau}_{y_m}$, which is even less stable than an estimator for first order statistics, is used in the estimation process. Note that this is again under the condition that the sub-population variances relate equally to the sub-population expectations and that the design P_D has the same impact on all sub-samples $S \cap U_m$.

As it was outlined in the previous section, GVFs can yield stabler variance predictions than the direct variance estimator in a model-design framework and therefore stabilize also small area estimators. Even in the original work, Fay and Herriot [1979] use a GVF prediction of the shape (3.23c) and plugged the estimate into (3.28c). Otto and Bell [1995], Maples et al. [2009] and Hawala and Lahiri [2010] use GVFs as well, mostly for poverty statistics in small areas.

3.4 SIMULATION STUDY

As the theoretical analysis of GVFs takes place under the probability law $P_{\mathcal{M}_\theta, D}$ although the application in practice is under P_D, the illustrative simulation study in this section has two steps: The data used in every simulation run are repeatedly generated by superpopulation models $\mathcal{M}_\theta$. In a second step, different shapes of GVFs are studied with respect to their predictive power under different laws $P_{\mathcal{M}_\theta, D}$. Using varying superpopulation models $\mathcal{M}_\theta$ (different parameters $\theta_1, \theta_2 \in \Theta$ and different shapes for $\mathcal{M}$), the sensitivity of the predictive power to the shape-parameter-choice can be studied. The impact of survey design, stated in Section 3.2, is also studied as three different designs are simulated.

Furthermore, it can be evaluated whether the GVF prediction is much harmed when the underlying data originates from different superpopulation models, i.e. when the assumption $Y^{(j)} \sim P_{\mathcal{M}_\theta}$ for all $j \in \mathcal{Q}$ is violated.

3.4.1 *DGP in the Model-design Framework*

MODEL-BASED FINITE POPULATION

This simulation study on GVFs deals with theoretical aspects of the set-up. Consequently, in each of the $r = 1, \dots, 1500$ simulation runs, a finite population U, and $|U| = N = 10000$, is generated under the following DGP for all $j \in U$

$$Y_i^{(j)} \sim_{\text{iid}} \mathcal{G}(3,2), \qquad j = 1, \dots, 50 \tag{3.29a}$$
$$Y_i^{(j)} \sim_{\text{iid}} \mathcal{G}(8,3), \qquad j = 51, \dots, 100 \tag{3.29b}$$
$$Y_i^{(j)} \sim_{\text{iid}} \text{Pois}(6), \qquad j = 101, \dots, 150 \tag{3.29c}$$
$$Y_i^{(j)} \sim_{\text{iid}} \text{Pois}(24), \qquad j = 151, \dots, 200 \tag{3.29d}$$

This set-up is interesting because $E_{\mathcal{M}_\theta}\left[Y_i^{(1)}\right] = E_{\mathcal{M}_\theta}\left[Y_i^{(101)}\right]$ and $E_{\mathcal{M}_\theta}\left[Y_i^{(51)}\right] = E_{\mathcal{M}_\theta}\left[Y_i^{(151)}\right]$ and thus the same holds for the HT under $P_{\mathcal{M}_\theta, D}$, too. Furthermore, the functional relation between variance and expectation in the DGP are not the same but similar between Poisson and Gamma distribution (cf. Equations 3.6 and 3.7).

First, these properties allow us to study the quality of the GVF, when the assumptions on the DGP hold (i.e. when one accounts for the different structures between Gamma and Poisson distributions and different parameters within these families). Second, the impact on GVFs can be studied when these assumptions only hold approximately (i.e. when either the variables do not originate from the same class of distributions or have different moments). Furthermore, it can be studied whether some GVF shapes are more sensitive to assumption violations.

SAMPLING RANDOMIZATION

Whilst the relational similarity between a random variable's $Y_i^{(j)}$ $P_{\mathcal{M}_\theta}$-expectation and variance is most easily transferred to the model-design setting under SRS (cf. Example 1), the impact of differing sampling designs is of interest in the simulation study, too. Therefore, not only SRS is implemented, but also a Stratified Random Sampling (StratRS) and a stratified Two-Stage Cluster Sampling (TSC). The assignment to strata and clusters is as follows: Units are assigned randomly to a partition $\{U_\mu\}_{\mu=1,\dots,25}$ of the finite population. The disjoint subsets are of variable size and the assignments to the partition sets is kept constant over all $R = 1500$ MC runs. A second partition $\{\tilde{U}_\nu\}_{\nu=1,\dots,5}$ is defined with

$\tilde{U}_1 = \biguplus_{\mu=1}^{5} U_\mu, \ldots, \tilde{U}_5 = \biguplus_{\mu=21}^{25} U_\mu$. A detailed overview about the allocations to the partitions can be found in Appendix B.

For the StratRS, units are drawn randomly from each sub-population U_μ, $\mu = 1, \ldots, 25$, with proportional allocation

$$n_\mu \triangleq \left\lceil n \cdot \frac{|\tilde{U}_\mu|}{|U|} \right\rceil \quad .$$

$n = \sum_{\mu=1}^{25} n_\mu = 253$ is the fixed sample size for SRS. For the stratified TSC the partition $\{\tilde{U}_\nu\}_\nu$ serves for stratification, and on the first sampling stage, two subsets U_μ are chosen from each $\tilde{U}_\nu$. Then, on the second stage, roughly 30% of the units (but at least one) in the drawn $\tilde{U}_\mu$ are selected at random and enter the survey sample S. The expected sample size for the TSC design is comparable (but not exactly the same) to the fixed sample size n in SRS and StratRS.

GVF PREDICTION

The GVF shapes under study are the Models (3.23a), (3.23b), (3.23c) and (3.23d). If it is accounted for the different original distributions of $Y_i^{(j)}$ in the GVF model, this is done by the inclusion of indicator variables, for example in shape (3.23b) as

$$\begin{aligned} \log \operatorname{Var}_{\mathcal{M}_\theta, D}\left[g_S^{(j)}\left(S, Y^{(j)}\right)\right] = & \\ \sum_{P \in \{\mathcal{G}(2,3), \mathcal{G}(8,3), \text{Pois}(6), \text{Pois}(24)\}} & \psi_{0,P} \cdot \mathbb{1}(Y^{(j)} \sim P) \\ + \psi_{1,P} \cdot \mathbb{1}(Y^{(j)} \sim P) \cdot \log g_S^{(j)}\left(S, Y^{(j)}\right) + \varepsilon_j, & \quad \varepsilon_j \sim_{\text{ind}} N(0, \sigma_j^2) \, . \end{aligned}$$

That means, that in this case, the assumption violation is only in terms of the variance of ε_j. Use of indicators in a model are indicated in the summarizing Figures by the hint 'indicators'. Models are designated by their dependent variable.

GVFs are estimated on a random subset $\tilde{Q} \subset Q$, $|\tilde{Q}| = 20$ of variables, where it is assured that from each distribution type, 5 variables are included so that the indicator models are always estimable. Due to the common use in practice, weighted OLS are run with the assumption of proportionality between σ_j^2 and the squared relative variance. Due to the possible instabilities mentioned in Section 3.3.3, the weighted OLS is compared to unweighted estimation.

3.4.2 *Objectives of the Study*

The predictive power of various GVF models is compared to the diret HT variance estimator given in (3.3a), and replicate weights. Replicate

weights are matrices in $\mathbb{R}_0^{+n\times B}$ and each column $b = 1,\dots,B$ summarizes a realization of a pre-specified resampling algorithm such as mentioned in Section 3.1. Here, the replicate weights are generated using the `R` package `survey` [Lumley, 2019]. For the GVF model estimation, though, the direct HT variance estimate is used and not the resampling variance estimate, although this would have been possible, too. All variance estimators/ predictors are considered under the criteria of Monte-Carlo relative bias (3.30) and relative MSE.

The relative bias is the expectation of

$$\text{RelErr}\left[\text{Var}'_{\mathcal{M}_\theta,D}\left[g_S^{(j)}(S,Y^{(j)})\right]\right] \triangleq \frac{\widetilde{\text{Var}}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}(S,Y^{(j)})\right]}{\text{Var}_{\mathcal{M}_\theta,D}\left[g_S^{(j)}(S,Y^{(j)})\right]} - 1 \tag{3.30}$$

where $\text{Var}'_{\mathcal{M}_\theta,D} \in \{\widetilde{\text{Var}}_{\mathcal{M}_\theta,D}, \widehat{\text{Var}}_D, \widehat{\text{Var}}_D^{\text{rw}}\}$ and superscript rw stands for replicate weight estimator. The relative MSE is then the expectation of the squared (3.30) expression.

In addition, the R^2 coefficient is stored (when applicable) to analyse its property as quality measure. Finally, the quality measure $\bar{\delta}_{1-\alpha}$ of Gershunskaya and Dorfman [2013] is calculated with $\alpha = 0.05$ and stored. Here, the question is whether the asymptotics already hold under the three different survey designs and allow the use of $\bar{\delta}_{1-\alpha}$ to evaluate and choose GVFs.

3.4.3 *Simulation Results*

For StratRS and the stratified TSC, some clusters and strata must be put together when sample sizes therein equal 1. Whilst this has no notable impact for StratRS, this leads to an overestimation of variance for stratified TSC as can be seen in Figure B.1 in Appendix B, where the MC distribution of the direct variance estimators are contrasted with the MC variance of the HTs.

The number of models in combination with the choice whether to include indicator variables or whether to use weights in OLS is too high to be exposed in a single figure. Thus, the GVF shapes will first be contrasted to each other in terms of the MC mean of the relative error. The best GVF models are then compared to the classical HT variance estimator and the resampling estimator.

The complete summary statistics are given in Tables B.4 to B.15 in Appendix B. The MC mean of (3.30) results when both the numerator and the denominator are MC outcomes: As the finite population is known in each simulation run, the true HT variances can be estimated for each generated finite population (and also the GVF for the generated sample from the finite population). Averaging over the finite populations yields an estimator for $\text{E}_{\mathcal{M}_\theta,D}\left[\text{Var}_D\left[\hat{\tau}_y\right]\right]$ (or of the GVF prediction respectively).

The nonlinear regression (3.23c) encountered several problems, especially convergence to an optimal parameter estimate was seldomly achieved when indicator variables were included as explanatories, which led to the exclusion of Model (3.23c) with indicator variables. But also when all variables of interest were merged in the regression model, the estimation sometimes would not always converge with arbitrary starting values given certain realizations $S = s$. If a GVF of that shape is used in practice, the estimation thus requires additional fine tuning that cannot be provided in a routine running 1500 times.

The other classical linear regression models (3.23a), (3.23b) and (3.23d) performed, on the other hand, very similar in terms of average relative error. This can be explained as follows: These shapes differ mainly on the underlying assumption of the error distribution of ε_j, $j \in \Omega$. However, potential heteroskedasticity in the linear model (that might be compensated using (3.23a)) does not affect neither the point estimation nor the prediction. The possible bias due to log-transformation is accounted for (cf. Section 3.3.1).

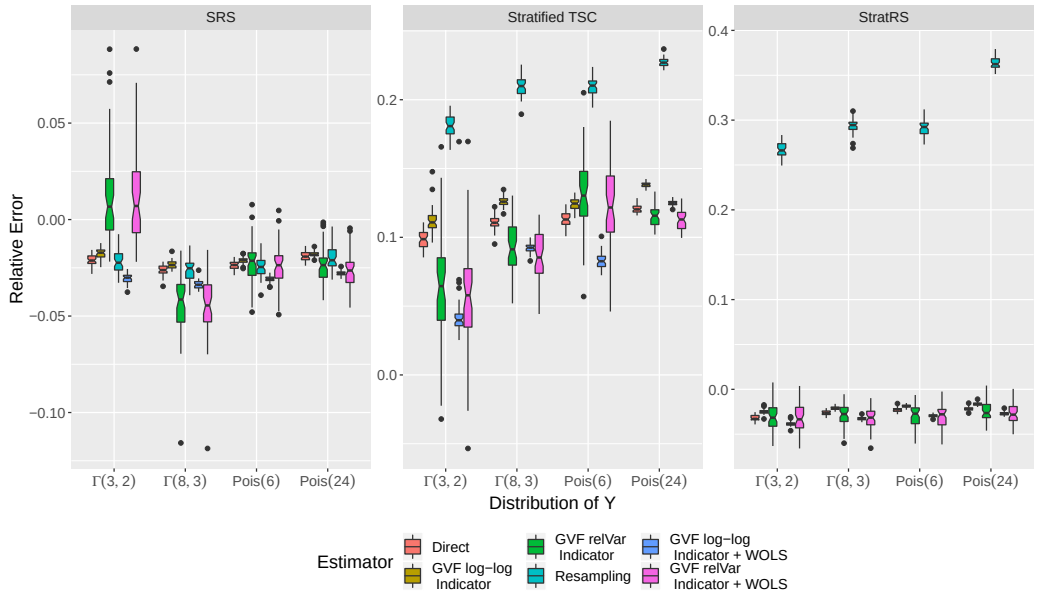
Concerning the average squared relative error, though, the log-log regression slightly outperforms Model (3.23a) and both of them have slightly lower squared error than the linear model, which can be checked in the electronic appendix.

Furthermore, we find in general that the GVFs are sensitive to the violation of assumptions, that is, when no indicator variables are included to account for the different DGPs of the original data. Additionally, weighting does not reduce (squared) prediction error for SRS and StratRS in average. In fact, there can be observed a slight deterioration in the median relative error in Figure 3.1 and Tables B.4, B.5, B.7 and B.8. For these designs, Figure B.1 shows also that the direct variance estimator (3.3a) returns very reliable estimates (i.e. a low relative error). For stratified TSC, however, the direct estimator yields higher errors due to the merging of strata. Consequently, downweighting high relative variance variables $j \in \tilde{\Omega}$ in weighted OLS yields a better regression model and thus better GVF predictions in average. Furthermore

In a next step, the GVF predictions of Models (3.23a) and (3.23b) – with indicator variables and estimated by weighted OLS – are compared to the performance of the direct estimator and the resampling estimator. The MC distribution of the relative error is plotted in Figure 3.1

Especially for the stratified designs, the resampling estimator perform surprisingly bad. This might be due to the fact that the finite population correction is quite significant in this set-up, and the bootstrap chosen for the generation of replicate weights cannot appropriately account for that. The fact that the bias is positive for the stratified designs is an argument for this hypothesis. On the other hand, alternative replicate weight methods in `survey` were not applicable to the design as only two units per

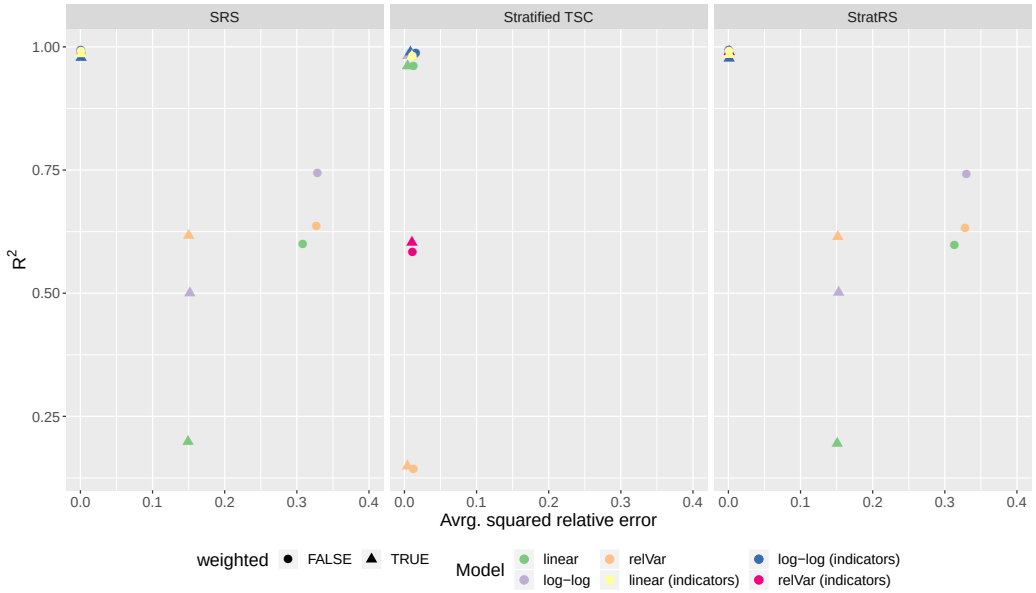
Figure 3.1: MC Relative Error of Variance Estimators



stratum were chosen in the stratified TSC and also for StratRS, sometimes only one unit per stratum was chosen. The log-log GVF performs quite similar to the direct estimator, there is only a slight bias in the prediction and variance of the prediction error is very small. The regression using the relative variance as endogeneous variable, on the other hand, yields predictions with higher variability. Though often comparable to the log-log GVF in average, the spread of prediction errors is larger, which, on the other hand, makes it more often cover the real HT model-design variance within the inter-quartile range. Furthermore, it is more easily theoretically justifiable. If the model assumptions are fulfilled, Figure 3.1 suggests that GVFs under a wisely chosen shape represent an alternative to the direct variance estimator, not only concerning the computational effort but also the prediction efficiency.

But the possible gain in efficiency, is hardly an advantage if there is no measure that allows to assess whether the chosen GVF shape yields such good predictions. Figure 3.2 plots the average share of explained variation (R^2) of the GVF regression of all GVF models against the average relative squared prediction error. Obviously, the previously named models that are considered best – the log-log regression and the regression using the relative variance, estimated by (weighted) OLS and accounting by indicator variables for the different distributions of the input variables – yield only partially high R^2 close to one: The model with the relative variance as dependent variable, Model (3.23a), has for both weighted and unweighted regression a medium R^2 for stratified TSC. On the other hand, the linear models always have high R^2, for stratified TSC this holds

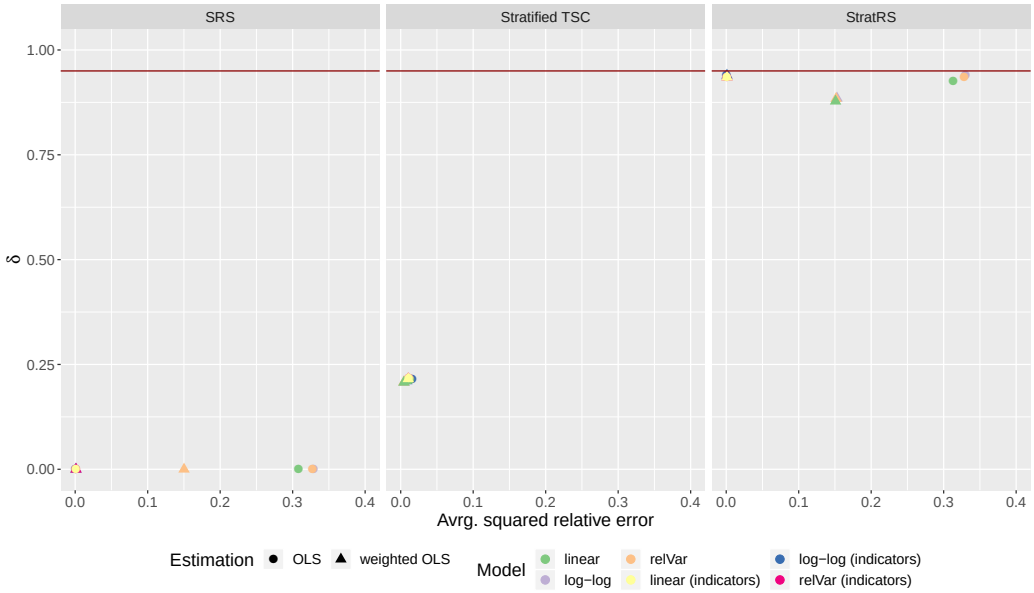
Figure 3.2: Average R^2 of the GVF Regression vs. Average Relative Prediction Error



even for the model without indicator variables. The distinction between good and bad shape decisions is thus not clear-cut. Note however, that this mainly holds for the stratified TSC. Under the simpler designs, the R^2 is capable to find the importance of indicator variables. Consequently, the share of explained variation gives an indication for good GVF shapes but is not always a good advice, especially when the designs become complexer.

Completely different results can be found for $\bar{\delta}_{0.95}$ suggested by Gershunskaya and Dorfman [2013], summarized in Figure 3.3. Calculating this indicator with the direct variance estimates yields a 0% and roughly 21% coverage with a nominal level of 95% for SRS and stratified TSC respectively which can be seen as an indication that the asymptotics may not yet hold for these survey designs. When the asymptotics do not hold, it is consequently not surprising that the average $\bar{\delta}_{0.95}$ are low although GVF predictions can be good. It is then interesting that the $\bar{\delta}_{0.95}$ of the different GVF predictions yield similar values like when the direct estimator (3.3a) is plugged in for SRS and stratified TSC, which are independent from the average relative MSE and not close to the nominal level of 0.95. For StratRS, on the other hand, where the asymptotics seem to hold ($\bar{\delta}_{0.95}$ is close to the nominal level when the direct HT variance estimator is used for the computation), the quality measure also achieves the nominal level for the GVFs. But the MC average $\bar{\delta}_{0.95}$ values are similar for all GVF shapes and therefore cannot help to choose between the shapes and weighted and unweighted estimation. Especially

Figure 3.3: Average $\bar{\delta}_{0.95}$ of the GVF Regression vs. Average Relative Prediction Error



for the logarithmic regression, the indicator $\bar{\delta}_{0.95}$ is not capable to find that the lack of indicator variables is inadequate. Nonetheless, it indicates that the linear model and Model (3.23a) without indicators is not appropriate. Nonetheless, it seems that – in contrast to R^2 – one can determine a priori whether $\bar{\delta}_{0.95}$ is usable by plugging the square root of the direct HT variance estimator (3.3a) into (3.26). If the nominal level is not achieved for the direct estimator, one cannot rely on the result of the quality measure for GVFs.

To summarize the simulation study, the GVF predictions are a stable and reliable alternative to direct variance estimation if the model assumptions hold, that is, when the input HT are generated from identically distributed random variables. Then, barely a difference between the shapes can be found, though the logarithmic regression and the regression using relative variances were slightly on top. When the survey design is easy, weighted OLS does not help in the estimation of GVFs as the input data are already reliable and the weights are generated using estimates. However, under stratified TSC, regression weighting helps to improve estimation, especially when the variances of the underlying DGP are large.

4

A MULTIVARIATE PROBABILITY DENSITY ESTIMATOR USING SOMS

4.1 INTRODUCTION TO SOMS

Both, Chapter 2 and 3 dealt not only with classical problems of applied statistics, but also solved them using classical approaches. Chapter 4 breaks this pattern: Though the problem of estimating unknown probabilities (non-parametrically) is old, the offered solution belongs to the upcoming field of machine learning. Therefore, Chapter 4 can be seen like a synthesis of classical statistical estimation theory and novel, computer-intensive methods that still lack sometimes a thorough theoretical foundation. This reconciliation is demonstrated here on the Self-organizing Map.

The Self-organizing Map (SOM) originates from Kohonen [1982] and was intended to mimic 'input-driven self organization' of brain cells to specialize for different cognitive functions [Kohonen, 2013]. Notwithstanding this originally biological motivation, SOMs are currently applied in exploratory data analysis, text structure analysis [Kohonen, 2013] and data visualization [Kangas and Kohonen, 1996]. They have been applied to impute survey data as well [Fessant and Midenet, 2002], which raises the question of how survey data should be treated in machine learning in general and in SOMs in particular when data involve a complex survey design P_D (cf. Section 4.7).

The SOM Algorithm is a competitive learning algorithm that aims at structuring the input data in an unsupervised manner. Like the k-means, SOMs aim at the identification of 'hidden' clusters in the data, that is, output sub-spaces in which random realizations of Y resemble more another than between categories. In contrast to usual vector quantization methods, though, SOMs have the advantage that the vector quantizers are arranged on an undirected graph $\mathcal{G} = (V_G, E_G)$ and the edges between indexed units indicate neighborhoods in the input data space [Alahakoon et al., 2000]. This gives additional information on the input data space than the mere assignment of Y to vector quantizers.

This makes the algorithm especially interesting when too few information about the data is available to adopt parametric distributional assumptions, i.e. the concept $Y \sim P_{\mathcal{M}_\theta}$ and $P_{\mathcal{M}_\theta} \in \left\{P_{\mathcal{M}_\vartheta} : \vartheta \in \Theta\right\}$: The algorithm does not only find representatives that summarize local information, but also arranges this local information within neighborhoods, which raises the question on the relation between the graph's topology

and that of the input data space. Kangas and Kohonen [1996] name the SOM a 'kind of nonlinear projection of a probability density function of high-dimensional input data onto a two-dimensional array'. If the information that gets lost on this projection consists mainly of noise, reconstruction of the high-dimensional probability distribution function could be feasible. The relatively efficient storage and easy training of SOMs then would turn the algorithm to an attractive instrument for multivariate probability distribution estimation. In this Chapter, SOMs are studied under this point of view.

Especially from Section 4.4 on, the focus turns to modifications of the original SOM algorithm that are favorable for probability estimation. We will present an estimator similar to the semi-parametric method of Gaussian mixtures. In fact, the provided estimator comes in the shape of a density estimator, thus at least requiring that the probability measure P is approximatable by a Lebesgue density.

In the following, we introduce the basic SOM algorithm and a batch variant thereof as well as important findings and shortcomings in the asymptotic behaviour, which might hinder its application in probability distribution estimation. Then, extensions of the algorithm are discussed that shall remedy these pitfalls.

4.2 THE BASIC SOM

4.2.1 *Algorithmic Description*

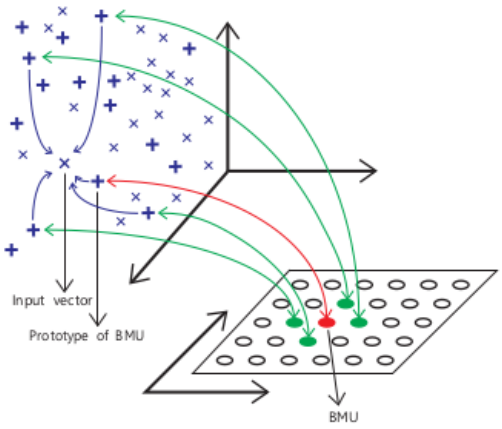
First, additional notation for this chapter must be introduced. For this introductory section, assume that the graph $\mathcal{G} = (V_G, E_G)$ is a finite grid on $\mathbb{Z}^k$ where k is usually a small positive integer, most often $k = 2$. The size of the grid is $|V_G| = N_G$ and the set E_G denotes the edges between to nodes $(m_1, \dots, m_l, \dots, m_k)$ and $(m_1, \dots, m_l \pm 1, \dots, m_k)$.

Let again $Y_i \in \mathcal{Y}$ be a p-dimensional real-valued random variable and $Y := (Y_1, \dots, Y_N)$ where N is the size of a finite index set U. Assume that Y_i are iid for $i \in U$ and their probability law P is of interest. As the generating model $\mathcal{M}$ is unknown, we omit here the subscript $\mathcal{M}_\theta$ for the DGP.

To each node j of the grid $\mathcal{G}$ on which a SOM for Y shall be trained, there is attributed in each training step $\kappa = 0, 1, \dots$ a feature vector $\mathbf{u}_j^{(\kappa)} \in \mathbb{R}^p$, and the ensemble be $U^{(\kappa)} = \left\{\mathbf{u}_j^{(\kappa)}\right\}_{j \in V_G}$. In each training step, a training data point $\mathbf{y}_i$, $i \in S$, that is considered to be a realization of Y_i, is presented to the elements of U and the best matching unit (bmu)

$$b(\cdot, U^{(\kappa)}) : \mathbb{R}^p \to \mathcal{G}, \quad b(\mathbf{y}_i, U^{(\kappa)}) = \arg\min_{j \in V_G} d\left(\mathbf{y}_i, \mathbf{u}_j^{(\kappa)}\right) \tag{4.1}$$

Figure 4.1: Concept of the SOM



Source: Hajjar and Hamdan [2011]

that is the node whose feature vector has minimal distance d to $\mathbf{y}_i$, and its neighbors are updated to better resemble the training data. Neighborhood of $j \in V_G$ is here defined by the set of nodes $\iota \in V_G$ for which the distance $\delta(j, \iota)$ is below a certain level. The distance is often the Eucledian distance [Kangas and Kohonen, 1996] but could also correspond to the length of the shortest path. The shortest path will prove to be a useful metric when we relate SOMs to MRFs. When $V_G \subset \mathbb{Z}^k$, the shortest path metric corresponds to the ℓ_1 metric.

The vectors' update is dynamic in the iteration step κ due to a learning rate $\alpha > 0$ and depends additionally on the neighborhood function $h \geqslant 0$ that incorporates the distance between the trained node ι and the Best Matching Unit (bmu) $\delta\left(\iota, b(\mathbf{y}_i, U^{(\kappa)})\right)$. Note that b is not necessarily well defined as it might happen that more than one feature vector has minimal distance. Implications thereof will be discussed in the next sections and the phenomenon is in this introductory section disregarded. This procedure, already described in words, is summarized in Algorithm 4.1.

Also Figure 4.1 illustrates the working of the SOM: There, data from a 3-dimensonal input data space that will be mapped nonlinearly to points in a two-dimensional grid. To each point in the grid, a feature vector is attributed indicated by the green and red colored arrows. A nonlinear map then assigns input data close to a feature vector to the corresponding element of the grid. The learning rule (4.2) trains at the same time feature vectors that belong to neighboured points in the grid such that those feature vectors shall become similar. In that sense, using eucledian distance, the feature vectors in Figure 4.1 do not yet seem well trained.

Algorithm 4.1 On-line SOM

Require: $\{\mathbf{y}_i\}_{i\in S}$, $\delta : \mathcal{G} \to [0,\infty)$, $d : \mathbb{R}^p \to [0,\infty)$, $\mathcal{G} \subset \mathbb{Z}^k$, $U^{(0)}$, $h : [0,\infty) \to [0,1]$, $\alpha : \mathbb{N}_0 \to [0,1]$

Ensure: A SOM that is completely described by $(U, \mathcal{G})$

for $\kappa = 1, 2, \ldots, K$ **do**

Draw randomly $i \in S$

Find the bmu $b_\kappa(\mathbf{y}_i) := b\left(\mathbf{y}_i, U^{(\kappa)}\right)$

Update for each $j \in V_G$

$$\mathbf{u}_j^{(\kappa)} = \mathbf{u}_j^{(\kappa-1)} + h\left(\delta\left(j, b_\kappa(\mathbf{y}_i)\right)\right) \cdot \alpha(\kappa) \cdot \left(\mathbf{y}_i - \mathbf{u}_j^{(\kappa-1)}\right) \tag{4.2}$$

if Convergence **then**

break

end if

end for

Set $U = \{\mathbf{u}_j\}_{j\in V_G} \leftarrow \left\{\mathbf{u}_j^{\kappa}\right\}_{j\in V_G}$

Note that Algorithm 4.1 is also applicable for a discrete but permanent input data stream, i.e. in each training step κ, a random realization $\mathbf{y}_\kappa$ could be presented to the algorithm. In accordance with the survey statistical context and the subsequent algorithms that require a finite training data set, however, we referred to a training data set S.

A priori, it is not known whether Algorithm 4.1 converges. In fact, convergence depends on the properties of P, h, α and the input data dimension p, and will be discussed in the next section.

The difference between the SOM and classical vector quantization is that in general, the neighborhood function has a support of more than one point, $h \not\equiv \mathbb{1}_{\{0\}}$ [Fort et al., 2001, Kohonen, 2013]. The neighborhood function h is a reinforcement rule, meaning that the adaption is bigger for nodes closer to $b(\mathbf{y}_i, U^{(\kappa)})$, yielding (approximate) neighborhood preservation, which is also often referred to as (approximate) topology-preservation. Though the Gaussian kernel is a common choice for h [Tolat, 1990, Kohonen, 2013, for example].

An alternative to the described on-line algorithm are batch-type variants of the SOM, and the one internally linked to Algorithm 4.1 is described in Algorithm 4.2 and can be found in Fort et al. [2001]. The batch algorithm is derived from an equilibrium state that the on-line algorithm reaches if convergent. Fort et al. [2001] argues that the Batch-SOM would also converge when Algorithm 4.1 converges (which still has to be discussed), the argument is given in the Appendix A, including some theoretical aspects that are missing in Fort et al. [2001]. It is important that Fort et al. [2001] states not that the equilibria agree.

Algorithm 4.2 Batch SOM

Require: $\{\mathbf{y}_i\}_{i\in S}$, $\delta : \mathcal{G} \to [0,\infty)$, $d : \mathbb{R}^p \to [0,\infty)$, $\mathcal{G} \subset \mathbb{Z}^k$, $\{\mathbf{u}_j^{(0)}\}_{j\in\mathcal{G}}$, $h : [0,\infty) \to [0,1]$
Ensure: SOM $\{\mathbf{u}_j\}_{j\in V_G}$
for $\kappa = 1, 2, \ldots, K$ **do**
Determine best matching units for all $i \in S$
Update each $j \in V_G$ by

$$\mathbf{u}_j^{(\kappa)} = \frac{\sum_{\iota\in V_G}\sum_{i\in S} h\left(\delta\left(b(\mathbf{y}_i), j\right)\right)\mathbf{y}_i}{\sum_{\iota\in V_G}\sum_{i\in S} h\left(\delta\left(b(\mathbf{y}_i), j\right)\right)} \tag{4.3}$$

end for

This is due to the fact that the learning rule (4.2) is not transferable to a globally convex energy function [Heskes, 1999] and thus, there are many local minima possible and the converged SOM need not agree with the Batch Self-organizing Map (BSOM). Fort et al. [2001] and Fort et al. [2002] observe a better ordering of the on-line variant because the stochasticity allows to leave local minima. However, as in each learning step, only one observation is presented to the SOM, learning is slower for the on-line algorithm when a given set S of observations exists and no additional information arrives during the training process [Fort et al., 2001].

The definition of the bmu in (4.1) – which is commonly used in vector quantization – is closely related to the concept of Voronoi tesselation [Kohonen, 2013]. The Voronoi region of a node j is defined as

$$V_j \triangleq \{\mathbf{x} \in \mathbb{R}^p : b(\mathbf{x}, U) = j\} \quad , \tag{4.4}$$

which depends on U and therefore varies for each training step κ. In fact, if the Algorithm 4.1 converges, it can be seen as a local minimization of the approximate energy function

$$L \triangleq \frac{1}{2}\sum_{j\in V_G} h\left(\delta\left(j, b(Y, U)\right)\right)\int_{V_j}\|Y - \mathbf{u}_j\|^2 \, dP \quad , \tag{4.5}$$

implying that the employed metric is the Eucledian distance. Remember that the Voronoi region V_j is a function of all feature vectors $\mathbf{u}_\iota$, $\iota \in V_G$, too, due to its definition via b. Thus, a correct differentiation of L with respect to $\mathbf{u}_\iota$ is not straight forward [Heskes, 1999] and the update rule is only an approximate gradient descent step.

Besides the choice between batch and on-line training, Algorithms 4.1 and 4.2 raise several questions that shall be picked up in the next sections: First, there are several possible choices for the neighborhood function (though the Gaussian kernel is very common, Kangas and Kohonen [1996], Kohonen [2013, for example]). However, in order to prove convergence, often a finite support is necessary [Sadeghi, 2001]. In the case

of the on-line algorithm, the learning rate needs to be determined, too. Yin and Allinson [1995] give some necessary properties concerning the decay rate of α for another proof of convergence. Sadeghi [2001], on the other hand, requires the learning rate to be constant for his proof for convergence in distribution.

In addition, the metrics d and δ and the size of the grid, N_G must be chosen prior to training. The first problem, the choice of metrics δ and d, can play a role in the training as they determine the neighborhoods that will be updated. Update rules (4.2) and (4.3) assume implicitly that d is the squared Eucledian distance. Otherwise, the update direction $\left(Y_i - \mathbf{u}_j^{(\kappa)}\right)$ must be replaced by $\frac{1}{2}\nabla_y d\left(Y_i, \mathbf{u}_j^{(\kappa)}\right)$ [Tolat, 1990]. As the chosen metrics play a major role for the interpretation of SOMs with alternative update rules, common metrics used for SOMs are discussed in Section 4.3. The second problem is discussed in Section 4.5, and the suggested dynamic solution is the main remedy to control magnification.

4.2.2 *Properties of the SOM*

There are two aspects when talking about (asymptotic) properties of SOMs. As we consider in the following finished algorithms, we leave the iteration superscript κ. First, the organization of the map, i.e. convergence related to its topology, which stems from $h \neq \mathbb{1}_{\{0\}}$. Kohonen [1982] and Tolat [1990] denote this organization a 'topologically correct' mapping. An ordered state can be stated at most when $p = k$, that is the dimension of input and output space are equal [Tolat, 1990, Erwin et al., 1992]. Tolat [1990] also names the option that $p > k$ but Y_i lies on a k-dimensional subspace. But already for $p = 2$, when an ordered state is achieved, the exit probability can be positive for particular neighborhood definitions [Cottrell et al., 2016]. When $p = k = 1$, however, the convergence to an ordered state can almost surely be assured. If $p > k$, the neighborhood preservation can only hold approximately. These results hold for lattices $\mathcal{G}$ and are not studied for general graphs.

Second, the vector quantization quality may be assessed by the average quantization error, that is for Eucledian distance, the average squared deviation of inputs within the Voronoi regions V_ι from the centers $\mathbf{u}_j$, $j, \iota \in V_G$,

$$\mathrm{E}\left[\mathrm{dist}(Y, U)\right] = \sum_{j \in V_G} \int_{V_j} d(Y, \mathbf{u}_j)\, \mathrm{d}\, P \quad . \tag{4.6}$$

When the neighborhood function has a compact support, i.e. when $h(\delta(j, \iota)) = 0$ if $\delta(j, \iota) > c$, $c \geqslant 0$, the expected loss function

$$\sum_{j \in V_G} \sum_{\iota \in V_G} h(\delta(j, \iota)) \int_{V_j} d(Y, \mathbf{u}_\iota)\, \mathrm{d}\, P \quad . \tag{4.7}$$

reduces to a weighted error between Y_i and $\mathbf{u}_\iota$ in the joint Voronoi region $\cup_{\iota \in \mathcal{N}_j} V_\iota$, and therefore its minimization contributes to the reduction of the quantization error (though the optimization problems are not equivalent).

Vesanto et al. [2003] call (4.7) the distortion measure and gives a decomposition of it into a sort of quantization error and a disorganization penalty. They argue that this reflects the trade-off between finding good representative feature vectors and the similarity of feature vectors within a neighborhood. We shall see a similar decomposition in Section 4.4. This trade-off is reflected by $h \not\equiv \mathbb{1}_{\{0\}}$ and its implication for density estimation is discussed in Heskes [2001] and will be picked up later. In the following, we will focus on convergence concerning the representative feature vectors U.

Rarely for SOMs, statements can be found for general P, α, d, δ, h, p and k. Due to the increasing complexity, early analyses were restricted to the case where $p = k = 1$ [Erwin et al., 1992]. Yin and Allinson [1995] and Sadeghi [2001] give some limited results for the higher dimensional input space.

Sadeghi [2001] shows that the on-line learning is an aperiodic T-chain and using Markov chain theory, the convergence of the feature vectors distribution to a finite invariant measure is demonstrated that depends on the constant learning rate α (which is in contrast to the convergence results of Yin and Allinson [1995]). Like in Yin and Allinson [1995], it is shown that a Central Limit Theorem (CLT) holds for the feature vectors. However, the findings are only valid for a neighborhood function h with finite support and $\operatorname{supp} Y_i$ must be bounded, Y_i being piecewise continuous on (multidimensional) intervals. In that case, elements in U take values in a bounded and convex set that is a superset of $\operatorname{supp} Y_i$. Besides these restrictions, these results are on properties of U that do not depend on P (except for the named conditions). Therefore, these results are only limitedly useful for probability estimation of P.

Yin and Allinson [1995], on the other hand, rewrite the on-line algorithm as a sum of (time dependently) weighted independent random variables (which is in disagreement with our assumption that $i \in S$ and therefore, draws in training steps are not independent). Under the assumption of not too fast decaying learning rates and h being a step function $\mathbb{1}_{\mathcal{N}_j}$ with shrinking neighborhood size $\mathcal{N}_j$ for each $j \in V_G$, they manage to show L^2-convergence of $\{\mathbf{u}_j\}_{j \in V_G}$ to the conditional expectations of Y_i if there are no zero-hit nodes. There are no explicit requirements on the properties of P and Y_i, besides the existence of all moments of Y_i. Note that some steps in their proof remain unclear, as discussed in Appendix A.

If the input variables are discrete, though, it can be shown that the loss function L in fact is an energy function for the update rule (4.2).

This is not the case for continuous random variables, though [Heskes, 1999]. The existence of a global energy function simplifies the analysis of the learning rule because then, the update can be interpreted as a (local) gradient descent step and known (local) convergence results become applicable.

As already mentioned, the convergence might depend on the choice between on-line and batch algorithm. The convergence of the on-line variant motivates the update rule in the BSOM. If the on-line algorithm converges in L^2, none of the feature vectors $\mathbf{u}_j$, $j \in \mathcal{G}$ would change in expectation, which can be translated into

$$\begin{aligned} \forall j \in V_G \quad & \sum_{\iota \in V_G} \int_{V_\iota} h\left(\delta(b(Y,U),j)\right)\left(Y-\mathbf{u}_j\right)\, dP \\ & = \sum_{\iota \in V_G} h(\delta(\iota,j)) \int_{V_\iota} (Y-\mathbf{u}_j)\, dP = 0 \end{aligned} \tag{4.8}$$

[Fort et al., 2001]. Fort et al. [2001] argues that a rearrangement yields (cf. Equation (A.9))

$$\begin{aligned} \mathbf{u}_j &= \frac{\sum_{\iota \in V_G} h\left(\delta(j,\iota)\right) \int_{V_\iota} Y\, dP}{\sum_{\iota \in V_G} h\left(\delta(j,\iota)\right) P(Y \in V_\iota)} \\ &= \frac{\sum_{\iota \in V_G} h\left(\delta(j,\iota)\right) P(Y \in V_\iota) E[Y|Y \in V_\iota]}{\sum_{\iota \in V_G} h\left(\delta(j,\iota)\right) P(Y \in V_\iota)} \quad , \end{aligned} \tag{4.9}$$

which leads to the empirical update rule (4.3) when empirical means are used

$$\begin{aligned} & \frac{\sum_{\iota \in V_G} h(\delta(j,\iota)) \sum_{i \in S} \mathbb{1}_{V_\iota}(Y_i) \cdot Y_i}{\sum_{\iota \in V_G} h(\delta(j,\iota)) \sum_{i \in S} \mathbb{1}_{V_\iota}(Y_i)} \\ = & \frac{\sum_{i \in S} \sum_{\iota \in V_G} \mathbb{1}_{V_\iota}(Y_i) \cdot h\left(\delta(j, b(Y_i,U))\right) \cdot Y_i}{\sum_{i \in S} \mathbb{1}_{V_\iota}(Y_i) \cdot h\left(\delta(j, b(Y_i,U))\right)} \quad . \end{aligned}$$

Note that this rearrangement assumes that $\mathbf{u}_j$ does not depend on P, which means pointwise convergence of $\mathbf{u}_j$ to a constant function. A correct re-formulation of the converged state in Fort et al. [2001] thus would be

$$\begin{aligned} & \sum_{\iota \in \mathcal{G}} h\left(\delta(j,\iota)\right) \cdot P(Y \in V_\iota) \cdot E\left[\mathbf{u}_j | Y \in V_\iota\right] \\ = & \sum_{\iota \in \mathcal{G}} h\left(\delta(j,\iota)\right) \cdot P(Y \in V_\iota) \cdot E\left[Y | Y \in V_\iota\right] \end{aligned} \tag{4.10}$$

and when Y is presented to the SOM after training, $E\left[\mathbf{u}_j | Y \in V_\iota\right] = E\left[\mathbf{u}_j\right]$. This implies that in the converged state, we have in fact that the expected feature vectors correspond to a locally re-weighted expectation of Y and convergence means L^1-convergence.

A link that seems not to have been made up to now is the identification of SOMs with Markov Random Field (MRF), though the process of weight updates was identified with Markov processes. When the neighborhood function is chosen such that supp h is compact, the impact of a random outcome of Y_i is locally bounded on the graph.

Then, possibly a joint probability measure on the σ-field over $\mathcal{G} \times \mathcal{Y}$ can be constructed that reflects the joint distribution of $\mathbb{1}_{V_j}(Y)$ and Y under certain conditions. Therefore, we must bring into accordance the terms used here with those of MRF theory. We start with the definition of *neighorhood* relations and *cliques*.

Definition 10 (Neighborhood). *On a graph $\mathcal{G} = (V_G, E_G)$ endowed with metric δ, a* c-neighborhood *($c > 0$) of a node j, $\mathcal{N}_j$, is defined as the union of nodes ι in V_G with $\delta(j, \iota) \leqslant c$*

$$\mathcal{N}_j := \{\iota \in V_G : \delta(j, \iota) \leqslant c\} \quad .$$

If $\iota \in \mathcal{N}_j$, we have an equivalence relation and write $\iota \sim j$.

When the feature vectors $\mathbf{u}_j$ are clearly identified with their corresponding nodes, fixed U are sites on the graph $\mathcal{G}$ with proximities given by the neighborhoods in $\mathcal{G}$.

Definition 11 (Clique). *Assume a finite set $\mathcal{G}$ and a measurable, symmetric and reflexive relation $\sim$ between elements $x, y \in \mathcal{G}$. A subset $\Gamma \subset \mathcal{G}$ is called a* clique *if $\forall x \in \Gamma : \ \forall y \in \Gamma : \ x \sim y$.*

For $c < \operatorname{diam}(\mathcal{G})$, we have pairwise different cliques. If we can then identify the distortion measure (4.7) with a nearest-neighbor potential with sites U, neighborhood relation $\sim$ and configurations Y there exists a Markov Random Field (MRF) (under further conditions to be specified later) [Isham, 1981]. Then, there exists a function f such that $P(Y|V_j) = f(\mathbf{u}_\iota : \iota \in \mathcal{N}_j)$ [Isham, 1981]. Difficulties arise as the configuration space $\mathcal{Y}$ is not necessarily finite and discrete.

4.3 DISTANCE MEASURES IN SOMS

The choice of δ determines the proximity of nodes in the output data space and in combination with the neighborhood function h, the metric on $\mathcal{G}$ is consequently essential for the training of the feature vectors. That the update depends on δ via $h \circ \delta$ makes the difference between SOMs and Neural Gas algorithms [Villmann and Claussen, 2006, Arnonkijpanich et al., 2008]. The dependence of the update rules (4.2) and (4.3) on δ impacts the problem of magnification error: As two nodes in $\mathcal{G}$ may be neighbored with a distance disproportional to their distance in the input space, the learning of their feature vectors on a training data

Y is not proportional to the original distance neither. This can lead to an over-representation of low-density areas by relatively many feature vectors [Ritter and Schulten, 1986, Villmann and Claussen, 2006].

The distance plugged into h is often the absolute distance between nodes j and ι in the two-dimensional space. This is in so far sensible as the arrangement of $\mathcal{G}$ is usually on a grid with edges linking those units j and ι with $|j - \iota| = 1$. However, note that there is no necessity to arrange the nodes in $\mathbb{Z}$. Therefore, it makes also sense to define δ based on the existence of edges between two nodes j and ι without focusing on the absolute length of the edge.

In Section 4.5, a dynamic SOM introduced by Fritzke [1994] will be discussed. As in this case a low-dimensional simplex defines a local neighborhood, the metric between j and ι can be defined as the minimal path length over the edges in E_G

$$\delta(j, \iota) \triangleq \min |\rho(j, \iota)| \quad , \tag{4.11}$$

which is also applicable on lattices. Alternatively, when $V_G \subset \mathbb{Z}^k$, the Eucledian distance might be used. This has the inconvenient that there exists a difference between a lattice in $\mathbb{Z}^k$ or, say $(2\mathbb{Z})^k$, as the edge length doubles though the relative position of nodes in V_G can still be the same. Output space metric (4.11) allows us to find

Remark 7. *In a two-dimensional, finite lattice $\mathcal{G}$ with $j = (j_1, j_2)$, $m = (m_1, m_2) \in V_G$ where $j \sim m$ if $(j_1 \pm 1, j_2) = (m_1, m_2)$ or $(j_1, j_2 \pm 1) = (m_1, m_2)$, all cliques in the graph are subsets $\Gamma \subset \mathcal{G}$ with $|\Gamma| \leqslant 2$ [Isham, 1981].*

For the SOM's input data space – that is, the image space of Y – and the corresponding metric d, on the other hand, things are different. Though the most often used metric is the (squared) Eucledian [Kohonen, 2013], this choice is not unproblematic: The squared Eucledian distance is not scale invariant, $d(\lambda \mathbf{y}_1, \lambda \mathbf{y}_2) = \lambda^4 d(\mathbf{y}_1, \mathbf{y}_2)$ for some $\lambda \in \mathbb{R}$. This might be critical, for possibly varying input measures (say cm or m). Especially when Y is p-variate with $p > 1$, different scales of vector elements in $Y = (Y_1, \ldots, Y_p)$ impact their influence on the training process [Pacifico and de Carvalho, 2011]. When the training is then stopped too early, adaptation of the feature vectors U to low-scale elements in Y might not be sufficient because training took first place for those vector elements with large scale. Therefore, sometimes the the input data is normalized. Blayo [1992], however, questions this common practice in the case that $p \to \infty$. For fixed p, on the other hand, his arguments are not valid and the normalization can be useful. If the data are normalized, an alternative and frequently used distance is the dot product, that is, the angle between Y and the elements of U in the input space [Tolat, 1990, Blayo, 1992, Kohonen, 2013], which simplifies calculus.

If the realizations of Y lie on a curved manifold with a dimension smaller than p, on the other hand, the Eucledian distance does not return the distance between these two points on the manifold. In addition, a mapping onto a finite subset of $\mathbb{Z}^k$ cannot be achieved, what yields the approximate neighborhood preservation discussed previously. Then, d should be the natural metric on that manifold [Ritter, 1999]. However, if the data analysis is explorative and barely something is known about P, unusual choices for the metric d are difficult to justify.

Notwithstanding the problems with the Eucledian distance, (squared) metrics based on scalar products in $\mathbb{R}^p$ ($d(\mathbf{y}_1, \mathbf{y}_2) = \|\mathbf{y}_1, \mathbf{y}_2\|_\Lambda^2 \triangleq \langle \mathbf{y}_1 - \mathbf{y}_2, \mathbf{y}_1 - \mathbf{y}_2 \rangle_\Lambda$ where Λ is a symmetric, positive definite matrix) have a good statistical interpretation such as the Mahalanobis distance. Therefore, in the following, we will focus on generalizations of the Eucledian distance.

A metric that corrects for the scaling problem of random variables is the Mahalanobis distance with the inverse covariance matrix of Y, say $\Sigma^{-1} = \Lambda$ in the previous description. This has been used in SOM training as well [Paul and Gupta, 2013, Hajjar and Hamdan, 2011]. If the normalization of the input data is $\Sigma^{-1/2}Y$, the normalization corresponds together with the squared Eucledian distance corresponds to the use of the Mahalanobis distance with non-standardized Y. Note that the batch update rule (4.3) does not change in that case when derived by the equilibrium condition.

A normalization by the multiplication from the right of Y with $\Sigma^{-1/2}$ has the inconvenient though, that it does not take into account neither the local behavior of Y in a Voronoi region V_j nor the (possibly still unordered) state of the SOM. Arnonkijpanich et al. [2008] therefore suggest to use adaptive distances that vary within Voronoi regions: For node j, we have then $d_{\Lambda_j}(Y, \mathbf{u}_j) = \langle Y - \mathbf{u}_j, Y - \mathbf{u}_j \rangle_{\Lambda_j}$ where Λ_j is a symmetric, positive definite matrix with $\det \Lambda_j = 1$ for all $j \in V_G$. To make the set $d \simeq \{d_{\Lambda_j}\}_{j \in V_G}$ a proper metric, define

$$
\begin{aligned}
&d : \mathbb{R}^p \times \mathbb{R}^p \to [0, \infty), \\
&d(\mathbf{x}, \mathbf{y}) = \begin{cases} d_{\Lambda_j}(\mathbf{x}, \mathbf{y}), \text{ if } \mathbf{x} \in V_j \wedge \mathbf{y} \in V_j \\ d_{\Lambda_j}(\mathbf{x}, \mathbf{u}_j) + d_{\Lambda_\iota}(\mathbf{y}, \mathbf{u}_\iota) + \|\mathbf{u}_j - \mathbf{u}_\iota\|^2, \\ \qquad\qquad \text{if } \mathbf{x} \in V_j \wedge \mathbf{y} \in V_\iota \text{ and } j \neq \iota \end{cases}
\end{aligned} \tag{4.12}
$$

where $\|\cdot\|$ denotes the Eucledian norm. The properties of a metric then persist for $\{d_{\Lambda_j}\}_{j \in V_G}$ as shown in Appendix A. A similar approach to Arnonkijpanich et al. [2008] was suggested by Pacifico and de Carvalho [2011]. Their algorithm however implies that all Λ_j, $j \in V_G$ must be diagonal, too.

This choice of metric has however a negative impact on the theoretical properties of the topographic map and the optimal feature vectors as

the centres of the Voronoi regions cannot be compared easily any more, as will be discussed later. Therefore, a compromise between the classical Mahalanobis and the local adaptive distances is suggested: $d(\mathbf{x}, \mathbf{y})$ is set to $\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle_\Lambda$ where Λ is a symmetric and positive definite matrix that minimizes the distortion measure given (U, V). This means that Λ must be updated during the training and the distance is therefore adaptive. This additional computational burden, though, allows a more flexible density estimator for $d\,P$ in later sections. As the proposed density estimator introduced below has similarity to Gaussian Mixture Model (GMM), it is furthermore attractive to fix the determinant of Λ proportionally to the determinant of the inverse covariance matrix of Y instead of 1. In addition, as will be discussed in Section 4.5 and summarized in Section 4.6, we will grow the graph $\mathcal{G}$ dynamically. With an increased number of nodes, the number of grid points increases and distribute on a convex superset of the support of Y (not necessarily bounded when $\operatorname{supp} P$ is not bounded). In order not to make Λ cover these richer details of $(\mathcal{G}, U)$, we will in the final algorithm fix the proportion between Λ and $\operatorname{Var}[Y]$ to $|V_G|$. The idea is that the average distance between the nodes and Y (hence, division by $|V_G|$) should be of the same size as $\operatorname{Var}[Y]$ and Λ is interpreted as proportional to the estimator's inverse covariance matrix.

4.4 VARIATIONS OF THE SOM

4.4.1 *Simplex Arrangements*

It is common to arrange the graph $\mathcal{G}$ as a lattice on $\mathbb{Z}^2$. However, as has already become clear by the general formulation, this needs not be the case. Kohonen [2013] for example illustrates a hexagonal neighborhood structure in $\mathbb{Z}^2$. Many convergence studies for example arrange $\mathcal{G}$ on a line which allows to identify easily the quality of the SOM's ordering when Y is univariate [Ritter and Schulten, 1986, Erwin et al., 1992].

In the following, we will focus on another arrangement based on simplices. Fritzke [1994] suggests to arrange the nodes in simplices as the number of new edges, when a new node is inserted, grows only linearly, in contrast to when $\mathcal{G}$ is a hypercube. This will become relevant in Section 4.5. Also Ritter [1999] suggests a triangulation instead of the 2-dimensional hypercube arrangement of $\mathcal{G}$. In consequence they suggest – in accordance with the triangulation of $\mathcal{G}$ and coordinates of the nodes on the hyperbolic plane – to use the Poincaré metric for δ and to set h to a multiple of $\mathbb{1}_{[0,\sigma]}$ where σ is called the 'neighborhood radius'. Besides the linear growth rate made possible by the triangulation, the two-dimensional simplex architecture has thus the advantage to mimic the distribution of random variables occuring in a non-Eucledian subspace

of $\mathbb{R}^p$, when used with the appropriate metric. Note that the arrangement in simplices does not change qualitatively the clique-architecture of the graph $\mathcal{G}$.

Remark 8. *In a two-dimensional, finite graph $\mathcal{G}$ where edges form triangles and neighbors are defined by $j \sim \iota \overset{def}{\Leftrightarrow} \min_\rho |\rho(j,\iota)| = 1$, all cliques in the graph are subsets $\Gamma \subset \mathcal{G}$ with $|\Gamma| \leqslant 3$.*

4.4.2 *Related Energies and Random Field Theory*

The main problem with the SOM is that the update rule generally does not correspond to a stochastic gradient descent step of a loss function [Heskes, 1999]. A bad adaptation of the final map to the input data thus indicates convergence to a (possibly bad) local minimum and requires to re-train the SOM with another start set of feature vectors $U^{(0)}$, which is a trial-and-error approach. Thus, formulations of a loss function that leads to similar update rules, retains the specific properties of the original SOM but with better mathematical properties represent an alternative with simplified theoretical analysis.

A naive guess for the (expected) loss function corresponding to the update (4.2) would be Equation (4.5) and is incorrect for continuous Y [Heskes, 1999]. Therefore, Heskes [1999] suggests either a reformulation of the bmu to

$$b(Y, U) = \arg\min_{j \in V_G} \sum_{\iota \in \mathcal{G}} h(\delta(j,\iota)) \cdot d(Y, \mathbf{u}_\iota) \tag{4.13}$$

or a soft assignment approach, i.e. replacing the indicator $\mathbb{1}_{V_j}(Y)$ by a fuzzy-assignment of Y to each node in V_G. The same soft assignments are used for a deterministic annealing approach in Graepel et al. [1998].

As the matching rule (4.13) yields the underlying energy function which shall be minimized, continuous, we adopt this rule in the following analysis. The new rule allows to define similary to (4.5) the empirical distortion as

$$\begin{aligned} \tilde{E}(Y; U, V) &:= \sum_{j \in V_G} h(\delta(j, b(Y, U))) \cdot d(Y, \mathbf{u}_j) \\ &= \sum_{j \in V_G} \mathbb{1}_{V_j}(Y) \cdot \sum_{\iota \in V_G} h(\delta(j,\iota)) \cdot d(Y, \mathbf{u}_\iota) \quad , \end{aligned} \tag{4.14}$$

which is no longer an approximation but the true loss function to whom the update rule (4.2) is a gradient descent step [Heskes, 1999] where $V_j = \{\mathbf{y} \in \mathbb{R}^p : b(\mathbf{y}, U) = j\}$ under (4.13).

Equation (4.14) is then easily differentiable with respect to the elements of U for the first order conditions on optimization, but has still the problem of non-convexity, which bears the risk of getting stuck into

local optima when using a gradient descent algorithm. Heskes [1999] suggests therefore to apply rather than hard assignments $\{\mathbb{1}_{V_j}(\cdot)\}_{j\in\mathcal{G}}$ soft assignments $\{p_j\}_{j\in\mathcal{G}}$ which are in the $|V_G|-1$ unit simplex. Graepel et al. [1998] suggests the same procedure. In combination with an entropy term and a temeperature $T>0$, this suggestion,

$$\sum_{j\in V_G} p_j \sum_{\iota\in V_G} h(\delta(j,\iota))\cdot d(Y,\mathbf{u}_\iota) + \frac{1}{T}\cdot p_j\cdot \log p_j \quad ,$$

leads the authors to derive expressions for p_j that remind of Gibbs measures. However, the properties of $\tilde{E}$ and $\mathcal{G}$ are never explicitely studied as Markov Random Field (MRF) that would imply the existence of such a measure. Furthermore, p_j are often erroneously taken as equivalents to $P(Y\in V_j)$. In general, this does not hold, though, as shall be demonstrated in the following. For that reason, we first integrate in the following the distortion (4.14) in the framework of MRFs, which then allows us to motivate the induced Gibbs measure and to relate it to the iterative solution in Heskes [1999] and Heskes [2001]. To the best of the author's knowledge, the topographic maps based on versions of the SOMs, have not yet been interpreted in the context of random fields. To start with, we give the (original) definition of a MRF for discrete random variables.

Definition 12 (Markov random field). *Let $x_1,\ldots,x_n$ be random variables on a discrete sample space Ω endowed with a neighborhood relation $\sim$. A probability measure P on Ω is a Markov Random Field (MRF) if the conditional probability of an arbitrary outcome ζ of a random variable x_i given all other random variables, $P\left(x_i=\zeta|x_j : j\neq i\right)$ equals the conditional probability given all neighbors of x_i, $P\left(x_i=\zeta|x_j : j\sim i\right)$ [Isham, 1981].*

Now given U and $V=\{V_j\}_{j\in V_G}$ (remember that $V=V(U)$), the feature vectors are arranged on a graph with a neighborhood relation and one could interpret $\{Y|_{V_j}\}_{j\in V_G}$ (assuming the measurability of the Voronoi regions) as the random variables taking some values. Ignoring the fact that Y is possibly continuous, we would have a MRF if for arbitrary $A\in\mathcal{A}_y$, we had $P(Y|_{V_j}\in A|Y|_{V_\iota}\in A : \iota\in V_G) = P(Y|_{V_j}\in A|Y|_{V_\iota}\in A : \iota\sim j)$.

For the classical MRF (that is, with Y being discrete) the MRF can be assured by the existence of a nearest neighbor potential [Isham, 1981]; and for traditional MRFs there exist Gibbs measures due to the Hammersley-Griffet theorem [Grimmett, 1973] that relate the nearest neighbor potential at a configuration with the probability of the configuration to occur, allowing us to state a formula for the distributions of $\{Y|_{V_j}\}_{j\in V_G}$ and therefore Y.

To advance with the MRF theory, define for the metric (4.11) the neighborhood

$$h : [0,\infty) \to [0,\infty), \qquad h(a) = \begin{cases} \gamma_1, & \text{if } a = 0 \\ \gamma_2, & \text{if } 0 < a \leqslant c \\ 0, & \text{else} \end{cases} \tag{4.15}$$

where a is the length of the shortest path between nodes $j, \iota \in V_G$:

$$a = \min_{\rho} |\rho(j,\iota)|$$

and $\gamma_1 \geqslant \gamma_2 > 0$, i.e. h is a step function. For $c = 1$, we approach the context of Pasurek [2007] and Kondratiev et al. [2010] as the distortion measure becomes

$$\sum_{j \in V_G} \sum_{\iota \sim j} h(\delta(j,\iota)) \cdot d\left(Y|_{V_j}, \mathbf{u}_\iota\right) \quad ,$$

we get something close to a nearest neighbor potential where we set $Y|_{V_j} := \mathbb{1}_{V_j}(Y) \cdot Y$. Note that we have to assume that there is no mass point of P at zero. Otherwise, we have for the $j \in V_G$ with $0 \in V_j$ that $P\left(Y|_{V_j} = 0\right) \neq 1 - P(Y \in V_j)$. Under this assumption, we have for any $A \in \mathcal{A}_y$ and $\iota \in V_G$ $P(Y|_{V_\iota} \in A) = P(Y \in A | Y \in V_\iota)$.

If there exists a Gibbs measure, it is then similar to the optimization problem that we consider in Section 4.4.3 based on the already mentioned soft assignments $\{p_j\}_{j \in V_G}$. Let us define the Gibbs measure in general.

Definition 13 (Gibbs Measure). *A probability measure* $\mu \in \text{Prob}\left((\mathcal{Y}, \mathcal{A}_y)^{V_G}\right)$ *is a* Gibbs measure *if it solves at temperature* T *the equation*

$$\mu(A) = \int_{\mathcal{Y}} \pi_\Gamma(A|\mathbf{y})\, d\,\mu(\mathbf{y})$$

where π_Γ *is defined for* $\Gamma \subset V_G$ *as*

$$\pi_\Gamma(A|\mathbf{y}) := Z_\Gamma^{-1}(\mathbf{y}) \int_{\mathcal{Y}_\Gamma} \mathbb{1}_A\left((\mathbf{y}_\Gamma, \mathbf{y}_{\Gamma^c})\right) \cdot \exp\left(-\frac{1}{T} H\left(\mathbf{y}_\Gamma, \mathbf{y}_{\Gamma^c}\right)\right) d\,\mathbf{y}_\Gamma,$$

if $\mathbf{y}$ *belongs to the* tempered *subset of* $\mathcal{Y}$ *(values that return for infinite graphs finite values of* H *– for* $|V_G| < \infty$ *the tempered sets equal* $\mathcal{Y}$*) and zero else.* $Z_\Gamma(\mathbf{y})$ *is a normalizing constant and* H *is a Hamiltonian to be specified later.* $\mathcal{Y}_\Gamma$ *is the projection from* $\mathcal{Y}^{V_G}$ *to the sample space for all nodes in* Γ*.* π_Γ *is called a local specification based on the Hamiltonian* H *[Pasurek, 2007, Definition 2.4 and Equation 2.21].*

The problem is that Y needs not be a discrete and bounded random variable (which would be necessary construct a classical MRF from which to conclude on P). When random field theory shall be applied to the modified topographic map, some modifications to the basic MRF are necessary. In the following, we rely for this on Pasurek [2007] and Kondratiev et al. [2010]. As these authors also allow infinite graphs, their results may be also conclusive as we will grow later on dynamically the SOM with the modified matching rule.

Kondratiev et al. [2010, Equation 1] defines the Hamiltonian (adapted to our notation) as a function of the shape

$$H(Y) := \sum_{j \in V_G} V\left(Y|_{V_j}\right) + \sum_{\iota \sim j} W\left(Y|_{V_j}, Y|_{V_\iota}\right) \quad . \tag{4.16}$$

We must thus define functions V and W such that the distortion measure (4.7) can be identified with a Hamiltonian in order to apply the results of Kondratiev et al. [2010]. In Appendix A, we check whether the assumptions in Kondratiev et al. [2010] hold for which type of random variables. As no interaction of the restricted random variables appear in $\tilde{E}$, we set $W \equiv 0$ and $\tilde{E} = V$ and take the feature vectors U and Voronoi regions $V = V(U)$ as given.

In the following, we consider the graph $\mathcal{G}$ to be finite, i.e. we do not consider the limit behavior of the dynamically growing Algorithm 4.3 that will be discussed later. In that case, requiring that given U and V and

1. Voronoi regions only overlap on null sets of P (meaning that b is almost everywhere well defined)
2. h is as in (4.15) with $c = 1$
3. δ is the metric of shortest paths on $\mathcal{G}$
4. P has no mass point at 0
5. all second moments of Y exist

we can apply Theorem 1 of Kondratiev et al. [2010] and conclude that a Gibbs measure exists (i.e. a Gibbs or equilibrium state) on $\tilde{E}(Y; U, V)$. Applying then Kondratiev et al. [2010, Equation 13] to P, we get then for $\Gamma = V_G$ (cf. Definition 13) and arbitrary $A \in \otimes_{j \in V_G} \mathcal{A}_y$

$$\pi_{V_G}(A) = \frac{1}{Z_{V_G}} \cdot \int_{y^{V_G}} \mathbb{1}_A\left(\mathbf{y} \cdot \mathbb{1}_{V_j}(\mathbf{y})\right) \cdot \exp\left(-\tilde{E}(\mathbf{y}; U, V)\right) \, d\lambda(\mathbf{y}) \tag{4.17}$$

and $\pi_{V_G}(A) = P\left(\{Y \cdot \mathbb{1}_{V_j}(Y)\}_j \in A\right)$ [Kondratiev et al., 2010, Definition 2]. If we have thus another $\tilde{A} \in \mathcal{A}_y$ and plug in $A = \tilde{A} \times y \times \cdot \times y$ we get

from (4.17) $\pi_{V_G}(A) = P(A|V_1)$ and other conditional probabilities based on the position of $\tilde{A}$.

This motivates a density estimator for the law P given that $\mathbf{y} \in V_j$

$$\widehat{\mathrm{d}P}(\mathbf{y}|\mathbf{y} \in V_j) \propto \exp\left(-\sum_{\iota \in V_G} h(\delta(j,\iota)) \cdot d(\mathbf{u}_\iota, \bar{\mathbf{u}}_j) - c_j \cdot d(\mathbf{y}, \bar{\mathbf{u}}_j)\right) \tag{4.18}$$

with

$$\bar{\mathbf{u}}_j := c_j^{-1} \cdot \sum_{\iota \in V_G} h(\delta(j,\iota)) \cdot \mathbf{u}_\iota, \qquad c_j := \sum_{\iota \in \mathcal{G}} h(\delta(j,\iota)) \quad . \tag{4.19}$$

and this piecewise density estimator can simply be stuck together for all components $V_j \in V$ when $P(Y \in V_j) \equiv \frac{1}{|V_G|}$, yielding as a global density estimator

$$\widehat{\mathrm{d}P}(\mathbf{y}) \propto \sum_{j \in V_G} \exp\left(-\sum_{\iota \in V_G} h(\delta(j,\iota)) \cdot d(\mathbf{u}_\iota, \bar{\mathbf{u}}_j) - c_j \cdot d(\mathbf{y}, \bar{\mathbf{u}}_j)\right). \tag{4.20}$$

This shape reminds of Gaussian kernels, which can get arbitrarily close to a wide class of probabilities with Lebesgue density [Nguyen and McLachlan, 2019], though a continuity requirement for P is not directly found in Kondratiev et al. [2010] and would also oppose the origins of MRFs from the discrete case.

Another aspect is that Kůrková [1992] points out that Kolmogorov's representation theorem for continuous functions on the p-dimensional hypercube is also applicable to three-layered neural networks. Though the original SOM is not a classical neural network (which is a supervised machine learning algorithm), the definition of a global energy function $\tilde{E}$ and a neighborhood function h of the type (4.15) may lead to its interpretation as a neural network. Under this viewpoint, the requirement on P for a density approximation by (4.20) is that P has a continuous Lebesgue density and a bounded support.

We can thus conclude that there are several arguments coming from different directions (MRFs, GMMs and neural network theory) that argue for the working of the proposed density estimator and due to the various origins of the arguments, the type of probability laws P, for which (4.20) is possibly applicable, is relatively broad – it must be assured only that $P(Y \in V_j) = \frac{1}{|V_G|}$ for all $j \in \mathcal{G}$, which we seek to achieve by a dynamic adaption of the graph $\mathcal{G}$ on the training realizations of Y and therefore the underlying distribution P. Such a growth rule is discussed in Section 4.5.

The existence of a Gibbs measure, and its interrelation to the original P, gives rise to the *variational principle*. The variational principle of thermodynamics states that an equilibrium state maximizes the entropy (under some conditions on the corresponding equilibrium measure). If there is another Gibbs measure, the relative entropy between those equals zero. The entropy here is maximal when there is uniform assignment to the Voronoi regions. The assignment to bmus depends here on U, but under the simplifying assumption that the assignment probabilities are a priori independent from U, we get an alternative, easier-to-solve optimization problem, which exhibits nonetheless similiarities to the Gibbs measure introduced here.

4.4.3 *Optimization of an Alternative Energy Function*

Besides the discussion about the existence of a Gibbs measure, there are still feature vectors to be optimized in the expected loss, i.e. the distortion measure (4.7). We study here first the locally optimal feature vectors to link them in a second step to the (locally) optimal feature vectors of an alternative optimization problem.

The alternative assignment rule (4.13) yields with a metric of type $d(\mathbf{y}, \mathbf{u}_j) = \langle \mathbf{y} - \mathbf{u}_j, \mathbf{y} - \mathbf{u}_j \rangle_\Lambda$ (cf. Section 4.3) the gradient

$$\nabla_{\mathbf{u}_j} E\left[\tilde{E}(Y; U, V)\right] = -2\Lambda \cdot \sum_{\iota \in V_G} P(Y \in V_\iota) \cdot E\left[h(\delta(j, \iota)) \cdot \left(Y - \mathbf{u}_j\right) | Y \in V_\iota\right]$$

[Heskes, 1999]. Setting the gradient to zero, this yields as optimal $\mathbf{u}_j$ Expression (4.9) (in expectations). Under the condition $P(Y \in V_j) \equiv \frac{1}{|V_G|}$, and a neighborhood function (4.15) with $\gamma_1 = \gamma_2$ and $c = 1$, these optimal feature vectors correspond to conditional expectations of joined Voronoi regions,

$$E\left[\mathbf{u}_j\right] = E\left[Y | \cup_{\iota \sim j} V_\iota\right] \quad , \tag{4.21}$$

where $\iota \sim j$ denotes a neighbor ι of node j, i.e. the existence of an arc (j, ι) in E_G. This then again makes $\bar{\mathbf{u}}_j$ (4.19) a conditional expectation, too.

The optimization problem

$$\begin{aligned} &\min_U \ \tilde{E}(Y; U, V) \\ &\text{s.t. } V = V(U) \end{aligned} \tag{4.22}$$

is non-trivial and can be stuck easily in local optima when gradient descent methods are applied to an arbitrary initialization. For that reason,

Graepel et al. [1998] and Heskes [1999] introduced for SOMs an alternative optimization problem

$$\begin{aligned} \min_{U,p} \quad & \sum_{j \in V_G} p_j \cdot \sum_{\iota \in \mathcal{G}} h(\delta(j,\iota)) \cdot d(Y, \mathbf{u}_\iota) + T(p_j \cdot \log p_j) \\ \text{s.t.} \quad & \sum_{j \in V_G} p_j = 1 \\ & p_j \in [0,1] \quad \forall j \in V_G \quad . \end{aligned} \tag{4.23}$$

Using the Lagrangien, this optimization problem yields as optimal solution for p_j, $j \in V_G$ given U

$$p_j = \frac{\exp\left(-\frac{1}{T} \sum_{\iota \in V_G} h(\delta(j,\iota)) \cdot d(Y, \mathbf{u}_\iota)\right)}{\sum_{l \in V_G} \exp\left(-\frac{1}{T} \sum_{\iota \in V_G} h(\delta(l,\iota)) \cdot d(Y, \mathbf{u}_\iota)\right)} \quad . \tag{4.24}$$

This unit simplex $\{p_j\}_{j \in V_G}$ is often referred to as Gibbs measure and it is assumed that $p_j = P(Y \in V_G)$ [Graepel et al., 1998], though it does not correspond to

$$\frac{1}{Z_{V_G}} \cdot \int_{V_j} \exp\left(-\sum_{\iota \in V_G} h(\delta(j,\iota)) \cdot d(\mathbf{y}, \mathbf{u}_\iota)\right) d\,\mathbf{y} \quad , \tag{4.25}$$

which would result from (4.17). Nonetheless, Equation (4.24) looks like an empirical version of (4.25) for $T = 1$. Furthermore, Graepel et al. [1998] argues that for $T \to 0$, the original optimization problem (4.22) is recovered whereas for $T \to \infty$, the optimization problem (4.23) becomes globally concave.

Though given the optimal $p = \{p_j\}_{j \in V_G}$, there can also be (locally) optimal feature vectors be derived, we shall first discuss the interpretation of p_j as assignment probability $P(Y \in V_j)$: Whilst $P(Y \in V_j)$ does not contain anymore the random variable Y, p_j does so. This suggests the interpretation of $p = \{p_j\}_{j \in V_G}$ as a set of a posteriori assignment probabilities: Given an a priori assignment probability implicitely by (4.23) set to uniformity, and the sample realization $Y = \mathbf{y}$, the assignment probability is updated as if the assignment did not depend a priori on the feature vectors. If the average $\bar{p}_j$ of p_j^i, $i \in S$, is taken where p_j^i corresponds to the a posteriori assignment probability (4.24) of independent Y_i with $Y_i \overset{d}{=} Y \sim P$ for all $i \in S$, $\bar{p}_j$ would, on the other hand, remind of an empirical integral for (4.25).

Furthermore, note that the density estimator (4.20) reminds of a Gaussian mixture with mixture components

$$\pi_j := \exp\left(-\frac{1}{T} \sum_{\iota \in V_G} h(\delta(j,\iota)) \cdot d(\mathbf{u}_\iota, \bar{\mathbf{u}}_j)\right) \tag{4.26}$$

for $T = 1$ where assignment to a mixture component $j \in V_G$ is not known for any realization of Y. Again with $T = 1$, these components are recovered as the optimal mixtures for the optimization problem (4.23) and are a compromise between the uniformity requirement (resulting from the T-weighted entropy therm) and the quantization error.

Any distribution P that is absolutely continuous on the Lebesgue measure can be approximated by continuous Gaussian mixtures [Nguyen and McLachlan, 2019, Corollary 6]. Here, we consider a subclass of distributions such that the second moment of Y exists and in that case, mixing need not be continuous and we have even an approximation on a sub-manifold of p-dimensional GMMs depending on U (and the metric-matrix Λ). Note that Corollary 6 in Nguyen and McLachlan [2019] has the continuous mixing component as a function of the mixtures' central parameter (i.e. the conditional expectations for Gaussian kernels), which is also the case here with (4.26) depending on $\{\bar{\mathbf{u}}_j\}_{j\in V_G}$. This is another indication that with the modified SOMs, we observe a melting pot of different theories such as GMMs and MRFs.

Second, let us study the optimal feature vectors for (4.23): Plugging in (4.24) into the Lagrangien of (4.23) yields

$$-\log\left(\sum_{j\in V_G}\exp\left(-\frac{1}{T}\sum_{\iota\in V_G}h(\delta(j,\iota))\cdot d(Y,\mathbf{u}_\iota)\right)\right) \quad , \tag{4.27}$$

which is the negative of a Gaussian mixture likelihood with mixing components (4.26) just as in the proposed density estimator (4.20). Optimizing (4.23) can therefore be considered as the direct optimization of the Gibbs measure underlying $(P, \mathcal{G})$ (which exists for every $(U, V(U))$ yielding almost surely a partition). This is also in accordance with the previously (and not further studied) variational principle claiming that the Gibbs measure maximizes the entropy under a given map $(U, \mathcal{G})$.

The (locally) optimal feature vector $\mathbf{u}_j$ for arbitrary $j \in V_G$ is consequently

$$\mathbf{u}_j = \frac{\sum_{\iota\in V_G} h(\delta(j,\iota))\cdot p_\iota \cdot Y}{\sum_{\iota\in V_G} h(\delta(j,\iota))\cdot p_\iota} \quad , \tag{4.28}$$

where the prefactor does not cancel out when there are $|S| > 2$ iid realizations of Y, that is, when the optimization problem (4.23) becomes

$$\min_{U,\{p^i\}_{i\in S}} \frac{1}{|S|}\sum_{i\in S}\sum_{j\in V_G} p_j^i \sum_{\iota\in V_G} h(\delta(j,\iota))\cdot d(Y_i,\mathbf{u}_\iota) + T\cdot p_j^i\cdot\log p_j^i \quad , \tag{4.29}$$

where $p^i = \{p_j^i\}_{i\in S}$.

Note that the optimal U and p have an internality problem due to the term p in (4.28). Usually, this problem is solved in statistics iteratively, i.e. using iteratively the optimal p from (4.24) to be plugged into (4.28) and the other way around, leading to an EM-type procedure, where the components p lie on a sub-manifold of the $(|V_G|-1)$-dimensional unit simplex, as they depend on U through a specific shape. Though, the risk function to be minimized is not a log-likelihood but a penalized log-likelihood (cf. Appendix A).

Taking expectations under the previously named neighborhood function with compact support, yields that $\bar{\mathbf{u}}_j$ corresponds to the conditional expectation given that $Y \in \cup_{\iota \in j} V_\iota$ when $p_j \to \mathbb{1}_{V_j}$ pointwise. As we assumed $P(Y \in V_j) \equiv \frac{1}{|V_G|}$, a good indication for this behavior is suggested as $\bar{p}_j \approx \frac{1}{|V_G|}$ for all $j \in V_G$.

Considering the empirical optimization problem (4.23) as an approximation to some expected loss, we consider in a third step the expected alternative energy function:

$$\begin{aligned} \min_{U,p} & \sum_{j \in V_G} p_j \cdot E\left[\sum_{\iota \in V_G} h(\delta(j,\iota)) \cdot d(Y, \mathbf{u}_\iota)\right] + T \cdot p_j \log p_j \\ \text{s.t.} & \sum_{j \in V_G} p_j = 1 \\ & p_j \in [0,1] \quad \forall j \in V_G \quad . \end{aligned} \tag{4.30}$$

Independent optimization of the variables U and p yields as optimal p_j, $j \in V_G$,

$$p_j = \frac{\exp\left(-\frac{1}{T} E\left[\sum_{\iota \in V_G} h(\delta(j,\iota)) \cdot d(Y, \mathbf{u}_\iota)\right]\right)}{\sum_{l \in V_G} \exp\left(-\frac{1}{T} E\left[\sum_{\iota \in V_G} h(\delta(l,\iota)) \cdot d(Y, \mathbf{u}_\iota)\right]\right)} \quad . \tag{4.31}$$

Due to Jensen's inequality and (4.25), we can thus conclude that $p_j \neq P(Y \in V_j)$ for the non-empirical optimization problem as well and that the independent optimization of U and p is also in theory an approximation. Nonetheless, note that we get p_j^i for each $i \in S$ and thus, averaging $\bar{p}_j$ converges only in probability to (4.31) but is conceptually closer to (4.25).

Next, we consider the optimal solution to the feature vectors. The optimal U under the optimization problem (4.30) can then be P almost surely defined by

$$\mathbf{u}_j = \frac{\sum_{\iota \in V_G} h(\delta(j,\iota)) \cdot p_\iota \cdot E[Y]}{\sum_{\iota \in V_G} h(\delta(j,\iota)) \cdot p_\iota} \quad , \tag{4.32}$$

which is presumably to much centered towards $E[Y]$ in contrast to the conditional expectations. We conclude therefore, that for the modified

SOMs, it is more worthwhile to study the properties of the empirical alternative energy from the perspective to simplify the original optimization problem based on the distortion measure (4.7) than to formulate an alternative risk function that is then sought to be optimized empirically.

Recall the important assumption that $P(Y \in V_j) \equiv \frac{1}{|V_G|}$, which links the alternative energy to the Gibbs measure. For given U and training data, this requirement cannot be assured, but by the empirical means $\{\bar{p}_j\}_{j \in V_G}$, it can be assessed to which extend it is possibly violated. This is for the reason that $\bar{p}_j$ resembles the empirical version of the Gibbs measure (4.25). It is thus essential to modify the Algorithm 4.2 besides the discussed metrics, the alternative bmu definition, and the alternative energy, in order to assure $\bar{p}_j \approx \frac{1}{|V_G|}$. This leads to a dynamically growing topographic map like it is discussed in Section 4.5.

4.5 GROWING SOMS

An important inconvenient of the classical SOM is that the graph size must be determined a priori, which is problematic when barely something of the data is known and the analysis shall be explorative. For example, if P is already originally a Gaussian mixture with a fixed number of components, less nodes are presumably required than otherwise. Also, as outlined in Section 4.4, for a mathematical statistical interpretation of the variations of the SOM, it is necessary that in the empirical optimization $\bar{p}_j \approx \frac{1}{|V_G|}$.

Due to the problem of a pre-defined graph size, dynamic algorithms that grow the graph during the training were proposed, for example by Fritzke [1994], Fritzke [1995] and Alahakoon et al. [2000]. Alahakoon et al. [2000] suggest to grow the SOM on the boundaries of the graph (which are defined when $\mathcal{G}$ is a lattice) such that a new node reduces the local quantization error when the latter trespasses a pre-defined threshold s. If – based on the quantization error criterion – the insertion could take place anywhere in the graph $\mathcal{G}$, this would allow to bound the L^2 distance between an estimated probability measure based on U and the original distribution of Y. However, for $V_G \subset \mathbb{Z}^2$, growth everywhere on the graph is hard to realize, because insertion *between* nodes is not possible. So Alahakoon et al. [2000] only grow the grid at its boundary, which does not prevent from high quantization errors in the interior. A growth rule exclusively on the boundary can therefore lead to a problem known as magnification error [Ritter and Schulten, 1986]: Low-density areas of P are over-represented in the SOM because the trained feature vectors U accumulate themselves in areas with a relatively low probability mass. This problem already exists with the classical SOM because the size of the update depends on the metric δ and not on d, but can be reinforced by the growth rule of Alahakoon et al. [2000].

Alternative growth rules are suggested by Fritzke [1994] and Fritzke [1995]. These dynamic extensions of the SOM aim at a uniform distribution of $b(Y, U)$ on V_G. The author proposes to grow the SOM around nodes that are overproportionally often hit, i.e. nodes $jl \in V_G$ with $P(b(Y, U) = j) \gg \frac{1}{|V_G|}$, and the insertion of new nodes is as follows. Assume that j is very often a bmu. A new node j' is then inserted between this most often bmu j and a neighbor ι. ι is chosen as the neighbor node with $d(\mathbf{u}_j, \mathbf{u}_\iota) = \min_{l \in \mathcal{N}_j} d(\mathbf{u}_j, \mathbf{u}_l)$, as between these feature vector values – presumably close in the input space due to approximate topology preservation – there is much agitation of the probability function assumed. On the other hand, nodes that are never bmu can be removed in Fritzke [1994], which allows to reflect zero density areas of P. The deletion rule is not implemented in the final modified mini-batch SOM Algorithm 4.3, but would be conceptually possible in future work.

Also this update rule reduces, in principle, the average distance between input and feature weights, because those nodes in 'high density areas' of the probability function become more and more similar to the input data. Nonetheless, these nodse are not necessarily those with the most distorted feature vectors. Fritzke [1994] claims that the dynamic growth rule maximizes the entropy but lacks a mathematical proof. Nota bene, the insertion of nodes in high density areas of P also counteracts the magnification problem. And the possibility to insert nodes in the interior of the graph allows to tackle magnification, too.

The way to re-arrange the neighborhood after the insertion of j' differs in Fritzke [1995] from Fritzke [1994]. In Fritzke [1995], the indices in V_G are like in the traditional algorithm arranged in a two-dimensional lattice. Inserting j' then requires the insertion of a completely new row or column respectively between j and ι. This leads to an exponential growth in the number of nodes [Fritzke, 1994]. In addition, nodes other than j' in the new row/ column are possibly not required and make the SOM training unnecessarily computer intensive.

Fritzke [1994], on the other hand, arranges $\mathcal{G}$ as a set of linked, low-dimensional simplices. When triangles are chosen as structure, this leads to an triangularization of Image(Y). Neighbours are those nodes that share a vertex, that is δ is of the type (4.11). The insertion of a new node j' thus means that simpleces that include the most-hit node j and its neighbor ι, must be dissolved. New simpleces that include j and j' and new simpleces involving j' and ι are created, leading to a slow, linear growth of the network structure.

Recall the alternative optimization problem (4.23) from Section 4.4.3. Like already mentioned, the concentrated empirical risk (4.27), where the optimal p is plugged in, reminds of a GMM with additional penalty terms. The iterative update of Equations (4.24) and (4.28) (cf. the Sum-

mary (A.10)) then yield an expected maximum penalized likelihood algorithm.

Geman and Hwang [1982] relate penalized ML to the method of sieves and for many infinite dimensional estimation problems, the method of sieves yields consistent estimators when the restrictions are relaxed 'sufficiently slowly' [Geman and Hwang, 1982]. Adding new mixture components changes the likelihood surface and the EM-type algorithm that often gets stuck in local optima may exit such local optima due to the change of surface [Priebe and Marchette, 1993]. Priebe and Marchette [1993] also introduces dynamic growth rules for Gaussian mixture model estimation and requires to 'wait long enough between [mixture component] creation', too. This cannot be assured for an exponential growth rate, at least when there is no maximum number of nodes defined. The relation to the method of sieves thus favors the simplex arrangement and growth rule in Fritzke [1994].

Fritzke [1994] notes that other update rules than high accumulation on a node can serve as an update rule, too. For example, the growth decision of Alahakoon et al. [2000] (i.e. high distortion of U) is also applicable with the growth rule in Fritzke [1994]. This makes the suggested dynamic extension more flexible than the one developped in Alahakoon et al. [2000]. In Priebe [1994], the author suggests to use a growth rule that either creates an additional node when the local standardized quantization error trespasses a threshold, which corresponds to the growth rule of Alahakoon et al. [2000]. Or a stochastic growth rule might be implemented similar to simulated annealing [Priebe, 1994]. This then has also similiarities to the chosen update rule in (4.33b): Let $\mathcal{G}_\kappa = (V_{G_\kappa}, E_{G_\kappa})$ denote the graph in training step κ. As we require $p_j \approx \frac{1}{|V_G|}$, we create an additional node if $\max_{j \in V_{G_\kappa}} \bar{p}_j$ deviates too much from $\frac{1}{|V_{G_\kappa}|}$. Denote by G the function describing the growth decision. We have, for a growth decision based on units $S' \subseteq S$

$$G\left(Y_{S'}; U^{(\kappa)}, p_{S'}^{(\kappa)}\right) = \begin{cases} 1, & \text{if positive growth decision} \\ 0, & \text{else} \end{cases} \tag{4.33a}$$

and in particular for Algorithm 4.3

$$\begin{aligned} &G\left(Y_{S'}; U^{(\kappa)}, p_{S'}^{(\kappa)}\right) \\ &\quad = \left(1 - \mathbb{1}_{[0,\varepsilon]}\left(\max_{j \in V_{G_\kappa}} \left| \frac{\frac{1}{|V_{G_\kappa}|} - \bar{p}_j^{(\kappa)}}{1 + \frac{1}{|V_{G_\kappa}|}} \right|\right)\right) \cdot \mathbb{1}_{[0, n_{\mathcal{G}_{max}})}(|V_{G_\kappa}|) \end{aligned} \tag{4.33b}$$

where

$$\bar{p}_j^{(\kappa)} = \frac{1}{|S'|} \sum_{i \in S'} p_j^{i(\kappa)} \quad .$$

That is, due to the structure of the optimal simplex $p^{(\kappa)}$ in step κ, the growth decision ressembles a sort of deterministic annealing rather than the simulated annealing suggested in Priebe [1994]. Note however the similiarties between the update rules suggested in Fritzke [1994], Alahakoon et al. [2000] and Priebe [1994] though the studies are in different fields of statistics/ machine learning. This underpins again that there is a link between MRFs, penalized ML of GMMs, the method of sieves for GMMs and SOMs. There remains the question on how many nodes should there be at maximum: The method of sieves requires a sufficiently slow growth [Geman and Hwang, 1982, Priebe and Marchette, 1993] and an unbounded growth of nodes (because the stopping criterion may not be met) might be possible. The similarity of the proposed density estimator to Gaussian mixture approximation suggests to apply a growth rule for the mixture components to the growth of the graph $\mathcal{G}$. Nguyen and McLachlan [2019, Remark 27] states that the optimal number of mixture components is $\mathcal{O}\left(\sqrt{n}\right)$ where $|S| = n$.

Like in Alahakoon et al. [2000], removal of nodes that are never addressed as bmu is possible in Fritzke [1994]. The removal has the advantage that if $\text{Image}(Y)$ is not convex, feature vectors that are not even close to $\operatorname{supp} P$ are removed. Such vectors might occur due to the averaging in the training (cf. Equation 4.3) or the initialization of the feature vector of a newly added node as a convex combination of persisting feature vectors.

4.6 SUMMARY OF THE FINAL ALGORITHM AND THEORY

As some modifications of the original batch Algorithm 4.2 were proposed in the preceeding sections, we summarize here all those that are finally applied in the simulation and application sections 4.8 and 4.9 respectively. The modifications encompass the alternative assignment rule (4.13) [Heskes, 1999] (Section 4.4.2), updates with the simplified alternative energy function [Heskes, 1999] (Section 4.4.3), the adaptive distances suggested in Arnonkijpanich et al. [2008] (Section 4.3) and the simplex arrangement and growth rule of Fritzke [1994] (Section 4.5). The updates of U done in each step are based on the alternative energy function discussed in Section 4.4 first proposed by Graepel et al. [1998] and Heskes [1999].

Furthermore, a mini-batch version is suggested, that allows to combine the improved convergence of the batch algorithm [Fort et al., 2002, Kohonen, 2013] with faster computations in each training step like in the on-line algorithm. That means, the calculus in each training step κ is done only for a random subset $S' \subset S$ with $|S'| < |S|$. Another argument for a mini-batch version is that the batch version ressembles an EM-type algorithm of a GMM [Heskes, 2001, Kostiainen and Lampinen, 2002, Ver-

beek et al., 2005], and a stochastic version thereof can help to overcome local optima [Jank, 2006, for pure EM].

The final algorithm is described in Algorithm 4.3. The growth decision G in training step κ is described by (4.33b). In words, if the maximum posterior assignment probability in training step κ, $\bar{p}_j^{(\kappa)}$, exceeds the uniform assignment probability $\frac{1}{|V_{G_\kappa}|}$, we have a positive growth decision as far as the number of nodes does not trespass the maximum number of nodes $n_{\mathcal{G}_{max}}$. Deviation from uniformity is measured by an error compromising between the relative and absolute error by a value greater than ε.

The convergence criterion checks whether the maximal relative change from $U^{(\kappa)}$ to $U^{(\kappa+1)}$ for nodes existing in V_{G_κ} and $V_{G_{\kappa+1}}$ trespasses a certain level. Alternatively, the Algorithm stops when the maximum number of iterations $K \in \mathbb{N}$ is reached. The idea to update the metric in the input space stems from the fact that elements in a p-dimensional random vector may vary in scale. In order not to overfit on one variable with a large variance, a sort of Mahalanobis distance is employed. However, we remark that not the sample mean is used as location parameter, but U, thus it is sensible to adjust the scale parameter with respect to these like in Arnonkijpanich et al. [2008].

The weighting components (that can be interpreted like posterior probabilities) $p^i = \{p_j^i\}_{j \in V_G}$, $i \in S'$, on the other hand, originate from the problem that the original update rule is generally not based on an energy function. As the definition of an alternative assignment rule solves the problem but still yields the optimization problematic due to involving indicator variables $\{\mathbb{1}_{V_j}\}_{j \in V_G}$ that rely on U, an easier energy function (4.23) has been defined. This easier energy function, however, shares under certain conditions optimality criteria similar to the original problem and is nonetheless easier to optimize (locally). Furthermore, the alternative energy function (4.23) returns the negative of a Gaussian mixture likelihood when the optimal weighting components p_j^i, $i \in S$, $j \in \mathcal{G}$ are plugged in.

This mixture model suggests a density estimator that corresponds to the one motivated by the Gibbs measure of the original optimization problem. $\mathcal{G}$ is a finite graph (with therefore also finite vertex degree). Therefore, the summands of the energy function (4.14) can be considered to be a potential with state space equal to the vertices in $\mathcal{G}$ and a configuration space equal to the sample space of Y conditioned on a Voronoi partition of $\mathcal{Y}$. The pair potential is then constantly set to zero (cf. Appendix A. Due to the finiteness of the graph, the energy is uniformly bounded from below by zero in any subset of $\mathcal{G}$. Having this global stability [Kondratiev et al., 2010], one can state that the set of tempered Gibbs measures on $\mathcal{G}$ with the corresponding energy function (4.14) is non-empty. For finite $\mathcal{G}$ and $Y \in L^2(\mathcal{Y}, \mathcal{A}_\mathcal{Y}, P)$, the tempered Gibbs mea-

Algorithm 4.3 Extended SOM Mini-Batch Algorithm

Require: Training data $\{\mathbf{y}_i\}_{i\in S}$, initial set of simplices $\mathcal{G}_0 = (V_{G_0}, E_{G_0})$, initial feature vectors $U^{(0)} := \{\mathbf{u}_j^{(0)}\}_{j\in V_{G_0}}$, metric δ on $\mathcal{G}$, neighborhood function h, weight matrix $\Lambda^{(0)} = \mathrm{Var}[Y]^{-1}$

Ensure: Topographic map $(\mathcal{G}, U)$

for $\kappa = 0, 1, 2, \ldots, K$ **do**

Draw randomly a subset $S' \subset S$

Calculate for all $j \in V_{G_\kappa}$ and $i \in S'$

$$p_j^{i(\kappa)} = \frac{\exp\left(-\frac{1}{T}\sum_{\iota\in V_{G_\kappa}} h(\delta(j,\iota))\cdot d_{\Lambda^{(\kappa)}}(\mathbf{y}_i, \mathbf{u}_\iota^{(\kappa)})\right)}{\sum_{l\in V_{G_\kappa}}\cdot\exp\left(-\frac{1}{T}\sum_{\iota\in V_{G_\kappa}} h(\delta(l,\iota))\cdot d_{\Lambda^{(\kappa)}}(\mathbf{y}_i, \mathbf{u}_\iota^{(\kappa)})\right)}$$

Update the feature vectors $j \in V_{G_\kappa}$ according to

$$\mathbf{u}_j^{(\kappa+1)} = \left(\sum_{i\in S'}\sum_{\iota\in V_{G_\kappa}} p_\iota^{i(\kappa)} h(\delta(\iota,j))\right)^{-1} \left(\sum_{i\in S'}\sum_{\iota\in V_{G_\kappa}} p_\iota^{i(\kappa)} h(\delta(\iota,j))\mathbf{y}_i\right)$$

Update the weighting matrix

$$D \leftarrow \sum_{i\in S'}\sum_{j\in V_{G_\kappa}} p_j^{i(\kappa)} \sum_{\iota\in V_{G_\kappa}} h(\delta(j,\iota))\cdot\left(\mathbf{y}_i - \mathbf{u}_\iota^{(\kappa+1)}\right)\left(\mathbf{y}_i - \mathbf{u}_\iota^{(\kappa+1)}\right)^T$$

$$\Lambda^{(\kappa+1)} = D^{-1}\cdot\left(\det D\cdot\frac{|V_{G_\kappa}|}{\det \mathrm{Var}[Y]}\right)^{1/p}$$

Calculate $p_{S'}^{(\kappa+1)}$ on $\mathcal{G}_\kappa$

if Positive growth decision (i.e. $G\left(\mathbf{y}_{S'}; U^{(\kappa+1)}, p_{S'}^{(\kappa+1)}\right) = 1$) **then**

Find $j = \arg\max_\iota \frac{1}{|S'|}\sum_{i\in S'} p_\iota^{i(\kappa)}$

Find $\tilde{j} = \arg\max_{\iota\in\mathcal{N}_j} \|\mathbf{u}_j - \mathbf{u}_\iota\|^2$

Dissolve simplices including both j and $\tilde{j}$

Initialize $\mathbf{u}_{j'}^{(\kappa+1)} = 0.5\cdot(\mathbf{u}_j^{(\kappa+1)} + \mathbf{u}_{\tilde{j}}^{(\kappa+1)})$

$\mathcal{G}_{\kappa+1} \leftarrow \left(V\cup j', E\setminus\{(j,\tilde{j})\}\cup\{(j,j'),(\tilde{j},j')\}\right)$

end if

$\kappa \leftarrow \kappa + 1$

if Convergence criterion is met **then**

break

end if

end for

$U \leftarrow U^{(\kappa)}$, $\mathcal{G} \leftarrow \mathcal{G}_\kappa$, $\Lambda \leftarrow \Lambda^{(\kappa)}$

sures correspond to the general Gibbs measures. If the feature vectors in U are such that $P(Y \in V_j(U)) = \frac{1}{|V_G|}$ for all $j \in V_G$, the Gibbs measure yields the same Lebesgue density as the Gaussian mixture model stemming from (4.27).

In order to ensure the uniformity of the assignment probabilities to Voronoi regions, the dynamic growth rule is established. This dynamic growth rule is that one suggested in Fritzke [1994], but ignores the deletion of unnecessary nodes. $P(Y \in V_j) = \frac{1}{|V_G|}$ (proxied by $\{\bar{p}_j\}_{j \in V_G}$) is required for the justification of the proposed density estimator. Therefore, the growth decision should rather rely on $P(Y \in V_j)$ and the weighting components in (4.24) than on quantization error. The number of nodes should be linked to the sample size $|S|$: Priebe and Marchette [1993] requires a slow growth for the method of sieves for mixture models, which can also be linked to the modified mini-batch SOM Algorithm 4.3. This similarity to Gaussian mixture approximation might give a hint for the maximal number of nodes. As Nguyen and McLachlan [2019, Remark 27] noted, the number of mixture components should be $\mathcal{O}\left(\sqrt{|S|}\right)$.

As a mini-batch training is employed, it is not always the case that for each Voronoi region there is at least one observation falling into it. Therefore, the non-parametric estimators $\frac{1}{|S'|} \cdot \sum_{i \in S'} \mathbb{1}_{V_j}(Y_i)$, $j \in V_G$, are unstable and not recommended for the evaluation of the assignment probabilities. Hence, the average of the parametric estimators (4.24) are used in order to stabilize the estimation process. The reasoning is that if $\mathbf{y}_i \notin \cup_{\iota \rightsquigarrow j} V_\iota$, the p_j^i gets close to 0 and close to 1 otherwise, which equals in fact the traditional empirical mean.

4.7 MACHINE LEARNING WITH SURVEY DATA

Like other machine learning algorithms, SOMs might encounter problems when applied to survey data: Basic machine learning algorithms such as artificial neural networks with one node can be identified with regression models [Marsland, 2015, Chapter 3] and therefore should share the same problems as in Pfeffermann [1993]. Furthermore, one major role of machine learning with survey data is imputation [Fessant and Midenet, 2002, Mallinson and Gammerman, 2003]. However, Reiter et al. [2006] states the importance to account for the sampling design P_D for the completion of missing data.

Nonetheless, survey design in machine learning seems to be a white spot on the research landscape. The only report known to the author that accounts for the survey design when using a machine learning algorithm is written by Amer [2007] and the design considered is classicalStratRS. Though, the report's type of publication remains unclear. These arguments concern mostly neural networks and support vector machines as

supervised learning algorithms. Although SOMs are an unsupervised algorithm, the identification of the suggested density estimator (4.20) with a Gaussian mixture model based on the negative log-likelihood (4.27) suggests that the same problem can occur with complex designs P_D generating S.

One possible way to approach the problem would be to weight the concentrated energy function (4.27) noting that it is equivalent to a negative log-likelihood. This would result in a pseudo-ML estimator like in Chapter 2. However, this approach would require a batch optimization of (4.27) or the corresponding EM-type algorithm, and increase the computational effort. Furthermore, it would be necessary to re-study the correspondence between the alternative optimization problem and the original one and the applicability of random field theory.

Alternatively, one could use the fact that stochastic optimization based on the random sub-sampling $S' \subset S$. If the stochastic optimization in the machine learning algorithm was run on the finite population U as the original sample from a superpopulation model P (cf. Section 1.3 and Chapter 2), the probability to be drawn into a mini-batch $S' \subset U$ should be constant for each $i \in U$ and equal to $\frac{|S'|}{N}$. When the probability P'_D of unit i to be drawn into a mini-batch S' given that i is in survey sample $S \subset U$, $S \sim P_D$ is set to

$$P'_D(i \in S' | i \in S) \propto \frac{1}{P_D(i \in S)} \quad ,$$

the joint probability that $i \in S' \cap S$ is again constant by $P'_D(i \in S' | i \in S) \cdot P_D(i \in S) = \text{const}$. Then, we get as marginal probability that $i \in S'$

$$P'_D(i \in S') = \sum_{s \in \mathcal{S}} \text{Prob}(i \in S' \cap s) = |\mathcal{S}| \cdot \text{const}.$$

where $\mathcal{S}$ is again the set of survey samples with $P_D(s) > 0$. From this derivation, it seems plausible to draw units into the mini-batch S' with a probability proportional to their design weights, which are the inverse of the inclusion probabilities. However, this is only a first order correction and variances will usually be different from a SOM applied to the finite population $\{Y_i\}_{i \in U}$.

4.8 SIMULATION STUDY

4.8.1 *Simulation Set-up*

The following simulation study aims at the evaluation of the elaborated Algorithm 4.3 and to compare it with established density estimation methods such as kernel density estimation [Chacón and Duong, 2018]

and classical GMMs [Scrucca et al., 2016]. Often, multivariate Gaussians or Gaussian mixtures are evaluated in comparative simulation studies for multivariate density estimation [Ćwik and Koronacki, 1996, 1997, Hwang et al., 1994, Wand and Jones, 1994, Duong and Hazelton, 2005, Zhang et al., 2006, Lu et al., 2013]. Sometimes, mixtures of multivariate t-distributions are used as well [Hwang et al., 1994, Zhang et al., 2006]. Recurring to these distributions is mostly based on the fact that the quality measures for density estimators need to evaluate functionals: The outcome $\widehat{\mathrm{d}\,\mathrm{P}}(\mathbf{x};\mathbf{y})$ depends on the evaluation point $\mathbf{x}$ and the training data $\mathbf{y}$. I.e. for almost every $\mathbf{y} \in \mathcal{Y}$ the random outcome is a function $\widehat{\mathrm{d}\,\mathrm{P}}(\cdot;\mathbf{y})$ on a superset of $\mathcal{Y}$. For the true density $\mathrm{d}\,\mathrm{P}$ such a quality measure is the Mean Integrated Squared Error (MISE)

$$\mathrm{MISE}\left[\widehat{\mathrm{d}\,\mathrm{P}}, \mathrm{d}\,\mathrm{P}\right] := \int\int \left(\widehat{\mathrm{d}\,\mathrm{P}}(\mathbf{x};\mathbf{y}) - \mathrm{d}\,\mathrm{P}(\mathbf{x})\right)^2 \mathrm{P}\,(\mathrm{d}\,\mathbf{x})\ \mathrm{P}\,(\mathrm{d}\,\mathbf{y}) \tag{4.34}$$

and the Kullback-Leibler divergence

$$\mathrm{KL}\left[\widehat{\mathrm{d}\,\mathrm{P}}, \mathrm{d}\,\mathrm{P}\right] := \int\int \log\left(\frac{\mathrm{d}\,\mathrm{P}(\mathbf{y})}{\widehat{\mathrm{d}\,\mathrm{P}}(\mathbf{x};\mathbf{y})}\right) \mathrm{P}\,(\mathrm{d}\,\mathbf{x})\ \mathrm{P}\,(\mathrm{d}\,\mathbf{y}) \quad . \tag{4.35}$$

These require integration that becomes burdensome in higher-dimensional settings and are difficult to compute for non-standard multivariate distributions. The standard scenario with multivariate Gaussian mixtures – though computationally convenient in evaluation – has the disadvantage that many of the density estimators rely on Gaussian kernels and therefore, simulation results can be misleading. Non-standard scenarios, though, make it difficult to evaluate (4.34) and (4.35) as in every simulation run $b = 1, \ldots, B$ one would still need to compute

$$\mathrm{MISE}^{(b)}\left[\widehat{\mathrm{d}\,\mathrm{P}}, \mathrm{d}\,\mathrm{P}\right] := \int \left(\widehat{\mathrm{d}\,\mathrm{P}}(\mathbf{x};\mathbf{y}^{(b)}) - \mathrm{d}\,\mathrm{P}(\mathbf{x})\right)^2 \mathrm{P}\,(\mathrm{d}\,\mathbf{x})$$

and

$$\mathrm{KL}^{(b)}\left[\widehat{\mathrm{d}\,\mathrm{P}}, \mathrm{d}\,\mathrm{P}\right] := \int \log\left(\frac{\mathrm{d}\,\mathrm{P}(\mathbf{x})}{\widehat{\mathrm{d}\,\mathrm{P}}(\mathbf{x};\mathbf{y}^{(b)})}\right) \mathrm{P}\,(\mathrm{d}\,\mathbf{x}) \quad .$$

Zhang et al. [2006] and Lu et al. [2013] solve this problem by Monte-Carlo integration. This means, in each of the $b = 1, \ldots, B$ simulation runs with a density estimate $\widehat{\mathrm{d}\,\mathrm{P}_b} := \widehat{\mathrm{d}\,\mathrm{P}\,(\cdot;\mathbf{y}^{(b)})}$, there are $\nu = 1, \ldots, n_B$ realizations from $X_\nu \sim_{iid} \mathrm{P}$ and the MC versions

$$\mathrm{MISE}^{(b)}_{\mathrm{MC}}\left[\widehat{\mathrm{d}\,\mathrm{P}_b}, \mathrm{d}\,\mathrm{P}\right] := \frac{1}{n_B}\sum_{\nu=1}^{n_B}\left(\widehat{\mathrm{d}\,\mathrm{P}_b}(X_\nu) - \mathrm{d}\,\mathrm{P}(X_\nu)\right)^2 \tag{4.36}$$

and

$$\mathrm{KL}_{\mathrm{MC}}^{(b)}\left[\widehat{\mathrm{d}\,\mathrm{P}_b}, \mathrm{d}\,\mathrm{P}\right] := \frac{1}{n_B}\sum_{\nu=1}^{n_B} \log\left(\frac{\mathrm{d}\,\mathrm{P}(X_\nu)}{\widehat{\mathrm{d}\,\mathrm{P}}(X_\nu)}\right) \tag{4.37}$$

are evaluated. The distribution over $b = 1, \ldots, B$ simulation runs then gives us information about the distribution of the Kullback-Leibler and MISE functionals

$$\int \left(\widehat{\mathrm{d}\,\mathrm{P}}(\mathbf{x};\cdot) - \mathrm{d}\,\mathrm{P}(\mathbf{y})\right)^2 \mathrm{P}\,(\mathrm{d}\,\mathbf{x})$$

and

$$\int \log\left(\frac{\mathrm{d}\,\mathrm{P}(\mathbf{x})}{\widehat{\mathrm{d}\,\mathrm{P}}(\mathbf{x};\cdot)}\right) \mathrm{P}\,(\mathrm{d}\,\mathbf{x})$$

In our simulation study, we set $n_B = 3000$ and $B = 500$.

In this simulation study, we investigate the performance of the proposed density estimator for three different bivariate densities. Like in previous studies, one of the distributions under study, P_1, is a GMM:

$$\begin{aligned}\mathrm{P}_1 := \frac{1}{6}\cdot \mathrm{N}\left(\underbrace{\begin{pmatrix}0\\0\end{pmatrix}}_{=:\mu_1}, \underbrace{\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}}_{=:\Sigma_1}\right) + \frac{1}{3}\cdot \mathrm{N}\left(\underbrace{\begin{pmatrix}1\\1\end{pmatrix}}_{=:\mu_2}, \underbrace{\begin{pmatrix}3 & 1\\1 & 2\end{pmatrix}}_{=:\Sigma_2}\right) + \\ + \frac{1}{2}\cdot \mathrm{N}\left(\underbrace{\begin{pmatrix}7\\7\end{pmatrix}}_{\mu_3}, \underbrace{\begin{pmatrix}9 & 4\\4 & 4\end{pmatrix}}_{=:\Sigma_3}\right).\end{aligned} \tag{4.38a}$$

The pdf $\mathrm{d}\,\mathrm{P}_1 =: \mathrm{f}_1$ is thus defined for $\mathbf{x} \in \mathbb{R}^2$ as

$$\mathrm{f}_1(\mathbf{x}) \triangleq \frac{1}{6}\cdot\phi\,(\mathbf{x}) + \frac{1}{3}\cdot\phi\left(\Sigma_2^{-1/2}\,(\mathbf{x}-\mu_2)\right) + \frac{1}{2}\cdot\phi\left(\Sigma_3^{-1/2}\,(\mathbf{x}-\mu_3)\right), \tag{4.38b}$$

where ϕ is the density of the multivariate standard normal distribution.

The second distribution under study is chosen due to the nature of business surveys: There, positive and skewed random variables that are correlated with each other are common, for example 'research and development investments' and 'return on investment'. Therefore, we combine the exponential distribution (with parameter $\lambda = 3$) with the log-normal distribution and have as second density $\mathrm{d}\,\mathrm{P}_2 =: \mathrm{f}_2$ for $\mathbf{x} = (x_1, x_2) \in (0,\infty) \times (0\infty)$

$$\mathrm{f}_2(x_1, x_2) = \frac{\lambda}{\sqrt{2\pi}}\cdot\frac{1}{x_2}\cdot \mathrm{e}^{-\frac{(x_2-x_1)^2}{2}}\cdot \mathrm{e}^{-\lambda x_1} \tag{4.39}$$

and zero else.

The third distribution P_3 is bivariate logistic (that is, a two-dimensonal extreme value distribution) with symmetric marginal densities $\mathrm{d}\,P_3(x_1 \times \mathbb{R}) = \mathrm{d}\,P_3(\mathbb{R} \times x)$, $x \in \mathbb{R}$, set to

$$\mathrm{d}\,P_3(x \times \mathbb{R}) = -e^{-x} \cdot e^{-e^{-x}} \quad , \tag{4.40}$$

which is the univariate Gumbel distribution with location parameter 0 and scale parameter 1. The data are generated by the R-package `evd` [Stephenson, 2002]. Also, the evaluation of the density $\mathrm{d}\,P_3 =: f_3$ is done using `evd`. The dependency parameter between the two Gumbel distribtuions is set to 0.3.

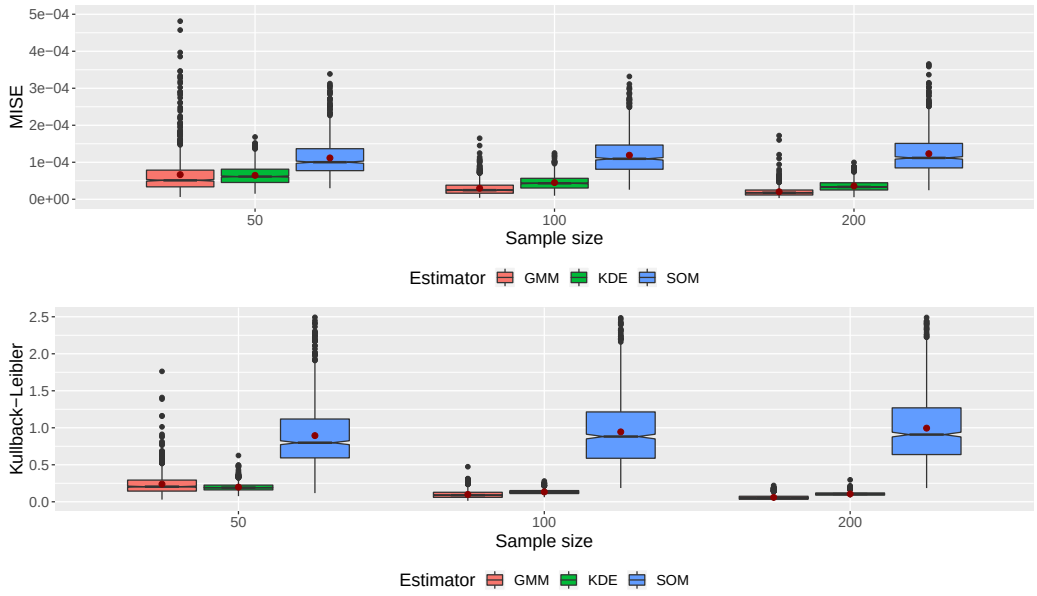
The performance of the proposed estimator (4.20) is evaluated using the distribution of MC-MISE (4.36) and MC-Kullback-Leibler divergence (4.37). The estimator is contrasted with classical GMMs where the number of components is chosen by the Bayesian Information criterion (R-package `mclust`, Scrucca et al. [2016]) and the multivariate kernel density estimator (R-package `ks`, Chacón and Duong [2018]). The standard settings are chosen for both competing estimators. In order to check whether the estimators converge weakly as foreseen by the theory, the size of the training data $Y \sim_{iid} P_m$, $m = 1, 2, 3$, increases from 50 over 100 to 150, i.e. $|S| \in \{50, 100, 150\}$. The number of maximum nodes for the proposed estimator is set to $\mathcal{O}\left(\sqrt{|S|}\right)$. The choice of temperature T depends on the distribution under study and was set after visual inspection of the marignal density estimators and can be learned from the electronic appendix. An automated choice of the temperature T could be topic of future research.

4.8.2 *Simulation Results*

The simulation results are presented in Figures 4.2 to 4.4. For the established estimators, we can observe that both the Kullback-Leibler divergence and the MISE reduce when sample size increases from 50 to 150. For the SOM based estimator, more differentiation is necessary.

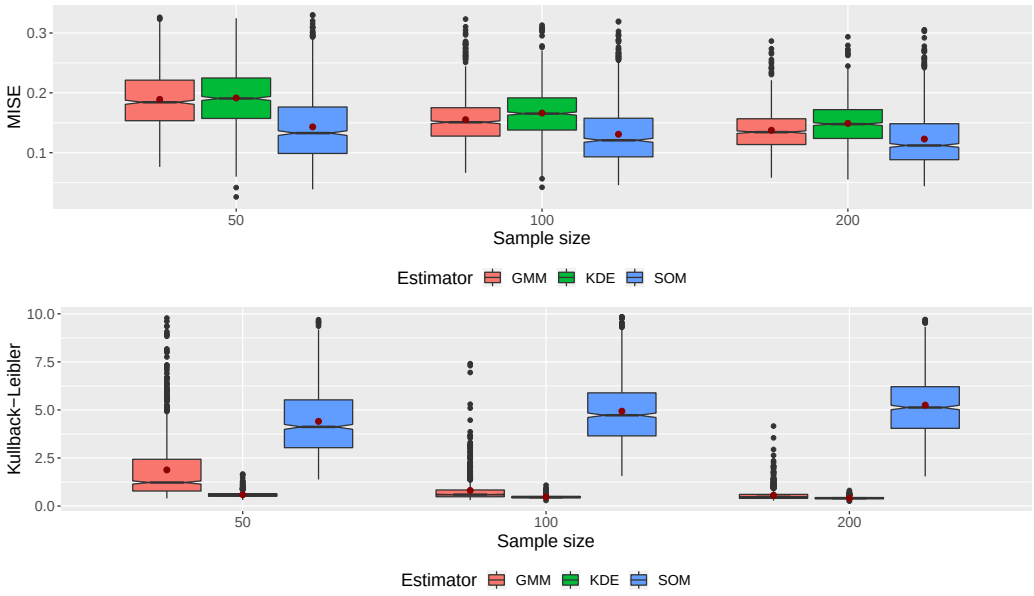
The proposed algorithm seems to have troubles with the decrease of MISE and Kullback-Leibler divergence as the sample size grows for the GMM model although the estimator was designed from a subclass of Gaussian mixtures, confer Figure 4.2. Contrary to the discussed theory, we cannot observe a decrease neither in MISE nor the Kullback-Leibler divergence with an increase of the sample size. First of all, it cannot be excluded that this lack of decrease is due to MC errors, as the Monte-Carlo averages for all sample sizes are still close to each other. Alternatively, this might be due to the topographic nature: Neither mixture components nor the conditional means $\{\bar{\mathbf{u}}_j\}_{j \in V_G}$ are independent from each other for the proposed estimator, though this is the case for the

Figure 4.2: MISE and KL of Estimators for P_1



DGP. Possibly there is also an interaction between sample size and hyperparameters (number of iterations, temperature) that would make it necessary to adjust the latter with growing sample size. Note however, that the proximity of the MISE to that of the GMM and kernel density estimator is promising. Furthermore, large outliers like for the GMM do not occur. For the second distribution under study, it can be observed, on the other hand, that the SOM based estimator is favorable in terms of MISE and the squared error also decreases with increasing sample size as foreseen by the theory. However, this cannot be verified for the Kullback-Leibler divergence. The Kullback-Leibler divergence as loss function for density estimation is known to be driven by the tail behavior of the underlying distribution (and the kernels, when the estimator is a kernel density estimator) [Hall, 1987]. As the problem of increasing Kullback-Leibler divergence especially occurs with P_2, which is very skew, this might be a possible explanation: The Kullback-Leibler divergence even goes to zero when $f_2(x_1, x_2) \to 0$ and the estimator $\widehat{d\,P_2}(x_1, x_2)$ does not, but penalizes divergence of f_2 and $\widehat{d\,P_2}$ in the tails. So if the proposed estimator adjusted better than GMM and kernel density estimation in high density areas, but performed worse in the tails, this would explain the contrary behavior of MISE and Kullback-Leibler divergence in Figure 4.3. For the bivariate extreme value distribution P_3, neither such boundary nor tail effects can be observed. For the GMM estimator, there are such outliers in the Kullback-Leibler divergence for $|S| = 50$ that the MC mean does not fit into the plot. Furthermore, we observe an advantage of the proposed estimator for small sample sizes and the traditional estima-

Figure 4.3: MISE and KL of Estimators for P_2



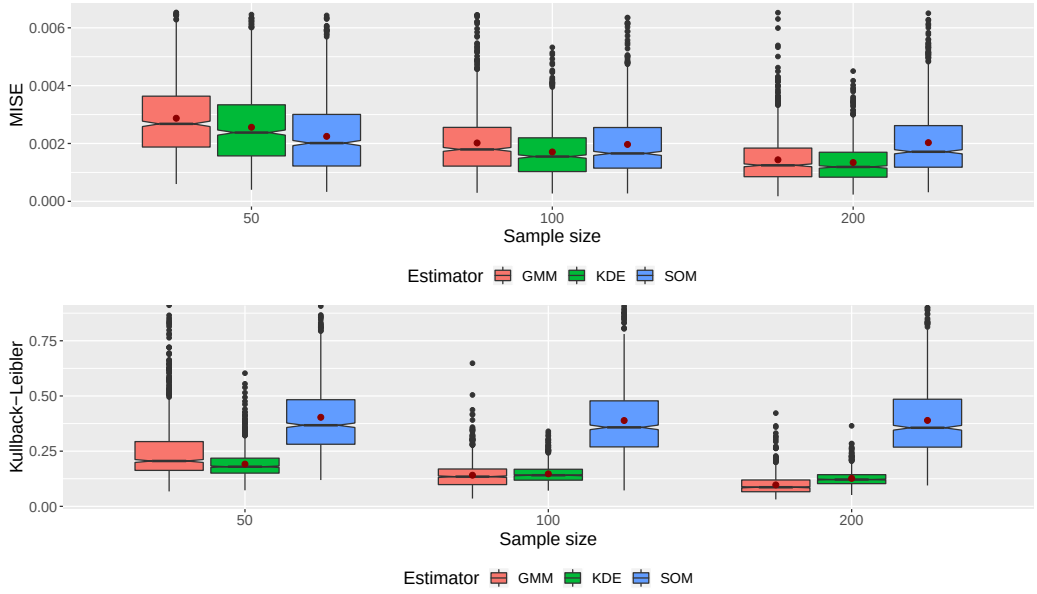
tors only catch up with increasing sample size. Nonetheless, we observe (almost) slight decreases in both the MISE and the Kullback-Leibler divergence for the proposed estimator like predicted by the theory.

4.8.3 *Some Remarks on the Simulation Study*

The choice of hyperparameters has proved to be essential for the performance of the proposed estimator in previous simulation runs. Beyond the scope of this work, their choice surely can be optimized or heuristics be developed. Especially with the Gaussian mixture model, the size of the mini-batch $|S'|$ appeared to be important as the likelihood of Gaussian mixtures have many local optima and a smaller batch size allows the algorithm to generalize from the training data. On the other hand, smaller mini-batches increase the number of iterations to get a stable topographic map and when stopped to early, the resulting estimator becomes more variable.

Setting Λ proportional to the inverse map size (cf. Section 4.3) leads to increased inversion problems that we observed to ease when setting the temperature T to a higher value. Also the choice of temperature gets important as it has an impact on the covariance matrix used in Estimator (4.20). The more nodes are used, the smaller the bandwidth or (inverse) covariance matrix can get without destabilizing the estimator. Furthermore, the composition of the energy function (4.23) suggests – identifying the first term as a sum of conditional squared errors – that the temperature may interrelate with the random variable's second moment. In

Figure 4.4: MISE and KL of Estimators for P_3



fact, we observe for $E[X_1|P_1] = \frac{67}{36}$ and $E[X_2|P_1] = \frac{53}{36}$ a temperature that makes the SOM based estimator in MISE terms competitive and lies in between the temperatures chosen for P_2 with $E[X_1|P_2] = \frac{1}{3}$ and $E[X_2|P_2] = \frac{\exp(9)}{9} - \exp\left(\frac{9}{2}\right)$ respectively and for P_3 with $E[X_1|P_3] = E[X_1|P_3] = \frac{\pi^2}{6}$. This suggests a negative interrelation between a 'good' temperature and the variance.

4.9 APPLICATION TO BUSINESS SURVEYS

The application of proposed new estimation methods to real world data is always an essential step for the establishment into the tool kit of proven methods. Real world data have often the defect that the DGP is unknown so they are not adequate to control for the reliability of methods in contrast to simulation studies. However, we can compare Estimator (4.20) again with other tools such as kernel density estimation and GMMs.

Like in Section 2.3, we choose for the difficulties of non-normal data a business survey; it is again the BEEPS. In the data set, variables for the total turnover from sales for the last fiscal year (variable `d2`) and three fiscal years ago (variable `n2`) are available and are supposedly highly correlated. In order to include as many observations as possible, the variables were transformed from local currency to US dollar so that the data set can contain firms from different countries. The World Bank's official exchange rate two years (or five years respectively) before the interview year were taken because the questionnaire for the 2009 survey usually

refers to 2007 and for the other years, we assume likewise. A more detailed data description can be found in Section 2.3.2. Data points that have a survey weight available were chosen in order to contrast classical estimation results with that one discussed in Section 4.7. Like in Chapter 2, the problem of Algorithm 4.3 is the stochasticity, this time due to the mini-batches. Therefore, the algorithm is run 100 times with a subsequent estimation of $\mathrm{d}\,\mathrm{P}$. The density is evaluated on a two-dimensional grid for each of the 100 runs and results presented here are the average of those 100 algorithm runs. Hyperparameters such as mini-batch size, maximum grid size $|V_G|$ and temperature were chosen to make the marginal density estimate similar to the univariate kernel density estimate. As convergence results do not hold for accumulation points and zero is often such an accumulation point in real-world data sets, only observations with a total of sales greater than zero in both fiscal years were used.

The results are presented in Figure 4.5 and 4.6. Whilst the density estimator (4.20) based on topographic maps has a similar shape and functional values to the kernel density estimator, the GMM estimator has a much steeper and skewer shape, such that for better interpretability, the plots had to be cut on the vertical axis. Knowing that the GMM seems to be prone to outliers (cf. especially Figures 4.2 and 4.4), the similarity between the kernel density and the SOM based estimator may be considered a hint for their reliability. On the other hand, the SOM based estimator flattens faster in the tails than the other two estimators as can be seen especially on the third panel – this is also in accordance with the interpretation of the Kullback-Leibler behavior in Figure 4.3. For better readability and to check whether there is positive correlation between the total sales three fiscal years ago and last fiscal year (approximately 0.6825), we plot the contours of the density estimators in Figure 4.5 as well. We see that all estimators reflect the correlation though the GMM estimator is extremely concentrated.

The use of survey weights showed to be problematic – given the hyperparameter choice for the Algorithm 4.3, none of the 100 runs succeeded for the weighted mini-batch selection. However, for another seed, we were able to finish one run of the Algorithm. The result is contrasted with the unweighted version from Figure 4.5 and plotted in Figure 4.6.

As the data reveal a negative correlation between survey weights and total sales (-0.15 for the last fiscal year), firms with less total sales are drawn with higher probability in each training step. The firm size is a stratifier in the survey design and correlates with total sales (correlation between number of permanent full-time employees and total sales: 0.3495). Therefore, the design can even be considered to be informative if the sampling fractions differ between the strata, because the stratifier variable is not taken into account in the estimation. Therefore, it is not

Figure 4.5: Density Estimation of Total Sales – I

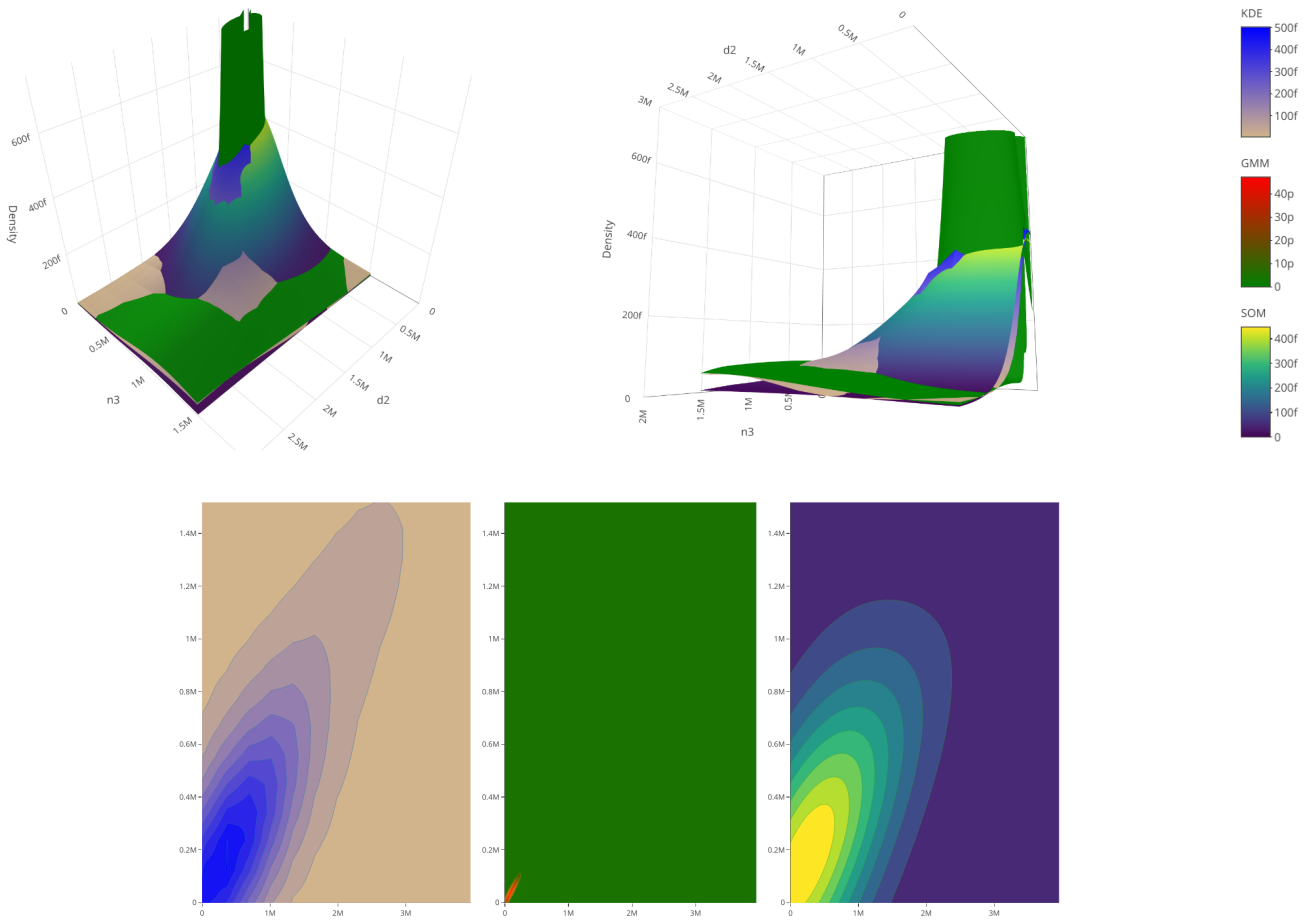
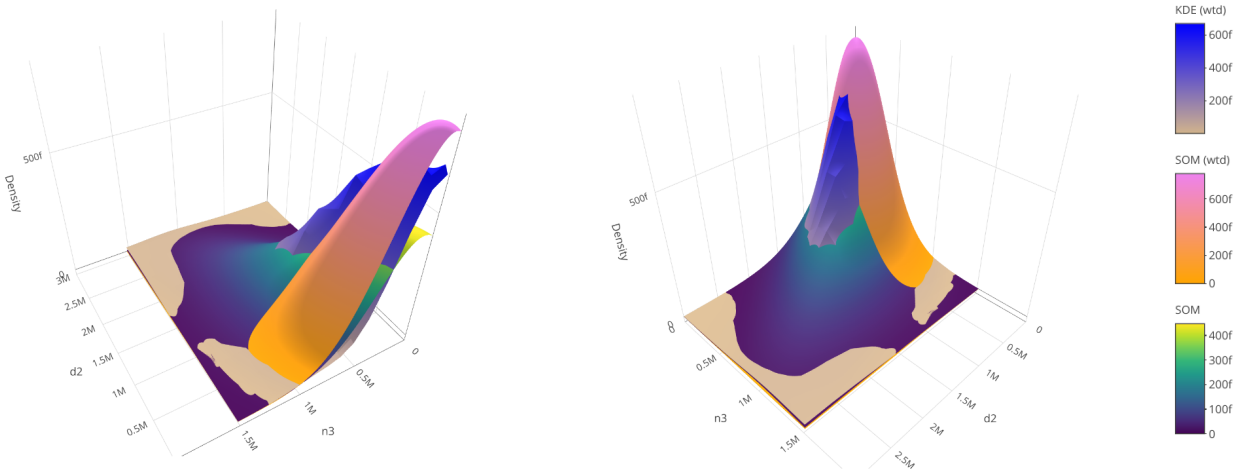


Figure 4.6: Density Estimation of Total Sales – Weighted



surprising that the weighted density estimator is steeper than the unweighted one, as can be seen in Figure 4.6. The package `ks` also allows the inclusion of (not clearer specified) weights. The kernel density estimator using these weights is also plotted in Figure 4.6. Interestingly, the SOM based estimator is in that case a little steeper than the kernel density estimator. Note, however, that the proportions of both estimators are again similar.

4.10 CONCLUSION

The in depth study of SOMs and derivations thereof have revealed through the theory of MRFs a link to the random variable's underlying probability measure that goes beyond previous studies [Graepel et al., 1998, Heskes, 2001, Kostiainen and Lampinen, 2002]. The proximity to the theory of Gaussian mixture approximations has furthermore given a hint about the weak convergence behavior of the suggested density estimator that was partly confirmed by a subsequent simulation study, especially also on the requirements on the distribution to be estimated (such as $Y \in \mathcal{L}^2$ and P^Y having a Lebesgue density). Nonetheless, a thorough analysis of convergence behavior should follow this work as the optimization for each grid size is iterative through an EM-type algorithm. In addition, one could also use the same argumentation as in Chapter 2: When the sample size increases, it would also become necessary to increase the mini-batch size because otherwise the constant MC error could hinder convergence. A dynamic growth rule allows the density estimator to adapt subsequently on the complexity of the training data. An overall convergence behavior thus depends on the interaction of the growth

rate of $|S'|$, the growth rate of nodes $|V_G|$, the number of training steps between node creation and the growth of the sample size $|S|$.

In terms of MISE, the new estimator has proved to perform competitively to well established density estimators, especially when the data is not Gaussian. The behavior of the Kullback-Leibler divergence on the other hand is only partly explainable by the theory given here. This and the appropriate choice of the hyperparameters $|S'|$, maximal graph size and T still require a closer inspection, though there seems to be an interdependency with the second moments of the random variable to be studied.

Finally, the discussed way to deal with survey design in machine learning (cf. Section 4.7) has returned a weighted density estimator similar to weighted kernel density estimation (where the type of weight is not clearly specified) and different from the unweighted version. This indicates that dealing with survey design is important and that the stochastic optimization offers a way to account for survey design (at least in point estimation).

5

SUMMARY AND OUTLOOK

Like many other survey data, business surveys serve for different purposes: Whilst econometricians may run regression analysis in order to reveal statistically relevant relationships between business variables, official statistics focus on the provision of summary statistics such as finite population totals. This work has discussed problems that arise when complex survey designs that are common for business surveys [Cox and Chinnappa, 1995, Burgard et al., 2014] are used and confounded with classical statistical inference. Besides the problem of extreme (and possibly informative) designs, challenges when data are non-Gaussian were discussed.

Chapter 2 has discussed extensions of the linear mixed model to other families of the exponential distribution and under power transformations to restablish normality assumptions. Furthermore we took into account the implementability of crossed random effects structures and the inclusion of survey weights. In terms of point estimation, the proposed algorithm showed to be competitive to existing software in a simulation study; though much more computer intensive. The simulation study further revealed that the survey weighting becomes necessary when it cannot be excluded that the survey design is informative for the aimed regression analysis. However, the proposed MCEM algorithm suffered nonetheless from slight stochastic and/or numeric instabilities that should be tackled in future. Previous research [Burgard and Dörr, 2019] demonstrated in addition that the Fisher information does not – at least under the chosen set-up – yield a reliable estimator for the standard error, though it was suggested in earlier literature [Booth and Hobert, 1999]. For useful application of the suggested algorithm in statistical inference, thus, the development of reliable standard error estimators is essential. Finally, the MCEM algorithm was applied on a PUM of the World Bank: the BEEPS. Accounting for the economic sectors and an establishment's location (country) as crossed random effects, the question was studied whether firms that were led by women suffer from higher obstacles for access to finance. The lack of reliable standard estimators led to an analysis that was combined with established software. No statistically significant impact of gender on the loan and credit approval was found, though the results have to be interpreted carefully, and we had partially results diverging from standard software. These were, however, consistent with previous theoretical studies of the topic.

Chapter 3 reconciled the use of GVFs in survey statistics with their motivation originating from classical randomization. This reconciliation also allowed an error decomposition that goes beyond the one already known [Cho et al., 2002]. The role of survey design for the performance of a GVF prediction – a component that is often neglected – was highlighted. First, the relation of a population total estimator (the HT or GREG) to its variance was elaborated under the model-design framework. The often cited design effect was put in this context, too. The role of design was also underlined by the concluding MC simulation that studied both the performance of common GVF shapes and whether the quality measures applied in practice really give information on the GVFs' prediction quality. Like common in business surveys, the estimation process was complicated by the non-normality of the data. It was found that due to the positivity and skew distributions of variance estimators, it is usually advantageous to use the logarithmic or the relative variance estimator respectively when fitting the GVF, though the predictions are often biased. The common quality measures, are not so reliable that in practice, when the true variance of a point estimator is not known, they barely indicate an appropriate choice for the shape of the GVF. Therefore, without a lot of experience or extensive prior Monte-Carlo research, it seems critical to apply GVFs in practice. A future project would thus be to develop more reliable quality measures for GVFs.

Chapter 4 finally, introduced the popular self-organizing maps and extended the theoretical knowledge about them as a link to MRFs was established. The MRF literature suggests the existence of a Gibbs measure, which in turn easily motivates a density estimator that could already by derived in Heskes [2001] and Kostiainen and Lampinen [2002] but only under simplifying assumptions. Like in Heskes [1999, 2001] and Kostiainen and Lampinen [2002], the link to Gaussian mixture models was made, which gives additional hints on both the weak convergence of the proposed estimator and the requirements on the underlying distribution. The original algorithm was extended to grow dynamically and to adapt its metric on the requirements of the data. In a simulation study, the behavior of the new estimator was studied and compared to established density estimation methods for different sampling distributions. Especially for skewed data like they occur in business surveys, the results were promising for the new method. However, the estimator's behavior for mixture distributions is complicated and the results for the Kullback-Leibler divergence disagree with the competitive performance in terms of MISE. This might be due to an interplay of sample size and hyperparameter choice which still has to be studied. Especially the choice of a good temperature is relevant for the estimator's performance and not clear a priori. Also, how to account for survey design remains an open question, though first thoughts were expressed and tried on a real world

data set. Again, using the BEEPS of the World Bank, the bivariate distribution of time-delayed sales turnover was estimated. The proposed estimator returned similar results like the kernel density estimator. The result changed into the expected direction when the design (that is possibly informative for the estimator) was accounted for and kept similarity to a weighted kernel density estimator.

To conclude, this work has discussed several fields where survey design and non-normality may play a major role in statistics. The promising results in all three chapters underline that statistical research has progressed so far that scientists in empirical research, official statistics or even machine learning can and should leave unrealistic assumptions on data normality and survey non-informativity and turn to more modern, possibly computer-intensive methods that are less restrictive on data requirements.

Part II

APPENDIX

A

PROOFS

SAMPLE INFORMATIVITY AND NON-ORTHOGONALITY

First we show that Definition 9 is equivalent to Pfeffermann [1993, Definition 3]. Let p_θ denote a density of $P_{\mathcal{M}_\theta}$ with respect to a measure μ. If both Y_i and Z_i are either continuous or discrete, μ equals either the Lebesgue or the count measure. If this is not the case, μ is a mixture measure composed of the two. We can then combine p_θ with P_D to get a joint mixture density $p_{\theta,D}$

$$p_{\theta,D}(s;\mathbf{y},\mathbf{z}) := P_D(s;(\mathbf{y},\mathbf{z}))\cdot p_\theta(\mathbf{y},\mathbf{z})$$

which yields with $A \in \mathcal{A}$ and $s \in 2^U$

$$\begin{aligned} P_{\mathcal{M}_\theta,D}(s\times A) &= \int_A P_D(s;(\mathbf{y},\mathbf{z}))\cdot p_\theta(\mathbf{y},\mathbf{z})\,\mathrm{d}\,\mu(\mathbf{y},\mathbf{z}) \\ &= \int_A P_D(s;(Y,Z))\,\mathrm{d}\,P^N_{\mathcal{M}_\theta} \quad , \end{aligned}$$

i.e. the definition of the joint model-design measure (1.13). Let $p_\theta(\mathbf{z})$ be the marginal density of Z at $\mathbf{z}$, i.e. the density of $P^N_{\mathcal{M}_\theta}(\mathrm{proj}_z^{-1} \in \cdot)$ and $p_\theta(\mathbf{y}|\mathbf{z})$ be the conditional density of Y at $\mathbf{y}$ given $Z=\mathbf{z}$, $p_\theta(\mathbf{y}|z) = p_\theta(\mathbf{y},\mathbf{z})/p_\theta(\mathbf{z})$. Then it is possible to define also the conditional probability of $Y|Z \in \mathrm{proj}_z(A)$ lying in $\mathrm{proj}_y(A)$ on $(\mathcal{Y}_N,\mathcal{A}_y)$ through the projections from the joint space Ω and arbitrary $A \in \mathcal{A}$

$$P^N_{\mathcal{M}_\theta}\left(Y \in \mathrm{proj}_y(A)|Z \in \mathrm{proj}_z(A)\right) = \frac{P^N_{\mathcal{M}_\theta}\left((Y,Z)\in A\right)}{P^N_{\mathcal{M}_\theta}\left((Y,Z)\in(\mathcal{Y}_N\times \mathrm{proj}_z(A))\right)}$$

cf. Equation (1.12) or equivalently through the density $p_\theta(\cdot|\mathbf{z})$

$$\begin{aligned} &P^N_{\mathcal{M}_\theta}\left(Y \in \mathrm{proj}_y(A)|Z \in \mathrm{proj}_z(A)\right) \\ &= \int_{\mathrm{proj}_z(A)} \int_{\mathrm{proj}_y(A)} p_\theta(\mathbf{y}|\mathbf{z})\,\mathrm{d}\,\mu(\mathbf{y})\,p_\theta(\mathbf{z})\,\mathrm{d}\,\mu(\mathbf{z}) \end{aligned} \tag{A.1}$$

for any $A \in \mathcal{A}$ with $\mathbf{z} \in \mathrm{proj}_z(A)$. For the model-design density conditioning on $Z=\mathbf{z}$ is analoguous.

Then, we have thanks to the product structure of the algebras for a $A \in \mathcal{A}$ and $\mathrm{proj}_y(A) =: A_y \in \mathcal{A}_y$ for almost every $\mathbf{z}$

$$P^N_{\mathcal{M}_\theta}\left(\mathrm{proj}^s(A_y)\times \mathcal{Y}_{U\setminus s}|Z=\mathbf{z}\right) = \int_{\left(\mathrm{proj}^s(A_y)\times \mathcal{Y}_{U\setminus s}\right)} p_\theta(\mathbf{y}|\mathbf{z})\,\mathrm{d}\,\mu(\mathbf{y}), \tag{A.2}$$

which corresponds to the marginal probability of Y_s given $Z = \mathbf{z}$ and yields the same joint density of Y_s and Z as Pfeffermann [1993, Equation 3.3] by multiplication with $p_\theta(\mathbf{z})$ because $A_y \in \mathcal{A}_y$ is arbitrary. On the other hand, we have

$$\begin{aligned} &P_{\mathcal{M}_\theta,D}(\{2^U \times (\text{proj}^s(A_y) \times \mathcal{Y}_{U\setminus s})\}|Z = \mathbf{z}) \\ &\quad = \sum_{s\in 2^U} \int_{(\text{proj}^s(A_y)\times\mathcal{Y}_{U\setminus s})} p_{\theta,D}(s;\mathbf{y}|\mathbf{z})\, d\,\mu(\mathbf{y}) \\ &\quad = \sum_{s\in 2^U} \int_{(\text{proj}^s(A_y)\times\mathcal{Y}_{U\setminus s})} P_D(s;\mathbf{y},\mathbf{z}) \cdot p_\theta(\mathbf{y}|\mathbf{z})\, d\,\mu(\mathbf{y}) \quad , \end{aligned} \tag{A.3}$$

which is identical to the marginal distribution of Y_s given $Z = \mathbf{z}$ under $P_{\mathcal{M}_\theta}$ if $P_D(\cdot;(Y,Z)) = P_D(\cdot;Z)$ almost surely. The term $\int_{(\text{proj}^s(A_y)\times\mathcal{Y}_{U\setminus s})} P_D(s;\mathbf{y},\mathbf{z}) \cdot p_\theta(\mathbf{y}|\mathbf{z}) \cdot p_\theta(\mathbf{z})\, d\,\mu(\mathbf{y})$ yields the joint density of Y_s and S like in Pfeffermann [1993, Equation 3.2]. The same inference occurs when integrating out S from that joint distribution yields the same marginal distribution for Y_s like above. This exactly the case if $P_D(\cdot;(Y,Z)) = P_D(\cdot;Z)$ which means $Y \perp S|Z$. Therefore, Remark 3 is proven.

□

PROOF OF REMARK 5

It holds for an estimator g_S

$$\begin{aligned} E_{\mathcal{M}_\theta,D}\left[g_S\left(S,Y,Z\right)\right)] &= \int_{\mathcal{S}\times\Omega} g_S(S,Y,Z)\, d\, P_{\mathcal{M}_\theta,D} \\ &= \sum_{s\in 2^U} \int_\Omega P_D(s;(Y,Z)) \cdot g_S(s,Y,Z)\, dP^N_{\mathcal{M}_\theta} \\ &= \int_\Omega \sum_{s\in 2^U} P_D(s;(Y,Z)) \cdot g_S(s,Y,Z)\, dP^N_{\mathcal{M}_\theta} \\ &= \int_\Omega E_D\left[g_S(S,Y,Z)\right]\, d\, P^N_{\mathcal{M}_\theta} \\ &= E_{\mathcal{M}_\theta}\left[E_D\left[g_S(S,Y,Z)\right]\right] \quad . \end{aligned} \tag{A.4}$$

The conditional expectation of g_S given (Y,Z) is therefore $E_{\mathcal{M}_\theta,D}\left[g_S(S,Y,Z)\,|(Y,Z)\right] = E_D\left[g_S(S,Y,Z)\right]$ and $\text{Var}_{\mathcal{M}_\theta,D}\left[g_S(S,Y,Z)\,|(Y,Z)\right] = \text{Var}_D\left[g_S(S,Y,Z)\right]$. We get by the law of total variance

$$\begin{aligned} \text{Var}_{\mathcal{M}_\theta,D}\left[g_S(S,Y,Z)\right] &= \text{Var}_{\mathcal{M}_\theta,D}\left[E_{\mathcal{M}_\theta,D}\left[g_S(S,Y,Z)\,|(Y,Z)\right]\right] \\ &\quad + E_{\mathcal{M}_\theta,D}\left[\text{Var}_{\mathcal{M}_\theta,D}\left[g_S(S,Y,Z)\,|(Y,Z)\right]\right] \\ &= \text{Var}_{\mathcal{M}_\theta,D}\left[E_D\left[g_S(S,Y,Z)\right]\right] + E_{\mathcal{M}_\theta,D}\left[\text{Var}_D\left[g_S(S,Y,Z)\right]\right] \\ &= \text{Var}_{\mathcal{M}_\theta}\left[E_D\left[g_S(S,Y,Z)\right]\right] + E_{\mathcal{M}_\theta}\left[\text{Var}_D\left[g_S(S,Y,Z)\right]\right] \end{aligned} \tag{A.5}$$

and for a design unbiased estimator g_S for the statistic $g_U(Y)$

$$= \mathrm{Var}_{\mathcal{M}_\theta}\left[g_U(Y)\right] + \mathrm{E}_{\mathcal{M}_\theta}\left[\mathrm{Var}_D\left[g_S(S,Y,Z)\right]\right] \quad . \tag{A.6}$$

□

UNIQUENESS OF ML VARIANCE-COVARIANCE MATRIX

Assume that the random vector G can be written as

$$G^T = (G_{11}^T, \ldots, G_{1q_1}^T, G_{21}^T, \ldots, G_{2q_2}^T, \ldots, G_{kq_k}^T)$$

where q_j represents the q_j-th repetition of a random vector G_{j1} (meaning that $G_{j,m_1} \sim G_{j,m_2}$, $m_1, m_2 = 1, \ldots, q_j$) and $\sum_{j=1}^{k} q_j \cdot |G_{j1}| = q$. Then, Σ has the following structure

$$\Sigma = \mathrm{diag}\left(\underbrace{\Sigma_1, \ldots, \Sigma_1}_{q_1\text{-times}}, \ldots, \underbrace{\Sigma_k, \ldots, \Sigma_k}_{q_k\text{-times}}\right)$$

where Σ_j has the dimension of $|G_{j1}| \times |G_{j1}|$. By choosing an adequate size of the sub-vectors G_{jm}, it is possible to assume them pairwise independent. Then, $\boldsymbol{\rho}^T \triangleq (\boldsymbol{\rho}_1^T, \ldots, \boldsymbol{\rho}_k^T)$ where $\boldsymbol{\rho}_j$ collects the unique elements from Σ_j of maximal size $\frac{(\dim \Sigma_j)\cdot(\dim \Sigma_j + 1)}{2}$. This dimension is achieved when no additional assumptions on Σ_j other than symmetry are imposed. Due to the independency, the problem thus reduces to ML estimation of $\Sigma_1, \ldots, \Sigma_k$ with random vector realizations of $G_{11}, \ldots, G_{1q_1}$ to $G_{k1}, \ldots, G_{kq_k}$ with known expectation o. The log-likelihood under multivariate normality for the random variables $G_{j1}, \ldots, G_{jq_j}$ is therefore

$$-\frac{q_j \cdot \dim \Sigma_j}{2}\log(2\pi) - \frac{q_j}{2}\log\det\Sigma_j - \frac{1}{2}\sum_{\nu=1}^{q_j} \boldsymbol{\gamma}_{j\nu}^T \Sigma_j^{-1} \boldsymbol{\gamma}_{j\nu} \quad .$$

A symmetric deviation $d\Sigma_j$ such that $\Sigma_j - d\Sigma_j$ remains positive definite yields the following differential change in the log-likelihood

$$-\frac{q_j}{2}\,\mathrm{tr}\left(\Sigma_j^{-1} d\Sigma_j\right) + \frac{1}{2}\sum_{\nu=1}^{q_j} \boldsymbol{\gamma}_{j\nu}^T \Sigma_j^{-1} d\Sigma_j \Sigma_j^{-1} \boldsymbol{\gamma}_{j\nu}$$

due to Jacobi's formula and the fact that $\mathrm{adj}(\Sigma_j) = (\det\Sigma_j)\Sigma_j^{-1}$. This differential equals

$$\frac{q_j}{2}\,\mathrm{tr}\left(\Sigma_j^{-1} d\Sigma_j \left(q_j I - \Sigma_j^{-1}\sum_{\nu=1}^{q_j} \boldsymbol{\gamma}_{j\nu}\boldsymbol{\gamma}_{j\nu}^T\right)\right)$$

and this term equals zero for arbitrary small $d\Sigma_j$ if and only if the second factor equals zero. Thus, the (symmetric) ML estimator $\frac{1}{q_j}\sum_{\nu=1}^{q_j}\sum_{\nu=1}^{q_j} \boldsymbol{\gamma}_{j\nu}\boldsymbol{\gamma}_{j\nu}^T$

is unique. We do not prove here that the second order conditions hold, too. Then, however, as the solution for $\hat{\rho}_j$ is unique for all $j = 1, \ldots, k$, the solution for $\hat{\rho} = (\hat{\rho}_1^T, \ldots, \hat{\rho}_k^T)^T$ is unique, too.

□

(STRONG) ML CONSISTENCY FOR MIXED MODELS UNDER BOX-COX AND DUAL TRANSFORMATIONS

Under both, the Box-Cox and the Dual transformation, the population joint log-likelihood $\mathcal{LL}$ meets the requirements of Theorem 1 in Rubin [1956] if the vector Y is splitted like in the previous proof and we assume $\theta \in \Theta$ with Θ compact:

For the Box-Cox transformation, nothing changes from the derivation in Hernandez and Johnson [1980a], except for the added term B (cf. Equation (2.12)), which is bounded by an integrable function and equicontinuous in ρ. Note that this requires to bound all parameters β, λ and σ^2 on (multivariate) intervals. A good definition, however, of these intervals can be compplicated as already mentioned for the parameter λ in the main text.

In a next step, we check the conditions required in Rubin [1956] for the Dual transformation. This is easily verified because Rubin's conditions (1956) were studied for Box-Cox transformations in Hernandez and Johnson [1980b] and for any compact Θ such that the compact interval for λ is symmetric around 0, we have for $\lambda geq 0$ that

$$h_D(\cdot;\lambda) = \begin{cases} \frac{h_{BC}(;\lambda) - h_{BC}(;-\lambda)}{2}, & \text{if } \lambda > 0 \\ h_{BC}(;\lambda), & \text{if } \lambda = 0 \end{cases} . \tag{A.7}$$

Obviously, $\mathcal{LL}$ is therefore continuous in θ and measurable in Y's sample space also under this transformation. The equicontinuity required in Rubin [1956] follows from the same argument like in Hernandez and Johnson [1980a]. As the likelihood differs from a classical LMM only by the sum of logarithmic first order derivatives of the transformation h, we get due to the triangle inequality as an uppder bound for $\mathcal{LL}$

$$\text{const.} + \sum_{i \in U} |\log \frac{\partial h_D(y_i;\lambda)}{\partial y_i}|$$

and this term is $P^N_{\mathcal{M}_\theta}$-integrable as the logarithm is concave and

$$\begin{aligned}
&\int \sum_{i\in U} \left| \log \frac{\partial h_D(Y_i;\lambda)}{\partial Y_i} \right| \, \mathrm{d}\, P_{\mathcal{M}_\theta} \\
\leqslant \lambda + &\int \sum_{i\in U} \left| \mathbb{1}_{\{Y_i<1\}}(-\lambda-1)\log Y_i + \mathbb{1}_{\{Y_i\geqslant 1\}}(\lambda-1)\log Y_i \right| \, \mathrm{d}\, P^N_{\mathcal{M}_\theta} \\
\leqslant &\int \sum_{i\in U} \left| \log Y_i \, \mathrm{d}\, P^N_{\mathcal{M}_\theta} + \int \mathbb{1}_{\{Y_i\geqslant 1\}}\lambda \log Y_i - \lambda \mathbb{1}_{\{Y_i<1\}} \log Y_i \right| \, \mathrm{d}\, P^N_{\mathcal{M}_\theta} \\
&\leqslant \int \sum_{i\in U} (1+\lambda) \int |\log Y_i| \, \mathrm{d}\, P_{\mathcal{M}_\theta}
\end{aligned}$$

and because

$$\begin{aligned}
&\mathrm{E}_{\mathcal{M}_\theta}\left[\frac{1}{N}\sum_{i\in U} \lambda \log Y_i\right] = \mathrm{E}_{\mathcal{M}_\theta}\left[\frac{1}{N}\sum_{i\in U} \log Y_i^\lambda\right] \\
\leqslant \mathrm{E}_{\mathcal{M}_\theta}&\left[\log\left(\frac{1}{N}\sum_{i\in U} Y_i^\lambda\right)\right] \leqslant \log \mathrm{E}_{\mathcal{M}_\theta}\left[\frac{1}{N}\sum_{i\in U} Y_i^\lambda\right],
\end{aligned}$$

the population joint log-likelihood is $P^N_{\mathcal{M}_\theta}$-integrable as the moments of Y_i^λ exist: We have correspondence between the λ moment of Y_i and the moments of Box-Cox transformed distributions (cf. Equation (A.7)). The latter exist [Freeman and Modarres, 2006]. Hence, Theorem 1 from Rubin [1956] is applicable and therefore, the average population log-likelihood converges (for an increasing number of groupings in G) almost surely to the expected joint log-likelihood.

As arg max is a continuous function for $\mathcal{LL}$ ($\mathcal{LL}$ is twice continuously differentiable and thus, the implicit function $\nabla_\theta \mathcal{LL} = 0$ which is equivalent to arg max is continuous), $\arg\max \mathcal{LL}$ converges with probability one, too, within compact Θ.

□

MODEL-DESIGN UNBIASEDNESS OF $\widehat{\mathcal{LL}}$

First, note that for any $\mathbf{a} \in \mathbb{R}^N$, $a_i \neq 0$ for all $i = 1, \ldots, N$, it holds that

$$\mathrm{E}_D\left[\mathbf{a}^\mathsf{T}\mathbf{1}_S\right] = \mathbf{a}^\mathsf{T}\mathrm{E}_D\left[\mathbf{1}_S\right] = \sum_{i\in U} a_i \mathrm{E}_D\left[\mathbb{1}_S(i)\right] = \sum_{i\in U} a_i \cdot P_D(i \in S)$$

and consequently, we get, according to the DGP (2.9) that $G \perp (X, Z)$,

$$\begin{aligned}
\mathrm{E}_{\mathcal{M}_\theta, D}\left[\widehat{\mathcal{LL}}|X, Z\right] &= \int_{\mathcal{S}\times\mathcal{Y}_N} \widehat{\mathcal{LL}}\, \mathrm{d}\, \mathrm{P}_{\mathcal{M}_\theta, D}(\cdot|X, Z) \\
&\overset{\text{Eq.(1.13)}}{=} \sum_{s\in\mathcal{S}} \int_{\mathcal{Y}_N} \mathrm{P}_D(s; Y, X, Z) \cdot \widehat{\mathcal{LL}}\, \mathrm{d}\, \mathrm{P}_{\mathcal{M}_\theta}(\cdot|X, Z) \\
&= \int_{\mathcal{Y}_N} \sum_{s\in\mathcal{S}} \left(\sum_{i\in U} \mathbb{1}_s(i) \cdot w_i \cdot \nu_i \right) \cdot \mathrm{P}_D(s; Y, X, Z)\, \mathrm{d}\, \mathrm{P}_{\mathcal{M}_\theta}(\cdot|X, Z) \\
&\quad + \int_{\mathcal{Y}_N} \sum_{s\in\mathcal{S}} B \cdot \mathrm{P}_D(s; Y, X, Z)\, \mathrm{d}\mathrm{P}_{\mathcal{M}_\theta}(\cdot|X, Z) \\
&= \int_{\Omega} \sum_{i\in U} \mathrm{E}_D\left[\mathbb{1}_S(i)\right] \cdot w_i \cdot \nu_i\, \mathrm{d}\, \mathrm{P}_{\mathcal{M}_\theta} + \int_{\mathbb{R}^N} B\, \mathrm{P}_{\mathcal{M}_\theta}(G|X, Z) \\
&= \int_{\mathcal{Y}_N} \sum_{i\in U} \nu_i\, \mathrm{d}\, \mathrm{P}_{\mathcal{M}_\theta}(\cdot|X, Z) + \mathrm{E}_{\mathcal{M}_\theta}[B] \\
&= \mathrm{E}_{\mathcal{M}_\theta}\left[\widehat{\mathcal{LL}}|X, Z\right] \qquad \text{(A.8)}
\end{aligned}$$

□

UNWEIGHTED ML OPTIMIZATION OF MIXED MODELS UNDER POWER TRANSFORMATIONS

Gurka et al. [2006] suggest for mixed model estimation under Box-Cox transformations a rescaling procedure in order to apply standard statistical software. In the following, we ignore survey design in order to focus on the problem of that rescaling procedure. The dependent variable Y is in that setting divided by $\left(\prod_{i=1}^{N} \frac{\mathrm{d}\, h(y_i;\lambda)}{\mathrm{d}\, y_i}\right) =: \tilde{Y}$ in order to make the log-likelihood criterion of the regression model resemble the classical LMM log-likelihood. This is, however, erronuous: The first order condition for $\hat{\boldsymbol{\beta}}$ are for the rescaled problem in Gurka et al. [2006, Equation (7)], using our notation

$$\mathbf{x}^\top \left(\mathbf{z}\Sigma\mathbf{z}^\top + \sigma^2 I_N\right)^{-1} (h(\mathbf{y};\lambda)/\tilde{y} - \mathbf{x}\boldsymbol{\beta}) = 0$$

and for the original problem Equation (4) in Gurka et al. [2006]

$$\mathbf{x}^\top \left(\mathbf{z}\Sigma\mathbf{z}^\top + \sigma^2 I_N\right)^{-1} (h(\mathbf{y}) - \mathbf{x}\boldsymbol{\beta}) = 0 \quad .$$

Conditional on λ, the estimator $\hat{\boldsymbol{\beta}}$ can therefore be restablished from the rescaled problem by multiplication with $\tilde{y}$. If therefore the optimal solutions for λ are also identical, we have the relation stated in Gurka et al.

[2006] and Rojas-Perilla et al. [2017]. However, the first order conditions for the estimation of λ are for the rescaled problem

$$\left(\sum_{j=1}^{N} \frac{\partial^2 h(y_j;\lambda)}{\partial y_j \partial \lambda} \prod_{\substack{i=1 \\ i\neq j}}^{N} \frac{\partial h(y_i;\lambda)}{\partial y_i}\right) \nabla_\lambda h(\mathbf{y};\lambda) \cdot \left(\mathbf{z}\Sigma\mathbf{z}^T + \sigma^2 I_N\right)^{-1} (h(\mathbf{y};\lambda)/\tilde{y} - \mathbf{x}\boldsymbol{\beta}) = 0$$

and for the original problem

$$\nabla_\lambda h(\mathbf{y};\lambda) \left(\mathbf{z}\Sigma\mathbf{z}^T + \sigma^2 I_N\right)^{-1} (h(\mathbf{y};\lambda)/\tilde{y} - \mathbf{x}\boldsymbol{\beta}) + \sum_{i=1}^{N} \frac{1}{\frac{\partial h(y_i;\lambda)}{\partial y_i}} \cdot \frac{\partial^2 h(y_i;\lambda)}{\partial y_i \partial \lambda} = 0$$

If λ_0 was a solution to the first and second equation, we would get - plugging in this into the first equation -

$$-\left(\sum_{j=1}^{N} \frac{\partial^2 h(y_j;\lambda_0)}{\partial y_j \partial \lambda_0} \prod_{\substack{i=1 \\ i\neq j}}^{N} \frac{\partial h(y_i;\lambda_0)}{\partial y_i}\right) \cdot \left(\sum_{i=1}^{N} \frac{1}{\frac{\partial h(y_i;\lambda_0)}{\partial y_i}} \cdot \frac{\partial^2 h(y_i;\lambda_0)}{\partial y_i \partial \lambda_0}\right) = -\sum_{k,j=1}^{N} \frac{\partial^2 h(y_j;\lambda_0)}{\partial y_j \partial \lambda_0} \cdot \frac{\partial^2 h(y_k;\lambda_0)}{\partial y_k \partial \lambda_0} \cdot \frac{1}{\frac{\partial h(y_k;\lambda_0)}{\partial y_k}} \cdot \prod_{\substack{i=1 \\ i\neq j}}^{N} \frac{\partial h(y_i;\lambda)}{\partial y_i}$$

which is generally unequal to zero for arbitrary realizations from the sample space and parameter λ_0 as the case $N = 2$ demonstrates: In that case the sum above equals

$$-\left(\left(\frac{\partial^2 h(y_1;\lambda_0)}{\partial y_1 \partial \lambda_0}\right)^2 \cdot \frac{\frac{\partial h(y_2;\lambda_0)}{\partial y_2}}{\frac{\partial h(y_1;\lambda_0)}{\partial y_1}} + 2\frac{\partial^2 h(y_1;\lambda_0)}{\partial y_1 \partial \lambda_0} \cdot \frac{\partial^2 h(y_2;\lambda_0)}{\partial y_2 \partial \lambda_0} + \left(\frac{\partial^2 h(y_2;\lambda_0)}{\partial y_2 \partial \lambda_0}\right)^2 \cdot \frac{\frac{\partial h(y_1;\lambda_0)}{\partial y_1}}{\frac{\partial h(y_2;\lambda_0)}{\partial y_2}}\right) \neq 0$$

and therefore λ_0 cannot be an interior solution to the rescaled optimization problem and it results a contradiction.

ON-LINE SOM CONVERGENCE IMPLIES BATCH SOM CONVERGENCE

Fort et al. [2001] considers convergence in expectation: Assume that $E\left[\mathbf{u}_j^{(\kappa+1)} - \mathbf{u}_j^{(\kappa)}\right]$ holds for all $j \in V_G$. This means for all nodes $j \in$

V_G and their neighbors ι defined by the criterion $h(\delta(j,\iota)) > 0$, $\mathcal{N}_j := \cup_{\iota \in V_G : h(\delta(j,\iota))>0}\{\iota\}$ that

$$0 = \sum_{\iota \in \mathcal{N}_j} h(\delta(j,\iota)) \cdot P\left(Y_i \in V_\iota^{(\kappa)}\right) \cdot E\left[Y_i - \mathbf{u}_j^{(\kappa)} | Y_i \in V_\iota^{(\kappa-1)}\right]$$

implying

$$\mathbf{u}_j^{(\kappa)} = \frac{\sum_{\iota \in \mathcal{N}_j} h(\delta(j,\iota)) \cdot P\left(Y_i \in V_\iota^{(\kappa)}\right) \cdot E\left[Y_i | Y_i \in V_\iota^{(\kappa-1)}\right]}{\sum_{\iota \in \mathcal{N}_j} h(\delta(j,\iota))} \cdot P\left(Y_i \in V_\iota^{(\kappa)}\right) \quad , \tag{A.9}$$

where the superscript on the Voronoi region indicates their dependence on $U^{(\kappa-1)}$. Fort et al. [2001] argues now that for bounded Y_i, the feature vectors are bounded and then there exists a convergent sub-sequence of (A.9). This convergent sub-sequence then goes to a stationary point of the on-line Batch algorithm. This, however, does not imply convergence of the overall sequence. But $\sum_{\iota \in \mathcal{N}_j} h(\delta(j,\iota)) \cdot E\left[Y_i - \mathbf{u}_j^{(\kappa)} | Y_i \in V_\iota^{(\kappa-1)}\right]$ can be seen as the local gradient of $\frac{1}{2}\sum_{\iota \in \mathcal{N}_j} h(\delta(j,\iota)) \cdot E\left[\|Y_i - \mathbf{u}_j^{(\kappa)}\|^2 | Y_i \in V_\iota^{(\kappa-1)}\right]$ and the latter is locally Lipschitz continuous in $E[Y_i | Y_i \in V_\iota]$. Then, the Picard-Lindelöf Theorem is applicable, proving the existence of a locally unique solution for $U^{(\kappa)}$, which then equals (A.9). The update rule (4.3) can then be considered to be an empirical version of (A.9).

□

DISCUSSION OF THEOREM 4.1 IN YIN AND ALLINSON [1995]

Theorem 4.1 in Yin and Allinson [1995] relies on Equation A.1 in Yin and Allinson [1995] that states for some $\beta \in (0,1)$ and $\delta > 0$

$$\exp(iX) = 1 + iX - \frac{X^2}{2} + \beta\frac{|X|^{2+\delta}}{2^\delta} \quad ,$$

where X is considered to be a univariate, real-valued random variable and $i = \sqrt{-1}$. This is, however, not true because $|\exp(iX)|^2 = 1$ for any value of X and

$$\left|1 + iX - \frac{X^2}{2} + \beta\frac{|X|^{2+\delta}}{2^\delta}\right|^2 = \left(1 - \frac{X^2}{2} + \beta\frac{|X|^{2+\delta}}{2^\delta}\right)^2 + X^2 \neq 1 \quad .$$

This questions their complete proof.

□

ADAPTIVE DISTANCES FOR SOMS

We show that $d = \{d_{\Lambda_j}\}_{j \in \mathcal{G}}$ meets the conditions of a metric. Take thus $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^p$.

1. The symmetry of d follows from the fact that each d_{Λ_j} is symmetric.

2. Obviously, for all $j \in \mathcal{G}$, $d_{\Lambda_j}(\mathbf{x}, \mathbf{y}) \geqslant 0$. If $\mathbf{x} = \mathbf{y}$, then they fall into the same Voronoi region ι and thus $d(\mathbf{x}, \mathbf{y}) = d_{\Lambda_\iota}(\mathbf{x}, \mathbf{y}) = 0$. If, on the other hand $d(\mathbf{x}, \mathbf{y}) = 0$, then at least for one $\iota \in \mathcal{G}$, $d_{\Lambda_\iota}(\mathbf{x}, \mathbf{y}) = 0$ and consequently, $\mathbf{x} = \mathbf{y}$.

3. Assume that $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ fall into the same Voronoi region j. Then the triangle inequality holds:

$$d(\mathbf{x}, \mathbf{y}) = d_{\Lambda_j}(\mathbf{x}, \mathbf{y}) \leqslant d_{\Lambda_j}(\mathbf{x}, \mathbf{z}) + d_{\Lambda_j}(\mathbf{z}, \mathbf{y}) = d(\mathbf{x}, \mathbf{y}) + d(\mathbf{z}, \mathbf{y})$$

Assume now that $\mathbf{x} \in V_j$ and $\mathbf{y} \in V_j$ and $\mathbf{z} \in V_\iota$, $j \neq \iota$.

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) = d_{\Lambda_j}(\mathbf{x}, \mathbf{y}) &\leqslant d_{\Lambda_j}(\mathbf{x}, \mathbf{u}_j) + d_{\Lambda_j}(\mathbf{u}_j, \mathbf{z}) + 2\|\mathbf{u}_j - \mathbf{u}_\iota\|^2 \\ &\quad + 2d_{\Lambda_\iota}(\mathbf{u}_\iota, \mathbf{z}) \\ &= d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}) \end{aligned}$$

and

$$\begin{aligned} d(\mathbf{x}, \mathbf{z}) &= d_{\Lambda_j}(\mathbf{x}, \mathbf{u}_j) + d_{\Lambda_\iota}(\mathbf{z}, \mathbf{u}_\iota) + \|\mathbf{u}_j - \mathbf{u}_\iota\|^2 \\ &\leqslant d_{\Lambda}(\mathbf{x}, \mathbf{y}) + d_{\Lambda_j}(\mathbf{y}, \mathbf{u}_j) + d_{\Lambda_\iota}(\mathbf{z}, \mathbf{u}_\iota) + \|\mathbf{u}_j - \mathbf{u}_\iota\|^2 \\ &= d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \quad . \end{aligned}$$

If, on the other hand, $\mathbf{x} \in V_j$, $\mathbf{y} \in V_l$ and $\mathbf{z} \in V_\iota$ with $j \neq \iota$ and $\iota \neq l$ and $j \neq l$

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= d_{\Lambda_j}(\mathbf{x}, \mathbf{u}_j) + d_{\Lambda_l}(\mathbf{y}, \mathbf{u}_l) + \|\mathbf{u}_j - \mathbf{u}_l\|^2 \\ &\leqslant d_{\Lambda_j}(\mathbf{x}, \mathbf{u}_j) + d_{\Lambda_l}(\mathbf{y}, \mathbf{u}_l) + 2d_{\Lambda_\iota}(\mathbf{z}, \mathbf{u}_\iota) + \|\mathbf{u}_j - \mathbf{u}_\iota\|^2 \\ &\quad + \|\mathbf{u}_\iota - \mathbf{u}_l\|^2 \\ &= d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}) \quad . \end{aligned}$$

Consequently, d is a metric.

□

PROPERTIES OF THE DISTORTION MEASURE IN TERMS OF MRFS

In this paragraph, we extend the Assumption 1 of Kondratiev et al. [2010] to the multivariate configuration space (as $\mathcal{Y} \subset \mathbb{R}^p$) and check whether these assumptions hold for the distortion measure. For finite graphs, part i) of Assumption 1 is trivial. In part ii), it is required that for every $u, v \in \mathcal{Y}$ it holds for some $I_w, J_w, r > 0$ that

$$|W(u, v)| \leqslant \frac{1}{2}\left(I_w + J_w\left(|u|^r + |v|^r\right)\right) \quad ,$$

where $|\cdot|$ denotes the Eucledian norm. Note that there is a misprint in Kondratiev et al. [2010, Equation 6], as can be learned from Kondratiev et al. [2010, Equation 22], Pasurek [2007, Equation 2.6] and Daletskii et al. [2014, Equation 3.1]. As here, $W \equiv 0$, this holds for arbitrary small r.

Then there remains part iii). For finite graphs, the edge degree is bounded and therefore the requirement translates to

$$\tilde{E}(Y; U, V) \geqslant a|Y|^q - c$$

for fixed $a, c > 0$ and $q > r$. This is, however, unproblematic for $q \geqslant 2$ as $\tilde{E}$ is quadratic for each site $j \in V_G$.

We state thus, that the only requirement that we have to pose on P is that there is no mass point at 0. The proofs in Kondratiev et al. [2010] require furthermore (applied to our Hamiltonian) that $Y \in L^p(\mathcal{Y}, \mathcal{A}_y, P)$ for every $p \in (0, 2)$, that is, the assumptions hold when the second moments of Y exist.

LOCAL CONVERGENCE FOR THE ITERATIVE SOLUTION ON (4.29)

Consider the optimization procedure based on (4.29) for all $i \in S$ and $j \in V_G$

$$\mathbf{u}_j^{(\kappa+1)} = \frac{\sum_{i\in S}\sum_{\iota\in V_G} h(\delta(j,\iota)) \cdot p_\iota^{i(\kappa)} \cdot Y_i}{\sum_{i\in S}\sum_{\iota\in V_G} h(\delta(j,\iota)) \cdot p_\iota^{i(\kappa)}} \tag{A.10a}$$

$$p_j^{i(\kappa+1)} = \frac{\exp\left(-\frac{1}{T}\sum_{\iota\in V_G} h(\delta(j,\iota)) \cdot d\left(Y_i, \mathbf{u}_\iota^{(\kappa+1)}\right)\right)}{\sum_{l\in V_G}\exp\left(-\frac{1}{T}\sum_{\iota\in V_G} h(\delta(l,\iota)) \cdot d\left(Y_i, \mathbf{u}_\iota^{(\kappa+1)}\right)\right)} \quad . \tag{A.10b}$$

Assume with metric $\|\cdot\|_\Lambda$ the DGP (up to a constant 'const.') to be the GMM

$$\sum_{j\in V_G} p_j \cdot \exp\left(-\frac{c_j}{T} d(Y, \bar{\mathbf{u}}_j)\right) \tag{A.11a}$$

meaning that

$$Y|\mathbb{1}_j \sim N\left(\bar{\mathbf{u}}_j, \frac{T}{c_j}\Lambda^{-1}\right) \tag{A.11b}$$

$$\{\mathbb{1}_j\}_{j\in V_G} \sim \text{Multi}(1, p_j : j \in V_G) \quad , \tag{A.11c}$$

where it is important that $\mathbb{1}_j \neq \mathbb{1}_{V_j}(Y)$ and the assignment $\mathbb{1}_j$ be independent from U and Definition of c_j and $\bar{\mathbf{u}}_j$ obey (4.19). With a penalty term

$$-\frac{1}{T}\sum_{j\in V_G}\mathbb{1}_j \sum_{\iota\in V_G} h(\delta(j,\iota)) \cdot d(\mathbf{u}_\iota, \bar{\mathbf{u}}_j) \quad , \tag{A.11d}$$

we get optimization problem (4.23) as the expected penalized log-likelihood. The update rules (A.10) then correspond to the steps of an expected maximum penalized likelihood algorithm. Hero and Fessler [1995, Theorem 1] discuss convergence properties of such EM type algorithms (which can be achieved locally under certain regularity conditions). This means that we achieve (local) minimization of the penalized log-likelihood. Note that due to (4.27), this minimum yields again with $T = 1$ the negative logarithm of the Gibbs measure's density.

B

ADDITIONAL INFORMATION ON (SIMULATION) DATA

Table B.1: Overview on the Electronic Appendices

The codes are available on https://github.com/s4padoer/design-estimation-business-surveys/

Folder	File	Description
chapter2	sim_mixed_effects_regression.R	Simulation set-up described in chapter 2
chapter2	merge_mixed_effects_regression.R	Editing of the simulation results from chapter 2
chapter2	optim_lmer.R	Function doing a grid search to find the best Box-Cox/ Dual transformation with respect to the data log-likelihood
chapter3	sim_modelbased_gvf.R	Simulation set-up described in chapter 3
chapter3	merge_modelbased_gvf.R	Editing of the simulation results from chapter 3
chapter3	varfu.R	Variance estimators for SRS, StratRS and TSC. Functions doing balanced half-sampling for the quality indicator of Gershunskaya and Dorfman [2013].
chapter4	sim_density_estimation.R	Simulation set-up described in chapter 4
chapter4	merge_density_estimation.R	Editing of the simulation results from chapter 4
chapter4	merged_density_estimation.RData	Data file storing the edited simulation results from chapter 4

Table B.1: Overview on the Electronic Appendices

Folder	File	Description
chapter4	som_ simplex.cpp	C++ functions to speed up training the topographic map (Algorithm 4.3)
chapter4	som_simplex.R	R function implementing the topographic map (Algorithm 4.3)
chapter4	distributions_ som_simplex.cpp	Function returning marginal and joint (and some conditional) densities based on Estimator (4.20)
application_beeps	edit_data_beeps.R	Editing of the PUM
application_beeps	beeps_pre_ regression.R	Model selection and data analysis using lme4
application_beeps	beeps_regression.R	Code for running the MCEM based estimation method 100 times
application_beeps	merge_beeps_ regression.R	Editing and plotting of the 100 regression estimates
application_beeps	AppliedLoans.RData	Data file storing the edited subsets of the PUM
application_beeps	lme4_ regressions.RData	Data file storing the chosen regressions estimated with lme4
application_beeps	BEEPS_Panel_2002_ 2005_2009_ 19 Aug 2010.dta	Original PUM
application_beeps	worldbank_exchange_ rate.csv	Official World Bank exchange rates from LCUs to US Dollar (downloaded 17 March 2020)
application_beeps	worldbank_exchange_ rate_edited.csv	Exchange rates from LCUs to US Dollar edited to merge with PUM
application_beeps	beeps_density.R	Code for runnig 100 times Estimator 4.20 on the BEEPS PUM

Table B.1: Overview on the Electronic Appendices

Folder	File	Description
application_beeps	beeps_pre_density.RData	Data file storing the bivariate kernel density and GMM density estimates for the bivariate total sales
application_beeps	beeps_density.RData	Data file storing all bivariate density estimates for the bivariate total sales
application_beeps	merge_beeps_density.R	Code for merging the 100 estimates of the bivariate density of total sales
lib	wmm	Self-authored R-package wmm running the weighted mixed effects model estimation

Table B.2: Partitions in the Simulation Study in Chapter 2

u_d \ u_g	1	2	3	4	5	6	7	8	9	10
1	3	9	3	3	1	9	8	7	7	9
2	2	2	6	4	3	4	8	6	13	7
3	4	2	3	1	3	9	8	9	5	10
4	1	5	2	4	1	11	9	11	8	8
5	5	1	1	0	3	7	5	6	6	9
6	5	4	3	5	2	7	4	5	2	9
7	1	2	1	2	1	4	9	10	9	11
8	5	3	2	3	3	5	11	8	4	11
9	1	4	3	3	2	7	10	6	11	8
10	1	1	4	2	5	5	5	6	4	9
11	4	2	1	4	4	7	4	7	6	4
12	5	2	2	0	3	6	7	7	11	6
13	1	1	3	1	1	7	6	3	7	10
14	2	2	1	0	2	10	12	7	7	5
15	0	0	3	2	1	3	5	8	5	11
16	3	5	3	0	2	8	6	11	5	10
17	4	4	2	2	2	12	12	10	6	6
18	1	0	2	2	4	8	4	11	11	5
19	1	4	3	3	0	9	8	5	8	8
20	4	4	2	2	2	8	3	6	12	10

Table B.3: Partitions in the Simulation Study in Chapter 3

μ mod5 \ ν	1	2	3	4	5
1	2	71	235	448	803
2	6	109	276	516	886
3	15	130	331	532	947
4	38	174	353	627	998
0	55	185	408	715	1140

Table B.4: MC Relative Error of GVF Prediction - Model (3.23a) - under SRS

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Using Indicators in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.022	-0.007	0.007	0.013	0.025	0.088
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.119	-0.053	-0.045	-0.045	-0.034	-0.016
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.049	-0.030	-0.024	-0.025	-0.019	0.005
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.046	-0.033	-0.026	-0.026	-0.022	-0.004
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.417	-0.414	-0.412	-0.412	-0.411	-0.409
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.616	-0.615	-0.614	-0.614	-0.614	-0.612
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.158	0.160	0.163	0.164	0.166	0.174
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.159	0.162	0.163	0.163	0.164	0.167
Using Indicators in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.022	-0.005	0.007	0.013	0.021	0.088
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.116	-0.053	-0.041	-0.044	-0.034	-0.016
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.048	-0.029	-0.021	-0.023	-0.017	0.008
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.042	-0.030	-0.024	-0.024	-0.020	-0.001
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.279	-0.274	-0.271	-0.272	-0.270	-0.267
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.354	-0.353	-0.352	-0.352	-0.351	-0.349
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.431	0.438	0.441	0.441	0.444	0.451
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.953	0.956	0.958	0.958	0.959	0.963

Table B.5: MC Relative Error of GVF Prediction - Model (3.23a) - under StratRS

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Using Indicators in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.066	-0.043	-0.033	-0.033	-0.020	0.004
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.066	-0.040	-0.032	-0.032	-0.024	-0.010
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.061	-0.040	-0.028	-0.031	-0.022	-0.002
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.050	-0.035	-0.028	-0.027	-0.019	0.001
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.427	-0.422	-0.421	-0.421	-0.420	-0.418
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.616	-0.615	-0.614	-0.614	-0.613	-0.611
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.155	0.159	0.161	0.161	0.163	0.167
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.157	0.160	0.161	0.162	0.162	0.167
Using Indicators in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.063	-0.041	-0.032	-0.032	-0.021	0.008
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.060	-0.036	-0.028	-0.028	-0.020	-0.005
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.060	-0.038	-0.027	-0.030	-0.021	-0.006
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.046	-0.032	-0.026	-0.025	-0.017	0.004
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.289	-0.282	-0.280	-0.280	-0.278	-0.274
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.354	-0.352	-0.351	-0.351	-0.350	-0.346
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.433	0.439	0.443	0.442	0.445	0.449
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.950	0.955	0.957	0.957	0.958	0.964

Table B.6: MC Relative Error of GVF Prediction - Model (3.23a) - under Stratified TSC

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Using Indicators in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.053	0.035	0.058	0.058	0.077	0.170
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.044	0.074	0.085	0.086	0.102	0.116
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.046	0.104	0.122	0.123	0.145	0.185
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.100	0.106	0.113	0.113	0.118	0.128
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	0.006	-0.003	0.000	-0.001	0.001	0.005
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.066	0.067	0.067	0.067	0.068	0.069
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.049	0.052	0.053	0.053	0.055	0.058
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.098	0.099	0.099	0.099	0.099	0.100
Using Indicators in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.032	0.040	0.065	0.062	0.085	0.166
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.052	0.080	0.091	0.093	0.108	0.130
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.057	0.115	0.130	0.131	0.148	0.205
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.102	0.109	0.116	0.116	0.120	0.133
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	0.070	0.073	0.076	0.075	0.077	0.082
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.099	0.100	0.101	0.101	0.101	0.102
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.129	0.132	0.133	0.133	0.134	0.138
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.132	0.133	0.133	0.133	0.134	0.134

Table B.7: MC Relative Error of GVF Prediction - Model (3.23b) - under SRS

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Using Indicators in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.038	-0.032	-0.030	-0.030	-0.029	-0.026
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.037	-0.035	-0.034	-0.034	-0.032	-0.026
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.035	-0.032	-0.031	-0.031	-0.030	-0.028
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.031	-0.029	-0.028	-0.028	-0.027	-0.024
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.413	-0.412	-0.412	-0.412	-0.412	-0.411
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.622	-0.622	-0.621	-0.621	-0.621	-0.621
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.172	0.173	0.173	0.173	0.174	0.175
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.144	0.144	0.144	0.144	0.145	0.145
Using Indicators in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.025	-0.019	-0.017	-0.018	-0.016	-0.012
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.027	-0.025	-0.024	-0.023	-0.022	-0.016
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.025	-0.022	-0.021	-0.021	-0.020	-0.018
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.021	-0.019	-0.018	-0.018	-0.017	-0.014
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.224	-0.223	-0.223	-0.223	-0.222	-0.221
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.369	-0.369	-0.369	-0.369	-0.368	-0.368
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.549	0.550	0.550	0.551	0.551	0.554
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.907	0.908	0.909	0.909	0.910	0.911

Table B.8: MC Relative Error of GVF Prediction - Model (3.23b) - under StratRS

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Using Indicators in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.046	-0.040	-0.038	-0.038	-0.037	-0.031
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.037	-0.034	-0.033	-0.033	-0.032	-0.027
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.033	-0.030	-0.030	-0.030	-0.028	-0.026
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.031	-0.028	-0.027	-0.027	-0.026	-0.021
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.421	-0.420	-0.420	-0.420	-0.419	-0.418
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.622	-0.621	-0.621	-0.621	-0.621	-0.621
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.170	0.171	0.172	0.171	0.172	0.173
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.142	0.142	0.143	0.143	0.143	0.143
Using Indicators in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.033	-0.026	-0.025	-0.025	-0.024	-0.017
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.025	-0.022	-0.021	-0.021	-0.020	-0.016
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.023	-0.020	-0.019	-0.019	-0.018	-0.015
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.020	-0.018	-0.016	-0.016	-0.015	-0.011
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.232	-0.231	-0.230	-0.230	-0.230	-0.228
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.368	-0.367	-0.367	-0.367	-0.367	-0.366
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.551	0.553	0.554	0.553	0.554	0.556
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.906	0.908	0.909	0.909	0.909	0.910

Table B.9: MC Relative Error of GVF Prediction - Model (3.23b) - under Stratified TSC

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Using Indicators in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	0.025	0.036	0.040	0.044	0.044	0.170
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.083	0.090	0.093	0.092	0.094	0.100
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.072	0.079	0.083	0.083	0.086	0.101
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.120	0.124	0.125	0.125	0.126	0.129
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.020	-0.019	-0.018	-0.018	-0.018	-0.016
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.075	0.076	0.077	0.077	0.077	0.078
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.033	0.034	0.035	0.035	0.035	0.037
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.107	0.108	0.109	0.109	0.109	0.109
Using Indicators in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	0.096	0.107	0.111	0.112	0.116	0.148
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.117	0.124	0.126	0.126	0.128	0.135
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.114	0.121	0.125	0.124	0.127	0.133
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.134	0.137	0.138	0.138	0.139	0.142
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	0.067	0.068	0.069	0.069	0.069	0.071
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.109	0.110	0.110	0.110	0.111	0.112
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.125	0.126	0.126	0.127	0.127	0.129
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.142	0.143	0.143	0.143	0.143	0.144

Table B.10: MC Relative Error of GVF Prediction - Model (3.23c) - under SRS

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.422	-0.422	-0.421	-0.421	-0.421	-0.420
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.617	-0.617	-0.617	-0.617	-0.616	-0.616
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.154	0.154	0.154	0.155	0.155	0.157
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.158	0.159	0.159	0.159	0.159	0.160
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-1.881	-0.268	-0.258	-0.286	-0.248	-0.165
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.95	-0.95	-0.95	-0.95	-0.95	-0.95
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.360	0.404	0.409	0.417	0.416	0.754
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.850	-0.849	-0.849	-0.849	-0.849	-0.848

Figure B.1: MC Performance of Direct Variance Estimators

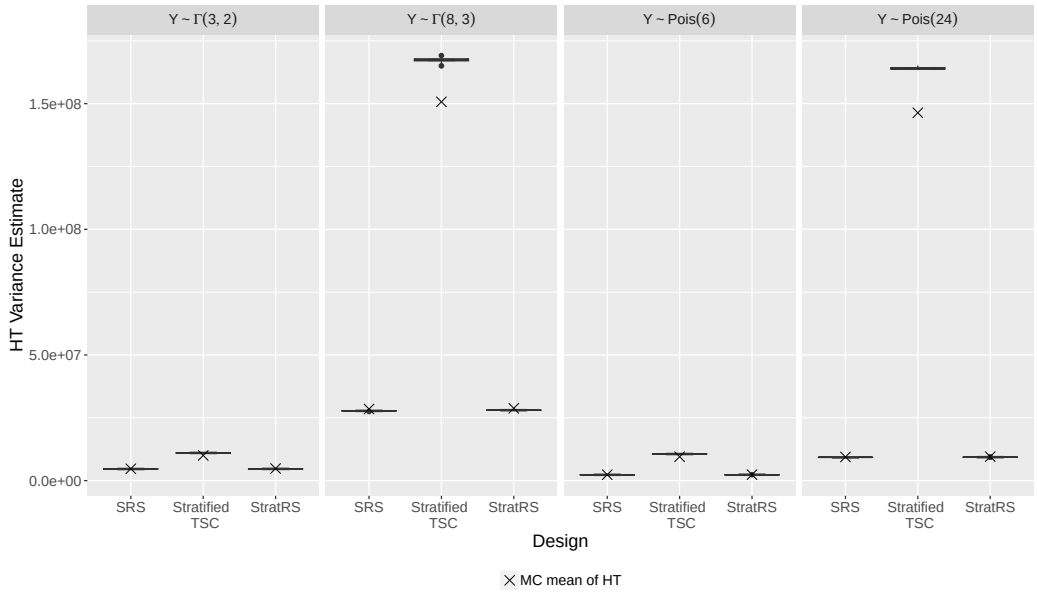


Table B.11: MC Relative Error of GVF Prediction - Model (3.23c) - under StratRS

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.429	-0.429	-0.429	-0.428	-0.428	-0.427
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.617	-0.616	-0.616	-0.616	-0.616	-0.616
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.152	0.153	0.154	0.153	0.154	0.155
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.156	0.157	0.157	0.157	0.158	0.158
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.637	-0.280	-0.266	-0.268	-0.257	-0.051
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.954	-0.954	-0.954	-0.954	-0.954	-0.953
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.335	0.400	0.407	0.404	0.413	0.435
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.861	-0.861	-0.861	-0.861	-0.860	-0.860

Table B.12: MC Relative Error of GVF Prediction - Model (3.23c) - under Stratified TSC

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.271	-0.105	-0.069	0.193	-0.036	11.759
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.863	-0.862	-0.862	-0.862	-0.862	-0.861
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.558	-0.074	-0.049	-0.058	-0.017	0.086
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.858	-0.858	-0.858	-0.858	-0.858	-0.857
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.495	-0.010	0.010	-0.003	0.028	0.237
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.894	-0.894	-0.894	-0.894	-0.894	-0.893
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.086	0.025	0.037	0.053	0.066	0.231
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.891	-0.891	-0.891	-0.891	-0.891	-0.890

Table B.13: MC Relative Error of GVF Prediction - Model (3.23d) - under SRS

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Using Indicators in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.032	-0.016	-0.007	-0.005	0.003	0.035
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.118	-0.053	-0.044	-0.045	-0.034	-0.016
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.049	-0.030	-0.024	-0.025	-0.018	0.005
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.046	-0.033	-0.027	-0.026	-0.022	-0.005
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.431	-0.419	-0.414	-0.415	-0.411	-0.402
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.611	-0.607	-0.606	-0.605	-0.604	-0.597
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.137	0.154	0.161	0.161	0.170	0.186
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.168	0.174	0.179	0.178	0.182	0.187
Using Indicators in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.022	-0.006	0.007	0.013	0.021	0.088
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.116	-0.053	-0.041	-0.044	-0.033	-0.016
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.048	-0.029	-0.021	-0.023	-0.017	0.008
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.042	-0.030	-0.024	-0.024	-0.020	-0.002
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.280	-0.270	-0.265	-0.265	-0.260	-0.247
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.369	-0.366	-0.364	-0.363	-0.361	-0.353
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.433	0.449	0.457	0.458	0.465	0.493
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.895	0.901	0.906	0.906	0.910	0.918

Table B.14: MC Relative Error of GVF Prediction - Model (3.23d) - under StratRS

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Using Indicators in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.066	-0.043	-0.033	-0.033	-0.020	0.004
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.065	-0.040	-0.032	-0.032	-0.024	-0.009
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.061	-0.040	-0.028	-0.031	-0.022	-0.002
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.050	-0.035	-0.028	-0.027	-0.019	0.001
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.441	-0.427	-0.424	-0.424	-0.420	-0.410
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.609	-0.606	-0.604	-0.604	-0.602	-0.597
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.139	0.156	0.162	0.163	0.167	0.191
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.168	0.177	0.181	0.181	0.185	0.189
Using Indicators in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.063	-0.041	-0.031	-0.032	-0.021	0.008
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.060	-0.036	-0.028	-0.028	-0.020	-0.005
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	-0.060	-0.038	-0.027	-0.030	-0.021	-0.006
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	-0.046	-0.032	-0.026	-0.025	-0.017	0.004
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.296	-0.278	-0.274	-0.274	-0.269	-0.253
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	-0.367	-0.362	-0.360	-0.360	-0.358	-0.351
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.429	0.454	0.460	0.462	0.469	0.497
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.900	0.910	0.914	0.914	0.918	0.924

Table B.15: MC Relative Error of GVF Prediction - Model (3.23d) - under Stratified TSC

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Using Indicators in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.053	0.035	0.058	0.058	0.077	0.169
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.045	0.074	0.085	0.086	0.102	0.116
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.046	0.104	0.122	0.123	0.145	0.185
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.100	0.106	0.113	0.113	0.118	0.128
Merging all Variables in the Model & Weighted OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.031	-0.019	-0.012	-0.012	-0.005	0.003
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.068	0.070	0.071	0.071	0.072	0.076
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.024	0.040	0.044	0.043	0.047	0.055
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.101	0.102	0.103	0.103	0.103	0.104
Using Indicators in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	-0.031	0.040	0.065	0.062	0.085	0.165
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.052	0.080	0.091	0.093	0.108	0.130
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.057	0.116	0.130	0.131	0.148	0.205
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.102	0.109	0.116	0.116	0.120	0.133
Merging all Variables in the Model & OLS						
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(3,2)\}$	0.051	0.069	0.073	0.074	0.082	0.092
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \mathcal{G}(8,3)\}$	0.098	0.099	0.100	0.101	0.102	0.1052
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(6)\}$	0.114	0.130	0.134	0.134	0.139	0.145
$\{j \in \mathcal{Q} : Y_i^{(j)} \sim \text{Pois}(24)\}$	0.131	0.132	0.133	0.133	0.133	0.134

BIBLIOGRAPHY

Damminda Alahakoon, Saman K. Halgamuge, and Bala Srinivasan. Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Transactions on Neural Networks*, 11(3):601–614, 2000.

James Alegria and Charles T. Scott. Generalized variance function applications in forestry. Research Note 345, US Department of Agriculture, 1991.

Safaa R. Amer. Neural network imputation in complex survey design, 2007. The year of authorship could not be verified.

Banchar Arnonkijpanich, Barbara Hammer, Alexander Hasenfuss, and Chidchanok Lursinsap. Matrix learning for topographic neural maps. In *International Conference on Artificial Neural Networks*, pages 572–582. Springer, 2008.

Douglas Bates. Computational methods for mixed models, 2018. URL https://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf. Vignette for lme4.

Marco Bee, Roberto Benedetti, and Giuseppe Espa. A framework for cut-off sampling in business survey design. Discussion Paper 6, Università degli Studi di Trento - Dipartimento di Economia, 2007.

Yves G. Berger. Rate of convergence to normal distribution for the horvitz-thompson estimator. *Journal of Statistical Planning and Inference*, 67(2):209 – 226, 1998. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(97)00107-9. URL http://www.sciencedirect.com/science/article/pii/S0378375897001079.

Francois Blayo. Kohonen self-organizing maps: Is the normalization necessary? *Complex Systems*, 6(6):105–123, 1992.

Hélène Boistard, Hendrik P. Lopuhaä, and Anne Ruiz-Gazen. Functional central limit theorems in survey sampling. *ArXiv e-prints*, 1509, 2015.

James G. Booth and James P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Methodology)*, 61(1):265–285, 1999. doi: 10.1111/1467-9868.00176. URL https://doi.org/10.1111/1467-9868.00176.

George E. P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodology)*, pages 211–252, 1964. doi: 10.2307/2984418. URL http://www.jstor.org/stable/2984418.

Jan Pablo Burgard and Patricia Dörr. Survey-weighted generalized linear mixed models. Research Papers in Economics 1, Trier University, 2018. URL https://www.uni-trier.de/fileadmin/fb4/prof/VWL/EWF/Research_Papers/2018-01.pdf.

Jan Pablo Burgard and Patricia Dörr. Survey-weighted unit-level small area estimation. In A. Abbruzzo, E. Brentari, M. Chiodi, and D. Piacentino, editors, *Book of Short Papers SIS 2018*, page 689. Pearson, 2018. ISBN 788891910233. URL http://meetings3.sis-statistica.org/index.php/sis2018/49th/paper/viewFile/1550/43.

Jan Pablo Burgard and Patricia Dörr. Survey-weighted generalized linear mixed models. Technical report, Trier University, 2019. Submitted to the Journal of Computational and Graphical Statistics.

Jan Pablo Burgard, Ralf Münnich, and Thomas Zimmermann. The impact of sampling designs on small area estimates for business data. *Journal of Official Statistics*, 30(4):749–771, 2014.

Ricardo Cao, José A. Vilar, and Juan M. Vilar. Generalised variance function estimation for binary variables in large-scale sample surveys. *Australian & New Zealand Journal of Statistics*, 54(3):301–324, 2012.

Sara Carter, Eleanor Shaw, Wing Lam, and Fiona Wilson. Gender, entrepreneurship, and bank lending: The criteria and processes used by bank loan officers in assessing applications. *Entrepreneurship Theory and Practice*, 31(3):427–444, 2007.

José E. Chacón and Tarn Duong. *Multivariate Kernel Smoothing and Its Applications*. Number 160 in Monographs on Statistics and Applied Probability. Taylor & Francis, 2018. `ks` R package version 1.11.7.

Hukum Chandra, Ray Chambers, and Nicola Salvati. Small area estimation of proportions in business surveys. Working paper series, University of Wollongong, 2009. URL http://ro.uow.edu.au/cssmwp/35.

Guillaume Chauvet. A note on the consistency of the narain-horvitz-thompson estimator. *arXiv preprint arXiv:1412.2887*, 2014. URL https://arxiv.org/pdf/1412.2887.pdf.

Elizabeth Chell and Susan Baines. Does gender affect business 'performance'? a study of microbusinesses in business services in the uk. *Entrepreneurship & Regional Development*, 10(2):117–135, 1998.

Moon J. Cho, John L. Eltinge, Julie Gershunskaya, and Larry Huff. Evaluation of generalized variance function estimators for the us current employment survey. In *Proceedings of the American Statistical Association Section on Survey Research Methods*, pages 534–539, 2002.

Moon J. Cho, John L. Eltinge, Julie Gershunskaya, and Larry Huff. Evaluation of generalized variance functions in the analysis of complex survey data. *Journal of Official Statistics*, 30(1):63–90, 2014. doi: 10.2478/jos-2014-0004.

Kennon R. Copeland, C. Gaughan, and Chris Boardman. Generalized variance functions to create stable and timely variance estimates for prescription count estimates. In *Proceedings of the American Statistical Association Section on Survey Research Methods*, 2006.

Marie Cottrell, Madalina Olteanu, Fabrice Rossi, and Nathalie Villa-Vialaneix. Theoretical and applied aspects of the self-organizing maps. In *Advances in Self-organizing Maps and Learning Vector Quantization*, pages 3–26. Springer, 2016.

Brenda G. Cox and B. Nanjamma Chinnappa. *Business Survey Methods*, chapter Unique Features of Business Surveys, pages 1–17. Probability and Mathematical Statistics. Wiley, 1995.

Jan Ćwik and Jacek Koronacki. Probability density estimation using a gaussian clustering algorithm. *Neural Computing & Applications*, 4(3): 149–160, 1996.

Jan Ćwik and Jacek Koronacki. Multivariate density estimation: A comparative study. *Neural Computing & Applications*, 6(3):173–185, 1997.

Alexei Daletskii, Yuri Kondratiev, Yuri Kozitsky, and Tanja Pasurek. Gibbs states on random configurations. *Journal of Mathematical Physics*, 55(8):083513, 2014.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodology)*, pages 1–38, 1977. doi: 10.2307/2984875. URL http://www.jstor.org/stable/2984875.

Jean-Claude Deville. Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25 (2):193–204, 1999.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418): 376–382, 1992.

Loredana Di Consiglio, Stefano Falorsi, and Ioannis Nikolaidis. *Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys*, chapter Possible Methods for Implementing the Integrated Approach of Variance Estimation. Eurostat, 2013.

Patricia Dörr and Jan Pablo Burgard. Data-driven transformations and survey-weighting for linear mixed models. Research Papers in Economics 16, Trier University, 2019. URL https://www.uni-trier.de/fileadmin/fb4/prof/VWL/EWF/Research_Papers/2019-16.pdf.

Tarn Duong and Martin L. Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506, 2005.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

John L. Eltinge and Amang Sukasih. Approximation methods for covariance matrix estimators used in analysis of diary and interview data from the us consumer expenditure survey. In *BLS document*. CiteSeer, 2001.

Erwin Erwin, Klaus Obermayer, and Klaus Schulten. Self-organizing maps: Ordering, convergence properties and energy functions. *Biological Cybernetics*, 67(1):47–55, 1992.

Lola Fabowale, Barbara Orser, and Allan Riding. Gender, structural factors, and credit terms between canadian small businesses and financial institutions. *Entrepreneurship Theory and Practice*, 19(4):41–65, 1995.

Enrico Fabrizi, Maria Rosaria Ferrante, and Carlo Trivisano. Bayesian small area estimation for skewed business survey variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(4):861–879, 2017.

Robert W. Fairlie and Alicia M. Robb. Gender differences in business performance: Evidence from the characteristics of business owners survey. *Small Business Economics*, 33(4):375, 2009.

Michael Fay and Lesley Williams. Gender bias and the availability of business loans. *Journal of Business Venturing*, 8(4):363–376, 1993.

Robert E. Fay and Roger A. Herriot. Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277, 1979.

F. Fessant and S. Midenet. Self-organising map for data imputation and correction in surveys. *Neural Computing & Applications*, 10(4):300–310, 2002.

Salvatore Filiberti, Edoardo Pizzoli, and Veronica Rondinelli. Estimating sampling errors in rea survey: A model approach for the synthetic presentation of results for main variables involved in economic analysis. *Acta Applicandae Mathematicae*, 96(1-3):203–214, 2007. doi: 10.1007/s10440-007-9109-y.

Jean-Claude Fort, Marie Cottrell, and Patrick Letremy. Stochastic on-line algorithm versus batch algorithm for quantization and self organizing maps. In *Neural Networks for Signal Processing XI: Proceedings of the 2001 IEEE Signal Processing Society Workshop (IEEE Cat. No.01TH8584)*, pages 43–52, 2001.

Jean-Claude Fort, Patrick Letremy, and Marie Cottrell. Advantages and drawbacks of the batch kohonen algorithm. In *ESANN*, volume 2, pages 223–230, 2002.

M. A. Foster, L. Tian, and L. J. Wei. Estimation for the box-cox transformation model without assuming parametric error distribution. *Journal of the American Statistical Association*, 96(455):1097–1101, 2001. ISSN 01621459. URL http://www.jstor.org/stable/2670255.

Jade Freeman and Reza Modarres. Inverse box–cox: The power-normal distribution. *Statistics & Probability Letters*, 76(8):764–772, 2006.

Bernd Fritzke. Growing cell structures —- a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 7(9):1441–1460, 1994.

Bernd Fritzke. Growing grid —- a self-organizing network with constant neighborhood range and aadaptation strength. *Neural Processing Letters*, 2(5):9–13, 1995.

Sirius Fuller and Anthony Tersine Jr. Analyzing generalized variances for the american community survey 2005 public use microdata sample. Final report, US Census Bureau, 2010. American Community Survey Research and Evaluation Program.

Wayne A. Fuller. Regression analysis for sample survey. *Sankhya*, 37(3): 117–132, 1975.

Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, pages 401–414, 1982.

German Federal Statistical Office. Zensusgesetz zensg 2011 51a, 2009. URL https://www.zensus2011.de/SharedDocs/Downloads/DE/Gesetze/Zensusgesetz_2011.pdf?__blob=publicationFile&v=12.

Julie Gershunskaya and Alan H. Dorfman. Calibration and evaluation of generalized variance functions. In *Proceedings of Joint Statistical Meetings, Survey Methods Section*, pages 2655–2669, 2013.

Thore Graepel, Matthias Burger, and Klaus Obermayer. Self-organizing maps: Generalizations and new optimization techniques. *Neurocomputing*, 21(1-3):173–190, 1998.

Geoffrey R. Grimmett. A theorem about random fields. *Bulletin of the London Mathematical Society*, 5(1):81–84, 1973.

Matthew J. Gurka, Lloyd J. Edwards, Keith E. Muller, and Lawrence L. Kupper. Extending the box-cox transformation to the linear mixed model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(2):273–288, 2006. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2005.00391.x.

Chantal Hajjar and Hani Hamdan. Self-organizing map based on hausdorff distance for interval-valued data. In *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1747–1752. IEEE, 2011.

Peter Hall. On kullback-leibler loss and density estimation. *The Annals of Statistics*, pages 1491–1519, 1987.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Sam Hawala and Partha Lahiri. Variance modeling in the us small area income and poverty estimates program for the american community survey. In *Proceedings of Joint Statistical Meetings, Survey Methods Section*, 2010.

David Haziza, Fulvia Mecatti, and John N. K. Rao. Evaluation of some approximate variance estimators under the rao-sampford unequal probability sampling design. *Metron*, 66(1):91–108, 2008.

Dan Hedlin, Hannah Falvey, Ray Chambers, and Phillip Kokic. Does the model matter for greg estimation? a business survey example. *Journal of Official Statistics*, 17(4):527–544, 2001.

Fabián Hernandez and Richard A. Johnson. The large sample behavior of transformations to normality. Technical report, University of Wisconsin, 1980a.

Fabián Hernandez and Richard A. Johnson. The large-sample behavior of transformations to normality. *Journal of the American Statistical Association*, 75(372):855–861, 1980b. doi: 10.1080/01621459.1980.10477563. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1980.10477563.

Alfred O. Hero and Jeffrey A. Fessler. Convergence in norm for alternating expectation-maximization (em) type algorithms. *Statistica Sinica*, pages 41–54, 1995.

Tom Heskes. Energy functions for self-organizing maps. In *Kohonen Maps*, pages 303–315. Elsevier, 1999.

Tom Heskes. Self-organizing maps, vector quantization, and mixture modeling. *IEEE Transactions on Neural Networks*, 12(6):1299–1305, 2001.

Paul E. Hinrichs. Consumer expenditure estimation incorporating generalized variance functions in hierarchical bayes models. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 2003.

Paul W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Jenq-Neng Hwang, Shyh-Rong Lay, and Alan Lippman. Nonparametric multivariate density estimation: A comparative study. *IEEE Transactions on Signal Processing*, 42(10):2795–2810, 1994.

Cary T. Isaki and Wayne A. Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96, 1982.

Valerie Isham. An introduction to spatial point processes and markov random fields. *International Statistical Review/ Revue Internationale de Statistique*, pages 21–43, 1981.

Wolfgang Jank. *The EM Algorithm, Its Randomized Implementation and Global Optimization: Some Challenges and Opportunities for Operations Research*, pages 367–392. Springer US, Boston, MA, 2006. ISBN 978-0-387-39934-8. doi: 10.1007/978-0-387-39934-8_21. URL https://doi.org/10.1007/978-0-387-39934-8_21.

Eugene G. Johnson and Benjamin F. King. Generalized variance functions for a complex sample survey. Research Report 87-6, Educational Testing Service, 1987.

Jari Kangas and Teuvo Kohonen. Developments and applications of the self-organizing map and related algorithms. *Mathematics and Computers in Simulation*, 41(1):3–12, 1996. ISSN 0378-4754. doi: https://doi.org/10.1016/0378-4754(96)88223-1. URL http://www.sciencedirect.com/science/article/pii/0378475496882231.

Jae Kwang Kim and Mingue Park. Calibration estimation in survey sampling. *International Statistical Review*, 78(1):21–39, 2010. doi: 10.1111/j.1751-5823.2010.00099.x. URL https://www.jstor.org/stable/27919793.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, 1965.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

Teuvo Kohonen. Essentials of the self-organizing map. *Neural Networks*, 37:52–65, 2013.

Yuri Kondratiev, Yuri Kozitsky, and Tanja Pasurek. Gibbs random fields with unbounded spins on unbounded degree graphs. *Journal of Applied Probability*, 47(3):856–875, 2010.

Timo Kostiainen and Jouko Lampinen. On the generative probability density model in the self-organizing map. *Neurocomputing*, 48(1):217–228, 2002.

Phillip S. Kott. A design-sensitive approach to fitting regression models with complex survey data. *Statistics Surveys*, 12:1–17, 2018. doi: 10.1214/17-SS118. URL https://doi.org/10.1214/17-SS118.

D. Krewski and John N. K. Rao. Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, pages 1010–1019, 1981.

Jan Kubacki and Alina Jędrzejczak. The comparison of generalized variance function with other methods of precision estimation for polish household budget survey. In *Survey Sampling in Economic and Social Research*, pages 58–69. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, 2012.

Věra Kůrková. Kolmogorov's theorem and multilayer neural networks. *Neural networks*, 5(3):501–506, 1992.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

Hyunshik Lee and James Croal. A simulation study of various estimators which use auxiliary data in an establishment survey. In *Proceedings of the Survey Research Methods Section, American Statistical Association, Survey Research Methods Section*, pages 336–341, 1989.

Erich L. Lehmann. *Theory of Point Estimation*. Mathematical Statistics. John Wiley & Sons, 1983.

Luo Lu, Hui Jiang, and Wing H. Wong. Multivariate density estimation by bayesian sequential partitioning. *Journal of the American Statistical Association*, 108(504):1402–1410, 2013.

Thomas Lumley. *Complex Surveys: A Guide to Analysis using R*. John Wiley & Sons, 2010. ISBN 9780470284308.

Thomas Lumley. Survey: Analysis of complex survey samples, 2019. R package version 3.35-1.

H. Mallinson and A. Gammerman. Imputation using support vector machines. Technical report, University of London, 2003. URL https://www.cs.york.ac.uk/euredit/_temp/_V2_%20Methods%20&%20Results%20D6.1/_Chapter%206%20SVM.pdf.

J. Maples, W. Bell, and Elizabeth T. Huang. Small area variance modeling with application to county poverty estimates from the american community survey. In *Proceedings of the American Statistical Association Section on Survey Research Methods*, pages 5056–5067, 2009.

Stephen Marsland. *Machine Learning – An Algorithmic Perspective*. Machine Learning & Pattern Recognition Series. Taylor & Francis, 2015.

Alina Matei and Yves Tillé. Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21(4):543–570, 2005.

Charles E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92 (437):162–170, 1997. doi: 10.1080/01621459.1997.10473613. URL https://doi.org/10.1080/01621459.1997.10473613.

Ronald C. Neath. On convergence properties of the monte carlo em algorithm. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, pages 43–62. Institute of Mathematical Statistics, 2013.

Hien D. Nguyen and Geoffrey McLachlan. On approximations via convolution-defined mixture models. *Communications in Statistics – Theory and Methods*, 48(16):3945–3955, 2019.

Patricia D. Olson, Virginia S. Zuiker, Sharon M. Danes, Kathryn Stafford, Ramona K. Z. Heck, and Karen A. Duncan. The impact of the family and the business on family business sustainability. *Journal of business venturing*, 18(5):639–666, 2003.

Mark C. Otto and William R. Bell. Sampling error modelling of poverty and income statistics for states. In *Proceedings of the American Statistical Association Section on Government Statistics*, pages 160–165, 1995.

Art B. Owen. Monte carlo theory, methods and examples, 2013. URL http://statweb.stanford.edu/~owen/mc/.

Luciano D. S. Pacifico and Francisco de A. T. de Carvalho. A batch self-organizing maps algorithm based on adaptive distances. In *The 2011 International Joint Conference on Neural Networks*, pages 2297–2304. IEEE, 2011.

Tetyana Pasurek. *Theory of Gibbs Measures with Unbounded Spins: Probabilistic and Analytical Aspects*. Habilitation, Universität Bielefeld, 2007.

Sourav Paul and Mousumi Gupta. Image segmentation by self organizing map with mahalanobis distance. *International Journal of Emerging Technology and Advanced Engineering*, 3(2):288–291, 2013.

A. Pavone and Aldo Russo. Generalized variance function: Theory and empirics. *parameters*, 2:1, 1999.

Danny Pfeffermann. The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, pages 317–337, 1993.

Danny Pfeffermann and Michail Sverchkov. Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102(480):1427–1439, 2007.

Danny Pfeffermann, Chris J. Skinner, David J. Holmes, Harvey Goldstein, and Jon Rasbash. Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: Series B (Methodology)*, 60(1):23–40, 1998.

José C. Pinheiro and Douglas M. Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35, 1995.

M. J. D. Powell. Updating conjugate directions by the bfgs formula. *Mathematical Programming*, 38(1):29, 1987. doi: 10.1007/BF02591850. URL https://doi.org/10.1007/BF02591850.

John Preston. Rescaled bootstrap for stratified multistage sampling. *Survey Methodology*, 35(2):227–234, 2009.

Carey E. Priebe. Adaptive mixtures. *Journal of the American Statistical Association*, 89(427):796–806, 1994.

Carey E. Priebe and David J. Marchette. Adaptive mixture density estimation. *Pattern Recognition*, 26(5):771–785, 1993.

Maurice H. Quenouille. Notes on bias in estimation. *Biometrika*, 43(3/4): 353–360, 1956.

Sophia Rabe-Hesketh and Anders Skrondal. Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):805–827, 2006.

John N. K. Rao and G. Hussain Choudhry. *Business Survey Methods*, chapter Small Area Estimation: Overview and Empirical Study, pages 527–542. Probability and Mathematical Statistics. Wiley, 1995.

Jerome P. Reiter, Trivellore E. Raghunathan, and Satkartar K. Kinney. The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32(2):143, 2006.

Allan L. Riding and Catherine S. Swift. Women business owners and terms of credit: Some empirical findings of the canadian experience. *Journal of Business Venturing*, 5(5):327–340, 1990.

Helge Ritter. Self-organizing maps on non-euclidean spaces. In *Kohonen Maps*, pages 97–109. Elsevier, 1999.

Helge Ritter and Klaus Schulten. On the stationary state of kohonen's self-organizing sensory mapping. *Biological cybernetics*, 54(2):99–106, 1986.

P. M. Robinson and Carl-Erik Särndal. Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 240–248, 1983.

Natalia Rojas-Perilla, Sören Pannier, Timo Schmid, and Nikos Tzavidis. Data-driven transformations in small area estimation. Discussion Paper 2017/30, Discussion Paper, School of Business & Economics: Economics, 2017. URL http://hdl.handle.net/10419/172326.

Peter Rosa, Sara Carter, and Daphne Hamilton. Gender as a determinant of small business performance: Insights from a british study. *Small Business Economics*, 8(6):463–478, 1996.

Richard M. Royall. Likelihood functions in finite population sampling theory. *Biometrika*, 63(3):605–614, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335741.

Herman Rubin. Uniform convergence of random functions with applications to statistics. *The Annals of Mathematical Statistics*, 27(1):200–203, 1956.

Susana Rubin-Bleuer and Ioana Schiopu Kratina. On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33(6):2789–2810, 2005.

Ali A. Sadeghi. Convergence in distribution of the multi-dimensional kohonen algorithm. *Journal of Applied Probability*, 38(1):136–151, 2001.

Sameena Salvucci, Stanley Weng, and Kaufman Steven. Design effects and generalized variance functions for the 1990-91 schools and staffing survey (sass). Technical Report 95-342-1, US Department of Education – National Center for Education Statistics, 1995.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer Science & Business Media, 1992. ISBN 0-387-40620-4.

Luca Scrucca, Michael Fop, Thomas Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2016. URL https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf.

Jun Shao and C. F. Jeff Wu. A general theory for jackknife variance estimation. *The Annals of Statistics*, 17(3):1176–1197, 1989.

Robert P. Sherman, Yu-Yun K. Ho, and Siddhartha R. Dalal. Conditions for convergence of monte carlo em sequences with an application to product diffusion modeling. *The Econometrics Journal*, 2(2):248–267, 1999.

A. G. Stephenson. evd: Extreme value distributions. *R News*, 2(2):0, June 2002. ISSN 1609-3631. URL https://CRAN.R-project.org/doc/Rnews/.

Yves Tillé and Alina Matei. Package 'sampling', 2016. URL https://cran.r-project.org/web/packages/sampling/index.html. Survey sampling (Version 2.8).

V. V. Tolat. An analysis of kohonen's self-organizing maps using a system of energy functions. *Biological Cybernetics*, 64(2):155–164, 1990.

Richard Valliant. Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, 82(398):499–508, 1987.

Richard Valliant. Smoothing variance estimates for price indexes over time. *Journal of Official Statistics*, 8(4):433, 1992.

Richard Valliant, A. H. Dorfman, and Richard M. Royall. *Finite Population Sampling and Inference - A Prediction Approach*. Probability and Statistics. Wiley, 2000. ISBN 0471293415.

Jakob J. Verbeek, Nikos Vlassis, and Ben J. A. Kröse. Self-organizing mixture models. *Neurocomputing*, 63:99–123, 2005.

Juha Vesanto, Mika Sulkava, and Jaakko Hollmén. On the decomposition of the self-organizing map distortion measure. In *Proceedings of the Workshop on Self-organizing Maps (WSOM'03)*, pages 11–16, 2003.

Thomas Villmann and Jens Christian Claussen. Magnification control in self-organizing maps and neural gas. *Neural Computation*, 18(2):446–469, 2006.

Matt P. Wand and M. Chris Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116, 1994.

R. W. M. Wedderburn. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63(1):27–32, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335080.

Greg C. G. Wei and Martin A. Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

Fiona Wilson, Sara Carter, Stephen Tagg, Eleanor Shaw, and Wing Lam. Bank loan officers' perceptions of business owners: The role of gender. *British Journal of Management*, 18(2):154–171, 2007.

Kirk Wolter. *Introduction to variance estimation*. Springer, 1985.

Ralph S. Woodruff. A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66(334):411–414, 1971.

World Bank. Enterprise survey and indicator surveys – sampling methodology. Technical report, World Bank, 2009. URL https://www.enterprisesurveys.org/content/dam/enterprisesurveys/documents/methodology/Sampling_Note.pdf.

C. F. Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, pages 95–103, 1983. doi: 10.1214/aos/1176346060. URL https://projecteuclid.org/euclid.aos/1176346060.

Zhenlin Yang. A modified family of power transformations. *Economics Letters*, 92(1):14–19, 2006.

Hujun Yin and Nigel M. Allinson. On the distribution and convergence of feature space in self-organizing maps. *Neural Computation*, 7(6): 1178–1187, 1995.

Yong You and John N. K. Rao. A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30(3):431–439, 2002.

Xibin Zhang, Maxwell L. King, and Rob J. Hyndman. A bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics & Data Analysis*, 50(11):3009–3031, 2006.

Vadim V. Zipunnikov and James G. Booth. Monte carlo em for generalized linear mixed models using randomized spherical radial integration. Working paper, Cornell University, 2006.