Adaptive Likelihood Weights and Mixtures of Empirical Distributions

by

Jean-François Plante

B.Sc., Université Laval, 2000
M.Sc., Université Laval, 2002

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

Doctor of Philosophy

in

The Faculty of Graduate Studies

(Statistics)

The University of British Columbia

July, 2007

Abstract

Suppose that you must make inference about a population, but that data from $m-1$ similar populations are available. The weighted likelihood uses exponential weights to include all the available information into the inference. The contribution of each datum is discounted based on its dissimilarity with the target distribution.

One could hope to elicitate the likelihood weights from scientific information, but using data-based weights is more pragmatic. To this day, no entirely satisfactory method has been found for determining likelihood weights from the data.

We propose a way to determine the likelihood weights based on data. The suggested "MAMSE" weights are nonparametric and can be used as likelihood weights, or as mixing probabilities to define a mixture of empirical distributions. In both cases, using the MAMSE weights allows strength to be borrowed from the $m-1$ similar populations whose distribution may differ from the target.

The MAMSE weights are defined for different types of data: univariate, censored and multivariate. In addition to their role for the likelihood, the MAMSE weights are used to define a weighted Kaplan-Meier estimate of the survival distribution and weighted coefficients of correlation based on ranks. The maximum weighted pseudo-likelihood, a new method to fit a family of copulas, is also proposed. All these examples of inference using the MAMSE weights are shown to be asymptotically unbiased. Furthermore, simulations show that inference based on MAMSE-weighted methods can perform better than their unweighted counterparts. Hence, the adaptive weights we propose successfully trade bias for precision.

Table of Contents

Abstract ii

Table of Contents iii

List of Tables vi

List of Figures xi

Acknowledgements xiii

1 Introduction 1

1.1 Review 1

1.2 Overview 3

2 Likelihood, Entropy and Mixture Distributions 5

2.1 Likelihood and Entropy 5

2.2 The Weighted Likelihood 6

2.3 The MAMSE Weights 7

2.4 An Algorithm to Compute the MAMSE Weights 10

3 MAMSE Weights for Univariate Data 17

3.1 Notation and Review 17

3.2 Definition of the MAMSE Weights 18

3.3 Computing the MAMSE Weights 19

3.4 Structural Properties of the MAMSE Weights 20
3.5 Strong Uniform Convergence of the Weighted Empirical CDF 22
3.6 Weighted Strong Law of Large Numbers 30
3.7 Strong Consistency of the Maximum Weighted Likelihood Estimate 36
3.8 Asymptotic Behavior of the MAMSE Weights 46
3.9 Issues for Asymptotic Normality 50
3.10 Simulations 51
3.10.1 Two Normal Distributions 52
3.10.2 Complementary Populations 54
3.10.3 Negative Weights 56
3.10.4 Earthquake Data 60

4 MAMSE Weights for Right-Censored Data 65
4.1 Notation and Review of the Kaplan-Meier Estimate 66
4.2 Definition of the MAMSE Weights for Right-Censored Data 69
4.3 Computing the MAMSE Weights for Right-Censored Data 72
4.4 Uniform Convergence of the MAMSE-Weighted Kaplan-Meier Estimate . . 73
4.5 Simulations 86
4.5.1 Survival in the USA 87
4.5.2 Survival in Canada 94

5 MAMSE Weights for Copulas 100
5.1 Review of Copulas 100
5.2 Notation and Empirical Estimation 106
5.3 Definition of the MAMSE Weights for Copulas 111
5.4 Computing the MAMSE Weights for Copulas 113
5.5 Uniform Convergence of the MAMSE-Weighted Empirical Copula 114
5.6 Weighted Coefficients of Correlation Based on Ranks 122

5.7 Weighted Strong Law of Large Numbers 126
5.8 Strong Consistency of the Weighted Pseudo-Likelihood 134
5.9 Simulations 143
5.9.1 Different Correlations for Bivariate Copulas 144
5.9.2 Different Bivariate Distributions with Common Correlation 147
5.9.3 Multivariate Normal 148

6 Summary and Future Research 159
6.1 Summary of the Work 159
6.2 Future Research 161

7 Conclusion 165

Bibliography 167

List of Tables

3.1 Average MAMSE weights for Population 1 when equal samples of size n are drawn from Normal distributions with unit variance and means 0 and Δ respectively. The results are averages over 10000 replicates. 52

3.2 Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE). Samples of equal size n are simulated from Normal distributions with unit variance and means 0 and Δ respectively. The results are averaged over 10000 replicates. 53

3.3 Average MAMSE weights for Population 1 when samples of size n and $10n$ are drawn from Normal distributions with unit variance and means 0 and Δ respectively. The results are averages over 40000 replicates. 54

3.4 Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE). Samples of sizes n and $10n$ are simulated from Normal distributions with unit variance and means 0 and Δ respectively. The results are averaged over 40000 replicates. 55

3.5 Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE) and average MAMSE weights allocated to samples of sizes n drawn from $\mathcal{N}(0,1)$, $|\mathcal{N}(0,1)|$ and $-|\mathcal{N}(0,1)|$ respectively. The results are averages over 10000 repetitions. 56

3.6 Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE) when the MAMSE weights are calculated without the constraints $\lambda_i \geq 0$. Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathcal{N}(0,1)$ and a $\mathcal{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions. . . . 59

3.7 Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE) when the usual MAMSE weights (i.e. constrained to positive values) are used. Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathcal{N}(0,1)$ and a $\mathcal{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions. . . . 59

3.8 Relative efficiency of the MWLE with and without the constraints $\lambda_i \geq 0$ as measured by 100 MSE(constrained MWLE)/MSE(unconstrained MWLE). Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathcal{N}(0,1)$ and a $\mathcal{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions. 60

3.9 Number of earthquakes in three Western Canadian areas between the 12^{th} of February 2001 and the 12^{th} of February 2006. The magnitude of these earthquakes is modeled by a Gamma distribution; the maximum likelihood estimates appear above and are used as the "true" parameters for this simulation. 62

3.10 Efficiency in estimating some probabilities about the magnitude of the next earthquake in the Lower Mainland – Vancouver Island area followed by the average of the actual estimates and their true values. Efficiency is measured by 100 MSE(plug-in MLE)/MSE(plug-in MWLE). The first four columns of probabilities should be multiplied by the corresponding multiplier. 63

4.1 Relative performance of the WKME as measured by $100A_1/A_{\boldsymbol{\lambda}}$. Both areas are calculated on the interval $[0,55]$ and various upper bounds U are used to determine the weights. Samples of equal size n are taken from each of four subpopulations, then used to estimate the survival of a white male living in the USA. Each scenario is repeated 20000 times. 89

4.2 Relative performance of the WKME as measured by $100A_1/A_{\boldsymbol{\lambda}}$. Areas are calculated on the interval $[0, U-5]$, where U is the upper bound used to determine the weights. Samples of equal size n are taken from each of four subpopulations, then used to estimate the survival of a white male living in the USA. Each scenario is repeated 20000 times. 90

4.3 Relative performance of the weighted Kaplan-Meier estimate compared to the usual Kaplan-Meier estimate for estimating $F_1(55) = 0.11976$ as measured by 100 MSE$\{\hat{F}_1(55)\}$/MSE$\{\hat{F}_{\boldsymbol{\lambda}}(55)\}$. Different choices of U and n are considered. Each scenario is repeated 20000 times. 93

4.4 Relative performance of the weighted Kaplan-Meier estimate compared to the usual Kaplan-Meier estimate for estimating $F_1^{-1}(0.10) = 52.081$ as measured by MSE$(\hat{q}_1)$/MSE$(\hat{q}_{\boldsymbol{\lambda}})$. Different choices of U and n are considered. Each scenario is repeated 20000 times. 93

4.5 Average MAMSE weights for different rates of censoring p and different sample sizes n. The right-hand side of the table presents the relative performance of the WKME as measured by $100A_1/A_{\boldsymbol{\lambda}}$. Figures are averaged over 20000 repetitions and the values $U = 80$ and $T = 75$ are used. 94

4.6 Relative performance of the weighted Kaplan-Meier estimate compared to the Kaplan-Meier estimate as measured by $100A_1/A_{\boldsymbol{\lambda}}$. Average MAMSE weights are also shown, but they are multiplied by a factor of 1000 for easier reading. In the simulation with males and females, the average weights in italics refer to the male populations. Note that $U = 90$, $T = 85$ and that all figures are based on 10000 repetitions. 97

5.1 Population value of different coefficients of correlation. 102

5.2 Empirical estimates of different coefficients of correlation. 109

5.3 Values of ρ under two different scenarios that are simulated for different families of copulas. 144

5.4 Performance of a MAMSE-weighted coefficient of correlation based on ranks as measured by 100 MSE($\hat{\rho}$)/MSE($\hat{\rho}_{\boldsymbol{\lambda}}$) for different scenarios and sample sizes $n \in \{20, 50, 100, 250\}$. Each figure is averaged over 10000 repetitions. 146

5.5 Performance of the maximum weighted pseudo-likelihood estimate as measured by 100 MSE($\hat{\theta}$)/MSE($\hat{\theta}_{\boldsymbol{\lambda}}$) for different scenarios and sample sizes $n \in \{20, 50, 100, 250\}$. Each figure is averaged over 10000 repetitions. 146

5.6 Distributions from which bivariate random variables are drawn. The choice of parameters in each population yields a Spearman correlation of $\rho = 0.20$. 147

5.7 Average MAMSE weights as well as the relative efficiency of the weighted correlation and of the MWPLE when compared to their non-weighted counterparts. Samples of size $n \in \{20, 50, 100, 250\}$ are drawn from five different bivariate distributions that have a common correlation of $\rho = 0.2$. Figures are averaged over 10000 repetitions. 148

5.8 Parameters of the simulation for 3-variate Normal variables. Population 1 is from the target distribution, but the other populations are affected by measurement errors. 153

5.9 Average weight allocated to each of four 3-variate Normal distributions. Population 1 is observed from the target distribution, but the other populations contain measurement errors. The values are averaged over 5000 repetitions. 153

5.10 Relative performance of the MWPLE when compared to the MPLE as measured by 100 MSE($\hat{\boldsymbol{\Gamma}}_1$)/MSE($\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\lambda}}$) or an equivalent ratio for the individual entries of $\boldsymbol{\Gamma}_1$. Population 1 is observed from a 3-variate Normal distribution and the other populations contain measurement errors. The values are averaged over 5000 repetitions. 154

5.11 Parameters of the simulations for 4-variate Normal variables. Population 1 comes from the target distribution, but the other populations are affected by measurement errors. 156

5.12 Average weight allocated to each of four 4-variate Normal distributions. Population 1 is observed from the target distribution and the other populations contain measurement errors. The values are averaged over 5000 repetitions. 157

5.13 Relative performance of the MWPLE when compared to the MPLE for 4-variate Normal distributions. Population 1 is observed from the target distribution and the other populations contain measurement errors. The values are averaged over 5000 repetitions. 157

List of Figures

3.1 Average values of 100× the MAMSE weights without the constraints $\lambda_i \geq 0$. Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathcal{N}(0,1)$ and a $\mathcal{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions. 58

3.2 Average values of 100× the usual MAMSE weights (with constraints $\lambda_i \geq 0$). Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathcal{N}(0,1)$ and a $\mathcal{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions. 61

3.3 Histograms of the magnitude of earthquakes measured between the 12^{th} of February 2001 and the 12^{th} of February 2006 for three different Western Canadian areas. The curves correspond to the fitted Gamma density. 62

4.1 Graphics representing the Riemann sums used in the proof of Case 2 (left panel) and Case 3 (right panel). 82

4.2 Survival functions for subgroups of the American population as taken from the life tables of the National Center for Health Statistics (1997). 88

4.3 Average value of the MAMSE weights for different upper bounds U and sample sizes n. The cells' area are proportional to the average weight allocated to each population. The numbers correspond to $100\bar{\lambda}_1$ and are averaged over 20000 repetitions. 91

4.4 Typical examples of the weighted Kaplan-Meier estimate (dashed line) and of the usual Kaplan-Meier estimate (plain black line) for different sample sizes. Note that $U = 80$ and $T = 75$. The true distribution is depicted by a smooth gray line. 92

4.5 Survival functions of Canadians as taken from the life tables of Statistics Canada (2006). Survival functions are available for males and females of each province. The curves for New Brunswick are repeated as gray dotted lines on each panel to facilitate comparisons since they are the populations of interest. 96

4.6 Typical examples of estimates under different scenarios. Note the increased number of steps of the weighted Kaplan-Meier estimate (dashed line) which makes it smoother than the usual Kaplan-Meier estimate (plain black line). The true distribution appears as a smooth gray curve. 99

5.1 Average value of the MAMSE weights for scenarios where samples of equal sizes n are drawn from 5 different populations. The cell areas are proportional to the average weight allocated to each population. The numbers correspond to $100\bar{\lambda}_1$ and are averaged over 10000 repetitions. 145

Acknowledgements

I would like to thank my supervisor Jim Zidek for the numerous inspiring discussions we had about the foundations of statistics. With an amazing intellectual tact, Jim gave me enough direction to keep me thinking hard, but enough liberty to let me design this thesis my own way. His guidance will definitely have a deep impact in the way I approach statistics in my future career. I would also like to acknowledge Jim's financial support.

I would like to thank the members of my supervisory committee, John Petkau and Matías Salibián-Barrera, for their insightful comments and suggestions. I would like to extend an additional thank to John for sharing his wisdom in many occasions when I was seeking an advice.

The Department of Statistics offers a friendly environment for development and growth. I did appreciate the opportunity to be exposed to different views of statistics. Department's faculty and staff members as well as fellow students have always been very helpful. I am thankful to them.

I made many very good friends during my stay in Vancouver. They brightened my days and will truly be missed. Thank you for the friendship, the support and for the fun!

I want to thank Christian Genest, my Master's supervisor, for his outstanding guidance while I was at Laval University. I learned so much under his supervision that his teachings are still making a significant difference in my work and in my understanding of statistics as a science and as a discipline.

Thanks to my family for their constant support and prayers. I had the chance to grow up in a healthy and supportive familial environment that forged my personality. I am thankful

for that.

Thank you Pascale for sharing my life through the second half of this endeavour.

For partial support of this work through graduate scholarships and research grants, thanks are due to the Natural Sciences and Engineering Research Council of Canada and to the Fonds québécois de la recherche sur la nature et les technologies.

Chapter 1

Introduction

Suppose that you must make inference about a population from which you obtained a sample and that additional data are available from populations whose distributions may be similar to the target, but not necessarily identical. The weighted likelihood allows use of the relevant information contained in that data and its incorporation into the inference. In this thesis, we propose a data-based criterion to determine likelihood weights and show that they can successfully trade bias for precision.

1.1 Review

The expression *weighted likelihood* refers to a few different statistical methods. In survey sampling theory, a method called *weighted likelihood* is used to adjust for a response-dependent sampling intensity; see Krieger & Pfeffermann (1992). The increased robustness of estimates obtained through *weighted likelihood equations* have been studied by several groups of statisticians including Markatou et al. (1997) and Ahmed et al. (2005). However, both these methods are only remotely related to the *relevance weighted likelihood* as defined by Hu (1994). In that context, the weighted likelihood is an extension to the likelihood that uses data from multiple populations for making inference.

Statistical models typically use information about data that come from the target distribution. When other possibly similar data are available, the relevant information they contain is ignored unless their similarity with the population of interest is precisely incorporated in the model.

The original work of Hu (1994), enhanced in Hu & Zidek (2001) and Hu & Zidek (2002), proposes use of the weighted likelihood to capture the relevant information from data whose distribution may differ from that of interest, hence trading bias for precision. A specific paradigm is derived from the work of Hu and Zidek by Wang (2001) and further developed in Wang, van Eeden & Zidek (2004) and in Wang & Zidek (2005). We adopt that paradigm throughout this thesis.

Suppose the data come from m distinct populations that have different yet similar distributions. More formally, for each fixed $i = 1, \ldots, m$,

$$X_{i1}, \ldots, X_{in_i} \overset{iid}{\sim} F_i$$

where F_i denotes the cumulative distribution function (CDF) of each population. We denote by f_i the corresponding density or mass function for population $i = 1, \ldots, m$. The family of density (or mass) functions $f(x|\theta)$ indexed by $\theta \in \Theta$ is used to model the data. Population 1 is of inferential interest, but the weighted likelihood

$$L_{\boldsymbol{\lambda}}(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(X_{ij}|\theta)^{\lambda_i/n_i}$$

lets other populations contribute to the inference so that the relevant information they contain is not lost. The vector of exponents $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^{\mathsf{T}}$ downweights data according to the degree to which they are thought to be dissimilar from Population 1. The expression for the weighted log-likelihood may be more intuitive:

$$\ell_{\boldsymbol{\lambda}}(\theta) = \sum_{i=1}^{m} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} \log f(X_{ij}|\theta).$$

The maximum weighted likelihood estimate (MWLE) is a value of θ that maximizes $L_{\boldsymbol{\lambda}}(\theta)$.

Ideally, the weights λ_i would be determined from scientific knowledge. However, using data-based weights seems more pragmatic. The ad-hoc methods proposed by Hu &

Zidek (2002) as well as the cross-validation weights of Wang (2001) and Wang & Zidek (2005) are the only adaptive weights proposed in the literature for the paradigm considered. However, none of these solutions is fully satisfactory.

In particular, the cross-validation weights (CVW) suffer from instability problems. The simulation results in Wang (2001) and Wang & Zidek (2005) can be reproduced only at the cost of fine-tuning the algorithms numerically. In the ideal situation where some of the populations are identical to Population 1, the weights may not even be defined.

Let $\hat{F}_i$ denote the empirical distribution function (EDF) based on the sample from Population i. Hu & Zidek (1993) call $\hat{F}_{\boldsymbol{\lambda}} = \sum_{i=1}^{m} \lambda_i \hat{F}_i$ the *relevance weighted empirical distribution* (REWED) when $\lambda_i \geq 0$ add to 1. Hu & Zidek (2002) show that the REWED is the nonparametric maximum weighted likelihood estimate of the target distribution function. In Chapter 2, we look at the link between the REWED and the weighted likelihood from a different angle.

Note that Ahmed (2000) uses a paradigm similar to the one retained in this thesis as he considers samples of different sizes from m populations to develop a Stein-type estimator of the quantiles of the distribution. The properties of his method and its links to the REWED and the weighted likelihood are however not the object of this work.

In this thesis, we propose a data-based criterion to determine likelihood weights which yields consistent estimates. Inferential methods based on our proposed weights perform at least as well as the CVW in comparable simulations, but without the instability problems. We also propose other weighted methods based on the newly proposed weights that are shown to perform well and to be asymptotically unbiased.

1.2 Overview

In Chapter 2, we present our heuristic interpretation of the weighted likelihood that links it to mixture distributions. This leads to the general definition of the MAMSE (Minimum Averaged Mean Squared Error) weights and we provide an algorithm to calculate them.

The ensuing three chapters are implementations of the idea of the MAMSE weights in three different contexts.

In Chapter 3, univariate data are treated using the empirical distribution function. The properties of the resulting nonparametric MAMSE weights are studied. Asymptotically, the MAMSE-weighted empirical distribution function converges to the target distribution and the maximum weighted likelihood estimate is strongly consistent. Finally, simulations show that improved performance is indeed possible on samples of realistic sizes.

Chapter 4 introduces MAMSE weights for right-censored data by using the Kaplan-Meier estimate to infer the CDF of each population. The weights are used to define a weighted Kaplan-Meier estimate that converges uniformly to the target distribution. Simulations show that improved performance is once again possible.

Finally, MAMSE weights for multivariate data are defined using the empirical copula. The MAMSE-weighted mixture of empirical copulas converges uniformly to the copula underlying the target distribution. We define MAMSE-weighted coefficients of correlation and show that they are consistent estimates of the correlation in the target population. To infer the copula underlying the target population, the maximum weighted pseudo-likelihood estimate (MWPLE) is proposed. We show that the MWPLE is consistent when calculated with MAMSE weights. Simulations confirm once more that MAMSE-weighted methods may perform better than their unweighted counterparts, even in subasymptotic ranges.

We conclude this thesis with a summary and some suggestions for future research.

Chapter 2

Likelihood, Entropy and Mixture Distributions

This chapter links the likelihood to the weighted likelihood by showing that both expressions can be derived as particular cases of the *entropy maximization principle* proposed by Akaike (1977).

After drawing the connection between relative entropy and the likelihood, we heuristically extend that link to the relationship between mixture distributions and the weighted likelihood. This leads us to formulate a general criterion to determine likelihood weights, called the "MAMSE" weights. We propose an algorithm to compute the MAMSE weights and prove its convergence.

2.1 Likelihood and Entropy

Consider first the one-sample situation where n independent data points, $Y_1, \ldots, Y_n$, come from a distribution whose unknown and unknowable density is $g(y)$. In his pioneering work, Akaike (1977) argues that the goal of inference should be the estimation of $g(y)$. When a parametric model $f(y|\theta)$ is to be used, Akaike proposes maximizing the relative entropy:

$$B(g, f) = -\int \frac{g(y)}{f(y|\theta)} \log\left\{\frac{g(y)}{f(y|\theta)}\right\} f(y|\theta)\, \mathrm{d}y.$$

The relative entropy is in fact minus the Kullback-Leibler divergence between f and g. In that sense, it is a measure of the proximity of the distributions f and g. The expression for $B(g, f)$ can be further simplified:

$$B(g,f) = -\int \log\left\{\frac{g(y)}{f(y|\theta)}\right\} g(y)\,\mathrm{d}y = \int \log\{f(y|\theta)\}g(y)\,\mathrm{d}y - \int \log\{g(y)\}g(y)\,\mathrm{d}y.$$

In particular, when the objective is to maximize $B(g, f)$ as a function of θ, the last term of the rightmost expression can be ignored since it does not depend on θ.

Calculating the entropy would require the knowledge of the unknown and unknowable true distribution g. We thus have to estimate it. Let

$$\hat{G}(y) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(Y_i \leq y)$$

be the empirical distribution function (EDF) of the dataset $Y_1, \ldots, Y_n$. The indicator variable $\mathbb{1}(\bullet)$ is equal to one if all the elements of its argument are true and equal to 0 otherwise. Using $\mathrm{d}\hat{G}(y)$ as an approximation to $\mathrm{d}G(y) = g(y)\,\mathrm{d}y$ yields

$$\int \log\{f(y|\theta)\}\,\mathrm{d}\hat{G}(y) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i|\theta),$$

the log-likelihood! Therefore, calculating the likelihood is equivalent to calculating the entropy where the true distribution is estimated by the empirical distribution of the data. Hence, the maximum likelihood estimate can be seen as a special case of Akaike's entropy maximization principle.

2.2 The Weighted Likelihood

Consider now the m-population paradigm of Wang (2001) introduced in Chapter 1. With appropriate weights, the mixture $F_{\boldsymbol{\lambda}} = \sum_{i=1}^{m} \lambda_i F_i$ can be arbitrarily close to F_1. Let $\hat{F}_i$ denote the EDF based on the sample from Population i. The weighted EDF, written

$\hat{F}_{\boldsymbol{\lambda}} = \sum_{i=1}^{m} \lambda_i \hat{F}_i$ with $\lambda_i \geq 0$ and called *relevance weighted empirical distribution* by Hu & Zidek (1993), may use more data than $\hat{F}_1$, and thus be less variable. Hu & Zidek (1993) note the implicit bias involved in defacto replacing F_1 by $F_{\boldsymbol{\lambda}}$, but do not investigate as we do here the possibility of trading bias for precision.

In the context of maximum entropy, consider using the weighted EDF as an estimate of F_1. Then,

$$\int \log f(x|\theta)\, \mathrm{d}\hat{F}_{\boldsymbol{\lambda}}(x) = \sum_{i=1}^{m} \lambda_i \int \log f(x|\theta)\, \mathrm{d}\hat{F}_i(x) = \sum_{i=1}^{m} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} \log f(X_{ij}|\theta),$$

the weighted log-likelihood. The maximum weighted likelihood can thus be derived from Akaike's entropy maximization principle.

2.3 The MAMSE Weights

Let $\hat{F}_i(x)$ represent an estimate of the CDF of Population i. Based on the heuristics presented in the previous section, the weighted likelihood is akin to using a mixture of the CDFs of the m populations from which data are available. Let

$$\hat{F}_{\boldsymbol{\lambda}}(x) = \sum_{i=1}^{m} \lambda_i \hat{F}_i(x)$$

with $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^{\mathsf{T}}$, $\lambda_i \geq 0$ and $\boldsymbol{\lambda}^{\mathsf{T}}\mathbf{1} = 1$ be such a mixture. A wise choice of likelihood weights should make:

- $\hat{F}_{\boldsymbol{\lambda}}(x)$ close to $F_1(x)$, the target distribution,
- $\hat{F}_{\boldsymbol{\lambda}}(x)$ less variable than $\hat{F}_1(x)$.

We combine the two requirements above in an objective function. Let $\mu(x)$ be a probability measure that has a domain comparable to the distribution of interest and define

$$P(\boldsymbol{\lambda}) = \int \left[\left\{ \hat{F}_1(x) - \hat{F}_{\boldsymbol{\lambda}}(x) \right\}^2 + \sum_{i=1}^{m} \lambda_i^2 \, \widehat{\text{var}} \left\{ \hat{F}_i(x) \right\} \right] \mathrm{d}\mu(x). \tag{2.1}$$

The term $\widehat{\text{var}} \left\{ \hat{F}_i(x) \right\}$ is a form of penalty to foster decreasing the variance by using as many of the m populations as necessary. In the coming chapters, we will use different estimates $\hat{F}_i$ depending on the type of data at hand. The exact expression for $\widehat{\text{var}} \left\{ \hat{F}_i(x) \right\}$ will depend on that choice; see Sections 3.1, 4.2 and 5.3 for specific examples. The probability measure $\mu(x)$ may also depend on the data and will be specified later.

We call MAMSE (minimum averaged mean squared error) weights any vector of values $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^\mathsf{T}$ that solves the program

$$\begin{aligned} &\text{minimize} && P(\boldsymbol{\lambda}) \\ &\text{subject to} && \{\lambda_i \geq 0, i = 1, \ldots, m\} \text{ and } \sum_{i=1}^{m} \lambda_i = 1. \end{aligned} \tag{2.2}$$

The name MAMSE is motivated by the conceptual resemblance of the integrand with the mean squared error (Bias^2 + Variance).

Note that the objective function $P(\boldsymbol{\lambda})$ is quadratic in $\boldsymbol{\lambda}$ and must be optimized under linear constraints. Substituting $\lambda_1 = 1 - \sum_{i=2}^{m} \lambda_i$ allows the constraint $\sum_{i=1}^{m} \lambda_i = 1$ to embedded into the objective function. Let us write

$$\tilde{\boldsymbol{\lambda}} = \begin{bmatrix} \lambda_2 \\ \vdots \\ \lambda_m \end{bmatrix}, \quad \mathcal{F}(x) = \begin{bmatrix} \hat{F}_1(x) - \hat{F}_2(x) \\ \vdots \\ \hat{F}_1(x) - \hat{F}_m(x) \end{bmatrix} \text{ and}$$

$$\mathbf{V}(x) = \begin{bmatrix} \widehat{\text{var}}\{\hat{F}_2(x)\} & & 0 \\ & \ddots & \\ 0 & & \widehat{\text{var}}\{\hat{F}_m(x)\} \end{bmatrix}.$$

Then, the function $P(\boldsymbol{\lambda})$ can be written as

$$\begin{aligned} P(\boldsymbol{\lambda}) &= \int \Bigg[\left\{ \hat{F}_1(x) - \sum_{i=2}^{m} \lambda_i \hat{F}_i(x) - \left(1 - \sum_{i=2}^{m} \lambda_i\right) \hat{F}_1(x) \right\}^2 \\ &\qquad + \left(1 - \sum_{i=2}^{m} \lambda_i\right)^2 \widehat{\text{var}}\left\{\hat{F}_1(x)\right\} + \sum_{i=2}^{m} \lambda_i^2 \widehat{\text{var}}\left\{\hat{F}_i(x)\right\} \Bigg] \mathrm{d}\mu(x) \\ &= \int \Bigg[\left[\sum_{i=2}^{m} \lambda_i \left\{ \hat{F}_1(x) - \hat{F}_i(x) \right\} \right]^2 \\ &\qquad + (1 - \tilde{\boldsymbol{\lambda}}^{\mathsf{T}} \mathbf{1})^2 \widehat{\text{var}}\left\{\hat{F}_1(x)\right\} + \sum_{i=2}^{m} \lambda_i^2 \widehat{\text{var}}\left\{\hat{F}_i(x)\right\} \Bigg] \mathrm{d}\mu(x) \\ &= \int \left[\{\tilde{\boldsymbol{\lambda}}^{\mathsf{T}} \mathcal{F}(x)\}^2 + (1 - 2\tilde{\boldsymbol{\lambda}}^{\mathsf{T}} \mathbf{1} + \tilde{\boldsymbol{\lambda}}^{\mathsf{T}} \mathbf{1}\mathbf{1}^{\mathsf{T}} \tilde{\boldsymbol{\lambda}}) \widehat{\text{var}}\left\{\hat{F}_1(x)\right\} + \tilde{\boldsymbol{\lambda}}^{\mathsf{T}} \mathbf{V}(x) \tilde{\boldsymbol{\lambda}} \right] \mathrm{d}\mu(x) \\ &= \int \Bigg[\tilde{\boldsymbol{\lambda}}^{\mathsf{T}} \left[\mathcal{F}(x)\mathcal{F}(x)^{\mathsf{T}} + \mathbf{V}(x) + \mathbf{1}\mathbf{1}^{\mathsf{T}} \widehat{\text{var}}\left\{\hat{F}_1(x)\right\} \right] \tilde{\boldsymbol{\lambda}} \\ &\qquad - 2\tilde{\boldsymbol{\lambda}}^{\mathsf{T}} \mathbf{1} \widehat{\text{var}}\left\{\hat{F}_1(x)\right\} + \widehat{\text{var}}\left\{\hat{F}_1(x)\right\} \Bigg] \mathrm{d}\mu(x) \\ &= \tilde{\boldsymbol{\lambda}}^{\mathsf{T}} \bar{\mathsf{A}} \tilde{\boldsymbol{\lambda}} - 2\tilde{\boldsymbol{\lambda}}^{\mathsf{T}} \mathbf{1} \bar{b} + \bar{b} \end{aligned} \tag{2.3}$$

where

$$\begin{aligned} \bar{\mathsf{A}} &= \int \left[\mathcal{F}(x)\mathcal{F}(x)^{\mathsf{T}} + \mathbf{V}(x) + \mathbf{1}\mathbf{1}^{\mathsf{T}} \widehat{\text{var}}\left\{\hat{F}_1(x)\right\} \right] \mathrm{d}\mu(x) \\ \bar{b} &= \int \widehat{\text{var}}\left\{\hat{F}_1(x)\right\} \mathrm{d}\mu(x). \end{aligned}$$

Hence, the objective function $P(\boldsymbol{\lambda})$ can be written as a quadratic function of $\tilde{\boldsymbol{\lambda}}$ involving matrices and vectors whose entries depend on $\hat{F}_i(x)$ and its estimated variance. Note that a similar quadratic development is possible with $\boldsymbol{\lambda}$ as well.

2.4 An Algorithm to Compute the MAMSE Weights

To improve inference about Population 1, we attempt to use the data from $m-1$ other populations. Suppose that one of the other populations is different enough that its sample does not even overlap with that of Population 1. Extracting useful information from that sample seems unlikely and it is nearly impossible to evaluate to which degree the populations are similar, except for suspecting that they are probably quite different.

We suggest to preprocess the samples from the m populations to discard those that are clearly different from the target population. Some samples may also be discarded because they are difficult to compare with the target. Specifics of the preprocessing are described for all three versions of the MAMSE weights in their respective chapters. Typically, the preprocessing steps will also insure that Assumption 2.1 is respected.

Assumption 2.1

$$\int \widehat{\mathrm{var}}\{\hat{F}_i(x)\}\,\mathrm{d}\mu(x) > 0$$

for any Population i that is considered for the optimization of (2.2).

All samples that fail the preprocessing are assigned weights of zero, hence they do not appear in the MAMSE objective function. For simplicity, the notation used in this section does not reflect the possibly reduced number of populations considered: we suppose that m populations remain after preprocessing.

If we ignore the constraints $\lambda_i \geq 0$, the MAMSE weights are the solution to the equation $\bar{\mathsf{A}}\tilde{\boldsymbol{\lambda}} = \bar{b}\mathbf{1}$. To ensure the weights are nonnegative, we apply the following algorithm and denote its solution by $\boldsymbol{\lambda}^*$ (or $\tilde{\boldsymbol{\lambda}}^*$).

1. Solve the equation $\bar{\mathsf{A}}\tilde{\boldsymbol{\lambda}} = \bar{b}\mathbf{1}$;

2. if all the weights obtained are nonnegative, stop. Otherwise set the negative weights to 0, ignore the corresponding samples and repeat from Step 1 with the reduced

system. The weight allocated to Population 1 from Step 1 cannot be negative (see Lemma 2.4). If no other samples are left, then $\tilde{\boldsymbol{\lambda}} = \mathbf{0}$ and $\lambda_1 = 1$.

The objective function $P(\boldsymbol{\lambda})$ is quadratic and positive (thus convex). Since the constraints form a convex set, intuition suggests that $\boldsymbol{\lambda}^*$ should be the global constrained minimum. We prove this more formally next.

Consider the generic program

$$\begin{aligned} \text{minimize} \quad & P(\boldsymbol{\lambda}) \\ \text{subject to} \quad & \mathbf{h}(\boldsymbol{\lambda}) \leq \mathbf{0} \end{aligned}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\mathbf{h}(\boldsymbol{\lambda}) = [h_1(\boldsymbol{\lambda}), \ldots, h_k(\boldsymbol{\lambda})]^\mathsf{T}$ is a vector of functions, each being from $\mathbb{R}^m$ to $\mathbb{R}$. Let $\boldsymbol{\nabla} P(\boldsymbol{\lambda})$ denote the gradient of P and $\mathbf{P}(\boldsymbol{\lambda})$ its Hessian. Similarly, $\boldsymbol{\nabla} h_i(\boldsymbol{\lambda})$ is the gradient of $h_i(\boldsymbol{\lambda})$ and $\mathbf{H}_i(\boldsymbol{\lambda})$ its Hessian. By definition, an $m \times m$ matrix B is positive definite (denoted $B \succ 0$) if $\mathbf{y}^\mathsf{T} B \mathbf{y} > 0$ for all $\mathbf{y} \in \mathbb{R}^m \backslash \{\mathbf{0}\}$. The Kuhn-Tucker conditions are necessary conditions that any solution to a program like (2.2) must respect. The second order version that follows (see for instance Luenberger (2003), page 316) are necessary and sufficient conditions to show that a given $\boldsymbol{\lambda}^*$ solves program (2.2).

Theorem 2.1 (Kuhn-Tucker Second Order Sufficiency Conditions) *Let $h_1, \ldots, h_k$ and P be continuous and twice differentiable functions from $\mathbb{R}^m$ to $\mathbb{R}$. Sufficient conditions that a point $\boldsymbol{\lambda}^*$ be a strict relative minimum point of the program above are that there exists $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]^\mathsf{T} \in \mathbb{R}^k$ such that $\boldsymbol{\mu} \geq 0$, $\boldsymbol{\mu}^\mathsf{T} \mathbf{h}(\boldsymbol{\lambda}^*) = 0$,*

$$\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*) + \sum_{i=1}^{k} \mu_i \boldsymbol{\nabla} h_i(\boldsymbol{\lambda}^*) = \mathbf{0} \tag{2.4}$$

and the matrix

$$\mathbf{P}(\boldsymbol{\lambda}^*) + \sum_{i=1}^{k} \mu_i \mathbf{H}_i(\boldsymbol{\lambda}^*) \succ 0. \tag{2.5}$$

Note that from Equation (2.3), we have $\boldsymbol{\nabla} P(\boldsymbol{\lambda}) = \bar{\mathsf{A}}\tilde{\boldsymbol{\lambda}} - \mathbf{1}\bar{b}$ and $\mathbf{P}(\boldsymbol{\lambda}) = \bar{\mathsf{A}}$. The function $P(\boldsymbol{\lambda})$ and its derivatives do not depend on λ_1 since it was replaced by $\lambda_1 = 1 - \mathbf{1}^\mathsf{T}\tilde{\boldsymbol{\lambda}}$. Consequently, it is implicitly understood in the following that P and h_i are functions of $\tilde{\boldsymbol{\lambda}} \in \mathbb{R}^{m-1}$, even when we write $P(\boldsymbol{\lambda})$ and $h_i(\boldsymbol{\lambda})$ rather than $P(\tilde{\boldsymbol{\lambda}})$ and $h_i(\tilde{\boldsymbol{\lambda}})$.

Lemma 2.1 *The Hessian matrix* $\mathbf{P}(\boldsymbol{\lambda})$ *is positive definite.*

Proof of Lemma 2.1. Remember that

$$\bar{\mathsf{A}} = \int \left[\mathcal{F}(x)\mathcal{F}(x)^\mathsf{T} + \mathbf{V}(x) + \mathbf{1}\mathbf{1}^\mathsf{T} \widehat{\mathrm{var}}\{\hat{F}_1(x)\} \right] \mathrm{d}\mu(x).$$

For any fixed x, each term of the integrand as written above are nonnegative definite. In particular, for any $\mathbf{y} \in \mathbb{R}^{m-1} \backslash \{\mathbf{0}\}$,

$$\mathbf{y}^\mathsf{T} \left\{ \mathcal{F}(x)\mathcal{F}(x)^\mathsf{T} \right\} \mathbf{y} \geq 0$$

and

$$\mathbf{y}^\mathsf{T}\mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{y} \int \widehat{\mathrm{var}}\{\hat{F}_1(x)\} \, \mathrm{d}\mu(x) \geq 0$$

because of Assumption 2.1. As another consequence of Assumption 2.1, the diagonal elements of $\int \mathbf{V}(x) \, \mathrm{d}\mu(x)$ are strictly positive, meaning that

$$\mathbf{y}^\mathsf{T} \left[\int \mathbf{V}(x) \, \mathrm{d}\mu(x) \right] \mathbf{y} \;\; = \;\; \sum_{i=1}^{m-1} y_i^2 \left[\int \widehat{\mathrm{var}}\{\hat{F}_i(x)\} \, \mathrm{d}\mu(x) \right] > 0$$

for any $\mathbf{y} \in \mathbb{R}^{m-1} \backslash \{\mathbf{0}\}$. Consequently, $\mathbf{y}^\mathsf{T}\bar{\mathsf{A}}\mathbf{y} > 0$, i.e. the Hessian of P, $\mathbf{P}(\boldsymbol{\lambda}) = \bar{\mathsf{A}}$, is positive definite. ∎

Corollary 2.1 *Equation (2.5) is satisfied.*

Proof of Corollary 2.1. In our implementation of the general Kuhn-Tucker conditions, $h_i(\boldsymbol{\lambda}) \equiv -\lambda_{i+1}$. Therefore, $\mathbf{H}_i(\boldsymbol{\lambda}) \stackrel{\Delta}{=} \boldsymbol{\nabla}^{\mathsf{T}}\boldsymbol{\nabla} h_i(\boldsymbol{\lambda}) = \mathbf{0}$ are null matrices. From Lemma 2.1, we know that $\mathbf{P}(\boldsymbol{\lambda}^*)$ is positive definite, hence Equation (2.5) is satisfied. ∎

Applying the algorithm above will change negative weights to $\lambda_{i+1} = 0$ for some $i \in I^C \subset \{1, \ldots, m-1\}$ where I^C may be null. The set I contains the remaining indices and may also be null.

Let $J \subset \{1, \ldots, m-1\}$ be a possibly null subset of indices, then $A_{I,J}$ is the sub-matrix of A for the rows $i \in I$ and the columns $j \in J$. We define the subvector $\boldsymbol{\lambda}_I$ similarly.

The proposed algorithm involves solving reduced systems where the rows and columns for $i \in I^C$ are excluded. The system of equations that has to be solved then involves the matrix

$$\mathcal{A}_I = \int \left[\mathcal{F}_I(x)\mathcal{F}_I(x)^{\mathsf{T}} + \mathbf{V}_{I,I}(x) + \mathbf{1}_I\mathbf{1}_I^{\mathsf{T}}\widehat{\mathrm{var}}\{\hat{F}_1(x)\}\right] \mathrm{d}\mu(x).$$

For convenience of exposition, suppose that the order of appearance of the λ_i in $\tilde{\boldsymbol{\lambda}}$ are such that all the λ_i that are "forced" to be zero are last. Then, with $\mathcal{F}(x) = \left[\begin{array}{c} \mathcal{F}_I(x) \\ \hdashline \mathcal{F}_{I^C}(x) \end{array}\right]$ we can write

$$\begin{aligned}
\bar{\mathsf{A}} &= \int \left[\left[\begin{array}{c} \mathcal{F}_I(x) \\ \hdashline \mathcal{F}_{I^C}(x) \end{array}\right]\left[\begin{array}{c} \mathcal{F}_I(x) \\ \hdashline \mathcal{F}_{I^C}(x) \end{array}\right]^{\mathsf{T}} + \mathbf{V}(x) + \mathbf{1}\mathbf{1}^{\mathsf{T}}\widehat{\mathrm{var}}\{\hat{F}_1(x)\}\right] \mathrm{d}\mu(x) \\
&= \int \left[\begin{array}{c:c} \mathcal{F}_I(x)\mathcal{F}_I(x)^{\mathsf{T}} + \mathbf{V}_{I,I}(x) & \mathcal{F}_I(x)\mathcal{F}_{I^C}(x)^{\mathsf{T}} \\ \hdashline \mathcal{F}_{I^C}(x)\mathcal{F}_I(x)^{\mathsf{T}} & \mathcal{F}_{I^C}(x)\mathcal{F}_{I^C}(x)^{\mathsf{T}} + \mathbf{V}_{I^C,I^C}(x) \end{array}\right] + \mathbf{1}\mathbf{1}^{\mathsf{T}}\widehat{\mathrm{var}}\{\hat{F}_1(x)\}\,\mathrm{d}\mu(x) \\
&= \left[\begin{array}{c:c} \bar{\mathsf{A}}_{I,I} & \bar{\mathsf{A}}_{I,I^C} \\ \hdashline \bar{\mathsf{A}}_{I^C,I} & \bar{\mathsf{A}}_{I^C,I^C} \end{array}\right].
\end{aligned}$$

In particular, $\mathcal{A}_I = \bar{\mathsf{A}}_{I,I}$ and $\mathcal{A}_{I^C} = \bar{\mathsf{A}}_{I^C,I^C}$. Therefore, the last step of the proposed algorithm is to solve the system of equations $\mathcal{A}_I\tilde{\boldsymbol{\lambda}}_I = \bar{\mathsf{A}}_{I,I}\tilde{\boldsymbol{\lambda}}_I = \mathbf{1}_I\bar{b}$.

Lemma 2.2 *If $I^C \neq \emptyset$ and $\mathbf{y} \in \mathbb{R}^{m-1} \backslash \{\mathbf{0}\}$ is any nonnegative vector with $\mathbf{y}_I = \mathbf{0}$ and $\mathbf{y}_{I^C} > \mathbf{0}$, then $\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)^\mathsf{T} \mathbf{y} > \mathbf{0}$.*

Proof of Lemma 2.2. First note that the expression $\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)^\mathsf{T} \mathbf{y}$ corresponds to the directional derivative of P at $\boldsymbol{\lambda}^*$ in the direction $\mathbf{y}$. Next consider the unit vector $\mathbf{e}_i \in \mathbb{R}^{m-1}$ whose i^{th} element is 1. For $i \in I^C$, the global unconstrained minimum of the convex function P is outside of the half-space $\lambda_{i+1} \geq 0$. Therefore, P increases in the direction $\mathbf{e}_i$ at $\boldsymbol{\lambda}^*$ and thus $\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)^\mathsf{T} \mathbf{e}_i > 0$.

Finally, the hypothesized vector $\mathbf{y}$ can be expressed as a linear combination of vectors $\{\mathbf{e}_i : i \in I^C\}$ with nonnegative coefficients y_i. Therefore,

$$\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)^\mathsf{T} \mathbf{y} = \sum_{i \in I^C} y_i \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)^\mathsf{T} \mathbf{e}_i > 0.$$

■

Although $I = \emptyset$ or $I^C = \emptyset$ may occur, the following proofs hold under these special cases.

Lemma 2.3 *The proposed algorithm solves the quadratic program*

$$\begin{aligned} &\textit{minimize} && P(\boldsymbol{\lambda}) \\ &\textit{subject to} && \{\lambda_i \geq 0, i = 2, \ldots, m\}. \end{aligned}$$

Proof of Lemma 2.3. To verify that the Kuhn-Tucker conditions are satisfied, first note that for $i = 1, \ldots, m-1$ the functions $h_i(\boldsymbol{\lambda}) \equiv -\lambda_{i+1}$ are continuous and twice differentiable. The quadratic objective function $P(\boldsymbol{\lambda})$ shares the same smoothness properties. Moreover, Corollary 2.1 establishes that Equation (2.5) holds.

At termination, the algorithm yields $\tilde{\boldsymbol{\lambda}}^*_I \geq \mathbf{0}$ and $\tilde{\boldsymbol{\lambda}}^*_{I^C} = \mathbf{0}$. The proposed solution $\boldsymbol{\lambda}^*$ is thus in the feasible set. It remains to show that there exists a $\boldsymbol{\mu}$ satisfying the Kuhn-Tucker conditions stated earlier. We will show that $\boldsymbol{\mu} = \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)$ satisfies the required properties. Expression (2.4) can be written

$$\boldsymbol{\nabla} P(\boldsymbol{\lambda}^*) + \sum_{i=1}^{m-1} \mu_i (-\mathbf{e}_i) = \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*) - \boldsymbol{\mu} = \mathbf{0}$$

and clearly holds for $\boldsymbol{\mu} = \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)$. The other Kuhn-Tucker conditions require that $\boldsymbol{\mu} \geq \mathbf{0}$ and $\boldsymbol{\mu}^{\mathsf{T}} \tilde{\boldsymbol{\lambda}}^* = \mathbf{0}$.

$\underline{\boldsymbol{\mu} \geq \mathbf{0}}$

The last step of the algorithm before termination is to solve $\bar{\mathsf{A}}_{I,I} \tilde{\boldsymbol{\lambda}}_I = \mathbf{1}_I \bar{b}$. Therefore,

$$\boldsymbol{\mu}_I = \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*)_I = \left[\bar{\mathsf{A}} \boldsymbol{\lambda}^* - \mathbf{1} \bar{b}\right]_I = \bar{\mathsf{A}}_{I,I} \tilde{\boldsymbol{\lambda}}_I^* + \bar{\mathsf{A}}_{I,I^C} \tilde{\boldsymbol{\lambda}}_{I^C}^* - \mathbf{1}_I \bar{b} = \mathbf{0}$$

since $\tilde{\boldsymbol{\lambda}}_{I^C}^* = \mathbf{0}$.

In addition, we have from Lemma 2.2 that $\mu_i = \boldsymbol{\mu}^{\mathsf{T}} \mathbf{e}_i = \boldsymbol{\nabla} P(\boldsymbol{\lambda}^*) \mathbf{e}_i > 0$ for all $i \in I^C$, and hence $\boldsymbol{\mu}_{I^C} > \mathbf{0}$. Therefore, $\boldsymbol{\mu} \geq \mathbf{0}$.

$\underline{\boldsymbol{\mu}^{\mathsf{T}} \tilde{\boldsymbol{\lambda}}^* = \mathbf{0}}$

We can write the condition $\boldsymbol{\mu}^{\mathsf{T}} \tilde{\boldsymbol{\lambda}}^* = 0$ as $\boldsymbol{\mu}_I^{\mathsf{T}} \tilde{\boldsymbol{\lambda}}_I^* + \boldsymbol{\mu}_{I^C}^{\mathsf{T}} \tilde{\boldsymbol{\lambda}}_{I^C}^* = 0$. It is shown above that $\boldsymbol{\mu}_I = \mathbf{0}$, hence $\boldsymbol{\mu}_I^{\mathsf{T}} \tilde{\boldsymbol{\lambda}}_I^* = 0$. Moreover, the definition of the set I implies that $\tilde{\boldsymbol{\lambda}}_{I^C}^* = \mathbf{0}$, thus $\boldsymbol{\mu}_{I^C}^{\mathsf{T}} \tilde{\boldsymbol{\lambda}}_{I^C}^* = 0$ and the condition is satisfied.

Consequently, the solution found by the proposed algorithm is a strict relative minimum since it satisfies the sufficient Kuhn-Tucker conditions. ■

Lemma 2.4 *The solution found by the proposed algorithm satisfies the additional constraint* $\sum_{i=2}^{m} \lambda_i < 1$, *or equivalently,* $\lambda_1 > 0$.

Proof of Lemma 2.4. The solution found by the algorithm satisfies $\bar{\mathsf{A}}_{I,I} \boldsymbol{\lambda}_I^* = \mathbf{1}_I \bar{b}$. Expanding $\bar{\mathsf{A}}_{I,I}$ in this equation yields

$$\left[\int \left[\mathcal{F}_I(x) \mathcal{F}_I(x)^{\mathsf{T}} + \mathbf{V}_{I,I}(x) + \mathbf{1}_I \mathbf{1}_I^{\mathsf{T}} \widehat{\text{var}} \left\{ \hat{F}_1(x) \right\} \right] \, \mathrm{d}\mu(x) \right] \tilde{\boldsymbol{\lambda}}_I^* = \mathbf{1}_I \bar{b}.$$

By subtracting $\bar{b} \mathbf{1}_I \mathbf{1}_I^{\mathsf{T}} \tilde{\boldsymbol{\lambda}}_I^*$ from both sides and multiplying the resulting equation by $\tilde{\boldsymbol{\lambda}}_I^{*\mathsf{T}}$ on

the left, we have

$$\tilde{\boldsymbol{\lambda}}_I^{*\mathsf{T}} \left[\int \left\{ \mathcal{F}_I(x)\mathcal{F}_I(x)^\mathsf{T} + \mathbf{V}_{I,I}(x) \right\} \mathrm{d}\mu(x) \right] \tilde{\boldsymbol{\lambda}}_I^* = \bar{b}\tilde{\boldsymbol{\lambda}}_I^{*\mathsf{T}} \left(\mathbf{1}_I - \mathbf{1}_I \mathbf{1}_I^\mathsf{T} \tilde{\boldsymbol{\lambda}}_I^* \right)$$

$$= \bar{b}\tilde{\boldsymbol{\lambda}}_I^{*\mathsf{T}} \mathbf{1}_I \left(1 - \mathbf{1}_I^\mathsf{T} \tilde{\boldsymbol{\lambda}}_I^* \right) = \bar{b} \left(1 - \sum_{i \in I} \lambda_{i+1}^* \right) \left(\sum_{i \in I} \lambda_{i+1}^* \right).$$

By the same argument as in Lemma 2.1, the matrix on the left hand-side is positive definite, and hence the expression itself is positive. Since $\bar{b}$ and $\tilde{\boldsymbol{\lambda}}_I^*$ are positive, we necessarily have $1 - \sum_{i \in I} \lambda_{i+1}^* > 0$. Hence, the solution to the program in Lemma 2.3 always satisfies the additional constraint $\sum_{i \in I} \lambda_{i+1}^* = \sum_{i=2}^m \lambda_i^* < 1$ (remember that $\tilde{\boldsymbol{\lambda}}_{I^C}^* = \mathbf{0}$). This inequality is equivalent to $\lambda_1^* > 0$.

Regarding the comment to the effect that λ_1 cannot be negative for intermediate steps, consider the development above for such steps where $\boldsymbol{\lambda}_I$ may still contain negative values. Note that the left-hand-side of the expression is still positive because of its positive definiteness. Moreover, the right-hand-side can be written as $\lambda_1(1-\lambda_1)\bar{b}$, meaning that $\lambda_1(1-\lambda_1)$ is positive. Therefore, $\lambda_1 \in (0,1)$, except if $I = \emptyset$ in which case $\lambda_1 = 1$ and $\tilde{\boldsymbol{\lambda}} = \mathbf{0}$. ■

Theorem 2.2 *The proposed algorithm solves the quadratic program*

$$\begin{array}{ll} \text{minimize} & P(\boldsymbol{\lambda}) \\ \textit{subject to} & \{\lambda_i \geq 0, i = 1, \ldots, m\} \textit{ and } \sum_{i=1}^m \lambda_i = 1. \end{array}$$

Proof of Theorem 2.2. The result follows from Lemmas 2.3 and 2.4. ■

We now define MAMSE weights for specific types of data and study their properties.

Chapter 3

MAMSE Weights for Univariate Data

We define MAMSE weights for univariate data by using the empirical distribution function (EDF) based on the sample from each population as an estimate of their CDF. This choice yields nonparametric MAMSE weights with interesting asymptotic properties that are studied. In particular, the MAMSE-weighted mixture of empirical distributions converges to the target distribution and the maximum weighted likelihood estimate (MWLE) is a strongly consistent estimate of the true parameter when the MAMSE weights are used. Simulation studies evaluate the performance of the MWLE on finite samples.

3.1 Notation and Review

Let $(\Omega, \mathcal{B}(\Omega), P)$ be the sample space on which the random variables

$$X_{ij}(\omega) : \Omega \to \mathbb{R}, \qquad i = 1, \ldots, m,\ j \in \mathbb{N}$$

are defined. The X_{ij} are assumed to be independent with continuous distribution F_i.

We consider samples of nondecreasing sizes: for any positive integer k, the random variables $\{X_{ij} : i = 1, \ldots, m,\ j = 1, \ldots, n_{ik}\}$ are observed. Moreover, the sequences of sample sizes are such that $n_{1k} \to \infty$ as $k \to \infty$. We do not require that the sample sizes of the other populations tend to ∞, nor do we restrict the rate at which they increase.

Let $\hat{F}_{ik}(x)$ be the EDF based on the sample X_{ij}, $j = 1, \ldots, n_{ik}$, i.e.

$$\hat{F}_{ik}(x) = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \mathbb{1}\left\{X_{ij}(\omega) \leq x\right\}.$$

The empirical measure $\mathrm{d}\hat{F}_{ik}(x)$ allocates a weight $1/n_{ik}$ to each of the observations X_{ij}, $j = 1, \ldots, n_{ik}$.

Under the assumption that data from each population are independent and identically distributed as is the case here, the empirical distribution is a nonparametric estimate of the CDF. Indeed, the Glivenko-Cantelli lemma (see e.g. Durrett (2005) page 58) states that

$$\sup_x \left|\hat{F}_{ik}(x) - F_i(x)\right| \to 0$$

almost surely as $n_{ik} \to \infty$.

For any fixed value of x, one may also note that $n_{ik}\hat{F}_{ik}(x)$ follows a Binomial distribution with parameters n_{ik} and $F_i(x)$. Hence, the asymptotic normality of the EDF at fixed x follows from the central limit theorem.

3.2 Definition of the MAMSE Weights

For univariate data, we define the MAMSE weights based on the EDFs of the m populations.

The MAMSE objective function assesses $\sum_{i=1}^{m} \lambda_i \hat{F}_{ik}(x)$ as an estimate of $\hat{F}_{1k}(x)$ in terms of bias and variance. As a preselection step, any sample whose range of values does not overlap with that of Population 1 will be discarded. For the remaining m samples, the MAMSE weights will be the vector $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^{\mathsf{T}}$ that minimizes

$$P_k(\boldsymbol{\lambda}) = \int \left[\left\{ \hat{F}_{1k}(x) - \sum_{i=1}^{m} \lambda_i \hat{F}_{ik}(x) \right\}^2 + \sum_{i=1}^{m} \frac{\lambda_i^2}{n_{ik}} \hat{F}_{ik}(x) \left\{1 - \hat{F}_{ik}(x)\right\} \right] \mathrm{d}\hat{F}_{1k}(x) \qquad (3.1)$$

under the constraints $\lambda_i \geq 0$ and $\sum_{i=1}^{m} \lambda_i = 1$.

Equation (3.1) is a special case of Equation (2.1) where $\mathrm{d}\mu(x) = \mathrm{d}\hat{F}_{1k}(x)$ and where the substitution

$$\widehat{\mathrm{var}}\{\hat{F}_i(x)\} = \frac{1}{n_{ik}}\hat{F}_{ik}(x)\left\{1 - \hat{F}_{ik}(x)\right\}$$

is based on the variance of the Binomial variable $n_{ik}\hat{F}_i(x)$.

The choice of $\mathrm{d}\mu(x) = \mathrm{d}\hat{F}_{1k}(x)$ allows to integrate where the target distribution $F_1(x)$ has most of its mass. Other choices for $\mu(x)$ could be explored in the future.

For $\omega \in \Omega$ and $k \in \mathbb{N}$, we denote the MAMSE weights by $\boldsymbol{\lambda}_k(\omega) = [\lambda_{1k}(\omega), \ldots, \lambda_{mk}(\omega)]^{\mathsf{T}}$. The MAMSE weights are used to define an estimate of the distribution $F_1(x)$,

$$\hat{G}_k(x) = \sum_{i=1}^{m} \lambda_{ik}(\omega)\hat{F}_{ik}(x),$$

the MAMSE-weighted EDF.

3.3 Computing the MAMSE Weights

The algorithm proposed in Section 2.4 applies to the MAMSE weights for univariate data. To prove the convergence of the algorithm, it is sufficient to show that Assumption 2.1 is respected when the MAMSE weights are defined using Expression (3.1).

Lemma 3.1 *Let x be any point within the range of the data set from Population i, i.e. let*

$$x \in \left(\min_{j=1,\ldots,n_{ik}} X_{ij} \, , \, \max_{j=1,\ldots,n_{ik}} X_{ij} \right).$$

Then,

$$\widehat{\mathrm{var}}\left\{\hat{F}_{ik}(x)\right\} = \frac{1}{n_{ik}}\hat{F}_{ik}(x)\left\{1 - \hat{F}_{ik}(x)\right\} > 0.$$

Proof of Lemma 3.1. Note that

$$0 < \hat{F}_{ik}(x) = \frac{\#\{X_{ij} \leq x : j \leq n_{ik}\}}{n_{ik}} < 1$$

since the numerator is at least one and at most $n_{ik} - 1$, or otherwise x would be outside the specified interval. Consequently,

$$\widehat{\text{var}}\left\{\hat{F}_{ik}(x)\right\} = \frac{1}{n_{ik}}\hat{F}_{ik}(x)\left\{1 - \hat{F}_{ik}(x)\right\} > 0$$ ■

Theorem 3.1 *Assumption 2.1 is respected for the definition of MAMSE weights suggested in Section 3.2.*

Proof of Theorem 3.1. The preprocessing step described in Section 3.2 requires that any sample considered overlaps with that of Population 1. In particular this means that at least one of the data points from Population 1 will fall within the range of the sample of Population i, meaning that

$$\begin{aligned}\int \widehat{\text{var}}\left\{\hat{F}_{ik}(x)\right\}\, \mathrm{d}\hat{F}_1(x) &= \int \frac{1}{n_{ik}}\hat{F}_{ik}(x)\left\{1 - \hat{F}_{ik}(x)\right\}\, \mathrm{d}\hat{F}_1(x) \\ &= \frac{1}{n_{ik}n_{1k}}\sum_{j=1}^{n_{1k}} \hat{F}_{ik}\{X_{1j}(\omega)\}\left[1 - \hat{F}_{ik}\{X_{1j}(\omega)\}\right] > 0\end{aligned}$$

since all the terms in the sum are nonnegative and at least one must be positive by Lemma 3.1. ■

3.4 Structural Properties of the MAMSE Weights

Choosing the empirical distribution function to define the MAMSE weights implies some invariance properties that are discussed next.

Theorem 3.2 *The MAMSE weights are invariant to a strictly increasing transformation of the data.*

Proof of Theorem 3.2. Let $X_{ij} = g(Y_{ij})$ where g is a strictly increasing function of the real line. Let H_i denote the cumulative distribution function of Y_{ij}. Then for all y, $x = g(y)$

and any $i = 1, \ldots, m$

$$\begin{aligned} \hat{H}_{ik}(y) &= \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \mathbb{1}\{Y_{ij} \leq y\} = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \mathbb{1}\{g(Y_{ij}) \leq g(y)\} \\ &= \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \mathbb{1}\{X_{ij} \leq x\} = \hat{F}_{ik}(x). \end{aligned}$$

Since $P(\boldsymbol{\lambda})$ is integrated with respect to $\mathrm{d}\hat{F}_{1k}$, a discrete measure, there is no Jacobian of transformation in the integral and replacing all $\hat{F}_{ik}$ by the corresponding $\hat{H}_{ik}$ will not change the expression $P(\boldsymbol{\lambda})$, nor its maximum. ■

Theorem 3.3 *The MAMSE weights do not depend on the parametric model $f(x|\theta)$ used in the weighted likelihood.*

Proof of Theorem 3.3. The result follows immediately from the definition of MAMSE weights and the choice of the nonparametric empirical distribution functions as estimates of F_i. ■

Theorem 3.4 *The MWLE based on MAMSE weights is invariant under a one-to-one reparametrization of $f(x|\theta)$ into $h(x|\tau) \stackrel{\Delta}{=} f\{x|h(\tau)\}$, i.e. $\hat{\theta}$ is a MWLE iff $\hat{\tau}$ is a MWLE.*

Proof of Theorem 3.4. By Theorem 3.3, the MAMSE weights $\boldsymbol{\lambda}_k(\omega) = [\lambda_{1k}(\omega), \ldots, \lambda_{mk}(\omega)]^{\mathsf{T}}$ are invariant to the choice of parametric model $f(x|\theta)$. If τ is such that $\theta = h(\tau)$ and h is a one-to-one mapping of the parameter space, then $\tau_{\max}$ is such that

$$\prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} f\left\{X_{ij}|h(\tau)\right\}^{\lambda_{ik}(\omega)/n_{ik}} \leq \prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} f\left\{X_{ij}|h(\tau_{\max})\right\}^{\lambda_{ik}(\omega)/n_{ik}}$$

for all τ if and only if $\theta_{\max} = h(\tau_{\max})$ is such that

$$\prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} f(X_{ij}|\theta)^{\lambda_{ik}(\omega)/n_{ik}} \leq \prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} f(X_{ij}|\theta_{\max})^{\lambda_{ik}(\omega)/n_{ik}}$$

for all θ. Hence, the MWLE possesses the same functional invariance property as the MLE if we use the MAMSE weights. ■

3.5 Strong Uniform Convergence of the Weighted Empirical CDF

This section explores the large sample behavior of the MAMSE-weighted EDF. Note that asymptotic results for adaptive weights are obtained by Wang, van Eeden and Zidek (2004), but they do not apply here because their assumption that the weight allocated to Population 1 tends to one as the sample size increases may not be satisfied (see Section 3.8 for more details on the asymptotic behavior of the MAMSE weights).

The proof of uniform convergence of $\hat{G}_k(x)$ is built as a sequence of lemmas showing that $\hat{G}_k$ is close to $\hat{F}_{1k}$, and ultimately that

$$\sup_x \left|\hat{G}_k(x) - F_1(x)\right| \to 0$$

almost surely as $k \to \infty$.

Whether a sample is rejected in the preprocessing or not may vary with k and ω. However, as the sample sizes increase, the probability that a sample is rejected tends to zero unless the domain of possible values of a Population does not overlap at all with that of Population 1, i.e. unless $P(X_{11} < X_{i1}) = 0$ or 1. Thus, without loss of generality, we suppose that no samples were excluded by the preprocessing.

Lemma 3.2 *For any $\omega \in \Omega$ and $k \in \mathbb{N}$,*

$$\int \left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right|^2 \, \mathrm{d}\hat{F}_{1k}(x) \leq \left(\frac{n_{1k}^2 - 1}{n_{1k}^2}\right) \frac{1}{6n_{1k}}.$$

Proof of Lemma 3.2. For any $\omega \in \Omega$ and $k \in \mathbb{N}$, consider

$$\begin{aligned} I &= \int \left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right|^2 \mathrm{d}\hat{F}_{1k}(x) \\ &\leq \int \left[\left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right|^2 + \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)^2}{n_{ik}} \hat{F}_{ik}(x) \left\{1 - \hat{F}_{ik}(x)\right\} \right] \mathrm{d}\hat{F}_{1k}(x) \end{aligned}$$

By the definition of MAMSE weights, the expression above is minimized in $\boldsymbol{\lambda}$. The sub-optimal choice of weights $[\lambda_1, \ldots, \lambda_m] = [1, 0, \ldots, 0]$ cannot lead to a smaller value of I, i.e.

$$\begin{aligned} I &\leq \int \left[\left|\hat{F}_{1k}(x) - \hat{F}_{1k}(x)\right|^2 + \frac{1}{n_{1k}} \hat{F}_{1k}(x) \left\{1 - \hat{F}_{1k}(x)\right\} \right] \mathrm{d}\hat{F}_{1k}(x) \\ &= \frac{1}{n_{1k}^2} \sum_{j=1}^{n_{1k}} \frac{j}{n_{1k}} \left(1 - \frac{j}{n_{1k}}\right) = \left(\frac{n_{1k}^2 - 1}{n_{1k}^2}\right) \frac{1}{6n_{1k}}. \end{aligned}$$

This bound is tight since the optimal $\boldsymbol{\lambda}$ could be arbitrarily close to $[1, 0, \ldots, 0]^{\mathsf{T}}$, making I arbitrarily close to the bound above. For instance, letting $n_{1k} \to \infty$ while the other n_{ik}'s are held constant will do the trick. ∎

Lemma 3.3 *There exists $\Omega_1 \subset \Omega$ with $P(\Omega_1) = 1$ such that for all $\omega \in \Omega_1$ and any fixed $k \in \mathbb{N}$,*

$$\max_x \left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right| \leq \frac{1}{n_{1k}} + \max_{j \in \{1,\ldots,n_{1k}\}} \left|\hat{G}_k\{X_{1j}(\omega)\} - \hat{F}_{1k}\{X_{1j}(\omega)\}\right|.$$

Proof of Lemma 3.3. Define

$$\Omega_0 = \left\{\omega \in \Omega : \exists i, i', j, j' \text{ with } (i,j) \neq (i',j') \text{ and } X_{ij}(\omega) = X_{i'j'}(\omega)\right\}.$$

Since the distributions F_i are continuous, $P(\Omega_0) = 0$. Fix $k \in \mathbb{N}$ and consider any fixed $\omega \in \Omega_1 = \Omega \backslash \Omega_0$. Note that for $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n_{ik}\}$, $[\min_{i,j} X_{ij}(\omega),\ \max_{i,j} X_{ij}(\omega)]$ is a compact set outside of which $D(x) = |\hat{G}_k(x) - \hat{F}_{1k}(x)| \equiv 0$. Let x_0 be a value maximiz-

ing the bounded function $D(x)$. We treat two cases.

Case 1: $\hat{G}_k(x_0) \leq \hat{F}_{1k}(x_0)$.

Define $x_1 = \max\{X_{1j}(\omega) : j = 1, \ldots, n_{1k},\ X_{1j}(\omega) \leq x_0\}$, the largest data point less than x_0 found in Population 1. The step function $\hat{F}_{1k}(x)$ is right-continuous, nondecreasing and has equal steps of size $1/n_{1k}$ at each observation $X_{1j}(\omega)$. By the choice of x_1, and the definition of the EDF,

$$\hat{F}_{1k}(x_1) = \hat{F}_{1k}(x_0).$$

The step function $\hat{G}_k(x)$ is nondecreasing, thus

$$\hat{G}_k(x_1) \leq \hat{G}_k(x_0).$$

Consequently,

$$\begin{aligned} |\hat{G}_k(x_0) - \hat{F}_{1k}(x_0)| &= \hat{F}_{1k}(x_0) - \hat{G}_k(x_0) \leq \hat{F}_{1k}(x_1) - \hat{G}_k(x_1) \\ &\leq \max_{j \in \{1,\ldots,n_{1k}\}} \left| \hat{F}_{1k}\{X_{1j}(\omega)\} - \hat{G}_k\{X_{1j}(\omega)\} \right| \\ &\leq \frac{1}{n_{1k}} + \max_{j \in \{1,\ldots,n_{1k}\}} \left| \hat{F}_{1k}\{X_{1j}(\omega)\} - \hat{G}_k\{X_{1j}(\omega)\} \right|. \end{aligned}$$

Case 2: $\hat{G}_k(x_0) \geq \hat{F}_{1k}(x_0)$.

Define $x_2 = \min\{X_{1j}(\omega) : j = 1, \ldots, n_{1k},\ X_{1j}(\omega) > x_0\}$, the smallest data point exceeding x_0 found in Population 1. The step function $\hat{F}_{1k}(x)$ is right-continuous, nondecreasing and has equal steps of size $1/n_{1k}$ at each observation $X_{1j}(\omega)$. Therefore,

$$\hat{F}_{1k}(x_2) = \frac{1}{n_{1k}} + \hat{F}_{1k}(x_0).$$

Since $\hat{G}_k(x)$ is nondecreasing,

$$\hat{G}_k(x_2) \geq \hat{G}_k(x_0).$$

Consequently,

$$\begin{aligned} |\hat{G}_k(x_0) - \hat{F}_{1k}(x_0)| &= \hat{G}_k(x_0) - \hat{F}_{1k}(x_0) \\ &\leq \hat{G}_k(x_2) - \hat{F}_{1k}(x_2) + \frac{1}{n_{1k}} \\ &\leq \frac{1}{n_{1k}} + \max_{j\in\{1,\ldots,n_{1k}\}} \left|\hat{F}_{1k}\{X_{1j}(\omega)\} - \hat{G}_k\{X_{1j}(\omega)\}\right| \end{aligned}$$

which completes the proof. ■

Lemma 3.4 *Let a_k be a sequence of numbers such that $\lim_{k\to\infty} a_k^3/n_{1k} = 0$. Then, there exists $\Omega_1 \subset \Omega$ with $P(\Omega_1) = 1$ such that for all $\epsilon > 0$, there exists a k_0 such that $\forall\omega \in \Omega_1$*

$$a_k \max_{j\in\{1,\ldots,n_{1k}\}} \left|\hat{G}_k\{X_{1j}(\omega)\} - \hat{F}_{1k}\{X_{1j}(\omega)\}\right| \leq \epsilon$$

for all $k \geq k_0$.

Proof of Lemma 3.4. Let $\epsilon > 0$. Consider an arbitrary but fixed $\omega \in \Omega_1 = \Omega\backslash\Omega_0$ where Ω_0 has the same definition as in the proof of Lemma 3.3.

Suppose that Lemma 3.4 is false. Then there exists an infinite sequence k_ℓ such that for $\ell \in \mathbb{N}$,

$$\left|\hat{G}_{k_\ell}\{X_{1(j_{0\ell})}(\omega)\} - \hat{F}_{1k_\ell}\{X_{1(j_{0\ell})}(\omega)\}\right| > \epsilon_{k_\ell} = \frac{\epsilon}{a_{k_\ell}} \tag{3.2}$$

for some $j_{0\ell} \in \{1, 2, \ldots, n_{1k_\ell}\}$, where parentheses in the index identify order statistics, i.e. $X_{1(j)}$ is the j^{th} smallest value among $X_{11}, \ldots, X_{1n_{1k_\ell}}$.

Consider a fixed value of ℓ. For simplicity, we drop the index ℓ and revert to k and j_0 that are fixed. Note that

1. $\hat{G}_k(x)$ is a nondecreasing function,

2. $\hat{F}_{1k}(x)$ is a right-continuous nondecreasing step function with equal jumps of $1/n_{1k}$.

We treat two cases:

Case 1: $\hat{G}_k\{X_{1(j_0)}(\omega)\} \geq \hat{F}_{1k}\{X_{1(j_0)}(\omega)\}$.

Note that

$$\hat{F}_{1k}\{X_{1(j_0)}(\omega)\} \leq \hat{G}_k\{X_{1(j_0)}(\omega)\} \leq 1$$

and hence, $\hat{F}_{1k}\{X_{1(j_0)}(\omega)\} \leq 1 - \epsilon_k$ or inequality (3.2) would not hold. Consequently, $j_0 \leq n_{1k} - \lfloor \epsilon_k n_{1k} \rfloor$ and for $i \in \{0, 1, \ldots, \lfloor \epsilon_k n_{1k} \rfloor\}$, we have

$$\begin{aligned} \hat{G}_k\{X_{1(j_0+i)}(\omega)\} &\geq \hat{G}_k\{X_{1(j_0)}(\omega)\}, \\ \hat{F}_{1k}\{X_{1(j_0+i)}(\omega)\} &= \hat{F}_{1k}\{X_{1(j_0)}(\omega)\} + \frac{i}{n_{1k}} \end{aligned}$$

and hence

$$\hat{G}_k\{X_{1(j_0+i)}(\omega)\} - \hat{F}_{1k}\{X_{1(j_0+i)}(\omega)\} \geq \hat{G}_k\{X_{1(j_0)}(\omega)\} - \hat{F}_{1k}\{X_{1(j_0)}(\omega)\} - \frac{i}{n_{1k}} \geq \epsilon_k - \frac{i}{n_{1k}}.$$

As a consequence,

$$\begin{aligned} &\int |\hat{G}_k(x) - \hat{F}_{1k}(x)|^2 \, \mathrm{d}\hat{F}_{1k}(x) \\ &\geq \frac{1}{n_{1k}} \sum_{i=0}^{\lfloor \epsilon_k n_{1k} \rfloor} |\hat{G}_k\{X_{1(j_0+i)}(\omega)\} - \hat{F}_{1k}\{X_{1(j_0+i)}(\omega)\}|^2 \\ &\geq \frac{1}{n_{1k}} \sum_{i=0}^{\lfloor \epsilon_k n_{1k} \rfloor} \left(\epsilon_k - \frac{i}{n_{1k}} \right)^2 = \frac{1}{n_{1k}^3} \sum_{i=0}^{\lfloor \epsilon_k n_{1k} \rfloor} (\epsilon_k n_{1k} - i)^2 \geq \frac{1}{n_{1k}^3} \sum_{i=0}^{\lfloor \epsilon_k n_{1k} \rfloor} i^2 \\ &= \frac{\lfloor \epsilon_k n_{1k} \rfloor (\lfloor \epsilon_k n_{1k} \rfloor + 1)(2\lfloor \epsilon_k n_{1k} \rfloor + 1)}{6 n_{1k}^3} \\ &\geq \frac{1}{3} \left(\frac{\lfloor \epsilon_k n_{1k} \rfloor}{n_{1k}} \right)^3 \end{aligned}$$

By Lemma 3.2, we thus have that

$$\begin{aligned}
\frac{1}{3}\left(\frac{\epsilon_k n_{1k}-1}{n_{1k}}\right)^3 \leq \frac{1}{3}\left(\frac{\lfloor \epsilon_k n_{1k}\rfloor}{n_{1k}}\right)^3 \leq \left(\frac{n_{1k}^2-1}{n_{1k}^2}\right)\frac{1}{6n_{1k}} \leq \frac{1}{6n_{1k}} \\
\Leftrightarrow \quad \left(\frac{\epsilon n_{1k}}{a_k}-1\right)^3 \leq \frac{n_{1k}^2}{2} \quad \Leftrightarrow \quad \frac{\epsilon n_{1k}}{a_k} \leq \frac{n_{1k}^{2/3}}{2^{1/3}}+1 \\
\Leftrightarrow \quad \frac{a_k}{\epsilon n_{1k}} \geq \frac{2^{1/3}}{n_{1k}^{2/3}+2^{1/3}} \quad \Leftrightarrow \quad a_k\,\frac{n_{1k}^{2/3}+2^{1/3}}{n_{1k}} \geq 2^{1/3}\epsilon,
\end{aligned}$$

a contradiction since $a_k^3/n_{1k} \to 0$ and $k_\ell \to \infty$ as $\ell \to \infty$, i.e. the left-hand term converges to 0. Therefore,

$$a_k \max_{j\in\{0,\ldots,n_{1k}\}} \hat{G}_k\{X_{1j}(\omega)\} - \hat{F}_{1k}\{X_{1j}(\omega)\} \leq \epsilon$$

Note that the bound does not depend on the choice of ω.

Case 2: $\hat{G}_k\{X_{1(j_0)}(\omega)\} \leq \hat{F}_{1k}\{X_{1(j_0)}(\omega)\}$.

Note that $\hat{F}_{1k}\{X_{1(j_0)}(\omega)\} \geq \hat{G}_k\{X_{1(j_0)}(\omega)\} \geq 0$. Since both functions are at least ϵ_k apart, $\hat{F}_{1k}\{X_{1(j_0)}(\omega)\} \geq \epsilon_k$ and thus $j_0 \geq \lfloor \epsilon_k n_{1k}\rfloor$. Then for $i \in \{0,1,\ldots,\lfloor \epsilon_k n_{1k}\rfloor\}$, we have

$$\begin{aligned}
\hat{G}_k\{X_{1(j_0-i)}(\omega)\} &\leq \hat{G}_k\{X_{1(j_0)}(\omega)\} \\
\hat{F}_{1k}\{X_{1(j_0-i)}(\omega)\} &= \hat{F}_{1k}\{X_{1(j_0)}(\omega)\} - \frac{i}{n_{1k}}
\end{aligned}$$

and hence

$$\hat{F}_{1k}\{X_{1(j_0-i)}(\omega)\} - \hat{G}_k\{X_{1(j_0-i)}(\omega)\} \geq \hat{F}_{1k}\{X_{1(j_0)}(\omega)\} - \hat{G}_k\{X_{1(j_0)}(\omega)\} - \frac{i}{n_{1k}} \geq \epsilon_k - \frac{i}{n_{1k}}.$$

Then,

$$\begin{aligned}
&\int |\hat{G}_k(x) - \hat{F}_{1k}(x)|^2 \,\mathrm{d}\hat{F}_{1k}(x) \\
&\geq \frac{1}{n_{1k}} \sum_{i=0}^{\lfloor \epsilon_k n_{1k}\rfloor} |\hat{G}_k\{X_{1(j_0-i)}(\omega)\} - \hat{F}_{1k}\{X_{1(j_0-i)}(\omega)\}|^2
\end{aligned}$$

$$\begin{aligned}
&\geq \frac{1}{n_{1k}} \sum_{i=0}^{\lfloor \epsilon_k n_{1k} \rfloor} \left(\epsilon_k - \frac{i}{n_{1k}} \right)^2 = \frac{1}{n_{1k}^3} \sum_{i=0}^{\lfloor \epsilon_k n_{1k} \rfloor} (\epsilon_k n_{1k} - i)^2 \geq \frac{1}{n_{1k}^3} \sum_{i=0}^{\lfloor \epsilon_k n_{1k} \rfloor} i^2 \\
&= \frac{\lfloor \epsilon_k n_{1k} \rfloor (\lfloor \epsilon_k n_{1k} \rfloor + 1)(2\lfloor \epsilon_k n_{1k} \rfloor + 1)}{6 n_{1k}^3} \\
&\geq \frac{1}{3} \left(\frac{\lfloor \epsilon_k n_{1k} \rfloor}{n_{1k}} \right)^3
\end{aligned}$$

By Lemma 3.2, we thus have that

$$\begin{aligned}
\frac{1}{3} \left(\frac{\epsilon_k n_{1k} - 1}{n_{1k}} \right)^3 \leq \frac{1}{3} \left(\frac{\lfloor \epsilon_k n_{1k} \rfloor}{n_{1k}} \right)^3 \leq \left(\frac{n_{1k}^2 - 1}{n_{1k}^2} \right) \frac{1}{6 n_{1k}} \leq \frac{1}{6 n_{1k}} \\
\Leftrightarrow \quad \left(\frac{\epsilon n_{1k}}{a_k} - 1 \right)^3 \leq \frac{n_{1k}^2}{2} \quad \Leftrightarrow \quad \frac{\epsilon n_{1k}}{a_k} \leq \frac{n_{1k}^{2/3}}{2^{1/3}} + 1 \\
\Leftrightarrow \quad \frac{a_k}{\epsilon n_{1k}} \geq \frac{2^{1/3}}{n_{1k}^{2/3} + 2^{1/3}} \quad \Leftrightarrow \quad a_k \frac{n_{1k}^{2/3} + 2^{1/3}}{n_{1k}} \geq 2^{1/3} \epsilon,
\end{aligned}$$

a contradiction since $a_k^3 / n_{1k} \to 0$ and $k_\ell \to \infty$ as $\ell \to \infty$, i.e. the left-hand term converges to 0. Therefore,

$$a_k \max_{j \in \{0, \ldots, n_{1k}\}} \hat{F}_{1k}\{X_{1j}(\omega)\} - \hat{G}_k\{X_{1j}(\omega)\} \leq \epsilon.$$

Combining the two cases, we know that $\forall \epsilon > 0$, $\exists k_0$ such that $\forall k \geq k_0$,

$$a_k \max_{j \in \{0, \ldots, n_{1k}\}} \left| \hat{G}_k\{X_{1j}(\omega)\} - \hat{F}_{1k}\{X_{1j}(\omega)\} \right| \leq \epsilon.$$

Since k_0 does not depend on $\omega \in \Omega_1$ and $P(\Omega_1) = 1$, the uniform convergence is almost sure. ■

Lemma 3.5 *There exists $\Omega_1 \subset \Omega$ with $P(\Omega_1) = 1$ such that for all $\epsilon > 0$, there exists a k_0 such that*

$$\max_x \left| \hat{G}_k(x) - \hat{F}_{1k}(x) \right| \leq \epsilon$$

for all $k \geq k_0$ with the same k_0 for all $\omega \in \Omega_1$.

Proof of Lemma 3.5. Consider the set Ω_1 defined in the proof of Lemma 3.3. For any fixed $\omega \in \Omega_1$,

$$\max_x \left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right| \leq \frac{1}{n_{1k}} + \max_{j\in\{1,\ldots,n_{1k}\}} \left|\hat{G}_k\{X_{1j}(\omega)\} - \hat{F}_{1k}\{X_{1j}(\omega)\}\right|.$$

By Lemma 3.4, $\forall \epsilon > 0$, $\exists k_1$ such that $\forall k \geq k_1$,

$$\max_{j\in\{1,\ldots,n_{1k}\}} \left|\hat{G}_k\{X_{1j}(\omega)\} - \hat{F}_{1k}\{X_{1j}(\omega)\}\right| \leq \frac{\epsilon}{2}$$

for all $\omega \in \Omega_1$. Moreover, $\exists k_2$ such that $\forall k \geq k_2$, $1/n_{1k} \leq \epsilon/2$. Therefore, for all $k \geq k_0 = \max(k_1, k_2)$, we have

$$0 \leq \max_x \left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \qquad \blacksquare$$

Theorem 3.5 *The random variable*

$$\sup_x \left|\hat{G}_k(x) - F_1(x)\right|$$

converges almost surely to 0.

Proof of Theorem 3.5. By Lemma 3.5, $\forall \epsilon > 0$, $\exists k_1$ such that

$$\max_x \left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right| < \frac{\epsilon}{2}$$

$\forall k \geq k_1$ and any $\omega \in \Omega_1$ with $P(\Omega_1) = 1$. The Glivenko-Cantelli theorem states that

$$\sup_x \left|\hat{F}_{1k}(x) - F_1(x)\right| \to 0$$

almost surely as $k \to \infty$. Hence, there exists $\Omega_2 \subset \Omega$ with $P(\Omega_2) = 1$ such that $\forall \epsilon > 0$ and

$\omega \in \Omega_2$, $\exists k_2(\omega)$ with

$$\sup_x \left|\hat{F}_{1k}(x) - F_1(x)\right| < \frac{\epsilon}{2}$$

$\forall k \geq k_2(\omega)$. Consider now $\Omega_0 = \Omega_1 \cap \Omega_2$ and $k_0(\omega) = \max\{k_1, k_2(\omega)\}$. Note that we have $P(\Omega_0) \geq P(\Omega_1) + P(\Omega_2) - 1 = 1$. For any fixed ω, k and x, the inequality

$$\left|\hat{G}_k(x) - \hat{F}_1(x)\right| \leq \left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right| + \left|\hat{F}_{1k}(x) - \hat{F}_1(x)\right|$$

holds, hence for any $\omega \in \Omega_0$ and all $k \geq k_0(\omega)$ we have

$$\sup_x \left|\hat{G}_k(x) - F_1(x)\right| \leq \sup_x \left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right| + \sup_x \left|\hat{F}_{1k}(x) - F_1(x)\right| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Therefore, $\sup_x \left|\hat{G}_k(x) - F_1(x)\right|$ converges almost surely to 0. ■

Corollary 3.1 *Let Y_k be a sequence of random variables with distribution $\hat{G}_k(y)$, then Y_k converge weakly to the variable Y with distribution $F_1(y)$ as $k \to \infty$.*

Proof of Corollary 3.1. The result is clear from Theorem 3.5 and the definition of weak convergence. ■

3.6 Weighted Strong Law of Large Numbers

By the weak convergence shown in Corollary 3.1,

$$\int g(x)\,\mathrm{d}\hat{G}_k(x) \to \int g(x)\,\mathrm{d}F_1 \tag{3.3}$$

as $k \to \infty$ for functions g that respect some regularity conditions. Examples of such results can be found in Section 2.2.b of Durrett (2005) for instance.

The left-hand side of Expression (3.3) can be written as

$$\sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{n_{ik}} \sum_{j=1}^{n_{ik}} g\{X_{ij}(\omega)\},$$

and is hence a form of weighted strong law of large numbers (WSLLN).

In Section 3.7, we use the WSLLN to prove the consistency of the weighted likelihood with MAMSE weights. For that, we need a WSLLN that holds for an unbounded function g with some discontinuities. We prove the needed theorem in this section as it could not be located in the literature.

Lemma 3.6 *Consider any two distribution functions F and G from $\mathbb{R}$ to $[0,1]$ such that $\sup_x |F(x) - G(x)| < \epsilon$ for some $\epsilon > 0$. Then, for any connected set $A \subset \mathbb{R}$,*

$$|\,\mathrm{d}F(A) - \,\mathrm{d}G(A)| \leq 2\epsilon.$$

Proof of Lemma 3.6. Let $B = [a, b] \subset \mathbb{R}$ and define $B_\delta = (a - \delta, b]$ for $\delta > 0$. Let

$$\begin{aligned} e_\delta = |\,\mathrm{d}F(B_\delta) - \,\mathrm{d}G(B_\delta)| &= |F(b) - F(a-\delta) - G(b) + G(a-\delta)| \\ &\leq |F(b) - G(b)| + |F(a-\delta) - G(a-\delta)| \leq 2\epsilon \end{aligned}$$

for all $\delta > 0$. Since δ can be arbitrarily small, $|\,\mathrm{d}F(B) - \,\mathrm{d}G(B)| \leq 2\epsilon$. The result holds for any combination of closed or open boundaries with minor changes to the proof. ∎

Theorem 3.6 *Let $g(x)$ be a function for which $\int |g(x)|\,\mathrm{d}F_1(x) < \infty$. The function $g(x)$ is continuous on $\mathbb{R}$ except possibly on a finite set of point $\{d_1, \ldots, d_L\}$. For each of populations $2, \ldots, m$ at least one of these two conditions hold*

1. *the sample size is bounded: $\forall k \in \mathbb{N}$, $n_{ik} \leq M_i$.*
2. *$\int |g(x)|\,\mathrm{d}F_i(x) < \infty$.*

Further suppose that the sequences of sample sizes are non-decreasing with k for all populations. Then,

$$\left| \int g(x)\, \mathrm{d}\hat{G}_k(x) - \int g(x)\, \mathrm{d}F_1(x) \right| \to 0$$

almost surely as $k \to \infty$.

Proof of Theorem 3.6. We show that for any $\epsilon > 0$, we can find a sequence of inequalities that imply that $\left| \int g(x)\, \mathrm{d}\hat{G}_k(x) - \int g(x)\, \mathrm{d}F_1(x) \right| < \epsilon$ for any large enough k. The inequalities come from truncating g and from approximating it by a step function.

For $t \in \mathbb{N}$, let $D_t = \cap_{\ell=1}^{L} (d_\ell - 2^{-t}, d_\ell + 2^{-t})^C$, $B_t = [-t, t] \cap D_t$ and

$$\tau_t(x) = \begin{cases} g(x) & \text{if } x \in B_t \\ 0 & \text{otherwise} \end{cases}.$$

Since $g(x)$ is continuous and B_t is a compact set, the image of τ_t is bounded, say $\tau_t(x) \in [L_t, U_t]$. By the Heine-Cantor Theorem, τ_t is uniformly continuous on B_t, i.e. $\forall \epsilon_{\tau,t} > 0$, $\exists \delta_{\tau,t} > 0$ such that

$$\forall x_1, x_2 \in B_t, \quad |x_1 - x_2| \le \delta_{\tau,t} \Longrightarrow |\tau_t(x_1) - \tau_t(x_2)| \le \epsilon_{\tau,t}.$$

Let $\epsilon_{\tau,t} = 2^{-t}$ and choose $0 < \delta_{\tau,t} < 2^{-t}$ accordingly. For $s = 1, \ldots, S_t$, where $S_t = \lceil 2t/\delta_{\tau,t} \rceil$, let

$$A_{st} = [-t + (s-1)\delta_{\tau,t},\ -t + s\delta_{\tau,t}) \cap B_t.$$

In the rare case where $2t/\delta_{\tau,t}$ is an integer, we let $A_{S_t,t} = [2t - \delta_{\tau,t}, 2t]$. The sets A_{st} form a partition of the compact set B_t. Note that the choice of D_t and $\delta_{\tau,t}$ ensures that A_{st} are connected, with the harmless exception of $A_{S_t,t}$ which could sometimes consist of two singletons. Define h_t by

$$h_t(x) = \sum_{s=1}^{S_t} b_{st} \mathbb{1}_{A_{st}}(x)$$

where

$$b_{st} = \inf_{y \in A_{st}} g(y) \qquad \text{and} \qquad \mathbb{1}_{A_{st}}(x) = \begin{cases} 1 & \text{if } x \in A_{st} \\ 0 & \text{otherwise} \end{cases}.$$

Then, by construction, $\sup_x |\tau_t(x) - h_t(x)| \leq 2^{-t}$ and

$$\left| \int g(x)\, \mathrm{d}\hat{G}_k(x) - \int g(x)\, \mathrm{d}F_1(x) \right| \leq T_1 + T_2 + T_3 + T_4 + T_5 \tag{3.4}$$

where

$$\begin{aligned} T_1 &= \left| \int g(x)\, \mathrm{d}\hat{G}_k(x) - \int \tau_t(x)\, \mathrm{d}\hat{G}_k(x) \right| \\ T_2 &= \left| \int \tau_t(x)\, \mathrm{d}\hat{G}_k(x) - \int h_t(x)\, \mathrm{d}\hat{G}_k(x) \right| \\ T_3 &= \left| \int h_t(x)\, \mathrm{d}\hat{G}_k(x) - \int h_t(x)\, \mathrm{d}F_1(x) \right| \\ T_4 &= \left| \int h_t(x)\, \mathrm{d}F_1(x) - \int \tau_t(x)\, \mathrm{d}F_1(x) \right| \\ T_5 &= \left| \int \tau_t(x)\, \mathrm{d}F_1(x) - \int g(x)\, \mathrm{d}F_1(x) \right|. \end{aligned}$$

We will now prove that for any $\epsilon > 0$ and ω in a subset of Ω with probability 1, we can choose t_ω such that the five terms above are less than $\epsilon/5$ for all $k \geq k_\omega(t_\omega)$.

To begin, note that

$$T_4 = \left| \int h_t(x) - \tau_t(x)\, \mathrm{d}F_1(x) \right| \leq \int |h_t(x) - \tau_t(x)|\, \mathrm{d}F_1(x) \leq 2^{-t}$$

by construction. The same bound applies for T_2 and does not depend on k or ω.

By Theorem 3.5, $\sup_x |\hat{G}_k(x) - F_1(x)|$ converges almost surely to 0. Therefore, $\exists \Omega_0 \subset \Omega$ with $P(\Omega_0) = 1$ such that for each $\omega \in \Omega_0$ and any t, $\exists k_{\omega,t}$ with

$$\sup_x |\hat{G}_k(x) - F_1(x)| < \frac{1}{S_t \max(|U_t|, |L_t|) 2^{t+1}}$$

for all $k \geq k_{\omega,t}$. For any such k and ω, Lemma 3.6 implies that

$$\left| \mathrm{d}\hat{G}_k(A_{st}) - \mathrm{d}F_1(A_{st}) \right| \leq \frac{2}{S_t \max(|U_t|, |L_t|) 2^{t+1}}$$

for any $s = 1, \ldots, S_t$. Developing T_3 yields

$$\begin{aligned} T_3 &= \left| \sum_{s=1}^{S_t} b_{st} \, \mathrm{d}\hat{G}_k(A_{st}) - \sum_{s=1}^{S_t} b_{st} \, \mathrm{d}F_1(A_{st}) \right| \\ &\leq \sum_{s=1}^{S_t} |b_{st}| \cdot \left| \mathrm{d}\hat{G}_k(A_{st}) - \mathrm{d}F_1(A_{st}) \right| \\ &\leq S_t \max(|U_t|, |L_t|) \frac{2}{S_t \max(|U_t|, |L_t|) 2^{t+1}} \\ &= \frac{1}{2^t}. \end{aligned}$$

Therefore, $\exists t_1$ such that $2^{-t} < \epsilon/5$ for all $t \geq t_1$, i.e. T_2, T_3 and T_4 are each bounded by $\epsilon/5$ for any $t \geq t_1$ and $k \geq k_{\omega,t}$.

We can write

$$T_5 = \left| \int g(x) \mathbb{1}_{B_t^c}(x) \, \mathrm{d}F_1(x) \right| \leq \int |g(x)| \mathbb{1}_{B_t^c}(x) \, \mathrm{d}F_1(x) \to 0$$

as $t \to \infty$ since the integrand goes to 0 for each $x \in \mathbb{R} \backslash \{d_1, \ldots, d_L\}$ by the dominated convergence theorem with bounding function $|g(x)|$. The integrand does not converge to 0 on $\{d_1, \ldots, d_L\}$, but that set has measure 0. Therefore, there exists t_2 such that $T_5 < \epsilon/5$ for all $t \geq t_2$.

Turning now to T_1, we denote by $I \subset \{1, \ldots, m\}$ the indices corresponding to the populations for which $n_{ik} \to \infty$ as $k \to \infty$. By the strong law of large numbers, for any fixed t, there exists $\Omega_{i,t} \subset \Omega$ with $P(\Omega_{i,t}) = 1$ such that for all $\omega \in \Omega_{i,t}$,

$$\sum_{j=1}^{n_{ik}} |g\{X_{ij}(\omega)\}| \mathbb{1}_{B_t^c}\{X_{ij}(\omega)\} \qquad \text{converges to} \qquad \int |g(x)| \mathbb{1}_{B_t^c}(x) \, \mathrm{d}F_i(x)$$

as $k \to \infty$. Consider a fixed

$$\omega \in \Omega_1 = \{\omega | X_{ij}(\omega) = d_\ell \text{ for some } i, j, \ell\}^C \cap \left\{ \bigcap_{i \in I, t \in \mathbb{N}} \Omega_{i,t} \right\}.$$

The intersection is over a countable number of sets of probability 1, hence $P(\Omega_1) = 1$. For any such $\omega \in \Omega_1$, T_1 is developed as

$$\begin{aligned} T_1 &= \left| \int g(x) \mathbb{1}_{B_t^c}(x) \, \mathrm{d}\hat{G}_k(x) \right| \leq \int |g(x)| \mathbb{1}_{B_t^c}(x) \, \mathrm{d}\hat{G}_k(x) \\ &= \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{n_{ik}} \sum_{j=1}^{n_{ik}} |g\{X_{ij}(\omega)\}| \, \mathbb{1}_{B_t^c}\{X_{ij}(\omega)\} \\ &\leq \sum_{i=1}^{m} \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} |g\{X_{ij}(\omega)\}| \, \mathbb{1}_{B_t^c}\{X_{ij}(\omega)\}. \end{aligned}$$

Since ω is fixed, $\exists t_\omega^*$ such that $\mathbb{1}_{B_t^c}\{X_{ij}(\omega)\} \equiv 0$, $\forall i \in I^C$, $j = 1, \ldots, n_{iM_i}$, $t \geq t_\omega^*$. For $i \in I$, the dominated convergence theorem says that there exists t_i^* such that

$$\int |g(x)| \mathbb{1}_{B_t^c}(x) \, \mathrm{d}F_i(x) < \frac{\epsilon}{10m}$$

for all $t \geq t_i^*$. Choose $t \geq t_3 = \max_{i \in I} t_i^*$. Since $\omega \in \Omega_1$, $\exists k_{i,t,\omega}$ such that for all $k \geq k_{i,t,\omega}$,

$$\frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} |g\{X_{ij}(\omega)\}| \mathbb{1}_{B_t^c}\{X_{ij}(\omega)\} \leq \int |g(x)| \mathbb{1}_{B_t^c}(x) \, \mathrm{d}F_i(x) + \frac{\epsilon}{10m} \leq \frac{\epsilon}{5m}.$$

Therefore, $\forall t \geq \max(t_3, t_\omega^*)$, there exists $k_{\omega,t}^* = \max_{i \in I} k_{i,t,\omega}$ such that

$$T_1 = \left| \int g(x) \, \mathrm{d}\hat{G}_k(x) - \int \tau_t(x) \, \mathrm{d}\hat{G}_k(x) \right| \leq \frac{\epsilon}{5}$$

for all $k \geq k_{\omega,t}^*$.

In conclusion, for any $\omega \in \Omega_0 \cap \Omega_1$ and any $\epsilon > 0$, we can choose $t_\omega = \max(t_1, t_2, t_3, t_\omega^*)$

that yields inequalities showing that

$$\left| \int g(x)\, \mathrm{d}\hat{G}_k(x) - \int g(x)\, \mathrm{d}F_1(x) \right| \leq \epsilon$$

for all $k \geq k_\omega(t_\omega) = \max(k_{\omega,t_\omega}, k^*_{\omega,t_\omega})$. In other words, the left hand side of Expression (3.4) converges to 0 for any $\omega \in \Omega_0 \cap \Omega_1$ with $P(\Omega_0 \cap \Omega_1) = 1$, i.e. that expression converges almost surely to 0. ■

Corollary 3.2 (Weighted Strong Law of Large Numbers) *Let X_i denote a variable with distribution F_i. Suppose $E|X_i| < \infty$ for $i = 1, \ldots, m$, then*

$$\sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{n_{ik}} \sum_{j=1}^{n_{ik}} X_{ij}(\omega) \to E(X_1)$$

almost surely as $k \to \infty$.

Proof of Corollary 3.2. Use Theorem 5 with $g(x) = x$. ■

Theorem 3.6 is useful for proving the consistency of the MWLE by extending the proof of Wald (1949). This extension is given next.

3.7 Strong Consistency of the Maximum Weighted Likelihood Estimate

In this section, we adapt the work of Wald (1949) to prove that the MWLE obtained with MAMSE weights is a strongly consistent estimate. For easier reference to his original work, the numbering of Wald is noted with the prefix W.

Wald's Assumptions

The assumptions of Wald (1949) are reproduced below and adapted as required to extend his proof to the MWLE.

Let $F(x|\theta)$ be a parametric family of distributions where θ is an element of Θ, a closed subset of a finite dimensional Cartesian space. We assume that $\exists\theta_0 \in \Theta$ such that $F(x|\theta_0) \equiv F_1(x)$. Wald (1949) does not assume that $F(x|\theta_0)$ is continuous in x, but we do and denote its corresponding density function by $f(x|\theta_0)$.

The following notation is used by Wald (1949):

$$\forall\theta \in \Theta, \rho > 0, \quad f(x,\theta,\rho) = \sup_{|\theta-\theta'|\leq\rho} f(x|\theta'), \quad f^*(x,\theta,\rho) = \max\{f(x,\theta,\rho), 1\},$$
$$\forall r > 0, \qquad \phi(x,r) = \sup_{|\theta|>r} f(x|\theta), \qquad \phi^*(x,r) = \max\{\phi(x,r), 1\}.$$

Assumption 3.1 (W1) *For all $\theta \in \Theta$, $F(x|\theta)$ is absolutely continuous for all x. Therefore, $F(x|\theta)$ admits a density function $f(x|\theta)$.*

Assumption 3.2 (W2) *For sufficiently small ρ and sufficiently large r, the expected values $\int \log f^*(x,\theta,\rho)\,\mathrm{d}F_1(x)$ and $\int \log \phi^*(x,r)\,\mathrm{d}F_1(x)$ are finite.*

Assumption 3.3 (W3) *If $\lim_{i\to\infty}\theta_i = \theta$, then $\lim_{i\to\infty} f(x|\theta_i) = f(x|\theta)$.*

Assumption 3.4 (W4) *If $\theta_1 \neq \theta_0$, then $F(x|\theta_0) \neq F(x|\theta_1)$ for at least one x.*

Assumption 3.5 (W5) *If $\lim_{i\to\infty}|\theta_i| = \infty$, then $\lim_{i\to\infty} f(x|\theta_i) = 0$.*

Assumption 3.6 (W6) *$\int |\log f(x|\theta_0)|\,\mathrm{d}F_i(x) < \infty$ for $i = 1,\ldots,m$.*

Assumption 3.7 (W7) *The parameter space Θ is a closed subset of a finite-dimensional Cartesian space.*

Assumption 3.8 (W8) *The functions $f(x,\theta,\rho)$ and $\phi(x,r)$ are measurable for any θ, ρ and r.*

Assumption 3.9 *The functions $f(x|\theta_0)$, $f(x,\theta,\rho)$ and $\phi(x,r)$ are continuous except possibly on a finite set of points $\{d_1,\ldots,d_L\}$. The set of discontinuities may depend on θ, ρ or r, but must be finite for any fixed values of these parameters.*

Assumptions W1 to W8 are from Wald (1949); only Assumption W6 is modified to cover the m populations of our paradigm. Assumption 3.9 is required to ensure that Theorem 3.6 applies. Note that these assumptions are mostly concerned with the family of distributions $F(x|\theta)$ (the model), rather than with the true distribution of the data.

Lemmas 3.10 and 3.11 are useful for determining if the family of distributions $F(x|\theta)$ satisfies Assumption 3.9.

Wald's Lemmas

Wald's lemmas do not need to be modified. We state them for completeness, but do not reproduce the proofs provided in Wald (1949).

For expectations, the following convention is adopted. Let U be a random variable. The expected value of U exists if $\mathrm{E}\{\max(U,0)\} < \infty$. If $E\{\max(U,0)\}$ is finite but $E\{\min(U,0)\}$ is not, we say that $E\{\min(U,0)\} = -\infty$.

Let a generic X represent a random variable with distribution $F_1(x) \equiv F(x|\theta_0)$.

Lemma 3.7 (W1) *For any $\theta \neq \theta_0$, we have* $\mathrm{E}\log f(X|\theta) < \mathrm{E}\log f(X|\theta_0)$.

Lemma 3.8 (W2) $\lim_{\rho\to 0} \mathrm{E}\log f(X,\theta,\rho) = \mathrm{E}\log f(X|\theta)$.

Lemma 3.9 (W3) *The equation* $\lim_{r\to\infty} \mathrm{E}\log\phi(X,r) = -\infty$ *holds.*

The next two lemmas are useful in determining if Assumption 3.9 is satisfied.

Lemma 3.10 *Let $f(x|\theta)$ be continuous for all $\theta \in \Theta$ and $x \in N_{x_1}$, a neighborhood of x_1. Then for θ_0 and ρ fixed, $f(x,\theta_0,\rho)$ is continuous at x_1.*

Proof of Lemma 3.10. Suppose that $f(x, \theta_0, \rho)$ has a discontinuity at $x = x_1$. Then, there exists $\epsilon > 0$ such that for all $\delta > 0$, there exists x_2 with $|x_1 - x_2| < \delta$ but

$$|f(x_1, \theta_0, \rho) - f(x_2, \theta_0, \rho)| > \epsilon. \tag{3.5}$$

Let $A \subset N_{x_1}$ be a compact set around x_1. Let $B = \{\theta : |\theta - \theta_0| \leq \rho\}$. The set $A \times B$ is compact and hence $f(x|\theta)$ is uniformly continuous on that domain by Heine-Borel. Therefore, for the ϵ chosen above, there exists a $\delta_1 > 0$ such that $x_1, x_2 \in A$ and $|x_1 - x_2| < \delta_1$ imply

$$|f(x_1|\theta) - f(x_2|\theta)| < \epsilon/2 \tag{3.6}$$

for all $\theta \in B$. Choose such an x_2 and define

$$\theta_1 = \arg \max_{|\theta - \theta_0| \leq \rho} f(x_1|\theta) \qquad \text{and} \qquad \theta_2 = \arg \max_{|\theta - \theta_0| \leq \rho} f(x_2|\theta).$$

The maxima are attained since $A \times B$ is compact and $f(x|\theta)$ continuous in θ. Therefore,

$$f(x_1, \theta_0, \rho) = f(x_1|\theta_1) \quad \text{and} \quad f(x_2, \theta_0, \rho) = f(x_2|\theta_2). \tag{3.7}$$

Consider the following two cases.

Case 1: $f(x_1|\theta_1) > f(x_2|\theta_2)$

By Equations (3.5) and (3.7), $f(x_1|\theta_1) \geq f(x_2|\theta_2) + \epsilon$. Furthermore, inequality (3.6) implies that

$$f(x_2|\theta_1) > f(x_1|\theta_1) - \frac{\epsilon}{2} \geq f(x_2|\theta_2) + \frac{\epsilon}{2},$$

a contradiction with the definition of θ_2.

Case 2: $f(x_1|\theta_1) < f(x_2|\theta_2)$

By Equations (3.5) and (3.7), $f(x_1|\theta_1) \leq f(x_2|\theta_2) - \epsilon$. Inequality (3.6) yields

$$f(x_1|\theta_2) > f(x_2|\theta_2) - \frac{\epsilon}{2} \geq f(x_1|\theta_1) + \frac{\epsilon}{2},$$

a contradiction with the definition of θ_1.

Therefore, we conclude that $f(x, \theta_0, \rho)$ is continuous at x_1. ■

By Lemma 3.10, if $f(x|\theta)$ is continuous in x and θ, then $f(x, \theta_0, \rho)$ is continuous in x for any fixed θ_0 and ρ.

Before introducing Lemma 3.11, define

$$\omega_g(\delta, x_0) = \sup_{|x - x_0| < \delta} |g(x) - g(x_0)|,$$

the modulus of continuity of the function $g(x)$ around x_0. Note that when it exists,

$$\lim_{\delta \to 0} \frac{\omega_g(\delta, x_0)}{\delta} = \left| g'(x_0) \right|.$$

Lemma 3.11 *Suppose that $f(x|\theta)$ is continuous in θ and that $\phi(x, r)$ has a discontinuity at x_0. Then, there exists $\epsilon > 0$ such that $\omega_{f(\cdot|\theta)}(\delta, x_0) > \epsilon$ for any $\delta > 0$ and some θ.*

Proof of Lemma 3.11. Fix $r > 0$. Since $\phi(x, r)$ is discontinuous at x_0, there exists $\epsilon > 0$ such that for any $\delta > 0$, $\exists x_1$ such that $|x_0 - x_1| < \delta$ but

$$|\phi(x_0, r) - \phi(x_1, r)| > 2\epsilon. \tag{3.8}$$

For any fixed δ and x_1, consider the following two cases.

Case 1: $\phi(x_0, r) > \phi(x_1, r) + 2\epsilon$.

By the continuity of $f(x|\theta)$, it is possible to choose $|\theta_0| > r$ such that $f(x_0|\theta_0)$ is arbitrarily close to $\phi(x_0, r)$, say less than ϵ apart, i.e.

$$f(x_0|\theta_0) \geq \phi(x_0, r) - \epsilon.$$

For that possibly suboptimal θ_0, $\phi(x_1, r) \geq f(x_1|\theta_0)$, hence

$$f(x_0|\theta_0) \geq \phi(x_0, r) - \epsilon > \phi(x_1, r) \geq f(x_1|\theta_0)$$

meaning that

$$|f(x_0|\theta_0) - f(x_1|\theta_0)| \geq f(x_0|\theta_0) - f(x_1|\theta_0) \geq \phi(x_0, r) - \epsilon - \phi(x_1, r) > \epsilon$$

because of Equation 3.8. Therefore, $\omega_{f(\cdot|\theta_0)}(\delta, x_0) > \epsilon$.

Case 2: $\phi(x_0, r) < \phi(x_1, r) - 2\epsilon$.

The continuity of $f(x|\theta)$ allows us to choose $|\theta_1| > r$ such that $f(x_1|\theta_1)$ is close to $\phi(x_1, r)$, say less than ϵ apart, i.e.

$$f(x_1|\theta_1) \geq \phi(x_1, r) - \epsilon.$$

Then, by the definition of ϕ, we have $\phi(x_0, r) \geq f(x_0|\theta_1)$, hence

$$f(x_1|\theta_1) \geq \phi(x_1, r) - \epsilon > \phi(x_0, r) \geq f(x_0|\theta_1).$$

Therefore,

$$|f(x_1|\theta_1) - f(x_0|\theta_1)| \geq f(x_1|\theta_1) - f(x_0|\theta_1) \geq \phi(x_1, r) - \epsilon - \phi(x_0, r) > \epsilon$$

by Equation 3.8. Therefore, $\omega_{f(\cdot|\theta_1)}(\delta, x_0) > \epsilon$.

By combining both cases, we can conclude that for all $\delta > 0$, there exists a θ such that $\omega_{f(\cdot|\theta)}(\delta, x_0) > \epsilon$ ■

Note that if $g(x)$ is continuous at x_0, $\lim_{\delta \to 0} \omega_g(\delta, x_0) = 0$. Having a modulus of continuity $\omega_{f(\cdot|\theta)}(\delta, x_0)$ bounded below is akin to having an infinite derivative $\partial f(x|\theta)/\partial x$ at x_0. This occurs when:

- there is a discontinuity in the function $f(x|\theta)$ at x_0,
- as $\theta \to \infty$, the slope of $f(x|\theta)$ around x_0 keeps increasing.

Therefore, discontinuities in $\phi(x, r)$ will occur if $f(x|\theta)$ is discontinuous itself, or if $f(x|\theta)$ has a peak that can become arbitrarily steep (i.e. its slope is not bounded for a fixed x as θ varies). The main result of this paper uses Theorem 3.6 which allows a finite number of discontinuities. Therefore, as long as $f(x|\theta)$ is a continuous model such that the set

$$\begin{aligned}\{x : &\text{ for an arbitrarily large } r > 0, \omega_{f(\cdot|\theta)}(\delta, x) \text{ is arbitrarily large for some } |\theta| > r\} \\ &= \{x : f(x|\theta) \text{ has an arbitrarily steep peak at } x \text{ for some } |\theta| > r\}\end{aligned}$$

is empty or made up of a finite number of singletons, Assumption 3.9 will hold.

Note the effect of the constraint $|\theta| > r$. Consider for instance the normal model with unknown mean and variance. As $\sigma^2 \to 0$, the normal density will have an arbitrarily steep peak close to μ. However, $|[\mu, \sigma^2]^\mathsf{T}| > r$ implies that there is a lower bound for σ^2, hence the steepness of the peak is bounded. Assumption 3.9 is satisfied under that model.

Main Result

We now turn to the main results of this section.

Theorem 3.7 (Theorem W1) *Let $\mathcal{T}$ be any closed subset of Θ that does not contain θ_0. Then,*

$$P\left[\lim_{k\to\infty}\frac{\sup_{\theta\in\mathcal{T}}\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}}f\{X_{ij}(\omega)|\theta\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}}{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}}f\{X_{ij}(\omega)|\theta_0\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}}=0\right]=1.$$

Proof of Theorem 3.7. Let X denote a random variable with distribution $F_1(x)\equiv F(x|\theta_0)$ and let r_0 be a positive number chosen such that

$$\mathrm{E}\{\log\phi(X,r_0)\}<\mathrm{E}\{\log f(X|\theta_0)\}. \tag{3.9}$$

The existence of such an r_0 follows from Lemma 3.9 and Assumption 3.6. Then, $\mathcal{T}_1=\{\theta:\theta\le r_0\}\cap\mathcal{T}$ is a compact set since it is a closed and bounded subset of a finite-dimensional Cartesian space. With each element $\theta\in\mathcal{T}_1$, we associate a positive value ρ_θ such that

$$\mathrm{E}\{\log f(X,\theta,\rho_\theta)\}<\mathrm{E}\{\log f(X|\theta_0)\}. \tag{3.10}$$

The existence of such ρ_θ follows from Lemma 3.7 and 3.8. Let $S(\theta,\rho)$ denote the sphere with center θ and radius ρ. The spheres $\{S(\theta,\rho_\theta)\}$ form a covering of the compact $\mathcal{T}_1$, hence there exists a finite sub-covering. Let $\theta_1,\ldots,\theta_h\in\mathcal{T}_1$ such that $\mathcal{T}_1\subset\bigcup_{s=1}^{h}S(\theta_s,\rho_{\theta_s})$. Clearly,

$$\begin{aligned}0 &\le \sup_{\theta\in\mathcal{T}}\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}}f\{X_{ij}(\omega)|\theta\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}\\ &\le \sum_{s=1}^{h}\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}}f\{X_{ij}(\omega),\theta_s,\rho_{\theta_s}\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}} \quad+\quad \prod_{i=1}^{m}\prod_{j=1}^{n_{ik}}\phi\{X_{ij}(\omega),r_0\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}.\end{aligned}$$

Therefore, to prove Theorem 3.7 if suffices to show that

$$P\left[\lim_{k\to\infty}\frac{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} f\{X_{ij}(\omega),\theta_s,\rho_{\theta_s}\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}}{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} f\{X_{ij}(\omega)|\theta_0\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}}=0\right]=1$$

for $s=1,\ldots,h$ and that

$$P\left[\lim_{k\to\infty}\frac{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} \phi\{X_{ij}(\omega),r_0\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}}{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} f\{X_{ij}(\omega)|\theta_0\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}}=0\right]=1.$$

The above equations can be rewritten as

$$\begin{aligned} P\Bigg[\lim_{k\to\infty} n_{1k}\Bigg[\sum_{i=1}^{m}\sum_{j=1}^{n_{ik}}\frac{\lambda_{ik}(\omega)}{n_{ik}}\log f\{X_{ij}(\omega),\theta_s,\rho_{\theta_s}\} \\ -\frac{\lambda_{ik}(\omega)}{n_{ik}}\log f\{X_{ij}(\omega)|\theta_0\}\Bigg]=-\infty\Bigg] \qquad (3.11) \\ = P\Bigg[\lim_{k\to\infty} n_{1k}\Bigg\{\int \log f(x,\theta_s,\rho_{\theta_s})\,\mathrm{d}\hat{G}_k(x) \\ -\int \log f(x|\theta_0)\,\mathrm{d}\hat{G}_k(x)\Bigg\}=-\infty\Bigg]=1 \end{aligned}$$

for $s=1,\ldots,h$ and

$$\begin{aligned} P\Bigg[\lim_{k\to\infty} n_{1k}\Bigg[\sum_{i=1}^{m}\sum_{j=1}^{n_{ik}}\frac{\lambda_{ik}(\omega)}{n_{ik}}\log \phi\{X_{ij}(\omega),r_0\} \\ -\frac{\lambda_{ik}(\omega)}{n_{ik}}\log f\{X_{ij}(\omega)|\theta_0\}\Bigg]=-\infty\Bigg] \qquad (3.12) \end{aligned}$$

$$= P\Bigg[\lim_{k\to\infty} n_{1k}\Bigg\{\int \log\phi(x,r_0)\,\mathrm{d}\hat{G}_k(x) - \int \log f(x|\theta_0)\,\mathrm{d}\hat{G}_k(x)\Bigg\} = -\infty\Bigg] = 1$$

respectively. Assumptions 3.6 and 3.9 insure that Theorem 3.6 applies to the integrals above, each of these converging almost surely to

$$\int \log f(x,\theta_s,\rho_{\theta_s})\,\mathrm{d}F_1(x),\ \int \log\phi(x,r_0)\,\mathrm{d}F_1(x)\ \text{ or }\ \int \log f(x|\theta_0)\,\mathrm{d}F_1(x).$$

Combining this result with Equations (3.10) and (3.9), we have that (3.11) and (3.12) hold. Hence the proof of Theorem 3.7 is complete. ■

Theorem 3.8 (Theorem W2) *Let $\hat{\theta}_k(\omega)$ be a sequence of random variables such that there exists a positive constant c with*

$$\frac{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} f\{X_{ij}(\omega)|\hat{\theta}_k(\omega)\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}}{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} f\{X_{ij}(\omega)|\theta_0\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}} \geq c > 0 \tag{3.13}$$

for all $k \in \mathbb{N}$ and all $\omega \in \Omega$. Then

$$P\left\{\lim_{k\to\infty}\hat{\theta}_k(\omega) = \theta_0\right\} = 1.$$

Proof of Theorem 3.8. Let $\epsilon > 0$ and consider the values of $\hat{\theta}_k(\omega)$ as k goes to infinity. Suppose that θ_ℓ is a limit point away from θ_0, such that $|\theta_\ell - \theta_0| > \epsilon$. Then,

$$\frac{\sup_{|\theta-\theta_0|\geq\epsilon}\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} f\{X_{ij}(\omega)|\theta\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}}{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} f\{X_{ij}(\omega)|\theta_0\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}} \geq c > 0$$

infinitely often. By Theorem 3.7, this event has probability 0 even with ϵ arbitrarily small. Therefore,

$$P\left\{\omega : \left|\lim_{k\to\infty} \hat{\theta}_k(\omega) - \theta_0\right| \leq \epsilon\right\} = 1$$

for all $\epsilon > 0$. ■

Corollary 3.3 (Corollary W1) *The MWLE is a strongly consistent estimate of* θ_0.

Proof of Corollary 3.3. The MWLE clearly satisfies Equation (3.13) with $c = 1$ because $\hat{\theta}_k(\omega)$ is chosen to maximize the numerator of (3.13). ■

3.8 Asymptotic Behavior of the MAMSE Weights

We study the asymptotic behavior of the MAMSE weights as $k \to \infty$ and its consequences in constructing a weighted central limit theorem. Let

$$\mathcal{L} = \left\{\boldsymbol{\lambda} : \sum_{i=1}^{m} \lambda_i F_i(x) \equiv F_1(x)\right\}$$

where $\equiv$ indicates that the functions are equal for all x. Clearly, $\mathcal{L}$ is a nonempty convex set with $[1, 0, \ldots, 0]^{\mathsf{T}} \in \mathcal{L}$. Moreover, if we consider the elements of $\mathcal{L}$ as elements of the normed space $([0,1]^m, ||\cdot||)$ where $||\cdot||$ stands for the Euclidean norm, then $\mathcal{L}^C$ is an open set.

We will show that for $k \in \mathbb{N}$, all accumulation points of the MAMSE weights will be in the set $\mathcal{L}$. In other words, the MAMSE weights can only converge to vectors that define a mixture distribution identical to the target distribution.

Theorem 3.9 *Suppose that* $\mathcal{L}^C \neq \emptyset$ *and let* $\boldsymbol{\lambda}^* \in \mathcal{L}^C$, *then for any* $\epsilon > 0$, *there exists a set* Ω_0 *of probability 1 such that*

$$||\boldsymbol{\lambda}^* - \boldsymbol{\lambda}_k(\omega)|| > \epsilon \quad i.o.$$

for all $\omega \in \Omega_0$ *and hence the MAMSE weights do not converge to* $\boldsymbol{\lambda}^*$ *as* $k \to \infty$.

Proof of Theorem 3.9. The Glivenko-Cantelli lemma shows that $\sup_x |\hat{F}_{ik}(x) - F_i(x)| \to 0$ almost surely as $k \to \infty$. Let Ω_i be the set of probability 1 where the convergence occurs.

The summand and the integrand of the following expressions are bounded by 1, thus

$$\frac{1}{n_{1k}} \sum_{j=1}^{n_{1k}} \left[\sum_{i=1}^{m} \lambda_i F_i\{X_{1j}(\omega)\} - F_1\{X_{1j}(\omega)\} \right]^2 \to \mathrm{E}\left[\left\{ \sum_{i=1}^{m} \lambda_i F_i(X_{11}) - F_i(X_{11}) \right\}^2 \right]$$

almost surely as $k \to \infty$ by the strong law of large numbers. Let Ω' be the set of probability 1 on which the convergence occurs. Note that the expectation in the expression above is taken over the random variable X_{11} which follows distribution F_1.

Consider now the set $\Omega_0 = \Omega' \cap \bigcap_{i=1}^{m} \Omega_i$ and let $\omega \in \Omega_0$ be any fixed element of that set. Note that by construction $P(\Omega_0) = 1$.

Let $B(\mathbf{x}, r)$ denote the open ball of radius r centered at $\mathbf{x}$. Since $\mathcal{L}^C$ is an open set, any small enough $\epsilon > 0$ will be such that $B(\boldsymbol{\lambda}^*, \epsilon) \cap \mathcal{L} = \emptyset$. Then, consider $P_k(\boldsymbol{\lambda})$ as defined in Equation (3.1) and for any $\boldsymbol{\lambda} \in B(\boldsymbol{\lambda}^*, \epsilon)$,

$$\begin{aligned}
P_k(\boldsymbol{\lambda}) &\geq \int \left| \sum_{i=1}^{m} \lambda_i \hat{F}_{ik}(x) - \hat{F}_{1k}(x) \right|^2 \mathrm{d}\hat{F}_{1k}(x) \\
&\geq \int \left\{ \left| \sum_{i=1}^{m} \lambda_i \hat{F}_{ik}(x) - F_1(x) \right|^2 - \left| F_1(x) - \hat{F}_{1k}(x) \right|^2 \right\} \mathrm{d}\hat{F}_{1k}(x) \\
&\geq \int \left\{ \left| \sum_{i=1}^{m} \lambda_i F_i(x) - F_1(x) \right|^2 \right. \\
&\qquad \left. - \left| \sum_{i=1}^{m} \lambda_i \hat{F}_{ik}(x) - \sum_{i=1}^{m} \lambda_i F_i(x) \right|^2 - \left| F_1(x) - \hat{F}_{1k}(x) \right|^2 \right\} \mathrm{d}\hat{F}_{1k}(x) \\
&\geq \frac{1}{n_{1k}} \sum_{j=1}^{n_{1k}} \left[\sum_{i=1}^{m} \lambda_i F_i\{X_{1j}(\omega)\} - F_1\{X_{1j}(\omega)\} \right]^2 \\
&\qquad - \int \left\{ \sum_{i=1}^{m} \lambda_i^2 \left| \hat{F}_{ik}(x) - F_i(x) \right|^2 + \left| \hat{F}_{1k}(x) - F_1(x) \right|^2 \right\} \mathrm{d}\hat{F}_{1k}(x)
\end{aligned}$$

$$\begin{aligned}
&\geq \frac{1}{n_{1k}} \sum_{j=1}^{n_{1k}} \left[\sum_{i=1}^{m} \lambda_i F_i\{X_{1j}(\omega)\} - F_1\{X_{1j}(\omega)\} \right]^2 \\
&\qquad - \sum_{i=1}^{m} \lambda_i^2 \sup_x \left| \hat{F}_{ik}(x) - F_i(x) \right|^2 - \sup_x \left| \hat{F}_{1k}(x) - F_1(x) \right|^2 \\
&\geq \frac{1}{2} \mathrm{E} \left[\left\{ \sum_{i=1}^{m} \lambda_i F_i(X_{11}) - F_i(X_{11}) \right\}^2 \right] = K > 0
\end{aligned}$$

for a large enough k.

The fact that $\boldsymbol{\lambda} \in \mathcal{L}^C$ implies that $\sum_{i=1}^{m} \lambda_i F_i(x) \neq F_1(x)$ for some x where $F_1(x)$ is not flat, i.e. some x with positive probability, thus

$$\mathrm{E} \left[\left\{ \sum_{i=1}^{m} \lambda_i F_i(X_{11}) - F_1(X_{11}) \right\}^2 \right] > 0.$$

Therefore, there exist $k_0(\omega)$ and $K > 0$ such that $P_k(\boldsymbol{\lambda}) > K$ for all $k \geq k_0(\omega)$. However, Lemma 3.2 shows that $P_k\{\boldsymbol{\lambda}_k(\omega)\} \to 0$ as $k \to \infty$. Therefore, $\boldsymbol{\lambda}_k(\omega) \in B(\boldsymbol{\lambda}^*, \epsilon)$ at most finitely many times. This is true of any $\boldsymbol{\lambda}^* \in \mathcal{L}^C$, meaning that for all $\omega \in \Omega_0$, $||\boldsymbol{\lambda}^* - \boldsymbol{\lambda}_k(\omega)|| > \epsilon$ at most finitely many times. ■

Corollary 3.4 *Consider the sequence of MAMSE weights $\boldsymbol{\lambda}_k(\omega)$ for ω fixed and $k \in \mathbb{N}$. Let $\boldsymbol{\lambda}$ be an accumulation point of the sequence $\boldsymbol{\lambda}_k(\omega)$, then $\boldsymbol{\lambda} \in \mathcal{L}$.*

Proof of Corollary 3.4. By Theorem 3.9, the neighborhood of any $\boldsymbol{\lambda} \in \mathcal{L}^C$ can be visited at most finitely many times. Hence, any accumulation point belongs to $\mathcal{L}$. ■

Corollary 3.5 *If $\mathcal{L}$ is a singleton, then $\mathcal{L} = \{[1, 0, \ldots, 0]^\mathsf{T}\}$ and*

$$\boldsymbol{\lambda}_k(\omega) \to [1, 0, \ldots, 0]^\mathsf{T}$$

almost surely as $k \to \infty$.

Proof of Corollary 3.5. The vector $[1, 0, \ldots, 0]^\mathsf{T}$ is always in $\mathcal{L}$. Therefore, $\mathcal{L}$ will be a

singleton only when $\mathcal{L} = \{[1, 0, \ldots, 0]^\mathsf{T}\}$. Let $\epsilon > 0$ and let

$$\mathcal{A} = [0, 1]^m \backslash B([1, 0, \ldots, 0]^\mathsf{T}, \epsilon)$$

where $B(\mathbf{x}, r)$ denote the open ball of radius r centered at $\mathbf{x}$. The set $\mathcal{A}$ is closed and bounded thus compact.

Let $\bar{B}(\mathbf{x}, r)$ be the closed ball of radius r centered at $\mathbf{x}$. Consider the sets $\bar{B}(\mathbf{x}, \epsilon/2)$ for $\mathbf{x} \in [0, 1]^m$; they form a covering of $\mathcal{A}$. Since $\mathcal{A}$ is a compact set, there exist a finite sub-covering with balls centered at $\mathbf{x}_s$ for $s = 1, \ldots, S$.

Consider now the sequence of MAMSE weights $\boldsymbol{\lambda}_k(\omega)$. By Theorem 3.9, for every fixed $\omega \in \Omega_1$ with $P(\Omega_1) = 1$, any of the balls $\bar{B}(\mathbf{x}_s, \epsilon/2)$ will contain at most finitely many $\boldsymbol{\lambda}_k(\omega)$, i.e.

$$\boldsymbol{\lambda}_k(\omega) \in \bigcup_{s=1}^{S} \bar{B}(\mathbf{x}_s, \epsilon/2) \qquad \text{finitely many times.} \tag{3.14}$$

Consequently,

$$\boldsymbol{\lambda}_k(\omega) \in \left\{ \bigcup_{s=1}^{S} \bar{B}(\mathbf{x}_s, \epsilon/2) \right\}^C \subset B([1, 0, \ldots, 0]^\mathsf{T}, \epsilon) \qquad i.o. \tag{3.15}$$

Expressions (3.14) and (3.15) imply that if it exists, the limit of $\boldsymbol{\lambda}_k(\omega)$ is in the set $B([1, 0, \ldots, 0]^\mathsf{T}, \epsilon)$. Since ϵ can be arbitrarily small and since the space is complete, we conclude that $\boldsymbol{\lambda}_k(\omega) \to [1, 0, \ldots, 0]^\mathsf{T}$ almost surely. ■

In the case where $\mathcal{L}$ is not a singleton, the MAMSE weights do not seem to converge to any particular point. Corollary 3.4 indicates that any accumulation point will be in $\mathcal{L}$, but it seems that the neighborhood of many points in $\mathcal{L}$ is visited infinitely often.

Describing the limit of the MAMSE weights in a general case seems rather tedious. In particular, the speeds at which the sample sizes increase in each population as well as the shape of each $\hat{F}_{ik}(x)$ compared to $\hat{F}_{1k}(x)$ will have an influence on the behavior of the

weights. The precise description thereof is left to future work.

3.9 Issues for Asymptotic Normality

In all simulations performed, we have not found any evidence showing that a MAMSE-weighted sum of random variables is not asymptotically normal. A formal proof showing the asymptotic distribution of such a sum remains however to be found.

Even in the realistic case where no mixture of the populations are exactly identical to the population of interest, i.e. $\mathcal{L} = \{[1, 0, \ldots, 0]^\mathsf{T}\}$, we cannot develop a central limit theorem for a MAMSE-weighted sum of variables without studying the speed of convergence of the MAMSE weights.

Let $\mu_i = \mathrm{E}(X_{i1})$ and $\sigma_i^2 = \mathrm{var}(X_{i1})$. Under reasonable assumptions, one could hope to show that the expression

$$\sqrt{n_{1k}}\, B_k = \sqrt{n_{1k}} \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{n_{ik}} \sum_{j=1}^{n_{ik}} (X_{ij} - \mu_1) \tag{3.16}$$

converges weakly to a Normal variable with mean 0.

To simplify things, suppose that the sample sizes from all populations increase at the same rate, that is $[n_{1k}, n_{2k}, \ldots, n_{mk}]/n_{1k} \to [\alpha_1, \ldots, \alpha_m]$ as $k \to \infty$ with $0 < \alpha_i < \infty$.

Note that

$$B_{ik} = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} (X_{ij} - \mu_i)$$

converges almost surely to 0 by the strong law of large numbers and that $\sqrt{n_{1k}} B_{ik}$ converges weakly to a Normal distribution with mean 0 and variance σ_i^2/α_i by the central limit theorem. However, developing B_k yields

$$B_k = \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{n_{ik}} \sum_{j=1}^{n_{ik}} (X_{ij} - \mu_1) \quad = \quad \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{n_{ik}} \sum_{j=1}^{n_{ik}} (X_{ij} - \mu_i) + \sum_{i=1}^{m} \lambda_{ik}(\omega)(\mu_i - \mu_1)$$

$$= \sum_{i=1}^{m} \lambda_{ik}(\omega) B_{ik} + \sum_{i=1}^{m} \lambda_{ik}(\omega)(\mu_i - \mu_1).$$

Recall that $B_{ik} \to 0$ and that Corollary 3.2 shows that $B_k \to 0$. Consequently, the expression $\sum_{i=1}^{m} \lambda_{ik}(\omega)\,(\mu_i - \mu_1)$ converges to 0 as well, but we do not know the rate of that convergence. If that rate is slower than $1/\sqrt{n_{1k}}$, Expression (3.16) will not converge to a Normal distribution even if the MAMSE weights converge almost surely to a fixed value.

To prove the asymptotic normality of (3.16) using the strategy sketched above, we need to show that the speed of convergence of the MAMSE weights is fast enough. This condition corresponds to the second half of Assumption 2.5 of Wang, van Eeden and Zidek (2004) who require the same rate of convergence, $1/\sqrt{n_{1k}}$, for the weights.

Note that the classical proof of asymptotic normality uses the moment generating function or the characteristic function, but it does not apply here because each datum is multiplied by a data-based weight. As a consequence, the terms $\lambda_{ik}(\omega)X_{ij}$ are not independent.

The study of the asymptotic distribution of expressions similar to (3.16) are left to future work. Any significant advances will probably come only after a thorough investigation of the speed of convergence of the MAMSE weights.

3.10 Simulations

In this section, the finite-sample performance of the MWLE with MAMSE weights is evaluated through simulations. Different cases of interest are considered.

The number of repetitions for each simulation study varies from 10000 to 40000. We used the bootstrap on a pilot simulation to evaluate the variability of the values presented throughout this section. Unless otherwise stated, the standard deviation of the error due to simulation is less than one unit of the last digit shown.

3.10.1 Two Normal Distributions

We first explore the merits of our weights for the ubiquitous Normal distribution. Samples of equal sizes n are drawn from

$$\text{Pop. } 1 : \mathcal{N}(0,1), \quad \text{Pop. } 2 : \mathcal{N}(\Delta,1)$$

for different values of Δ, each scenario being repeated 10000 times. Table 3.1 shows the average MAMSE weights under different circumstances.

	Average Values of $100\lambda_1$									
	$n=5$	10	15	20	25	50	100	200	1000	10000
$\Delta = 0$	72	71	72	71	71	72	72	72	72	72
0.001	72	71	71	72	72	72	72	71	72	72
0.01	72	72	71	72	72	72	72	72	72	74
0.10	72	72	73	73	73	73	74	76	86	98
0.25	74	74	75	76	76	79	83	88	97	100
0.50	77	79	80	82	83	88	93	96	99	100
0.75	80	83	86	88	89	94	97	98	100	100
1.00	84	87	90	92	93	96	98	99	100	100
1.50	89	92	94	95	96	98	99	99	100	100
2.00	93	94	96	97	97	99	99	100	100	100

Table 3.1: Average MAMSE weights for Population 1 when equal samples of size n are drawn from Normal distributions with unit variance and means 0 and Δ respectively. The results are averages over 10000 replicates.

From Table 3.1, we notice that the average weight of Population 1 does not seem to go below 0.7 for these scenarios. As n increases, the weight of Population 1 approaches 1, hence the MAMSE weights detect that the distributions are different and ultimately discard Population 2. Note that this convergence to 1 does not seem to occur for $\Delta = 0$ and seems very slow when Δ is tiny. The average weight for Population 1 increases as well when the discrepancy between the populations increases while n is kept fixed.

Table 3.2 shows the performance obtained for the MWLE with MAMSE weights when compared to the MLE. The ratio of the mean squared errors, 100 MSE(MLE)/MSE(MWLE)

is shown; a value greater than 100 meaning that the MWLE is preferable. This ratio is akin to the relative efficiency of the MLE with respect to the MWLE.

	Efficiency of the MWLE									
	$n = 5$	10	15	20	25	50	100	200	1000	10000
$\Delta = 0$	146	145	144	144	143	143	144	144	144	143
0.001	147	146	145	144	143	143	142	143	143	144
0.01	146	146	145	144	143	143	144	143	141	127
0.10	143	143	142	140	139	135	128	118	89	94
0.25	139	134	131	125	123	110	96	87	91	99
0.50	127	117	108	104	97	88	88	90	97	100
0.75	114	103	95	91	89	87	91	95	99	100
1.00	103	94	90	88	88	90	94	97	99	100
1.50	89	88	89	91	91	94	98	98	100	100
2.00	84	87	91	92	93	96	98	99	100	100

Table 3.2: Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE). Samples of equal size n are simulated from Normal distributions with unit variance and means 0 and Δ respectively. The results are averaged over 10000 replicates.

The MWLE performs better than the MLE for small n and Δ. When n and Δ increase, the methods' performances are eventually equivalent. For the cases in between however, the MLE is a better choice than the MWLE. Fortunately, the loss (at most 16%) seems to be smaller than the potential gain (up to 47%). When the two populations are identical, a steady improvement of about 43% is observed. Note that we cannot expect to improve uniformly over the MLE since the mean is an admissible estimator.

The weighted likelihood could be especially useful in situations where a large population is available to support a few observations from the population of interest. For the next simulation, 40000 replicates of each scenario are produced with the same Normal distributions as before, but with samples of size n and $10n$ for Population 1 and 2 respectively. Table 3.3 shows the average weight allocated to Population 1; Table 3.4 shows the relative efficiency of the methods as measured by 100 MSE(MLE)/MSE(MWLE).

The general behavior of the weights is similar to that in the previous simulation, except that their minimal average value is below 0.5 this time around. As a consequence of its

	Average Values of $100\lambda_1$							
	$n = 5$	10	15	20	25	50	100	200
$\Delta = 0$	51	50	49	49	49	49	49	48
0.001	51	50	49	49	49	49	49	48
0.01	52	50	50	49	49	49	49	49
0.10	54	53	52	53	53	54	57	62
0.25	58	59	60	61	62	69	78	86
0.50	66	70	73	76	79	87	93	96
0.75	74	79	83	86	88	94	97	98
1.00	80	86	89	91	93	96	98	99
1.50	87	92	94	95	96	98	99	99
2.00	91	94	96	97	97	99	99	100

Table 3.3: Average MAMSE weights for Population 1 when samples of size n and $10n$ are drawn from Normal distributions with unit variance and means 0 and Δ respectively. The results are averages over 40000 replicates.

larger size, the sample from Population 2 gets a heavier weight.

It appears that a larger Population 2 magnifies the gains or losses observed previously. Fortunately however, the magnitude of the further improvements seem to exceed that of the extra losses.

Note that the MAMSE weights are invariant to a common transformation of the data in all populations. Therefore, simulation results would be identical (less simulation error) for Normal populations with variance σ^2 and with means 0 and $\Delta\sigma$ respectively.

Overall, the MWLE works very well under the suggested scenarios.

3.10.2 Complementary Populations

We explained in Section 2.2 how the likelihood weights can be seen as mixing probabilities. Can the MAMSE weights detect and exploit the fact that Population 1 has the same distribution as a mixture of some of the other populations? Would the quality of the inference then be improved?

	Efficiency of the MWLE							
	$n = 5$	10	15	20	25	50	100	200
$\Delta = 0$	223	223	223	222	222	221	222	221
0.001	223	225	223	221	222	223	221	220
0.01	223	222	222	220	221	221	220	218
0.10	216	209	203	197	191	169	142	113
0.25	187	165	147	135	125	100	83	78
0.50	139	111	97	90	85	79	83	89
0.75	111	91	85	82	82	85	90	94
1.00	98	85	84	83	85	90	94	97
1.50	88	86	88	89	90	94	97	98
2.00	86	89	91	92	93	96	98	99

Table 3.4: Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE). Samples of sizes n and $10n$ are simulated from Normal distributions with unit variance and means 0 and Δ respectively. The results are averaged over 40000 replicates.

Pseudo-random samples of equal sizes n are drawn from the distributions

$$\text{Pop. } 1 : \mathcal{N}(0,1), \quad \text{Pop. } 2 : |\mathcal{N}(0,1)|, \quad \text{Pop. } 3 : -|\mathcal{N}(0,1)|$$

where $|\cdot|$ denotes absolute values. Hence Population 2 has a Half-Normal distribution and Population 3 follows the complementary distribution.

We consider different sample sizes, each scenario being repeated 10000 times. The results are summarized in Table 3.5. The first column shows 100 MSE(MLE)/MSE(MWLE); the other columns show the average MAMSE weights allocated to each of the three populations.

First observe that the combined average MAMSE weight of Population 2 and 3 accounts for at least half of the total weight for all sample sizes. The MAMSE weights thus detect that an equal mixture of Population 2 and 3 shares the same distribution as Population 1. Note also that the relative efficiency is uniformly greater than 100, meaning that the MWLE with MAMSE weights is preferable to the MLE in these situations.

n	Efficiency	$100\bar{\lambda}_1$	$100\bar{\lambda}_2$	$100\bar{\lambda}_3$
5	115	50	19	30
10	121	46	23	30
15	120	46	25	29
20	118	45	25	29
25	118	45	26	29
50	117	45	27	28
100	116	44	27	28
200	116	44	28	28
1000	115	44	28	28
10000	116	44	28	28

Table 3.5: Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE) and average MAMSE weights allocated to samples of sizes n drawn from $\mathcal{N}(0,1)$, $|\mathcal{N}(0,1)|$ and $-|\mathcal{N}(0,1)|$ respectively. The results are averages over 10000 repetitions.

3.10.3 Negative Weights

In most cases, the unconstrained optimization of $P(\boldsymbol{\lambda})$ yields positive weights. In some cases such as the one that we are going to explore, negative weights systematically occur. Some previous work such as van Eeden & Zidek (2004) showed that allowing negative weights may sometimes boost the performance of the MWLE. We explore the possibility of such improvements here.

Imagine a situation where a measurement of interest is cheaply obtained, but it is costly to determine whether a patient is diseased or not. We want to study the measurement of interest on the diseased patients. Suppose we have two small samples (one diseased, one not) as well as a larger sample where the health status of patients is unknown. If we allow negative values for MAMSE weights, would they adapt by including the larger population in the inference and allocating a negative weight to the small healthy population?

To represent the hypothetical situation above we simulate from the following distributions:

$$\text{Pop. } 1 : \mathcal{N}(0,1), \quad \text{Pop. } 2 : 0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(\Delta,1), \quad \text{Pop. } 3 : \mathcal{N}(\Delta,1),$$

where Population 1 and 3 have equal sample sizes of n, but Population 2 has a sample size of $10n$. Each scenario is repeated 40000 times.

Although we allow weights to be negative, we still apply the preprocessing step and set the weight of a population to 0 when it does not overlap with the sample from Population 1. If the preprocessing were ignored, there would be a possibility that $\bar{\mathsf{A}}$ would be nonnegative definite and that the MAMSE weights would not be unique.

Applying the preprocessing does not affect the pertinence of this example: if the distributions in the populations of diseased and healthy are so different that the samples are often disjoint, there is no point in using the weighted likelihood to include Population 2 as the measurements are in fact a cheap diagnostic test. Moreover, previous simulations without preprocessing yielded results that are not better than those presented here.

Figure 3.1 shows the average values of the unconstrained MAMSE weights for different scenarios. Negative weights do appear, hence the MAMSE criterion detects that Population 2 is a mixture of the other two populations and removes the component which is not of interest.

For a large Δ, notice how the negative weights are closer to 0 for smaller samples. In such cases, there is a higher probability that the sample from Population 3 will be disjoint of the sample from Population 1. As a result, the weight allocated to Population 3 is more often forced to 0 by the preprocessing step. As the sample size increases, the samples overlap more frequently.

Table 3.6 shows the performances obtained by the MWLE with unconstrained MAMSE weights. The MWLE performs better than the MLE in most cases, being almost twice as good in many cases. Unfortunately, the performances for large Δ are very poor, especially in the cases where the difference between the populations is so large that they overlap very lightly.

Using a weighted likelihood with negative weights provides an improvement over the MLE, but a similar improvement may be obtainable when the constraints are enforced.

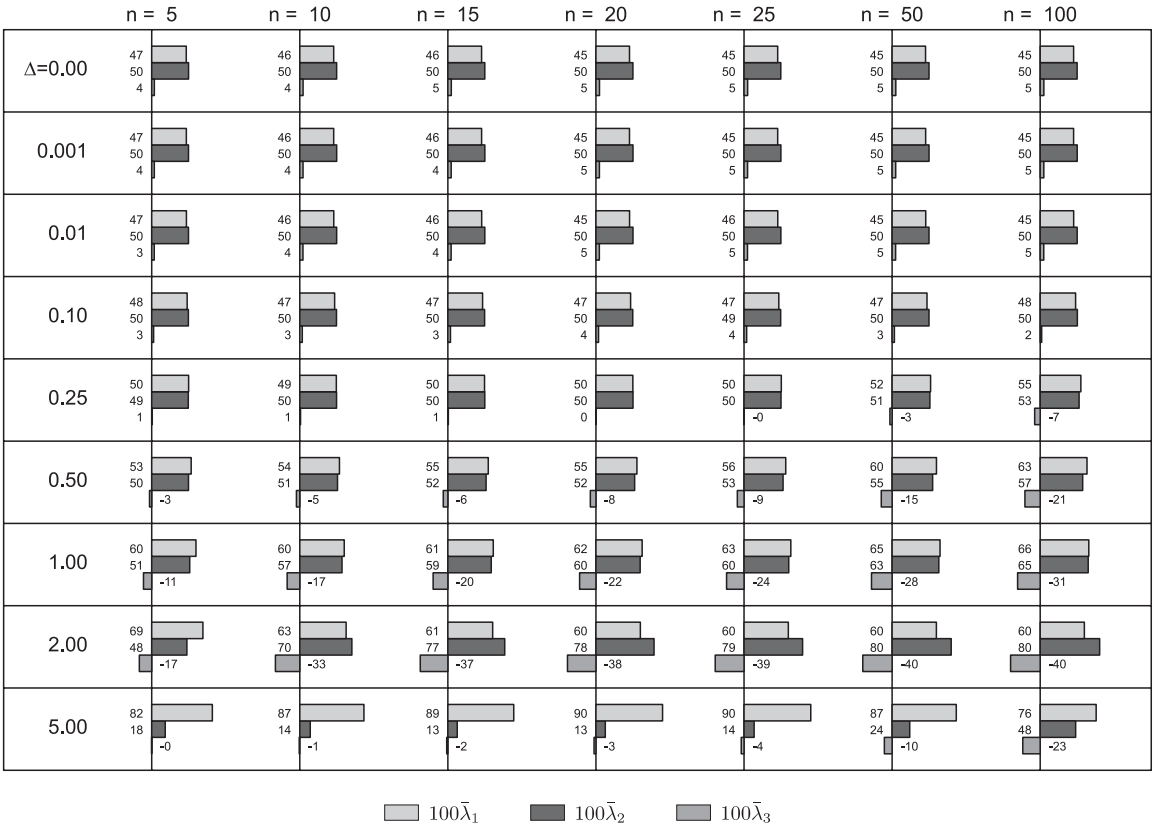

Figure 3.1: Average values of 100× the MAMSE weights without the constraints $\lambda_i \geq 0$. Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathcal{N}(0,1)$ and a $\mathcal{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions.

Table 3.7 shows the performance of the MWLE when the usual MAMSE weights are used. Figure 3.2 shows the average values of the weights obtained in that case. Using the MWLE with positively constrained MAMSE weights also provides an improvement over the MLE. This improvement is sometimes larger than that obtained with unconstrained weights. To discern between the two versions of MAMSE weights, Table 3.8 compares their relative efficiency; values above 100 favor the unconstrained weights. Note that the standard deviation of the error due to simulation in Table 3.8 can be more than one unit, but does not exceed 1.3 units.

	100 MSE(MLE)/MSE(MWLE)						
	$n = 5$	10	15	20	25	50	100
$\Delta = 0$	195	196	197	198	197	197	198
0.001	196	196	197	197	198	198	197
0.01	196	196	197	197	198	198	197
0.10	195	194	194	194	192	184	172
0.25	190	182	176	170	165	144	121
0.50	173	153	140	131	124	107	97
1.00	137	113	105	101	100	97	96
2.00	116	92	86	84	84	84	84
5.00	51	49	51	54	57	62	55

Table 3.6: Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE) when the MAMSE weights are calculated without the constraints $\lambda_i \geq 0$. Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathcal{N}(0, 1)$ and a $\mathcal{N}(\Delta, 1)$ distribution. All results are averages over 40000 repetitions.

	100 MSE(MLE)/MSE(MWLE)						
	$n = 5$	10	15	20	25	50	100
$\Delta = 0$	211	209	210	210	209	208	208
0.001	212	210	209	209	210	209	208
0.01	212	210	210	209	210	209	208
0.10	212	209	207	206	203	194	180
0.25	207	196	187	180	173	146	118
0.50	186	161	144	131	122	98	82
1.00	139	111	97	89	86	79	82
2.00	97	82	79	78	79	84	90
5.00	51	48	50	53	57	68	79

Table 3.7: Relative efficiency as measured by 100 MSE(MLE)/MSE(MWLE) when the usual MAMSE weights (i.e. constrained to positive values) are used. Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathcal{N}(0, 1)$ and a $\mathcal{N}(\Delta, 1)$ distribution. All results are averages over 40000 repetitions.

It seems that allowing negative weights further improves the performances only in a few cases. In fact, Figure 3.2 shows that Population 2 by itself can be used and Table 3.7 shows it has a positive impact. Table 3.8 suggests that the constrained MAMSE weights are to

	100 MSE(constrained)/MSE(negative)						
	$n = 5$	10	15	20	25	50	100
$\Delta = 0$	92	94	94	94	95	95	95
0.001	92	93	94	94	94	95	95
0.01	92	93	94	94	94	95	95
0.10	92	93	94	94	94	95	96
0.25	92	93	94	95	96	98	102
0.50	93	95	98	100	102	109	119
1.00	99	102	109	114	117	123	117
2.00	119	112	109	107	107	100	94
5.00	100	101	102	102	101	91	69

Table 3.8: Relative efficiency of the MWLE with and without the constraints $\lambda_i \geq 0$ as measured by 100 MSE(constrained MWLE)/MSE(unconstrained MWLE). Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathcal{N}(0, 1)$ and a $\mathcal{N}(\Delta, 1)$ distribution. All results are averages over 40000 repetitions.

be preferred more often than not. If we consider other complications arising from allowing negative weights, (e.g. making the weighted EDF non-monotone) keeping the constraints $\lambda_i \geq 0$ in the definition of the MAMSE weights seems a better option.

A different prevalence of diseased in Population 2 could affect the simulation results. If major differences were observed, the conclusion above could be revisited.

3.10.4 Earthquake Data

We now use a model whose weighted likelihood estimate does not have a simple form, i.e. it is not a weighted average of the MLE of each population.

Natural Resources Canada `http://earthquakescanada.nrcan.gc.ca/` maintains an educational website with resources about earthquakes. From their website, it is possible to download data about recent Western Canadian earthquakes. The histograms in Figure 3.3 show the magnitude of the earthquakes that occurred in the 5 year period from the 12^{th} of February 2001 to the 12^{th} of February 2006. Events are divided into 3 groups depending on the geographical location of their epicenter. For the purpose of this example, we make

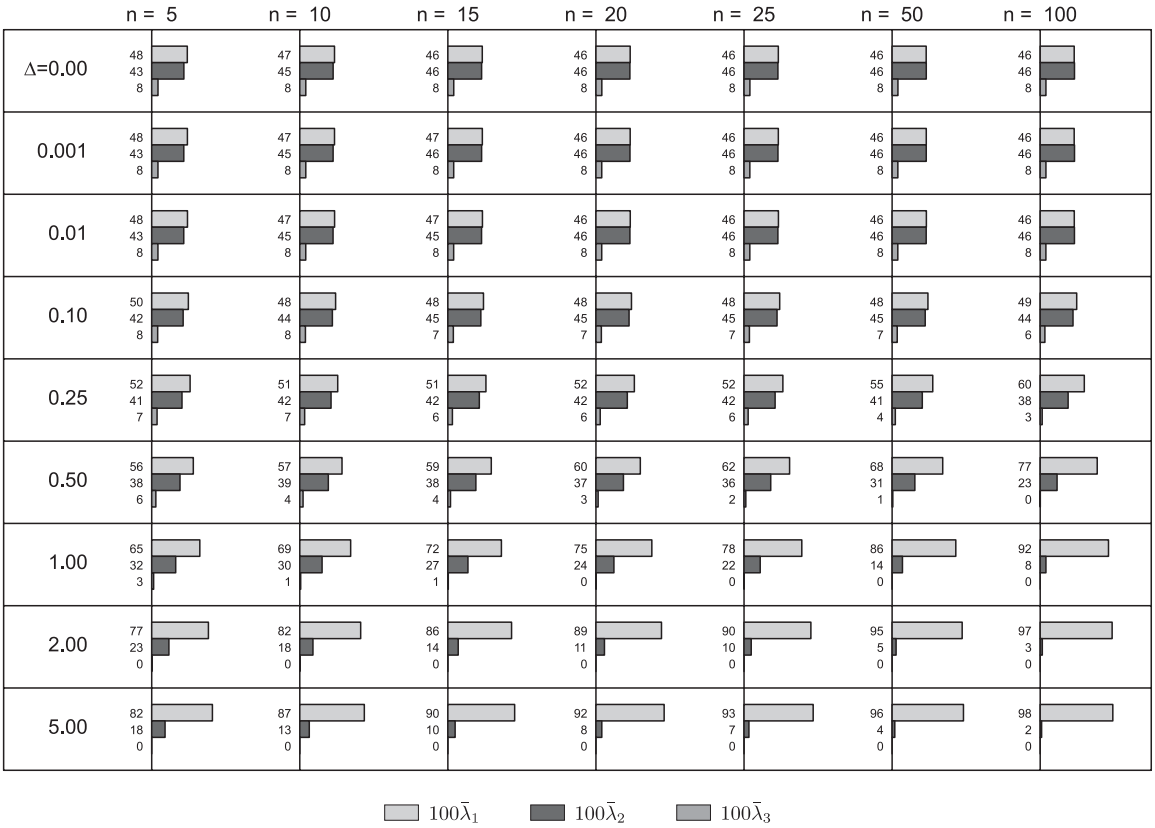

Figure 3.2: Average values of $100\times$ the usual MAMSE weights (with constraints $\lambda_i \geq 0$). Samples of size n, $10n$ and n are taken from each population. Population 2 is an equal mixture of Populations 1 and 3 that respectively follow a $\mathcal{N}(0,1)$ and a $\mathcal{N}(\Delta,1)$ distribution. All results are averages over 40000 repetitions.

the assumption that the magnitude of the earthquakes are independent random variables and fit a gamma distribution to each of the three populations using maximum likelihood. The fitted curves appear on Figure 3.3 and the estimated values of their parameters are shown in Table 3.9 along with the number of observations in each area. The gamma model is parametrized as

$$f(x|\beta,\mu) = \frac{\beta^{\beta\mu}}{\Gamma(\beta\mu)} x^{\beta\mu-1} e^{-\beta x}$$

for $\beta, \mu, x > 0$.

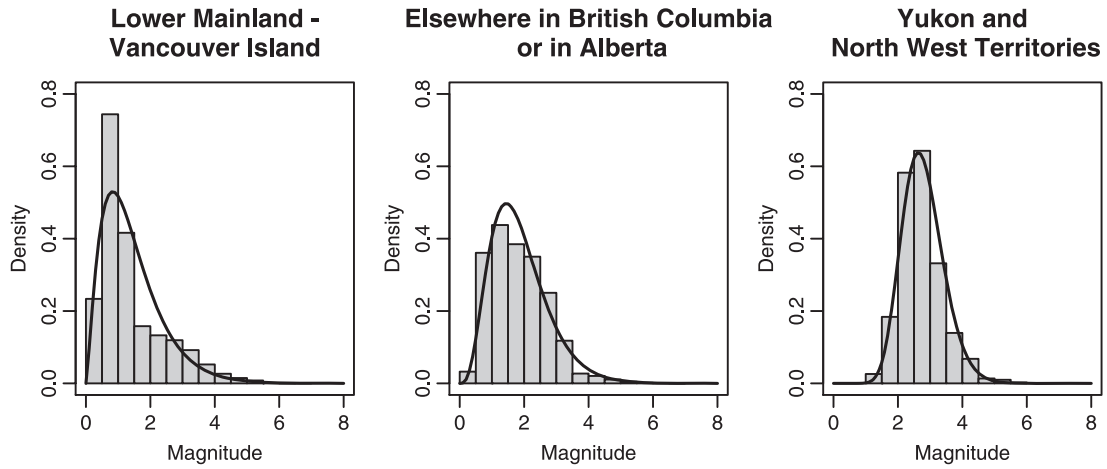

Figure 3.3: Histograms of the magnitude of earthquakes measured between the 12^{th} of February 2001 and the 12^{th} of February 2006 for three different Western Canadian areas. The curves correspond to the fitted Gamma density.

	Lower Mainland – Vancouver Island	Elsewhere in BC or in Alberta	Yukon and North West Territories
β	1.654	2.357	6.806
μ	1.437	1.869	2.782
n	4743	4866	1621

Table 3.9: Number of earthquakes in three Western Canadian areas between the 12^{th} of February 2001 and the 12^{th} of February 2006. The magnitude of these earthquakes is modeled by a Gamma distribution; the maximum likelihood estimates appear above and are used as the "true" parameters for this simulation.

We focus our interest on the magnitude of the next earthquake with epicenter in the Lower Mainland – Vancouver Island area. Suppose that only the 50 most recent events from each of the three regions are available. Would the MWLE that uses data from all three regions provide a better estimate than the MLE? To investigate the question, we produce 10000 pseudo-random samples of earthquakes based on the fitted gamma models shown above.

The average MAMSE weights are 0.959 for the Lower Mainland – Vancouver Island

area, 0.041 for the rest of British Columbia and Alberta and finally, nearly 0 for Yukon and North West Territories. Although it looks like a small contribution, the MSE of the MWLE for the vector (β, μ) was smaller with

$$100 \text{ MSE(MLE)/MSE(MWLE)}=107.$$

We also considered other values of possible interest, namely some probabilities about the magnitude (M) of the next earthquake that are all obtained by plugging the MLE or MWLE in the Gamma model. Table 3.10 summarizes these results.

		Probabilities				
Prob	Efficiency	MLE	MWLE	Model	Data	Multiplier
$P(M > 1)$	123	62	63	68	51	$\times 10^{-2}$
$P(M > 2)$	114	22	24	40	22	$\times 10^{-2}$
$P(M > 3)$	112	66	73	174	98	$\times 10^{-3}$
$P(M > 4)$	113	19	21	51	26	$\times 10^{-3}$
$P(M > 5)$	112	51	59	99	53	$\times 10^{-4}$
$P(M > 6)$	80	14	17	12	6	$\times 10^{-4}$

Table 3.10: Efficiency in estimating some probabilities about the magnitude of the next earthquake in the Lower Mainland – Vancouver Island area followed by the average of the actual estimates and their true values. Efficiency is measured by 100 MSE(plug-in MLE)/MSE(plug-in MWLE). The first four columns of probabilities should be multiplied by the corresponding multiplier.

The first column of Table 3.10 corresponds to the relative efficiency of using the MWLE compared to using the MLE as plug-in parameters for the gamma model in order to evaluate the probability of interest. The numbers shown are 100 MSE(plug-in MLE)/MSE(plug-in MWLE) followed by the estimated values of $P(M > k)$ using the MLE and the MWLE as plug-in parameters. For comparison purposes, the columns *Model* and *Data* contain respectively the true probabilities (from the simulated model) and the empirical proportions in the complete dataset. All probabilities are scaled for easier reading; using the corresponding multiplier will yield the original value. Note that discrepancies with the empirical probabilities reveal weaknesses of the gamma model to perfectly represent the magnitude of earthquakes rather than an advantage for one method over the other.

Interestingly enough, the MSE of the estimates is almost always smaller with the MWLE. Improved performance is hence possible by using the MWLE with MAMSE weights in this situation with distributions copied from real life.

Chapter 4

MAMSE Weights for Right-Censored Data

Situations arise where the exact value of some data may not be observed. In engineering, a test of the lifetime of light bulbs, say, may not last long enough to see the last bulb fail. In health science, patients may move away while participating in a study and the researcher does not know when the symptoms stopped or when the death of the patient occurred. In these two examples, the data are right-censored: some times of failure are not observed, but they are bounded below.

If censored data were ignored, large values of the response variable would be more often excluded from the inference than early deaths (or failures), resulting in an underestimation of the lifetimes. Some methods exist to account for censored data, including a nonparametric estimate of the cumulative distribution function proposed by Kaplan & Meier (1958). We suggest use of that estimate to define a version of the MAMSE weights for censored data.

The paradigm still involves m populations, one of which is of inferential interest. This situation may arise in practice when:

- interest concerns a subgroup of the population studied,
- data come from different studies with different schemes for censoring,

and possibly under other circumstances.

We first introduce the notation for this section and review the Kaplan-Meier estimate and its properties. We then define the MAMSE weights and show that the algorithm for

calculating them converges. We propose the MAMSE-weighted Kaplan-Meier estimate and prove its uniform convergence to the target distribution. Finally, simulations explore the performance of our proposed estimate in finite samples.

4.1 Notation and Review of the Kaplan-Meier Estimate

We introduce a notation that comprises m possibly right-censored populations and accounts for increasing sample sizes. For simplicity, we will adopt a "survival analysis" terminology where measurement of interest is the survival of individuals.

For Population i, let

$$\begin{aligned} X_{ij} &= \text{time of death of individual } j, \\ V_{ij} &= \text{censoring time of individual } j. \end{aligned}$$

The independent positive random variables $X_{ij}(\omega)$ and $V_{ij}(\omega)$ are defined on a common probability space $(\Omega, \mathcal{B}(\Omega), P)$. We denote their distributions by F_i and G_i respectively; the distributions F_i are assumed to be continuous, but the G_i do not need to satisfy that assumption. Instead of the values X_{ij}, we rather observe

$$Z_{ij} = \min(X_{ij}, V_{ij}) \qquad \text{and} \qquad \delta_{ij} = \begin{cases} 0 & \text{if } V_{ij} < X_{ij} \\ 1 & \text{if } V_{ij} \geq X_{ij} \end{cases}.$$

For any fixed $k \in \mathbb{N}$, we observe (Z_{ij}, δ_{ij}) for $i = 1, \ldots, m$ and $j = 1, \ldots, n_{ik}$. The index k will be useful to express the asymptotic results of Section 4.4. We assume that the sample sizes are non-decreasing with k and that $n_{ik} \to \infty$ as $k \to \infty$.

Let us define

$$N_{ik}(s) = \sum_{j=1}^{n_{ik}} \mathbb{1}_{\{Z_{ij} \leq s, \delta_{ij}=1\}} = \#\text{ of deaths up to time } s,$$

$$\begin{aligned} \mathrm{d}N_{ik}(s) &= N_{ik}(s) - N_{ik}(s^-) = \#\text{ of deaths at time } s, \\ Y_{ik}(s) &= \sum_{j=1}^{n_{ik}} \mathbb{1}_{\{Z_{ij} \geq s\}} = \#\text{ at risk just before time } s, \\ \mathrm{d}Y_{ik}(s) &= Y_{ik}(s) - Y_{ik}(s^+) = \#\text{ of deaths and censored at time } s. \end{aligned}$$

For Population i, the Kaplan-Meier estimate (KME) of the probability of dying at time t or earlier is written (see Kaplan & Meier 1958)

$$\hat{F}_{ik}(t) = 1 - \prod_{0 \leq s \leq t} \left\{ 1 - \frac{\mathrm{d}N_{ik}(s)}{Y_{ik}(s)} \right\}.$$

Note that the factors of the product differ from 1 only at times of death.

Let $H_i(t) = P(Z_{i1} \leq t)$ and let $\tau_{H_i} = \sup\{t : H_i(t) < 1\}$ be the largest value that Z_{ij} can attain. The possibility that $\tau_{H_i} = \infty$ is not ruled out although it is unlikely to occur in practice. In addition, let

$$H_i^*(t) = P(Z_{i1} \leq t, \delta_{i1} = 1) = \int_0^t \{1 - G_i(x)\} \, \mathrm{d}F_i(x)$$

be the distribution of observed death times for Population i.

The Kaplan-Meier estimate is an increasing step function with a jump at each of the times of death. By the definition of τ_{H_1}, the number of deaths observed in Population 1 is $\mathcal{N}_k = N_{1k}(\tau_{H_1})$. For $k \in \mathbb{N}$, let $t_{k1} < \cdots < t_{k\mathcal{N}_k}$ be the ordered times of these deaths. The times of death are distinct by the continuity of F_i. If in addition, we use the convention that $t_{k0} = 0$, the jumps of $\hat{F}_{1k}(t)$ are $J_{kj} = \hat{F}_{1k}(t_{kj}) - \hat{F}_{1k}(t_{k(j-1)})$ for $j \in \{1, \ldots, \mathcal{N}_k\}$ and we have that $\sum_{j=1}^{\mathcal{N}_k} J_{kj} \leq 1$.

Efron (1967) discusses how the Kaplan-Meier estimate redistributes weight of the censored data to the observations on the right. Consequently,

$$J_{k1} \leq J_{k2} \leq \cdots \leq J_{k\mathcal{N}_k}. \tag{4.1}$$

We will consider the Kaplan-Meier estimate on a bounded interval $[0, U]$ with $U < \tau_{H_1}$.

Theorem 4.1 (Winter, Földes & Rejtö)

$$\sup_{t \leq U} |\hat{F}_{ik}(t) - F_i(t)| \to 0$$

almost surely as $n_{ik} \to \infty$.

Földes & Rejtö (1981) study the rate of convergence of $\sup_{t \leq U} |\hat{F}_{ik}(t) - F_i(t)|$. To get a better rate, they assume that the distribution G_i is continuous. However, they also mention that they proved the result of Theorem 4.1 without making any assumptions about F_i and G_i in some earlier work published as Winter, Földes & Rejtö (1978).

Efron (1967) and Breslow & Crowley (1974) assume that the distribution G_i is continuous and show that $\hat{F}_{ik}(t)$ is approximately normal with mean $F_i(t)$ and variance

$$\frac{1}{n}\{1 - F_i(t)\}^2 \int_0^t \{1 - H_i(x)\}^{-2} \, \mathrm{d}H_i^*(x).$$

This expression for the variance can be estimated using Greenwood's formula (see e.g. Chen & Lo (1997), page 1069), yielding

$$\widehat{\mathrm{var}}\{\hat{F}_{ik}(t)\} \approx \widetilde{\mathrm{var}}\{\hat{F}_{ik}(t)\} = \{1 - \hat{F}_{ik}(t)\}^2 \sum_{0 \leq s \leq t} \frac{\mathrm{d}N_{ik}(s)}{Y_{ik}(s)Y_{ik}(s^+)}. \tag{4.2}$$

This expression becomes less reliable as t approaches τ_{H_i}, but for a large enough n_{ik}, it should be reliable on any interval $[0, T]$ with $T < U$. Note that the terms in the sum above are 0 except when s is a time of death.

Defining MAMSE weights based on the Kaplan-Meier estimate involves using an estimate of its variance. Equation (4.2) is thus used for the definition of the MAMSE weights in Equation (4.3). Since the variance term is used as a penalty to foster the inclusion of many populations into the inference, we take the liberty of using $\widetilde{\mathrm{var}}\{\hat{F}_{ik}(t)\}$ even though we do not make the assumption that G_i is continuous.

In the next section, we build a version of the MAMSE weights based on the Kaplan-Meier estimate. The weights are then used to define a MAMSE-weighted Kaplan-Meier estimate (WKME).

4.2 Definition of the MAMSE Weights for Right-Censored Data

We suggest using an expression of the form $\sum_{i=1}^{m} \lambda_i \hat{F}_{ik}(t)$ to estimate $F_1(t)$. To find the weights $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^\mathsf{T}$ adaptively, we use a version of the MAMSE criterion, see Equation (2.1), where the distribution functions are estimated with the corresponding Kaplan-Meier estimates.

When censored observations are found after the last time of death, the Kaplan-Meier estimate is not a proper cumulative distribution function on the positive real line because it never reaches a value of 1. For that reason, we assume that we can specify an upper bound $U < \tau_{H_1}$ and limit our study of the survival function to the interval $[0, T]$ where $T < U$ is such that $H_1^*(T) < H_1^*(U)$. The last inequality means that there is a non-null probability that a death is observed in the interval $(T, U]$. This will be the case whenever the probability of death (observed or not) is non-null in that interval since the cumulative probability of being censored before T is less than one by the definition of $\tau_{H_1} > U$.

Preprocessing

For a fixed k and $i \in \{2, \ldots, m\}$, let

$$m_{ik} = \min_{\{j \le n_{ik} : \delta_{ij} = 1\}} Z_{ij} \qquad \text{and} \qquad M_{ik} = \max_{\{j \le n_{ik} : \delta_{ij} = 1\}} Z_{ij}$$

be the smallest and largest times of death observed in Population i.

The weights allocated to the sample from Population i are set to 0 if it fails to satisfy the following two conditions:

1. $U \in [m_{ik}, M_{ik}]$, i.e. at least one observed death from Population i is in the interval $[0, U]$ and at least one observed death occurs after U;

2. $\sum_{\{j \leq n_{1k} : \delta_{1j} = 1\}} \mathbb{1}\{X_{1j} \in [m_{ik}, \min(M_{ik}, U)]\} \geq 1$, i.e. at least one observed death from Population 1 which occurred in $[0, U]$ falls within the range of the observed times of death in Population i.

Condition 1 ensures that Formula (4.2) is well defined on $[0, U]$ and not null everywhere on that interval. Condition 2 means that the same formula will be strictly positive for at least one of the times of death from Population 1 in $[0, U]$, ensuring the unicity of the MAMSE weights and the convergence of the algorithm used to calculate them. These consequences are explained in greater detail in Section 4.3.

The preprocessing requirements appear to be technical, but they avoid using samples yielding unreliable Kaplan-Meier estimates. After the last time of death, the KME remains constant forever. If Condition 1 failed, we could be relying on such a plateau where subtle differences in the distribution of the lifetimes may be hidden by the censoring scheme. If Condition 2 failed, the whole range where the KME of Population i increases would be compared against a constant function from Population 1. Hence, the conditions also have some intuitive foundations.

If very few deaths from Population 1 fall in $[0, U]$, Condition 2 is likely to fail and the other populations may see their weights set to 0. In such cases, the little information available about Population 1 makes it hard to compare it to the other samples. It is then better to be more conservative and to avoid relying on the other populations.

In particular, if less than 2 deaths from Population 1 fall in the interval $[0, U]$, we allocate no weight to the other populations. The lack of knowledge about Population 1 makes the comparison to other populations too uncertain.

Let $\mathcal{M}_k \subset \{1, \ldots, m\}$ be a set of indices corresponding to the population whose samples satisfy the preprocessing conditions. We always have $1 \in \mathcal{M}_k$ since Population 1 is never

removed from the pool of populations.

Objective Function

Let

$$P_k(\boldsymbol{\lambda}) = \int_0^U \left[\left\{ \hat{F}_{1k}(t) - \sum_{i=1}^m \lambda_i \hat{F}_{ik}(t) \right\}^2 + \sum_{i=1}^m \lambda_i^2 \widetilde{\mathrm{var}} \left\{ \hat{F}_{ik}(t) \right\} \right] \mathrm{d}\hat{F}_{1k}(t) \tag{4.3}$$

be a special case of Equation (2.1) where $\widehat{\mathrm{var}}\{\hat{F}_{ik}(t)\}$ is estimated by $\widetilde{\mathrm{var}}\{\hat{F}_{ik}(t)\}$ from Equation (4.2) and $\mathrm{d}\mu(x)$ is replaced by $\mathrm{d}\hat{F}_{1k}(t)$.

Note that none of the preprocessing steps can remove Population 1 since it is the population of interest. In cases where the last observed death in Population 1 is contained in $[0, U]$ (i.e. when $M_{1k} = t_{k\mathcal{N}_k} < U$), the expression for $\widetilde{\mathrm{var}}\{\hat{F}_{1k}(t_{k\mathcal{N}_k})\}$ involves a division by 0 since $Y_{ik}(t^+_{k\mathcal{N}_k}) = 0$. In that case, we substitute the ill-defined term by its value just before time $t_{k\mathcal{N}_k}$. Although this solution would not be acceptable for constructing confidence intervals, it should do no harm here where our purpose is to construct a penalty term that fosters using the data from all populations. In particular, this adjustment will affect at most one term of the integral $P_k(\boldsymbol{\lambda})$.

The weights are chosen to

$$\begin{aligned} \text{minimize} \quad & P_k(\boldsymbol{\lambda}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq 0 \text{ and } \sum_{i \in \mathcal{M}_k} \lambda_i = 1. \end{aligned}$$

The solution to that program will be referred to as the survival MAMSE weights and written $\boldsymbol{\lambda}_k(\omega) = [\lambda_{1k}(\omega), \ldots, \lambda_{mk}(\omega)]^\mathsf{T}$ to represent their dependence on ω and k. For values of t in the interval $[0, T]$, the weighted Kaplan-Meier estimate (WKME) of the lifetime's CDF is then defined by

$$\hat{G}_k(t) = \sum_{i=1}^m \lambda_{ik}(\omega) \hat{F}_{ik}(t). \tag{4.4}$$

4.3 Computing the MAMSE Weights for Right-Censored Data

The algorithm suggested in Section 2.4 applies to the MAMSE weights defined with the Kaplan-Meier estimate. To prove that the algorithm converges, we must show that Assumption 2.1 is satisfied.

Lemma 4.1

$$\int \widetilde{\text{var}}\left\{\hat{F}_{1k}(x)\right\}\, \mathrm{d}\hat{F}_{1k}(x) > 0$$

Proof of Lemma 4.1. To calculate the MAMSE weights, we require that at least two times of death from Population 1 fall in the interval $[0, U]$. Let t^* be the smaller of the two times of death. By the definition of the Kaplan-Meier estimate, $\hat{F}_{1k}(t^*) < 1$ since the larger time of death has some mass. Hence, Expression (4.2) is positive at $t = t^*$, i.e.

$$\widetilde{\text{var}}\left\{\hat{F}_{1k}(t^*)\right\} > 0.$$

Since t^* is a time of death, the Kaplan-Meier estimate for Population 1 makes a jump at that time. Hence $\mathrm{d}\hat{F}_{1k}(t^*) > 0$ as well and consequently,

$$\int \widetilde{\text{var}}\left\{\hat{F}_{1k}(x)\right\}\, \mathrm{d}\hat{F}_{1k}(x) > 0.$$ ■

Lemma 4.2 *For all external populations remaining after prescreening:* $i = \{2, \ldots, m\}$,

$$\int \widetilde{\text{var}}\left\{\hat{F}_{ik}(x)\right\}\, \mathrm{d}\hat{F}_{1k}(x) > 0.$$

Proof of Lemma 4.2. Let m_{ik} and M_{ik} be the smallest and largest time of death respectively observed in Population i. Note that the sum in Equation (4.2) is cumulative. It is thus positive for all $x \geq m_{ik}$.

The first condition for preprocessing requires that $m_{ik} \leq U \leq M_{ik}$. Since the Kaplan-Meier estimate for Population i jumps at M_{ik}, the first part of Equation (4.2), $\{1-\hat{F}_{ik}(t)\}^2$, is positive for all $t < M_{ik}$. Consequently,

$$\widetilde{\text{var}}\left\{\hat{F}_{ik}(x)\right\} > 0$$

for all $x \in [m_{ik}, M_{ik})$.

The second requirement for preprocessing ensures that a death occurs in Population 1 in the interval $[m_{ik}, \min(M_{ik}, U)]$. Consequently, the discrete measure $\mathrm{d}\hat{F}_{1k}(x)$ gives positive mass to at least one point in that interval where $\widetilde{\text{var}}\left\{\hat{F}_{ik}(x)\right\} > 0$. Therefore,

$$\int \widetilde{\text{var}}\left\{\hat{F}_{ik}(x)\right\}\, \mathrm{d}\hat{F}_{1k}(x) > 0. \qquad \blacksquare$$

Lemmas 4.1 and 4.2 are sufficient to show that the algorithm in Section 2.4 converges for this new application to the MAMSE weights.

4.4 Uniform Convergence of the MAMSE-Weighted Kaplan-Meier Estimate

We prove that the WKME, $\hat{G}_k(t)$, converges uniformly in probability to $F_1(t)$. The proof is built as a sequence of lemmas that appear next and follow a logic akin to that of Section 3.5.

Remember that we assumed in Section 4.1 that the distribution of the times of death is continuous, but the distribution of the times of censoring need not be.

Lemma 4.3

$$\sum_{0 \leq s \leq U} \frac{\mathrm{d}N_{1k}(s)}{Y_{1k}(s)Y_{1k}(s^+)} \xrightarrow{P} 0$$

as $k \to \infty$.

Proof of Lemma 4.3. Notice that

$$\sum_{0\le s\le U} \frac{\mathrm{d}N_{1k}(s)}{Y_{1k}(s)Y_{1k}(s^+)} \le \sum_{0\le s\le U} \frac{\mathrm{d}Y_{1k}(s)}{Y_{1k}(s)Y_{1k}(s^+)} = \sum_{0\le s\le U} \frac{Y_{1k}(s)-Y_{1k}(s^+)}{Y_{1k}(s)Y_{1k}(s^+)}.$$

The first inequality holds since $\mathrm{d}N_{1k}(s) \le \mathrm{d}Y_{1k}(s)$ for all s.

Suppose that the Z_{1j}'s are distinct and for a fixed k, let $Z_{1(j)}$ denote the j^{th} order statistic of the sample from Population 1, i.e. the j^{th} smallest value in the list $Z_{11},\ldots,Z_{1n_{ik}}$. Let $j_{0k} = \max\{j : j \le n_{1k},\, Z_{1(j)} \le U\}$ Then, there is at most one censored datum or death at any given time and the expression above can be rewritten as

$$\begin{aligned}\sum_{0\le s\le U}\left\{\frac{1}{Y_{1k}(s^+)}-\frac{1}{Y_{1k}(s)}\right\} &= \sum_{j=1}^{j_{0k}}\left\{\frac{1}{Y_{1k}(Z_{1(j+1)})}-\frac{1}{Y_{1k}(Z_{1(j)})}\right\} \\ &= \frac{1}{Y_{1k}(Z_{1(j_{0k})})}-\frac{1}{n_{1k}} \le \frac{1}{Y_{1k}(U)} \qquad (4.5)\end{aligned}$$

since $Y_{1k}(s)$ is decreasing in s and the series telescopes.

Concurrent death times are impossible, but concurrent censoring times are possible since the continuity of their underlying distribution is not assumed. Moreover, censoring cannot occur at a time of death since the two times are independent random variables and at least one of them has a continuous distribution. Even if multiple censoring occurs at one time, inequality (4.5) will still hold. Indeed, the term for concurrent censoring times in the summation is equal to the sum of the individual terms if the times were different but consecutive. For instance, if $Y_{1k}(t) - Y_{1k}(t^+) = 2$, we have

$$\begin{aligned}\frac{Y_{1k}(t)-Y_{1k}(t^+)}{Y_{1k}(t)Y_{1k}(t^+)} &= \frac{1}{Y_{1k}(t)-2}-\frac{1}{Y_{1k}(t)} \\ &= \frac{1}{Y_{1k}(t)-2}-\frac{1}{Y_{1k}(t)-1}+\frac{1}{Y_{1k}(t)-1}-\frac{1}{Y_{1k}(t)}.\end{aligned}$$

This extends for an arbitrary number of concurrent censoring times.

Let us now show that the bound $1/Y_{1k}(U)$ converges to 0. Since the Z_{1j}'s are indepen-

dent, $Y_{1k}(U)$ has a Binomial distribution with parameters n_{1k} and $1 - H_1(U)$. Let $\epsilon > 0$, then

$$\begin{aligned} P\left\{\sum_{0\le s\le U}\frac{\mathrm{d}N_{1k}(s)}{Y_{1k}(s)Y_{1k}(s^+)} > \epsilon\right\} &\le P\left\{\frac{1}{Y_{1k}(U)} > \epsilon\right\} = P\left\{Y_{1k}(U) < 1/\epsilon\right\} \\ &\approx \Phi\left[\frac{1/\epsilon - n_{1k}\{1-H_1(U)\}}{\sqrt{n_{1k}H_1(U)\{1-H_1(U)\}}}\right] \to 0 \end{aligned} \tag{4.6}$$

as $k \to \infty$ since the argument inside the standard normal CDF Φ tends to $-\infty$. The approximation is from the central limit theorem and becomes exact as $n_{1k} \to \infty$. ■

Whether a sample is rejected in the preprocessing or not may vary with k and ω. Remember that m populations are available before preprocessing and that weights of populations in $\mathcal{M}_k^C$ are forced to 0, but Population 1 is never excluded from the optimization problem. Hence preprocessing does not affect the distribution of probabilities calculated in expressions such as (4.6). Moreover, preprocessing does not change the fact that $\boldsymbol{\lambda} = [1, 0, \ldots, 0]\mathsf{T}$ is a suboptimal choice of weights, which we use in the proof of the next result.

Lemma 4.4

$$\int_0^U \left\{\hat{F}_{1k}(t) - \hat{G}_k(t)\right\}^2 \mathrm{d}\hat{F}_{1k}(t) \xrightarrow{P} 0$$

as $k \to \infty$, where $\hat{G}_k(t)$ is defined in Equation (4.4).

Proof of Lemma 4.4. By the definition of the MAMSE weights,

$$P_k\{\boldsymbol{\lambda}_k(\omega)\} \le P_k\{[1, 0, \ldots, 0]^\mathsf{T}\}$$

since $[1, 0, \ldots, 0]^\mathsf{T}$ is a suboptimal choice of weights. Following Equations (4.2) and (4.3), we thus have

$$\int_0^U \left\{\hat{F}_{1k}(t) - \hat{G}_k(t)\right\}^2 \mathrm{d}\hat{F}_{1k} \le P_k\{\boldsymbol{\lambda}_k(\omega)\} \le P_k\{[1, 0, \ldots, 0]^\mathsf{T}\}$$

$$
\begin{aligned}
&= \int_0^U \left[\{1 - \hat{F}_{1k}(t)\}^2 \sum_{0 \le s \le t} \frac{\mathrm{d}N_{1k}(s)}{Y_{1k}(s)Y_{1k}(s^+)} \right] \mathrm{d}\hat{F}_{1k}(t) \\
&\le \left\{ \sum_{0 \le s \le U} \frac{\mathrm{d}N_{1k}(s)}{Y_{1k}(s)Y_{1k}(s^+)} \right\} \int_0^U \{1 - \hat{F}_{1k}(t)\}^2 \, \mathrm{d}\hat{F}_{1k}(t) \\
&\le \sum_{0 \le s \le U} \frac{\mathrm{d}N_{1k}(s)}{Y_{1k}(s)Y_{1k}(s^+)} \xrightarrow{P} 0
\end{aligned}
$$

as $k \to \infty$ by Lemma 4.3. ■

Let $\nu_k = \max\{j \le n_{1k} : t_{kj} \le U\}$ be the index of the largest time of death observed at or before time U in the sample from Population 1. Since the steps of $\hat{F}_{1k}(t)$ are increasing in size (see Equation 4.1), we may define

$$\mathcal{J}_k = \max_{t \le U} |\hat{F}_{1k}(t) - \hat{F}_{1k}(t^-)| = J_{k\nu_k},$$

the biggest jump of $\hat{F}_{1k}(t)$ on the interval $[0, U]$.

Lemma 4.5 *$\mathcal{J}_k \to 0$ almost surely as $k \to \infty$.*

Proof of Lemma 4.5. The result follows from Theorem 4.1 since

$$\mathcal{J}_k \le 2 \sup_{0 \le t \le U} \left| \hat{F}_{1k}(t) - F_1(t) \right| \to 0$$

almost surely as $k \to \infty$. ■

Recalling that $T < U$ is such that $H_1^*(T) < H_1^*(U)$, we let

$$D_k = N_{1k}(U) - N_{1k}(T) = \sum_{j=1}^{n_{1k}} \mathbb{1}_{\{Z_{1j} \in (T,U],\, \delta_{1j}=1\}}$$

be the number of deaths observed in the interval $(T, U]$ among individuals sampled from

Population 1. Since the Z_{1j} are independent, D_k follows a Binomial distribution with parameters n_{1k} and $H_1^*(U) - H_1^*(T)$.

Let $\ell_k = N_{1k}(T)$ be the number of deaths observed in the interval $[0, T]$, and their corresponding times of death $t_{k1} < \ldots < t_{k\ell_k} \leq T$. By convention, we set $t_{k(\ell_k+1)} = \tau_{H_1}$ if no death is observed after $t_{k\ell_k}$.

Lemma 4.6

$$P\left\{\max_{0\leq t\leq T}\left|\hat{F}_{1k}(t) - \hat{G}_k(t)\right| \leq \mathcal{J}_k + \max_{t\in\{t_{k1},\ldots,t_{k(\ell_k+1)}\}}\left|\hat{F}_{1k}(t) - \hat{G}_k(t)\right|\right\}$$

converges to 1 as $k \to \infty$.

Proof of Lemma 4.6. Fix $k \in \mathbb{N}$ and $\omega \in \Omega$, and let $x_0 \in [0, T]$ be the value maximizing $|\hat{F}_{1k}(t) - \hat{G}_k(t)|$. That maximum exists since $|\hat{F}_{1k}(t) - \hat{G}_k(t)|$ is a bounded function being optimized on a compact set. Three disjoint cases need to be considered:

Case 1: $\hat{G}_k(x_0) \leq \hat{F}_{1k}(x_0)$ and $D_k \geq 1$.
Let $j_1 = \max\{j \in \{1, \ldots, \ell_k\} : t_{kj} \leq x_0\}$ be the index of the largest time of death from Population 1 inferior to x_0. By the choice of j_1, t_{kj_1} belongs to the same step as x_0 and hence

$$\hat{F}_{1k}(t_{kj_1}) = \hat{F}_{1k}(x_0).$$

Moreover,

$$\hat{G}_k(t_{kj_1}) \leq \hat{G}_k(x_0).$$

since $\hat{G}_k(t)$ is a monotone nondecreasing function. Recalling that x_0 maximizes the difference between $\hat{F}_{1k}(t)$ and $\hat{G}_k(t)$, we can write

$$\begin{aligned}\max_{0\leq t\leq T}|\hat{F}_{1k}(t) - \hat{G}_k(t)| &= \hat{F}_{1k}(x_0) - \hat{G}_k(x_0)\\ &\leq \hat{F}_{1k}(t_{kj_1}) - \hat{G}_k(t_{kj_1})\end{aligned}$$

$$\begin{aligned} &\leq \max_{t\in\{t_{k1},\ldots,t_{k\ell_k}\}} \left|\hat{F}_{1k}(t) - \hat{G}_k(t)\right| \\ &\leq \mathcal{J}_k + \max_{t\in\{t_{k1},\ldots,t_{k(\ell_k+1)}\}} \left|\hat{F}_{1k}(t) - \hat{G}_k(t)\right|, \end{aligned}$$

meaning that the maximum will always occur at a time of death from Population 1, where $\hat{F}_{1k}(t)$ has a step.

Case 2: $\hat{G}_k(x_0) > \hat{F}_{1k}(x_0)$ and $D_k \geq 1$.

Let $j_2 = \min\{j \in \{1,\ldots,\ell_k+1\} : t_{kj} \geq x_0\}$ be the index of the smallest time of death greater than x_0. The condition $D_k \geq 1$ ensures that $t_{k(\ell_k+1)}$ exists, hence j_2 is well defined. The choice of j_2 ensures that it belongs to the step of $\hat{F}_{1k}(t)$ that immediately follows x_0, hence

$$\hat{F}_{1k}(t_{k(j_2-1)}) = \hat{F}_{1k}(x_0).$$

The function $\hat{G}_k(t)$ is a right-continuous nondecreasing function and $t_{kj_2} \geq x_0$, thus

$$\hat{G}_k(t_{kj_2}) \geq \hat{G}_k(x_0).$$

Recalling that x_0 is the point maximizing the difference between $\hat{F}_{1k}(t)$ and $\hat{G}_k(t)$, we write

$$\begin{aligned} \max_{0\leq t\leq T} |\hat{F}_{1k}(t) - \hat{G}_k(t)| &= \hat{G}_k(x_0) - \hat{F}_{1k}(x_0) \\ &\leq \hat{G}_k(t_{kj_2}) - \hat{F}_{1k}(t_{k(j_2-1)}) \\ &= \{\hat{F}_{1k}(t_{kj_2}) - \hat{F}_{1k}(t_{k(j_2-1)})\} + \hat{G}_k(t_{kj_2}) - \hat{F}_{1k}(t_{kj_2}) \\ &\leq \mathcal{J}_k + \max_{t\in\{t_{k1},\ldots,t_{k(\ell_k+1)}\}} \left|\hat{F}_{1k}(t) - \hat{G}_k(t)\right|, \end{aligned}$$

meaning that under Case 2, the maximum will occur immediately before a jump of $\hat{F}_{1k}(t)$.

Case 3: $D_k = 0$.

This event has probability $[1 - \{H_1^*(U) - H_1^*(T)\}]^{n_{1k}} \to 0$ as $k \to \infty$.

The combination of Cases 1 and 2 implies that the desired result is true at least whenever $D_k \geq 1$. Consequently,

$$P\left\{\max_{0\leq t\leq T}\left|\hat{F}_{1k}(t)-\hat{G}_k(t)\right| \leq \mathcal{J}_k + \max_{t\in\{t_{k1},\ldots,t_{k(\ell_k+1)}\}}\left|\hat{F}_{1k}(t)-\hat{G}_k(t)\right|\right\}$$

$$\geq P(D_k \geq 1) = 1 - P(D_k = 0) \to 1$$

as $k \to \infty$ since the probability of Case 3 converges to 0 as $k \to \infty$. ■

Lemma 4.7

$$\max_{t\in\{t_{k1},\ldots,t_{k(\ell_k+1)}\}}\left|\hat{F}_{1k}(t)-\hat{G}_k(t)\right| \xrightarrow{P} 0$$

as $k \to \infty$.

Proof of Lemma 4.7. Let $\epsilon > 0$ be such that $\epsilon < H_1^*(U) - H_1^*(T)$. We will show that

$$P\left\{\max_{t\in\{t_{k1},\ldots,t_{k(\ell_k+1)}\}}\left|\hat{F}_{1k}(t)-\hat{G}_k(t)\right| > \epsilon\right\} \to 0$$

as $k \to \infty$.

For a large k, let $x_k \in \{1,\ldots,\ell_k+1\}$ be the index of a time of death where the difference $\left|\hat{F}_{1k}(t)-\hat{G}_k(t)\right|$ is maximized. We define the following three events:

$$\begin{aligned} A_k &= \left\{\omega\in\Omega : \hat{F}_{1k}(t_{kx_k}) - \hat{G}_k(t_{kx_k}) > \epsilon\right\} \\ B_k &= \left\{\omega\in\Omega : \hat{G}_k(t_{kx_k}) - \hat{F}_{1k}(t_{kx_k}) > \epsilon\right\} \\ C_k &= \left\{\omega\in\Omega : D_k \geq \epsilon n_{1k} + 1\right\}. \end{aligned}$$

Then,

$$P\left\{\max_{t\in\{t_{k1},\ldots,t_{k(\ell_k+1)}\}}\left|\hat{F}_{1k}(t)-\hat{G}_k(t)\right| > \epsilon\right\} \leq P\{C_k^C \cup (A_k\cap C_k)\cup(B_k\cap C_k)\}$$

$$\leq \quad P(C_k^C) + P(A_k \cap C_k) + P(B_k \cap C_k).$$

We next show that each of the three probabilities on the right hand side go to zero as $k \to \infty$. Note that the event C_k is used to remove an event of probability 0 that would otherwise complicate the proof that $P(B_k) \to 0$.

Case 1: $P(C_k^C) \to 0$.

Recalling that D_k follows a Binomial distribution with n_{1k} trials and probability of success $\{H_1^*(U) - H_1^*(T)\}$, we have

$$\begin{aligned} P(C_k^C) &= P\{D_k < \epsilon n_{1k} + 1\} \\ &\approx \Phi\left(\frac{\epsilon n_{1k} + 1 - n_{1k}\{H_1^*(U) - H_1^*(T)\}}{\sqrt{n_{1k}\{H_1^*(U) - H_1^*(T)\}\{1 - H_1^*(U) + H_1^*(T)\}}} \right) \\ &= \Phi(-c\sqrt{n_{1k}} + d/\sqrt{n_{1k}}) \to 0 \end{aligned}$$

as $k \to \infty$ since the choice of ϵ implies that $c > 0$, hence the argument inside the standard normal CDF Φ tends towards $-\infty$. We use the central limit theorem to approximate the Binomial by the Normal, but this comparison becomes exact as n_{1k} approaches ∞.

Case 2: $P(A_k \cap C_k) \to 0$.

Let $u_k = \min\{u : \sum_{i=u+1}^{x_k} J_{ki} \leq \epsilon\}$. This index exists when k is large enough since

- $J_{kx_k} \leq \mathcal{J}_k \to 0$ by Lemma 4.5 and
- $\sum_{i=1}^{x_k} J_{ki} = \hat{F}_{1k}(t_{kx_k}) > \hat{G}_k(t_{kx_k}) + \epsilon \geq \epsilon$.

For a large enough k, $J_{kx_k} < \epsilon$ and hence $u_k \leq x_k - 1$. For $j \in \{u_k, \ldots, x_k - 1\}$, we have

$$\hat{G}_k(t_{kj}) \leq \hat{G}_k(t_{kx_k})$$

since the function $\hat{G}_k(t)$ is monotone nondecreasing and

$$\hat{F}_{1k}(t_{kj}) = \hat{F}_{1k}(t_{kx_k}) - \sum_{i=j+1}^{x_k} J_{ki}$$

since the function $\hat{F}_{1k}(t)$ makes a jump of size J_{ki} at time t_{ki}. Combining these inequalities yields

$$\hat{F}_{1k}(t_{kj}) - \hat{G}_k(t_{kj}) \geq \hat{F}_{1k}(t_{kx_k}) - \hat{G}_k(t_{kx_k}) - \sum_{i=j+1}^{x_k} J_{ki} \geq \epsilon - \sum_{i=j+1}^{x_k} J_{ki} \geq 0.$$

The last inequality holds because of the choice of u_k. The function $\hat{F}_{1k}(t)$ gives a mass of J_{kj} to the point t_{kj}, and hence

$$\begin{aligned}
\int_0^U |\hat{G}_k(t) - \hat{F}_{1k}(t)|^2 \, \mathrm{d}\hat{F}_{1k}(t) &\geq \sum_{j=1}^{\ell_k} J_{kj} |\hat{G}_k(t_{kj}) - \hat{F}_{1k}(t_{kj})|^2 \\
&\geq \sum_{j=u_k}^{x_k} J_{kj} |\hat{G}_k(t_{kj}) - \hat{F}_{1k}(t_{kj})|^2 \\
&\geq J_{kx_k}\epsilon^2 + \sum_{j=u_k}^{x_k-1} J_{kj} \left(\epsilon - \sum_{i=j+1}^{x_k} J_{ki} \right)^2 . \qquad (4.7) \\
&\to \int_0^\epsilon (\epsilon - x)^2 \, \mathrm{d}x = \int_0^\epsilon x^2 \, \mathrm{d}x = \frac{\epsilon^3}{3}
\end{aligned}$$

since the summation corresponds to the Riemann sum for the integral $\int_0^\epsilon (\epsilon - x)^2 \, \mathrm{d}x$ depicted in Figure 4.1. The sum converges as $k \to \infty$ because the width of the columns J_{kj} tends to zero by Lemma 4.5.

To clarify the link between the Riemann sum and the integral, consider the change of variable $p = x_k - j$ and let

$$c_{kp} = \begin{cases} 0 & p = 0 \\ \sum_{i=1}^{p} J_{k(x_k - i + 1)} & p = 1, \ldots, x_k - u_k \end{cases} .$$

Note that $c_{k(p+1)} - c_{kp} = J_{k(x_k-p)} = J_{kj}$ and with respect to the variable j, c_{kp} equals $\sum_{i=j+1}^{x_k} J_{ki}$ when $p > 0$. We can thus write the expression in (4.7) as

$$\sum_{p=0}^{x_k-u_k-1} (c_{k(p+1)} - c_{kp})(\epsilon - c_{kp})^2 \to \int_0^\epsilon (\epsilon - x)^2 \, \mathrm{d}x = \int_0^\epsilon x^2 \, \mathrm{d}x = \frac{\epsilon^3}{3} \tag{4.8}$$

Consequently, there exists a k_0 such that

$$\int_0^U |\hat{G}_k(t) - \hat{F}_{1k}(t)|^2 \, \mathrm{d}\hat{F}_{1k}(t) > \frac{\epsilon^3}{6}$$

for all $k \geq k_0$, an event of probability 0 according to Lemma 4.4. We conclude that $P(A_k \cap C_k) \to 0$ as $k \to \infty$.

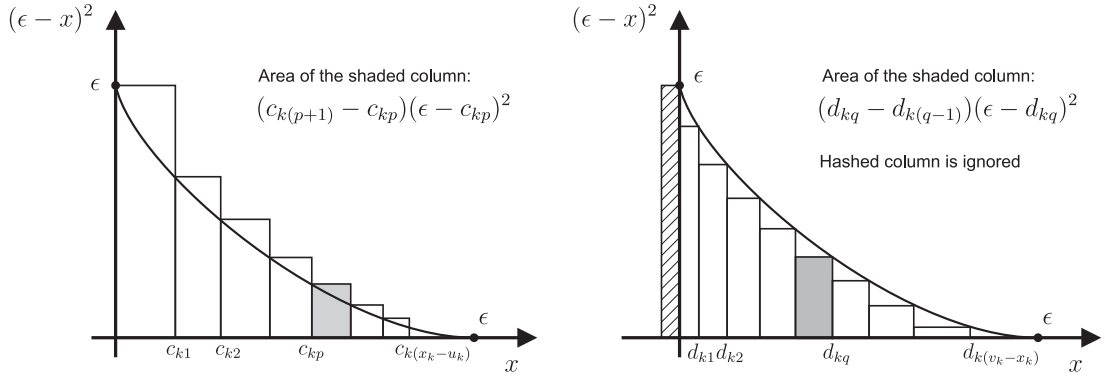

Figure 4.1: Graphics representing the Riemann sums used in the proof of Case 2 (left panel) and Case 3 (right panel).

Case 3: $P(B_k \cap C_k) \to 0$.

Recall Equation (4.1) and note that the smallest possible size of a jump in $\hat{F}_{1j}(t)$ is $1/n_{1k}$. Hence $J_{kj} \geq 1/n_{1k}$ and $D_k \geq \epsilon n_{1k} + 1$ implies that

$$\sum_{\{j:t_{kj} \in (T,U], j > \ell_k + 1\}} J_{kj} \geq \frac{\epsilon n_{1k} + 1}{n_{1k}} > \epsilon.$$

Let $v_k = \max\{v : \sum_{j=x_k+1}^{v} J_{kj} \leq \epsilon\}$. For a large enough k, Lemma 4.5 implies that

$J_{k(x_k+1)} \leq \mathcal{J}_k < \epsilon$ and thus $v_k \geq x_k + 1$. For $j \in \{x_k + 1, \ldots, v_k\}$, the monotonicity of the nondecreasing function $\hat{G}_k(t)$ implies that

$$\hat{G}_k(t_{kj}) \geq \hat{G}_k(t_{kx_k})$$

In addition, the function $\hat{F}_{1k}(t)$, a Kaplan-Meier estimate, makes a jump of size J_{ki} at each time of death t_{ki}. Therefore,

$$\hat{F}_{1k}(t_{kj}) = \hat{F}_{1k}(t_{kx_k}) + \sum_{i=x_k+1}^{j} J_{ki}.$$

Combining these two inequalities yields

$$\hat{G}_k(t_{kj}) - \hat{F}_{1k}(t_{kj}) \geq \hat{G}_k(t_{kx_k}) - \hat{F}_{1k}(t_{kx_k}) - \sum_{i=x_k+1}^{j} J_{ki} \geq \epsilon - \sum_{i=x_k+1}^{j} J_{ki} \geq 0,$$

the last inequality holding because of the choice of v_k. Using again the fact that $\mathrm{d}\hat{F}_{1k}(t)$ allocates a mass of J_{kj} to t_{kj}, we find that

$$\begin{aligned}
\int_0^U |\hat{G}_k(t) - \hat{F}_{1k}(t)|^2 \,\mathrm{d}\hat{F}_{1k}(t) &\geq \sum_{j=1}^{\ell_k} J_{kj} |\hat{G}_k(t_{kj}) - \hat{F}_{1k}(t_{kj})|^2 \\
&\geq \sum_{j=x_k}^{v_k} J_{kj} |\hat{G}_k(t_{kj}) - \hat{F}_{1k}(t_{kj})|^2 \\
&\geq \sum_{j=x_k+1}^{v_k} J_{kj} \left(\epsilon - \sum_{i=x_k+1}^{j} J_{ki} \right)^2 . \qquad (4.9) \\
&\to \int_0^\epsilon (\epsilon - x)^2 \,\mathrm{d}x = \int_0^\epsilon x^2 \,\mathrm{d}x = \frac{\epsilon^3}{3} \qquad (4.10)
\end{aligned}$$

since the summation corresponds to the Riemann sum for the integral $\int_0^\epsilon (\epsilon - x)^2 \,\mathrm{d}x$ depicted in Figure 4.1. The term $J_{kx_k}\epsilon^2$ ignored in Equation (4.9) corresponds to the hashed column. The sum converges as $k \to \infty$ because the width of the columns J_{kj} tend to zero by Lemma 4.5.

To clarify the link between the Riemann sum and the integral, consider the change of variable $q = j - x_k$ and let

$$d_{kq} = \begin{cases} 0 & q = 0 \\ \sum_{i=1}^{q} J_{k(x_k+i)} & q = 1, \ldots, v_k - x_k \end{cases}.$$

Note that $d_{kq} - d_{k(q-1)} = J_{k(x_k+q)} = J_{kj}$ and that with respect to the variable j, d_{kq} corresponds to $\sum_{i=x_k+1}^{j} J_{ki}$ when $q > 0$. We can thus rewrite Expression (4.9) as

$$\sum_{q=1}^{v_k - x_k} (d_{kq} - d_{k(q-1)})(\epsilon - d_{kq})^2 \to \int_0^{\epsilon} (\epsilon - x)^2 \, \mathrm{d}x = \int_0^{\epsilon} x^2 \, \mathrm{d}x = \frac{\epsilon^3}{3} \tag{4.11}$$

Therefore, there exists a k_0 such that

$$\int_0^U |\hat{G}_k(t) - \hat{F}_{1k}(t)|^2 \, \mathrm{d}\hat{F}_{1k}(t) > \epsilon^3/6$$

for all $k \geq k_0$, an event of probability 0 according to Lemma 4.4. We conclude that $P(B_k \cap C_k) \to 0$ as $k \to \infty$.

Combining the three cases yields the desired result. ■

Lemma 4.8 *For $T < U < \tau_{H_1}$ with $H_1^*(T) < H_1^*(U)$ and any sufficiently small $\epsilon > 0$,*

$$P\left\{ \sup_{t \leq T} |\hat{G}_k(t) - \hat{F}_{1k}(t)| > \epsilon \right\} \to 0$$

as $k \to \infty$.

Proof of Lemma 4.8. For an arbitrary $\epsilon > 0$, let

$$A_k = \left\{ \omega \in \Omega : \max_{0 \leq t \leq T} \left| \hat{F}_{1k}(t) - \hat{G}_k(t) \right| \leq \mathcal{J}_k + \max_{t \in \{t_{k1}, \ldots, t_{k(\ell_k+1)}\}} \left| \hat{F}_{1k}(t) - \hat{G}_k(t) \right| \right\}$$

$$\begin{aligned} B_k &= \left\{\omega \in \Omega : \max_{t\in\{t_{k1},\ldots,t_{k(\ell_k+1)}\}} \left|\hat{F}_{1k}(t) - \hat{G}_k(t)\right| < \frac{\epsilon}{2}\right\} \\ C_k &= \left\{\omega \in \Omega : \mathcal{J}_k < \frac{\epsilon}{2}\right\}. \end{aligned}$$

Clearly, $A_k \cap B_k \cap C_k$ implies that $\sup_{t\leq T}|\hat{G}_k(t) - \hat{F}_{1k}(t)| < \epsilon$. Therefore,

$$\begin{aligned} P\left(\sup_{t\leq T}|\hat{G}_k(t) - \hat{F}_{1k}(t)| \leq \epsilon\right) &\geq P(A_k \cap B_k \cap C_k) \\ &\geq P(A_k) + P(B_k) + P(C_k) - 2 \to 1 \end{aligned}$$

as $k \to \infty$ by Lemmas 4.5, 4.6 and 4.7. ■

Theorem 4.2 *Let $0 < T < U < \tau_{H_1}$ with $H_1^*(T) < H_1^*(U)$. For all sufficiently small $\epsilon > 0$,*

$$P\left\{\sup_{t\leq T}\left|\hat{G}_k(t) - F_1(t)\right| \leq \epsilon\right\} \to 1$$

as $k \to \infty$, i.e. $\sup_{t\leq T}\left|\hat{G}_k(t) - F_1(t)\right| \xrightarrow{P} 0$.

Proof of Theorem 1. Let us define

$$\begin{aligned} A_k &= \left\{\omega \in \Omega : \sup_{0\leq t\leq T}\left|\hat{G}_k(t) - F_1(t)\right| \leq \epsilon\right\}, \\ B_k &= \left\{\omega \in \Omega : \sup_{0\leq t\leq T}\left|\hat{G}_k(t) - \hat{F}_{1k}(t)\right| \leq \frac{\epsilon}{2}\right\}, \\ C_k &= \left\{\omega \in \Omega : \sup_{0\leq t\leq T}\left|\hat{F}_{1k}(t) - F_1(t)\right| \leq \frac{\epsilon}{2}\right\}. \end{aligned}$$

By the triangular inequality,

$$\left|\hat{G}_k(t) - F_1(t)\right| \leq \left|\hat{G}_k(t) - \hat{F}_{1k}(t)\right| + \left|\hat{F}_{1k}(t) - F_1(t)\right|$$

for any fixed t. In particular

$$\begin{aligned}\sup_{0\le t\le T}\left|\hat{G}_k(t)-F_1(t)\right| &\le \sup_{0\le t\le T}\left\{\left|\hat{G}_k(t)-\hat{F}_{1k}(t)\right|+\left|\hat{F}_{1k}(t)-F_1(t)\right|\right\}\\ &\le \sup_{0\le t\le T}\left|\hat{G}_k(t)-\hat{F}_{1k}(t)\right|+\sup_{0\le t\le T}\left|\hat{F}_{1k}(t)-F_1(t)\right|.\end{aligned}$$

Consequently, $(B_k \cap C_k) \subset A_k$ and

$$P(A_k) \ge P(B_k \cap C_k) \ge P(B_k) + P(C_k) - 1 \to 1$$

as $k \to \infty$. Lemma 4.8 implies that $P(B_k) \to 1$ and Theorem 4.1 implies that $P(C_k) \to 1$. Therefore,

$$\sup_{t\le T}\left|\hat{G}_k(t)-F_1(t)\right| \xrightarrow{P} 0\,. \qquad \blacksquare$$

The WKME converges uniformly in probability to the lifetime distribution for Population 1 in the interval $[0, T]$. The MAMSE weights, although they use data from different distributions, provide an asymptotically unbiased estimate of $F_1(t)$.

4.5 Simulations

This section presents the results of simulations performed to evaluate the finite-sample performance of the MAMSE-weighted Kaplan-Meier estimate compared to the usual Kaplan-Meier estimate. The two examples are based on real survival functions to mimic reality.

Simulations use between 10000 and 20000 repetitions. Unless otherwise stated, this number is large enough to make the standard deviation of the simulation error smaller than the last digit shown in the tables or on the figures.

4.5.1 Survival in the USA

The *Centers for Disease Control and Prevention* maintains a website which includes a section called *National Center for Health Statistics*. That section contains the decennial life tables published by the National Center for Health Statistics (1997) at the address `http://www.cdc.gov/nchs/products/pubs/pubd/lftbls/decenn/1991-89.htm` .

From this publication, we obtain the survival curves for four subgroups of the population in the United States:

- White males;
- White females;
- Males other than white;
- Females other than white.

The life tables have a resolution of one year from birth to the age of 109 years. We simulate the age at death based on these tables; the day and time of death is chosen uniformly during the year of death. The CDF of the survival time for each of the four populations is depicted in Figure 4.2.

Our purpose here is to explore the potential value of our new method rather than to study its behavior extensively under all possible conditions. Hence, we use a simple definition for the distribution of the censoring times based on the distribution of the lifetimes.

Lemma 4.9 *Let $X_1, \ldots, X_{r+1}$ be $r+1$ independent and identically distributed variables with common distribution F. Then,*

$$P\{\max(X_1, \ldots, X_r) \leq X_{r+1}\} = \frac{1}{r+1}$$

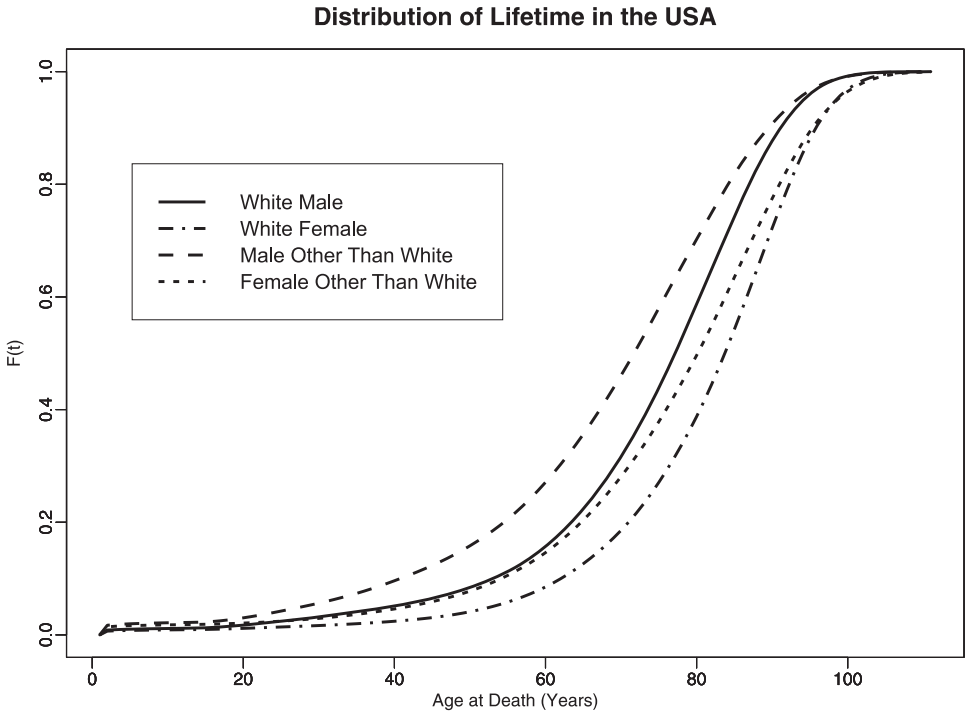

Figure 4.2: Survival functions for subgroups of the American population as taken from the life tables of the National Center for Health Statistics (1997).

Proof of Lemma 4.9.

$$P\{\max(X_1,\ldots,X_r) \le X_{r+1}\} = \int_0^\infty P\left\{\max(X_1,\ldots,X_r) \le X_{r+1} | X_{r+1} = t\right\} \mathrm{d}F(t)$$

$$= \int_0^\infty \{F(t)\}^r \,\mathrm{d}F(t) = \int_0^1 u^r \,\mathrm{d}u = \frac{1}{r+1}. \qquad \blacksquare$$

Let $X_1, ..., X_r$ be r random variables with the same distribution as the survival time of an individual. The censoring time is defined as $V = \max(X_1, \ldots, X_r)$, yielding a censoring rate of $1/(r+1)$. Throughout this section, $r = 4$ is used yielding a censoring rate of 20%. The last example however involves different rates of censoring, from 16% to 34%, obtained with $r \in \{2, 3, 4, 5\}$.

We restrict our goal to inferring the survival distribution of white males based on equal samples drawn from the four demographic groups mentioned above. We investigate if the

MAMSE-weighted Kaplan-Meier estimate is a better choice than the usual Kaplan-Meier estimate based on the data from white males.

For different values of the upper bound $U \in \{60, 70, 80, 90, 100\}$, we generate samples of equal sizes $n \in \{10, 25, 100, 1000\}$ from each of the four populations. Each scenario is repeated 20000 times.

We calculate both the weighted Kaplan-Meier estimate $\hat{F}_{\boldsymbol{\lambda}}(t)$ and the usual Kaplan-Meier estimate $\hat{F}_1(t)$. To evaluate the quality of the estimators, we compare the area that separates them from the real survival curve $F_1(t)$, more precisely, we use

$$A_{\boldsymbol{\lambda}} = \int_0^T |\hat{F}_{\boldsymbol{\lambda}}(t) - F_1(t)| \, \mathrm{d}t \quad \text{and} \quad A_1 = \int_0^T |\hat{F}_1(t) - F_1(t)| \, \mathrm{d}t.$$

Table 4.1 shows the ratio $100A_1/A_{\boldsymbol{\lambda}}$ for different choices of n, U and with $T = 55$. The interval of interest is thus $[0, T]$. Values above 100 mean that the weighted Kaplan-Meier estimate performs better.

	Ratio $100A_1/A_{\boldsymbol{\lambda}}$, with $T = 55$				
	$U = 60$	70	80	90	100
$n = 10$	114	135	142	118	100
25	137	148	149	128	101
100	135	143	140	128	102
1000	121	120	108	105	103

Table 4.1: Relative performance of the WKME as measured by $100A_1/A_{\boldsymbol{\lambda}}$. Both areas are calculated on the interval $[0, 55]$ and various upper bounds U are used to determine the weights. Samples of equal size n are taken from each of four subpopulations, then used to estimate the survival of a white male living in the USA. Each scenario is repeated 20000 times.

The weighted Kaplan-Meier estimate seems to be a better estimate under all scenarios considered. No apparent trends in the magnitude of the improvement against n and U are observed. The MAMSE criterion evaluates the dissimilarity between the populations on the interval $[0, U]$. It is thus not surprising that no clear trend is observed as U varies. Note that for the largest sample size ($n = 1000$), the advantage of the WKME seems more

modest.

The white females have the longest survival time and nearly 25% of them reach the age of 90. However, less than 3% survive long enough to celebrate their 100^{th} birthday. For $U = 100$, the samples from other populations will frequently fall short of the upper bound and be ignored, especially for small sample sizes. This might partially explain the abrupt change in performance from $U = 90$ to $U = 100$.

The improvements observed in Table 4.1 appear again in Table 4.2 where the interval on which the functions are compared varies with the upper bound U. Once again, the newly proposed weighted Kaplan-Meier estimate performs better than the classical one under all the scenarios.

	Ratio $100A_1/A_{\boldsymbol{\lambda}}$, $T = U - 5$				
	$U = 60$	70	80	90	100
$n = 10$	114	132	137	116	100
25	137	141	137	122	101
100	135	134	128	118	101
1000	121	115	103	101	101

Table 4.2: Relative performance of the WKME as measured by $100A_1/A_{\boldsymbol{\lambda}}$. Areas are calculated on the interval $[0, U - 5]$, where U is the upper bound used to determine the weights. Samples of equal size n are taken from each of four subpopulations, then used to estimate the survival of a white male living in the USA. Each scenario is repeated 20000 times.

Figure 4.3 illustrates the average weights allocated to each of the four subpopulations. Notice that the weight allocated to Population 1 is close to one when $U = 100$, especially for small sample sizes. This supports the explanations regarding the sudden drop in performance observed in Table 4.1 and 4.2: samples from other populations are often dismissed because no individual reaches 100 years of age.

Unless a mixture of the survival distributions of the other populations is identical to that of the survival distribution for Population 1, Theorem 4.2 predicts that the weight allocated to Population 1 will converge to 1. A tendency to increase towards 1 is observed for $U \in \{70, 80, 90\}$, but not for $U = 60$. It is expected that for larger sample sizes, $\bar{\lambda}_1$

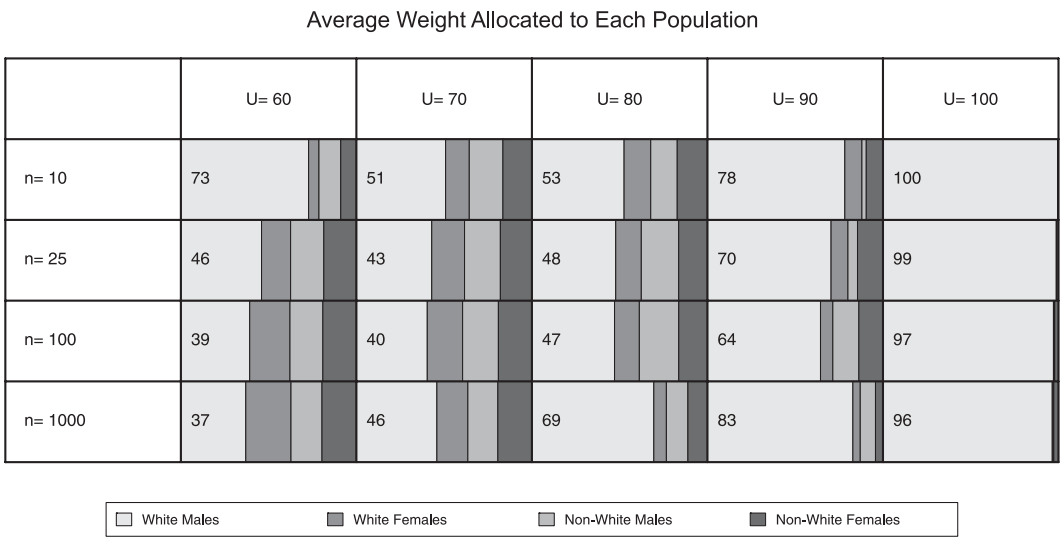

Figure 4.3: Average value of the MAMSE weights for different upper bounds U and sample sizes n. The cells' area are proportional to the average weight allocated to each population. The numbers correspond to $100\bar{\lambda}_1$ and are averaged over 20000 repetitions.

would eventually converge to 1 even in that latter case. The large weight allocated to the three other subpopulations for samples as large as 1000 should be interpreted as a sign that a mixture of these 3 distributions is extremely close to the true survival distribution in Population 1 and that does not seem unreasonable based on Figure 4.2.

Figure 4.4 depicts examples of estimates of the survival functions. While the smooth gray line shows the true distribution of the lifetime of a white male in the United States, the plain black line shows the Kaplan-Meier estimate based on a sample of size n and the dashed line corresponds to the MAMSE-weighted Kaplan-Meier estimate that we propose. The numbers on each panel correspond to $A_{\boldsymbol{\lambda}}$ and A_1 respectively with $T = 75$.

As we may expect from the good performances noticed in the previous tables, $A_{\boldsymbol{\lambda}}$ is typically smaller than A_1. Some exceptions arise: A_1 will occasionally be smaller than $A_{\boldsymbol{\lambda}}$, as it is the case for $n = 1000$ in Figure 4.4. A close look at the dashed line shows one advantage of the weighted Kaplan-Meier estimate: using more populations involves having more steps in the estimate since jumps may occur at a time of death from any of the samples

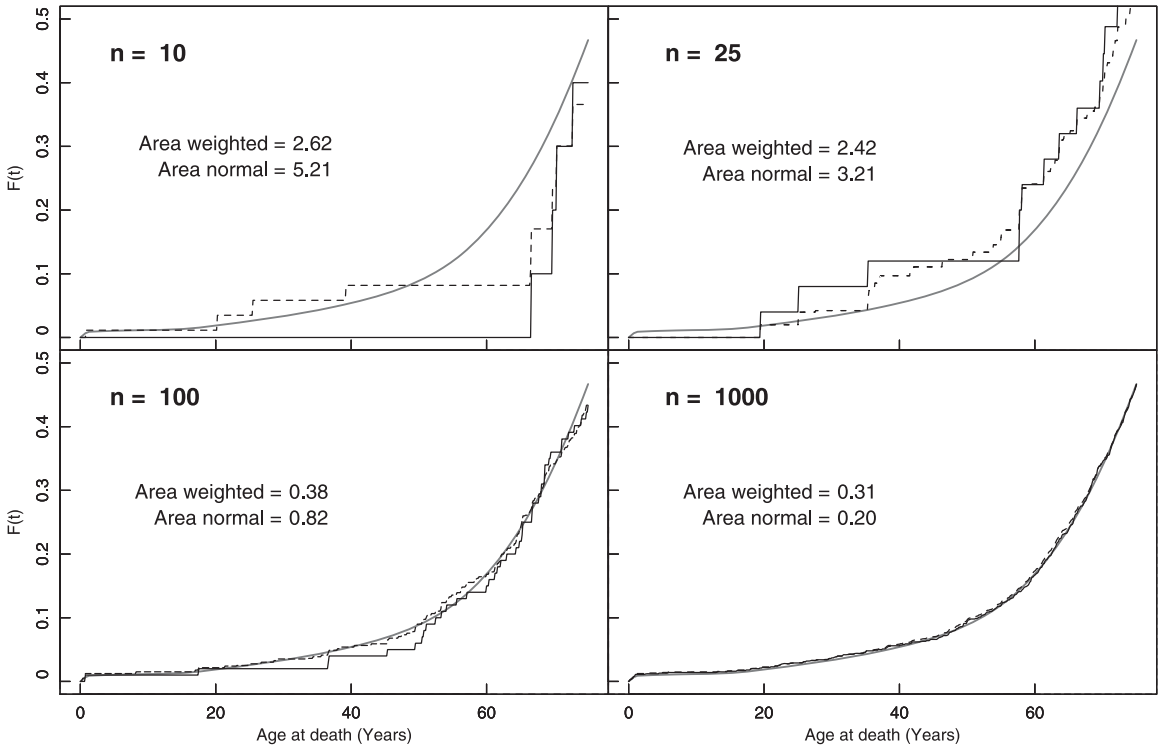

Figure 4.4: Typical examples of the weighted Kaplan-Meier estimate (dashed line) and of the usual Kaplan-Meier estimate (plain black line) for different sample sizes. Note that $U = 80$ and $T = 75$. The true distribution is depicted by a smooth gray line.

considered. This results in a smoother step function.

An estimated survival function will typically be used to answer further questions about the population of interest. Tables 4.3 and 4.4 show the performances of the weighted Kaplan-Meier in estimating $F_1(55) = 0.11976$ or $F_1^{-1}(0.10) = 52.081$. Note that we write $\hat{q}_1 = \hat{F}_1^{-1}(0.10)$ and $\hat{q}_{\boldsymbol{\lambda}} = \hat{F}_{\boldsymbol{\lambda}}^{-1}(0.10)$.

The estimates obtained by the weighted Kaplan-Meier estimate feature a smaller MSE in almost all cases. Moreover, the magnitude of the gains seems to outweigh that of the occasional losses, especially when we consider that such losses occur when n is large, not the cases where our method would be most useful.

The relative performances of the WKME for estimating a quantile is similar to the results obtained for the probability $F_1(55)$. The resemblance between the tables is not

	$100\,\frac{\text{MSE}\{\hat{F}_1(55)\}}{\text{MSE}\{\hat{F}_{\boldsymbol{\lambda}}(55)\}}$				
	$U = 60$	70	80	90	100
$n = 10$	117	151	172	134	101
25	137	159	170	149	102
100	125	142	142	134	104
1000	110	107	84	86	103

Table 4.3: Relative performance of the weighted Kaplan-Meier estimate compared to the usual Kaplan-Meier estimate for estimating $F_1(55) = 0.11976$ as measured by 100 $\text{MSE}\{\hat{F}_1(55)\}/\text{MSE}\{\hat{F}_{\boldsymbol{\lambda}}(55)\}$. Different choices of U and n are considered. Each scenario is repeated 20000 times.

	$100\ \text{MSE}(\hat{q}_1)/\text{MSE}(\hat{q}_{\boldsymbol{\lambda}})$				
	$U = 60$	70	80	90	100
$n = 10$	120	140	161	141	100
25	153	173	172	133	101
100	124	145	139	126	102
1000	119	113	86	86	106

Table 4.4: Relative performance of the weighted Kaplan-Meier estimate compared to the usual Kaplan-Meier estimate for estimating $F_1^{-1}(0.10) = 52.081$ as measured by $\text{MSE}(\hat{q}_1)/\text{MSE}(\hat{q}_{\boldsymbol{\lambda}})$. Different choices of U and n are considered. Each scenario is repeated 20000 times.

surprising since they both use the estimated curves around their 10% quantile.

For Table 4.5, we fix $U = 80$ and $T = 75$, then try different distributions for the time of censoring that yield an average fraction $p \in \{1/3, 1/4, 1/5, 1/6\}$ of censored data.

The proportion of censored data has little or no effect on the relative performance of the WKME compared to the KME. A closer look at the raw data shows that the precision of both the KME and the WKME are affected by a larger p, but it appears that the magnitude of this effect is the same for both methods. As a result, their relative performance seems unaffected by the rate of censoring.

Overall, the weighted Kaplan-Meier estimate seems to outperform the usual Kaplan-Meier estimate in almost all the cases studied. Next, we look at another population with more subgroups.

	Average of $100\lambda_1$				$100A_1/A_{\boldsymbol{\lambda}}$			
	$p = \frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$
$n = 10$	57	54	53	52	133	135	137	136
25	50	48	48	47	136	137	137	138
100	47	47	47	47	127	127	128	126
1000	68	69	69	69	102	103	103	102

Table 4.5: Average MAMSE weights for different rates of censoring p and different sample sizes n. The right-hand side of the table presents the relative performance of the WKME as measured by $100A_1/A_{\boldsymbol{\lambda}}$. Figures are averaged over 20000 repetitions and the values $U = 80$ and $T = 75$ are used.

4.5.2 Survival in Canada

Statistics Canada periodically publishes life tables for the survival of Canadians. We use the life tables published for the reference period of 2000 to 2002 that are available online at the address `http://www.statcan.ca/bsolc/english/bsolc?catno=84-537-X`.

The life tables from Statistics Canada (2006) provide the survival functions of Canadians with a resolution of one year, from birth to over 105 years. Distinct tables are provided for males and females from each of the 10 Canadian provinces. Due to its smaller population, Prince Edward Island is the only exception with a resolution of 5 years. It is excluded from our simulations for that reason.

We suppose that n males and n females from each province (except PEI) are followed and that their time of death is observed or censored. Censorship was determined the same way as before, following Lemma 4.9 with $r = 4$, which yields a censoring rate of 20%.

We perform three simulations.

1. Males: We estimate the survival function of a male in New Brunswick when data sets of males across the country are available (total 9 populations).

2. Females: We estimate the survival function of a female in New Brunswick when data sets of females across the country are available (total 9 populations).

3. Males and Females: We estimate the survival function of a female in New Brunswick

when data sets of males and females across the country are available (total 18 populations).

Figure 4.5 depicts the survival functions for the different provinces and genders. The survival curves for New Brunswick are repeated on each panel for comparison purposes since they are the target distributions; they appear as gray dotted lines.

In the following, we express the distributions in term of survival functions. These functions are obtained by simple arithmetic as $S_\bullet(t) = 1 - F_\bullet(t)$ and $\hat{S}_\bullet(t) = 1 - \hat{F}_\bullet(t)$ where $\bullet$ stands for any common index defined previously.

Throughout this section, we choose the upper bound $U = 90$ and we fix $T = 85$. Denote by $S_1(t)$ the target distribution and let $\hat{S}_1(t)$ and $\hat{S}_{\boldsymbol{\lambda}}(t)$ be the KME and the WKME respectively. The measure of performance considered will use the area between our estimate and the true survival function:

$$A_{\boldsymbol{\lambda}} = \int_0^T |\hat{S}_{\boldsymbol{\lambda}}(t) - S_1(t)| \, \mathrm{d}t \quad \text{and} \quad A_1 = \int_0^T |\hat{S}_1(t) - S_1(t)| \, \mathrm{d}t.$$

Table 4.6 shows $100A_1/A_{\boldsymbol{\lambda}}$, the ratio of these areas. Values above 100 mean that the WKME performs better than the KME. Table 4.6 also shows the average weight allocated to each population under the three scenarios considered. Simulations are performed for $n \in \{10, 25, 100\}$ and each scenario is based on 10000 repetitions. For Simulation 3, we can distinguish between the weights allocated to men and women by their font: the weights for males appear in italics.

Under all scenarios, the WKME yields a more precise estimate than the usual Kaplan-Meier estimate. The improvement is small for the males, but substantial for estimating the survival of females from New Brunswick, with a gain exceeding 50% in some cases. This might mean that the survival of the women are more similar across provinces than that of men, although it is not easily seen from Figure 4.5.

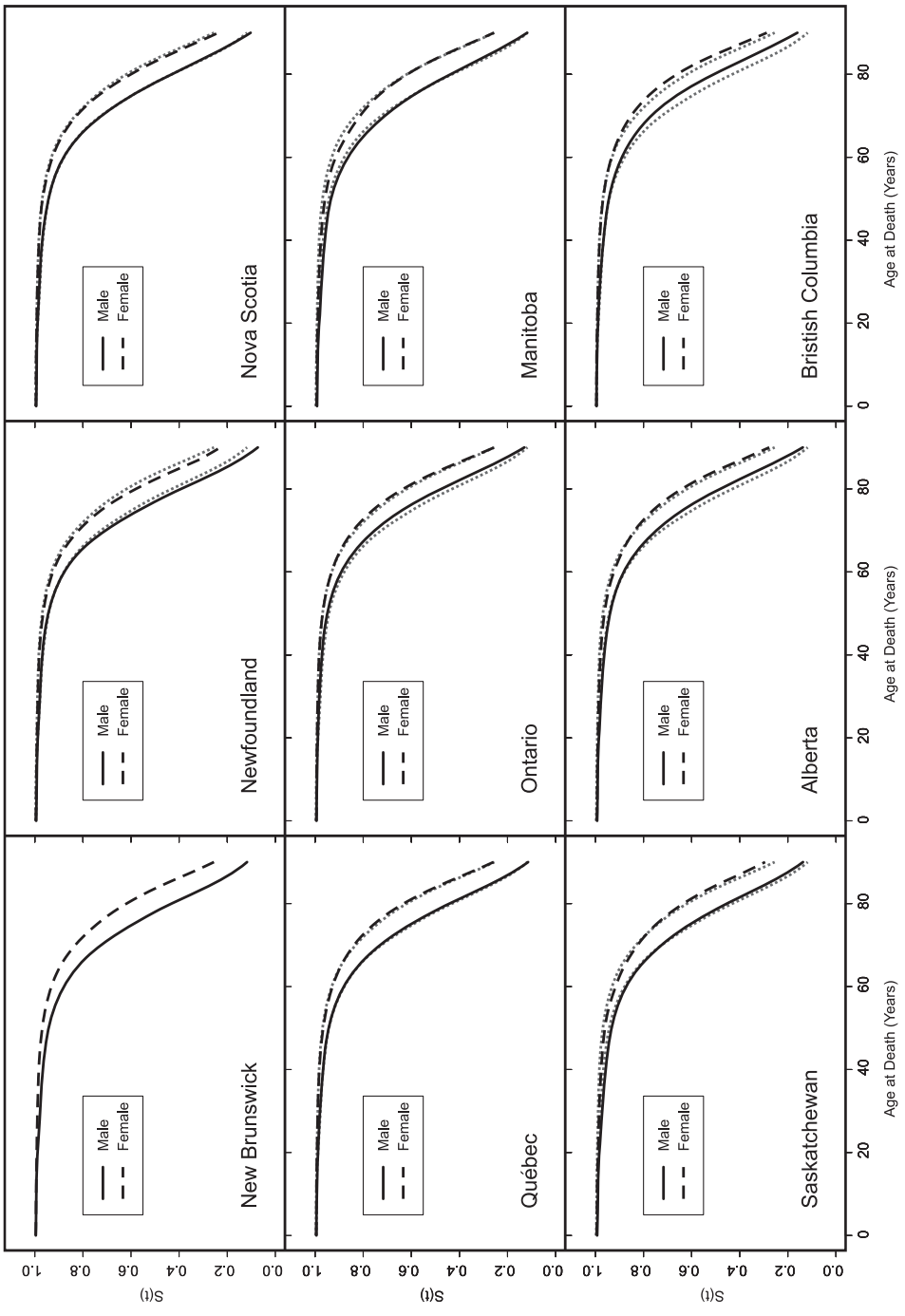

Figure 4.5: Survival functions of Canadians as taken from the life tables of Statistics Canada (2006). Survival functions are available for males and females of each province. The curves for New Brunswick are repeated as gray dotted lines on each panel to facilitate comparisons since they are the populations of interest.

	n	$100\frac{A_{\bar{\boldsymbol{\lambda}}}}{A_1}$	$1000\times$ $\bar{\lambda}_{NB}$	$\bar{\lambda}_{NF}$	$\bar{\lambda}_{NS}$	$\bar{\lambda}_{QC}$	$\bar{\lambda}_{ON}$	$\bar{\lambda}_{MB}$	$\bar{\lambda}_{SK}$	$\bar{\lambda}_{AB}$	$\bar{\lambda}_{BC}$
Males	10	101	349	82	83	80	77	85	84	81	79
	25	101	346	91	84	83	76	86	82	78	75
	100	101	350	109	87	82	71	87	79	70	65
Females	10	146	438	65	66	68	67	72	78	73	73
	25	153	376	82	79	78	76	80	77	77	75
	100	151	359	98	87	77	78	79	75	74	74
Males & Females	10	146	365	47	51	52	53	55	60	58	59
			22	*22*	*22*	*22*	*22*	*22*	*22*	*22*	*22*
	25	148	284	49	54	61	59	57	64	64	69
			25	*25*	*25*	*25*	*25*	*25*	*25*	*25*	*25*
	100	144	271	48	59	73	76	64	79	81	94
			15	*15*	*15*	*15*	*15*	*15*	*15*	*15*	*15*

Table 4.6: Relative performance of the weighted Kaplan-Meier estimate compared to the Kaplan-Meier estimate as measured by $100A_1/A_{\bar{\boldsymbol{\lambda}}}$. Average MAMSE weights are also shown, but they are multiplied by a factor of 1000 for easier reading. In the simulation with males and females, the average weights in italics refer to the male populations. Note that $U = 90$, $T = 85$ and that all figures are based on 10000 repetitions.

Note the difference in performance between the Females and the Males & Females simulations: using more populations did not seem to improve the performance and may even have worsened it. Intuitively, calculating the MAMSE weights has a cost in effective sample size. When the populations' distributions are close to each other, this cost is recovered and the quality of inference is improved. Otherwise, the performances may degrade. Figure 4.5 shows how the survival functions of men and women differ. The additional information contained in the males survival seems insufficient to recover the cost of using them. Their dissimilarity is also visible through the smaller weights allocated to the male populations on average when compared to the females in that simulation.

It is interesting to note that the weight allocated to female populations of interest seems to decrease as the sample size increases. In the Males & Females scenario, notice also the small magnitude of the weights allocated to men. As n increases, the dissimilarity between the survival distributions becomes more certain.

Figure 4.6 depicts typical estimates obtained in the simulations. The WKME is typically more precise than the usual KME, however, some exceptions arise where the area between the KME and the true distribution is smaller than $A_{\boldsymbol{\lambda}}$. Notice once again that using the weighted Kaplan-Meier estimate produces a smoother curve than the KME since the function has more jumps.

The MAMSE weights allow us to build a weighted Kaplan-Meier estimate that exploits information from similar populations. For the problem of estimating the survival of Canadians from New Brunswick, the weighted Kaplan-Meier estimate features better performance than the usual Kaplan-Meier estimate in all scenarios considered. The WKME may be an alternative to consider when data from similar populations are easily available, or when one is interested in a subpopulation of a larger group as is the case for this simulation.

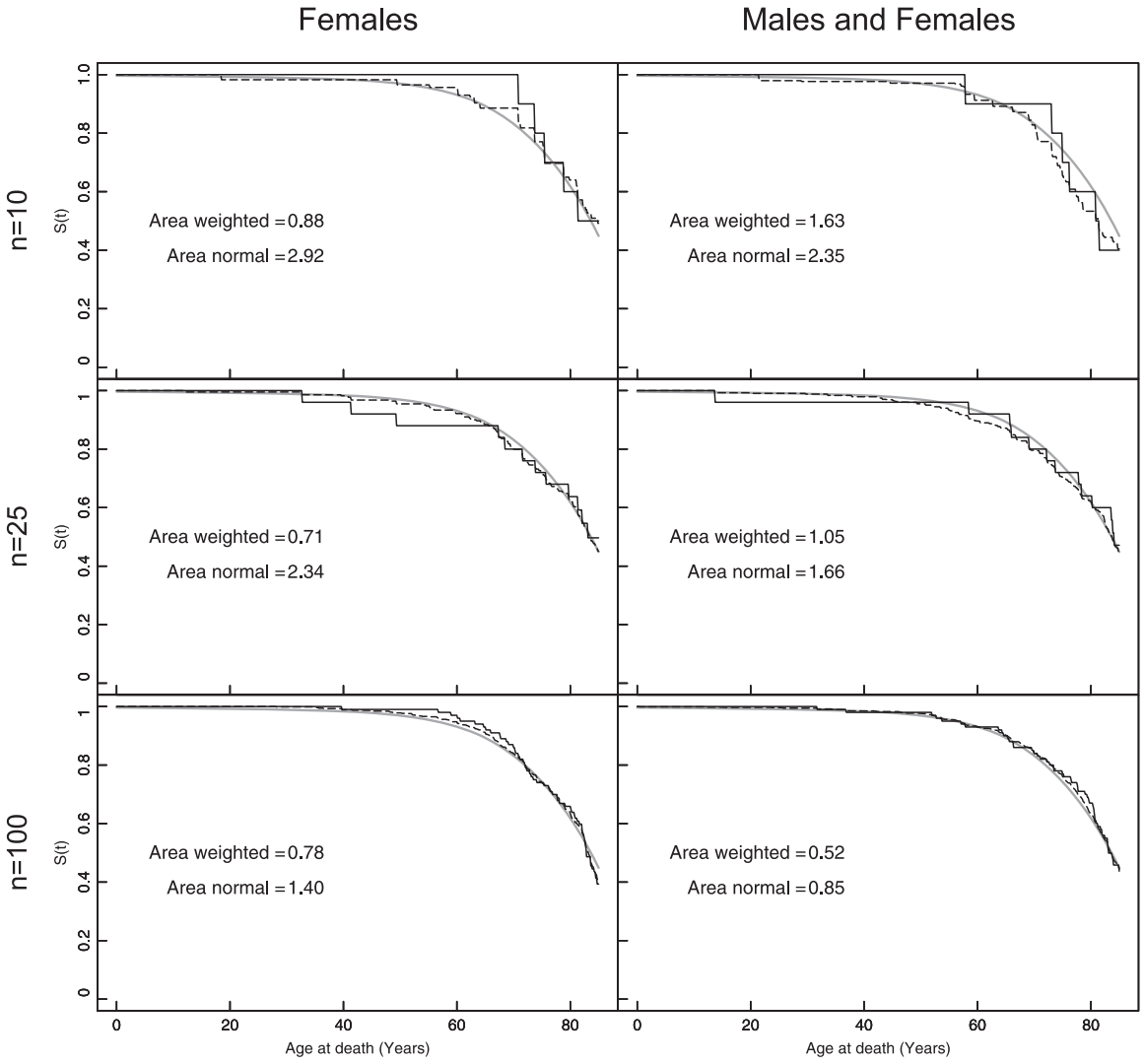

Figure 4.6: Typical examples of estimates under different scenarios. Note the increased number of steps of the weighted Kaplan-Meier estimate (dashed line) which makes it smoother than the usual Kaplan-Meier estimate (plain black line). The true distribution appears as a smooth gray curve.

Chapter 5

MAMSE Weights for Copulas

Copulas are distribution functions with uniform margins, but most importantly, they are a tool for expressing the dependence structure of multivariate distributions.

In this chapter, the MAMSE weights are extended to multivariate data using copulas. The empirical copula, a nonparametric estimate of the copula based on the ranks of the data, is used for that purpose.

After reviewing the theory related to copulas, we define the MAMSE weights using the empirical copula. The nonparametric weights obtained are then used to define a mixture of empirical copulas which yields weighted coefficients of correlation. The weighted pseudo-likelihood, an extension of the pseudo-likelihood, is proposed and shown to produce consistent estimates when used in conjunction with the MAMSE weights. Simulations evaluate the performances of the maximum weighted pseudo-likelihood estimate and of the weighted coefficients of correlation.

5.1 Review of Copulas

Let $\mathbf{X}$ be a p-dimensional vector with continuous distribution $H(\mathbf{x})$ and continuous marginal distributions $G_1(x_1), \ldots, G_p(x_p)$. The dependence between the elements of $\mathbf{X}$ is best described by its underlying copula, a cumulative distribution function with uniform marginals. Sklar (1959) showed that all multivariate distributions admit the representation

$$H\left([x_1, \ldots, x_p]^\mathsf{T}\right) = C\left\{G_1(x_1), \ldots, G_p(x_p)\right\}$$

where C is a CDF with uniform margins called a copula. Any continuous distribution H is associated with a unique copula C.

If $h_1, \ldots, h_p$ are one-to-one increasing functions, then the unique copula associated with the distribution of $\mathbf{Y} = [h_1(X_1), \ldots, h_p(X_p)]^\mathsf{T}$ is the same as the copula underlying the distribution of $\mathbf{X} = [X_1, \ldots, X_p]^\mathsf{T}$.

Coefficients of Correlation

The expression "coefficient of correlation" typically refers to an empirical estimate of the dependence between two variables. These statistics however estimate population values that are also called "coefficients of correlation". For instance, the Pearson correlation of a bivariate distribution $H(x, y)$ with marginal distributions $F(x)$ and $G(y)$ is given by

$$r = \frac{1}{\sigma_x \sigma_y} \int \int (x - \mu_x)(y - \mu_y) \, \mathrm{d}H(x, y)$$

where μ_x and σ_x^2 are the mean and the variance of $F(x)$ and similarly for $G(y)$.

Pearson's correlation is a parameter of the multivariate normal model and it is an efficient measure of dependence under that model. However, Pearson's correlation is not a good measure of dependence in a general setting. In particular, the range of values for Pearson's coefficient is limited by the distribution of the margins F and G. How can one interpret a correlation of 0.80, say, when it could be the maximal value for r given the marginal distributions at hand? Moreover, a monotone transformation of the variables will typically affect r.

Better measures of correlation are invariant to monotone transformations of the data. Spearman's ρ and Kendall's τ are well-known such coefficients. Let $C(u, v)$ denote the unique copula associated with $H(x, y)$. Then Table 5.1 gives the population value of different coefficients of correlation, including ρ and τ. These coefficients depend only on the copula C, hence they are invariant to a monotone increasing transformation of the margins. The

empirical estimates of the coefficients of correlation in Table 5.1 are based on ranks; they are presented in Section 5.2

Usual Name	Population Value
Spearman	$\rho = 12 \int\int uv \, \mathrm{d}C(u,v) - 3$
Kendall	$\tau = 4 \int\int C(u,v) \, \mathrm{d}C(u,v) - 1$
Gini	$\gamma = \int\int \lvert u+v-1 \rvert - \lvert u-v \rvert \, \mathrm{d}C(u,v)$
Blomqvist	$\beta = 4\, C\left(\frac{1}{2}, \frac{1}{2}\right) - 1$
Blest	$\nu = 2 - 12 \int\int (1-u)^2 v \, \mathrm{d}C(u,v)$
Symmetrized Blest	$\xi = -2 + 12 \int\int uv(4-u-v) \, \mathrm{d}C(u,v)$

Table 5.1: Population value of different coefficients of correlation.

More details on Spearman's ρ, Kendall's τ, Gini's γ and Blomqvist's β can be found in Nelsen (1999). Blest's coefficients were first introduced by Blest (2000), then further developed by Genest & Plante (2003). Pinto da Costa & Soares (2005) studied the same coefficients of correlation and rediscovered independently some of the results published by Genest & Plante (2003).

Families of Copulas

A number of families of copulas have been studied in the literature. The books of Joe (1997), Nelsen (1999) and Cherubini et al. (2004) present the most common ones, and discuss methods for constructing copulas. Note in particular that building a copula in more than 2 dimensions is not always an easy endeavor as not all lower dimensional copulas are compatible

with each other.

Let us introduce the families of copulas that will be used for simulations in Section 5.9.

Normal Copula

The copula underlying the Normal distribution does not have a closed form but can be expressed as follow. Let $H_\Sigma(\mathbf{x})$ be the CDF of a multivariate Normal with mean $\mathbf{0}$ and covariance matrix Σ. In addition, suppose that the diagonal of Σ contains only ones, hence Σ is a correlation matrix and the margins of H_Σ are $\mathcal{N}(0,1)$. If $\Phi(x) = P(Z \leq x)$ for $Z \sim \mathcal{N}(0,1)$, then

$$C_\Sigma(\mathbf{u}) = H_\Sigma\left\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)\right\}$$

is the copula underlying H_Σ. In p dimensions, the Normal copula depends on $p(p-1)/2$ parameters which make Σ positive definite. In 2 dimensions, it depends on one parameter, $r \in (-1,1)$. The limiting cases where $r = \pm 1$ corresponds to perfect linear concordance or discordance. Independence is given by $r = 0$.

Under the bivariate Normal model, the population values of ρ and τ are

$$\rho = \frac{6}{\pi} \arcsin\left(\frac{r}{2}\right) \quad \text{and} \quad \tau = \frac{2}{\pi} \arcsin(r).$$

Simulating the Normal copula can be done by generating a multivariate Normal, then applying the inverse transformation Φ^{-1} to its marginal values.

Farlie-Gumbel-Morgenstern Copula

The Farlie-Gumbel-Morgenstern (FGM) copula (see Nelsen (1999) page 68) is parametrized by $\theta \in [-1,1]$ and written

$$C_\theta(u,v) = uv + \theta uv(1-u)(1-v).$$

Its simple closed form expression is convenient, but it does not span a great range of dependence. In particular, $\rho = \theta/3$ under that model, hence the absolute value of Spearman's ρ is at most $1/3$.

With the simplicity of its closed-form expression, the FGM copula is often used for illustrative purposes in the literature. However, it is not a very flexible model since it features a limited range of dependence. Hence, it is rarely used as an inferential model in practice.

Simulating a datum (U, V) from a FGM copula can be done by generating a first uniform variate (V), then transforming a second uniform variate (W) by inverting the CDF of the conditional distribution of U given V. This yields a quadratic equation whose root of interest is

$$U = \frac{-\{1 + \alpha(1-2V)\} + \sqrt{\{1 + \alpha(1-2V)\}^2 - 4\alpha(1-2V)W}}{-2\alpha(1-2V)}.$$

Genest & MacKay (1986) present a recipe to build numerous families of copulas that are called Archimedean copulas. Many important families of copulas are Archimedean. Chapter 4 of Nelsen (1999) is devoted to the construction of such copulas and to the study of their properties.

The next three families of copulas are particular cases of Archimedean copulas. An algorithm to produce pseudo-random variates from these distributions is given by Genest & MacKay (1986).

Clayton Copula

The Clayton copula was introduced by Clayton (1978) and is often used in survival analysis (see Oakes (1986) for instance). The bivariate Clayton family is parametrized by $\theta \in [-1, 0) \cup (0, \infty)$, the limiting case where $\theta \to 0$ corresponding to independence. The

expression for the copula is

$$C_\theta(u,v) = \max\left\{\left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}, 0\right\}.$$

Frank Copula

This family was first discussed by Frank (1979) and is indexed by one real parameter, $\theta \in (-\infty, 0) \cup (0, \infty)$. The limiting case where $\theta \to 0$ corresponds to independence. Its CDF is

$$C_\theta(u,v) = -\frac{1}{\theta}\log\left\{1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right\}.$$

The Frank copula is the only radially symmetric Archimedean copula. Its shape is akin to the Normal copula.

Gumbel-Hougaard

The Gumbel-Hougaard family is one of three copulas that can be a limit distribution for extremes and is hence used to model data of that type. Indexed by $\theta \in [1, \infty)$ it is written

$$C_\theta(u,v) = \exp\left[-\left\{(-\log u)^\theta + (-\log v)^\theta\right\}^{1/\theta}\right]$$

and the choice $\theta = 1$ corresponds to independence.

The families of copulas presented in this section are absolutely continuous and hence admit a density function, as long as we omit limiting cases of perfect concordance or discordance where the distributions collapse to a line.

The existence of a density function is less clear for the Clayton, Frank and Gumbel-Hougaard copulas. Genest & MacKay (1986) show that all bivariate Archimedean copulas can be factorized as a mixture of two components, one absolutely continuous, the second singular on a line that they describe explicitly. From their result, it is straightforward to

verify that the singular part of the three Archimedean copulas presented in this section has mass 0, hence that they are absolutely continuous.

5.2 Notation and Empirical Estimation

Suppose that p-dimensional data are available from m different populations believed to have similar dependence structures (i.e. similar copulas). For any fixed k, we observe n_{ik} data points from Population i. Explicitly,

$$\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_{ik}} \overset{iid}{\sim} F_i$$

are observed where $\mathbf{X}_{ij} = [X_{ij1}, \ldots, X_{ijp}]^\mathsf{T}$ is a vector in p dimensions. By Sklar's (1959) Theorem, there exists a unique copula underlying the distribution F_i; we denote it by $C_i(\mathbf{u})$, $\mathbf{u} = [u_1, \ldots, u_p]^\mathsf{T}$ being a vector in $[0,1]^p$. That unique copula is a cumulative distribution function defined on the unit cube such that

$$F_i(\mathbf{x}) = C_i\left\{G_{i1}(x_1), \ldots, G_{ip}(x_p)\right\}$$

where $G_{i1}, \ldots, G_{ip}$ are the marginal distributions of F_i.

Let $\mathbf{R}_{ij}^k = [R_{ij1}^k, \ldots, R_{ijp}^k]$ be the ranks associated with the vectors $\mathbf{X}_{ij}$, $j = 1, \ldots, n_{ik}$. For fixed i and ℓ, the list of values $X_{i1\ell}, \ldots, X_{in_{ik}\ell}$ is sorted and $R_{ij\ell}^k$ is the rank of $X_{ij\ell}$ in that list. Assume that the $\{F_i\}$ are continuous, and hence ties cannot occur with probability 1.

The empirical copula, defined on the ranks of a sample, is

$$\hat{C}_{ik}(\mathbf{u}) = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \prod_{\ell=1}^{p} \mathbb{1}\left(\frac{R_{ij\ell}^k}{n_{ik}} \leq u_\ell\right)$$

for $\mathbf{u} = [u_1, \ldots, u_p]^\mathsf{T}$. The indicator variable $\mathbb{1}(\bullet)$ is equal to one if all the elements of its argument are true and equal to 0 otherwise. The empirical copula puts a weight of $1/n_{ik}$

on the points of the grid

$$\left\{\frac{1}{n_{ik}}, \frac{2}{n_{ik}}, \ldots, 1\right\} \times \cdots \times \left\{\frac{1}{n_{ik}}, \frac{2}{n_{ik}}, \ldots, 1\right\}$$

corresponding to an observed combination of ranks. There is exactly one such point in every $(p-1)$-dimensional slice of the grid (rows and columns in 2 dimensions). Consequently, the univariate marginals of the empirical copula $\hat{C}_{ik}$ are uniformly distributed on the points $\{1/n_{ik}, 2/n_{ik}, \ldots, 1\}$.

Suppose the $\{n_{ik}\}$ are strictly increasing with k. Deheuvels (1979) shows that

$$\sqrt{\frac{n_{ik}}{\log\log n_{ik}}} \sup_{\mathbf{u}\in[0,1]^p} |\hat{C}_{ik}(\mathbf{u}) - C_i(\mathbf{u})| \to 0$$

almost surely as $k \to \infty$. Fermanian et al. (2004) show that $\sqrt{n_{ik}}\{\hat{C}_{ik}(\mathbf{u}) - C_i(\mathbf{u})\}$ converges weakly to a Brownian sheet with mean zero and a variance that depends on C_i and its partial first-order derivatives. Although they hold for an arbitrary number of dimensions, the results of Fermanian et al. (2004) are presented for bivariate copulas only. Tsukahara (2005) credits Fermanian et al. (2004) for the discovery and expresses the same results in p dimensions.

Remark 5.1 *Let $\mathcal{U}_i(\mathbf{u})$ be a p-dimensional centered Gaussian random field with covariance function*

$$C_i(\mathbf{u} \wedge \mathbf{v}) - C_i(\mathbf{u})C_i(\mathbf{v}),$$

where $\wedge$ is the component-wise minimum. Such a random field is called a p-dimensional pinned C_i-Brownian sheet.

Theorem 5.1 (Tsukahara (2005)) *Assume that $C_i(\mathbf{u})$ is differentiable with continuous partial derivatives $\partial C_i(\mathbf{u})/\partial u_\ell$ for $\ell = 1, \ldots, p$ and let $[\mathbf{1}, u_\ell, \mathbf{1}]^\mathsf{T}$ represent a vector of ones,*

except for the ℓ^{th} element who is equal to the ℓ^{th} element of $\mathbf{u}$. *Then the random variable*

$$\sqrt{n_{ik}}\{\hat{C}_{ik}(\mathbf{u}) - C_i(\mathbf{u})\}$$

converges weakly to the random field

$$\mathcal{U}_i(\mathbf{u}) - \sum_{\ell=1}^{p} \left\{ \frac{\partial}{\partial u_\ell} C_i(\mathbf{u}) \right\} \mathcal{U}_i \left([\mathbf{1}, u_\ell, \mathbf{1}]^\mathsf{T} \right)$$

as $k \to \infty$.

Remark 5.2 *Let* $\mathbf{u} \in [0,1]^p$ *and* $\mathbf{v}_\ell = [\mathbf{1}, u_\ell, \mathbf{1}]^\mathsf{T}$ *for some* $\ell \in \{1, \ldots, p\}$. *For the Brownian sheet defined in the previous remark and* $1 \leq l < \ell \leq p$, *we have:*

$$\begin{aligned}
\operatorname{var}\{\mathcal{U}_i(\mathbf{u})\} &= C_i(\mathbf{u}) - C_i(\mathbf{u})^2 = C_i(\mathbf{u})\{1 - C_i(\mathbf{u})\} \\
\operatorname{var}\{\mathcal{U}_i(\mathbf{v}_\ell)\} &= C_i(\mathbf{v}_\ell)\{1 - C_i(\mathbf{v}_\ell)\} = u_\ell(1 - u_\ell) \\
\operatorname{cov}\{\mathcal{U}_i(\mathbf{u}), \mathcal{U}_i(\mathbf{v}_\ell)\} &= C_i(\mathbf{u}) - C_i(\mathbf{u}) u_\ell = C_i(\mathbf{u})(1 - u_\ell) \\
\operatorname{cov}\{\mathcal{U}_i(\mathbf{v}_l), \mathcal{U}_i(\mathbf{v}_\ell)\} &= C_i\left([\mathbf{1}, u_l, \mathbf{1}, u_\ell, \mathbf{1}]^\mathsf{T} \right) - u_l u_\ell
\end{aligned}$$

where $[\mathbf{1}, u_l, \mathbf{1}, u_\ell, \mathbf{1}]^\mathsf{T}$ *is a vector of ones, except for the elements* l *and* ℓ *that are equal to the* l^{th} *and* ℓ^{th} *element of* $\mathbf{u}$ *respectively.*

To define the MAMSE weights in Section 5.3, we need an estimate of the asymptotic variance of the empirical copula.

Remark 5.3 *From Theorem 5.1, we have that*

$$\begin{aligned}
&\operatorname{var}\left[\sqrt{n_{ik}}\{\hat{C}_{ik}(\mathbf{u}) - C_i(\mathbf{u})\} \right] \to \operatorname{var}\left\{ \mathcal{U}_i(\mathbf{u}) - \sum_{\ell=1}^{p} \frac{\partial}{\partial u_\ell} C_i(\mathbf{u}) \mathcal{U}_i(\mathbf{v}_\ell) \right\} \\
&= \operatorname{var}\{\mathcal{U}_i(\mathbf{u})\} + 2 \sum_{1 \leq l < \ell \leq p} \left\{ \frac{\partial}{\partial u_\ell} C_i(\mathbf{u}) \right\} \left\{ \frac{\partial}{\partial u_l} C_i(\mathbf{u}) \right\} \operatorname{cov}\{\mathcal{U}_i(\mathbf{v}_l), \mathcal{U}_i(\mathbf{v}_\ell)\} \\
&\quad + \sum_{\ell=1}^{p} \left\{ \frac{\partial}{\partial u_\ell} C_i(\mathbf{u}) \right\}^2 \operatorname{var}\{\mathcal{U}_i(\mathbf{v}_\ell)\} - 2 \sum_{\ell=1}^{p} \left\{ \frac{\partial}{\partial u_\ell} C_i(\mathbf{u}) \right\} \operatorname{cov}\{\mathcal{U}_i(\mathbf{u}), \mathcal{U}_i(\mathbf{v}_\ell)\}
\end{aligned}$$

The expressions for the variances and covariances from Remark 5.2 can be substituted in the expression above. Note that the only term that does not depend on derivatives of $C_i(\mathbf{u})$ *is*

$$\text{var}\,\{\mathcal{U}_i(\mathbf{u})\} = C_i(\mathbf{u})\{1 - C_i(\mathbf{u})\}.$$

Coefficients of Correlation Based on Ranks

The measures of dependence presented in Table 5.1 can be estimated from a sample by using ranks. Their classical estimates appear in Table 5.2. Note that we use a simplified notation where for a fixed population i and for a fixed k, we write $n = n_{ik}$, $\hat{C}(u_1, u_2) \equiv \hat{C}_{ik}(\mathbf{u})$ and $(R_j, S_j) = \mathbf{R}_{ij}^k$. That simplified notation is the most commonly seen in the literature.

Usual Name	Empirical Estimate
Spearman	$\hat{\rho}_n = -3\,\frac{n+1}{n-1} + \frac{12}{n(n+1)(n-1)}\sum_{i=1}^n R_i S_i$
Kendall	$\hat{\tau}_n = \binom{n}{2}^{-1}\sum_{1\le i<j\le n} sign(R_i - R_j)sign(S_i - S_j)$
Gini	$\hat{\gamma}_n = \frac{1}{\lfloor n^2/2\rfloor}\sum_{i=1}^n \lvert R_i + S_i - n - 1\rvert - \lvert R_i - S_i\rvert$
Blomqvist	$\hat{\beta}_n = 4\hat{C}\left(\frac{1}{2}, \frac{1}{2}\right) - 1$
Blest	$\hat{\nu}_n = \frac{2n+1}{n-1} - \frac{12}{n(n+1)^2(n-1)}\sum_{i=1}^n (n+1-R_i)^2 S_i$
Symmetrized Blest	$\hat{\xi}_n = -\frac{4n+5}{n-1} + \frac{6}{n(n+1)(n-1)}\sum_{i=1}^n R_i S_i\left(4 - \frac{R_i+S_i}{n+1}\right)$

Table 5.2: Empirical estimates of different coefficients of correlation.

Rescaled Ranks

The empirical copula allocates a non-null weight to points that are on the boundary of $[0,1]^p$ even though that set has probability 0 under a continuous copula. In some applications, we need to evaluate functions that are well defined on $(0,1)^p$, but that may have asymptotes

close to the boundaries of the p-dimensional cube.

To avoid such problems, one can use a rescaled empirical copula where the mass of $\mathbf{R}_{ij}^k/n_{ik}$ is rather allocated to

$$\mathbf{Y}_{ij}^{k*} = \frac{\mathbf{R}_{ij}^k - \frac{1}{2}}{n_{ik}} \qquad \text{or} \qquad \mathbf{Y}_{ij}^{k} = \frac{\mathbf{R}_{ij}^k}{n_{ik}+1}.$$

Hence, the rescaled empirical copula allocates an equal mass to some points of the grid

$$\left\{\frac{0.5}{n_{ik}}, \ldots, \frac{n_{ik}-0.5}{n_{ik}}\right\} \times \cdots \times \left\{\frac{0.5}{n_{ik}}, \ldots, \frac{n_{ik}-0.5}{n_{ik}}\right\}$$

or

$$\left\{\frac{1}{n_{ik}+1}, \ldots, \frac{n_{ik}}{n_{ik}+1}\right\} \times \cdots \times \left\{\frac{1}{n_{ik}+1}, \ldots, \frac{n_{ik}}{n_{ik}+1}\right\},$$

where the function of interest is well-defined.

Remark 5.4 *Let $\hat{C}_{ik}^*$ denote a rescaled empirical copula. The asymptotic behavior of $\hat{C}_{ik}^*$ is typically identical to that of $\hat{C}_{ik}$ since*

$$\sup_{\mathbf{u}\in[0,1]^p} |\hat{C}_{ik}^*(\mathbf{u}) - \hat{C}_{ik}(\mathbf{u})| \leq \frac{1}{n_{ik}} \to 0$$

as $k \to \infty$.

The Pseudo-Likelihood

Methods have been proposed for fitting a family of copulas to the data. A review of such methods can be found in Cherubini et al. (2004). Let us consider the pseudo-likelihood originally proposed by Genest et al. (1995).

Suppose that the family of copulas $C(\mathbf{u}|\theta)$, $\theta \in \Theta$ admits a density function $c(\mathbf{u}|\theta)$. To fit that family of distributions to the data from Population i, Genest et al. (1995) suggest

maximizing the pseudo-likelihood,

$$L(\theta) = \prod_{j=1}^{n_{ik}} c\left(\mathbf{Y}_{ij}^k \,\middle|\, \theta\right),$$

where $\mathbf{Y}_{ij}^k$ are the rescaled ranks described above. The resulting maximum pseudo-likelihood estimate is a consistent estimate of the true parameter θ_0 for which $C(\mathbf{u}|\theta_0) \equiv C_i(\mathbf{u})$.

Alternatively, the log-pseudo-likelihood can be written

$$\ell(\theta) = \sum_{j=1}^{n_{ik}} \log c\left(\mathbf{Y}_{ij}^k \,\middle|\, \theta\right) = \int \log c(\mathbf{u}|\theta)\, \mathrm{d}\hat{C}_{ik}^*(\mathbf{u})$$

where $\hat{C}_{ik}^*(\mathbf{u})$ denotes the empirical copula defined with the rescaled ranks $\mathbf{Y}_{ij}^k$.

5.3 Definition of the MAMSE Weights for Copulas

Univariate MAMSE weights are designed to yield a mixture of distributions close to the target distribution F_1, but less variable than its empirical distribution function. In the context of copulas, an extended version of the MAMSE weights can aim at choosing a mixture of the empirical copulas,

$$\hat{C}_{\boldsymbol{\lambda}k}(\mathbf{u}) = \sum_{i=1}^{m} \lambda_i \hat{C}_{ik}(\mathbf{u}) \tag{5.1}$$

with $\lambda_i \geq 0$ and $\mathbf{1}^\mathsf{T}\boldsymbol{\lambda} = 1$, that is close to C_1, but less variable than $\hat{C}_{1k}$. Let us define

$$P_k(\boldsymbol{\lambda}) = \int \left[|\hat{C}_{1k}(\mathbf{u}) - \hat{C}_{\boldsymbol{\lambda}k}(\mathbf{u})|^2 + \sum_{i=1}^{m} \lambda_i^2 \widehat{\mathrm{var}}\{\hat{C}_{ik}(\mathbf{u})\} \right] \mathrm{d}M_k(\mathbf{u}) \tag{5.2}$$

where M_k is a discrete probability measure allocating a weight of $1/n_{1k}^p$ to every point of the grid

$$\mathcal{G}_k = \left\{ \frac{1}{n_{1k}}, \frac{2}{n_{1k}}, \ldots, 1 \right\} \times \cdots \times \left\{ \frac{1}{n_{1k}}, \frac{2}{n_{1k}}, \ldots, 1 \right\}.$$

Remark 5.3 provides an approximation to the variance of $\sqrt{n_{ik}}\,\hat{C}_{ik}(\mathbf{u})$. However, the expression depends on the derivatives of $C_i(\mathbf{u})$, the true unknown and unknowable copula. To estimate these derivatives, we could use the data and either assume a parametric model for each population, or use smoothing techniques. Either of these choices involves many degrees of freedom in the selection of families of distributions or basis functions for instance.

Our goal is to show that the concept of the MAMSE weights has the potential to improve the quality of the inference while at the same time keeping their computation feasible. Thus at this stage we choose on the basis of Remark 5.3, the simple compromise approximation,

$$\widehat{\mathrm{var}}\{\hat{C}_{ik}(\mathbf{u})\} \approx \widetilde{\mathrm{var}}\{\hat{C}_{ik}(\mathbf{u})\} = \frac{1}{n_{ik}}\hat{C}_{ik}(\mathbf{u})\{1-\hat{C}_{ik}(\mathbf{u})\}, \tag{5.3}$$

which corresponds to the only term in Remark 5.3 that does not involve a derivative. We recognize that incorporating other terms in that approximation may improve the performance of the weights but leave that and the investigation of the computational feasibility of doing so to future work. As we will show, the choice made above does demonstrate the potential value of the MAMSE weights in this context.

Note that the theoretical results in the following sections hold with a different penalty term as long as the property expressed in Lemma 5.1 is preserved, i.e. as long as

$$\int \widehat{\mathrm{var}}\{\hat{C}_{1k}(\mathbf{u})\}\,\mathrm{d}M_k(\mathbf{u}) \to 0$$

as $k \to \infty$. The property stated in Theorem 5.2 should also hold to insure the proper convergence of the algorithm calculating the MAMSE weights.

The value of $\boldsymbol{\lambda}$ minimizing the objective function $P_k(\boldsymbol{\lambda})$ defined in (5.2) with the substitution (5.3) are called the MAMSE weights and denoted $\boldsymbol{\lambda}_k(\omega) = [\lambda_{1k}(\omega), \ldots, \lambda_{mk}(\omega)]^{\mathsf{T}}$. The weighted empirical copula obtained using MAMSE weights will be written

$$\hat{\mathcal{C}}_k(\mathbf{u}) = \sum_{i=1}^{m} \lambda_{ik}(\omega)\hat{C}_{ik}(\mathbf{u}). \tag{5.4}$$

5.4 Computing the MAMSE Weights for Copulas

The algorithm proposed in Section 2.4 applies to the MAMSE weights for copulas.

All copulas are constrained to be on the $[0,1]^p$ cube and have uniform marginals. For that reason, there is no need for preprocessing in this case as the domain of all distributions overlap at all times.

To prove the convergence of the algorithm, it is sufficient to show that Assumption 2.1 is satisfied when the MAMSE weights are defined as in Equation (5.2).

Theorem 5.2 *Assumption 2.1 is satisfied for the definition of MAMSE weights suggested in Section 5.3, that is*

$$\int \widetilde{\mathrm{var}}\{\hat{C}_{ik}(\mathbf{u})\}\,\mathrm{d}M_k(\mathbf{u}) > 0$$

for $i \in \{1,\ldots,m\}$.

Proof of Theorem 5.2. Consider the term

$$\int \widetilde{\mathrm{var}}\{\hat{C}_{ik}(\mathbf{u})\}\,\mathrm{d}M_k(\mathbf{u}) = \int \frac{1}{n_{ik}}\,\hat{C}_{ik}(\mathbf{u})\{1-\hat{C}_{ik}(\mathbf{u})\}\,\mathrm{d}M_k(\mathbf{u})$$

for some $i = 1,\ldots,m$. Note that the measure $M_k(\mathbf{u})$ gives a weight $1/n_{1k}^p$ to each point of the grid $\mathcal{G}_k$. The empirical copula $\hat{C}_i(\mathbf{u})$ is greater than 0 and less than 1 for all points of the grid $\mathcal{G}_k$, except for the point $\mathbf{1}$ and possibly for a few "slices" (rows and columns in 2 dimensions) of points close to the axes when $n_{ik} < n_{1k}$. In all cases, $0 < \hat{C}_i(\mathbf{u}) < 1$ for many $\mathbf{u} \in \mathcal{G}_k$, meaning that $\hat{C}_{ik}(\mathbf{u})\{1-\hat{C}_{ik}(\mathbf{u})\} > 0$ for those $\mathbf{u}$. Consequently,

$$\frac{1}{n_{ik}}\int \hat{C}_{ik}(\mathbf{u})\{1-\hat{C}_{ik}(\mathbf{u})\}\,\mathrm{d}M_k(\mathbf{u}) > \frac{1}{n_{1k}^p n_{ik}}\sum_{\mathbf{u}\in\mathcal{G}_k}\hat{C}_{ik}(\mathbf{u})\{1-\hat{C}_{ik}(\mathbf{u})\} > 0,$$

the desired result. ∎

5.5 Uniform Convergence of the MAMSE-Weighted Empirical Copula

We now prove a sequence of lemmas that will build up to showing that the MAMSE-weighted empirical copula converges to the copula underlying Population 1. For that purpose, we suppose that the $\{n_{ik}\}$ are increasing with k and that $n_{ik} \to \infty$ as $k \to \infty$.

Lemma 5.1

$$\int \left\{\hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u})\right\}^2 \, \mathrm{d}M_k(\mathbf{u}) \to 0$$

almost surely as $k \to \infty$ where $\hat{\mathcal{C}}_k(\mathbf{u})$ is defined as in (5.4).

Proof of Lemma 5.1. The choice of $\boldsymbol{\lambda} = [1, 0, \ldots, 0]^\mathsf{T}$ yields a larger value of $P_k(\boldsymbol{\lambda})$ than the MAMSE weights because of the optimization involved in the definition of the latter. Hence, for any $\omega \in \Omega$,

$$\int \left\{\hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u})\right\}^2 \, \mathrm{d}M_k(\mathbf{u}) \leq P_k\{\boldsymbol{\lambda}_k(\omega)\} \leq P_k([1, 0, \ldots, 0]^\mathsf{T})$$

$$= \frac{1}{n_{1k}} \int \hat{C}_{1k}(\mathbf{u})\{1 - \hat{C}_{1k}(\mathbf{u})\} \, \mathrm{d}M_k(\mathbf{u}) \leq \frac{1}{4n_{1k}} \to 0$$

as $k \to \infty$. ∎

Lemma 5.2 *Let $\mathbf{u}, \mathbf{v} \in [0,1]^p$ be such that $v_\ell \leq u_\ell$ for $\ell = 1, \ldots, p$. Then*

$$0 \leq \hat{C}_{ik}(\mathbf{u}) - \hat{C}_{ik}(\mathbf{v}) \leq \sum_{\ell=1}^{p} \frac{\lceil n_{ik}(u_\ell - v_\ell) \rceil}{n_{ik}}$$

where $\lceil x \rceil$ denotes the smallest integer greater or equal to x.

Proof of Lemma 5.2. Consider the vectors $\mathbf{u}, \mathbf{v} \in [0,1]^p$ with $v_\ell \leq u_\ell$ for $\ell = 1, \ldots, p$. Then,

$$\begin{aligned} 0 &\leq \hat{C}_{ik}(\mathbf{u}) - \hat{C}_{ik}(\mathbf{v}) \\ &= \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \left[\mathbb{1}\left[\frac{\mathbf{R}_{ij}^k}{n_{ik}} \in \{[0, u_1] \times \cdots \times [0, u_p]\} \right] - \mathbb{1}\left[\frac{\mathbf{R}_{ij}^k}{n_{ik}} \in \{[0, v_1] \times \cdots \times [0, v_p]\} \right] \right] \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \sum_{\ell=1}^{p} \left\{ \mathbb{1}\left(\frac{R_{ij\ell}^k}{n_{ik}} \leq u_\ell \right) - \mathbb{1}\left(\frac{R_{ij\ell}^k}{n_{ik}} \leq v_\ell \right) \right\} \\
&= \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \sum_{\ell=1}^{p} \mathbb{1}\left(\frac{R_{ij\ell}^k}{n_{ik}} \in (v_\ell, u_\ell] \right) \\
&= \sum_{\ell=1}^{p} \left\{ \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \mathbb{1}\left(\frac{R_{ij\ell}^k}{n_{ik}} \in (v_\ell, u_\ell] \right) \right\} \\
&\leq \sum_{\ell=1}^{p} \frac{\lceil n_{ik}(u_\ell - v_\ell) \rceil}{n_{ik}}
\end{aligned}
$$

for any such vectors. ■

Let $\mathcal{G}_k^*$ be an extended grid that includes the axes:

$$\mathcal{G}_k^* = \left\{ 0, \frac{1}{n_{1k}}, \frac{2}{n_{1k}}, \ldots, 1 \right\} \times \cdots \times \left\{ 0, \frac{1}{n_{1k}}, \frac{2}{n_{1k}}, \ldots, 1 \right\}.$$

Lemma 5.3

$$\sup_{\mathbf{u} \in [0,1]^p} \left| \hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u}) \right| \leq \frac{p}{n_{1k}} + \sup_{\mathbf{u} \in \mathcal{G}_k^*} \left| \hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u}) \right|$$

for all k and $\omega \in \Omega$.

Proof of Lemma 5.3. For a fixed k, $|\hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u})|$ is a bounded function on the compact set $[0,1]^p$ and hence its maximum is attained. Let $\mathbf{v} \in [0,1]^p$ be a point where that maximum is achieved. We treat two cases.

Case 1: $\hat{C}_{1k}(\mathbf{v}) \geq \hat{\mathcal{C}}_k(\mathbf{v})$.

Let $\mathbf{v}^* = [v_1^*, \ldots, v_p^*]^\mathsf{T}$ be defined by $v_\ell^* = \lfloor n_{1k} v_\ell \rfloor / n_{1k}$, where $\lfloor x \rfloor$ denotes the largest integer smaller or equal to x. Then $\mathbf{v}^* \in \mathcal{G}_k^*$ is on the same "plateau" of the multivariate

step function $\hat{C}_{1k}(\mathbf{u})$ as $\mathbf{v}$, meaning that

$$\hat{C}_{1k}(\mathbf{v}^*) \;=\; \hat{C}_{1k}(\mathbf{v}).$$

Moreover, since $\mathcal{C}_k(\mathbf{u})$ is a nondecreasing function and $\mathbf{v}^* \leq \mathbf{v}$, we have

$$\hat{\mathcal{C}}_k(\mathbf{v}^*) \;\leq\; \hat{\mathcal{C}}_k(\mathbf{v}).$$

Recalling that $\mathbf{v}$ is the point where the difference between $\hat{\mathcal{C}}_k(\mathbf{u})$ and $\hat{C}_{1k}(\mathbf{u})$ is maximized, we can write

$$\begin{aligned} |\hat{C}_{1k}(\mathbf{v}) - \hat{\mathcal{C}}_k(\mathbf{v})| &= \hat{C}_{1k}(\mathbf{v}) - \hat{\mathcal{C}}_k(\mathbf{v}) \leq \hat{C}_{1k}(\mathbf{v}^*) - \hat{\mathcal{C}}_k(\mathbf{v}^*) \\ &\leq \sup_{\mathbf{u}\in\mathcal{G}_k^*} \left|\hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u})\right| \\ &\leq \frac{p}{n_{1k}} + \sup_{\mathbf{u}\in\mathcal{G}_k^*} \left|\hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u})\right|, \end{aligned}$$

meaning that the maximum occurs at a point of the grid $\mathcal{G}_k^*$.

Case 2: $\hat{C}_{1k}(\mathbf{v}) \leq \hat{\mathcal{C}}_k(\mathbf{v})$.

Let $\mathbf{v}^* = [v_1^*, \ldots, v_p^*]^\mathsf{T}$ be defined by $v_\ell^* = \lceil n_{1k} v_\ell \rceil / n_{1k}$, where $\lceil x \rceil$ denotes the smallest integer greater or equal to x. Then, $\mathbf{v}^* \in \mathcal{G}_k^*$ and

$$\hat{\mathcal{C}}_k(\mathbf{v}^*) \geq \hat{\mathcal{C}}_k(\mathbf{v})$$

since $\mathcal{C}_k(\mathbf{u})$ is a nondecreasing function and $\mathbf{v}^* \geq \mathbf{v}$. By Lemma 5.2,

$$\hat{C}_{1k}(\mathbf{v}^*) - \hat{C}_{1k}(\mathbf{v}) \leq \sum_{\ell=1}^{p} \frac{\lceil n_{1k}(v_\ell^* - v_\ell) \rceil}{n_{1k}} \leq \frac{p}{n_{1k}}.$$

Recalling that $\mathbf{v}$ maximizes the difference between $\hat{C}_{1k}(\mathbf{u})$ and $\mathcal{C}_k(\mathbf{u})$, we can write

$$\begin{aligned} |\hat{C}_{1k}(\mathbf{v}) - \hat{\mathcal{C}}_k(\mathbf{v})| &= \hat{\mathcal{C}}_k(\mathbf{v}) - \hat{C}_{1k}(\mathbf{v}) \\ &\leq \hat{\mathcal{C}}_k(\mathbf{v}^*) - \hat{C}_{1k}(\mathbf{v}^*) + \frac{p}{n_{1k}} \\ &\leq \frac{p}{n_{1k}} + \sup_{\mathbf{u}\in\mathcal{G}_k^*} \left|\hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u})\right|. \end{aligned}$$

Combining Cases 1 and 2 yields the desired result. ■

Lemma 5.4 *Let* $\mathbf{u} = [u_1, \ldots, u_p]^\mathsf{T} \in [0,1]^p$*, then*

$$0 \leq \hat{C}_{ik}(\mathbf{u}) \leq \min_{\ell=1,\ldots,p} u_\ell.$$

Proof of Lemma 5.4. For each $\ell \in \{1, \ldots, p\}$, we have

$$\hat{C}_{ik}(\mathbf{u}) = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \prod_{\ell=1}^{p} \mathbb{1}\left(\frac{R_{ij\ell}^k}{n_{ik}} \leq u_\ell\right) \leq \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \mathbb{1}\left(\frac{R_{ij\ell}^k}{n_{ik}} \leq u_\ell\right) \leq u_\ell.$$

We get the desired result by noting that these p inequalities have to be satisfied simultaneously. ■

Lemma 5.5 *We have*

$$\sup_{\mathbf{u}\in\mathcal{G}_k^*} \left|\hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u})\right| \to 0$$

almost surely as $k \to \infty$*.*

Proof of Lemma 5.5. Let $\epsilon > 0$. For any given $k \in \mathbb{N}$, let $\mathbf{u}_k = [u_{k1}, \ldots, u_{kp}]^\mathsf{T}$ be the point of the grid $\mathcal{G}_k^*$ where $|\hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u})|$ is maximized. Let

$$\begin{aligned} A_k &= \left\{\omega \in \Omega : \hat{C}_{1k}(\mathbf{u}_k) - \hat{\mathcal{C}}_k(\mathbf{u}_k) > \epsilon\right\}, \\ B_k &= \left\{\omega \in \Omega : \hat{\mathcal{C}}_k(\mathbf{u}_k) - \hat{C}_{1k}(\mathbf{u}_k) > \epsilon\right\}, \\ C_k &= \left\{\omega \in \Omega : \mathbf{u}_k \in \left[\frac{\epsilon}{2}, 1\right]^p\right\}. \end{aligned}$$

The negation of Lemma 5.5 is

$$A_k \cup B_k \quad i.o.$$

which will happen if and only if

$$\left(A_k \cup B_k \cap C_k^C\right) \cup \left(A_k \cup B_k \cap C_k\right) \quad i.o..$$

We will show that neither of the two events in the decomposition above can occur infinitely often.

Case 1: $A_k \cup B_k \cap C_k^C$.

We have

$$\begin{aligned} \left|\hat{C}_{1k}(\mathbf{u}_k) - \hat{\mathcal{C}}_k(\mathbf{u}_k)\right| &\leq \left|\hat{C}_{1k}(\mathbf{u}_k)\right| + \left|\hat{\mathcal{C}}_k(\mathbf{u}_k)\right| \\ &= \hat{C}_{1k}(\mathbf{u}_k) + \sum_{i=1}^{m} \lambda_{ik}(\omega)\hat{C}_{ik}(\mathbf{u}_k) \\ &\leq 2 \min_{\ell \in \{1,\ldots,p\}} u_{k\ell} \leq \epsilon \end{aligned}$$

by Lemma 5.4 since the MAMSE weights sum to 1. Consequently, $A_k \cup B_k \cap C_k^C = \emptyset$ for all k.

Case 2: $A_k \cup B_k \cap C_k$.

Let $\mathbf{v}$ be a vector of integers such that $\mathbf{v}/n_{1k} = \mathbf{u}_k$; we temporarily omit the index k for notational simplicity. Let also $\mathbf{w} = [w_1, \ldots, w_p]^\mathsf{T}$ be a point from the set

$$\mathcal{W} = \left\{0, 1, \ldots, \left\lfloor \frac{n_{1k}}{2p}\,\epsilon \right\rfloor\right\} \times \cdots \times \left\{0, 1, \ldots, \left\lfloor \frac{n_{1k}}{2p}\,\epsilon \right\rfloor\right\}.$$

The points $(\mathbf{v} - \mathbf{w})/n_{1k}$ belong to $\mathcal{G}_k$ since $\mathbf{u}_k \in [\epsilon/2, 1]^p$. Next, we show that

$$\left|\hat{\mathcal{C}}_k\left(\frac{\mathbf{v}-\mathbf{w}}{n_{1k}}\right) - \hat{C}_{1k}\left(\frac{\mathbf{v}-\mathbf{w}}{n_{1k}}\right)\right| \geq \frac{\epsilon}{2} - \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}} \geq 0$$

by treating two subcases. Note that the last inequality holds because

$$\frac{1}{n_{1k}} \sum_{\ell=1}^{p} w_\ell \leq \frac{p}{n_{1k}} \left\lfloor \frac{n_{1k}}{2p} \epsilon \right\rfloor \leq \frac{\epsilon}{2}.$$

Subcase A: $A_k \cap C_k$.

From the fact that the copulas are monotone functions, we get:

$$\hat{\mathcal{C}}_k \left(\frac{\mathbf{v} - \mathbf{w}}{n_{1k}} \right) \leq \hat{\mathcal{C}}_k \left(\frac{\mathbf{v}}{n_{1k}} \right) = \hat{\mathcal{C}}_k(\mathbf{u}_k).$$

Then from Lemma 5.2,

$$\hat{C}_{1k} \left(\frac{\mathbf{v} - \mathbf{w}}{n_{1k}} \right) \geq \hat{C}_{1k} \left(\frac{\mathbf{v}}{n_{1k}} \right) - \frac{\sum_{\ell=1}^{p} \left\lceil n_{1k} \frac{w_\ell}{n_{1k}} \right\rceil}{n_{1k}} = \hat{C}_{1k}(\mathbf{u}_k) - \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}}.$$

Combining the two inequalities yields

$$\begin{aligned} \hat{C}_{1k} \left(\frac{\mathbf{v} - \mathbf{w}}{n_{1k}} \right) - \hat{\mathcal{C}}_k \left(\frac{\mathbf{v} - \mathbf{w}}{n_{1k}} \right) &\geq \hat{C}_{1k}(\mathbf{u}_k) - \hat{\mathcal{C}}_k(\mathbf{u}_k) - \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}} \\ &\geq \epsilon - \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}} \\ &\geq \frac{\epsilon}{2} - \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}} \geq 0. \end{aligned}$$

Subcase B: $B_k \cap C_k$.

By Lemma 5.2, we have

$$\begin{aligned} \hat{\mathcal{C}}_k \left(\frac{\mathbf{v}}{n_{1k}} \right) - \hat{\mathcal{C}}_k \left(\frac{\mathbf{v} - \mathbf{w}}{n_{1k}} \right) &= \sum_{i=1}^{m} \lambda_{ik}(\omega) \left\{ \hat{C}_{ik} \left(\frac{\mathbf{v}}{n_{1k}} \right) - \hat{C}_{ik} \left(\frac{\mathbf{v} - \mathbf{w}}{n_{1k}} \right) \right\} \\ &\leq \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{n_{ik}} \sum_{\ell=1}^{p} \left\lceil \frac{n_{ik} w_\ell}{n_{1k}} \right\rceil \leq \sum_{i=1}^{m} \frac{1}{n_{ik}} \sum_{\ell=1}^{p} \left\lceil \frac{n_{ik} w_\ell}{n_{1k}} \right\rceil \end{aligned}$$

$$\leq \sum_{i=1}^{m} \frac{1}{n_{ik}} \sum_{\ell=1}^{p} \left(\frac{n_{ik} w_\ell}{n_{1k}} + 1 \right) = \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}} + \sum_{i=1}^{m} \frac{p}{n_{ik}}.$$

Hence,

$$\hat{\mathcal{C}}_k \left(\frac{\mathbf{v} - \mathbf{w}}{n_{1k}} \right) \geq \hat{\mathcal{C}}_k \left(\frac{\mathbf{v}}{n_{1k}} \right) - \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}} - \sum_{i=1}^{m} \frac{p}{n_{ik}} = \hat{\mathcal{C}}_k(\mathbf{u}_k) - \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}} - \sum_{i=1}^{m} \frac{p}{n_{ik}}.$$

Since the empirical copula is a monotone function, we also have

$$\hat{C}_{1k} \left(\frac{\mathbf{v} - \mathbf{w}}{n_{1k}} \right) \leq \hat{C}_{1k} \left(\frac{\mathbf{v}}{n_{1k}} \right) = \hat{C}_{1k}(\mathbf{u}_k).$$

Let us consider only k that are large enough to make $\sum_{i=1}^{m} p/n_{ik} < \epsilon/2$. From the two previous inequalities, we obtain

$$\begin{aligned} \hat{\mathcal{C}}_k \left(\frac{\mathbf{v} - \mathbf{w}}{n_{1k}} \right) - \hat{C}_{1k} \left(\frac{\mathbf{v} - \mathbf{w}}{n_{1k}} \right) &\geq \hat{\mathcal{C}}_k(\mathbf{u}_k) - \hat{C}_{1k}(\mathbf{u}_k) - \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}} - \sum_{i=1}^{m} \frac{p}{n_{ik}} \\ &\geq \epsilon - \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}} - \sum_{i=1}^{m} \frac{p}{n_{ik}} \\ &\geq \frac{\epsilon}{2} - \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}} \geq 0. \end{aligned}$$

Combining subcases A and B yields

$$P_k(\boldsymbol{\lambda}) \geq \int \{\hat{\mathcal{C}}_k(\mathbf{u}) - \hat{C}_{1k}(\mathbf{u})\}^2 \, \mathrm{d}M_k(\mathbf{u}) \geq \frac{1}{n_{1k}^p} \sum_{\mathbf{w} \in \mathcal{W}} \left(\frac{\epsilon}{2} - \frac{\sum_{\ell=1}^{p} w_\ell}{n_{1k}} \right)^2.$$

The sum above corresponds to the Riemann sum of the multiple integral

$$\int_0^{\frac{\epsilon}{2p}} \cdots \int_0^{\frac{\epsilon}{2p}} \left(\frac{\epsilon}{2} - \sum_{\ell=1}^{p} y_\ell \right)^2 \mathrm{d}y_1 \cdots \mathrm{d}y_p = K_p.$$

The number K_p is a fixed positive constant for any fixed p.

As a consequence, there exists a k_0 such that for all $k \geq k_0$, $P_k(\boldsymbol{\lambda}) > K_p/2 > 0$, a contradiction with Lemma 5.1. We must thus conclude that $A_k \cup B_k \cap C_k$ occurs at most a finite number of times.

Since the two cases above do not occur infinitely often, we conclude that $A_k \cup B_k$ occurs at most a finite number of times. Hence,

$$\sup_{\mathbf{u} \in \mathcal{G}_k^*} \left| \hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u}) \right| \to 0$$

almost surely as $k \to \infty$. ■

Lemma 5.6

$$\sup_{\mathbf{u} \in [0,1]^p} \left| \hat{C}_{1k}(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u}) \right| \to 0$$

almost surely as $k \to \infty$.

Proof of Lemma 5.6. The result follows from Lemma 5.3 and Lemma 5.5. ■

Theorem 5.3 *We have uniform convergence of the MAMSE weighted empirical copula:*

$$\sup_{\mathbf{u} \in [0,1]^p} \left| \hat{\mathcal{C}}_k(\mathbf{u}) - C_1(\mathbf{u}) \right| \to 0$$

almost surely as $k \to \infty$.

Proof of Theorem 5.3. Consider the decomposition

$$\left| \hat{\mathcal{C}}_k(\mathbf{u}) - C_1(\mathbf{u}) \right| \leq \left| \hat{\mathcal{C}}_k(\mathbf{u}) - \hat{C}_{1k}(\mathbf{u}) \right| + \left| \hat{C}_{1k}(\mathbf{u}) - C_1(\mathbf{u}) \right|$$

that holds for all $\mathbf{u} \in [0,1]^p$. Then

$$\sup_{\mathbf{u} \in [0,1]^p} \left| \hat{\mathcal{C}}_k(\mathbf{u}) - C_1(\mathbf{u}) \right| \quad \leq \quad \sup_{\mathbf{u} \in [0,1]^p} \left\{ \left| \hat{\mathcal{C}}_k(\mathbf{u}) - \hat{C}_{1k}(\mathbf{u}) \right| + \left| \hat{C}_{1k}(\mathbf{u}) - C_1(\mathbf{u}) \right| \right\}$$

$$\leq \sup_{\mathbf{u}\in[0,1]^p} \left|\hat{\mathcal{C}}_k(\mathbf{u}) - \hat{C}_{1k}(\mathbf{u})\right| + \sup_{\mathbf{u}\in[0,1]^p} \left|\hat{C}_{1k}(\mathbf{u}) - C_1(\mathbf{u})\right| \to 0$$

almost surely as $k \to \infty$. Indeed, the first term goes to 0 almost surely by Lemma 5.6 and the expression

$$\sup_{\mathbf{u}\in[0,1]^p} \left|\hat{C}_{1k}(\mathbf{u}) - C_1(\mathbf{u})\right|$$

converges almost surely to 0 as $k \to \infty$, see e.g. Expression 3.1 of Deheuvels (1979). ■

Corollary 5.1 *Let* $\mathbf{U}_k$ *be a sequence of random vectors with distribution* $\hat{\mathcal{C}}_k(\mathbf{u})$ *and* $\mathbf{U}$ *a random vector with distribution* $C_1(\mathbf{u})$*. Then* $\mathbf{U}_k$ *converges to* $\mathbf{U}$ *in distribution.*

Proof of Corollary 5.1. For almost every $\omega \in \Omega$, the sequence of distributions $\hat{\mathcal{C}}_k(\mathbf{u})$ converges to $C_1(\mathbf{u})$ for all $\mathbf{u}$, hence the definition of weak convergence is respected with probability 1. ■

Corollary 5.2 *If* $g(\mathbf{u})$ *is a bounded function, then*

$$\int g(\mathbf{u})\,\mathrm{d}\hat{\mathcal{C}}_k(\mathbf{u}) = \mathrm{E}\{g(\mathbf{U}_k)\} \to \mathrm{E}\{g(\mathbf{U})\} = \int g(\mathbf{u})\,\mathrm{d}C_1(\mathbf{u})$$

almost surely as $k \to \infty$*.*

Proof of Corollary 5.2. The result is a well known consequence of the weak convergence proved in Corollary 5.1, see page 164 of Durrett (2005) for more details. ■

5.6 Weighted Coefficients of Correlation Based on Ranks

Coefficients of correlation based on ranks typically estimate a measure of dependence that depends on the copula of the underlying distribution. Examples of such coefficient are provided in Tables 5.1 and 5.2.

Suppose that bivariate data from m populations are available. We suggest the use of the MAMSE weights to create a weighted coefficient of correlation and show in this section that it converges almost surely to the true correlation of Population 1.

Consider first Blomqvist's β and let $\hat{\beta}_{ik}$ be the Blomqvist coefficient based on the ranks of the sample $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_{ik}}$. The weighted Blomqvist's β is given by

$$\hat{\beta}_{\boldsymbol{\lambda}k} = \sum_{i=1}^{m} \lambda_{ik}(\omega)\hat{\beta}_{ik}.$$

Theorem 5.4 *$\hat{\beta}_{\boldsymbol{\lambda}k} \to \beta_1$ almost surely as $k \to \infty$.*

Proof of Theorem 5.4. As a direct consequence of Corollary 5.3,

$$\begin{aligned}
\hat{\beta}_{\boldsymbol{\lambda}k} &= 4\sum_{i=1}^{m} \lambda_{ik}(\omega)\hat{C}_{ik}\left(\frac{1}{2}, \frac{1}{2}\right) - 1 \\
&= 4\hat{\mathcal{C}}_k\left(\frac{1}{2}, \frac{1}{2}\right) - 1 \\
&\to 4C_1\left(\frac{1}{2}, \frac{1}{2}\right) - 1 = \beta_1
\end{aligned}$$

almost surely as $k \to \infty$ ∎

The remainder of this section will cover a general case that includes the other coefficients in Table 5.1 and 5.2, except for Kendall's τ. The population version of Kendall's τ depends on $\int C(u, v)\, \mathrm{d}C(u, v)$. A weighted version of Kendall's τ is achievable, but replacing both $C(u, v)$ in this theoretical expression with mixture distributions will not yield a linear combination like the other coefficients considered. The weighted Kendall's τ hence requires a special treatment that is left to future work.

Replacing the copulas by their empirical counterparts in Table 5.1 yields estimates of the correlation based on ranks. However, the expressions obtained typically differ from the usual estimates based on ranks and do not necessarily span the interval $[-1, 1]$, but rather an interval whose bounds tend to ± 1 as the sample size goes to infinity.

The classic formulas for the coefficients of correlation based on the ranks of the sample

$\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_{ik}}$ will thus admit a representation of the form

$$\hat{\kappa}_{ik} = a_k \int\int g(\mathbf{u})\, \mathrm{d}\hat{C}_{ik}(\mathbf{u}) + b_k \tag{5.5}$$

that estimates

$$\hat{\kappa}_{ik} = a \int\int g(\mathbf{u})\, \mathrm{d}C_i(\mathbf{u}) + b.$$

The coefficients $a_n \to a$ and $b_n \to b$ as $n \to \infty$ are chosen to ensure that $\hat{\kappa}_{ik} \in [-1, 1]$ for all sample sizes n_{ik}. Moreover, the values ± 1 occur only for perfect concordance or discordance, i.e. when the ranks are identical ($R^k_{ij1} = R^k_{ij2}$) or antithetic ($R^k_{ij1} = n_{ik} + 1 - R^k_{ij2}$).

The function $g(\mathbf{u})$ is typically bounded and its functional form defines the coefficient of correlation at hand (see Chapter 2 of Plante (2002) for detailed explanations).

Consider a particular case: for Population i and a fixed k, the empirical estimate of Spearman's coefficient, can be written

$$\begin{aligned}
\hat{\rho}_{ik} &= -3\,\frac{n_{ik}+1}{n_{ik}-1} + \frac{12 n_{ik}}{(n_{ik}+1)(n_{ik}-1)} \sum_{j=1}^{n_{ik}} \frac{R^k_{ij1}}{n_{ik}}\,\frac{R^k_{ij2}}{n_{ik}} \\
&= -3\,\frac{n_{ik}+1}{n_{ik}-1} + \frac{12 n_{ik}^2}{(n_{ik}+1)(n_{ik}-1)} \int u_1 u_2\, \mathrm{d}\hat{C}_{ik}(\mathbf{u})
\end{aligned}$$

and clearly has the same asymptotic behavior as $-3 + 12 \int u_1 u_2\, \mathrm{d}\hat{C}_{ik}(\mathbf{u})$.

Let

$$\hat{\kappa}_{\boldsymbol{\lambda} k} = \sum_{i=1}^{m} \lambda_{ik}(\omega) \hat{\kappa}_{ik}$$

be a MAMSE-weighted coefficient of correlation based on ranks.

Remark 5.5 *The MAMSE weighted coefficients of correlations are invariant under monotone transformations of the data. Indeed, both the MAMSE weights and the individual correlations are based on the ranks of the data which are invariant to such transformations.*

Next, we prove that the MAMSE-weighted coefficients of correlations are strongly consistent estimates of the correlation in Population 1.

Lemma 5.7

$$\frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} g\left(\frac{\mathbf{R}_{ij}^k}{n_{ik}}\right) = \int g(\mathbf{u}) \, \mathrm{d}\hat{C}_{ik}(\mathbf{u}) \to \int g(\mathbf{u}) \, \mathrm{d}C_i(\mathbf{u})$$

almost surely as $k \to \infty$ for any bounded continuous function g.

Proof of Lemma 5.7. By the uniform convergence of $\hat{C}_{ik}$ proved by Deheuvels (1979), the sequence of distributions $\hat{C}_{ik}(\mathbf{u})$ converges to $C_1(\mathbf{u})$ for all $\mathbf{u}$ and almost every $\omega \in \Omega$. Consequently, $\hat{C}_{ik}$ converges weakly to C_i. Since the function g is bounded, the almost sure convergence of the expectation follows (see e.g. Durrett (2005), page 164). ■

Lemma 5.8 *The coefficient $\hat{\kappa}_{ik}$ converges almost surely to*

$$\kappa_i = a \int g(\mathbf{u}) \, \mathrm{d}C_i(\mathbf{u}) + b,$$

its associated population value.

Proof of Lemma 5.8. Note that

$$\begin{aligned}
|\hat{\kappa}_{ik} - \kappa_i| &\leq \left|\hat{\kappa}_{ik} - \left\{a \int g(\mathbf{u}) \, \mathrm{d}\hat{C}_{ik}(\mathbf{u}) + b\right\}\right| + \left|\left\{a \int g(\mathbf{u}) \, \mathrm{d}\hat{C}_{ik}(\mathbf{u}) + b\right\} - \kappa_i\right| \\
&= \left|(a_k - a) \int g(\mathbf{u}) \, \mathrm{d}\hat{C}_{ik}(\mathbf{u}) + (b_k - b)\right| \\
&\qquad \left|a \int g(\mathbf{u}) \, \mathrm{d}\hat{C}_{ik}(\mathbf{u}) - a \int g(\mathbf{u}) \, \mathrm{d}C_i(\mathbf{u}) + b - b\right| \\
&\leq |a_k - a| \left|\int g(\mathbf{u}) \, \mathrm{d}\hat{C}_{ik}(\mathbf{u})\right| + |b_k - b| \\
&\qquad + a \left|\int g(\mathbf{u}) \, \mathrm{d}\hat{C}_{ik}(\mathbf{u}) - \int g(\mathbf{u}) \, \mathrm{d}C_i(\mathbf{u})\right| \\
&\to 0
\end{aligned}$$

almost surely as $k \to \infty$ since $\int g(\mathbf{u}) \, \mathrm{d}\hat{C}_{ik}(\mathbf{u})$ is bounded and by Lemma 5.7, the other term involving integrals tends to 0. ■

Theorem 5.5 *The weighted coefficient of correlation*

$$\hat{\kappa}_{\boldsymbol{\lambda}k} = \sum_{i=1}^{m} \lambda_{ik}(\omega)\hat{\kappa}_{ik} \to \kappa_1$$

almost surely as $k \to \infty$.

Proof of Theorem 5.5. The proof follows the same steps as that of Lemma 5.8

$$\begin{aligned}
|\hat{\kappa}_{\boldsymbol{\lambda}k} - \kappa_1| &\leq \left|\hat{\kappa}_{\boldsymbol{\lambda}k} - \left\{a\int g(\mathbf{u})\,\mathrm{d}\hat{\mathcal{C}}_k(\mathbf{u}) + b\right\}\right| + \left|\left\{a\int g(\mathbf{u})\,\mathrm{d}\hat{\mathcal{C}}_k(\mathbf{u}) + b\right\} - \kappa_1\right| \\
&= \left|(a_k - a)\int g(\mathbf{u})\,\mathrm{d}\hat{\mathcal{C}}_k(\mathbf{u}) + (b_k - b)\right| \\
&\qquad \left|a\int g(\mathbf{u})\,\mathrm{d}\hat{\mathcal{C}}_k(\mathbf{u}) - a\int g(\mathbf{u})\,\mathrm{d}C_1(\mathbf{u}) + b - b\right| \\
&\leq |a_k - a|\left|\int g(\mathbf{u})\,\mathrm{d}\hat{\mathcal{C}}_k(\mathbf{u})\right| + |b_n - b| \\
&\qquad +a\left|\int g(\mathbf{u})\,\mathrm{d}\hat{\mathcal{C}}_k(\mathbf{u}) - \int g(\mathbf{u})\,\mathrm{d}C_1(\mathbf{u})\right| \\
&\to 0
\end{aligned}$$

almost surely as $k \to \infty$ since g is bounded and $\int g(\mathbf{u})\,\mathrm{d}\hat{\mathcal{C}}_k(\mathbf{u})$ converges to $\int g(\mathbf{u})\,\mathrm{d}C_1(\mathbf{u})$ by Corollary 5.2. ■

Corollary 5.3 *The MAMSE weighted versions of Spearman's* ρ*, Gini's* γ *or Blest's* ν *and* ξ *are strongly consistent estimates of their corresponding population value.*

Proof of Corollary 5.3. Direct consequence of Theorem 5.5. ■

5.7 Weighted Strong Law of Large Numbers

In Chapter 3, we proved a WSLLN for unbounded functions in order to prove the consistency of the weighted likelihood with MAMSE weights. In the next section, we propose the maximum weighted pseudo-likelihood estimate and prove its consistency. To do that, we first prove a weighted strong law of large number more general than Corollary 5.2 as it

holds for functions g continuous on $(0,1)^p$ with asymptotes at the boundary of the unit hyper-cube.

Let $\hat{C}^*_{ik}$ denote a rescaled empirical copula based on transformed ranks $\mathbf{Y}^k_{ij}$ as defined in Section 5.2. We let $\hat{\mathcal{C}}^*(\mathbf{u}) = \sum_{i=1}^m \lambda_{ik}(\omega)\hat{C}^*_{ik}$ and propose to show that for a function g satisfying some regularity conditions,

$$\sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{n_{ik}} \sum_{j=1}^{n_{ik}} g\left(\mathbf{Y}^k_{ij}\right) = \int g(\mathbf{u})\,\mathrm{d}\hat{\mathcal{C}}^*_k(\mathbf{u}) \to \int g(\mathbf{u})\,\mathrm{d}C_1(\mathbf{u})$$

almost surely as $k \to \infty$.

Since we are now using rescaled copulas, let us first prove the following result.

Lemma 5.9

$$\sup_{\mathbf{u}\in[0,1]^p} |\hat{\mathcal{C}}^*_k(\mathbf{u}) - C_1(\mathbf{u})| \to 0$$

almost surely as $k \to \infty$.

Proof of Lemma 5.9. Remark 5.4 points out that

$$\sup_{\mathbf{u}\in[0,1]^p} |\hat{C}^*_{ik}(\mathbf{u}) - \hat{C}_{ik}(\mathbf{u})| < \frac{1}{n_{ik}} \to 0$$

as $k \to \infty$. Therefore,

$$\begin{aligned}
\sup_{\mathbf{u}\in[0,1]^p} |\hat{\mathcal{C}}^*_k(\mathbf{u}) - C_1(\mathbf{u})| &\leq \sup_{\mathbf{u}\in[0,1]^p} \left\{ |\hat{\mathcal{C}}^*_k(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u})| + |\hat{\mathcal{C}}_k(\mathbf{u}) - C_1(\mathbf{u})| \right\} \\
&\leq \sup_{\mathbf{u}\in[0,1]^p} |\hat{\mathcal{C}}^*_k(\mathbf{u}) - \hat{\mathcal{C}}_k(\mathbf{u})| + \sup_{\mathbf{u}\in[0,1]^p} |\hat{\mathcal{C}}_k(\mathbf{u}) - C_1(\mathbf{u})| \\
&\leq \sum_{i=1}^{m} \lambda_{ik}(\omega) \sup_{\mathbf{u}\in[0,1]^p} |\hat{C}^*_{ik}(\mathbf{u}) - \hat{C}_{ik}(\mathbf{u})| + \sup_{\mathbf{u}\in[0,1]^p} |\hat{\mathcal{C}}_k(\mathbf{u}) - C_1(\mathbf{u})| \\
&\leq \sum_{i=1}^{m} \sup_{\mathbf{u}\in[0,1]^p} |\hat{C}^*_{ik}(\mathbf{u}) - \hat{C}_{ik}(\mathbf{u})| + \sup_{\mathbf{u}\in[0,1]^p} |\hat{\mathcal{C}}_k(\mathbf{u}) - C_1(\mathbf{u})| \to 0
\end{aligned}$$

as $k \to \infty$ by Remark 5.4 and Theorem 5.3. ∎

Proposition A·1(i) of Genest et al. (1995) provides a strong law of large numbers that we use to prove the weighted strong law of large numbers that we need. To simplify their notation, Genest et al. (1995) present their result with bivariate copulas, but we do not make this simplification here.

Theorem 5.6 (Genest et al. (1995)) *Let $r(u) = u(1-u)$, $\delta > 0$, $\{q_\ell, \ell = 1, \ldots, p\}$ be positive numbers satisfying $\sum_{\ell=1}^{p} 1/q_\ell = 1$ and $g(\mathbf{u})$ a continuous function from $(0,1)^p$ to $\mathbb{R}$ such that $\mu_i = \int g(\mathbf{u})\,\mathrm{d}C_i(\mathbf{u})$ exists. If $M < \infty$ is a positive real number such that*

$$|g(\mathbf{u})| \leq M \prod_{\ell=1}^{p} r(u_\ell)^{a_\ell}$$

with $a_\ell = (-1+\delta)/q_\ell$, then

$$R_n = \int g(\mathbf{u})\,\mathrm{d}\hat{C}^*_{ik}(\mathbf{u}) \to \mu_i$$

almost surely as $k \to \infty$.

A bounded function always satisfies the assumptions of Theorem 5.6 because $\prod_{\ell=1}^{p} r(u_\ell)^{a_\ell}$ has a lower bound. Lemma 5.7 is thus a particular case of Theorem 5.6.

Lemma 5.10 *Let A be a cube in p dimensions whose opposite corners are $\mathbf{u}_1$ and $\mathbf{u}_2$ with $\mathbf{u}_1 < \mathbf{u}_2$, i.e.*

$$A = (u_{11}, u_{21}] \times \cdots \times (u_{1p}, u_{2p}].$$

Let $C_1(\mathbf{u})$ and $C_2(\mathbf{u})$ be p-dimensional copulas such that

$$\sup_{\mathbf{u} \in [0,1]^p} |C_1(\mathbf{u}) - C_2(\mathbf{u})| < \epsilon.$$

Then,

$$|\,\mathrm{d}C_1(A) - \,\mathrm{d}C_2(A)| < \frac{2^p}{\epsilon}.$$

Proof of Lemma 5.10. The coordinates of the corners of the cube can be re-expressed by replacing some elements of $\mathbf{u}_1$ by the corresponding elements of $\mathbf{u}_2$. Let $\mathcal{S} \subset \{1, \ldots, p\}$ be a set of indices and $\mathbf{v}_{\mathcal{S}}$ any vector such that

$$v_i = \begin{cases} u_{1i} & \text{if } i \in \mathcal{S} \\ u_{2i} & \text{if } i \in \mathcal{S}^C \end{cases} .$$

To calculate the probability of the cube, we can use the following development which is akin to the formula for the intersection of multiple sets:

$$\mathrm{d}C_1(A) = C_1\left(\mathbf{v}_{\emptyset}\right) - \sum_{i=1}^{p} C_1\left(\mathbf{v}_{\{i\}}\right) + \sum_{1 \le i < j \le p} C_1\left(\mathbf{v}_{\{i,j\}}\right) - \cdots \pm C_1\left(\mathbf{v}_{\{1,\ldots,p\}}\right).$$

Consider now the difference

$$\begin{aligned} |\,\mathrm{d}C_1(A) - \mathrm{d}C_2(A)| &= \Bigg| C_1\left(\mathbf{v}_{\emptyset}\right) - C_2\left(\mathbf{v}_{\emptyset}\right) - \sum_{i=1}^{p} \left\{C_1\left(\mathbf{v}_{\{i\}}\right) - C_2\left(\mathbf{v}_{\{i\}}\right)\right\} \\ &\quad + \sum_{1 \le i < j \le p} \left\{C_1\left(\mathbf{v}_{\{i,j\}}\right) - C_2\left(\mathbf{v}_{\{i,j\}}\right)\right\} \\ &\quad - \cdots \pm \left\{C_1\left(\mathbf{v}_{\{1,\ldots,p\}}\right) - C_2\left(\mathbf{v}_{\{1,\ldots,p\}}\right)\right\} \Bigg| \\ &\le \left|C_1\left(\mathbf{v}_{\emptyset}\right) - C_2\left(\mathbf{v}_{\emptyset}\right)\right| + \sum_{i=1}^{p} \left|C_1\left(\mathbf{v}_{\{i\}}\right) - C_2\left(\mathbf{v}_{\{i\}}\right)\right| \\ &\quad + \sum_{1 \le i < j \le p} \left|C_1\left(\mathbf{v}_{\{i,j\}}\right) - C_2\left(\mathbf{v}_{\{i,j\}}\right)\right| \\ &\quad + \cdots + \left|C_1\left(\mathbf{v}_{\{1,\ldots,p\}}\right) - C_2\left(\mathbf{v}_{\{1,\ldots,p\}}\right)\right| \\ &\le 2^p \epsilon. \end{aligned}$$

Since the sum above has

$$\sum_{i=1}^{p} \binom{p}{i} = 2^p$$

terms each bounded by ϵ. ∎

Corollary 5.4 *Let B be a closed cube,*

$$B = [u_{11}, u_{21}] \times \cdots \times [u_{1p}, u_{2p}].$$

Let $C_1(\mathbf{u})$ and $C_2(\mathbf{u})$ be p-dimensional copulas such that

$$\sup_{\mathbf{u} \in [0,1]^p} |C_1(\mathbf{u}) - C_2(\mathbf{u})| < \epsilon.$$

Then,

$$|\, \mathrm{d}C_1(B) - \, \mathrm{d}C_2(B)| < \frac{2^p}{\epsilon}.$$

Proof of Corollary 5.4. Apply Lemma 5.10 to the sequence of half-open cubes

$$A_\delta = (u_{11} - \delta, u_{21}] \times \cdots \times (u_{1p} - \delta, u_{2p}]$$

with $\delta \to 0$. A similar proof holds for a mix of closed and open intervals. ∎

Theorem 5.7 *Let $g(\mathbf{u})$ be any continuous function on $(0,1)^p$ that satisfies the assumptions of Theorem 5.6. Suppose that the sample sizes $n_{ik} \to \infty$ for all populations. Then,*

$$\left| \int g(\mathbf{u}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) - \int g(\mathbf{u}) \, \mathrm{d}C_1(\mathbf{u}) \right| \to 0$$

almost surely as $k \to \infty$.

Proof of Theorem 5.7. For $t \in \mathbb{N}$, let $B_t = [2^{-t}, 1 - 2^{-t}]^p$ and

$$\tau_t(\mathbf{u}) = \begin{cases} g(\mathbf{u}) & \text{if } \mathbf{u} \in B_t \\ 0 & \text{otherwise} \end{cases}.$$

Since $g(\mathbf{u})$ is continuous and B_t is a compact set, the image of τ_t is bounded. Suppose that $\tau_t(\mathbf{u}) \in [L_t, U_t]$. By the Heine-Cantor Theorem, τ_t is uniformly continuous on B_t,

i.e. $\forall \epsilon_{\tau,t} > 0$, $\exists \delta_{\tau,t} > 0$ such that

$$\forall \mathbf{u}, \mathbf{v} \in B_t, \quad |\mathbf{u} - \mathbf{v}| \leq \delta_{\tau,t} \Longrightarrow |\tau_t(\mathbf{u}) - \tau_t(\mathbf{v})| \leq \epsilon_{\tau,t}.$$

Let $\epsilon_{\tau,t} = 2^{-t}$ and choose $0 < \delta_{\tau,t} < 2^{-t}$ accordingly. Let $\mathcal{A}_t$ be a partition of the interval $[2^{-t}, 1 - 2^{-t}]$,

$$\mathcal{A}_t = \{[2^{-t}, (2)2^{-t}]\} \cup \left\{(s2^{-t}, (s+1)2^{-t}], s = 2, \ldots, 2^t - 2\right\}.$$

Then, the elements of

$$\mathcal{A}_t \times \cdots \times \mathcal{A}_t = \{A_{1t}, \ldots, A_{S_t t}\}$$

form a partition of B_t of cardinality $S_t = (2^t - 2)^p$. Define

$$h_t(\mathbf{u}) = \sum_{s=1}^{S_t} b_{st} \mathbb{1}_{A_{st}}(\mathbf{u})$$

where

$$b_{st} = \inf_{y \in A_{st}} g(\mathbf{u}) \qquad \text{and} \qquad \mathbb{1}_{A_{st}}(\mathbf{u}) = \begin{cases} 1 & \text{if } \mathbf{u} \in A_{st} \\ 0 & \text{otherwise} \end{cases}.$$

Then, by construction $\sup_{\mathbf{u} \in [0,1]^p} |\tau_t(\mathbf{u}) - h_t(x)| \leq 2^{-t}$ and

$$\left| \int g(\mathbf{u}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) - \int g(\mathbf{u}) \, \mathrm{d}C_1(\mathbf{u}) \right| \leq T_1 + T_2 + T_3 + T_4 + T_5 \tag{5.6}$$

where

$$\begin{aligned} T_1 &= \left| \int g(\mathbf{u}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) - \int \tau_t(\mathbf{u}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) \right| \\ T_2 &= \left| \int \tau_t(\mathbf{u}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) - \int h_t(\mathbf{u}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) \right| \\ T_3 &= \left| \int h_t(\mathbf{u}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) - \int h_t(\mathbf{u}) \, \mathrm{d}C_1(\mathbf{u}) \right| \end{aligned}$$

$$\begin{aligned} T_4 &= \left| \int h_t(\mathbf{u})\, \mathrm{d}C_1(\mathbf{u}) - \int \tau_t(\mathbf{u})\, \mathrm{d}C_1(\mathbf{u}) \right| \\ T_5 &= \left| \int \tau_t(\mathbf{u})\, \mathrm{d}C_1(\mathbf{u}) - \int g(\mathbf{u})\, \mathrm{d}C_1(\mathbf{u}) \right|. \end{aligned}$$

We now prove that for any $\epsilon > 0$ and ω in a subset of Ω with probability 1, we can choose t_ω such that the five terms above are less than $\epsilon/5$ for all $k \geq k_\omega(t_\omega)$.

First note that

$$T_4 = \left| \int h_t(\mathbf{u}) - \tau_t(\mathbf{u})\, \mathrm{d}C_1(\mathbf{u}) \right| \leq \int |h_t(\mathbf{u}) - \tau_t(\mathbf{u})|\, \mathrm{d}C_1(\mathbf{u}) \leq 2^{-t}$$

by construction. The same bound applies for T_2 and does not depend on k or ω.

By Lemma 5.9, $\sup_{\mathbf{u}\in[0,1]^p} |\hat{C}_k^*(\mathbf{u}) - C_1(\mathbf{u})|$ converges almost surely to 0. Therefore, $\exists \Omega_0 \subset \Omega$ with $P(\Omega_0) = 1$ such that for each $\omega \in \Omega_0$ and any t, $\exists k_{\omega,t}$ with

$$\sup_{\mathbf{u}\in[0,1]^p} |\hat{C}_k^*(\mathbf{u}) - C_1(\mathbf{u})| < \frac{1}{S_t \max(|U_t|, |L_t|) 2^{t+p}}$$

for all $k \geq k_{\omega,t}$. For any such k and ω, Lemma 5.10 implies that

$$\left| \mathrm{d}\hat{C}_k^*(A_{st}) - \mathrm{d}C_1(A_{st}) \right| \leq \frac{2^p}{S_t \max(|U_t|, |L_t|) 2^{t+p}}$$

for any s, t. Developing T_3 yields

$$\begin{aligned} T_3 &= \left| \sum_{s=1}^{S_t} b_{st}\, \mathrm{d}\hat{\mathcal{C}}_k^*(A_{st}) - \sum_{s=1}^{S_t} b_{st}\, \mathrm{d}C_1(A_{st}) \right| \\ &\leq \sum_{s=1}^{S_t} |b_{st}| \cdot \left| \mathrm{d}\hat{\mathcal{C}}_k^*(A_{st}) - \mathrm{d}C_1(A_{st}) \right| \\ &\leq S_t \max(|U_t|, |L_t|) \frac{2^p}{S_t \max(|U_t|, |L_t|) 2^{t+p}} \\ &= \frac{1}{2^t}. \end{aligned}$$

Therefore, $\exists t_1$ such that $2^{-t} < \epsilon/5$ for all $t \geq t_1$, i.e. T_2, T_3 and T_4 are each bounded by

$\epsilon/5$ for any $t \geq t_1$ and $k \geq k_{\omega,t}$.

We can write

$$T_5 = \left| \int g(\mathbf{u}) \mathbb{1}_{B_t^C}(\mathbf{u}) \, \mathrm{d}C_1(\mathbf{u}) \right| \leq \int |g(\mathbf{u})| \mathbb{1}_{B_t^C}(\mathbf{u}) \, \mathrm{d}C_1(\mathbf{u}).$$

The integrand on the right-hand side goes to 0 as $t \to \infty$ for each $\mathbf{u} \in (0,1)^p$. Hence, the dominated convergence theorem ensures that the right-hand-side expression goes to 0 as $t \to \infty$ (the bounding function is $|g(\mathbf{u})|$). Therefore, there exists t_2 such that $T_5 < \epsilon/5$ for all $t \geq t_2$.

Turning now to T_1, Theorem 5.6 means that there exists $\Omega_{i,t} \subset \Omega$ with $P(\Omega_{i,t}) = 1$ such that for all $\omega \in \Omega_{i,t}$,

$$\frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} g\left(\mathbf{Y}_{ij}^k\right) = \int g(\mathbf{u}) \, \mathrm{d}\hat{C}_{ik}^*(\mathbf{u}) \to \int g(\mathbf{u}) \, \mathrm{d}C_i(\mathbf{u})$$

as $k \to \infty$. Consider a fixed

$$\omega \in \Omega_1 = \bigcap_{i \in \{1,\ldots,m\},\, t \in \mathbb{N}} \Omega_{i,t}.$$

The intersection is over a countable number of sets of probability 1, hence $P(\Omega_1) = 1$.

For any $\omega \in \Omega_1$, T_1 is developed as

$$\begin{aligned} T_1 &= \left| \int g(\mathbf{u}) \mathbb{1}_{B_t^C}(x) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) \right| \leq \int |g(\mathbf{u})| \mathbb{1}_{B_t^C}(\mathbf{u}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) \\ &= \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{n_{ik}} \sum_{j=1}^{n_{ik}} \left| g\left(\mathbf{Y}_{ij}^k\right) \right| \mathbb{1}_{B_t^C}\left(\mathbf{Y}_{ij}^k\right) \leq \sum_{i=1}^{m} \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \left| g\left(\mathbf{Y}_{ij}^k\right) \right| \mathbb{1}_{B_t^C}\left(\mathbf{Y}_{ij}^k\right). \end{aligned}$$

The dominated convergence theorem says that $\exists t_i^*$ such that

$$\int |g(\mathbf{u})| \mathbb{1}_{B_t^C}(\mathbf{u}) \, \mathrm{d}C_i(\mathbf{u}) < \epsilon/10m$$

for all $t \geq t_i^*$. Choose $t \geq t_3 = \max_{1 \leq i \leq m} t_i^*$. Since $\omega \in \Omega_1$, $\exists k_{i,t,\omega}$ such that for all $k \geq k_{i,t,\omega}$,

$$\frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \left| g\left(\mathbf{Y}_{ij}^k\right) \right| \mathbb{1}_{B_t^C}\left(\mathbf{Y}_{ij}^k\right) \leq \int |g(\mathbf{u})| \mathbb{1}_{B_t^C}(\mathbf{u}) \, \mathrm{d}C_i(\mathbf{u}) + \frac{\epsilon}{10m} \leq \frac{\epsilon}{5m}.$$

Therefore, $\forall t \geq \max(t_3, t_\omega^*)$, there exists $k_{\omega,t}^* = \max_{1 \leq i \leq m} k_{i,t,\omega}$ such that

$$T_1 = \left| \int g(\mathbf{u}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) - \int \tau_t(\mathbf{u}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) \right| \leq \frac{\epsilon}{5}$$

for all $k \geq k_{\omega,t}^*$.

In conclusion, for any $\omega \in \Omega_0 \cap \Omega_1$ and any $\epsilon > 0$, we can choose $t_\omega = \max(t_1, t_2, t_3, t_\omega^*)$ that yields inequalities showing that

$$\left| \int g(\mathbf{u}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) - \int g(\mathbf{u}) \, \mathrm{d}C_1(\mathbf{u}) \right| \leq \epsilon$$

for all $k \geq k_\omega(t_\omega) = \max(k_{\omega,t_\omega}, k_{\omega,t_\omega}^*)$. In other words, the left hand side of expression (5.6) converges to 0 for any $\omega \in \Omega_0 \cap \Omega_1$ with $P(\Omega_0 \cap \Omega_1) = 1$, i.e. it converges almost surely. ■

5.8 Strong Consistency of the Weighted Pseudo-Likelihood

Section 5.2 introduces the pseudo-likelihood as proposed by Genest et al. (1995). When the goal of inference is to study the dependence structure underlying the data, the heuristics of Section 2.1 can be used with distributions replaced by copulas. Therefore, the maximum pseudo-likelihood estimate can be seen as a particular case of the entropy maximization principle where the true copula is replaced by its empirical counterpart.

Suppose that p-dimensional data are available from m populations. It is natural to extend the pseudo-likelihood following the heuristics of Section 2.2. If we consider the family of copulas $c(\mathbf{u}|\theta)$ and use a mixture of empirical copulas such as $\mathcal{C}_k^*(\mathbf{u})$ as an approximation to the true underlying copula, the entropy maximization principle yields the weighted

pseudo-likelihood:

$$L(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} c\left(\mathbf{Y}_{ij}^{k} \,\middle|\, \theta\right)^{\lambda_i / n_{ik}} \tag{5.7}$$

where $\mathbf{Y}_{ij}^{k}$ are rescaled ranks as presented in Section 5.2. The maximum weighted pseudo-likelihood estimate (MWPLE) is a value of θ maximizing $L(\theta)$.

To our knowledge, the weighted pseudo-likelihood has never been suggested in the literature. This section shows that when MAMSE weights are used in conjunction with (5.7), the MWPLE is a strongly consistent estimate of the true parameter θ_0. The proof is adapted from the work of Wald (1949).

Families of Copulas

The weighted pseudo-likelihood is based on the density function of a family of copulas. However, not all copulas have a density function. In fact, any copula can be factorized as the sum of an absolutely continuous part and a singular term (see Nelsen (1999) page 23). For instance, the bivariate Marshall-Olkin copula, $C(\mathbf{u}|\alpha,\beta) = \min(u_1^{1-\alpha}u_2, u_1 u_2^{1-\beta})$, is parametrized by $0 < \alpha, \beta < 1$ and gives a positive probability to the line $u^\alpha = v^\beta$, its singular component. Such a family cannot be fitted with the pseudo-likelihood or its weighted counterpart as it does not admit a density function with respect to Lebesgue measure. As mentioned previously, all the families presented in Section 5.1 admit a density.

Let $\{c(\mathbf{u}|\theta) : \theta \in \Theta\}$ be the family of copulas that we wish to fit to the data at hand. A family of copulas that does not cover the complete range of dependence (i.e. which does not approach perfect positive or negative correlation) will typically have a compact parameter space. This is the case for the FGM copula.

Perfect concordance or discordance between two (or more) variables corresponds to a limit case where the copula places all of its mass on the lines $u_i = u_j$ or $u_i = -v_j$. For such a limit case, the copula does not admit a density anymore and its parameter space will typically be infinite or open, and hence not compact. Other limiting cases may have to

be omitted because the definition of the copula presents an undefined form; independence under the Clayton model is such an example as it would involve dividing by 0.

Despite this fact, we make the assumption that Θ is a compact set. In practice, this means that the proof of consistency that we provide will hold for (possibly constrained) families of copulas that do not include limiting cases such as perfect concordance or discordance between any 2 marginals. Relaxing Assumption 5.1 may be possible but is left to future work.

Assumption 5.1 *The set Θ is a compact subset of a finite-dimensional Cartesian space.*

Due to the constraints on their margins, copulas that admit a density will typically be smooth, hence little generality is lost with Assumption 5.2.

Assumption 5.2 *The density function of $C(\mathbf{u}|\theta)$, $c(\mathbf{u}|\theta)$, is jointly continuous for $\mathbf{u} \in (0,1)^p$ and $\theta \in \Theta$.*

Assumption 5.3 *Suppose the true copula underlying the true unknown distribution is $C(\mathbf{u}|\theta_0) \equiv C_1(\mathbf{u})$ for some $\theta_0 \in \Theta$.*

In particular, this latter assumption means that the copula underlying the true distribution admits a continuous density function.

Revised Wald's Assumptions

For all $\theta \in \Theta$ and $\rho > 0$, let us define

$$c(\mathbf{u}, \theta, \rho) = \sup_{|\theta - \theta'| \leq \rho} c(\mathbf{u}|\theta').$$

Assumption 5.4 *For a sufficiently small ρ, the expected values*

$$\int \log\left[\max\{c(\mathbf{u}, \theta, \rho), 1\}\right] \, \mathrm{d}C_1(\mathbf{u})$$

are finite.

Assumption 5.5 *If* $\theta_1 \neq \theta_0$*, then* $C(\mathbf{u}|\theta_0) \neq C(\mathbf{u}|\theta_1)$ *for at least one* $\mathbf{u}$*.*

Assumption 5.6 $\int |\log c(\mathbf{u}|\theta_0)| \, \mathrm{d}C_i(\mathbf{u}) < \infty$ *for* $i = 1, \ldots, m$*.*

Assumption 5.7 *The functions* $c(\mathbf{u}, \theta, \rho)$ *are measurable for any* θ *and* ρ*.*

Assumption 5.8 *Suppose that the functions* $\log c(u|\theta_0)$ *and* $\log c(\mathbf{u}, \theta, \rho)$ *satisfy the assumptions of Theorem 5.6 for any* θ *and* ρ*.*

Wald (1949) makes additional assumptions, but they are not required here because the copula is defined on a bounded set and because of the stronger assumptions on continuity that are made above.

Remark 5.6 *If* $\lim_{i\to\infty} \theta_i = \theta$*, then* $\lim_{i\to\infty} c(\mathbf{u}|\theta_i) = c(\mathbf{u}|\theta)$ *from the continuity of* $c(\mathbf{u}|\theta)$*.*

Remark 5.7 *The function* ϕ *of Wald also found in Section 3.7 is not required here because we assume that the set* Θ *is compact.*

Remark 5.8 *The functions* $c(x, \theta, \rho)$ *are continuous from the continuity of* $c(\mathbf{u}|\theta)$ *and Lemma 3.10.*

An Example That Satisfies the Assumptions

Before proving the main result of this section, we show that the assumptions made above are not vacuous by verifying that at least one often-used family of distributions satisfies all the stated conditions.

For this example, we use a simplified notation where $\mathbf{u} = [u, v]^{\mathsf{T}}$.

Consider the Clayton family of distributions whose density function

$$c(\mathbf{u}|\theta) = \frac{\theta + 1}{(uv)^{\theta+1}} \left(\frac{1}{u^\theta} + \frac{1}{v^\theta} - 1 \right)^{-(1/\theta+2)}$$

is indexed by $\theta \in \Theta = [a, b]$ with $0 < a < b < \infty$. Choosing a compact Θ for this family means that the limiting cases of independence and of perfect concordance are omitted.

Assumptions 5.1, 5.2 and 5.5 are satisfied. All distributions considered are defined on $(0, 1)^2$ and its associated Borelian, hence Assumption 5.7 is also satisfied. Assumption 5.3 is necessary in the context of proving consistency.

For the model considered, we show that Assumptions 5.4 and 5.6 are consequences of the bound implied by Assumption 5.8 to which we now turn our attention.

Remark 5.9 *For $u, v \in (0, 1)$, there exists a finite constant $M > 0$ such that*

$$|\log u| \leq \frac{M}{u^{1/4}} \leq \frac{M}{\{r(u)r(v)\}^{1/4}}$$

where $r(u) = u(1 - u)$ since $\log u$ is finite for all $u \in (0, 1)$, except when $u \to 0$ and

$$\lim_{u\to 0} \frac{-\log u}{u^{-1/4}} = \lim_{u\to 0} \frac{u^{-1}}{\frac{1}{4}u^{-5/4}} = 4 \lim_{u\to 0} u^{1/4} = 0$$

by l'Hospital's rule.

Remark 5.10 *For $u, v \in (0, 1)$, the function $r(u)$ is bounded between 0 and $1/4$. Therefore,*

$$\frac{1}{r(u)^{1/4}} \geq 4^{1/4} = \sqrt{2}$$

meaning that for any constant M_1, there exists a constant $M_2 = M_1/\sqrt{2}$ such that

$$M_1 \leq \frac{M_2}{r(u)^{1/4}} \leq \frac{M_2}{\{r(u)r(v)\}^{1/4}}.$$

Lemma 5.11 *The functions of Assumption 5.8 satisfy the bound implied by Theorem 5.6 for the Clayton family of copulas with $\Theta = [a, b]$ where $0 < a < b < \infty$.*

Proof of Lemma 5.11. Using the notation of Theorem 5.6, set $p = q = 2$ and $\delta = 1/2$. We

show that there exists positive finite constants M_{θ_0} and $M^*_{\rho,\theta}$ such that

$$|\log c(\mathbf{u}|\theta_0)| \leq \frac{M_{\theta_0}}{\{r(u)r(v)\}^{1/4}}$$

and

$$|\log c(\mathbf{u},\theta,\rho)| \leq \frac{M^*_{\rho,\theta}}{\{r(u)r(v)\}^{1/4}}.$$

We can write

$$\begin{aligned}
|\log c(\mathbf{u}|\theta)| &= \left|\log(\theta+1) - \left(\frac{1}{\theta}+2\right)\log\left(u^{-\theta}+v^{-\theta}-1\right) - (\theta+1)\log uv\right| \\
&\leq |\log(\theta+1)| + \left|\left(\frac{1}{\theta}+2\right)\log\left(u^{-\theta}+v^{-\theta}-1\right)\right| \\
&\qquad + |(\theta+1)\log u| + |(\theta+1)\log v| \\
&= \log(\theta+1) + (1+2\theta)\frac{1}{\theta}\log\left(u^{-\theta}+v^{-\theta}-1\right) \\
&\qquad +(\theta+1)|\log u| + (\theta+1)|\log v| \\
&\leq \frac{\frac{\log(\theta+1)}{\sqrt{2}} + (1+2\theta)M'_\theta + 2(\theta+1)M}{\{r(u)r(v)\}^{1/4}} \\
&= \frac{M_\theta}{\{r(u)r(v)\}^{1/4}}
\end{aligned}$$

by Remark 5.9 and the fact that

$$\begin{aligned}
\frac{1}{\theta}\log\left(u^{-\theta}+v^{-\theta}-1\right) &\leq \frac{1}{\theta}\log\left\{2\max(u^{-\theta},v^{-\theta})-1\right\} \\
&\leq \frac{1}{\theta}\log\left(2u^{-\theta}-1\right) + \frac{1}{\theta}\log\left(2v^{-\theta}-1\right) \\
&\leq \frac{1}{\theta}\log\left(2u^{-\theta}\right) + \frac{1}{\theta}\log\left(2v^{-\theta}\right) \\
&= |\log u| + |\log v| + \frac{2}{\theta}\log 2 \\
&\leq \frac{M}{\{r(u)r(v)\}^{1/4}} + \frac{M}{\{r(u)r(v)\}^{1/4}} + \frac{\frac{\sqrt{2}\log 2}{\theta}}{\{r(u)r(v)\}^{1/4}} \\
&= \frac{M'_\theta}{\{r(u)r(v)\}^{1/4}}
\end{aligned}$$

by Remarks 5.9 and 5.10. Note that the constant M_θ is continuous in θ for parameters θ in any compact subset of $[a, b]$.

The inequalities above holds for $\theta = \theta_0$, which is the first part of Assumption 5.8. The second part is also satisfied since

$$\log \sup_{|\theta-\theta'|\le\rho} c(\mathbf{u}|\theta') = \sup_{|\theta-\theta'|\le\rho} \log c(\mathbf{u}|\theta') \le \sup_{|\theta-\theta'|\le\rho} \frac{M_{\theta'}}{\{r(u)r(v)\}^{1/4}} \le \frac{M^*_{\rho,\theta'}}{\{r(u)r(v)\}^{1/4}}$$

because M_θ is a continuous function on the compact set $\{\theta : |\theta - \theta'| \le \rho\} \cap \Theta$, it thus achieves its maximum denoted $M^*_{\rho,\theta'}$. ■

Corollary 5.5 *Under the assumptions of Lemma 5.11, Assumptions 5.4 and 5.6 are satisfied.*

Proof of Corollary 5.5. Any of the integrals in Assumptions 5.4 and 5.6 are bounded by a positive constant times

$$\begin{aligned}
\int \frac{1}{\{r(u)r(v)\}^{1/4}} \,\mathrm{d}C_i(\mathbf{u}) &\le \int \frac{1}{\min\{r(u), r(v)\}^{1/2}} \,\mathrm{d}C_i(\mathbf{u}) \\
&\le \int_{\{\mathbf{u}:r(u)\le r(v)\}} \frac{1}{r(u)^{1/2}} \,\mathrm{d}C_i(\mathbf{u}) + \int_{\{\mathbf{u}:r(u)>r(v)\}} \frac{1}{r(v)^{1/2}} \,\mathrm{d}C_i(\mathbf{u}) \\
&\le \int \frac{1}{r(u)^{1/2}} \,\mathrm{d}C_i(\mathbf{u}) + \int \frac{1}{r(v)^{1/2}} \,\mathrm{d}C_i(\mathbf{u}) \\
&= 2\int \frac{1}{r(u)^{1/2}} \,\mathrm{d}u = \pi < \infty.
\end{aligned}$$

Hence we obtain the desired result. ■

Remark 5.11 *Corollary 5.5 guarantees that*

$$\int \log c(\mathbf{u}|\theta_0) \,\mathrm{d}C_i(\mathbf{u}) \qquad \textit{and} \qquad \int \log c(\mathbf{u}, \theta, \rho) \,\mathrm{d}C_i(\mathbf{u})$$

exist. Therefore, the assumptions of Theorem 5.6 are satisfied, i.e. Assumption 5.8 is satisfied.

Wald's Lemmas

For expectations, the following convention is adopted. Let Y be a random variable. The expected value of Y exists if $\mathrm{E}\{\max(Y,0)\} < \infty$. If $E\{\max(Y,0)\}$ is finite but $E\{\min(Y,0)\}$ is not, we say that $E\{\min(Y,0)\} = -\infty$. Moreover, a generic $\mathbf{U}$ represents a random variable with distribution $C_1(\mathbf{u}) \equiv C(\mathbf{u}, \theta_0)$.

The following lemmas are equivalent to those found in Section 3.7.

Lemma 5.12 *For any $\theta \neq \theta_0$, we have* $\mathrm{E} \log c(\mathbf{U}|\theta) < \mathrm{E} \log c(\mathbf{U}|\theta_0)$.

Lemma 5.13 $\lim_{\rho \to 0} \mathrm{E} \log c(\mathbf{u}, \theta, \rho) = \mathrm{E} \log c(\mathbf{U}|\theta)$.

Main Result

Let now turn to the main result of this section. Throughout this subsection, we suppose that Assumptions 5.1 to 5.8 are satisfied.

Theorem 5.8 *Let $\mathcal{T}$ be any closed subset of Θ that does not contain θ_0. Then,*

$$P\left[\lim_{k\to\infty} \frac{\sup_{\theta\in\mathcal{T}} \prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} c\left(\mathbf{Y}_{ij}^{k} \,\middle|\, \theta\right)^{\lambda_{ik}(\omega) n_{1k}/n_{ik}}}{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} c\left(\mathbf{Y}_{ij}^{k} \,\middle|\, \theta_0\right)^{\lambda_{ik}(\omega) n_{1k}/n_{ik}}} = 0\right] = 1.$$

Proof of Theorem 5.8: Let $\mathbf{U}$ denote a random variable with distribution $C_1(\mathbf{u}) \equiv C(\mathbf{u}|\theta_0)$. With each element $\theta \in \mathcal{T}$, we associate a positive value ρ_θ such that

$$\mathrm{E}\{\log c(\mathbf{U}, \theta, \rho_\theta)\} < \mathrm{E}\{\log c(\mathbf{U}|\theta_0)\}. \tag{5.8}$$

The existence of such ρ_θ follows from Lemmas 5.12 and 5.13. Let $S(\theta, \rho)$ denote the sphere with center θ and radius ρ. The spheres $\{S(\theta, \rho_\theta) : \theta \in \mathcal{T}\}$ form a covering of the compact set $\mathcal{T}$, hence there exists a finite sub-covering. Let $\theta_1, \ldots, \theta_h \in \mathcal{T}$ such that $\mathcal{T} \subset \bigcup_{s=1}^{h} S(\theta_s, \rho_{\theta_s})$.

Clearly,

$$0 \leq \sup_{\theta \in \mathcal{T}} \prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} c\left(\mathbf{Y}_{ij}^k \,\middle|\, \theta\right)^{\lambda_{ik}(\omega) n_{1k}/n_{ik}} \leq \sum_{s=1}^{h} \prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} c\left(\mathbf{Y}_{ij}^k, \theta_s, \rho_{\theta_s}\right)^{\lambda_{ik}(\omega) n_{1k}/n_{ik}}.$$

Therefore, to prove Theorem 5.8 if suffices to show that

$$P\left[\lim_{k\to\infty} \frac{\displaystyle\prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} c\left(\mathbf{Y}_{ij}^k, \theta_s, \rho_{\theta_s}\right)^{\lambda_{ik}(\omega) n_{1k}/n_{ik}}}{\displaystyle\prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} c\left(\mathbf{Y}_{ij}^k \,\middle|\, \theta_0\right)^{\lambda_{ik}(\omega) n_{1k}/n_{ik}}} = 0\right] = 1$$

for $s = 1, \ldots, h$. These equations can be rewritten as

$$\begin{aligned} P\Bigg[\lim_{k\to\infty} n_{1k} \Bigg[\sum_{i=1}^{m} \sum_{j=1}^{n_{ik}} \frac{\lambda_{ik}(\omega)}{n_{ik}} \log c\left(\mathbf{Y}_{ij}^k, \theta_s, \rho_{\theta_s}\right) \\ - \frac{\lambda_{ik}(\omega)}{n_{ik}} \log c\left(\mathbf{Y}_{ij}^k \,\middle|\, \theta_0\right) \Bigg] = -\infty \Bigg] \\ = P\Bigg[\lim_{k\to\infty} n_{1k} \Bigg\{ \int \log c(\mathbf{u}, \theta_s, \rho_{\theta_s}) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) \\ - \int \log c(\mathbf{u}|\theta_0) \, \mathrm{d}\hat{\mathcal{C}}_k^*(\mathbf{u}) \Bigg\} = -\infty \Bigg] = 1 \end{aligned} \tag{5.9}$$

for $s = 1, \ldots, h$.

The integrals in (5.9) converge almost surely to $\mathrm{E} \log c(\mathbf{U}, \theta_s, \rho_{\theta_s})$ and $\mathrm{E} \log c(\mathbf{U}|\theta_0)$ by Theorem 5.7. For large k, the expression inside the curly brackets is thus negative by (5.8). Hence the proof of Theorem 5.8 is complete. ■

Theorem 5.9 *Let $\hat{\theta}_k(\omega)$ be a sequence of random variables such that there exists a positive*

constant c with

$$\frac{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} c\left\{\mathbf{Y}_{ij}^{k} \,\middle|\, \hat{\theta}_k(\omega)\right\}^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}}{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} c\left(\mathbf{Y}_{ij}^{k} \,\middle|\, \theta_0\right)^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}} \geq c > 0 \tag{5.10}$$

for all $k \in \mathbb{N}$ and all $\omega \in \Omega$. Then

$$P\left\{\lim_{k\to\infty} \hat{\theta}_k(\omega) = \theta_0\right\} = 1.$$

Proof of Theorem 5.9. Let $\epsilon > 0$ and consider the values of $\hat{\theta}_k(\omega)$ as k goes to infinity. Suppose that θ_ℓ is a limit point away from θ_0, such that $|\theta_\ell - \theta_0| > \epsilon$. Then,

$$\frac{\sup_{|\theta-\theta_0|\geq\epsilon}\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} c\left(\mathbf{Y}_{ij}^{k} \,\middle|\, \theta\right)^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}}{\prod_{i=1}^{m}\prod_{j=1}^{n_{ik}} c\left(\mathbf{Y}_{ij}^{k} \,\middle|\, \theta_0\right)^{\lambda_{ik}(\omega)n_{1k}/n_{ik}}} \geq c > 0$$

infinitely often. By Theorem 5.8, this event has probability 0 even with ϵ arbitrarily small. Therefore,

$$P\left\{\omega : \left|\lim_{k\to\infty} \hat{\theta}_k(\omega) - \theta\right| \leq \epsilon\right\} = 1$$

for all $\epsilon > 0$. ∎

Corollary 5.6 *The MWPLE with MAMSE weights is a strongly consistent estimate of θ.*

Proof of Corollary 5.6. The MWPLE clearly respects Equation (5.10) with $c = 1$ since $\hat{\theta}_k(\omega)$ is then chosen to maximize the numerator of (5.10). ∎

5.9 Simulations

We study the performance of the weighted coefficients of correlations and of the weighted pseudo-likelihood with MAMSE weights in finite samples through simulations. Unless oth-

erwise specified, the number of repetitions for each simulation is sufficient to make the standard deviation of the simulation less than the last digit shown in the tables or on the figures.

5.9.1 Different Correlations for Bivariate Copulas

The bivariate families of copulas presented in Section 5.1 all depend on a single parameter and the population value of Spearman's ρ is a monotone function of that parameter. To use these families on a comparable scale, we use parameter values that are associated with a specified ρ. Four different families of copulas are considered: Normal, Clayton, Frank and Gumbel-Hougaard. Equal samples of size $n \in \{20, 50, 100, 250\}$ are simulated from five bivariate populations with different correlations. Two scenarios are adopted; they are described in Table 5.3. Each situation considered is repeated 10000 times.

	Scenario A	Scenario B
Pop. 1	0.35	0.25
Pop. 2	0.25	0.30
Pop. 3	0.30	0.35
Pop. 4	0.40	0.40
Pop. 5	0.45	0.45

Table 5.3: Values of ρ under two different scenarios that are simulated for different families of copulas.

Figure 5.1 shows the average weight allocated to each of the five populations by the MAMSE approach. Within a given scenario, the average weights do not seem to depend strongly on the actual distribution of the data. This is not very surprising since populations share the same correlation under such circumstances, and hence the shape of their copulas is similar even if they come from different families. As one would expect, the difference between Scenario A and B is bigger than the difference between two distributions within the same scenario. Under Scenario A, bias can cancel out since the correlation of the target population sits squarely within the range of correlations from the other populations. Consequently, more weight is given to Population 1 under Scenario B where that situation cannot occur.

Finally, note that under Scenario A populations 2 to 5 receive an approximately equal share of the weight, but under Scenario B, populations whose correlation is closer to the target receive a larger mass than the others. This behavior conforms with intuition.

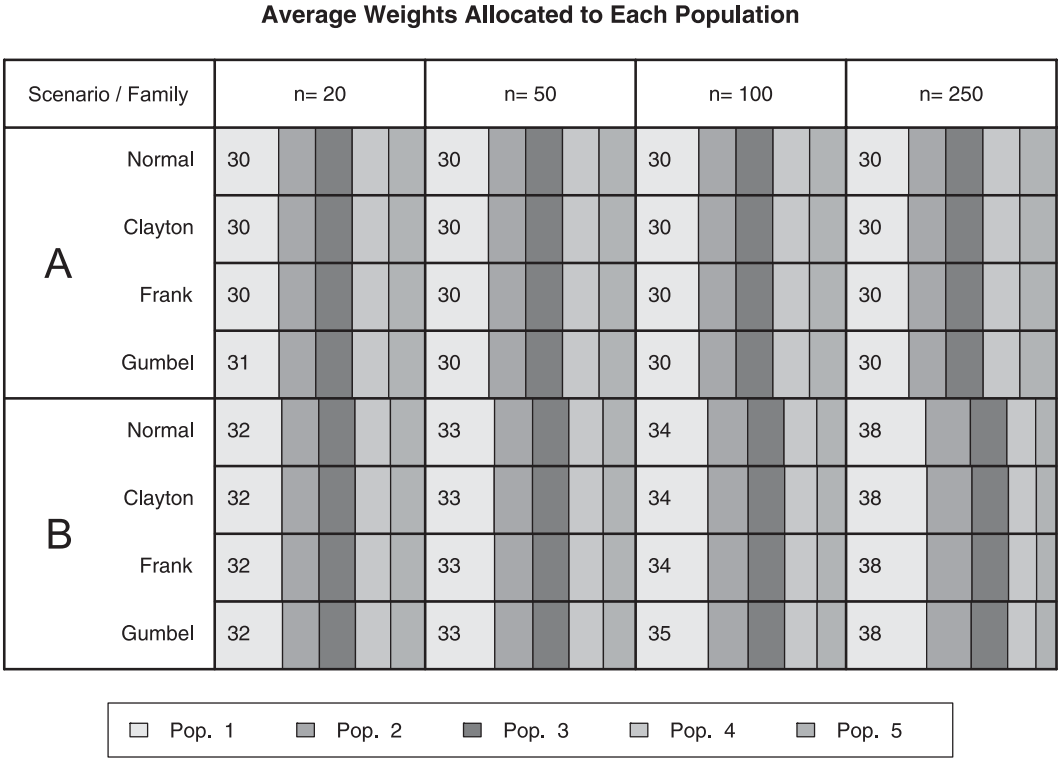

Figure 5.1: Average value of the MAMSE weights for scenarios where samples of equal sizes n are drawn from 5 different populations. The cell areas are proportional to the average weight allocated to each population. The numbers correspond to $100\bar{\lambda}_1$ and are averaged over 10000 repetitions.

Table 5.4 shows a ratio of mean squared errors comparing the usual estimate of Spearman's ρ with its MAMSE-weighted counterpart. Table 5.5 shows a similar ratio for estimating the parameter of the underlying copula with the pseudo-likelihood or the MAMSE-weighted pseudo-likelihood. Note that the error due to simulation in Table 5.5 has a standard deviation from below 6 units for $n = 20$ to less than 3 units for $n = 250$.

Under Scenario A, the performance of the MAMSE-weighted methods is impressive, featuring in most cases a MSE three times smaller than the equivalent non-weighted method.

	Family	n =20	50	100	250
Scenario A	Normal	333	331	315	271
	Clayton	335	329	314	271
	Frank	338	333	314	271
	Gumbel	331	327	312	271
Scenario B	Normal	275	197	131	72
	Clayton	284	197	131	81
	Frank	283	195	131	77
	Gumbel	286	196	136	78

Table 5.4: Performance of a MAMSE-weighted coefficient of correlation based on ranks as measured by 100 MSE($\hat{\rho}$)/MSE($\hat{\rho}_{\boldsymbol{\lambda}}$) for different scenarios and sample sizes $n \in \{20, 50, 100, 250\}$. Each figure is averaged over 10000 repetitions.

	Family	n =20	50	100	250
Scenario A	Normal	305	310	304	266
	Clayton	407	349	327	278
	Frank	398	356	325	273
	Gumbel	389	341	315	276
Scenario B	Normal	193	139	97	57
	Clayton	184	110	72	45
	Frank	245	164	112	68
	Gumbel	188	117	83	51

Table 5.5: Performance of the maximum weighted pseudo-likelihood estimate as measured by 100 MSE($\hat{\theta}$)/MSE($\hat{\theta}_{\boldsymbol{\lambda}}$) for different scenarios and sample sizes $n \in \{20, 50, 100, 250\}$. Each figure is averaged over 10000 repetitions.

Under that scenario, the correlation of the population of interest is in the middle of the range of correlations from the other populations, hence the bias can cancel out. Under Scenario B however, Populations 2 to 5 have larger correlations than the target population. Despite that fact, the methods of inference based on MAMSE weights still achieve appreciable gains in performance in many cases. The performance of the weighted methods is sometimes clearly inferior to their unweighted equivalents. Although the counter-performances seem to occur for larger sample sizes, a situation where the weighted methods are less needed, the performance of the method should be studied further before recommending its use in practice to avoid encountering such losses.

5.9.2 Different Bivariate Distributions with Common Correlation

In this simulation, we explore the possible advantages of the MAMSE-weighted methods when populations have copulas of different yet similar shapes. Five populations are used, all of which have a theoretical Spearman correlation of $\rho = 0.2$. However, the actual shape varies since the populations come from different copulas that are described in Table 5.6.

Pop. 1	:	Normal Copula, $r = 0.209$
Pop. 2	:	Frank Copula, $\theta = 1.22$
Pop. 3	:	Farlie-Gumbel-Morgenstern Copula, $\theta = 0.600$
Pop. 4	:	Gumbel-Hougaard Copula, $\theta = 1.16$
Pop. 5	:	Clayton Copula, $\theta = 0.310$

Table 5.6: Distributions from which bivariate random variables are drawn. The choice of parameters in each population yields a Spearman correlation of $\rho = 0.20$.

Note that Frank, Farlie-Gumbel-Morgenstern and the Normal copulas are all radially symmetric while those of Clayton and Gumbel-Hougaard are not. Hence, they do differ in shape although they are chosen to share a common theoretical correlation. For each sample size $n \in \{20, 50, 100, 250\}$ considered, 10000 replicates are produced. The MSE of estimating the correlation $\rho = 0.2$ and the MSE of estimating the parameter of the target distribution, a Normal model with $\theta = r = 2\sin(0.2\pi/6) \approx 0.209$, are evaluated.

Table 5.7 shows the average MAMSE weights allocated to each of the five populations considered as well as the efficiency calculated by the ratios 100 $\text{MSE}(\hat{\rho}_1)/\text{MSE}(\hat{\rho}_{\boldsymbol{\lambda}})$ and 100 $\text{MSE}(\hat{\theta}_1)/\text{MSE}(\hat{\theta}_{\boldsymbol{\lambda}})$. Note that the standard deviation of the error due to simulation can reach nearly 4 units for the efficiency at estimating θ_1. All other figures in the table respect the quoted margin of error of one standard deviation.

First note that a substantial weight is allocated to the external populations, the sample from the target population contributing a mere 31% in the inference. Note also that the remaining weight is spread rather evenly between the other populations. The gain in efficiency is clear and impressive. In fact, it corresponds approximately to the inverse of the weights allocated to Population 1 in average: $1/0.31 \approx 3.23$. When the populations

	Efficiency		$100\times$				
n	$\hat{\rho}_1$	$\hat{\theta}_1$	$\bar{\lambda}_1$	$\bar{\lambda}_2$	$\bar{\lambda}_3$	$\bar{\lambda}_4$	$\bar{\lambda}_5$
20	328	299	31	17	17	17	17
50	336	334	31	17	17	17	17
100	338	345	31	17	17	17	17
250	345	356	31	17	17	17	17

Table 5.7: Average MAMSE weights as well as the relative efficiency of the weighted correlation and of the MWPLE when compared to their non-weighted counterparts. Samples of size $n \in \{20, 50, 100, 250\}$ are drawn from five different bivariate distributions that have a common correlation of $\rho = 0.2$. Figures are averaged over 10000 repetitions.

considered are similar, the MAMSE-weighted correlation and the MWPLE clearly improve the inference. The relevant information contained in the samples from Population 2 to 5 has a clear value; that information should not be ignored.

The average weight allocated to each population does not seem to vary with n. This unexpected behavior cannot hold for an arbitrarily large sample size as it would contradict Theorem (5.3). In Table 3.1, it was noticed that the weights seem to converge very slowly when the distributions are very close to each other. The same behavior may be observed here since the populations that were simulated share a common Spearman ρ, i.e. they are quite similar to each other.

5.9.3 Multivariate Normal

The multivariate Normal copula is parametrized by its correlation matrix. In the following simulations, we will use Normal distributions in 3 and 4 dimensions, hence depending respectively on 3 and 6 parameters.

We simulate a scenario where the population of interest follows the target distribution and 3 other populations are available with the same underlying copula, but with measurement error that changes the correlations associated with their underlying copula.

The sample sizes simulated are unduly small as the MAMSE weights suffer from the curse of dimensionality: the measure M_k contains n_{1k}^p points for which operations of non-trivial

time must be performed. The time required to calculate the equations to determine the MAMSE weights increases rapidly to the hardware limitations even for moderate n_{1k} and p. There are ways in which the calculation of the MAMSE weights might be accelerated. The integrals with respect to $\mathrm{d}M_k(\mathbf{u})$ could be evaluated on a grid sparser than that associated with $M_k(\mathbf{u})$, whether it is through sampling or by a new definition of $M_k(\mathbf{u})$. Proving that the algorithm and the weights converge properly with such practical simplifications is left to future work.

Maximum Pseudo-Likelihood Estimate and its Weighted Counterpart

To calculate the pseudo-likelihood of the data, we apply the inverse CDF of a standard Normal to the rescaled ranks. For this section, we choose to use $\mathbf{Y}_{ij}^{k*} = (\mathbf{R}_{ij}^{k} - 1/2)/n_{ik}$, for a fixed k, $i = 1, \ldots, m$ and $j = 1, \ldots, n_{ik}$. We thus transform the data into

$$\mathbf{Z}_{ij} = \left[\Phi^{-1}\left(Y_{ij1}^{k*}\right), \ldots, \Phi^{-1}\left(Y_{ijp}^{k*}\right)\right]^{\mathsf{T}}.$$

By construction, the mean of the vectors $\mathbf{Z}_{ij}$, $j = 1, \ldots, n_{ik}$ is exactly $\mathbf{0}$ and the MLE of the marginal variances are

$$\frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \left\{ \Phi^{-1}\left(\frac{Y_{ij\ell}^{k} - \frac{1}{2}}{n_{ik}} \right) \right\}^2 \approx \int_0^1 \left\{\Phi^{-1}(x)\right\}^2 \mathrm{d}x = \int z^2 \, \mathrm{d}\Phi(z) = 1.$$

The actual value of the sample variance differs from 1 by a fixed amount that depends on the sample size since it determines the number of terms in the sum used above to approximate the corresponding integral.

The expression for the pseudo-likelihood based on the rescaled ranks $\mathbf{Y}_{ij}^{k*}$ is identical to the likelihood of a centered normal based on the corresponding $\mathbf{Z}_{ij}$. The maximum

likelihood estimate of Σ_i, the covariance matrix in Population i, is given by

$$\hat{\Sigma}_i = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \mathbf{Z}_{ij} \mathbf{Z}_{ij}^{\mathsf{T}}.$$

The diagonal of $\hat{\Sigma}_i$ already contains a numerical approximation of 1. Define then $\hat{\Sigma}_i^* = \hat{\Sigma}_i$, except for $diag(\hat{\Sigma}_i^*) = \mathbf{1}$, i.e. we replace the diagonal of $\hat{\Sigma}_i$ by true ones. We use $\hat{\Sigma}_i^*$ as the maximum pseudo-likelihood estimate of Σ.

The change of diagonal does not impact the positive definiteness of the matrix because the elements of the diagonal of $\hat{\Sigma}_i$ are equal and slightly smaller than 1, hence the replacement by 1 is equivalent to adding a small positive multiple of the identity matrix which is itself positive definite. We prefer the approach above to brute force numerical maximization of the pseudo-likelihood.

Now, turn to the weighted pseudo-likelihood whose expression is equivalent to the weighted likelihood for centered Normal variates with common covariance matrix Σ based on the $\mathbf{Z}_{ij}$:

$$\prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} \left\{ \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\mathbf{Z}_{ij}^{\mathsf{T}} \Sigma^{-1} \mathbf{Z}_{ij}/2} \right\}^{\frac{\lambda_{ik}(\omega)}{n_{ik}}} \propto \frac{1}{|\Sigma|^{1/2}} \prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} \exp\left(\frac{-\lambda_{ik}(\omega)}{2n_{ik}} \mathbf{Z}_{ij}^{\mathsf{T}} \Sigma^{-1} \mathbf{Z}_{ij} \right),$$

where $|\Sigma|$ denotes the determinant of the matrix Σ. The corresponding weighted log-likelihood is

$$\begin{aligned} -\frac{1}{2} \log|\Sigma| - \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{2n_{ik}} \sum_{j=1}^{n_{ik}} \mathbf{Z}_{ij}^{\mathsf{T}} \Sigma^{-1} \mathbf{Z}_{ij} &= -\frac{1}{2} \log|\Sigma| - tr\left\{ \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{2n_{ik}} \sum_{j=1}^{n_{ik}} \mathbf{Z}_{ij}^{\mathsf{T}} \Sigma^{-1} \mathbf{Z}_{ij} \right\} \\ &= -\frac{1}{2} \log|\Sigma| - tr\left\{ \Sigma^{-1} \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{2n_{ik}} \sum_{j=1}^{n_{ik}} \mathbf{Z}_{ij}^{\mathsf{T}} \mathbf{Z}_{ij} \right\} \\ &= -\frac{1}{2} \log|\Sigma| - \frac{1}{2} tr\left\{ \Sigma^{-1} A \right\} \end{aligned} \tag{5.11}$$

where

$$A = \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{n_{ik}} \sum_{j=1}^{n_{ik}} \mathbf{Z}_{ij}^{\mathsf{T}} \mathbf{Z}_{ij}.$$

Theorem 5.10 (Wichern & Johnson (2002), Result 4.10) *Given a $p \times p$ symmetric positive definite matrix B and a scalar $b > 0$, it follows that*

$$-b \log |\Sigma| - \frac{1}{2} tr(\Sigma^{-1} B) \leq -b \log |B| + pb \log(2b) - bp$$

for all positive definite $p \times p$ matrix Σ, with equality holding only for $\Sigma = (1/2b)B$.

The calculation of the MLE of the covariance matrix of a multivariate normal distribution typically uses a result akin to Theorem (5.10). Its proof can be found from different sources and is thus not reproduced here.

Applying Theorem (5.10) to Expression (5.11) with $b = 1/2$ and $B = A$, we conclude that the maximum weighted likelihood estimate of the covariance matrix Σ is given by

$$\hat{\Sigma}_{\boldsymbol{\lambda}} = \sum_{i=1}^{m} \frac{\lambda_{ik}(\omega)}{n_{ik}} \sum_{j=1}^{n_{ik}} \mathbf{Z}_{ij}^{\mathsf{T}} \mathbf{Z}_{ij} = \sum_{i=1}^{m} \lambda_{ik}(\omega) \hat{\Sigma}_i.$$

To avoid brute force optimization involving constraints on the positive definiteness of the estimate, we use the same approach as before and use

$$\hat{\Sigma}^*_{\boldsymbol{\lambda}} = \sum_{i=1}^{m} \lambda_{ik}(\omega) \hat{\Sigma}^*_i.$$

as the maximum weighted pseudo-likelihood estimate.

3-D Multivariate Normal

We suppose a multivariate normal model with measurement error. Let

$$\Sigma_A = \begin{bmatrix} 1 & 0.4 & 0.3 \\ 0.4 & 1 & 0.4 \\ 0.3 & 0.4 & 1 \end{bmatrix} \quad , \quad \Sigma_B = \begin{bmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.8 \\ 0.6 & 0.8 & 1 \end{bmatrix},$$

$$S_2 = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix} \quad , \quad S_3 = \begin{bmatrix} 0.2 & 0.1 & 0 \\ 0.1 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix}$$

$$\text{and} \qquad S_4 = \begin{bmatrix} 0.2 & -0.1 & 0 \\ -0.1 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix}.$$

We draw random vectors from multivariate Normals with means $\mathbf{0}$ and covariances that are defined in terms of the matrices above according to the formulas in the column *Covariance* of Table 5.8. Measurement errors affect the dependence structure of the populations. Let

$$\Sigma_i = \begin{bmatrix} 1 & \Gamma_{i1} & \Gamma_{i2} \\ \Gamma_{i1} & 1 & \Gamma_{i3} \\ \Gamma_{i2} & \Gamma_{i3} & 1 \end{bmatrix}$$

be the covariance matrix of Population $i = 1, \ldots, m$. The actual value of Σ_i is defined by three parameters explicitly written in Table 5.8.

Equal samples of size $n \in \{20, 35, 50\}$ are drawn from each population. The maximum pseudo-likelihood estimate of $\mathbf{\Gamma}_1 = [\Gamma_{11}, \Gamma_{12}, \Gamma_{13}]^\mathsf{T}$ and its MAMSE-weighted equivalent are calculated.

Table 5.9 shows the average weight allocated to each population. The inference relies

	Pop.	Covariance	100× Γ_{i1}	100× Γ_{i2}	100× Γ_{i3}
Scenario A	1	Σ_A	40	30	40
	2	$\Sigma_A + S_2/4$	38	29	38
	3	$\Sigma_A + S_3/4$	40	29	38
	4	$\Sigma_A + S_4/4$	38	29	36
Scenario B	1	Σ_B	80	60	80
	2	$\Sigma_B + S_2$	67	50	67
	3	$\Sigma_B + S_3$	75	50	67
	4	$\Sigma_B + S_4$	67	50	58

Table 5.8: Parameters of the simulation for 3-variate Normal variables. Population 1 is from the target distribution, but the other populations are affected by measurement errors.

strongly on the populations with measurement errors since they have a total weight of about 60% in all cases.

The weights do not seem very affected by the sample size n. The small range of values for n might be partially responsible. Recalling a similar comment from the previous simulation, it is also possible that the measurement error model produced distributions that are similar enough to make the convergence of the weights rather slow.

	n	100× $\bar{\lambda}_1$	100× $\bar{\lambda}_2$	100× $\bar{\lambda}_3$	100× $\bar{\lambda}_4$
Scenario A	20	41	19	20	20
	35	41	20	20	20
	50	41	20	20	19
Scenario B	20	37	21	22	21
	35	38	21	22	20
	50	39	20	22	19

Table 5.9: Average weight allocated to each of four 3-variate Normal distributions. Population 1 is observed from the target distribution, but the other populations contain measurement errors. The values are averaged over 5000 repetitions.

To evaluate the performance of the MWPLE, we compare its MSE to that of the MPLE based on Population 1 only. Let $\hat{\mathbf{\Gamma}}_1$ denote the maximum pseudo-likelihood estimate of $\mathbf{\Gamma}_1$

and $\hat{\mathbf{\Gamma}}_{\boldsymbol{\lambda}}$ denote its weighted counterpart. We estimate the mean squared errors

$$\mathrm{MSE}(\hat{\mathbf{\Gamma}}_1) = \mathrm{E}||\hat{\mathbf{\Gamma}}_1 - \mathbf{\Gamma}_1||^2 \qquad \text{and} \qquad \mathrm{MSE}(\hat{\mathbf{\Gamma}}_{\boldsymbol{\lambda}}) = \mathrm{E}||\hat{\mathbf{\Gamma}}_{\boldsymbol{\lambda}} - \mathbf{\Gamma}_1||^2$$

with 5000 replicates; $||\cdot||$ denoting Euclidean distance. Table 5.10 shows the values of

$$100\,\frac{\mathrm{MSE}(\hat{\mathbf{\Gamma}}_1)}{\mathrm{MSE}(\hat{\mathbf{\Gamma}}_{\boldsymbol{\lambda}})},$$

as well as an equivalent ratio of the MSE for estimating each element of the vector $\mathbf{\Gamma}_1$. Note that the standard deviation due to simulation error may reach almost 5 units under Scenario A and nearly 2.5 units under Scenario B.

		Efficiency			
	n	$\mathbf{\Gamma}_1$	Γ_{11}	Γ_{12}	Γ_{13}
Scenario A	20	243	240	260	231
	35	255	262	261	241
	50	257	264	266	243
Scenario B	20	74	68	123	47
	35	60	53	110	35
	50	52	44	105	27

Table 5.10: Relative performance of the MWPLE when compared to the MPLE as measured by 100 $\mathrm{MSE}(\hat{\mathbf{\Gamma}}_1)/\mathrm{MSE}(\hat{\mathbf{\Gamma}}_{\boldsymbol{\lambda}})$ or an equivalent ratio for the individual entries of $\mathbf{\Gamma}_1$. Population 1 is observed from a 3-variate Normal distribution and the other populations contain measurement errors. The values are averaged over 5000 repetitions.

Under Scenario A, the performances of the MWPLE are very good as its MSE is less than half that of the MPLE. However, this excellent performance does not obtain under Scenario B where the pseudo-likelihood is the winner. We can only speculate about the causes of these results at this stage, but at least two possible explanations should be explored in future research.

- The variance term in the definition of the MAMSE weights is a very rough estimate of the actual asymptotic variance of the empirical copula. This choice appears to be

reasonable for mild correlations as in Scenario A, but may degenerate as the correlation increases which is the case under Scenario B. A better penalty term can certainly be found.

- When the correlations get larger, the data points tend to cluster around a singular subset of the unit cube. The choice of the uniform measure M_k in the definition of the MAMSE weight might not be optimal in that situation.

Despite poor performances under Scenario B, the results of this section show that the MAMSE weights offer great potential to improve the quality of the inference, at least for moderate correlation and possibly for more cases if the MAMSE criterion is improved.

4-D Multivariate Normal

We suppose another multivariate normal model with measurement error. Let

$$\Sigma_A = \begin{bmatrix} 1 & 0.4 & 0.3 & 0.2 \\ 0.4 & 1 & 0.4 & 0.3 \\ 0.3 & 0.4 & 1 & 0.4 \\ 0.2 & 0.3 & 0.4 & 1 \end{bmatrix}, \quad \Sigma_B = \begin{bmatrix} 1 & 0.8 & 0.6 & 0.4 \\ 0.8 & 1 & 0.8 & 0.6 \\ 0.6 & 0.8 & 1 & 0.8 \\ 0.4 & 0.6 & 0.8 & 1 \end{bmatrix},$$

$$S_2 = \begin{bmatrix} 0.2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.2 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 0.2 & 0.1 & 0 & 0 \\ 0.1 & 0.2 & 0 & 0 \\ 0 & 0 & 0.2 & 0.1 \\ 0 & 0 & 0.1 & 0.2 \end{bmatrix}$$

$$\text{and} \qquad S_4 = \begin{bmatrix} 0.2 & -0.1 & 0 & 0 \\ -0.1 & 0.2 & 0 & 0 \\ 0 & 0 & 0.2 & -0.1 \\ 0 & 0 & -0. & 0.2 \end{bmatrix}.$$

We draw random vectors from multivariate Normal distributions with means $\mathbf{0}$ and covariances that are defined in terms of the matrices above according to the formulas in the column *Covariance* of Table 5.11. Measurement errors affect the dependence structure of the populations. The covariance matrix of Population i is written

$$\begin{bmatrix} 1 & \Gamma_{i1} & \Gamma_{i2} & \Gamma_{i3} \\ \Gamma_{i1} & 1 & \Gamma_{i4} & \Gamma_{i5} \\ \Gamma_{i2} & \Gamma_{i4} & 1 & \Gamma_{i6} \\ \Gamma_{i3} & \Gamma_{i5} & \Gamma_{i6} & 1 \end{bmatrix}$$

and depends on a vector of six parameters, $\mathbf{\Gamma}_i = [\Gamma_{i1}, \Gamma_{i2}, \Gamma_{i3}, \Gamma_{i4}, \Gamma_{i5}, \Gamma_{i6}]^\mathsf{T}$, whose values are explicitly written in Table 5.11.

			$100\times$					
	Pop.	Covariance	Γ_{i1}	Γ_{i2}	Γ_{i3}	Γ_{i4}	Γ_{i5}	Γ_{i6}
Scenario A	1	Σ_A	40	30	20	40	30	40
	2	$\Sigma_A + S_2/4$	38	29	19	38	29	38
	3	$\Sigma_A + S_3/4$	40	29	19	38	29	40
	4	$\Sigma_A + S_4/4$	36	29	19	38	29	36
Scenario B	1	Σ_B	80	60	40	80	60	80
	2	$\Sigma_B + S_2$	67	50	33	67	50	67
	3	$\Sigma_B + S_3$	75	50	33	67	50	75
	4	$\Sigma_B + S_4$	58	50	33	67	50	58

Table 5.11: Parameters of the simulations for 4-variate Normal variables. Population 1 comes from the target distribution, but the other populations are affected by measurement errors.

We simulate 5000 samples of size $n = 20$ from each of the populations and calculate both the maximum pseudo-likelihood estimate of $\mathbf{\Gamma}_1$ and its weighted counterpart.

The choice of such a small sample size is dictated by the curse of dimensionality. The measure M_k used to average the MAMSE criterion puts an equal weight on n_{1k}^p points. Doubling the sample size thus multiplies the run time by a factor greater than 16.

To accelerate the calculation of the multivariate MAMSE weights, one could try to

approximate the integrals of the MAMSE criterion by choosing a sparser M_k. Using $\mathrm{d}\hat{C}_{1k}$ instead of $\mathrm{d}M_k$ is another option to consider as it would not suffer from the curse of dimensionality in terms of run time. It might however be too sparse to detect differences between the copulas. In particular, it is not clear if the uniform convergence of the MAMSE-weighted empirical copula would hold with such a definition of the MAMSE weights. These investigations are however left to future research.

Table 5.12 shows the average weight allocated to each of the populations. Less than half of the weight is allocated to Population 1 meaning that the contribution of the other populations is quite substantial.

		$100\times$			
	n	$\bar{\lambda}_1$	$\bar{\lambda}_2$	$\bar{\lambda}_3$	$\bar{\lambda}_4$
Scenario A	20	46	18	18	18
Scenario B	20	41	20	20	19

Table 5.12: Average weight allocated to each of four 4-variate Normal distributions. Population 1 is observed from the target distribution and the other populations contain measurement errors. The values are averaged over 5000 repetitions.

Table 5.13 compares the performance of the MPLE to that of its weighted counterpart. The ratios in the table are calculated as those in the last section. The standard deviation of the error due to simulation is less than 4 units in that table.

		Efficiency						
	n	$\boldsymbol{\Gamma}_1$	Γ_{11}	Γ_{12}	Γ_{13}	Γ_{14}	Γ_{15}	Γ_{16}
Scenario A	20	235	232	234	259	225	234	229
Scenario B	20	98	58	118	214	59	130	62

Table 5.13: Relative performance of the MWPLE when compared to the MPLE for 4-variate Normal distributions. Population 1 is observed from the target distribution and the other populations contain measurement errors. The values are averaged over 5000 repetitions.

Once again, the improvement from using the weighted pseudo-likelihood is very substantial under Scenario A, but it suffers a small loss under Scenario B. The results tend to confirm that the proposed implementation of the MAMSE weights performs better for

moderate correlations. Nonetheless, the potential for improvement is clear and in this case, at least outweighs the performance losses that occur under Scenario B.

The simulations in this section show that using the MAMSE weights can improve the mean squared error in many cases, but as with most methods, it is not uniformly better over all scenarios that were considered.

Chapter 6

Summary and Future Research

This chapter presents a list of the most important original contributions contained in this thesis and sketches some of the directions that will be explored in future research.

6.1 Summary of the Work

The heuristic justification of the weighted likelihood presented in Section 2.2 provided the genesis of this thesis. That interpretation does not appear to have been exploited previously in the literature.

With these heuristics in mind, intuition suggests choosing likelihood weights that make a mixture of the m distributions close to the target distribution, but less variable than its natural estimate. The MAMSE weights are a specific implementation of that idea.

The MAMSE weights are not only likelihood weights. As a consequence of their definition, they can also be used to define a mixture of estimates of the CDF.

The general idea of the MAMSE weights involves using an estimate of the CDF of each population. Specific properties are studied for three different kinds of data. These three cases are built from nonparametric estimates of the CDF, meaning that the resulting MAMSE weights are nonparametric as well.

Univariate Data

We use the empirical distribution function to define the MAMSE criterion and prove the following properties.

- Invariance of the MWLE to a reparametrization of the model.
- The strong uniform convergence of the MAMSE-weighted empirical CDF.
- A weighted strong law of large numbers.
- The strong consistency of the MWLE with MAMSE weights.

In addition, simulations show that using the MAMSE weights allows for superior performances in many scenarios.

Censored Data

We use the Kaplan-Meier estimate to accommodate right-censored data. The main contributions are:

- the weighted Kaplan-Meier estimate, a fully nonparametric estimate of the survival function that uses data from the m populations at hand, and
- the uniform convergence thereof.

Simulations show possible improvements for inference on finite samples when using the MAMSE-weighted Kaplan-Meier estimate.

Multivariate Data Through Copulas

We treat the dependence structure of multivariate data through copulas. The empirical copula, an estimate based on the ranks of the data, is used to that end. Important contributions are as follows.

- Definition of the weighted empirical copula and a proof of its strong uniform convergence to the target copula.
- A weighted strong law of large numbers for bounded multivariate rank order statistics.

- A weighted coefficient of correlation based on MAMSE weights; the proof that this estimate is strongly consistent.
- The weighted pseudo-likelihood for fitting a family of copulas on data from m populations.
- A proof that the MWPLE is a strongly consistent estimate of the parameter of interest.

Here again, simulations showed that the methods proposed are powerful tools when appropriately applied.

The sum of these results show that the MAMSE weights succeed in trading bias for precision for different types of data.

6.2 Future Research

The idea of MAMSE weights and the intuitive understanding of the weighted likelihood open many directions for future research. We list some of these below.

Principle of the MAMSE Weights

- Using different functions to combine bias and variance. For instance, it was recently brought to my attention that statistics of the form $\int\{F(x)-G(x)\}^2\Psi(x)\,\mathrm{d}F(x)$ have been studied for goodness-of-fit. The choice $\Psi(x)\equiv 1$ that corresponds to the bias term in the MAMSE criterion is also known as the Cramer-von Mises test statistic. However, it seems that the choice $\Psi(x)=F(x)\{1-F(x)\}$ which corresponds to Anderson-Darling goodness-of-fit test typically offers a more powerful alternative. Hence, we intend to explore the properties of a different MAMSE criterion where the bias term adopts the form of the Anderson-Darling statistic.
- Using parametric estimates of the distributions in the general formulation of the

MAMSE weights (rather than nonparametric estimates) may allow one to account for nuisance parameters or for covariates.

Univariate Data

- The rate of convergence of the MAMSE weights needs further study.
- It would be useful to describe the asymptotic distribution of a MAMSE-weighted sum of variables and even more useful to determine the asymptotic distribution of the MWLE.
- The weights implied by the Stein-type estimate of Ahmed (2000) could be studied as possible likelihood weights. Alternatively, the approach of Ahmed (2000) may be extended to the empirical distribution functions to define Stein-like likelihood weights.

Censored Data

- MAMSE weights could be used as likelihood weights. The fact that the weighted Kaplan-Meier estimate may not converge on the complete real line is a challenge. Defining the MAMSE weights based on parametric models rather than on the Kaplan-Meier estimates may help in building weights that will make the MWLE consistent.
- The idea of the MAMSE weights might be extendable to a Cox proportional hazard model, something that warrants further investigations given the great importance of that model.

Copulas

- Defining and studying the properties of a weighted version of Kendall's τ.
- Exploring different definitions of the MAMSE weights that suffer less from the curse of dimensionality.

- Studying the effect of using different estimates for the variance of the empirical copula. In particular, verifying the hypothesis that it may improve the performance of the MWPLE in the presence of strong dependence.

- Using the weighted pseudo-likelihood in other contexts where the weights need not be estimated from the data. For instance, if data were available from populations with the same dependence structure, but different marginals (e.g. different currencies), or if m sets of rankings are the only information available.

- Proposing weighted coefficients of correlation under scenarios where the weights do not need to be determined from the data.

- Using the MWPLE and univariate MWLEs in order to build an estimate of a multivariate distribution function by combining the estimated copula with estimated marginal distributions. An important advantage of such an approach is the possibility to use a different number of populations for each of the marginal distributions. Suppose that data are available from some studies who did not record all the variables of interest. The information from these studies could be used for the inference on the marginals, but only the complete data could be used to determine the dependence structure.

- One approach to building multivariate distributions may be obtained by extending the work of Cox & Reid (2004) where the likelihoods would be replaced by MAMSE-weighted likelihoods. Such an extension may however require a definition of MAMSE weights for multivariate distributions that are not copulas.

Bootstrap

In this document, we propose three weighted nonparametric asymptotically unbiased empirical estimates of a target distribution. Resampling techniques such as the bootstrap could be developed from these weighted mixtures of empirical functions. To use such methods

with confidence, we should first demonstrate that the bootstrap estimators obtained that way are consistent.

One advantage of bootstrapping from a MAMSE-weighted mixture of empirical functions is the ability to draw from a larger number of values, resulting in a bootstrap sample with fewer ties. One could even try including a distribution as an external population, i.e. to allow for an infinite sample, hence further reducing the number of ties in the bootstrap sample.

The Stein-like quantile estimate of Ahmed (2000) implies an empirical distribution that could also be considered for the definition of an alternative bootstrap method.

Foundations of Statistics

In addition to these directions for research, this work may have shed a new light on likelihood methods for some readers; writing it definitely changed my understanding of this important statistical tool. In particular, the likelihood can be seen as an estimate of the entropy between the fitted model and the EDF. The convergence properties of the CDF can then be used to show the consistency of the MLE. By showing that the MAMSE-weighted EDF converges, we could also show that a MWLE based on MAMSE weights is consistent.

A question arises from that interpretation of the likelihood: Is the likelihood a good method because the empirical distribution function is a good estimate, or is the likelihood a more general principle?

Can any model be fitted successfully using the likelihood? Using the likelihood for linear model based on the normal distribution for instance is definitely reasonable, especially if when we consider that maximizing the normal likelihood is equivalent to a form of least squares. But what about complex models, possibly involving nonlinear covariates? Should we avoid using the likelihood blindly or is it safe in all situations?

Chapter 7

Conclusion

The weighted likelihood as proposed by Hu, Wang and Zidek suggests that information from populations similar to that of interest should not be ignored, but rather incorporated into the inference.

When I started working on the weighted likelihood, the question that everybody asked first was: "But how do you choose the weights?" The MAMSE weights provide one answer to that question which seemed to be of most concern to those who were not acquainted with the weighted likelihood.

This thesis does not say how to choose optimal likelihood weights. In fact, the answer to that question will depend on the purpose of the inference, more particularly on the loss function that one wants to use.

This thesis however offers a practical and effective solution to determine likelihood weights. These weights are shown to produce consistent estimates, yet they successfully trade bias for precision in many cases that were simulated.

The different cases of the MAMSE weights studied in detail in this document are all based on nonparametric estimates of the distribution functions. Hence, the proposed MAMSE weights are themselves nonparametric.

For all cases studied, the MAMSE-weighted empirical distributions are shown to converge uniformly to the target distribution without assuming any particular shape and minimal regularity conditions for the other populations. Hence, the estimates suggested are asymptotically unbiased even though they take into consideration unknown data.

With MAMSE weights, the maximum weighted likelihood estimate and the maximum

weighted pseudo-likelihood estimate converge strongly under conditions that are akin to those imposed on their unweighted counterparts.

Finally, simulations show that the MAMSE weights can indeed improve the quality of inference on finite samples in many situations.

The results presented in this document show that when different sources of data are available, it is possible to use them to improve the quality of the inference. Hence it is possible with the MAMSE weights to use the data from m populations in order to evaluate their level of dissimilarity and then to use that information to successfully trade bias for precision.

Bibliography

E. S. Ahmed, A. I. Volodin & A. A. Hussein (2005). Robust weighted likelihood estimation of exponential parameters, *IEEE Transactions on Reliability*, **54**, 389–395.

E. S. Ahmed (2000). Stein-type shrinkage quantile estimation, *Stochastic Analysis and Applications*, **18**, 475–492.

H. Akaike (1977). On entropy maximization principle, *Applications of Statistics*, 27–42.

D. Blest (2000). Rank correlation—an alternative measure, *Australian and New Zealand Journal of Statistics*, **42**, 101–111.

N. E. Breslow & J. Crowley (1974). A large sample study of the life table and product limit estimates under random censorship, *The Annals of Statistics*, **2**, 437–453.

K. Chen & S.-H. Lo (1997). On the rate of uniform convergence of the product-limit estimator: Strong and weak laws, *The Annals of Statistics*, **25**, 1050–1087.

U. Cherubini, E. Luciano & W. Vecchiato (2004). *Copula Methods in Finance*. Wiley Finance, Wiley, Hoboken.

D. G. Clayton (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, **65**, 141–151.

D. R. Cox & N. Reid (2004). A note on pseudolikelihood constructed from marginal densities, *Biometrika*, **91**, 729–737.

P. Deheuvels (1979). La fonction de dépendance empirique et ses propriétés: un test non paramétrique d'indépendance, *Académie royale de Belgique–Bulletin de la classe des sciences*, **65**, 274–292.

R. Durrett (2005). *Probability: Theory and Examples, Third Edition*. Duxbury Advanced Series, Thomson, Belmont.

B. Efron (1967). The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California, **4**, 831–853.

J.-D. Fermanian, D. Radulovic, M. Wegkamp (2004). Weak convergence of empirical copula processes, *Bernoulli*, **10**, 847-860.

A. Földes & L. Rejtö (1981). Strong uniform consistency for nonparametric survival curve estimators from randomly censored data. *The Annals of Statistics*, **9**, 122–129.

M. J. Frank (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae*, **19**, 194–226.

C. Genest, K. Ghoudi & L.-P. Rivest (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, **82**, 543–552.

C. Genest & J. MacKay (1986). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canadian Journal of Statistics*, **14**, 145–159.

C. Genest & J.-F. Plante (2003). On Blest's measure of rank correlation, *The Canadian Journal of Statistics*, **31**, 35–52.

F. Hu (1994). *Relevance Weighted Smoothing and a New Bootstrap Method*, unpublished doctoral dissertation, Department of Statistics, The University of British Columbia.

F. Hu & J. V. Zidek (1993). *A Relevance Weighted Nonparametric Quantile Estimator*. Technical report no. 134, Department of Statistics, The University of British Columbia, Vancouver.

F. Hu & J. V. Zidek (2001). The relevance weighted likelihood with applications, *Empirical Bayes and Likelihood Inference*, 211–235.

F. Hu & J. V. Zidek (2002). The weighted likelihood, *The Canadian Journal of Statistics*, **30**, 347–371.

H. Joe (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

R. A. Johnson & D. W. Wichern (2002). *Applied Multivariate Statistical Analysis, Fifth Edition*, Prentice Hall, Upper Saddle River.

E. L. Kaplan & P. Meier (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457–481.

A. M. Krieger & D. Pfeffermann (1992). Maximum likelihood estimation from complex sample surveys, *Survey Methodology*, **18**, 225–239.

D. G. Luenberger (2003). *Linear and Nonlinear Programming, Second Edition*, Kluwer, Norwell.

M. Markatou, A. Basu & B. Lindsay (1997). Weighted likelihood estimating equations: The discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference*, **57**, 215-232.

National Center for Health Statistics (1997). *U.S. Decennial Life Tables for 1989–91*, vol. 1, no. 1, Hyattsville, Maryland.

R. B. Nelsen (1999). *An Introduction to Copulas*, Lecture Notes in Statistics No. 139, Springer, Berlin.

D. Oakes (1986). Semiparametric inference in a model for association in bivariate survival data, *Biometrika*, **73**, 353–361.

J. Pinto da Costa & C. Soares (2005). A weighted rank measure of correlation, *Australian and New Zealand Journal of Statistics*, **47**, 515–529.

J.-F. Plante (2002). *À propos d'une mesure de corrélation des rangs de Blest*, unpublished Master's Thesis, Département de mathématiques et de statistique, Université Laval.

A. Sklar (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, **8**, 229–231.

Statistics Canada (2006). *Life Tables, Canada, Provinces and Territories. Reference Period: 2000 to 2002.*, catalog number 84-537-XIE, Ottawa, Canada.

H. Tsukahara (2005). Semiparametric estimation in copula models. *The Canadian Journal of Statistics*, **33**, 357–375.

C. van Eeden & J. V. Zidek (2004). Combining the data from two normal populations to estimate the mean of one when their means difference is bounded, *Journal of Multivariate Analysis*, **88**, 19–46.

A. Wald (1949). Note on the consistency of the maximum likelihood estimate, *The Annals of Mathematical Statistics*, **20**, 595–601.

X. Wang (2001). *Maximum Weighted Likelihood Estimation*, unpublished doctoral dissertation, Department of Statistics, The University of British Columbia.

X. Wang, C. van Eeden & J. V. Zidek (2004). Asymptotic properties of maximum weighted likelihood estimators, *Journal of Statistical Planning and Inference*, **119**, 37–54.

X. Wang & J. V. Zidek (2005). Selecting likelihood weights by cross-validation, *The Annals of Statistics*, **33**, 463–501.

B. B. Winter, A. Földes & L. Rejtö (1978). Glivenko-Cantelli theorems for the PL estimate. *Problems of Control and Information Theory*, **7**, 213–225.