Asymptotic Properties of the MAMSE Adaptive Likelihood Weights

Jean-François Plante, *University of Toronto*

Abstract. The weighted likelihood is a generalization of the likelihood designed to borrow strength from similar populations while making minimal assumptions. If the weights are properly chosen, the maximum weighted likelihood estimate may perform better than the maximum likelihood estimate. In a previous article, the minimum averaged mean squared error (MAMSE) weights are proposed and simulations show that they allow to outperform the MLE in many cases. In this paper, we study the asymptotic properties of the MAMSE weights. In particular, we prove that the MAMSE-weighted mixture of empirical distribution functions converges uniformly to the target distribution and that the maximum weighted likelihood estimate is strongly consistent. A short simulation illustrates the use of bootstrap in this context.

MSC: primary 62F1;62G20, secondary 62F10.

Keywords: Statistical Inference, Weighted Likelihood, Maximum Weighed Likelihood Estimate, Nonparametrics, Asymptotics, Strong Consistency, Borrowing Strength, Bootstrap.

1 Introduction

The weighted likelihood is a frequentist method that allows to borrow strength from datasets that do not follow the target distribution exactly. This work is a sequel to that of Hu (1994), later published as Hu & Zidek (2002), who designed the weighted likelihood in order to take advantage of the relevant information contained in such samples. In the formulation of the weighted likelihood, an exponential weight discounts the contribution of each datum based on the discrepancy of its distribution with that of the target population.

In the context of dependent data (e.g. smoothing), Hu et al. (2000b) use covariates to determine likelihood weights, but not the response variables themselves. In a different setting where the distribution of data stabilizes through time, Hu & Rosenberg (2000a) use weights that are determined by a function whose parameter is set by minimizing the mean squared error of the resulting estimate.

Although the initial paradigm of the weighted likelihood allows each datum to come from a different population, we rather adopt the same framework as Wang (2001), Wang & Zidek (2005) and Wang et al. (2004) where data come as samples from m populations. In this context, one could hope to set the weights based on scientific information, but it is more pragmatic and less arbitrary to determine them based on the data.

Under this paradigm, neither an ad-hoc method suggested by Hu & Zidek (2002) nor the cross-validation method explored by Wang & Zidek (2005) provide a satisfactory recipe for finding likelihood weights. The cross-validation weights, for instance, lack numerical stability. Recently, Plante (2008) suggested nonparametric adaptive weights whose formulation is based on heuristics showing that the weighted likelihood is a special case of the entropy maximization principle. Simulations show that the so-called MAMSE (minimum averaged mean squared error) weights allow to outperform the likelihood under many scenarios.

Competing methods that borrow strength from a fixed number of samples typically rely on a hierarchical model. By opposition, the MAMSE-weighted likelihood does not require to model the extra populations and hence cannot be negatively affected by model misspecification on a population of secondary interest. In situations where no hierarchical model arise naturally, this may constitute a major advantage.

The asymptotic properties of the weighted likelihood are studied by Hu (1997) for weights that do not depend on the data. Asymptotics for adaptive weights are developed by Wang et al. (2004) under the assumption that the weights asymptotically shift towards Population 1 at a certain rate. As Plante (2008) points out, the MAMSE weights do not follow this behavior and hence require a special treatment.

In this paper, we study the asymptotic properties of the MAMSE weights, the MAMSE-weighted mixture of empirical distribution functions and of the corresponding maximum weighted likelihood estimate (MWLE). In Section 2, we introduce the weighted likelihood and the MAMSE weights formally. A sequence of lemmas is presented in Section 3 to show that a MAMSE-weighted mixture of empirical distribution functions converges uniformly to the target distribution. In Section 4, we prove that the MWLE is a strongly consistent estimate by generalizing the proof of Wald (1949) for the likelihood. Section 5 discusses the asymptotic behavior of the MAMSE weights themselves. The use of bootstrap methods is illustrated through simulations in Section 6. The MAMSE-weighted MWLE offers better performances than the maximum likelihood estimate (MLE) in many cases, yielding good coverage for shorter bootstrap confidence intervals.

2 The Weighted Likelihood and the MAMSE Weights

We introduce a notation that allows for increasing sample sizes as it will be useful for the remaining of this manuscript.

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.

Suppose that Population 1 is of inferential interest. If we denote by $f(x|\theta)$ the family of distributions used to model Population 1, the weighted likelihood and the weighted log-likelihood are written

$$L(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(X_{ij}|\theta)^{\lambda_i/n_i} \qquad \text{and} \qquad \ell(\theta) = \sum_{i=1}^{m} \frac{\lambda_i}{n_i} \sum_{j=1}^{n_i} \log f(X_{ij}|\theta)$$

where the $\lambda_i \geq 0$ are likelihood weights such that $\sum_{i=1}^m \lambda_i = 1$.

Let $\hat{F}_{ik}(x) = (1/n_{ik}) \sum_{j=1}^{n_{ik}} \mathbb{1}(X_{ij} \leq x)$ be the empirical distribution function (EDF) based on the sample X_{ij}, $j = 1, \ldots, n_{ik}$. The empirical measure associated with $\hat{F}_{ik}(x)$ allocates a weight $1/n_{ik}$ to each of the observations X_{ij}, $j = 1, \ldots, n_{ik}$.

Plante (2008) shows heuristically that maximizing the weighted likelihood is comparable to maximizing the proximity between the model $f(x|\theta)$ and a mixture of the m empirical distribution functions obtained from the samples at hand. Such a mixture was considered before by Hu & Zidek (1993), Hu (1994) and Hu & Zidek (2002) who called it relevance weighted empirical distribution function (REWED). By comparison, the usual likelihood is akin to maximizing the entropy between $f(x|\theta)$ and $\hat{F}_{1k}(x)$.

Inspired by the heuristics briefly described above, Plante (2008) tries to find weights that make the mixture of EDFs $\sum_{i=1}^m \lambda_i \hat{F}_{ik}(x)$ close to $\hat{F}_{1k}(x)$, but less variable. He proposes the MAMSE objective function.

Some preprocessing steps first discard any sample whose range of values does not overlap with that of Population 1. For the remaining m samples, we write $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^\mathsf{T}$ and minimize

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

as a function of $\boldsymbol{\lambda}$ under the constraints $\lambda_i \geq 0$ and $\sum_{i=1}^m \lambda_i = 1$. We proceed to the substitution

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

in Equation (1) based on the variance of the Binomial variable $n_{ik}\hat{F}_i(x)$ for fixed x. The choice of $\mathrm{d}\hat{F}_{1k}(x)$ allows to integrate where the target distribution $F_1(x)$ has most of its mass.

The MAMSE weights are the solution to the constrained minimization of Equation (1), we denote them by $\boldsymbol{\mu}_k = [\mu_{1k}, \ldots, \mu_{mk}]^\mathsf{T}$, hence $\boldsymbol{\mu}_k$ is a random variable on the probability space $(\Omega, \mathcal{B}(\Omega), P)$.

The MAMSE weights are used to define an estimate of the distribution $F_1(x)$,

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

the MAMSE-weighted EDF.

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 are excluded by the preprocessing.

Note that the objective function $P_k(\boldsymbol{\lambda})$ is quadratic and may also be written as

$$P_k(\boldsymbol{\lambda}) = \tilde{\boldsymbol{\lambda}}^\mathsf{T} \bar{\mathsf{A}}_k \tilde{\boldsymbol{\lambda}} - 2\tilde{\boldsymbol{\lambda}}^\mathsf{T} \mathbf{1} \bar{b}_k + \bar{b}_k \tag{2}$$

where

$$\tilde{\boldsymbol{\lambda}} = \begin{bmatrix} \lambda_2 \\ \vdots \\ \lambda_m \end{bmatrix}, \quad \mathcal{F}_k(x) = \begin{bmatrix} \hat{F}_{1k}(x) - \hat{F}_{2k}(x) \\ \vdots \\ \hat{F}_{1k}(x) - \hat{F}_{mk}(x) \end{bmatrix}, \mathbf{V}_k(x) = \begin{bmatrix} \widehat{\mathrm{var}}\{\hat{F}_{2k}(x)\} & & 0 \\ & \ddots & \\ 0 & & \widehat{\mathrm{var}}\{\hat{F}_{mk}(x)\} \end{bmatrix},$$

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

The constraint $\sum_{i=1}^m \lambda_i = 1$ is embedded in Equation (2), but the constraints $\lambda_i \geq 0$ must be controlled manually. See Plante (2007) or Plante (2008) for more details on the calculation of the MAMSE weights.

3 Asymptotic Properties

This section explores the large sample behavior of the weighted EDF. The proofs are deferred to the Appendix.

Lemma 3.1. *For any $\omega \in \Omega$ and $k \in \mathbb{N}$, $\int \left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right|^2 \mathrm{d}\hat{F}_{1k}(x) \leq \left(\frac{n_{1k}^2 - 1}{n_{1k}^2}\right)\frac{1}{6n_{1k}}$.*

Lemma 3.2. *There exists $\Omega_1 \subset \Omega$ with $P(\Omega_1) = 1$ such that for all $\omega \in \Omega_1$ and any fixed $k \in \mathbb{N}$,*

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

Lemma 3.3. *Let a_k be an infinite sequence of positive 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 $\forall \omega \in \Omega_1$, $a_k \max_{j\in\{1,\ldots,n_{1k}\}} \left|\hat{G}_k(X_{1j}) - \hat{F}_{1k}(X_{1j})\right| \leq \epsilon$ for all $k \geq k_0$.*

Lemma 3.4. *There exists $\Omega_1 \subset \Omega$ with $P(\Omega_1) = 1$ such that for all $\epsilon > 0$, there exists 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 $\omega \in \Omega_1$.*

Theorem 3.1. *The random variable $\sup_x \left|\hat{G}_k(x) - F_1(x)\right|$ converges almost surely to 0.*

The MAMSE-weighted mixture of EDF thus converges uniformly to the target distribution under rather weak assumptions. In particular, Populations 2 to m are only required to have continuous distributions. We see next that this implies a weighted strong law of large numbers.

Lemma 3.5. *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 $A \subset \mathbb{R}$, $|\,\mathrm{d}F(A) - \mathrm{d}G(A)| \leq 2\epsilon$.*

Theorem 3.2. *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 points $\{d_1,\ldots,d_L\}$. 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 $k \to \infty$.*

Corollary 3.1. The 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 $\sum_{i=1}^{m}(\mu_{ik}/n_{ik}) \sum_{j=1}^{n_{ik}} X_{ij} \to E(X_1)$ almost surely as $k \to \infty$.*

Theorem 3.2 permits to prove the consistency of the MWLE by extending the proof of Wald (1949). This extension is given next.

4 Consistency of the MWLE

In this section, we adapt the work of Wald (1949) to prove that the MWLE obtained with MAMSE weights is a strongly consistent estimate.

4.1 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 $\theta \in \Theta$, 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): $\forall \theta \in \Theta, \rho > 0$, $f(x, \theta, \rho) = \sup_{|\theta - \theta'| \le \rho} f(x|\theta')$ and $\forall r > 0$, $\phi(x, r) = \sup_{|\theta| > r} f(x|\theta)$. In addition, $f^*(x) = \max\{f(x), 1\}$.

Assumption 1. 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 2. For sufficiently small ρ and sufficiently large r, the expressions $\int \log f^*(x, \theta, \rho)\, \mathrm{d}F_1(x)$ and $\int \log \phi^*(x, r)\, \mathrm{d}F_1(x)$ are finite.

Assumption 3. If $\lim_{i\to\infty} \theta_i = \theta$, then $\lim_{i\to\infty} f(x|\theta_i) = f(x|\theta)$.

Assumption 4. If $\theta_1 \neq \theta_0$, then $F(x|\theta_0) \neq F(x|\theta_1)$ for at least one x.

Assumption 5. If $\lim_{i\to\infty} |\theta_i| = \infty$, then $\lim_{i\to\infty} f(x|\theta_i) = 0$.

Assumption 6. $\int |\log f(x|\theta_0)|\, \mathrm{d}F_i(x) < \infty$ for $i = 1, \ldots, m$.

Assumption 7. The parameter space Θ is a closed subset of a finite-dimensional Cartesian space.

Assumption 8. The functions $f(x, \theta, \rho)$ and $\phi(x, r)$ are measurable for any θ, ρ and r.

Assumption 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 1 to 8 are from Wald (1949); only Assumption 6 is modified to cover the m populations of our paradigm. Assumption 9 is required to ensure that Theorem 3.2 applies. Lemmas 4.4 and 4.5 of Section 4.2 help in determining if the family of distributions $F(x|\theta)$ respects this new assumption.

Note that the assumptions above are mostly concerned with the family of distributions $F(x|\theta)$, the model, rather than the true distribution of the data.

4.2 Wald's Lemmas

Wald's lemmas (4.1, 4.2 and 4.3) do not need to be modified. We state them for completeness, but do not reproduce their proofs already 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 $E\{\min(U,0)\} = -\infty$. Moreover, a generic X represents a random variable with distribution $F_1(x) \equiv F(x|\theta_0)$.

Lemma 4.1. *For any $\theta \neq \theta_0$, we have $\mathrm{E}\log f(X|\theta) < \mathrm{E}\log f(X|\theta_0)$.*

Lemma 4.2. $\lim\limits_{\rho\to 0} \mathrm{E}\log f(X,\theta,\rho) = \mathrm{E}\log f(X|\theta)$.

Lemma 4.3. *The equation $\lim_{r\to\infty} \mathrm{E}\log\phi(X,r) = -\infty$ holds.*

The next two lemmas are useful in determining if Assumption 9 is respected; their proof is found in the Appendix.

Lemma 4.4. *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.*

By Lemma 4.4, 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 4.5, 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, $\lim_{\delta\to 0} \omega_g(\delta,x_0)/\delta = |g'(x_0)|$.

Lemma 4.5. *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.*

Note that if $g(x)$ is continuous at x_0, $\lim_{\delta\to 0}\omega_g(\delta,x_0) = 0$. In fact, the modulus of continuity $\omega_{f(\cdot,\theta)}(\delta,x_0)$ will have a lower bound as $\delta \to 0$ when: (1) there is a discontinuity in the function $f(x|\theta)$ at x_0, or (2) 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 3.2 which allows a finite number of discontinuities. Therefore, as long as $f(x|\theta)$ is a continuous model such that the set

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

is empty or made up of a finite number of singletons, Assumption 9 will hold.

4.3 Main Result

We now turn to the main results of this section.

Theorem 4.1. *Let $\mathcal{T}$ be any closed subset of Θ that does not contain θ_0. Then,*

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

Proof of Theorem 4.1. 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

$$\mathrm{E}\{\log\phi(X,r_0)\}<\mathrm{E}\{\log f(X|\theta_0)\}. \tag{3}$$

The existence of such an r_0 follows from Lemma 4.3 and Assumption 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

$$\mathrm{E}\{\log f(X,\theta,\rho_\theta)\}<\mathrm{E}\{\log f(X|\theta_0)\}. \tag{4}$$

The existence of such ρ_θ follows from Lemma 4.1 and 4.2. 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 $\theta_1,\dots,\theta_h\in\mathcal{T}_1$ such that $\mathcal{T}_1\subset\bigcup_{s=1}^h S(\theta_s,\rho_{\theta_s})$. Clearly,

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

Therefore, to prove Theorem 4.1 if suffices to show that

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

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

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

The above equations can be rewritten as

$$\begin{aligned}&P\left[\lim_{k\to\infty}n_{1k}\left\{\sum_{i=1}^m\sum_{j=1}^{n_{ik}}\frac{\mu_{ik}}{n_{ik}}\log f(X_{ij},\theta_s,\rho_{\theta_s})-\frac{\mu_{ik}}{n_{ik}}\log f(X_{ij}|\theta_0)\right\}=-\infty\right] \\ &=\ P\left[\lim_{k\to\infty}n_{1k}\left\{\int\log f(x,\theta_s,\rho_{\theta_s})\,\mathrm{d}\hat{G}_k(x)-\int\log f(x|\theta_0)\,\mathrm{d}\hat{G}_k(x)\right\}=-\infty\right]=1\end{aligned}\tag{5}$$

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

$$\begin{aligned}&P\left[\lim_{k\to\infty}n_{1k}\left\{\sum_{i=1}^m\sum_{j=1}^{n_{ik}}\frac{\mu_{ik}}{n_{ik}}\log\phi(X_{ij},r_0)-\frac{\mu_{ik}}{n_{ik}}\log f(X_{ij}|\theta_0)\right\}=-\infty\right] \\ &=\ P\left[\lim_{k\to\infty}n_{1k}\left\{\int\log\phi(x,r_0)\,\mathrm{d}\hat{G}_k(x)-\int\log f(x|\theta_0)\,\mathrm{d}\hat{G}_k(x)\right\}=-\infty\right]=1\end{aligned}\tag{6}$$

respectively. Assumptions 6 and 9 insure that Theorem 3.2 applies to the integrals above, each of these converging almost surely to $\int \log f(x, \theta_s, \rho_{\theta_s})\, \mathrm{d}F_1(x)$, $\int \log \phi(x, r_0)\, \mathrm{d}F_1(x)$ or $\int \log f(x|\theta_0)\, \mathrm{d}F_1(x)$. Combining this result with Equations (3) and (4), we have that (5) and (6) hold. Hence the proof of Theorem 4.1 is complete. □

Theorem 4.2. *Let $\hat{\theta}_k(\omega)$ be a sequence of random variables such that there exists a positive constant c with*

$$\frac{\prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} f\{X_{ij}|\hat{\theta}_k(\omega)\}^{\mu_{ik} n_{1k}/n_{ik}}}{\prod_{i=1}^{m} \prod_{j=1}^{n_{ik}} f(X_{ij}|\theta_0)^{\mu_{ik} n_{1k}/n_{ik}}} \geq c > 0 \tag{7}$$

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

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

infinitely often. By Theorem 4.1, this event has probability 0 even with ϵ arbitrarily small. Therefore, $P\{\omega : |\lim_{k\to\infty} \hat{\theta}_k(\omega) - \theta_0| \leq \epsilon\} = 1$ for all $\epsilon > 0$. □

Corollary 4.1. *The MWLE is a strongly consistent estimate of θ.*

Proof of Corollary 4.1. The MWLE clearly satisfies Equation (7) with $c = 1$ because $\hat{\theta}_k(\omega)$ is then chosen to maximize the numerator of (7). □

5 Asymptotic Behavior of the 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 5.1. *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 $||\boldsymbol{\lambda}^* - \boldsymbol{\mu}_k|| > \epsilon$ i.o. for all $\omega \in \Omega_0$ and hence the MAMSE weights do not converge to $\boldsymbol{\lambda}^*$ as $k \to \infty$.*

Corollary 5.1. *Consider the sequence of MAMSE weights $\boldsymbol{\mu}_k$ for ω fixed and $k \in \mathbb{N}$. Let $\boldsymbol{\lambda}$ be an accumulation point of the sequence $\boldsymbol{\mu}_k$, then $\boldsymbol{\lambda} \in \mathcal{L}$.*

Corollary 5.2. *If $\mathcal{L}$ is a singleton, then $\mathcal{L} = \{[1, 0, \ldots, 0]^\mathsf{T}\}$ and $\boldsymbol{\mu}_k \to [1, 0, \ldots, 0]^\mathsf{T}$ almost surely as $k \to \infty$.*

In the case where $\mathcal{L}$ is not a singleton, the MAMSE weights do not seem to converge to any particular point. Corollary 5.1 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.

Based on simulations, 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. The fact that the weights depend on the data complicates the situation, but there is more.

When a mixture of the distributions of populations $2, \ldots, m$ is identical to $F_1(x)$, accumulation points of $\boldsymbol{\mu}_k$ will be in $\mathcal{L}$, but they may remain random on $\mathcal{L}$ even for infinitely large sample sizes. This behavior is observed in some simulations of Plante (2008). The lack of convergence to a singleton eliminates many strategies of proof, the use of Slutsky's theorem for instance.

One could hope that proving normality would at least be possible under the hypothesis of Corollary 5.1 when the MAMSE weights converge to $[1, 0, \ldots, 0]^\mathsf{T}$. If the convergence occurs at a rate of $1/\sqrt{n_{1k}}$ or faster, the results of Wang et al. (2004) may apply. Determining the rate of convergence of the MAMSE weights is however not straightforward and may require the characterization of the shapes of each $F_{ik}(x)$ compared to $F_{1k}(x)$ as well as limitations on the speeds of incrementation of the sample sizes.

Until somebody finds a strategy to bypass these complications, we suggest to use resampling methods to produce confidence intervals or tests. Such methods are illustrated in the next section.

6 Bootstrap Simulation

The simulations of Plante (2006) show that the mean squared error (MSE) of the WMLE is often smaller than that of the MLE under different scenarios. We do not repeat such simulations here, but rather illustrate the construction of confidence intervals (CI) for the MWLE using parametric and nonparametric bootstrap. We also calculate bootstrap intervals based on the sample from Population 1 alone and compare their lengths and coverage probabilities.

For nonparametric bootstrap, sampling with replacement is performed on each of the m populations, thus creating a pseudo-sample for each of them. The MWLE is calculated on a large number (set to 1000) of such pseudo-samples and confidence intervals are built by taking quantiles from these simulated MWLEs.

Nonparametric bootstrap involves numerous ties in all samples. While our theoretical results assume the continuity of the distributions, the calculation of the weights themselves and hence of the MWLE hold in the presence of ties.

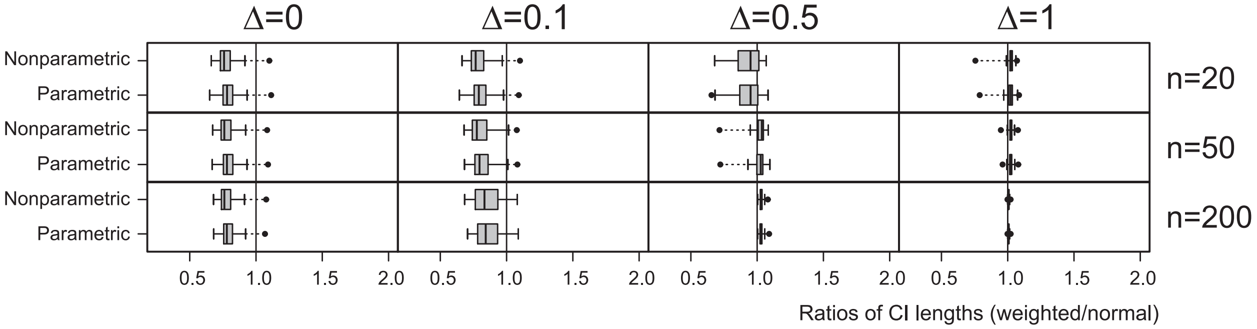

Figure 1: Relative length of the confidence intervals (weighted/normal) for parametric and nonparametric bootstrap. Ratios below one indicate that the weighted methods yielded a shorter interval. Each boxplot is based on 10000 simulated sample.

Parametric bootstrap allows to avoid ties, but requires modeling each population. The MLE of each Population is determined and pseudo-samples are drawn from the fitted distribution rather than through resampling.

6.1 Normal Distribution

We first consider the ubiquitous normal case. Samples of size n and $10n$ respectively are generated from:

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

Confidence intervals for the MLE and the MWLE are built using parametric and nonparametric bootstrap. The process is repeated for 10000 samples and hence yields 10000 intervals.

Figure 1 displays the relative length (length of CI for the MWLE/length of CI for the MLE) of the 10000 intervals obtained using the two different bootstrap methods. Values below 1 indicate a shorter interval for the MAMSE-weighted method. To improve the readability of the boxplots in the presence of extreme values, only the minimum and maximum values are drawn as points; dotted lines links them to the whiskers.

The weighted method often yields a shorter interval than the equivalent bootstrap method based on Population 1 alone, especially for small n and Δ. For larger n and Δ, the performance of the weighted methods are comparable to that of their unweighted counterparts, with occasional small losses.

The estimated coverage probabilities of the different methods appear in Table 1. The coverage of the CI for the MWLE is similar to that obtained for the MLE with the corresponding method. The shorter intervals observed in Figure 1 do not seem to be the consequence of systematically biased intervals.

We also use bootstrap to test the null hypothesis that $\mu = 0$. We base that test on a 2-sided confidence interval built using bootstrap. Figure 2 displays power graphs for that test. The power of the test is evaluated for different values of μ. To compensate for the added computations, each point is based on 1000 replicates rather than 10000. The weighted method corresponds to the dotted line. Note that the curves are smoothed.

	n	Nonparametric Normal	 Weighted	Parametric Normal	 Weighted
	20	92.6	93.0	93.3	93.8
$\Delta = 0$	50	94.0	94.0	94.0	94.4
	200	94.8	95.1	95.0	95.4
	20	92.4	92.4	93.3	93.0
$\Delta = 0.1$	50	93.8	92.8	94.1	93.0
	200	94.7	92.1	94.9	91.5
	20	92.9	91.0	93.6	91.4
$\Delta = 0.5$	50	94.0	92.5	94.5	92.7
	200	94.8	94.3	94.8	94.3
	20	92.5	91.0	93.2	91.9
$\Delta = 1$	50	94.2	93.3	94.4	93.5
	200	94.7	94.5	94.9	94.5

Table 1: Coverage probability of bootstrap confidence intervals for different methods. Each proportion is based on 10000 simulated samples; 1000 pseudo-samples are used on each of them to build the confidence intervals.

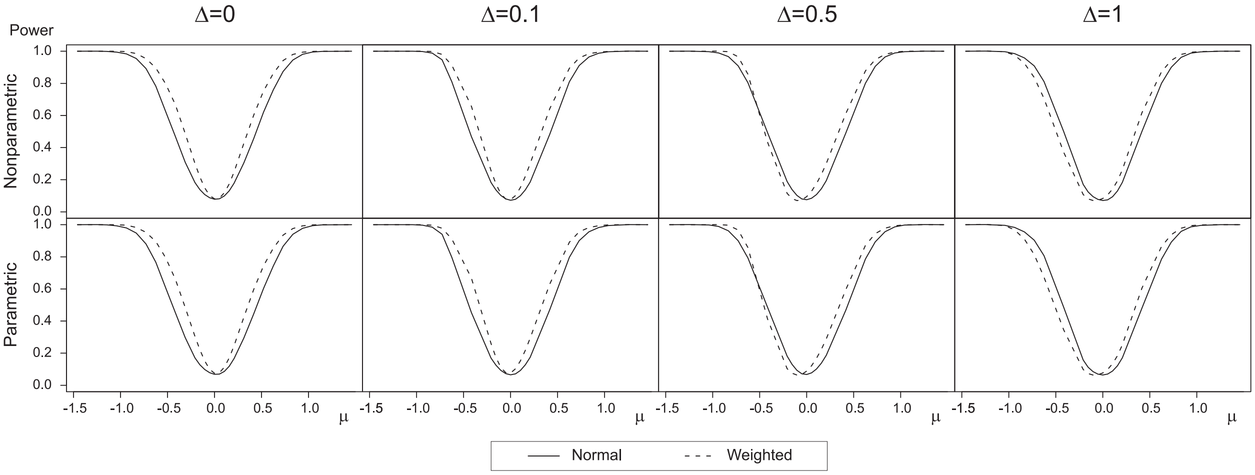

Figure 2: Power of a bootstrap test for $H : \mu = 0$. For different values of μ, 1000 samples were drawn and bootstrapped. Each bootstrap simulation is based on 1000 pseudo-samples. The curves are smoothed.

For small values of Δ, the test based on the weighted method is uniformly more powerful. The MAMSE weights attempt to trade bias for precision. The gain in precision can be seen with the narrower curve allowing a more powerful test for small values of Δ. When Δ gets larger, bias increases and the power curve starts shifting. Eventually, the test on the MWLE is less powerful for a range of values of μ.

6.2 Gamma Distribution

We reuse the scenario of Simulation 5.4 from Plante (2006) where the magnitude of earthquakes from three neighborhooding areas of Western Canada are considered. These three areas are respectively: Lower Mainland – Vancouver Island, rest of British Columbia and Alberta, Yukon and North West Territories. The magnitudes are modeled as independent gamma variables whose parameters are determined from 5 years of observed data. The gamma distribution is parametrized as

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

for $\beta, \mu, x > 0$ and the estimated parameters are

$$\text{Pop. } 1: \begin{cases} \mu = 1.437 \\ \beta = 1.654 \end{cases}, \quad \text{Pop. } 2: \begin{cases} \mu = 1.869 \\ \beta = 2.357 \end{cases}, \quad \text{Pop. } 3: \begin{cases} \mu = 2.782 \\ \beta = 6.806 \end{cases}.$$

Equal samples of size $n = 50$ are drawn from each populations. The goal being to infer about Population 1, the MLE is based on Population 1 alone, but the MWLE uses the samples from the 3 populations.

We first estimate the vector of parameters (μ, β). For the purpose of this simulation, we build non-parametric bivariate 95% confidence sets for $[\log(\mu), \log(\beta)]$ based on Tukey's depth as described by Yeh & Singh (1997). The log transformation allows a better fit since the method yields convex sets. The algorithm developed by Rousseeuw & Ruts (1996) is used to calculate Tukey's depth. The coverage provided by each method is evaluated and the area of the confidence sets (on the log scale) is use to compare their sizes.

Earthquakes are usually not felt unless their magnitude reaches about 3 on the Richter scale. Hence, we also estimate the probability that the magnitude of an earthquake in the Lower Mainland – Vancouver Island area reaches that threshold by plugin the estimated parameters in the Gamma model. We use different bootstrap methods to build confidence intervals for that probability that we denote $P(M > 3)$.

The estimated coverage probabilities appear in Table 2. They are all a bit below the 95% target. Interestingly, the weighted methods all yield coverages closer to 95% than the bootstrap based on Population 1 alone. Figure 4 shows the size of the confidence sets obtained in this simulation.

The confidence regions for the vector of parameters $[\log\beta, \log\mu]$ seem smaller when the weighted bootstrap is used. For the prediction of $P(M > 3)$, the weighted method seems to yield negligibly longer intervals but achieves a more accurate coverage.

More extensive simulations could be performed to describe the advantages and limitations of using bootstrap to determine the variance of the MWLE. The simulations presented above however show that bootstrap

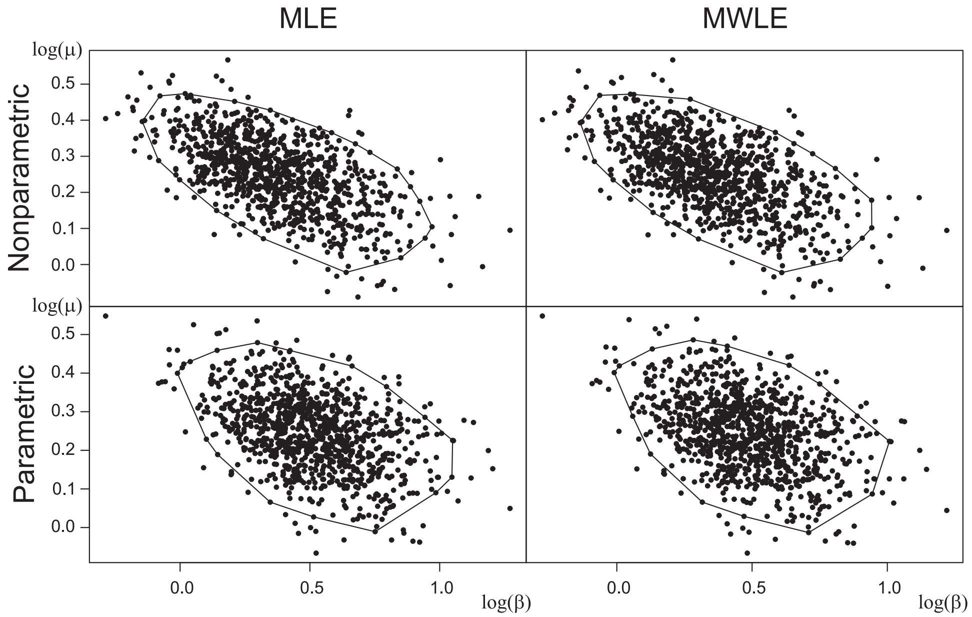

Figure 3: Examples of bivariate 95% confidence sets based on Tukey's depth. The confidence sets are built using log scales to offer a better fit with the convexity of Tukey's depth contours. The points on the graphs above are 1000 bootstrap replicates of $[\log \hat{\beta}, \log \hat{\mu}]$ derived from a single set of samples ($n = 50$ for all three populations).

		Coverage in %	
Bootstrap	Estimate	(μ, β)	$P(M > 3)$
Nonparametric	Normal	86.1	90.8
	Weighted	86.9	93.0
Parametric	Normal	89.3	92.9
	Weighted	89.7	94.5

Table 2: Coverage probability of the different methods of constructing confidence sets for a bivariate gamma parameter and for an estimate of $P(M > 3)$. Each proportion is based on 10000 replicates each calculated from 1000 pseudo-samples.

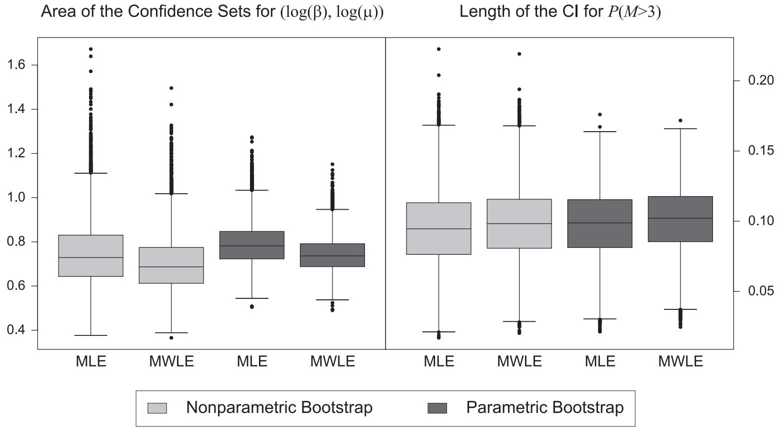

Figure 4: Area or length of the bootstrap confidence sets for estimating $[\log\beta, \log\mu]$ or $P(M > 3)$. Each boxplot is based on 10000 replicates.

confidence intervals or confidence sets seem to behave well with MAMSE-weighted methods. Ideally, the convergence of bootstrap estimates should be proved theoretically, but this is left to future work

7 Conclusion

The weighted likelihood allows to borrow strength from similar population while making parametric assumptions on the population of interest only. The method has been in the literature for a few years already, but the lack of a reliable method to determine likelihood weights seemed to limit its use.

Plante (2008) proposes a heuristic justification of the weighted likelihood that links the weights to a mixture of empirical distribution functions and leads to defining the nonparametric adaptive MAMSE likelihood weights. Simulations therein show that the MWLE with MAMSE weights has a lower mean squared error than the MLE in many cases of interest.

This paper studies the asymptotic properties implied by the MAMSE weights. In particular, we show that the MAMSE-weighted mixture of empirical distributions converges uniformly to the target distribution and that the MWLE is a strongly consistent estimate when used in conjunction with the MAMSE weights. These consistency results hold with very weak assumptions on the distributions underlying the m populations. Hence, the MAMSE weights succeed in determining what population may be useful to the inference or not without relying on a parametric statistical model.

The asymptotic distribution of the MWLE with MAMSE weights is not known at this point. In the simulations presented, the bootstrap behaved well with the weighted methods. Until more research allows to determine an approximate distribution for the MWLE, bootstrap can be used for practical applications based on the MAMSE weights.

Natural extensions of this work include the study of the rate of convergence of the MAMSE-based statistics and the asymptotic distributions thereof. The theory for discrete distributions could also be developed. Extensions of the MAMSE weights to censored and multivariate data have also been studied by Plante (2007) and could be further developed.

A Proofs

Below are the proofs of the results of sections 3 and 4.2.

Proof of Lemma 3.1. For any $\omega \in \Omega$ and $k \in \mathbb{N}$, consider

$$I \triangleq \int \left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right|^2 \mathrm{d}\hat{F}_{1k}(x) \leq \int \left|\hat{G}_k(x) - \hat{F}_{1k}(x)\right|^2 + \sum_{i=1}^{m} \frac{\mu_{ik}^2}{n_{ik}} \hat{F}_{ik}(x)\left\{1 - \hat{F}_{ik}(x)\right\} \mathrm{d}\hat{F}_{1k}(x).$$

By definition, the MAMSE weights minimize the expression above. The suboptimal choice of weights $[\lambda_1, \ldots, \lambda_m] = [1, 0, \ldots, 0]$ cannot lead to a smaller value of I, i.e.

$$I \leq \int \frac{1}{n_{1k}} \hat{F}_{1k}(x)\left\{1 - \hat{F}_{1k}(x)\right\} \mathrm{d}\hat{F}_{1k}(x) = \frac{1}{n_{1k}^2} \sum_{j=1}^{n_{1k}} \frac{j}{n_{1k}}\left(1 - \frac{j}{n_{1k}}\right) = \left(\frac{n_{1k}^2 - 1}{n_{1k}^2}\right) \frac{1}{6n_{1k}}.$$

This bound is tight since the optimal $\boldsymbol{\lambda}$ could be arbitrarily close to the vector $[1, 0, \ldots, 0]^\mathsf{T}$, 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. □

Proof of Lemma 3.2. Define $\Omega_0 = \{\omega \in \Omega : \exists i, i', j, j' \text{ with } (i,j) \neq (i',j') \text{ and } X_{ij} = X_{i'j'}\}$. 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 $j \in \{1, \ldots, n_{ik}\}$, $[\min_{i,j} X_{ij},\ \max_{i,j} X_{ij}]$ 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 maximizing the bounded function $D(x)$. We treat two cases.

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

Define $x_1 = \max\{X_{1j} : j = 1, \ldots, n_{1k},\ X_{1j} \leq x_0\}$, 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}. 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,

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

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

Define $x_2 = \min\{X_{1j} : j = 1, \ldots, n_{1k},\ X_{1j} > x_0\}$, 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}. Therefore, $\hat{F}_{1k}(x_2) = 1/n_{1k} + \hat{F}_{1k}(x_0)$. Since $\hat{G}_k(x)$ is nondecreasing, $\hat{G}_k(x_2) \geq \hat{G}_k(x_0)$.

Consequently,

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

which completes the proof. □

Proof of Lemma 3.3. 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 3.2.

Suppose that Lemma 3.3 is false. Then there exists an infinite sequence k_ℓ such that for $\ell \in \mathbb{N}$,

$$\left| \hat{G}_{k_\ell}\{X_{1(j_{0\ell})}\} - \hat{F}_{1k_\ell}\{X_{1(j_{0\ell})}\} \right| > \epsilon_{k_\ell} = \frac{\epsilon}{a_{k_\ell}} \tag{8}$$

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

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 $\hat{G}_k(x)$ is a nondecreasing function, ans that $\hat{F}_{1k}(x)$ is a right-continuous nondecreasing step function with equal jumps of $1/n_{1k}$. We treat two cases:

Case 1: $\hat{G}_k\{X_{1(j_0)}\} \ge \hat{F}_{1k}\{X_{1(j_0)}\}$.
Note that $\hat{F}_{1k}\{X_{1(j_0)}\} \le \hat{G}_k\{X_{1(j_0)}\} \le 1$ and hence, $\hat{F}_{1k}\{X_{1(j_0)}\} \le 1 - \epsilon_k$ or inequality (8) would not hold. Consequently, $j_0 \le n_{1k} - \lfloor \epsilon_k n_{1k} \rfloor$ and for $i \in \{0, 1, \dots, \lfloor \epsilon_k n_{1k} \rfloor\}$, we have $\hat{G}_k\{X_{1(j_0+i)}\} \ge \hat{G}_k\{X_{1(j_0)}\}$ and $\hat{F}_{1k}\{X_{1(j_0+i)}\} = \hat{F}_{1k}\{X_{1(j_0)}\} + \imath/n_{1k}$ and hence

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

Case 2: $\hat{G}_k\{X_{1(j_0)}\} \le \hat{F}_{1k}\{X_{1(j_0)}\}$.
Note that $\hat{F}_{1k}\{X_{1(j_0)}\} \ge \hat{G}_k\{X_{1(j_0)}\} \ge 0$. Since both functions are at least ϵ_k apart, $\hat{F}_{1k}\{X_{1(j_0)}\} \ge \epsilon_k$ and thus $j_0 \ge \lfloor \epsilon_k n_{1k} \rfloor$. Then for $i \in \{0, 1, \dots, \lfloor \epsilon_k n_{1k} \rfloor\}$, we have $\hat{G}_k\{X_{1(j_0-i)}\} \le \hat{G}_k\{X_{1(j_0)}\}$ and $\hat{F}_{1k}\{X_{1(j_0-i)}\} = \hat{F}_{1k}\{X_{1(j_0)}\} - i/n_{1k}$ and hence

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

Then, for both cases,

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

By Lemma 3.1, we thus have that

$$\frac{1}{3}\left(\frac{\epsilon_k n_{1k} - 1}{n_{1k}}\right)^3 \le \frac{1}{3}\left(\frac{\lfloor \epsilon_k n_{1k} \rfloor}{n_{1k}}\right)^3 \le \frac{1}{6n_{1k}} \Leftrightarrow \left(\frac{\epsilon n_{1k}}{a_k} - 1\right)^3 \le \frac{n_{1k}^2}{2} \Leftrightarrow a_k \frac{n_{1k}^{2/3} + 2^{1/3}}{n_{1k}} \ge 2^{1/3}\epsilon,$$

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, we know that $\forall \epsilon > 0$, $\exists k_0$ such that $\forall k \geq k_0$, $a_k \max_{j \in \{0,\ldots,n_{1k}\}} \left| \hat{G}_k(X_{1j}) - \hat{F}_{1k}(X_{1j}) \right| \leq \epsilon$. Since k_0 does not depend on $\omega \in \Omega_1$ and $P(\Omega_1) = 1$, the uniform convergence is almost sure. □

Proof of Lemma 3.4. Consider the set Ω_1 defined from Lemma 3.2. By Lemma 3.3, $\forall \epsilon > 0$, $\exists k_1$ such that $\forall k \geq k_1$, $\max_{j \in \{1,\ldots,n_{1k}\}} \left| \hat{G}_k(X_{1j}) - \hat{F}_{1k}(X_{1j}) \right| \leq \epsilon/2$ for all $\omega \in \Omega_1$. Moreover, $\exists k_2$ such that $\forall k \geq k_2$, $1/n_{1k} \leq \epsilon/2$. Then by Lemma 3.2, $\forall \omega \in \Omega_1$, we have $0 \leq \max_x \left| \hat{G}_k(x) - \hat{F}_{1k}(x) \right| \leq \epsilon/2 + \epsilon/2 = \epsilon$ for all $k \geq k_0 = \max(k_1, k_2)$. □

Proof of Theorem 3.1. By Lemma 3.4, $\forall \epsilon > 0$, $\exists k_1$ such that $\max_x \left| \hat{G}_k(x) - \hat{F}_{1k}(x) \right| < \epsilon/2$, $\forall k \geq k_1$ 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 $\sup_x \left| \hat{F}_{1k}(x) - F_1(x) \right| < \epsilon/2$, $\forall k \geq k_2(\omega)$. 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

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

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

Proof of Lemma 3.5. 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. □

Proof of Theorem 3.2. We show that for any $\epsilon > 0$, we can find a sequence of inequalities that 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 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 $\tau_t(x) = g(x) \mathbb{1}_{B_t}(x)$ where $\mathbb{1}_B(x)$ is an indicator function equal to 1 if $x \in B$ and otherwise null. 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

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

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 $h_t(x) = \sum_{s=1}^{S_t} b_{st} \mathbb{1}_{A_{st}}(x)$ where $b_{st} = \inf_{y \in A_{st}} g(y)$. Then, by construction, $\sup_x |\tau_t(x) - h_t(x)| \leq 2^{-t}$ and

$$\left| \int g(x)\,\mathrm{d}\hat{G}_k(x) - \int g(x)\,\mathrm{d}F_1(x) \right| \leq T_1 + T_2 + T_3 + T_4 + T_5 \tag{9}$$

where

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

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 3.1, $\sup_x |\hat{G}_k(x) - F_1(x)|$ 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 $\sup_x |\hat{G}_k(x) - F_1(x)| < 1/\{S_t \max(|U_t|, |L_t|) 2^{t+1}\}$ for all $k \geq k_{\omega,t}$. For any such k and ω, Lemma 3.5 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 $\omega \in \Omega_{i,t}$, $\sum_{j=1}^{n_{ik}} |g\{X_{ij}\}| \mathbb{1}_{B_t^c}(X_{ij})$ converges to $\int |g(x)| \mathbb{1}_{B_t^c}(x)\,\mathrm{d}F_i(x)$ as $k \to \infty$. Consider a fixed $\omega \in \Omega_1 = \{\omega | X_{ij} = d_\ell \text{ for some } i, j, \ell\}^C \cap \{\cap_{i \in I, t \in \mathbb{N}} \Omega_{i,t}\}$. The intersection is over a countable number of sets of probability 1, hence $P(\Omega_1) = 1$. For any such $\omega \in \Omega_1$, T_1 is developed as

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

Since ω is fixed, $\exists t_\omega^*$ such that $\mathbb{1}_{B_t^c}(X_{ij}) \equiv 0$, $\forall i \in I^C$, $j = 1, \ldots, n_{iM_i}$, $t \geq t_\omega^*$. For $i \in I$, the dominated convergence theorem says that there exists t_i^* such that $\int |g(x)| \mathbb{1}_{B_t^c}(x)\,\mathrm{d}F_i(x) < \epsilon/(10m)$ 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 $k \geq k_{i,t,\omega}$,

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

Therefore, $\forall t \geq \max(t_3, t_\omega^*)$, there exists $k_{\omega,t}^* = \max_{i \in I} k_{i,t,\omega}$ such that $T_1 \leq \epsilon/5$ 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 (9) 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. □

Proof of Corollary 3.1. Use Theorem 3.2 with $g(x) = x$. □

Proof of Lemma 4.4. 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

$$|f(x_1, \theta_0, \rho) - f(x_2, \theta_0, \rho)| > \epsilon. \tag{10}$$

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

$$|f(x_1|\theta) - f(x_2|\theta)| < \epsilon/2 \tag{11}$$

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,

$$f(x_1, \theta_0, \rho) = f(x_1|\theta_1) \quad \text{and} \quad f(x_2, \theta_0, \rho) = f(x_2|\theta_2). \tag{12}$$

Consider the following two cases.

<u>Case 1:</u> $f(x_1|\theta_1) > f(x_2|\theta_2)$
By Equations (10) and (12), $f(x_1|\theta_1) \geq f(x_2|\theta_2) + \epsilon$. Furthermore, inequality (11) implies that $f(x_2|\theta_1) > f(x_1|\theta_1) - \epsilon/2 \geq f(x_2|\theta_2) + \epsilon/2$, a contradiction with the definition of θ_2.

<u>Case 2:</u> $f(x_1|\theta_1) < f(x_2|\theta_2)$
By Equations (10) and (12), $f(x_1|\theta_1) \leq f(x_2|\theta_2) - \epsilon$. Inequality (11) yields $f(x_1|\theta_2) > f(x_2|\theta_2) - \epsilon/2 \geq f(x_1|\theta_1) + \epsilon/2$, a contradiction with the definition of θ_1.

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

Proof of Lemma 4.5. 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

$$|\phi(x_0, r) - \phi(x_1, r)| > 2\epsilon. \tag{13}$$

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 (13). 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 (13). 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 $\omega_{f(\cdot|\theta)}(\delta, x_0) > \epsilon$ □

Proof of Theorem 5.1. The Glivenko-Cantelli lemma shows that as $k \to \infty$, $\sup_x |\hat{F}_{ik}(x) - F_i(x)| \to 0$ almost surely. Let Ω_i be the set of probability 1 where the convergence occurs.

For any fixed $\boldsymbol{\lambda} \in [0, 1]^m$ with $\sum_{i=1}^m \lambda_i = 1$, the summand and the integrand of the following expressions are bounded by 1, thus $(1/n_{1k}) \sum_{j=1}^{n_{1k}} \{\sum_{i=1}^m \lambda_i F_i(X_{1j}) - F_1(X_{1j})\}^2 \to \mathrm{E}[\{\sum_{i=1}^m \lambda_i F_i(X_{11}) - F_i(X_{11})\}^2]$ 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 (1) and for any $\boldsymbol{\lambda} \in B(\boldsymbol{\lambda}^*, \epsilon)$, $P_k(\boldsymbol{\lambda})$ is greater than or equal to

$$\begin{aligned}
\int \left| \sum_{i=1}^m \lambda_i \hat{F}_{ik}(x) - \hat{F}_{1k}(x) \right|^2 \mathrm{d}\hat{F}_{1k}(x) &\geq \int \left\{ \left| \sum_{i=1}^m \lambda_i \hat{F}_{ik}(x) - F_1(x) \right|^2 - \left| F_1(x) - \hat{F}_{1k}(x) \right|^2 \right\} \mathrm{d}\hat{F}_{1k}(x) \\
&\geq \int \left\{ \left| \sum_{i=1}^m \lambda_i F_i(x) - F_1(x) \right|^2 - \left| \sum_{i=1}^m \lambda_i \hat{F}_{ik}(x) - \sum_{i=1}^m \lambda_i F_i(x) \right|^2 - \left| F_1(x) - \hat{F}_{1k}(x) \right|^2 \right\} \mathrm{d}\hat{F}_{1k}(x) \\
&\geq \frac{1}{n_{1k}} \sum_{j=1}^{n_{1k}} \left\{ \sum_{i=1}^m \lambda_i F_i(X_{1j}) - F_1(X_{1j}) \right\}^2 - \sum_{i=1}^m \lambda_i^2 \sup_x \left| \hat{F}_{ik}(x) - F_i(x) \right|^2 - \sup_x \left| \hat{F}_{1k}(x) - F_1(x) \right|^2
\end{aligned}$$

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

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

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 3.1 shows that $P_k\{\boldsymbol{\mu}_k\} \to 0$ as $k \to \infty$. Therefore, $\boldsymbol{\mu}_k \in B(\boldsymbol{\lambda}^*, \epsilon)$ at most finitely many times. This is true of any $\boldsymbol{\lambda}^* \in \mathcal{L}^C$, meaning that for all $\omega \in \Omega_0$, $||\boldsymbol{\lambda}^* - \boldsymbol{\mu}_k|| > \epsilon$ at most finitely many times. □

Proof of Corollary 5.1. By Theorem 5.1, the neighborhood of any $\boldsymbol{\lambda} \in \mathcal{L}^C$ can be visited at most finitely many times. Hence, any accumulation point belongs to $\mathcal{L}$. □

Proof of Corollary 5.2. The vector $[1, 0, \ldots, 0]^\mathsf{T}$ always belongs to $\mathcal{L}$. 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 $\boldsymbol{\mu}_k$. By Theorem 5.1, 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 $\boldsymbol{\mu}_k$, i.e.

$$\boldsymbol{\mu}_k \in \bigcup_{s=1}^{S} \bar{B}(\mathbf{x}_s, \epsilon/2) \qquad \text{finitely many times.} \tag{14}$$

Consequently,

$$\boldsymbol{\mu}_k \in \left\{ \bigcup_{s=1}^{S} \bar{B}(\mathbf{x}_s, \epsilon/2) \right\}^C \subset B([1, 0, \ldots, 0]^\mathsf{T}, \epsilon) \qquad \textit{i.o.} \tag{15}$$

Expressions (14) and (15) imply that if it exists, the limit of $\boldsymbol{\mu}_k$ 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 $\boldsymbol{\mu}_k \to [1, 0, \ldots, 0]^\mathsf{T}$ almost surely. □

Acknowledgements

This paper constitutes part of the author's doctoral thesis supervised by Professor James V. Zidek at the University of British Columbia. I am thankful to Professor Zidek for his guidance and support in discovering the weighted likelihood and in developing the results presented in this paper. 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.

References

Hu, F. and Zidek, J. (1993). *A relevance weighted nonparametric quantile estimator.* Technical report no.134, Department of Statistics, The University of British Columbia, Vancouver.

Hu, F. (1994). *Relevance weighted smoothing and a new bootstrap method*, unpublished doctoral dissertation, Department of Statistics, The University of British Columbia, 177 pp.

Hu, F. (1997). The asymptotic properties of the maximum-relevance weighted likelihood estimators, *The Canadian Journal of Statistics*, **25**, 45–59.

Hu, F. and Rosenberg, W. F. (2000a). Analysis of time trends in adaptive designs with application to a neurophysiology experiment, *Statistics in Medicine*, **19**, 2067–2075.

Hu, F., Rosenberg, W. F. and Zidek, J. V. (2000b). Relevance weighted likelihood for dependent data, *Metrika*, **51**, 223–243.

Hu, F. and Zidek, J. (2002). The weighted likelihood, *The Canadian Journal of Statistics*, **30**, 347–371.

Plante, J.-F. (2007). Adaptive likelihood weights and mixtures of empirical distributions. Unpublished doctoral dissertation, Department of Statistics, Some University, 171 pp.

Plante, J.-F. (2008). Nonparametric adaptive likelihood weights, *The Canadian Journal of Statistics*, **36**, 443–461.

Rousseeuw, P.J. and Ruts, I. (1996). Algorithm AS 307: Bivariate location depth, *Applied Statistics (JRSS-C)*, **45**, 516–526.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate, *The Annals of Mathematical Statistics*, **20**, 595–601.

Wang, X. (2001). *Maximum weighted likelihood estimation*, unpublished doctoral dissertation, Department of Statistics, The University of British Columbia, 151 pp.

Wang, X., van Eeden, C. and Zidek, J. V. (2004). Asymptotic properties of maximum weighted likelihood estimators, *Journal of Statistical Planning and Inference*, **119**, 37–54.

Wang, X. and Zidek, J. V. (2005). Selecting likelihood weights by cross-validation, *The Annals of Statistics*, **33**, 463–501.

Yeh, A. B. and Singh, K. (1997). Balanced confidence regions based on Tukey's depth and the bootstrap, *Journal of the Royal Statistical Society. Series B (Methodological)*, **59**, 639–652.