Statistica Sinica **23** (2013), 467-487
doi:http://dx.doi.org/10.5705/ss.2011.255

A UNIVERSALLY CONSISTENT MODIFICATION OF MAXIMUM LIKELIHOOD

Byungtae Seo and Bruce G. Lindsay

Sungkyunkwan University and Pennsylvania State University

Abstract: In some models, both parametric and not, maximum likelihood estimation fails to be consistent. We investigate why the maximum likelihood method breaks down with some examples and notice the paradox that, in those same models, maximum likelihood estimation would have been consistent if the data had been measured with error. With this motivation we define doubly-smoothed maximum likelihood as a natural mechanism for adding measurement error without bias. We show the proposed estimation procedure gives universal consistency in independent and identically distributed data. Our method of proof is new. The same arguments can show maximum likelihood itself is universally consistent in a discrete sample space. It is shown that the asymptotic efficiency can be quite high even when the bandwidth parameters are held fixed. Practical guidelines for the choice of kernel and tuning parameter are also given.

Key words and phrases: Consistency, efficiency, measurement error, MLE, spectral decomposition, NPMLE.

1. Introduction

Although the many successes of the maximum likelihood method make it seem like a nearly foolproof way to create good estimators, there are important models where the estimators fail to be consistent, even with independent and identically distributed (IID) data. These models are both parametric and nonparametric. We here consider a simple amendment to maximum likelihood that makes it universally consistent in IID data. By this we mean that the consistency does not depend on any regularity conditions about the model under investigation. The simple amendment to maximum likelihood involves kernel smoothing, but the estimator is consistent for any kernel with any fixed bandwidth. Moreover, it can be made arbitrarily close to maximum likelihood by moving the bandwidth to zero. This estimator will be called *doubly-smoothed maximum likelihood estimator* (DS-MLE). This paper is concerned with the consistency and efficiency of DS-MLE. A companion paper Seo and Lindsay (2010) is available that deals with computation and implementation in a particular model.

Many results about consistency are focused on the consistency of parameter estimators. Results of this type always depend on a series of regularity conditions because the parametrization of a class of distributions is just a way to label the distributions, and so is essentially arbitrary. Our notion of universal consistency requires that we separate the concept of consistency from the concept of parametrization.

To explain this, let us first suppose that $\{Y_1, \ldots, Y_n\}$ is a random sample from some unknown probability measure M_τ on $\mathbb{R}^d$. Suppose further that M_τ is one element of a class of probability distributions, $\mathcal{M}$. If we suppose that $\mathcal{M}$ is indexed by a parameter θ, so $\mathcal{M} = \{M_\theta\}$, then estimation of θ by $\hat{\theta}_n$ provides an estimator of the true parameter θ_τ. Translated into the world of distributions, the true parameter corresponds to some true distribution $M_\tau = M_{\theta_\tau}$ and $\hat{\theta}_n$ to an estimator $\hat{M}_n$ of M_τ, where $\hat{M}_n = M_{\hat{\theta}_n}$. If the method of estimation is parametrization invariant, like maximum likelihood, then $\hat{M}_n$ does not depend on the method of parametrization θ. If we say that a parametrically invariant method of estimation is consistent whenever $\hat{M}_n$ converges to M_τ, in some suitable metric, then consistency is a question free of parametrization. We will call this *distributional consistency*.

This consistency notion is independent of the dimension of the parameter space or a choice of metrics on the parameter space. We here consider models both parametric and nonparametric, and because of that point of view, we call θ the *model index* rather than the parameter, recognizing that there are many possible ways to index the class of models.

Of course, one could well also be interested whether the index estimator $\hat{\theta}$ is consistent under a particular choice of index θ. If we establish distributional consistency, it could imply the consistency of the index estimator, but this ultimately depends on whether the map between the index space and the model distributions is suitably continuous. We consider this question as well, although it goes beyond our main point. Our results for universal consistency apply to maximum likelihood itself in any discrete sample space, and so provide the strongest results possible in that setting.

The results of this paper can be summarized as follows. In Section 2, we illustrate the need for this methodology by considering three examples of inconsistency, including a parametric model and two nonparametric models. These examples are also used to motivate our methodology. In Section 3, we describe our methodology, which is based on kernel smoothing, and then show its universal consistency in Section 4 and 5. We also discuss the theoretical parametric and nonparametric efficiency of the proposed methodology in Section 6 and 7, showing that it can be fully efficient in a number of important settings. Although our methodology does not require that the bandwidth of the kernel should go to

zero for the consistency of estimators, we propose a general guideline to determine a reasonable range of bandwidths in Section 8. We offer further discussion in Section 9.

2. Examples

To motivate our estimation method, we present two examples in which maximum likelihood fails to give consistent estimators even though consistent estimators exist. Our first example is a parametric model.

Example 1 (Normal mixture model)**.** Consider a two-component normal mixture with unknown means μ_1, μ_2, variances σ_1^2, σ_2^2 and class probability ρ. Then the likelihood of a sample from this density is

$$L(\theta; \boldsymbol{x}) = \prod_i \left[\frac{\rho}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right) + \frac{1-\rho}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right) \right], \tag{2.3}$$

where $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. If we do not assume $\sigma_1^2 = \sigma_2^2$, this likelihood is unbounded and its global maximum is ∞ : let $\mu_2 = x_1$ and let σ_2^2 go to zero. Therefore $L(\theta; \boldsymbol{x})$ is not bounded and the parameter values that give the infinite spikes cannot be used to construct a consistent sequence of estimators Kiefer and Wolfowitz (1956).

As a simple amendment, suppose that each X_i was replaced by $X_i^* = X_i + hZ_i$, where Z_i's are i.i.d. $N(0,1)$. In this measurement error model, the distribution for X_i^* can be explicitly calculated as $\rho N(\mu_1, \sigma_1^2 + h^2) + (1-\rho) N(\mu_2, \sigma_2^2 + h^2)$ and the likelihood based on X_i^*'s is

$$\prod_i \left[\frac{\rho}{\sqrt{2\pi(\sigma_1^2 + h^2)}} \exp\left(-\frac{(x_i^* - \mu_1)^2}{2(\sigma_1^2 + h^2)} \right) + \frac{1-\rho}{\sqrt{2\pi(\sigma_2^2 + h^2)}} \exp\left(-\frac{(x_i^* - \mu_2)^2}{2(\sigma_2^2 + h^2)} \right) \right]. \tag{2.4}$$

Then we can see that this likelihood is bounded above, showing that adding measurement error is a means to remove the infinite spikes from a parametric likelihood function. However, if (2.4) is used instead of the original likelihood (2.3) for the same data, it certainly involves a bias due to adding artificial measurement error. The basic idea of this paper relies on adding measurement error while avoiding the bias caused by its addition.

Our next example involves a nonparametric maximum likelihood estimator. A consistent estimator of the nonparametric type is the empirical distribution function $\hat{F}$, which can be derived as the maximum likelihood estimator of a completely unknown distribution. If one were to allow arbitrary continuous densities, then the likelihood would again be unbounded. However, if we allow only

discrete densities $p(x)$, then there is a unique global maximum $\hat{p}$ which satisfies $\hat{p}(x_i) = 1/n$, assuming the data has no ties. In this same sense, the Kaplan-Meier estimator is the nonparametric MLE for censored univariate survival data, and is consistent Kaplan and Meier (1958). In multivariate censored data, however, the method can fail, and so becomes our second example.

Example 2 (Bivariate survivor function)**.** Let $\boldsymbol{T} = (T_1, T_2)$ be the pair of survival times with distribution $F(t_1, t_2)$ and let $\boldsymbol{C} = (C_1, C_2)$ be the pair of censoring times with distribution $G(c_1, c_2)$. Assuming $\boldsymbol{T}$ and $\boldsymbol{C}$ are independent, suppose that we can only observe $(\tilde{T}_1, \tilde{T}_2) = (\min(T_1, C_1), \min(T_2, C_2))$ and $(\delta_1, \delta_2) = (I(T_1 < C_1), I(T_2 < C_2))$, instead of $\boldsymbol{T}$ and $\boldsymbol{C}$. The nonparametric MLE of the joint survival function is not unique and not consistent in general. This non-uniqueness is caused by singly censored observations of which one of the survival times in (T_1, T_2) is exactly observed while the other is censored.

The problem with singly censored observations can be easily explained using the redistribution-to-the-right algorithm for maximum likelihood introduced by Efron (1967). In Figure (a), the point A is doubly right censored and other points are not censored, and the algorithm would equally redistribute the mass of point A to the data points (B, C) found in the upper right quadrant of the point A. However, in Figure (b), the T_1-coordinate of point A is observed but the T_2-coordinate is right censored. In this case, the mass of point A can be redistributed to any point on the dotted line since they produce the same likelihood. Thus the NPMLE is not unique.

If the distribution of (T_1, T_2) is continuous, then with probability one we gain no further observations along the dotted line and so the ambiguity persists. If there is positive probability of single censoring, then there exist a multitude of inconsistent MLEs. The literature contains several methods to fix this inconsistency Dabrowska (1988); Prentice and Cai (1992); van der Laan (1996); Akritas and Keilegom (2003).

A singly censored data point could be described as having one coordinate observed precisely and the other vaguely. Our third example has this same characteristic.

Example 3 (Measurement error problem)**.** Consider a bivariate continuous random variable (X, Z) with an unknown distribution function $G(x, z)$. Suppose Z has no measurement error but random variable X cannot be directly observed, instead one observes W which is X perturbed by some measurement error. Suppose further that the measurement error distribution of $W|X = x$ is completely known, say $f(w|x)$, and that W and Z are independent given X. Then the joint density of (X, W, Z) is $f(w|x, z)g(x, z) = f(w|x)g(x, z)$. Now consider the estimation of the nonparametric distribution G. As in the preceding argument, we

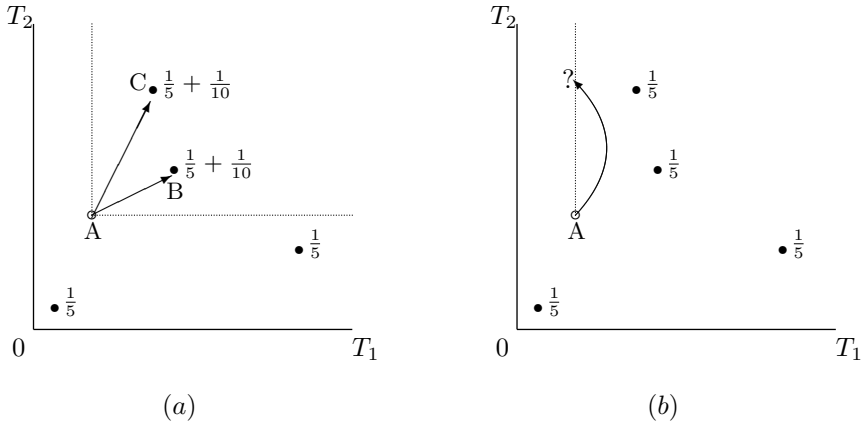

Figure 1. Estimated marginal cumulative distribution of X. Dashed, dotted, and solid line represent the true distribution, the MLE, and the DS-MLE, respectively

restrict attention to G discrete. Since X is not observed, the observed likelihood is

$$L(G) = \prod_{i=1}^{n} \int f(w_i|x)g(x, z_i)dx = \prod_{i=1}^{n} \int f(w_i|x = \xi)I(z = z_i)dG(\xi, z). \quad (2.5)$$

In this case, if the data have no ties and $f(w_i|x = \xi)$ is a completely known unimodal density $f(w - x)$ with mode 0, then it is known that the ML estimate for G is the empirical distribution of (W, Z), which clearly converges to the wrong distribution Roeder, Carroll, and Lindsay (1996); Gaydos (1997). This failure is due to the integrand in (2.5), $f(w_i|x = \xi)I(z = \eta)$, which is continuous in W but discrete in Z. Because of this mixed form of continuous and discrete densities, when the ML procedure estimates the conditional distribution of $X|Z = z_i$, it fails to pool information across Z observations. However, if both X and Z had been measured with error, there would be no inconsistency Roeder, Carroll, and Lindsay (1996).

In these examples we can see that maximum likelihood failed due to inhomogeneity in measurement accuracy. In every case, if we blurred the data by adding artificial measurement error, the inconsistency would disappear. In a similar vein, Luo, Stefanski, and Boos (2006) suggest adding noise for variable selection in a regression setting. Of course, the problem of using maximum likelihood after adding artificial measurement error to data is that the answer one attains would not only cause bias but also lose information. The method we consider removes these problems. Example 1 is extensively discussed in our companion paper Seo and Lindsay (2010).

3. Description of a Doubly Smoothed Likelihood Method

Suppose $X_1, \cdots, X_n$ is a random sample from an unknown probability measure M_τ on $\mathbb{R}^d$. Now, using a kernel $K_h(x,t)$, we can construct a nonparametric kernel density estimator

$$\hat{f}_n^*(t) = \int K_h(x,t) d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i,t), \tag{3.1}$$

where h is a tuning parameter that controls smoothness and $\hat{F}_n(x)$ is the empirical distribution based on $X_1, \ldots, X_n$. By applying the same kernel to the model density, the smoothed model density is defined as

$$m^*(t; M_\theta) = \int K_h(x,t) dM_\theta(x), \tag{3.2}$$

where $M_\theta(x)$ is the distribution function of a model density $m_\theta(x)$ indexed by θ. We can think of m^* as the density of a new variable T that arises from viewing X with the measurement error density $K_h(x,t)$. In this case, the smoothed kernel density can be considered as a nonparametric estimator for the density of the new variable T.

For our methodology, we rely on two basic assumptions for the kernel.

(A1) Kernel regularity: The kernel $K_h(x,t)$ defined on $\mathbb{R}^d \times \mathbb{R}^d$ is bounded above and is continuous in x for each t with $K_h(x,t) \to 0$ for each $t \in \mathbb{R}^d$ as $|x| \to \infty$.

(A2) Kernel identifiability: If $\int K_h(x,t) dM_1(x) = \int K_h(x,t) dM_2(x)$ except for a set of t of Lebesgue measure zero, then $M_1 = M_2$ *a.e.*

In addition, we assume a finite entropy condition on the true smoothed model density.

(A3) $\int m^*(t, M_\tau) |\log (m^*(t, M_\tau))| dt < \infty$

(A1) is a common assumption in the literature. (A2) is needed in our consistency proof, as it assures that the weak convergence of kernel smoothed probability measure implies the convergence of the original probability measure. When any kernel in the exponential family is used, this assumption is easily verified using the uniqueness of the Laplace transform. (A3) assumes the finite entropy of the smoothed model.

Under (A1) and (A2), we can see that the smoothed kernel density $\hat{f}_n^*(t)$ converges, for each t, to the smoothed model density $m^*(t; M_\theta)$ on a set of probability one by the Strong Law Of Large Numbers This convergence is independent

of the value of the tuning parameter h as long as the same kernel and tuning parameter are used for both $\hat{f}_n^*$ and m^*.

The doubly-smoothed maximum likelihood estimator (DS-MLE) of θ is defined as the minimizer of the Kullback-Leibler divergence between m^* and $\hat{f}_n^*$:

$$\hat{\theta}_n = \operatorname*{argmin}_\theta KL(\hat{f}_n^*(t), m^*(t; M_\theta)) = \operatorname*{argmin}_\theta \int \log\Big(\frac{\hat{f}_n^*(t)}{m^*(t; M_\theta)}\Big)\hat{f}_n^*(t)dt. \quad (3.3)$$

Clearly the method is invariant to the choice of the model index θ. The corresponding DS-MLE of the distribution is $M_{\hat{\theta}_n}$. We can also easily verify that minimizing (3.3) is equivalent to maximizing

$$l^*(\theta) = \int \log m^*(t; M_\theta)\hat{f}_n^*(t)dt = \frac{1}{n}\sum_{i=1}^n \int K(x_i, t)\log m^*(t; M_\theta)dt. \quad (3.4)$$

We call (3.4) the doubly-smoothed log-likelihood function because we smooth both data and model, and (3.4) approaches the usual log-likelihood function as the tuning parameter goes to zero. Moreover, in a discrete model with degenerate kernel smoothing, minimizing $KL(\hat{f}_n(t), m^*(t; M_\theta))$ exactly yields the maximum likelihood estimator of θ. Although the results of this paper are generally expressed in terms of nondegenerate kernel smoothing, our methods of proof, and therefore results, apply to any discrete maximum likelihood estimator.

If the model index θ is vector-valued, then solving (3.3) is typically equivalent to solving the estimating equation

$$\int \nabla_\theta \log m^*(t; M_\theta)\hat{f}_n(t)dt = 0. \quad (3.5)$$

The statistical theory of estimating equations then leads to the consistency and asymptotic normality of this minimum distance estimator of θ Basu and Lindsay (1994). However, in the case that the model index θ contains nonparametric components, as in Examples 2 and 3, the consistency of the estimator has not been established. In the next section, we show $M_{\hat{\theta}_n}$ corresponding to the DSMLE $\hat{\theta}_n$ is generally consistent for an essentially arbitrary model.

4. Consistency of $\hat{M}_n$

Our proof is based on almost sure convergence so we need a formal probability framework. We consider a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ with elements ω and a sequence $\{X_n\}$ of random vectors defined on Ω. The basic result we need is that the empirical distribution function $\hat{F}_n^\omega(x) = (1/n)\sum I(X_i(\omega) \le x)$ converges weakly to the true model distribution M_{θ_τ} for ω in a set $\Omega_0 \subset \Omega$ satisfying $P(\Omega_0) = 1$. For IID data this holds due to the Glivenko-Cantelli Theorem. The

reader should note that for a fixed ω, the sequences we consider in this section are not stochastic so we are able to use non-stochastic limiting results.

For simplicity of notation, we write $\hat{M}_n$ instead of $M_{\hat{\theta}_n}(x)$. Similarly, M_τ means $M_{\theta_\tau}(x)$. By *distributional consistency* we mean that $\hat{M}_n^\omega$ converges weakly to M_τ for a set of ω having probability one. This corresponds to showing that $d(\hat{M}_n^\omega, M_\tau) \to 0$ for a set of ω having probability one for any metric $d(\cdot,\cdot)$ on the space of probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$ that metricizes weak convergence Billingsley (1995).

From the the boundedness of the kernel K_h in (A1), it is easy to show that $m^*(t;M)$ is bounded above, and that there exists a positive constant U_h satisfying $m^*(t;M_n) < U_h$ for all n and almost all t.

Lemma 1. *Assuming* (A1) *and* (A3)*, for $\omega \in \Omega_0$, a set having probability one, and for $\hat{f}_n^*(t) = \int K_h(t,x) d\hat{F}_n^\omega(x)$, we have*

$$\lim_{n\to\infty} \int \log\left(\frac{m^*(t;M_\tau)}{U_h}\right) \hat{f}_n^*(t)dt = \int \log\left(\frac{m^*(t;M_\tau)}{U_h}\right) m^*(t;M_\tau)dt. \qquad (4.1)$$

Proof. See the Appendix.

For the next theorem, we assume that the maximizer of the doubly smoothed log-likelihood function exists. To ensure the existence of such a maximum, one might need to consider the closure of the model space $\mathcal{M}$ in the vague topology. This closure, $\overline{\mathcal{M}}$, could include subprobability distributions. However, it would be compact, and so $l^*(\theta)$ in (3.4) viewed as $l^*(M)$ would attain a maximum in $\overline{\mathcal{M}}$. Moreover, it is clear that the maximum would occur at a genuine probability distribution, as otherwise $l^*(M)$ could be increased by scalar multiplication of M. Even without appealing to existence, the following theorem still applies to any sequence of $\hat{M}_n$'s such that $l^*(\hat{M}_n) \geq l^*(M_\tau)$, and there always exists such a sequence as long as $M_\tau \in \mathcal{M}$. In the theorem, the maximizer $\hat{M}_n$ of l^*(or the minimizer of $KL(\hat{f}_n^*, m^*)$) can be interpreted as either the global maximizer or any sequence satisfying $l^*(\hat{M}_n) \geq l^*(M_\tau)$.

Theorem 1. *Let $\mathcal{M} = \{M_\theta\}$ be a class of probability measures on $\mathbb{R}^d$ indexed by θ. Suppose that $X_1, \cdots, X_n$ are IID random vectors from the true distribution $M_\tau \in \mathcal{M}$. Assuming* (A1)–(A3)*, the minimizer $\hat{M}_n$ of $KL(\hat{f}_n^*, m^*)$ converges weakly to M_τ on a set of probability one.*

A proof of Theorem 1 is given in Appendix. The essence of the proof is very simple. Like most consistency proofs for maximum likelihood, the concept of compactness plays an important role. Our proof uses the method of subsequences along with the properties of vague convergence of probability measures on $\mathbb{R}^d$. Given any sequence of probability measures M_n on $\mathbb{R}^d$, one can find a subsequence

M_k that converges vaguely to some subprobability measure M_0. After showing this implies convergence of certain terms, one can use the information inequality to show that M_0 must be the true distribution M_τ. This proof and result also apply to maximum likelihood estimators on a discrete sample space only with Assumption (A3).

Theorem 1 establishes the distributional consistency of a chosen estimated probability measure $\hat{M}_n$ but not the consistency of the model index $\hat{\theta}_n$. We consider the consistency of the estimated model index $\hat{\theta}_n$ for the true θ_τ in the next section.

5. Consistency of Index θ

The consistency of a model index θ can often be easily established by using the distributional consistency result. To establish consistency of $\hat{\theta}_n$, we need first to identify a metric for convergence, say $d(\theta_0, \theta_1)$, which would ordinarily be Euclidean distance when θ is a vector. We then need two model index assumptions.

(M1) Model identifiability : The model index θ is identifiable in the probability measure M_θ.

(M2) Model continuity : $M_{\theta_n}(x) \xrightarrow{v} M_{\theta_0}(x)$ implies that $d(\theta_n, \theta_0) \to 0$.

Corollary 1. *If* (A1)−(A3) *and* (M1)−(M2) *hold, then the minimizer* $\hat{\theta}_n$ *of* $KL(\hat{f}_n^*, m^*)$ *is consistent for true* θ_τ.

For a vector-valued model index θ, (M1) and (M2) directly imply Corollary 1, as in Example 1. The natural metrics $d(\cdot, \cdot)$ to apply to model indices that are themselves distributions, as in Examples 2 and 3, are those that metricize weak convergence. For these one can often apply subsequence arguments such as is used in our distributional consistency proof to prove consistency for the model index $\hat{\theta}$.

For a simple example, suppose one wishes to prove consistency of G estimation in Example 3 when G_τ is the true distribution. Given any subsequence of G_n, say G_m, there exists a further subsequence G_k such that G_k converges vaguely to a subdistribution, say G_1. But this implies that M_{G_k} converges vaguely to M_{G_1}. However, Theorem 4.2 implies M_{G_n} converges weakly to M_{G_τ}, so $G_1 = G_\tau$. This implies that M_{G_n} converges to M_{G_τ}. This, together with identifiability and continuity of model index θ_n proves the convergence of G_n. A similar technique can be applied when model index θ includes both vector valued parameters and distributions.

Some remarks on our approach are needed. Unlike the usual ML estimator that requires several regularity conditions on the parameters and model, if we smooth both the model and the data, the kernel smoothed model is automatically

quite regular. This explains how our methodology can cure an inconsistent ML estimator. Moreover, since we separately address distributional and index consistency, our proof does not require any specific structure for θ. That is, θ can be a set of parameters or nonparametric distributions, or both. So it can be easily applied to other consistency studies for the nonparametric or semi-parametric model. Finally, this proof shows that consistency does not depend on the choice of tuning parameter.

6. Parametric Efficiency of DS-MLE

In this section we consider the relative efficiency of the DS-MLE in arbitrary parametric models that are smoothly parameterized, where relative means relative to the theoretical efficiency of the MLE. We also examine efficiency from the fixed h point of view. We do so because this is the most stringent point of view: in any real problem, one would use an actual h. Doing asymptotics with that value fixed is more conservative than assuming that h is part of a sequence h_n that goes to zero.

6.1. Efficiency loss matrix

We assume that the parametric model m_θ has a smooth vector of scores $\mathbf{u}_\theta$ such that the usual asymptotic theory holds with information matrix $J(\theta) = E(\mathbf{u}\mathbf{u}')$. The information in the DS scores $\mathbf{u}^*$ can then be calculated as

$$E(\mathbf{u}\mathbf{u}^{*\prime})E(\mathbf{u}^*\mathbf{u}^{*\prime})^{-1}E(\mathbf{u}^*\mathbf{u}'),$$

the inverse of the usual asymptotic variance formula. We then take the loss of information is to be

$$E(\mathbf{u}\mathbf{u}') - E(\mathbf{u}\mathbf{u}^{*\prime})E(\mathbf{u}^*\mathbf{u}^{*\prime})^{-1}E(\mathbf{u}^*\mathbf{u}'), \tag{6.7}$$

which can be written as

$$E(\mathbf{u} - R\mathbf{u}^*)(\mathbf{u} - R\mathbf{u}^*)',$$

where $R = E(\mathbf{u}\mathbf{u}^{*\prime})E(\mathbf{u}^*\mathbf{u}^{*\prime})^{-1}$ is the matrix of regression coefficients that minimizes the last displayed equation in the Loewner ordering of matrices. From this representation it is clear that $\mathbf{u}^*$ is fully efficient at θ_0 if and only if the span of $\{u_1^*, \ldots, u_p^*\}$ equals (in the $L_2(\theta_0)$ sense) the span of $\{u_1, \ldots, u_p\}$.

Some initial results on the efficiency of the (fixed h) DS-MLE can be found in Basu and Lindsay (1994). An important point from that paper is that there is not necessarily any loss in efficiency. Moreover, the loss of information that occurs in using the DS-MLE (as compared to the theoretical efficiency of the MLE) is always less than the loss in efficiency one would obtain from simply adding artificial noise directly to the original data.

6.2. Kernel operators and efficiency

A deeper understanding of possible efficiency loss in DS-MLE can be found using the spectral theory of kernel operators. Our theoretical calculations are based on a representation of the DS score function. We start with a vector parameter model m_θ, with vector score function $u_\theta = \nabla_\theta \log m_\theta(x)$. The DS score vector can be written as

$$\begin{aligned} u^*_\theta(x_1) &= \int K_h(x_1, y) \frac{\int u_\theta(x_2) m_\theta(x_2) K_h(x_2, y)\ dx_2}{m^*_\theta(y)} dy \\ &= \int S_{\theta,h}(x_1, x_2)\ u_\theta(x_2) m_\theta(x_2) dx_2, \end{aligned} \tag{6.8}$$

where the symmetric kernel operator $S_{\theta,h}(x_1, x_2)$ is defined by

$$S_{\theta,h}(x_1, x_2) = \int \frac{K_h(x_1, y) K_h(y, x_2)}{m^*_\theta(y)} dy. \tag{6.9}$$

Representation (6.8) shows that an analysis of the kernel $S_{\theta,h}$ can be informative about the transformation that takes us from the score function u to the DS score u^*.

As tools to do so, we draw extensively from Lindsay et al. (2008). That paper considered quadratic distances based on general positive definite kernel functions (such as (6.9)). A general form of their quadratic distances between two probability measures F and G (or discrete probability densities f and g) with kernel operator S_G (or S_g) is given by

$$\begin{aligned} d_{S_G}(F, G) &= \iint S_G(x, y) d(F - G)(x) d(F - G)(y), \\ d_{S_g}(f, g) &= \sum_x \sum_y S_g(x, y)(f(x) - g(x))(f(y) - g(y)). \end{aligned} \tag{6.10}$$

With a specific choice of $S_g(x, y)$, (6.10) can produce some important statistical distances. For example, the L_2 and Pearson distances between two discrete probability densities, f and g, are quadratic distances with $S_g(x, y) = \mathbb{I}(x = y)$ and $S_g(x, y) = \mathbb{I}(x = y)/\sqrt{g(x)g(y)}$, respectively. Lindsay et al. (2008) presented more examples with discussion, and used an eigenanalysis to decompose the limiting distributions of quadratic distances.

Back to the kernel $S_{\theta,h}(x_1, x_2)$ in (6.9), we now have two observations. First, the presence of the smoothing kernel K in this formula means that S itself tends to be a smoother. Second, when used in a quadratic distance, the kernel function (6.9) would generate an analogue of Pearson's chi square distance. For this reason we will call this *Pearson's kernel*. The key fact that we wish to draw upon can be found in Lindsay et al. (2008, Thm. 3.1).

Theorem 2. *A nonnegative definite kernel S satisfying $\iint S(x,y)^2 dM(x)dM(y) < \infty$ can be written as $S(x,y) = \sum_{i=0}^{\infty} \lambda_i \gamma_i(x)\gamma_i(y)$, where the λ_i's and γ_i's are eigenvalues and corresponding normalized eigenvectors of K under baseline measure M. Moreover, $\sum_{i=0}^{n} \lambda_i \gamma_i(x)\gamma_i(y)$ converges strongly to S; that is, for every $g \in L_2$,*

$$\lim_{n\to\infty} \int \left(\int S(x,y)g(y)dM(y) - \sum_{i=1}^{n} \int \lambda_i \gamma_i(x)\gamma_i(y)g(y)dM(y) \right)^2 dM(x) = 0.$$

To illustrate the methodology, suppose that $m_\theta(x)$ is the $N(\theta_1, \theta_2)$ model and that the smoothing kernel $K_h(x,y)$ is $N(x; y, h^2)$. We can then explicitly calculate (6.9) to be

$$\frac{(\theta_2 + h^2)}{h\sqrt{2\theta_2 + h^2}} \exp\left[\frac{2\theta_1^2 h^2 - \theta_2(x_1^2 + x_2^2) + 2(\theta_2 + h^2)x_1 x_2 - 2\theta_1 h^2 (x_1 + x_2)}{2h^2(2\theta_2 + h^2)} \right], \tag{6.11}$$

and derive the spectral representation of Pearson's kernel under the normal model.

Proposition 1. *Pearson's kernel, $S_{\theta,h}(x_1, x_2)$, under the normal model, $N(\theta_1, \theta_2)$, has the spectral representation*

$$S_{\theta,h}(x_1, x_2) = \sum_{n=0}^{\infty} \lambda_{\theta n} \gamma_n(x_1; \theta_1, \theta_2)\gamma_n(x_1; \theta_1, \theta_2), \tag{6.12}$$

where $\lambda_{\theta n} = \left\{\theta_2/(\theta_2 + h^2)\right\}^n$, $\gamma_n(x; \theta_1, \theta_2) = (1/\sqrt{2^n n!}) H_n\left((x - \theta_1)/\sqrt{2\theta_2}\right)$, and $H_n(x)$ is the Hermite polynomial.

Proof. This can be proved using Mehler's formula; see the Appendix.

6.3. Spectral conditions for full efficiency with one scalar parameter

We now consider the information lost in a model. Assume that the score function u can be represented in the eigenfunction basis as $u = \sum a_j \gamma_j(x)$. We can then explicitly calculate $u^* = \sum a_j \lambda_j \gamma_j(x)$, and find that the relative efficiency of u^* is

$$\frac{\left(\sum a_j^2 \lambda_j\right)^2}{\left(\sum a_j^2\right)\left(\sum a_j^2 \lambda_j^2\right)}.$$

If we create a discrete density for the index j by setting $\pi_j = a_j^2 / \sum a_j^2$, then relative efficiency can be written as

$$\frac{E^2(\lambda_J)}{E(\lambda_J^2)}. \tag{6.13}$$

From this representation it is clear that the relative efficiency is one if and only if the random variable λ_J is degenerate at some nonzero value; that is, if a_j is nonzero on a set of eigenfunctions that have the same nonzero value of λ_j.

A similar rule for full efficiency of DS-MLE applies when the dimension of the score space is p, bigger than one. Now the representation of the scores in the Pearson kernel basis must give nonzero weights only to a set of eigenfunctions that correspond to p or fewer distinct eigenvalues. If so, the DS scores can be represented with positive weights on the same eigenfunctions, and the linear transformation R makes them equivalent to the likelihood scores.

As an example of this, consider the illustrative normal model. The ML and DS score for $N(\theta_1, \theta_2)$ can be represented in two terms of eigenfunctions of Pearson's kernel:

$$\mathbf{u}(\theta_1, \theta_2) = \left(\frac{x - \theta_1}{\theta_2}, \frac{(x-\theta_1)^2 - 1}{2\theta_2} \right) = \left(\frac{\gamma_1(x; \theta_1, \theta_2)}{\sqrt{\theta_2}}, \frac{\gamma_2(x; \theta_1, \theta_2)}{\theta_2 \sqrt{2}} \right),$$
$$\mathbf{u}^*(\theta_1, \theta_2) = \left(\frac{\omega}{\sqrt{\theta_2}} \gamma_1(x; \theta_1, \theta_2), \frac{\omega^2}{\theta_2 \sqrt{2}} \gamma_2(x; \theta_1, \theta_2) \right),$$

where $\omega = \theta_2/(\theta_2 + h^2)$. Then we can see that this two-dimensional score space is equivalent to that of the DS score, so full efficiency holds.

6.4. Worst case scenarios

We now turn from the best case scenario to the worst case. Consider again the scalar case, with efficiency calculated in (6.13), and the setting in which a_j is nonzero on a subset of the values $\lambda_1, \ldots, \lambda_m$. Elementary optimization considerations show that (6.13) attains its minimum, over all densities $\pi_1, \ldots, \pi_m$, when $\pi_1 + \pi_m = 1$. One can then use calculus to show that the minimizing value of π_1 is $\pi_1 = \lambda_m/(\lambda_1 + \lambda_m)$. This leads to a minimized relative information of $4r/(1+r)^2$, where $r = \lambda_m/\lambda_1$.

We conclude that we can always create score functions whose DS scores give arbitrarily small efficiency simply by choosing r sufficiently small. However, the original score functions must be rather unsmooth, with a small weight on a large eigenvalue and a large weight on a small eigenvalue. Curiously, though, full efficiency does occur if all the weight π_1 is on the small eigenvalue λ_m.

When the DS-MLE is used in models with parameter dimension $p > 1$, the richness of the model may increase the presence of scores with such "bad" spectral representations. This is counterbalanced by the fact that information loss is measured by how well a model score u_j is approximated by the best regression on the full set of u^* scores. That is, one can always come closer to approximating u_j by $R\mathbf{u}^*$ than by u_j^* alone. The answer is therefore fairly complicated; in the next section, we discuss this with nonparametric models.

7. Nonparametric Full Efficiency

In a non- or semi-parametric model, determining the efficiency of DS-MLE can become simpler in cases where the model is essentially nonparametic. To explain this, if the closed linear subspace generated by the set of ML scores of the model at the truth τ is complete in $L_2(\tau)$, we say that the model is *essentially nonparametric*. In this case, every $L_2(\tau)$ function $h(x)$ (or $\sum n^{-1}h(x_i)$ in a sample) is a fully efficient estimator of $E_\tau[h(X)] = \mu(\tau)$ when the expectation is viewed as a function of the model index. The logic is straightforward: for any reparametrization (μ, γ) of the index, we have that $E_{(\mu,\gamma)}[h(x) - \mu] = 0$ for all γ, and so $h(x) - \mu$ is the efficient score for μ Bickel et al. (1993); see also Lindsay (1983) and Tierney and Lambert (1984).

In an essentially nonparametric model, if one can show that the corresponding DS score space is also complete in $L_2(\tau)$, it is clear that we can approximate any ML score to an arbitrary high level of precision, and so the information loss at equation (6.7) can be made arbitrarily small for any finite dimensional score u. That is, the DS-MLE is theoretically fully efficient.

We start by illustrating these arguments in the normal mixture model, then extend the arguments to Example 3. Consider the normal mixture model, $m(x; Q, \sigma^2) = \int N(x; \mu, \sigma^2) dQ(\mu)$ where Q is an unknown mixing distribution and $N(x; \mu, \sigma^2)$ is the normal density. The usual directional ML score function at μ is

$$u_\mu(x) = \frac{N\left(x; \mu, \sigma^2\right)}{m(x; Q_\tau, \sigma^2)} - 1, \tag{7.1}$$

and the corresponding DS score function is

$$u^*_\mu(x) = \int K_h(x, t) \frac{N^*(t; \mu, \sigma^2)}{m^*(t; Q_\tau, \sigma^2)} dt - 1, \tag{7.2}$$

where

$$N^*(t; \mu, \sigma^2) = \int K_h(y, t) N(y; \mu, \sigma^2) dy,$$

$$m^*(t; Q_\tau, \sigma^2) = \int K_h(y, t) m(y; Q_\tau, \sigma^2) dy.$$

Technically, these are one-sided scores, as we have taken directional derivatives. The one-sided nature of these derivatives means that the scores generate a closed convex cone. If it is not a closed linear space, there will be non-standard asymptotics. Examples of this include binomial mixtures when the number of components in the mixture Q_τ is too small. However, if the mixing distribution Q_τ is sufficiently rich, then all the directional scores become two-sided and we have a closed linear space. In this case we call τ an *interior point* of the mixture space.

In the normal mixture model, the class of interior points is quite rich. For example, suppose that Q_τ is absolutely continuous with density $q_\tau(\mu)$. It is then sufficient that q_τ be continuous and positive valued everywhere. Then for any fixed ϵ sufficiently small that $q_\tau(\mu)$ is bounded below by B for $\mu \in [\mu_0 - \epsilon, \mu_0 + \epsilon]$, we have that the function

$$\frac{U(\mu; \mu_0 - \epsilon, \mu_0 + \epsilon)}{q_\tau(\mu)},$$

where $U(x; a, b)$ is the uniform density on (a, b), is bounded above for all μ. It follows that, for α sufficiently small and positive,

$$(1 + \alpha)\, q_\tau(\mu) - \alpha U(\mu; \mu_0 - \epsilon, \mu_0 + \epsilon)$$

is a density function. The directional score for this density, at τ, is

$$1 - \frac{\int U(\mu; \mu_0 - \epsilon, \mu_0 + \epsilon) N(x; \mu, \sigma_0^2) d\mu}{m(x; Q_\tau, \sigma^2)}.$$

Clearly this approaches $-u_\mu(x)$ as ϵ approaches 0, and so both $u_\mu(x)$ and $-u_\mu(x)$ are included in the closure of the space of score functions.

Next, we show that the closed linear space $\mathbb{M}$ generated by the scores is the zero mean subspace of $L_2(\tau)$, when τ is such an interior point, by showing that the orthogonal complement of $\mathbb{M}$ is the zero function. Consider any mean zero function $\rho(x)$ satisfying $E_\tau[u_\mu(x)\rho(x)] = 0$ for almost all μ. This can be written as $\int \rho(x) N(x; \mu, \sigma^2) dx = 0$. By the statistical completeness of the normal density in the parameter μ Lehmann and Casella (1998), we have that $\rho(x)$ is zero almost everywhere.

Finally, we consider conditions for the DS scores to be complete in this problem. Let $\rho(x)$ be a mean zero function in L_2 that is orthogonal to $u^*_\mu(x)$ for all μ:

$$\begin{aligned} 0 &= \int \rho(x) u^*_\mu(x) m(x; Q_\tau, \sigma^2) dx \\ &= \int \rho(x) \left[\int K_h(y, x) \frac{N^*(y; \mu, \sigma^2)}{m^*(y; Q_\tau, \sigma^2)} dy - 1 \right] m(x; Q_\tau, \sigma^2) dx \\ &= \int \int \rho(x) K_h(y, x) \frac{N^*(y; \mu, \sigma^2) m(x; Q_\tau, \sigma^2)}{m^*(y; Q_\tau, \sigma^2)} dx dy. \end{aligned} \tag{7.3}$$

If $N^*(y; x, \sigma^2)$ is a complete parametric family of densities in parameter x, the inner argument

$$\rho^*(y) = \int K_h(y, x) \frac{\rho(x) m(x; Q_\tau, \sigma^2)}{m^*(y; Q_\tau, \sigma^2)} dx \tag{7.4}$$

is zero almost surely. Now suppose that the kernel $K_h(y, x)$ is also a complete parametric family of densities in parameter y. It then follows that $\rho(x)$ is zero almost surely.

These arguments imply that the nonparametric efficiency of the DS-MLE depends on the kernel and whether the given model is essentially nonparametric. If the given model is essentially nonparametric, then we need only check if the kernel preserves the completeness of the score space.

Example 4. From Example 3, consider the nonparametric estimation of the distribution of an unobserved variable X with a normal additive measurement error. Suppose that the measurement error distribution $f(w|x)$ is $N(x, \sigma^2)$ with known σ^2. Then the model density for the observable variable W is $m_G(w) = \int N(w; x, \sigma^2) dG(x)$, where G is an unknown distribution function of X. The ML score function for G is the gradient function of $m_G(w)$ at x:

$$u_x(w) = D_G(x) = \frac{N\left(w; x, \sigma^2\right)}{m_G(w)} - 1.$$

For the normal kernel with variance h^2, the smoothed model density for W is $m^*_G(w) = \int N(w; x, \sigma^2 + h^2) dG(x)$ and the DS score function is

$$u^*_x(w) = \int K_h(y, w) \frac{N\left(w; x, \sigma^2 + h^2\right)}{m^*_G(w)} dy - 1.$$

As discussed, the ML scores and DS scores are both complete in $L_2(G_\tau)$ provided G_τ is contained in a sufficiently rich family of models.

8. Choice of Kernel and Tuning Parameter

Although, theoretically, any kernel satisfying (K1) and (K2) can be used in our estimation for the consistency purpose, it is desirable to choose a kernel which gives an algebraic formula for m^*, the smoothed model density. For instance, in Example 1, the smoothed model density is the normal density for the normal kernel due to the convolution property of normal densities. In Example 3, the model density is not specified, so any kernel can be used. For example, using a normal kernel can give us a convenient way to estimate the nonparametric distribution using algorithms for nonparametric normal mixture models.

The most popular approach to choosing a tuning parameter is a model-specific method that selects the tuning parameter which makes the mean squared error for some function of the parameters small. Given our very general formulation of the problem, it is difficult to be specific on the implementation of this. Given our focus on distributional consistency, one approach might be to attempt to minimize a measure of risk based on $d(\hat{M}_n, M_\tau)$.

Note that our method reduces to maximum likelihood if h goes to zero, so if the MLE is inconsistent, small values of h must generate estimators with considerable bias. On the other hand, using h too large could reduce the information in a problem and so result in large variance. The optimal choice of h must have some tradeoff between these two extremes.

Examination of our examples has led us to suspect that reasonable strategies for choosing h depend on the structure of the model. Fortunately, in our method the consistency of estimation does not depend on the choice of the tuning parameter, and therefore it is not surprising that the DS-MLE is quite robust to the choice of h. For this reason, in this section we propose a simple general method that gives a reasonable range for the tuning parameter.

As a general guideline, we propose to use the spectral degrees of freedom of the kernel Lindsay et al. (2008); Ray and Lindsay (2008). Notice that there is no bias in using $\hat{f}_n^*$ to estimate m^*, so the quality of the estimation process is determined by the variability of the kernel density estimator. If one were to use a histogram density estimator in our problem, one would naturally measure the degree of smoothing (and hence the variance reduction) by the number of bins used in the histogram. The spectral degrees of freedom measure provides a natural analogue of the "number of bins" for use with other kernel density estimators.

The limiting distribution of the L_2 distance between the smoothed model density $m^*(t; M_\theta)$ and the smoothed kernel density $\hat{f}_n^*$, namely $\int (\hat{f}_n^*(t) - m_\theta^*(t))^2 dt$, is $\sum_{i=1}^{\infty} \lambda_i Z_i^2$, where Z_i's are independent $N(0,1)$ and $\lambda_i \geq 0, \sum_{i=1}^{\infty} \lambda_i < \infty$. Here the λ_i's are the eigenvalues from a functional spectral decomposition. Now, the limiting distribution of the distance is an infinite sum of the independent scaled χ_1^2. This can be approximated with a scaled chi-square distribution with an appropriate degrees of freedom, sDOF: $\sum_{i=1}^{\infty} \lambda_i Z_i^2 \approx c\chi^2_{\text{sDOF}}$. By matching the first two moments Satterthwaite (1941), the *Pearson scale factor*, c, and the *spectral degrees of freedom*, called sDOF, can be obtained. This approximation generally becomes more accurate as the degrees of freedom increase. Thus, sDOF has a similar meaning to usual degrees of freedom in χ^2 goodness-of-fit test. One nice feature of sDOF is that it is easy to estimate empirically. For details, see Lindsay et al. (2008). While there might be some advantages to constructing a degrees of freedom measure based on Kullback-Leibler distance rather than L_2, it seems to us that the simple and easy to implement sDOF is adequate for our purposes.

Of course, there are no magic rules for choosing the number of bins to use, nor the spectral degrees of freedom, but we can eliminate certain unreasonable choices. For a given tuning parameter h, if the estimated sDOF is less than 5, say, we are probably oversmoothing. If it is greater than $n/5$, corresponding to

five observations per bin, we might suppose the amount of smoothing is becoming too small.

When it is difficult to find a more appropriate method for choosing the tuning parameter, sDOF always gives simple and dimensionless information for the tuning parameter just as the usual degrees of freedom does. Even if we have a more specific method of choosing h, it might require grid search over different tuning parameters. In such a case, sDOF has proved to be a very useful tool to narrow the range of values to consider. The effectiveness of sDOF and how we can incorporate it with a specific bandwidth selection rule are discussed in the next section.

9. Discussion

In some ways, our consistency results are quite strong. We show almost sure convergence to the true distribution for virtually any statistical model, completely without regularity conditions. However, there is a price paid in the strength of the convergence results. For example, the empirical CDF $\hat{F}_n$ converges to the true distribution in other strong metrics such as Kolmogorov-Smirnov measure. Our "weak convergence" results cannot be strengthened because of the kernel smoothing, which blurs the exact data locations.

The consistency result in Theorem 1 relies on the weak convergence of $\hat{F}_n^*$ to F_τ^*, where $\hat{F}_n^*$ and F_τ^* are distribution functions of $\hat{f}_n^*$ and f_τ^*, respectively. This consistency proof can be strengthened or simplified if the convergence also holds in Mallows distance. In this case, one can prove the convergence of $m^*(t, \hat{M}_n)$ to $m^*(t, \hat{M}_\tau)$ in total variation distance by applying similar arguments to those in Dümbgen, Samworth, and Schuhmacher (2011). The weak convergence of $\hat{M}_n$ can then be readily obtained using the Strong Law of Large Numbers.

Acknowledgement

The research of Byungtae Seo was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0014607). The research of Bruce Lindsay was partially supported by the National Science Foundation (NSF-DMS 0714839).

Appendix: Proofs

A.1. Proof of Lemma 1

To prove (4.1), first we apply Fubini's theorem for nonnegative functions:

$$\int \log\Big(\frac{m^*(t;M_\tau)}{U_h}\Big)\hat{f}_n^*(t)dt = -\int \log\Big(\frac{U_h}{m^*(t;M_\tau)}\Big)\int K_h(t,x)d\hat{F}_n^\omega(x)dt$$

$$= -\iint \log\Big(\frac{U_h}{m^*(t;M_\tau)}\Big) K_h(t,x)dtd\hat{F}_n^\omega(x)$$
$$= -\frac{1}{n}\sum_i \int \log\Big(\frac{U_h}{m^*(t;M_\tau)}\Big) K_h(t,x_i)dt. \quad \text{(A.1)}$$

Using the Strong Law of Large Numbers under (A3) and Fubini's theorem again, $(A.1)$ converges to

$$-\iint \log\Big(\frac{U_h}{m^*(t;M_\tau)}\Big) K_h(t,x)dtdM_\tau(x)$$
$$= \int \log\Big(\frac{m^*(t;M_\tau)}{U_h}\Big) \int K_h(t,x)dM_\tau(x)dt$$
$$= \int \log\Big(\frac{m^*(t;M_\tau)}{U_h}\Big) m^*(t;M_\tau)dt.$$

on a set Ω_0 of probability one.

A.2. Proof of Theorem 1

Fix $\omega \in \Omega_0$. Since $\hat{M}_n = \hat{M}_n^\omega$ is a sequence of distributions on $\mathbb{R}^d$, for any subsequence $\{m\} \subset \{n\}$ by Helly's selection principle we can select a further subsequence $\{k\} \subset \{m\}$ such that $\hat{M}_k$ is vaguely convergent to a subprobability measure M_0 Kallenberg (1997). If we can show that $M_0 = M_\tau$, then we are done by the method of subsequences (Chung, 1974, Thm. 4.3.4). One can easily justify the following sequence of inequalities.

$$0 \geq \liminf_k \int \log\Big(\frac{m^*(t;M_\tau)}{m^*(t;\hat{M}_k)}\Big) \hat{f}_k^*(t)dt$$
$$\geq \liminf_k \int \log\Big(\frac{m^*(t;M_\tau)}{U_h}\Big) \hat{f}_k^*(t)dt + \liminf_k \int -\log\Big(\frac{m^*(t;\hat{M}_k)}{U_h}\Big) \hat{f}_k^*(t)dt$$
$$\geq \lim_k \int \log\Big(\frac{m^*(t;M_\tau)}{U_h}\Big) \hat{f}_k^*(t)dt + \int \liminf_k \log\Big(\frac{U_h}{m^*(t;\hat{M}_k)}\Big) \hat{f}_k^*(t)dt$$
$$= \int \log\Big(\frac{m^*(t;M_\tau)}{U_h}\Big) m^*(t;M_\tau)dt + \int \log\Big(\frac{U_h}{m^*(t;M_0)}\Big) m^*(t;M_\tau)dt$$
$$= \int \log\Big[\frac{m^*(t;M_\tau)}{m^*(t;M_0)}\Big] m^*(t;M_\tau)dt \geq 0. \quad \text{(A.2)}$$

We have the first inequality because $\hat{M}_k$ is a minimizer of $KL(\hat{f}_k^*, m^*)$. The second inequality holds because $\liminf_k\{a_k + b_k\} \geq \liminf_k\{a_k\} + \liminf_k\{b_k\}$, and the third inequality holds by Fatou's Lemma. Note that $\log(U_h/m^*(t;\hat{M}_k))$ is nonnegative for almost all t.

The last inequality comes from the information inequality and the fact that M_0 is a subprobability measure. Therefore, equality holds in the information inequality, which means $m^*(t; M_\tau) = m^*(t; M_0)$ on a set of t-values with probability one under $m^*(t; M_0)$. From the kernel identifiability condition (K2), $m^*(t; M_\tau) = m^*(t; M_0)$ implies $M_\tau = M_0$. Therefore, every vaguely convergent subsequence of $\hat{M}_m$ vaguely converges to M_τ. This implies $\hat{M}_m$ weakly converges to M_τ (Chung, 1974, Thm. 4.3.4). This also implies that $\hat{M}_n$ weakly converges to M_τ.

A.3. Proof of Proposition 1

Using calculus, Pearson's kernel (6.11) can be expressed as

$$\frac{h^2}{\sqrt{1-\omega^2}} \exp\Big[\frac{1}{1-\omega^2}\Big\{2\omega\Big(\frac{x_1-\theta_1}{\sqrt{2\theta_2}}\Big)\Big(\frac{x_2-\theta_1}{\sqrt{2\theta_2}}\Big) - \omega^2\Big(\Big(\frac{x_1-\theta_1}{\sqrt{2\theta_2}}\Big)^2 + \Big(\frac{x_2-\theta_1}{\sqrt{2\theta_2}}\Big)^2\Big)\Big\}\Big],$$

where $\omega = \theta_2/(\theta_2 + h^2)$. In Mehler's formula,

$$\sum_{n=0}^{\infty} \frac{\omega^n}{2^n n!} H_n\Big(x\Big) H_n\Big(y\Big) = \frac{1}{\sqrt{1-\omega^2}} \exp\Big(\frac{2\omega xy - \omega^2(x^2+y^2)}{1-\omega^2}\Big),$$

if we replace x and y by $(x_1-\theta_1)/\sqrt{2\theta_2}$ and $(x_2-\theta_1)/\sqrt{2\theta_2}$, then (6.11) is

$$\begin{aligned} K_\theta(x_1, x_2) &= \sum_{n=0}^{\infty} \frac{\omega^n}{2^n n!} H_n\Big(\frac{x_1-\theta_1}{\sqrt{2\theta_2}}\Big) H_n\Big(\frac{x_2-\theta_1}{\sqrt{2\theta_2}}\Big) \\ &= \sum_{n=0}^{\infty} \omega^n \gamma_n(x_1; \theta_1, \theta_2)\gamma_n(x_1; \theta_1, \theta_2), \end{aligned}$$

where $\gamma_n(x; \theta_1, \theta_2) = (1/\sqrt{2^n n!}) H_n\big((x-\theta_1)/\sqrt{2\theta_2}\big)$. To prove $\gamma_n(x; \theta_1, \theta_2)$ terms are orthogonal under $N(\theta_1, \theta_2)$, we use the identity

$$\int H_n(x) H_m(x) \exp(-x^2) dx = I[m=n] 2^n n! \sqrt{\pi}. \tag{A.3}$$

Using the change variables $(x-\theta_1)/\sqrt{2\theta_2}$, (A.3) implies

$$\int \gamma_n(x; \theta_1, \theta_2)\gamma_m(x; \theta_1, \theta_2) \frac{1}{\sqrt{2\pi\theta_2}} \exp\Big(-\frac{(x-\theta_1)^2}{2\theta_2}\Big) dx = I[m=n].$$

References

Akritas, M. and Keilegom, I. (2003). Estimation of bivariate and marginal distribution with censored data. *J. Roy. Statist. Soc. Ser. B.* **65**, 457-471.

Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Ann. Inst. Statist. Math.* **46**, 683-705.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Billingsley, P. (1995). *Probability and Measure*. 3rd edition. Wiley, New York.

Chung, K. L. (1974). *A Course in Probability Theory*. 2nd edition. Academic Press, New York.

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *Ann. Statist.* **16**, 1475-1489.

Dümbgen, L., Samworth, R. and Schuhmacher, D. (2011). Approximation by log-concave distributions, with applications to regression. *Ann. Statist.* **39**, 702-730.

Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the* 5*th Berkeley Symposium* (Vol 4), Berkeley: University of California Press, 831-853.

Gaydos, B. L. (1997). The semiparametric likelihood method and its extensions with application to errors-in-variables. PhD thesis, Penn State.

Kallenberg, O. (1997). *Foundations of Modern Probability*. Springer, New York.

Kaplan, E. L. and Meier, P. (1958). Non-parametric estimation from incomplete obervations. *J. Amer. Statist. Assoc.* **53**, 457-481.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 886-906.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. 2nd edition. Springer, New York.

Lindsay, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11**, 486-497.

Lindsay, B. G., Markatou, M., Ray, S., Yang, K. and Chen, S. (2008). Quadratic distances on probabilities: A unified foundation. *Ann. Statist.* **36**, 983-1006.

Luo, X., Stefanski, L. A. and Boos, D. D. (2006). Tuning variable selection procedures by adding noise. *Technometrics* **48**, 165-175.

Prentice, R. L. and Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika* **79**, 495-512.

Ray, S. and Lindsay, B. G. (2008). Model selection in high-dimensions: A quadratic-risk based approach. *J. Roy. Statist. Soc. Ser. B.* **70**, 95-118.

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariates. *J. Amer. Statist. Assoc.* **91**, 722-732.

Satterthwaite, F. W. (1941). Synthesis of variance. *Psychometrika* **6**, 309-316.

Seo, B. and Lindsay, B. G. (2010). A computational strategy for doubly smoothed MLE exemplified by the normal mixture model. *Comput. Statist. Data Anal.* **54**, 1930-1941.

Tierney, L. and Lambert, D. (1984). Asymptotic efficiency of estimators of functionals of mixed distributions. *Ann. Statist.* **12**, 1380-1387.

van der Laan, M. J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *Ann. Statist.* **24**, 596-627.

Department of Statistics, Sungkyunkwan University, Seoul, Korea.

E-mail: seobt@skku.edu

Department of Statistics, Pennsylvania State University, University Park, Pennsylvania, USA.

E-mail: bgl@psu.edu

(Received October 2011; accepted March 2012)