The Pennsylvania State University

The Graduate School

Department of Statistics

DOUBLY-SMOOTHED MAXIMUM LIKELIHOOD ESTIMATION

A Thesis in

Statistics

by

Byungtae Seo

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2007

The thesis of Byungtae Seo was read and approved* by the following:

Bruce G. Lindsay
Willaman Professor of Statistics
Head of the Department of Statistics
Thesis Adviser, Chair of Committee

Bing Li
Professor of Statistics

David R. Hunter
Associate Professor of Statistics

Hung-mo Lin
Associate Professor of Health Evaluation Sciences

*Signatures on file in the Graduate School.

Abstract

In some cases, the maximum likelihood method fails to yield a consistent estimator. We describe why the ML method breaks down with some examples and explore how usual MLE can be modified to get consistency. The doubly-smoothed maximum likelihood estimation (DSMLE) is proposed based on kernel smoothing and minimum distance estimation. We show how it works and prove its universal consistency. Some computational aspects are discussed with fundamental guidelines for the choice of a kernel and a tuning parameter. Under this theoretical basis, the proposed method is applied to some important statistical models such as normal mixture models, measurement error models.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my sincere appreciation to my advisor, Dr. Bruce Lindsay. I am very fortunate to have been able to study under his training. His guidance, patience, and encouragement allow me not only to complete this dissertation but also to learn what a statistician is. I also thank Dr. Bing Li, Dr. David Hunter and Dr. Hung-mo Lin for all the help they have given to me.

I would like to give heart-felt thanks to my parents. Their endless support, sacrifice, and love have been my reason for being. I especially thank my wife, Hyekyung Jung, for her visible and invisible support and having my daughter who will be born in the next month. She has been my best friend since I have met her in State College.

Chapter 1

Background

1.1 Research Objectives

For a long time, the maximum likelihood method has gained its popularity in the sense that corresponding estimators are consistent, asymptotically efficient and normally distributed under some regularity conditions. However, in most cases, the ML estimator is not robust to outliers and in some cases, it is not even consistent when a given model does not satisfy some regularity conditions.

In this thesis, we study how we can modify the maximum likelihood method in a general sense in order to produce a reasonable estimator when the ML method breaks down. We suggest the minimum distance estimation with kernel smoothing and prove its universal consistency. Although one of the good aspects in the distance based estimation is to give us robust estimators, in this thesis we will concentrate on why the usual ML method does not work using some examples and how the proposed method can repair MLE.

There are several known cases that show the ML estimator is not consistent in different statistical problems. For example, Roeder et al. (1996) indicated that the semiparametric mixture model in a measurement error model produces inconsistent ML estiamtes when there is an additional error-free covariate. van der Laan (1996) discussed

the non-uniqueness of the non-parametric ML estimate for the bivariate survivor function. Kiefer and Wolfowitz (1956) also indicated the unboundedness of the likelihood in a finite normal mixture model.

These examples imply that the ML method is not always a good choice for a reasonable estimator. If the ML procedure does not work, first we should understand its reason and if one understands why the usual MLE does not work, then one can hope to find a natural choice for the transformation of data or model. There are some reasons why the ML estimate breaks down. One of the reasons could be discreteness of data or discontinuity of a model. Another possible reason could be an unbounded likelihood. Therefore, we can naturally think about kernel smoothing as a tool to regularize the model or data so that we can remove the irregular part of the model or likelihood. In this case, the initial estimation problem will be distorted and this could cause a bias or information loss. The question is how we can construct estimating procedure without distorting the initial statistical problem.

In Part I, we address the failure of the ML method using several well-known examples and discuss why these failures occur. Then we describe the proposed method called *doubly-smoothed maximum likelihood estimation* (DSMLE) with its universal consistency. We also discuss some computational aspects of DSMLE and a simple example to show how DSMLE rectifies the failure of the ML procedure. In Part II, we introduce the semiparametric mixture approach in the measurement error problem. A long lasting unsolved problem in this approach is the inconsistency of the ML method when there are additional error free covariates as well as covariates measured with error. We apply the proposed method to solve this inconsistency problem and show its consistency.

1.2 Minimum distance estimation for the parametric models

In a distance point of view, the ML method is to compare data and a certain family of models then it finds the closest model to the data in a given family of models. Thus the distance based method can be viewed as the generalized version of the ML approach and in some case, the MLE is a minimizer of the Kullback-Leibler distance between a model and observed data. The minimum distance estimation has been used under various distances and purposes. In this section, we review the minimum distance estimation in the parametric model.

1.2.1 Discrete model

Suppose the sample space of a random variable X is discrete and the model density for X, $m_\theta(x)$, is a family of discrete parametric probability densities on X where $\theta \in \Omega$. Let $d_n(x)$ be the proportion of the n observations which have value x. Then $d_n(x)$ is the empirical density function.

Based on this framework, the minimum distance estimators are constructed by minimizing a certain statistical distance between the empirical density d_n and the parametric model density $m_\theta(x)$ over $\theta \in \Omega$. There are some well-known statistical distances such as Kullback-Leibler (KL), Hellinger (HD), and Pearson's chi-square (PCS). The forms of these distances are

$$KL(d_n, m_\theta) = \sum_x m_\theta(x) \ln\left(\frac{m_\theta(x)}{d_n(x)}\right)$$

$$PCS(d_n, m_\theta) = \sum_x d_n(x) \frac{(d_n(x) - m_\theta(x))^2}{m_\theta(x)}$$

$$HD(d_n, m_\theta) = \sum_x \left(\sqrt{d_n(x)} - \sqrt{m_\theta(x)}\right)^2$$

In addition to these well-known distances, Cressie and Read (1984) introduced a class of disparity measures called family of power divergences (PWD). This family is defined as

$$PWD(d_n, m_\theta) = \sum_x d(x) \frac{(d(x)/m_\theta(x))^\lambda - 1}{\lambda(\lambda + 1)}$$

Lindsay (1994) also introduced blended families of disparities such as blended weighted Hellinger distances (BWHD) and blended weighted chi-square distances (BWCS). They are defined by

$$BWCS(d_n, m_\theta) = \sum \frac{\{d_n(x) - m_\theta(x)\}^2}{2\{\alpha d_n(x) + (1-\alpha) m_\theta(x)\}}$$

$$BWHD(d_n, m_\theta) = \sum \frac{\{d_n(x) - m_\theta(x)\}^2}{2\{\alpha \sqrt{d_n(x)} + (1-\alpha)\sqrt{m_\theta(x)}\}^2}$$

Although we call them *distance*, we do not require them to be actual metrics in the mathematical sense. For instance, Kullback-Leibler distance is not symmetric and does not satisfy the triangle inequality. However, all distances should be nonnegative and zero-distance should mean two densities are same with probability one. With these two properties, statistical distances can be used in important statistical purposes.

If we interpret the statistical distance as a loss between data and a model, statistical distance can be used as an important model selection tool. Another important usage of statistical distance is an estimation. In this case, we could get robustness and first

order efficiency. Lindsay (1994) studied these robustness and the first order efficiency by introducing residual adjustment function.

1.2.2 Continuous model

When we have a continuous model, the aforementioned distance based estimation cannot directly applied to the continuous model because the data is always discrete while the model is not.

The minimum distance estimation for the continuous case is first studied by Beran (1977). He proposed to estimate an unknown parameter θ by minimizing the Hellinger distance between the model density $m_\theta(x)$ and the nonparametric density estimator $\hat{f}_n$. In his paper, he showed the minimum Hellinger distance estimator (MHDE) could obtain robustness with first order efficiency. Since then, several authors followed up on his work such as Tamura and Boos (1989) and Simpson (1987, 1989).

These methods require that the estimated kernel density $\hat{f}_n$ should be consistent for the true model density. To get this consistency property, we need some complicated conditions for the kernel and tuning parameter in estimating the kernel density estimate $\hat{f}_n$. However, Basu and Lindsay (1994) showed that if m_θ is replaced with m_θ^* which is m_θ smoothed with the same kernel used in $\hat{f}_n$, we do not need complicated conditions. Moreover, we still get robustness and, at least in some cases, first order efficiency. Here we briefly describe their methodology.

Suppose $m_\theta(x)$ is a continuous model density and $X_1, \cdots, X_n$ are random sample from $m_\theta(x)$. Using a kernel $K_h(\cdot)$, we can construct a kernel density estimator

$$\hat{f}_n(t) = \int K_h(t-x) d\hat{F}_n(x) \tag{1.1}$$

where h is a tuning parameter and $\hat{F}_n(x)$ is the empirical distribution obtained from the sample. By applying the same kernel with the same tuning parameter h to the model density, we have a smoothed model density

$$m^*_\theta(t) = \int K_h(t-x) m_\theta(x) dx. \tag{1.2}$$

Now, find θ that minimizes density based distance between $m^*_\theta(t)$ and $\hat{f}_n(t)$ such as the squared Hellinger distance

$$HD(\hat{f}_n(t), m^*_\theta(t)) = \int \left(\sqrt{\hat{f}_n(t)} - \sqrt{m^*_\theta(t)} \right)^2 dt$$

and the Kullback-Leibler distance

$$KL(\hat{f}_n(t), m^*_\theta(t)) = \int \ln \left(\frac{\hat{f}_n(t)}{m^*_\theta(t)} \right) \hat{f}_n(t) dt.$$

One of the advantages of this method is that it does not require to let the tuning parameter h go to zero. That is, for a fixed h, the sequence of the estimators based on such a statistical distance is consistent.

1.3 Quadratic distance

Lindsay et al. (2007) introduced a new family of distances called *quadratic distance*. This section presents their foundational work on quadratic distance. Since the quadratic distance is based on the nonnegative definite kernel, here we first define the nonnegative kernel and then the quadratic distance.

DEFINITION 1.1. *If the quadratic form* $\iint K(s,t)d\sigma(s)d\sigma(t)$ *is nonnegative for all bounded signed measure* σ*, the kernel* $K(s,t)$ *is called* nonnegative definite *(NND). Moreover, if this nonnegativity holds for all* σ *satisfying the condition* $\int d\sigma(s) = 0$*,* $K(s,t)$ *is called conditionally nonnegative definite (CNND).*

DEFINITION 1.2. *Given a CNND* $K_{G}(s,t)$*, possibly depending on* G*, the* K*-based quadratic distance between two probability measures* F *and* G *is defined as*

$$D_{K}(F,G) = \iint K_{G}(s,t)d(F-G)(s)d(F-G)(t).$$

Many statistical distances have the quadratic distance form. For example, if we let $A_1, \cdots, A_n$ be a partition of the sample space into m bins, the Pearson's chi-square can be represented as

$$\sum_{i} \frac{(F(A_i) - G(A_i))^2}{G(A_i)},$$

if we define the kernel as

$$K_{G}(x,y) = \sum_{i=1}^{m} \frac{I(x \in A_i)I(y \in A_i)}{G(A_i)}.$$

If we define the kernel as $K(s,t) = I(s = t)$, the L_2 distance can be represented by

$$\iint I(x = y)(f(x) - g(x))(f(y) - g(y))dxdy.$$

Although some of statistical distances are not the quadratic distance form, such as Kullback-Leibler and Hellinger, most smooth distances are at least locally quadratic(see Lindsay et al. (2007)). Thus this new family of distances can be viewed as a generalized version of many statistical distances.

DEFINITION 1.3. *The G-centered kernel K, denoted by $\tilde{K}^G$, is defined as*

$$\tilde{K}^G(x,y) = K(x,y) - \int K(x,y)dG(x) - \int K(x,y)dG(y) + \iint K(x,y)dG(x)dG(y)$$

An important property of the centered kernel is that the K-based quadratic distance between F and G can be written as

$$D_K(F,G) = \iint \tilde{K}^G(x,y)dF(x)dF(y). \tag{1.3}$$

This representation plays an important role in the spectral decomposition theorem and enables one to estimate the empirical spectral degrees of freedom that will explained in the following subsections.

1.3.1 Spectral decomposition and spectral degrees of freedom

The spectral decomposition theory illuminates an important asymptotic property of the quadratic distance. That is, for the true model τ and the empirical distribution

$\hat{F}_n$, we have

$$nD_K(\hat{F},\tau) \xrightarrow{w} \sum_{i=1}^{\infty} \lambda_i Z_i^2$$

where Z_i's are independent $N(0,1)$ and $\lambda_i \geq 0, \sum_{i=1}^{\infty} \lambda_i < \infty$ Thus, the limiting distribution of $nD_K(\hat{F},\tau)$ is an infinite sum of the independent scaled χ_1^2. Based on this spectral decomposition, we approximate this to one scaled χ^2_{sDOF} with an appropriate degrees of freedom sDOF.

$$nD_K(\hat{F},\tau) \xrightarrow{w} \sum_{i=1}^{\infty} \lambda_i Z_i^2 \approx c\chi^2_{\text{sDOF}}$$

By matching the first two moments (Satterthwaite, 1941), we can find c and *spectral degrees of freedom*, sDOF.

$$c = \frac{\sum \lambda_i^2}{\sum \lambda_i}, \qquad sDOF = \frac{\left(\sum \lambda_i\right)^2}{\sum \lambda_i^2} \tag{1.4}$$

For detail, see Lindsay et al. (2007). Thus, sDOF has a similar meaning to the usual degrees of freedom in χ^2 goodness-of-fit test.

1.3.2 Estimating spectral degrees of freedom

From the spectral decomposition of the τ-centered kernel (Lindsay et al., 2007, Theorem 3.1), we have

$$\int \tilde{K}_h^{\tau}(x,x)d\tau(x) = \sum_{i=1}^{\infty} \lambda_i \text{ and } \iint \left(\tilde{K}_h^{\tau}(x,y)\right)^2 d\tau(x)d\tau(y) = \sum_{i=1}^{\infty} \lambda_i^2. \tag{1.5}$$

Combining (1.4) and (1.5), the sDOF can be written as

$$\text{sDOF} = \frac{\left(\sum_{i=1}^{\infty} \lambda_i\right)^2}{\sum_{i=1}^{\infty} \lambda_i^2} = \frac{\left(\int \tilde{K}_h^{\tau}(x,x) d\tau(x)\right)^2}{\iint \left(\tilde{K}_h^{\tau}(x,y)\right)^2 d\tau(x) d\tau(y)}.$$

However, we don't know the true distribution τ. For a given data set, one can estimate sDOF as follows. First the numerator $\int \tilde{K}_h^{\tau}(x,x) d\tau(x)$ can be estimated by $\frac{1}{n}\sum_{i=1}^{n} \tilde{K}_h^{\hat{F}}(x_i, x_i)$. Second, using a U-statistic, $\iint \left(\tilde{K}_h^{\tau}(x,y)\right)^2 d\tau(x) d\tau(y)$ can be estimated by

$$\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j<i} \left(\tilde{K}_h^{\hat{F}}(x_i, x_j)\right)^2.$$

Therefore sDOF can be empirically estimated by

$$\frac{\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{K}_h^{\hat{F}}(x_i, x_i)\right)^2}{\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j<i} \left(\tilde{K}_h^{\hat{F}}(x_i, x_j)\right)^2}$$

1.4 Mixtures

In section 1.2.2, Basu and Lindsay (1994)'s estimator does not often have the closed form as a result of kernel smoothing. Thus in many cases, we need a numerical algorithm to estimate parameters. When the model includes only parametric components, the minimizing problem in the distance based estimation can be implemented using some standard minimizing methods such as the Newton-Raphson method. However, if we need to estimate non-parametric component, the estimation would be complicated.

In this case, if we express the smoothed model density (1.2) as a mixture form,

$$m_\theta^*(t) = \int K_h(t-x)m_\theta(x)dx = \int K_h(t-x)dM_\theta(x)$$

where $M_\theta(x)$ is the distribution function of $m_\theta(x)$, some well known mixture algorithms can be easily applied.

In this section, we briefly introduce the mixture NPMLE theorem (Lindsay, 1995) in the nonparametric mixture models and several algorithms to estimate the NPMLE.

1.4.1 The mixture NPMLE theorem

Suppose that we have a mixture density

$$f(x;Q) = \int f(x;\phi)dQ(\phi),$$

where Q is a mixing distribution and $f(x;\phi)$ is a atomic density. The mixture likelihood function is then $L(Q) = \prod_{i=1}^n f(x_i;Q)$, which we want to maximize. The first part of the mixture MLE theorem mentions that (1) the nonparametric maximum likelihood estimator $\hat{Q}$ exists, (2) $\hat{Q}$ is necessarily discrete and (3) the number of support points for $\hat{Q}$ is no more than sample size n.

The second part is a gradient characterization. Maximization of $L(Q)$ over all possible Q is a very difficult problem, especially for the computation because we do not even know the number of support points for Q as well as the location of support points and their weights. The gradient characterization gives us a simpler way to find the NPMLE $\hat{Q}$. Suppose that Q_0 is a candidate for the NPMLE $\hat{Q}$, and that Q_1 is any

other distribution. By letting $Q_\alpha = (1-\alpha)Q_0 + \alpha Q_1$, we can construct an intermediate distribution between Q_0 and Q_1, where $0 \leq \alpha \leq 1$. Now the nonparametric likelihood becomes a one-parameter likelihood function $L^*(\alpha) = L(Q_\alpha)$. If $\frac{d}{d\alpha} \ln L^*(\alpha)|_{\alpha=0}$ is positive, then we know not only there exists α such that $L(Q_\alpha) > L(Q_0)$ but also Q_0 is not the NPMLE of Q. We define the directional derivative of $\ln L^*(\alpha)$ at $\alpha = 0$ to be

$$D_{Q_0}(Q_1) = \sum_{i=1}^{n} \left(\frac{f(x_i; Q_1)}{f(x_i; Q_0)} - 1 \right) = \int D_{Q_0}(\delta_\phi) dQ_1(\phi),$$

where δ_ϕ is Dirac delta function whose support is on ϕ. The second part of FTNPMLE says a much stronger result. Q_0 is MLE if and only if $D_{Q_0}(\delta_\phi) \leq 0$ for all ϕ.

The third part says the support point properties. That is, if ξ is a support point for $\hat{Q}$ which maximizes likelihood, then $D_{\hat{Q}}(\xi) = 0$. The last part is about the uniqueness of the fitted vector of likelihood values. If we define the fitted vector of likelihood values as $L(\hat{Q}) = (L_1(\hat{Q}), \cdots, L_n(\hat{Q}))$, then $L(\hat{Q})$ is uniquely determined. This result does not directly mean that $\hat{Q}$ is unique, it can be however used to prove the uniqueness of $\hat{Q}$ in various situations.

1.4.2 EM algorithm

Because a mixture model can be viewed as a component missing problem, we can apply the EM algorithm. To apply the EM algorithm, we assume k-finite mixture for $f(x; Q)$, where Q is a mixing distribution. Let $\phi_1, \cdots, \phi_k$ be the support points of Q

and let $\pi_1, \cdots, \pi_k$ be the corresponding weights. The mixture density is then

$$f(x; Q) = \sum_{j=1}^{k} \pi_j f(x; \phi_j).$$

If we define a multinomial indicator vector V_i such that

$$V_{ij} = \begin{cases} 1 & \text{if } x_i \text{ comes from } f(x; \phi_j) \\ 0 & \text{otherwise} \end{cases}$$

where $i = 1, \cdots, n$ and $j = 1, \cdots, k$, then V_i has a multinomial distribution with parameters $(\pi_1, \cdots, \pi_k)$, and the joint distribution of (x_i, z_i) and the conditional distribution v_i given x_i each are

$$p(x_i, v_i) = \prod_{j=1}^{k} \pi_j f(x_i; \phi_j)^{v_{ij}}$$

$$p(v_i | x_i) = \prod_{j=1}^{k} \left(\frac{\pi_j f(x_i; \phi_j)}{\sum_{m=1}^{k} \pi_m f(x_i; \phi_m)} \right)^{v_{ij}}$$

Since we do not observe V_{ij} directly, we need the EM algorithm to maximize log-likelihood function with two steps, the E-step and the M-step. To simplify notation, let us define $V_{ij}^{(t)}$ as

$$V_{ij}^{(t)} := E(V_{ij} | x_i; \pi^{(t)}, \phi^{(t)}) = \frac{\pi_j^{(t)} f(x_i; \phi_j^{(t)})}{\sum_{m=1}^{k} \pi_m^{(t)} f(x_i; \phi_m^{(t)})}$$

In the E-step, we calculate the conditional expectation of the full log-likelihood function given x_i and $(\phi^{(t)}, \pi^{(t)})$:

$$\begin{aligned}\sum_{i=1}^{n} E(\ln p(x_i, v_i)|x_i; \pi^{(t)}, \phi^{(t)}) &= \sum_{i=1}^{n} E\left[\sum_{j=1}^{k} z_{ij} \ln\left(\pi_j^{(t)} f(x_i|\phi_j^{(t)})\right) \mid x_i\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} v_{ij}^{(t)} \ln\left(\pi_j^{(t)} f(x_i|\phi_j^{(t)})\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} v_{ij}^{(t)} \ln \pi_j^{(t)} + \sum_{i=1}^{n} v_{ij}^{(t)} \ln f(x_i|\phi_j^{(t)})\end{aligned}$$

In the M-step, $(\pi^{(t)}, \phi^{(t)})$ is updated to $(\pi^{(t+1)}, \phi^{(t+1)})$ by maximizing the expected log-likelihood over (π, ϕ) :

$$\begin{aligned}\pi_j^{(t+1)} &= \frac{\sum_{i=1}^{n} v_{ij}^{(t)}}{n} \\ \phi_j^{(t+1)} &= \arg\max_{\phi_j} \sum_{i=1}^{n} v_{ij}^{(t)} \ln f(x_i|\phi_j^{(t)})\end{aligned}$$

The EM algorithm is usually used in the finite mixture models. However, the EM algorithm can be also used in the estimation of NPMLE assuming the mixing distribution has finite number of support points hence multinomial distribution. In this case, we need to determine the number of support points beforehand.

1.4.3 Gradient based algorithms

To compute the NPMLE of the mixing distribution, there are several well-known gradient based methods such as *vertex direction method* (VDM) (Bohning, 1985), *vertex exchange method* (VEM) (Bohning, 1986), and *intra-simplex direction method* (ISDM)

(Lesperance and Kalbfleisch, 1992). All these methods detect the violation of the second part of the mixture NPMLE theorem: Q_0 is NPMLE if and only if $D_{Q_0}(\Delta_\phi) \leq 0$ for all ϕ.

VDM and VEM find the global maximizer $\hat{\phi}$ of the gradient function. If the gradient function is greater than zero at $\hat{\phi}$, then these algorithms add the global maximum as a new support point and determine the corresponding weight in each step. In this case, one typical difficulty is that the gradient function has several modes, and sometimes it is too bumpy. This feature of the gradient function makes it difficult to find the global maximizer. So, a fine grid is needed to find the global maximum which takes much time, especially for the large sample size and a finer grid.

ISDM uses a more clever way to accelerate these algorithms. That is, ISDM finds all local maxima and adds these all local maximizers as new support points. However, we still need grid search. To lessen this computational difficulty in finding new set of support points, Yang (2004) suggested *splitting rule*. Suppose that we have a current mixing distribution with support points $\{\phi_1, \cdots, \phi_n\}$. Then for the gradient function $D_{\hat{Q}_n}(\phi)$ we have $D_{\hat{Q}_n}(\phi_j) = 0$, $D'_{\hat{Q}_n}(\phi_j) = 0$, and $D''_{\hat{Q}_n}(\phi_j) \leq 0$ from the third part of the mixture NPMLE theorem. Thus, if $D''_{\hat{Q}_n}(\phi_j) > 0$, we would say that a local violation occurs at ϕ_j. When this violation occurs at ϕ_j, we can know $D_{\hat{Q}_n}(\phi)$ is greater than 0 in the neighborhood of ϕ_j, which implies that at least one more support point in the neighborhood of ϕ_j is required. To do so, we split ϕ_j into ϕ_{j1} and ϕ_{j2} and assign half the weight of ϕ_j. This splitting rule could greatly reduce the computing time in updating the support points for NPMLE.

Part I

A universally consistent modification of maximum likelihood

Chapter 2

Introduction

Although the many successes of the maximum likelihood method make it seem like a nearly foolproof way to create good estimators, there are reasonable models where the estimators fail to be consistent. These models are both parametric and semiparametric. In this Part I, we will give several examples of this and investigate why it occurs. We here consider a simple amendment to maximum likelihood that makes it universally consistent. 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; moreover, the estimators can be made arbitrarily close to maximum likelihood by moving the bandwidth to zero.

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 it is essentially arbitrary. That is, there exist arbitrarily silly ways to parametrize. 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 $\{Y_1, \cdots, Y_n\}$ is a random sample from some unknown probability measure M_τ. 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 the 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 will here consider models both parametric and semiparametric, and from that point of view, it is best to call θ the model index rather than the parameter, recognizing that there are many possible ways to index the class of models.

Before we present our main result, in the next section, we show some motivating examples that lead us to inconsistent MLE's.

2.1 Motivating examples

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

Example 2.1. Two-component normal mixture (Kiefer and Wolfowitz, 1956)

Consider a two-component normal mixture with unknown means, variances and class

probability $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$. Then the likelihood of a sample from this density is given by :

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

If we do not assume equal variance, this likelihood is unbounded and its global maximum is ∞ : let $\mu_2 = x_1$ and let σ_2^2 go to zero. Therefore likelihood is not bounded and the parameter values that give the infinite spikes cannot be used to construct a consistent sequence of estimators.

As a simple amendment, now suppose that each X_i was replaced by $X_i^* = X_i + \sqrt{h}Z_i$, where Z_i's are i.i.d. $N(0,1)$. In this measurement error model, from the convolution property of normal densities the density for X_i^* is

$$\int N(x^*; x, h) \left\{ \rho N(x; \mu_1, \sigma_1^2) + (1-\rho) N(x; \mu_2, \sigma_2^2) \right\} dx$$

$$= \rho N(x^*; \mu_1, \sigma_1^2 + h) + (1-\rho) N(x^*; \mu_2, \sigma_2^2 + h)$$

and the likelihood is then

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

Then we can see that this likelihood is bounded above, showing that adding measurement error is a means to remove infinite spikes from parametric likelihood functions. The basic idea of the proposed method in this thesis achieves this. Because the main reason

of this inconsistency came from the irregularity of model, we need to regularize the model. In this example, we can see adding a measurement error to the original variable or equivalently kernel smoothing can regularize the given model. We will discuss this example in chapter 6 with a simplified normal mixture model after describing our main result.

Our next example involves a nonparametric maximum likelihood estimator. A consistent example 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 2.2. The bivariate Kaplan-Meier estimator

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 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 $\boldsymbol{C}$. In this case, the usual Kaplan-Meier estimator is not unique. The Kaplan-Meier estimator is inconsistent due to the singly censored data points.

The problem with singly censored points can be easily explained using a redistribution-to-the-right algorithm for maximum likelihood introduced by Efron (1967). In Figure 2.1(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 quadrant of the point A. However, in Figure 2.1(b), the T_1-coordinate of point A is observed but T_2-coordinate is right censored. In this case, the mass of point A can be redistributed to any point on the dotted line. Therefore, 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.

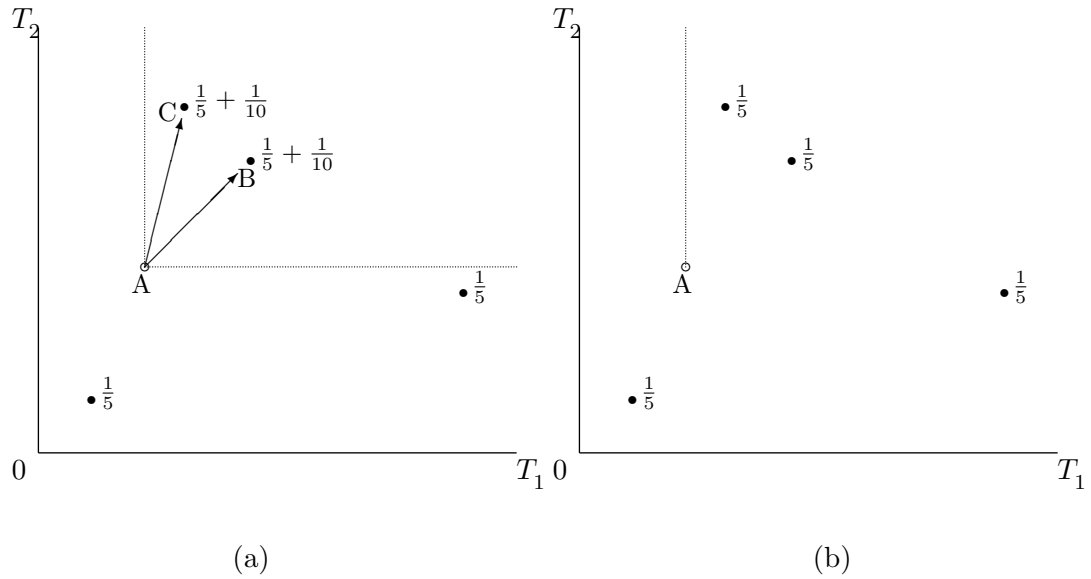


Fig. 2.1. NPMLE of survival time when (a) bivariate data only includes doubly censored or doubly uncensored data (b) bivariate data includes singly censored censored data

A singly censored data point could be described as having one coordinate observed precisely and the other vaguely.

EXAMPLE 2.3. Error-in-variables

Consider a bivariate random variable (X, Z) with unknown distribution function $G(x, z)$. Suppose Z has no measurement error but random variable X can not be directly observed, instead one observes W which is X perturbed by some measurement error. Suppose 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 let us consider the estimation of the nonparametric distribution G. As in the preceding example, we restrict attention to G discrete. Because X is not observed the observed likelihood is

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

Assuming that G is discrete we can rewrite the likelihood of the sample as

$$L(G) = \prod_{i=1}^{n} \int f(w_i|x = \xi)I(z = z_i)dG(\xi, z). \tag{2.1}$$

In this case, if the data have no ties and $f(w_i|x = \xi)$ is completely known unimodal density $f(w - x)$ with mode 0, then the ML estimate for G is the empirical distribution of (W, Z), which clearly converges to the wrong distribution (Roeder et al., 1996; Gaydos, 1997).

An explanation for this inconsistency is that the joint conditional density, $f(w_i|x = \xi)I(z = \eta)$, is continuous for W but discontinuous for Z. Because of this mixed form of continuous and discrete variables, when the ML procedure estimates the conditional distribution of $X|Z = z_i$, it fails to pool information across Z observations. This can be seen in Figure 2.2. In Figure 2.2(a), if Z observations are discrete, given each z_i, there will be several W observations. This allows ML method to consistently estimate the conditional distribution of X given Z. However, in Figure 2.2(b), if Z is continuous, there will be only one observation for each Z observation with probability one even for infinite sample size. So ML method fails to pool information across different Z observations. Consequently, the NPMLE of G is unique but inconsistent. However, if both X and Z had been measured with error, there would have no inconsistency. We will study this example in Part II to show how this inconsistency can be solved using the doubly-smoothed maximum likelihood method.

In all three of our 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. Recently, Luo et al. (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 lose information but also be simulation dependent for the same data set. The method we consider removes this problem. In the next chapter, we describe the doubly-smoothed maximum likelihood procedure and show its universal consistency.

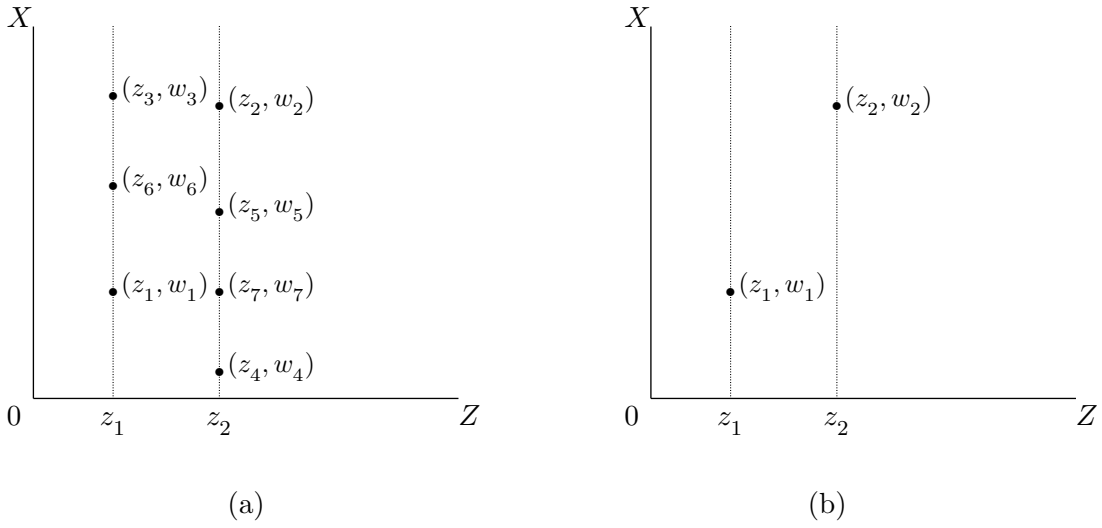


Fig. 2.2. Estimation of conditional distribution of X given $Z = z_i$ when (a) Z is discrete and (b) Z is continuous

Chapter 3

Doubly-smoothed maximum likelihood estimation

In the aforementioned examples, the inconsistency of MLE is due to the failure of some regularity conditions that are typically used in the general consistency proof. We ask how to regularize the model while losing little efficiency.

One simple amendment is to smooth the model or the data so that the model or data is regularized. The risk is that this could cause serious bias and extra variation depending on the degree of blurring. Thus determining the optimal degree of blurring is another major issue across various models. To minimize this undesirable blurring effect from having regularized either model or data, we suggest smoothing both model and data with the same degree of smoothing. By this means, we change data and model in a parallel way in order to not only cure the defect of the ML method but also reduce the bias caused by blurring. Moreover, it can make the choice of smoothing parameter a less difficult problem. That is, the consistency of our method does not depend on the choice of a tuning parameter and the proposed method is quite robust to the choice of a tuning parameter.

3.1 Description of method

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

estimator

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

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

$$m^*(t; M_\theta) = \int K_h(x,t)dM(x;\theta) \tag{3.1}$$

where $M(x;\theta)$ is a distribution function of $m(x;\theta)$. 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, smoothed kernel density can be considered as a nonparametric estimator for the density of new variable T. For our methodology, We will rely on the following basic assumptions for the kernel.

(K1) 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 $|x| \to \infty$.

(K2) 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.*

The first kernel assumption is very common assumption in the literature. The second kernel assumption is needed in our consistency proof, as it assures that the weak convergence of kernel smoothed probability measure will imply 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 transformation.

Under these assumptions, we can see that 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 for each t because the empirical distribution $\hat{F}_n$ converges weakly to $M(x;\theta)$ on a set of probability one. This convergence is independent of the value of the tuning parameter h as long as the same kernel and tuning parameter are used for data and model.

Now, the doubly-smoothed maximum likelihood estimator of θ (DSMLE) is defined as the minimizer of the Kullback-Leibler distance between the smoothed model density and the smoothed kernel density:

$$\hat{\theta}_n = \arg\min_\theta KL(\hat{f}_n^*(t), m^*(t; M_\theta)) = \arg\min_\theta \int \ln\left(\frac{\hat{f}_n^*(t)}{m^*(t; M_\theta)}\right) \hat{f}_n^*(t)dt. \tag{3.2}$$

We can also easily verify that minimizing (3.2) is equivalent to maximizing

$$l^*(\theta) = \int \ln m^*(t; M_\theta)\hat{f}_n^*(t)dt = \int \ln m^*(t; M_\theta)d\hat{F}_n^*(t). \tag{3.3}$$

We call (3.3) doubly-smoothed log-likelihood function because (3.3) will approach the usual log-likelihood function as the tuning parameter goes to zero.

Now, the corresponding doubly-smoothed maximum likelihood estimator of the distribution is $M_{\hat{\theta}_n}$. Although we could consider other statistical distances such as

Hellinger, chi-square, and so on, in this thesis we consider the Kullback-Leibler distance because of its relationship to the maximum likelihood method. Generally speaking, if we let h tend to zero the DSMLE will approach the maximum likelihood estimator. 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.

If the model index θ is vector-valued, then solving (3.2) is often equivalent to the solving estimating equation

$$\int \nabla_\theta \ln m^*(t; M_\theta) \hat{f}_n(t) dt = 0. \tag{3.4}$$

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) studied the consistency and efficiency of this estimator.

However, in the case that the model index θ contains nonparametric components as in Example 2.2 and 2.3, the consistency of the estimator has not been established. In the next section, we show the DSMLE $\hat{\theta}_n$ is very generally consistent for an essentially arbitrary model.

3.2 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 $\hat{F}_n^\omega(x) = \frac{1}{n} \sum I(X_i(\omega) \leq x)$ converges weakly to M_{θ_τ} for ω in a set

$\Omega_0 \subset \Omega$ satisfying $P(\Omega_0) = 1$ due to 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 the proof of consistency, we will first show the weak convergence of $M(x;\hat{\theta}_n)$ to $M(x;\theta_\tau)$ on a set of probability one. We consider the consistency of the estimated model index $\hat{\theta}_n$ to the true θ_τ in the next section. For the convenience of notation, from now on we use M_n instead of $M(x;\hat{\theta}_n)$. Similarly M_τ means $M(x;\theta_\tau)$.

By consistency we will 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 any metric $d(\cdot,\cdot)$ on the space of probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$ that metricizes weak convergence (Billingsley, 1995). We will need the following lemma in our main theorem. Let $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$.

LEMMA 3.1. *Assuming (K1), for $\omega \in \Omega_1$, a set having probability one, and for $\hat{f}_n^*(t) = \int K_h(t,x)d\hat{F}_n^\omega(x)$, we have*

$$\limsup_n \int \left[\ln m^*(t;M_\tau)\right]^- \hat{f}_n^*(t)dt \to \int \left[\ln m^*(t;M_\tau)\right]^- m^*(t;M_\tau)dt \tag{3.5}$$

$$\limsup_n \int \left[\ln m^*(t;M_k)\right]^+ \hat{f}_n^*(t)dt \to \int \left[\ln m^*(t;M_0)\right]^+ m^*(t;M_\tau)dt \tag{3.6}$$

Proof : To prove (3.5), first we apply Fubini's theorem for the nonnegative function.

$$\int \left[\ln m^*(t;M_\tau)\right]^- \hat{f}_n^*(t)dt = \int \left[\ln m^*(t;M_\tau)\right]^- \int K_h(t,x)d\hat{F}_n^\omega(x)dt$$

$$= \iint \left[\ln m^*(t;M_\tau)\right]^- K_h(t,x)dtd\hat{F}_n^\omega(x) \tag{3.7}$$

$$= \frac{1}{n}\sum_i \int \left[\ln m^*(t;M_\tau)\right]^- K_h(t,x_i)dt \tag{3.8}$$

If $\int \left[\ln m^*(t;M_\tau)\right]^- m^*(t;M_\tau)dt = \infty$, equation (3.5) holds from SLLN (Chung, 1974, Theorem 5.4.2). Now suppose $\int \left[\ln m^*(t;M_\tau)\right]^- m^*(t;M_\tau)dt < \infty$. Using the strong law of large numbers and Fubini theorem again (3.8) converges to

$$\iint \left[\ln m^*(t;M_\tau)\right]^- K_h(t,x)dtdM_\tau(x) = \int \left[\ln m^*(t;M_\tau)\right]^- \int K_h(t,x)dM_\tau(x)dt$$

$$= \int \left[\ln m^*(t;M_\tau)\right]^- m^*(t;M_\tau)dt$$

on a set Ω_1' of probability one. For equation (3.6), we apply the extended version of the dominated convergence theorem. The boundedness of the kernel K_h implies that $m^*(t;M)$ is bounded above, and so there exists positive number U_h such that $\left[\ln m^*(t;M_n)\right]^+ < U_h$ for all n. Then, for each n, $\left[\ln m^*(t;M_n)\right]^+ \hat{f}_n^*(t) < U_h\hat{f}_n^*(t)$, so we use $U_h\hat{f}_n^*(t)$ as a dominating sequence. It satisfies $U\hat{f}_n^*(t) \to U_h m^*(t;M_\tau)$ for all $t \in \mathbb{R}^d$ on Ω_0 from the consistency of $\hat{F}_n^\omega(t)$, where $\int U_h m^*(t;M_\tau)dt = U_h < \infty$. Now, the extended version of dominated convergence theorem implies (3.6). Let $\Omega_1 = \Omega_0 \cap \Omega_1'$ to finish the result. □

For the next theorem, we assume that the maximizer of the doubly-smoothed log-likelihood function exists. To assure the existence of this maximizer, we may need

some assumptions on the class of model distribution $\mathcal{M}$ such as the compactness of $\mathcal{M}$ in the weak topology. However, this theorem still applies to any sequence of $\hat{M}_n$'s such that $l^*(M_n) \geq l^*(M_\tau)$; there always exists such a sequence as long as $M_\tau \in \mathcal{M}$. Hence, in the next theorem, the maximizer M_n of l^*(or the minimizer of $KL(\hat{f}_n^*, m^*)$) can be interpreted as either the global maximizer or a sequence satisfying $l^*(M_n) \geq l^*(M_\tau)$.

THEOREM 3.1. *Let $\mathcal{M} = \{M_\theta\}$ be a class of model distributions indexed by θ. Suppose that $(X_1, \cdots, X_n)$ is a random sample from true distribution $M_\tau \in \mathcal{M}$. Suppose (K1) and (K2) are satisfied, then the minimizer $\hat{M}_n$ of $KL(\hat{f}_n^*, m^*)$ weakly converges to M_τ on a set of probability one.*

Proof : Fix $\omega \in \Omega_1$. 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 always select a further subsequence $\{k\} \subset \{m\}$ such that $\hat{M}_k$ is vaguely convergent to a subprobability measure M_0. If we can show that $M_0 = M_\tau$, then we are done by the method of subsequences (Chung, 1974, Theorem 4.3.4). One can easily justify the following sequence of inequalities.

$$
\begin{aligned}
0 &\geq \liminf_k \int \ln\left(\frac{m^*(t;M_\tau)}{m^*(t;\hat{M}_k)}\right) \hat{f}_k^*(t)dt \\
&= \liminf_k \int \left(\left[\ln m^*(t;M_\tau)\right]^+ - \left[\ln m^*(t;M_\tau)\right]^- - \left[\ln \hat{m}^*(t;\hat{M}_k)\right]^+ + \left[\ln m^*(t;\hat{M}_k)\right]^- \right) \hat{f}_k^*(t)dt \\
&\geq \liminf_k \int \left(\left[\ln m^*(t;M_\tau)\right]^+ + \left[\ln m^*(t;\hat{M}_k)\right]^- \right) \hat{f}_k^*(t)dt \\
&\quad - \limsup_k \int \left(\left[\ln m^*(t;M_\tau)\right]^- + \left[\ln m^*(t;\hat{M}_k)\right]^+ \right) \hat{f}_k^*(t)dt \\
&\geq \int \liminf_k \left(\left[\ln m^*(t;M_\tau)\right]^+ + \left[\ln m^*(t;\hat{M}_k)\right]^- \right) \hat{f}_k^*(t)dt \\
&\quad - \limsup_k \int \left[\ln m^*(t;M_\tau)\right]^- \hat{f}_k^*(t)dt - \limsup_k \int \left[\ln m^*(t;\hat{M}_k)\right]^+ \hat{f}_k^*(t)dt \qquad (3.9)
\end{aligned}
$$

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. The first integral in the last expression of (3.9) is equal to

$$\int \left(\left[\ln m^*(t; M_\tau) \right]^+ + \left[\ln m^*(t; M_0) \right]^- \right) m^*(t; M_\tau) dt$$

because for the given ω, $m^*(t; \hat{M}_k)$ converges to $m^*(t; M_0)$ and $f_k^*(t)$ converges to $m^*(t; M_\tau)$. From the lemma 3.1, the second integral converges to

$$\int \left(\left[\ln m^*(t; M_\tau) \right]^- + \left[\ln m^*(t; M_0) \right]^+ \right) m^*(t; M_\tau) dt$$

So the last expression in (3.9) converges to

$$\int \left(\left[\ln m^*(t; M_\tau) \right]^+ - \left[\ln m^*(t; M_\tau) \right]^- - \left[\ln m^*(t; M_0) \right]^+ + \left[\ln m^*(t; M_0) \right]^- \right) m^*(t; M_\tau) dt$$

$$= \int \ln \left[\frac{m^*(t; M_\tau)}{m^*(t; M_0)} \right] m^*(t; M_\tau) dt \geq 0 \tag{3.10}$$

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, Theorem 4.3.4). This also implies that $\hat{M}_n$ weakly converges to M_τ. □

3.3 Consistency of index θ

Theorem 3.1 establishes the consistency of the estimated probability measure $\hat{M}_n$ but not the consistency of the model index $\hat{\theta}_n$. However, using this theorem the consistency of the model index is often easily established. For the consistency of $\hat{\theta}_n$, we need first to identify a metric for convergence, say $d(\theta_0, \theta_1)$, which would ordinarily be Euclidian 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 : $M(x;\theta_n) \to M(x;\theta_0)$ implies that $d(\theta_n, \theta_0) \to 0$.

COROLLARY 3.1. *If the kernel assumptions and model index assumptions hold, then the minimizer $\hat{\theta}_n$ of $KL(\hat{f}^*_n, m^*)$ is consistent.*

The natural metrics $d(\cdot,\cdot)$ to apply to model indices which are themselves distributions, as in examples 2.2 and 2.3, are those metricizing weak convergence. One can then apply subsequence arguments to prove consistency. For a simple example, suppose one wishes to prove consistency of G estimation in example 3 when G_τ is the true distribution. If the sequence G_n is not weakly convergent to G_τ, then we could find a subsequence $\{m\} \subset \{n\}$ such that G_m converges vaguely to subdistribution G_1 with $G_1 \neq G_\tau$.

This implies M_{G_m} converges vaguely to M_{G_1}. However, Theorem 4.2 implies M_{G_n} converges weakly to M_{G_τ}. This, together with identifiability and continuity of

model index θ_n raises a contradiction to the convergence of $\hat{M}_n$. Similar technique can be applied when model index θ includes both vector valued parameters and distributions.

Unlike the usual ML estimator that requires several regularity conditions on the model, if we smooth both the model and the data, we do not require any regularity conditions such as continuity and boundedness of the model. This explains how our methodology can cure an inconsistent ML estimator. Moreover, this proof does not require a specific form of θ. That is, θ can be a set of parameters or non-parametric distributions or both. So it can be easily applied to other consistency studies for the nonparametric model or semi-parametric model. Finally, this proof implies that the consistency does not depend on the choice of tuning parameter.

Chapter 4

Choice of kernel and tuning parameter

4.1 Choice of kernel

As we have discussed, kernel smoothing can be used as a means of regularizing a model which is irregular in some sense. Although any kernel satisfying the kernel assumptions (K1) and (K2) can be used for our estimation, in practice there may exist the most appropriate kernel for any given problem. However, since the objective of using a kernel is not for estimating densities but for obtaining two regularized densities based on the model and the data, the corresponding tuning parameter plays a more important role than the kernel does.

If one thinks the choice of the kernel does not affect estimation as much as tuning parameter selection, we may choose a kernel which makes the calculations easier because numerical integration is required otherwise. That is, it might be a good idea to use the kernel that makes the smoothed model density closed form just as one often uses a conjugate prior in Bayesian world. This is not always possible, but in many cases, such a kernel can be found. For example, if we use the normal kernel with the normal model density to construct a smoothed model density, we know that smoothed model density is also normal because of the convolution property of normal densities.

4.2 Choice of tuning parameter

In this section, we suggest a reasonable range for the tuning parameter using spectral degrees of freedom suggested by Lindsay et al. (2007); Ray and Lindsay (2007). Note that our objective of this section is not to give an optimal choice of a tuning parameter in a general sense but to give a method to construct an appropriate range for the tuning parameter because the choice of a tuning parameter should be different over different problems.

The most popular approach to choosing a tuning parameter would be a model-specific method which selects the tuning parameter that makes the mean squared error small. However, if our purpose is to find the consistent solution from a certain estimation problem which has an inconsistent MLE, we should focus on the tuning parameter that can correct the failure of MLE. Of course, the consistency will hold for any fixed tuning parameter, but there will exist an appropriate range of a tuning parameter based on a given finite sample problem. If one can find the range that produce a reasonable solution, the smallest value in the range should be the answer because we do not want to be far away from MLE as long as the problem of MLE is fixed. Now the question is how small it should be.

Before we answer this question, let us consider the role of the tuning parameter in the χ^2 goodness-of-fit test. Choosing the tuning parameter in our estimation problem is analogous to choosing the number of bins in a histogram. In DSMLE, we smooth both data and model and then compare two smoothed densities to minimize the Kullback-Leibler distance because we think the smoothing can be used as a tool to prevent this

failure. If we choose a large tuning parameter, the resulting estimators could be robust and successfully cure the failure of MLE, but they would not detect some important discrepancies between original model density and empirical density, which lead us a bias or information loss. On the other hand, if we choose a small tuning parameter, the resulting estimators could be more efficient than those with a large tuning parameter but we could fail to construct a consistent sequence of estimators for the model index θ because the model or data is not sufficiently regularized.

Although choosing the tuning parameter should play a minor role in our estimation problem, for above reasons, it is desirable to suggest, at the least, a reasonable range. And sDOF, introduced in section 1.3.1, can give a rough guideline for this purpose with a good theoretical and intuitive bases.

4.2.1 Connection between quadratic distance and DSMLE

To illustrate the connection between the quadratic distance and DSMLE, let $\hat{f}_n^*$ be the smoothed kernel density based on sample and let m_θ^* be the smoothed model density based on the model density m. In the minimum distance estimation, suppose we use the doubly-smoothed L_2 distance between $\hat{f}_n^*$ and m_θ^*:

$$\int \left(\hat{f}_n^*(t) - m^*(t)\right)^2 dt \tag{4.1}$$

Then, from Lindsay et al. (2007, Proposition 1), this L_2 distance can be written as the quadratic distance form:

$$D_H(\hat{F}_n, M_\theta) = \int H(x,y)d(\hat{F}_n - M_\theta)(x)d(\hat{F}_n - M_\theta)(y) \tag{4.2}$$

where $H(x,y) = \int K_h(t,x)K_h(t,y)dt$. Hence the L_2 distance between two smoothed densities also has the form of a quadratic distance with the kernel H.

Now the quadratic distance (4.2) between the empirical distribution $\hat{F}_n$ based on data and the model distribution M_θ represents the departure of the model from the data. The minimum distance estimation based on L_2 distance finds the model that minimizes this departure in an assumed class of models. Minimum distance estimation based on the Kullback-Leibler distance, while not identical, might be expected to have similar behavior under possible choices of a tuning parameter.

Roughly speaking, if the quadratic distance captures the discrepancy between data and a model, minimum distance estimation should also be reasonable. For this reason, we hypothesize that if an appropriate tuning parameter is chosen for the quadratic distance, the minimum distance estimator is also appropriate, at least, in a distance point of view. Since there is a simple and easy degrees of freedom calculation for the L_2 distance, we will use it here.

4.2.2 General strategy

In the χ^2 goodness-of-fit test, a rough rule for the degrees of freedom in one dimensional data is that it should be greater than 5 and less than $n/5$, where n is the

number of observations. Applying this rough rule to our sDOF, a reasonable range for the sDOF can be obtained. Thus for a given tuning parameter h, if estimated sDOF is less than the lower bound of this range, h is too large. If sDOF is greater than the upper bound of this range, h is too small. If we accept this rule, this rule can also give a guideline for the choice of the tuning parameter in our estimation.

When it is very difficult to find an appropriate method to choose the tuning parameter, sDOF always gives simple and dimensionless information for the tuning parameter just as the usual degrees of freedom does. Although the sDOF gives the range of a tuning parameter rather than one optimal value, this is not so sensitive issue in our estimation because DSMLE is quite robust to the choice of the tuning parameter, at least within the chosen range of the tuning parameter based on sDOF.

Probably, in some cases, there would be some clues to determine an optimal tuning parameter based on a given problem. However, even in this case, it might require numerical search on the predetermined grid points to choose a tuning parameter because the estimator is usually not an explicit form. If the area of grid is large or a fine grid is required, the computing time will dramatically increase. In such a case, sDOF is also a very useful tool to narrow the range of values to consider.

Chapter 5

Computation

DSMLE can be obtained by minimizing Kullback-Leibler distance between smoothed model density and smoothed kernel density. However, an explicit expression for (3.2) is often difficult to derive due to some of integrals. In this case, numerical integration methods are required such as Simpson's rule, Gaussian quadrature formula and so on. However, these numerical integrations often increase computational complexity especially for the high dimensional data and nonparametric estimation. In this chapter, we suggest two useful ways to carry out the proposed estimation procedure.

5.1 Simulation based integration

Regarding numerical integration, the easiest way would be Monte-Carlo integration because it is intuitively easy to understand and program. For DSMLE, using Monte-Carlo integration can turn minimizing the Kullback-Leibler distance problem into a likelihood maximizing problem, but with a Monte-Carlo sample and the smoothed model density. In this case we can borrow some well-known optimizing techniques from ML estimation such as Newton-Raphson type methods and EM type methods.

To explain this, let us consider maximizing (3.3) that is equivalent to minimizing Kullback-Liebler distance between the smoothed model density and the smoothed kernel

density:

$$l^*(\theta) = \int \ln m^*(t; M_\theta)\hat{f}_n^*(t)dt = \int \ln\left(m^*(t, M_\theta)\right)\frac{1}{n}\sum_{i=1}^{n} K_h(x_i, t)dt$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int \ln\left(m^*(t, M_\theta)\right) K_h(x_i, t)dt \tag{5.1}$$

If we apply the Monte-Carlo integration to each integral, for each x_i, we generate b Monte-Carlo observation $(t_{i1}, \cdots, t_{ib})$ from $K_h(x_i, t)$. Then (5.1) can be approximated by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}\sum_{j=1}^{b}\ln(m^*(t_{ij}, M_\theta)) \tag{5.2}$$

Under this numerical integration in (5.2), the initial minimizing problem is reduced to finding MLE when a model is $m^*(t; M_\theta)$ and data is given by $\{t_{ij}; i = 1, \cdots, n, j = 1, \cdots, b\}$. Hence standard optimization methods can be used. Similar methods can be found in Wang (2004) and Mcfadden (1989).

One good aspect of this method is that we can use known statistical packages to estimate $\hat{\theta}^{DSMLE}$ assuming the model is m^* and the data is $\{t_{ij}; i = 1, \cdots, n, j = 1, \cdots, b\}$ as long as the smoothed model density has a known explicit form. The fact that it is equivalent to finding MLE with m^* and a Monte-Carlo sample enables us to easily estimate DSMLE of nonparametric distribution. For instance, if M_θ is just an unknown nonparametric distribution M, the smoothed model density is

$$m^*(t; M) = \int \left(K_h(x, t)\right) dM(x)$$

and the smoothed likelihood $l^*(M)$ can be rewritten as

$$\frac{1}{nb}\sum_{i=1}^{n}\sum_{j=1}^{b}\ln(m^*(t_{ij};M)) = \frac{1}{nb}\sum_{i=1}^{n}\sum_{j=1}^{b}\ln\left(\int K_h(x,t_{ij})dM(x)\right) \tag{5.3}$$

Then we can see that the initial problem is reduced to finding NPMLE of a mixing distribution $M(x)$ with a mixture density $\int K_h(t,x)dM(x)$ based on the Monte-Carlo sample $\{t_{ij}; i = 1,\cdots,n, j = 1,\cdots,b\}$. In this case, the EM-algorithm (Laird, 1978) will be the simplest programming approach to estimate MMLE of $M(x)$. Other approaches involve gradient based methods such as *vertex exchange method* (VEM) (Bohning, 1985), and *intra-simplex direction method* (ISDM) (Lesperance and Kalbfleisch, 1992).

Although applying the Monte-Carlo method can give a simple way to estimate MMLE, the computing time depends on the size of Monte-Carlo sample for each datum. Since we generate b Monte-Carlo samples for each data point, the computer program runs as if the total number of data points increases by b times. If we need to use very slow algorithm like EM, the computing time could become huge. This can be worse for high dimensional data. In the next section, we suggest another method to approximate the integral that avoids this simulation based integration.

5.2 Local Laplace approximation

Another good tool which does not depend on simulation based integration is Laplace approximation. This basically relies on the Taylor expansion of the logarithmic function of the integrand. In maximizing $l^*(\theta)$, we need to calculate each integral for a fixed x_i in the summand. In order to approximate this integral, we borrow the idea

of Laplace approximation. Unlike usual Laplace approximation, we do not use Taylor expansion for whole integrand, but instead use Taylor expansion only for $\ln(m^*(t;M_\theta))$ at each x_i.

Then, for each data point x_i, the smoothed model density $\ln\left(m^*(t,M_\theta)\right)$ can be approximated for t near x_i by

$$c_{0i} + c_{1i}(t-x_i) + c_{2i}(t-x_i)^2$$

where

$$c_{0i} = \ln\left(m^*(x_i,M_\theta)\right) \qquad c_{1i} = \frac{\partial}{\partial t}\ln\left(m^*(t,M_\theta)\right)|_{t=x_i} \qquad c_{2i} = \frac{1}{2}\frac{\partial^2}{\partial t^2}\ln\left(m^*(t,M_\theta)\right)|_{t=x_i}$$

Now, the each summand in (5.1) is approximated by

$$\begin{aligned}\int \ln\left(m^*(t,M_\theta)\right)K_h(x_i-t)dt &\approx \int \left(c_{0i}+c_{1i}(t-x_i)+c_{2i}(t-x_i)^2\right)K_h(x_i-t)dt \\ &= c_{0i} + c_{1i}E_{K_h}(T) + c_{2i}Var_{K_h}(T) \qquad (5.4)\end{aligned}$$

Since in many cases, the moment of K_h can be explicitly calculated, without any further numerical calculation, (5.4) can be calculated explicitly. For instance, if normal kernel is applied for $K_h(t-x)$ with variance h, (5.4) is simply $c_{0i}+c_{2i}h$ and the smoothed likelihood (5.1) is approximated by

$$\frac{1}{n}\sum_{i=1}^{n}\left(c_{0i}+c_{2i}h\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\ln\left(m^*(x_i,M_\theta)\right) + \frac{h}{2}\frac{\partial^2}{\partial t^2}\ln\left(m^*(t,M_\theta)\right)|_{t=x_i}\right). \qquad (5.5)$$

Therefore, the objective function to maximize can be approximated by (5.5) and the first term in the summand represents the log-likelihood based on the smoothed model density and unsmoothed data and the second term represents the correction term due to data smoothing.

Maximizing (5.5) with respect to θ can be done using general optimization techniques. However, unlike Monte-Carlo integration, the approximate objective function can not be interpreted as a likelihood function. So we might not use known statistical packages. Moreover if the EM-type algorithm is required, it is not clear how to incorporate EM algorithm with (5.5) due to the correction term. If this is the case, we might need to stick to Monte-Carlo integration or use other types of algorithm to maximize (5.5) such as MM algorithms (Hunter, 2003; Hunter and Lange, 2004).

Chapter 6

Illustrative example and conclusion

In this chapter, we show the applicability of the DSMLE using a mixture model announced in Example 2.1. Studying this example has some difficulties because the likelihood function is unbounded and has several modes. Moreover, due to possible label switching problems, it is hard to see the efficiency of estimators based on simulation. Since there are five parameters, it is also difficult to investigate the shape of likelihood function in a graphical way. For these reasons, we consider a simplified normal mixture model whose likelihood is unbounded like Example 2.1.

6.1 Simulation study with a simplified normal mixture

Let us consider a two-component normal mixture model assuming the mean and variance of the first component are known to be zero and one. Assuming further we know the mixing proportion is 0.5, the simplified normal mixture density is

$$f(x;\mu,\sigma^2) = \frac{0.5}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) + \frac{0.5}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{6.1}$$

Now, one can easily show the likelihood function of μ and σ^2 is not bounded similar to that of Example 2.1.

Hathaway (1985) suggested the constrained MLE that restricted the parameter space to $\Omega = \{(\mu,\sigma^2) : -\infty < \mu < \infty, \sigma^2 \geq c\}$ to resolve this unbounded likelihood

problem. But if the true parameter is not included in Ω, the constrained MLE can lie on the boundary of Ω. Tan et al. (2006) suggested the sequentially constrained MLE by letting c go to zero and proved that it is consistent when $\log c_n > -k(\log^2 n)$. Although the constrained MLE gives a simple amendment to remove infinite spike in the likelihood, determining the sequence c_n or c is still a big practical issue.

In the doubly-smoothed maximum likelihood method, if a normal kernel is used with tuning parameter h, the smoothed model density is

$$m^*(x;\mu,\sigma^2) = \frac{0.5}{\sqrt{2\pi(1+h)}} \exp\left(-\frac{x^2}{2(1+h)}\right) + \frac{0.5}{\sqrt{2\pi(\sigma^2+h)}} \exp\left(-\frac{(x-\mu)^2}{2(\sigma^2+h)}\right)$$

Therefore the likelihood function based on $m^*(x;\mu,\sigma^2)$ is now bounded above. Using smoothed model density has similar effect to that of the constrained MLE. However, because the doubly-smoothed maximum likelihood method also requires kernel smoothed data, the problem in the constrained MLE does not arise.

Figure 6.1 shows the log-likelihood function of (μ,σ^2). As we expected, there are infinite spikes near $\sigma^2 = 0$ when μ is at each data point. Moreover, as Day (1969) indicated, we can see some spurious maximizers near $\sigma^2 = 0$ which make the likelihood very irregular. Figure 6.2 shows the smoothed log-likelihood function of (μ,σ^2) by applying normal kernel with $h = 0.025$. This contour plot looks much more regular in the sense that not only all infinite spikes are removed but also most spurious maximizers vanish.

As we mentioned earlier, the constrained MLE has a risk to exclude true σ^2 as well as infinite spikes and spurious maximizers if the constrained parameter space does not contain true parameter. Figure 6.3 and 6.4 show this does not happen under DSMLE.

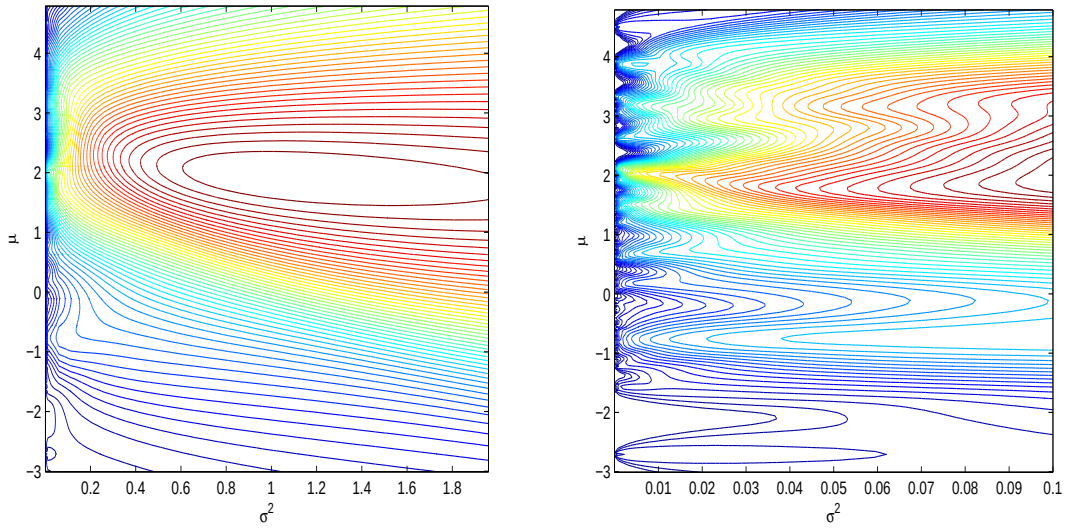


Fig. 6.1. Contour plot of log-likelihood function with true $\mu = 2$ and $\sigma^2 = 1$, shown in two different scales

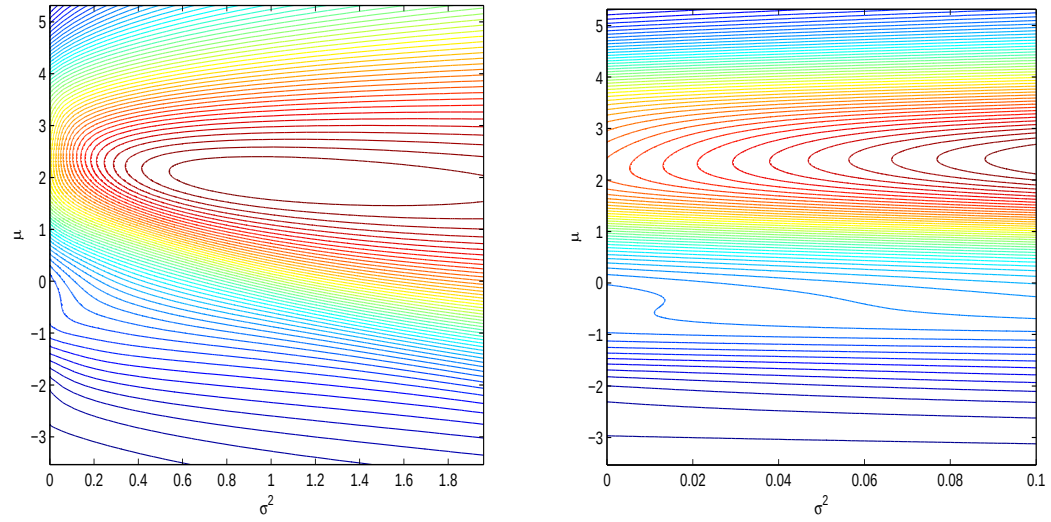


Fig. 6.2. Contour plot of smoothed log-likelihood function with tuning parameter $h = 0.025$ and true $\mu = 2$ and $\sigma^2 = 1$, shown in two different scales

Figure 6.3 shows the likelihood based on the simulated sample from (6.1) with $\mu = 2$ and $\sigma^2 = 0.1$. The maximum occurs around the true parameter $\mu = 2$ and $\sigma^2 = 0.1$. But this is not the actual maximizer because likelihood is not bounded. Figure 6.4 shows the doubly-smoothed log-likelihood using normal kernel with $h = 0.2$. Unlike the constrained MLE smoothed log-likelihood preserves its true maximizer removing the irregular likelihood part near $\sigma^2 = 0$ even though the tuning parameter h is greater than true σ^2.

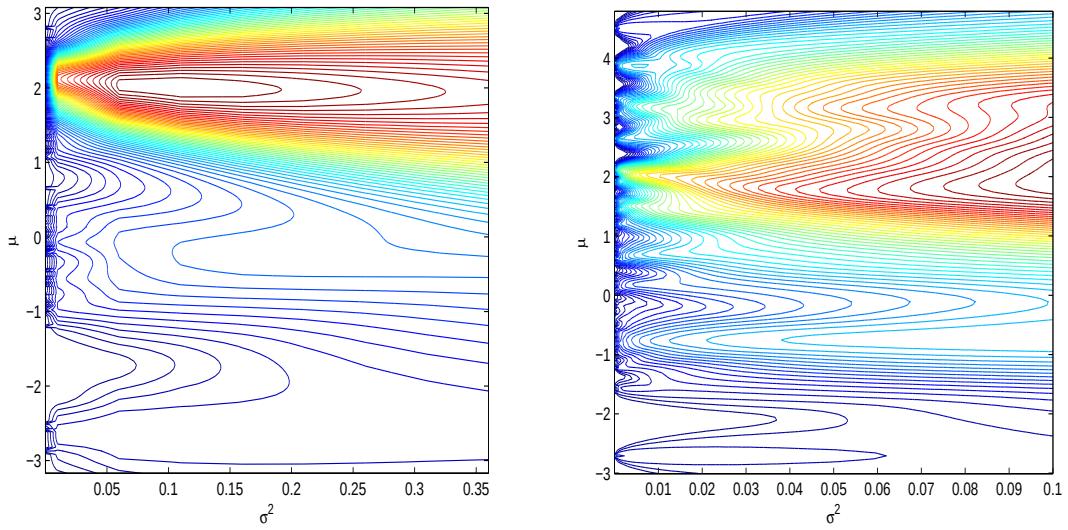


Fig. 6.3. Contour plot of log-likelihood function with true $\mu = 2$ and $\sigma^2 = 0.1$, shown in two different scales

Now in order to investigate the effect of the tuning parameter on the bias and variance, we did another simulation. For this simulation, $n = 100$, $n = 300$, and $n = 500$ samples are drawn from (6.1) with $\mu = 2$ and $\sigma^2 = 0.1$. For various tuning parameters, EM algorithm was applied. Table 6.1 and 6.2 show the bias with standard deviation of

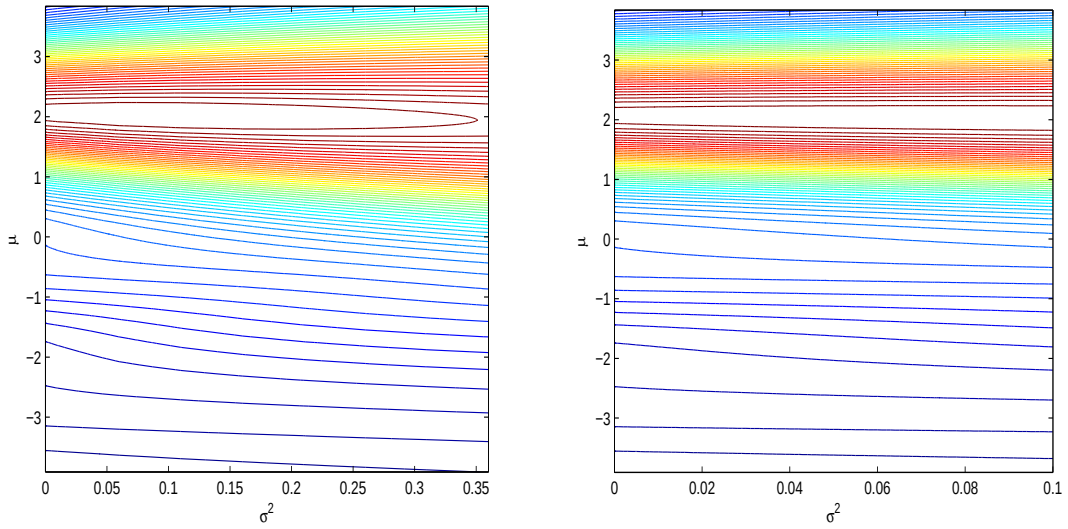


Fig. 6.4. Contour plot of smoothed log-likelihood function with tuning parameter $h = 0.2$ and true $\mu = 2$ and $\sigma^2 = 0.1$, shown in two different scales

$\hat{\mu}$ and $\hat{\sigma}^2$ based on 1000 replications. Note that for $h = 0$ we use true parameter values as a starting value of the algorithm, so the presented estimators are not actual MLE but they can be believed to be the consistent roots of the likelihood function.

In these tables, we can see that the bias and standard deviation decrease as n increases and this implies the consistency of DSMLE for each fixed h. We can also notice that the estimators are quite robust over different tuning parameters and that smoothing both model and data causes a very small amount of information loss.

For the complete investigation, we also increased the tuning parameter up to 10 though it is not shown in the tables. This is very large degree of smoothing because true σ^2 is 1. In this case, we experienced that the biases do not decrease as n increases. This phenomenon comes from the error in numerical integration because the magnitude of numerical error increases as the tuning parameter increases. Therefore we need to be

careful in numerical integration if a large tuning parameter is used. In this case, the number of Monte-Carlo sample should be increased for the simulation based integration or higher order approximation is required in the local Laplace approximation discussed in chapter 5.

Table 6.1. The bias(std) of $\hat{\mu}$ based on different tuning parameters

	Tuning						
	0	0.001	0.01	0.05	0.1	0.3	0.5
$n = 100$	-0.0158	-0.0158	-0.0160	-0.0166	-0.0173	-0.0183	-0.0179
	(0.2129)	(0.2129)	(0.2128)	(0.2124)	(0.2125)	(0.2147)	(0.2188)
$n = 300$	-0.0038	-0.0039	-0.0040	-0.0044	-0.0067	-0.0051	-0.0031
	(0.1180)	(0.1180)	(0.1180)	(0.1180)	(0.1151)	(0.1197)	(0.1221)
$n = 500$	-0.0026	-0.0026	-0.0026	-0.0028	-0.0029	-0.0020	0.0007
	(0.0895)	(0.0895)	(0.0895)	(0.0895)	(0.0897)	(0.0910)	(0.0928)

6.2 Conclusion and future work

Throughout Part I, we studied several examples that cause inconsistent ML estimators and a general modification tool for the ML method. The main reason of this inconsistency of MLE would be the inhomogeneity of measurement accuracy. In such cases, we showed kernel smoothing both data and model can resolve this undesirable feature of MLE for any statistical model. We also discussed the choice of a kernel and a tuning parameter as a rough guideline and potential computational strategies.

In some ways, our consistency results are quite strong. They show almost sure convergence for virtually any statistical model, completely without regularity conditions. However, we did pay a price in the weakness of the measure we used for convergence. For

Table 6.2. The bias(std) of $\hat{\sigma}^2$ based on different tuning parameters

	Tuning						
	0	0.001	0.01	0.05	0.1	0.3	0.5
$n = 100$	0.0037	0.0038	0.0041	0.0052	0.0066	0.0092	0.0091
	(0.3164)	(0.3165)	(0.3167)	(0.3179)	(0.3199)	(0.3272)	(0.3352)
$n = 300$	-0.0036	-0.0036	-0.0034	-0.0026	0.0012	-0.0011	-0.0046
	(0.1713)	(0.1713)	(0.1713)	(0.1717)	(0.1713)	(0.1756)	(0.1808)
$n = 500$	0.0002	0.0002	0.0003	0.0006	0.0009	-0.0006	-0.0057
	(0.1325)	(0.1325)	(0.1325)	(0.1327)	(0.1330)	(0.1354)	(0.1393)

example, the empirical CDF $\hat{F}_n$ converges to the true distribution in other strong metrics such as Kolmogorov-Smirnov measure but DSMLE can not guarantee the convergence to the true distribution under other strong metrics based on our proof. Hence we need more research for the convergence of DSMLE under other metrics.

We suggest two ways to carry out the estimation numerically: Monte-Carlo approximation and local Laplace approximation. The Monte-Carlo approximation method would be painful for a large data set and the local Laplace approximation can not be used for the model that includes missing values or requires EM optimization, like the mixture model. In the latter case, we believe MM algorithms (Hunter, 2003; Hunter and Lange, 2004) could resolve this problem; this will be another interesting future work. In addition, for the fast and accurate computation, further research is required.

In a simplified mixture example, we suggested DSMLE to solve unbounded likelihood but we did not mention the choice of a tuning parameter for the kernel. Although the simulation study shows DSMLE is quite robust to the choice of a tuning parameter, we should make the tuning parameter as small as we can because the purpose of this

modification is not to construct a new estimation procedure but to repair the ML estimation. Hence we should answer how small it should be. We can think that the tuning parameter that removes all the infinite spikes at $\sigma^2 = 0$ would be good enough. However, there are also several local modes near $\sigma^2 = 0$ generated from a set of close points as Day (1969) indicated. As we see in the simulation study, a large tuning parameter can also remove these spurious maxima, but we do not know how large it should be. We should study the nature of these spurious maxima in order to determine the optimal tuning parameter.

Part II

Measurement error problem

Chapter 7

Introduction

In many scientific studies, researchers are interested in finding the functional relationship between response variable Y and covariate X. The measurement error problem arises when the true covariate X is not observed and instead another covariate W is available such that W represents X with an error. In this case, our objective is still finding the functional relationship between the true response variable Y and the unobserved true covariate X, not Y and W.

For example, suppose that we want to study the effect of low-density lipoprotein (LDL) on the probability of heart disease, as previously discussed by Roeder et al. (1996). Because measuring LDL is very expensive, we measure total cholesterol instead of LDL. Now, LDL is a true predictor that we can not measure, and total cholesterol is another predictor which represents LDL.

As a second example, the NHANES-I Epidemiologic Study Cohort data set (Jones et al., 1987) is a cohort study originally consisting of 8,596 women who were interviewed about their nutrition habits and later examined for evidence of cancer. The response variable Y indicates the presence of breast cancer and the predictor variable X is "long-term" saturated fat intake. Because X is not observed, instead of observing X, 24-hour recall W was measured, that is, each participant's diet in the previous 24 hours was recalled and nutrition variables computed.

Under these settings, if we ignore measurement error and fit Y directly on W, the result could be very misleading. So the model that can take the measurement error into account is required. We call the model that specifies the relationship between the true covariate X and another observed variable W as a *measurement error model*. The measurement error model can be known from outside studies or fitted from a complete data set. In the next section, we briefly introduce some commonly used measurement error assumptions.

7.1 Measurement error models

The measurement error model explains the relationship between the true covariate X and the observable covariate W. For the simplest model, let us consider the relationship $W = X + U$ where U is called a measurement error. In this measurement error model, the measurement error U is added to the true covariate X so it is called the additive measurement error model. Similarly, the multiplicative measurement error model is $W = XU$. A more general model is a regression type measurement error model, $W = \alpha_0 + \alpha_1 X + U$.

In this thesis, we do not restrict attention to any specific form for the measurement error model. However, there are two important types of assumptions on the measurement error structure: classical versus Berkson type measurement error modeling and differential versus non-differential measurement error.

7.1.1 Berkson type and Classical measurement error model

The Berkson type measurement error model describes the conditional distribution of X given W. When W is fixed by design, it is more reasonable to model X given W. For example, if W is fixed amount of herbicide applied to a plant and X is the actual amount of herbicide absorbed by the plant. Then W is fixed by design and the true unknown X varies due to both the application and the plant absorption process. Hence it is natural to model $X|W$.

The classical measurement error model determines the conditional distribution of W given X. For example if we assume $W = X + U$, where U follows $N(0, \sigma_u^2)$, then the conditional distribution W given X is the normal distribution with mean X and variance σ_u^2. That is, we assume that W is a measure of X with error. In this thesis we assume this classical measurement error model.

7.1.2 Differential and Non-differential measurement error

Another important measurement error model choice is between differential and non-differential measurement error. We say that *non-differential measurement error* occurs when W and Y are independent given X. In other words, if the true predictor X is observed, then W does not give any additional information for prediction. Thus measurement error U and the other variables in the model are independent. In this case, we call W a *surrogate* for X. Otherwise, the model has *differential error*.

One advantage of the non-differential measurement assumption is that it enables one to estimate parameters without the true covariates X. Throughout this thesis, non-differential measurement error is assumed.

7.2 Attenuation in linear regression

In this section, we discuss the effect of measurement errors in the linear regression problem. To simplify our discussion, we consider a simple linear regression model in the next example.

EXAMPLE 7.1. In a simple linear regression model

$$Y = \beta_0^x + \beta_1^x X + \epsilon,$$

suppose that X is not observed instead we observe W which is related to X. We assume an additive measurement error model: $W = X + U$, where the random variable U has mean zero and variance σ_u^2 and U is independent to other variables. So we assume an additive non-differential classical measurement error model. If we ignore the measurement error U and regress Y directly on W, $Y = \beta_0^w + \beta_1^w W$, then the estimated slope coefficient is

$$\begin{aligned}
\hat{\beta}_1^w &= \frac{\sum w_i y_i - n\bar{w}\bar{y}}{\sum (w_i - \bar{w})^2} = \frac{\sum (x_i - \bar{x})^2}{\sum (w_i - \bar{w})^2} \times \frac{\sum w_i y_i - n\bar{w}\bar{y}}{\sum (x_i - \bar{x})^2} && (7.1)\\
&= \frac{\sum (x_i - \bar{x})^2}{\sum (w_i - \bar{w})^2} \times \frac{\sum (x_i + u_i) y_i - n(\bar{x} + \bar{u})\bar{y}}{\sum (x_i - \bar{x})^2} && (7.2)\\
&= \frac{\sum (x_i - \bar{x})^2}{\sum (w_i - \bar{w})^2} \times \frac{\sum x_i y_i - n\bar{x}\bar{y} + \sum u_i y_i - n\bar{u}\bar{y}}{\sum (x_i - \bar{x})^2} && (7.3)\\
&= \frac{\sigma_x^2}{\sigma_w^2} \left(\hat{\beta}_1^x + \frac{\sum u_i y_i - n\bar{u}\bar{y}}{\sum (x_i - \bar{x})^2} \right) && (7.4)
\end{aligned}$$

where $\hat{\beta}_1^x$ is the least square estimator of β_1^x based on $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$. The second term in the parenthesis of (7.4) strongly converges to zero becuase of the independence of Y and U. Thus (7.4) strongly converges to

$$\frac{\sigma_x^2}{\sigma_w^2} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\beta_1^x$$

Therefore the naive estimator $\hat{\beta}_1^w$ is not consistent for β_1^x unless the measurement error vanishes. More specifically, the absolute value of the estimated slope parameter attenuates to zero as long as $\sigma_u^2 > 0$.

This attenuation is a typical phenomenon in the linear regression problems and some nonlinear regression problems. However, this attenuation does not always happen and its amount of attenuation is hard to study in more complex situations such as the model with non-additive measurement error structure and non-linear models. In the next section, we categorize some known general methods and discuss their limitation.

7.3 Functional modeling versus structural modeling

In the measurement error problem, the essential concern is how we can model the unobserved variable X with the other variables. There are two representations of the unobserved covariate X which lead to two types of measurement error modeling (Carroll et al., 2006): *functional modeling* and *structural modeling*. In functional modeling, the unobserved X's are considered either as unknown parameters or as random variables from an unknown distribution. This assumption enables us to make no or minimal assumptions on the unobserved true covaraite X. If X_i's are considered as unknown

nuisance parameters, the likelihood is then

$$L(\theta, X_1, \cdots, X_n) = \prod_{i=1}^{n} f(Y_i, W_i | X_i, \theta) \tag{7.5}$$

Although this model makes minimal assumptions about the set of unobserved true covariates, maximizing likelihood with respect to θ and X_i's is often difficult because the number of parameters increases as the sample size increases. Moreover, it is well known that the resulting estimate for θ is not consistent (Neyman and Scott, 1948). For these reasons, with this model, the maximum likelihood estimation can not be directly used.

On the other hand, if X_i's are viewed as a random sample from an unknown distribution G_X, then it is natural to factor the joint density of all relevant variables as

$$f(w, y, x) = f(w|y, x) f(y|x) g(x). \tag{7.6}$$

where $g(x)$ is the density function induced by G_X. The joint density of observable variable (Y, W) becomes then

$$f(w, y) = \int f(w|y, x) f(y|x) g(x) dx. \tag{7.7}$$

The first term of the integrand in (7.7) is called the *measurement error model*, the second term is the *outcome model* or *response model* which is the model of interest, and the third term is the model for the population of the true covariate X. If we do not specify the distribution of X, this modeling makes no assumption for X.

In structural modeling, one assumes a specific parametric model for X. In this case, there is a general concern about the parametric specification of the unknown covariate X. Heckman and Singer (1984) used some parametric distributions and they found that estimated θ's can be quite different across different parametric distributional assumptions. That means the estimator is not robust to the parametric assumptions. Recently, Huang et al. (2006) devised a diagnostic tool to assess the effects of model missspecification so that we can choose an appropriate parametric model for the covariate distribution.

Because of these robustness problems in structural modeling, some authors recently have developed flexible models to alleviate a strong assumption on the covariate distribution; see, for example, Carroll et al. (1999); Richardson et al. (2002); Gustafson et al. (2002). These developments blur the distinction between functional modeling and structural modeling because even though these are distributional assumptions on the covariate distribution, the estimators' properties do not much depend heavily on those assumptions. Just the same, these models need the specification of a distribution for X and are determined by that choice of X.

If we want to completely remove this modeling issue for X, the best idea would be functional modeling assuming that X's are the random sample from an unspecified distribution. Roeder et al. (1996) combined a parametric model for the response model and a nonparametric model for $f(x)$ by leaving $f(x)$ unspecified. In a similar line, Schafer (2001) discussed the general applicability of this semiparametric model. Roeder et al. (1996)'s approach completely removes the misspecifination problems and its performance

is at least as good as other proposals. In the next section, we briefly introduce Roeder et al. (1996)'s approach, and then the limitation that becomes the motif of this part II.

7.4 Semi-parametric mixture approach

To study the relationship between the response variable Y and true but unobserved predictor X, we will assume a parametric conditional distribution for Y given X, denoted by $f_\beta(y|x)$, and a parametric conditional distribution for W given X, denoted by $f_\alpha(w|x)$.

Under the non-differential measurement error assumption, the conditional joint distribution of (Y, W), given X, is $f_\theta(y, w|x) = f_\beta(y|x)f_\alpha(w|x)$, where $\theta = (\alpha, \beta)$. Then the joint density of (Y, W) is

$$\begin{aligned} f_\theta(y, w) &= \int f_\alpha(w|y, x)f_\beta(y|x)g(x)dx \\ &= \int f_\alpha(w|y, x)f_\beta(y|x)dG_X(x) \end{aligned}$$

where $G_X(x)$ and $g_X(x)$ are the distribution function and density function for X. This can be also viewed as the mixture density with a mixing distribution $G_X(x)$. If we do not specify the distribution $G_X(x)$, this model involves a parametric component θ and a nonparametric component G_X, and hence is a semiparametric mixture density.

Now, let us consider more general situation. Suppose that we have an additional error free covariate Z measured without error as well as X measured with error. Thus, there are two types of covariates; one is measured with error and the other is measured without error. Now the joint density of (Y, W, Z) can be also written as a mixture form

using an indicator function:

$$
\begin{aligned}
f_\theta(y, w, z; G) &= \iint f_\theta(y, w, z | x = \xi, z = \zeta) dG_{X,Z}(\xi, \zeta) \\
&= \iint f_\beta(y | x = \xi, z = \zeta) f(w, z | x = \xi, z = \zeta) dG_{X,Z}(\xi, \zeta) \\
&= \iint f_\beta(y | x = \xi, z = \zeta) f_\alpha(w | x = \xi) I(z = \zeta) dG_{X,Z}(\xi, \zeta) \qquad (7.8)
\end{aligned}
$$

where $G_{X,Z}$ is the joint distribution function of (X, Z). Hence this model does not depend on specific modeling for the unobserved covariate distribution X.

Although this semiparametric mixture approach can completely remove the misspecification problem and show good performance when there is no error free covariate Z, as notified in Roeder et al. (1996), if there is an additional error free covariate Z and Z is a continuous random variable, the nonparametric ML estimate for the joint distribution of (X, Z) converges to $G_{W,Z}$, not $G_{X,Z}$. Curiously, this inconsistency problem does not happen if Z is also measured with error or is discrete.

Unfortunately, since Roeder et al. (1996) indicated the inconsistency problem of this model, to the best of our knowledge, there has been no literature which discussed this inconsistency problem. As a consequence, this semiparametric mixture approach cannot be directly applied to the practical usage even though their method is fully robust to the specification of the covariate distribution.

In the next chapter, we will dwell on this inconsistency problem more carefully and explain how we can settle down this long lasting problem.

Chapter 8

Doubly-smoothed maximum likelihood

8.1 Inconsistency of ML estimate

In this section, we follow up on Gaydos (1997)'s work to show the inconsistency of ML estimate in a simple model. Suppose that we want to estimate the joint distribution of (X, Z) where X is not observed directly. Now, suppose $\{(w_i, z_i); n = 1, \cdots, n\}$ is an IID sample of size n from the joint distribution of (W, Z), where W is a surrogate. The joint density of (W, Z) has then a simpler form than that of (7.8).

$$f_{W,Z}(w, z) = \iint f(w|x = \xi) I(z = \zeta) dG_{X,Z}(\xi, \zeta)$$

THEOREM 8.1. *If random variable (W, Z) has the joint density*

$$\iint f(w|x = \xi) I(z = \zeta) dG_{X,Z}(\xi, \zeta)$$

and $f(w|x) = f(w - x)$ is a unimodal density with mode 0. Then, the maximum likelihood estimate of $G_{X,Z}(x, z)$, the joint distribution of (X, Z), converges weakly to the distribution of (W, Z) with probability one.

Proof : Let $Q_{X|Z}$ and P_X be the distribution of X given Z and marginal distribution of Z. Also let $q_{X|Z}$ and p_X be the probability mass function of $Q_{X|Z}$ and P_X. Due to Lindsay(1983), we know that the nonparametric estimator of $G_{X,Z}$ is necessarily discrete.

So without loss of generality, we can think $q_{X|Z}$ and p_X are discrete probability mass function. Then, the likelihood can be written as

$$\prod_{i=1}^{n} P(W = w_i, Z = z_i) = \prod_{i=1}^{n} \sum_{x} f(w_i - x) q_{X|Z=z_i}(x) p_Z(z_i)$$

Now, this likelihood can be broken into two factors, $\prod_{i=1}^{n} \sum_x f(w_i - x) q_{X|Z=z_i}(x)$ and $\prod_{i=1}^{n} p_Z(z_i)$. For the second factor, we know the ML estimate for P_Z is simply the empirical distribution of $Z_1, \cdots, Z_n$. For the first factor, since Z is a continuous random variable, each z_i is distinct. For a fixed $Z = z_i$, there is only one term in the likelihood depending on $q_{X|Z=z_i}$. Therefore maximization problem is reduced to maximizing $\sum_x f(w_i - x) q_{X|Z=z_i}(x)$ for each i. From the unimodality with mode 0 assumption of $f(w|x = \xi)$, it is maximized when we put all mass at $x = w_i$. Therefore, conditional distribution of X given $Z = z_i$ has all mass at $x = w_i$. The resulting ML estimator of $G_{X,Z}$ is then just the empirical distribution of $(W_1, Z_1), \cdots, (W_n, Z_n)$. Consequently, ML estimate for $G_{X,Z}$ converges to the distribution of (W, Z), not (X, Z). □

This proof gives us several implications. First, we can see that the inconsistency of the ML procedure is caused by the inability of the ML method to pool the information available in the w_i across the z_i categories. Second, this inability of the ML method comes mainly from the discreteness of z_i observations even when they have a common continuous distribution. That is, z_i observation is too sharp to pool information in estimating $q_{X|Z}$. This suggest that, for a remedy of this phenomenon, we may need to use a smoothed density to avoid the sharpness of Z observation.

8.2 Doubly-smoothed maximum likelihood method

In part I, we explained the doubly-smoothed maximum likelihood method with its universal consistency property. In the following two subsections, we discuss its applicability for the measurement error problem.

8.2.1 Nonparametric estimation of covariate distribution

In this subsection, we will focus only on the estimation of the covariate distribution G and then we will describe full estimating procedure in the next subsection. Consider a bivariate random variable (X, Z) with an unknown distribution function $G(x, z)$. Suppose the random variable X can not be directly observed instead another random variable W is observed with measurement error while Z is measured without error. Assuming the measurement error distribution is completely known to be $f(w|x)$, the joint density of (X, W, Z) is $f(w|x, z)g(x, z) = f(w|x)g(x, z)$ under the non-differential measurement error assumption. The marginal density for (W, Z) is then

$$m_{_G}(w, z) = \int f(w|x)g(x, z)dx = \iint f(w|x)I(z' = z)dG(x, z') \tag{8.1}$$

and this is the model density.

Now, first we construct a smoothed model density by applying a kernel density $K_{_h}(\cdot, \cdot)$ with a tuning parameter h that controls smoothness of $K(\cdot, \cdot)$ to the model density $m_{_G}(w, z)$:

$$
\begin{aligned}
m^{*}_{G}(t_1, t_2) &= \iint K_h(t_1 - w, t_2 - z) m_G(w, z) dw dz \\
&= \iint K_h(t_1 - w, t_2 - z) \iint f(w|x) I(z^{'} = z) dG(x, z^{'}) dw dz \\
&= \iiint K_h(t_1 - w, t_2 - z) f(w|x) I(z^{'} = z) dw dz dG(x, z^{'}) \\
&= \iiint K_h(t_1 - w, t_2 - z^{'}) f(w|x) dw dG(x, z^{'})
\end{aligned}
$$

and we construct a smoothed kernel density by applying the same kernel density $K_h(\cdot, \cdot)$ with the same tuning parameter h to the given data:

$$
\hat{f}^{*}_{n}(t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} K_h(t_1 - w_i, t_2 - z_i)
$$

By this means, we construct two bounded and continuous densities; $m^{*}_{G}(t_1, t_2)$ and $f^{*}_{n}(t_1, t_2)$. The smoothed model density is based on the model $m_G(t_1, t_2)$ and the smoothed kernel density is based on the observed data $(w_1, z_1), \cdots, (w_n, z_n)$. As we indicated in section 8.1, the inconsistency of $\hat{G}$ is caused by the sharpness of z observations. Smoothing data could enable us to pool information across different z observations so that we can create a consistent sequence of $\hat{G}_n$'s.

However, smoothing cause bias if we smooth only data. Therefore we smooth not only the data but also the model with the same kernel and tuning parameter. By this means, we blur both the data and the model but in a parallel way and this blurring generate two regularized densities in the sense of continuous and bounded density.

From another point of view, the smoothed model density and smoothed kernel density can be viewed as two possible densities for the new random variable (T_1, T_2) which is generated by adding an error to the original random variable (W, Z). Thus for the random variable (T_1, T_2), there are two possible densities which are the model based density $m^*_G(t_1, t_2)$ and the data based density $\hat{f}^*_n(t_1, t_2)$.

With these two newly generated densities, the doubly-smoothed log-likelihood function is

$$\iint \ln \left[m^*(t_1, t_2; G) \right] \hat{f}^*_n(t_1, t_2) dt_1 dt_2. \tag{8.2}$$

The DSMLE of G is the maximizer of (8.2) and from the universal consistency property of DSMLE, it is consistent.

8.2.2 Estimating both parametric and nonparametric components

Now let us consider the full estimation with parametric components as well as nonparametric components. The joint density for (Y, X, W, Z) is $f(w|y, x, z) f_\theta(y|x, z) g(x, z) = f(w|x) f_\theta(y|x, z) g(x, z)$ under the non-differential measurement error assumption. Similarly to section 8.2.1, the joint marginal density of (Y, W, Z) is

$$m_{\theta,G}(y, w, z) = \int f(w|x) f_\theta(y|x, z) g(x, z) dx = \iint f(w|x) f_\theta(y|x, z) I(z' = z) dG(x, z'),$$

and the smoothed kernel density $\hat{f}^*_n(y, t_1, t_2)$ is

$$\hat{f}^*_n(y, t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} K_h(t_1 - w_i, t_2 - z_i) I(y = y_i) \tag{8.3}$$

where K_h is a kernel density and h is a tuning parameter for K_h. Note that we do not use smoothing for the Y variable. So the smoothed kernel density $\hat{f}_n^*$ is continuous in t_1 and t_2, but discrete in y. As we discussed in section 8.1, the failure of ML procedure mainly comes from the nonparametric estimation of the covariate distribution of X and Z. For this reason, smoothing Y variable doesn't appear necessary, although does provide us with a consistent estimator, using the consistency of DSMLE. Because smoothing a variable means adding blurring error to the original variable, it could cause efficiency loss and increase computational difficulty. Therefore we do not want to smooth all variables unless it is necessary.

Next, applying the same kernel to the model density, we construct a smoothed model density. The smoothed model density $m^*_{\theta,G}(y, t_1, t_2)$ is

$$m^*_{\theta,G}(y, t_1, t_2) = \iint K_h(t_1 - w, t_2 - z) m_{\theta,G}(y, w, z) dw dz \tag{8.4}$$

$$= \iiint K_h(t_1 - w, t_2 - z') f(w|x) f_\theta(y|x, z') dw dG(x, z') \tag{8.5}$$

where $G(x, z)$ is the distribution function for (X, Z). In this case, the Kullback-Leibler distance between two densities can not be established because the smoothed kernel density is the hybrid form of discrete and continuous variables. For this partial smoothing, we suggest the new objective function that is similar to the Kullback-Leibler distance:

$$Q(\hat{f}_n^*, m^*) = \sum_i \iint \ln \left(\frac{\hat{f}_n^*(t_1, t_2, y_i)}{m^*(t_1, t_2, y_i; \theta, G)} \right) \hat{f}_n^*(t_1, t_2, y_i) dt_1 dt_2. \tag{8.6}$$

Now minimizing Q with respect to (θ, G) can lead us to a hybrid version of DSMLE. Note that this is also equivalent to maximizing

$$\int \sum_i \ln \left(m^*(t_1, t_2, y_i; \theta, G) \right) \hat{f}_n^*(t_1, t_2, y_i) dt_1 dt_2 \tag{8.7}$$

The objective function Q is not exactly the Kullback-Leibler distance between m^* and $\hat{f}_n^*$. However, for each y_i the summand represents the Kullback-Leibler distance between $\hat{f}^*$ and m^*. On the other hand, (8.7) does not have the same form as (8.2), but if we rewrite (8.7) with a counting measure $P(y) = \frac{1}{n} \sum_i \delta_{y_i}(y)$, it can be also expressed as the form of (8.2).

Though it looks using Q is a natural extension of the Kullback-Leibler distance, the universal consistency in Part I cannot directly applied due to its hybrid form. In the next section, we show that the proposed method with partial smoothing could still give a consistent estimator but with some regularity conditions which are involved in unsmoothed variables.

8.3 Consistency under partial smoothing

Suppose $(X_1, Y_1), \cdots, (X_n, Y_n)$ is a random sample from an unknown probability measure M_τ on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ with the corresponding true probability density $m(x, y; \theta_\tau)$ with the model index θ_τ. If we smooth only X variable, the smoothed model density is

$$m^*(t, y; M_\theta) = \int K_h(t, x) m(x, y; \theta) dx$$

and the smoothed kernel density is

$$\hat{f}^*_n(t,y) = \frac{1}{n}\sum_i K_h(t,x_i)\delta_{y_i}(y) = \int K_h(t,x)\delta_{y_i}(y)d\hat{F}_n(x,y)$$

where $\delta_{y_i}(y)$ is the Dirac measure with unit mass at y_i and $\hat{F}_n(x,y)$ is the empirical distribution based on sample. Under this partial smoothing, the objective function is

$$Q(\hat{f}^*_n, m^*) = \sum_i \int \ln\left(\frac{\hat{f}^*_n(t,y_i)}{m^*(t,y_i;M_\theta)}\right)\hat{f}^*_n(t,y_i)dt,$$

and estimating procedure can be done by minimizing $Q(\hat{f}^*_n, m^*)$ with respect to θ or equivalently maximizing

$$\sum_i \int \ln\left(m^*(t,y_i;M_\theta)\right)\hat{f}^*_n(t,y_i)dt.$$

For the proof of consistency, we split the proof into two cases: (1) Y is discrete with finite support (2) Y is continuous. In this section, we prove the model consistency like Theorem 3.1 for each case and then Corollary 3.1 can be applied to finish the consistency of the hybrid version of DSMLE. Throughout the following section, we will assume the two kernel assumptions, (K1) and (K2) from section 3.1.

8.3.1 Y is discrete with finite support

In this section we consider the case where the unsmoothed variable Y is discrete with finite support points $y^{(1)}, \cdots, y^{(s)}$. The consistency proof is essentially same as that of section 3.2. Moreover, we still do not need any regularity condition for a given

model. However, we need some technical lemmas that enable us to interchange limits and integrals in the main theorem.

LEMMA 8.1. *If* $\hat{M}_n \xrightarrow{v} M_0$, *then for each* t

$$\frac{1}{n}\sum_{i=1}^{n}\left[\ln m^*(t,y_i;\hat{M}_n)\right]^- K_h(t,x_i) \xrightarrow{a.s.} \iint \left[\ln m^*(t,y;M_0)\right]^- K_h(t,x)dM_\tau$$

Proof :

$$\begin{aligned}
&\lim_n \sum_{j=1}^{s}\frac{1}{n}\sum_{i=1}^{n} I\left(y_i = y^{(j)}\right)\left[\ln m^*(t,y_i;\hat{M}_n)\right]^- K(t,x_i) \\
&= \sum_{j=1}^{s}\lim_n \left[\ln m^*(t,y^{(j)};\hat{M}_n)\right]^- \frac{1}{n}\sum_{i=1}^{n} I\left(y_i = y^{(j)}\right) K(t,x_i) \\
&= \sum_{j=1}^{s}\lim_n \left[\ln m^*(t,y^{(j)};\hat{M}_n)\right]^- \lim_n \frac{1}{n}\sum_{i=1}^{n} I\left(y_i = y^{(j)}\right) \lim_n \frac{\sum_{i=1}^{n} I\left(y_i = y^{(j)}\right) K(t,x_i)}{\sum_{i=1}^{n} I\left(y_i = y^{(j)}\right)}
\end{aligned} \tag{8.8}$$

In the last equality, $\sum_{i=1}^{n} I\left(y_i = y^{(j)}\right)$ could be zero for a finite n. However, with probability tending to one, we can ignore this situation because for each j $\{y^{(j)} : j = 1, \cdots, s\}$ is the set of support. Now, the first limit in (8.8) converges to $\left[\ln m^*(t,y^{(j)};M_0)\right]^-$ because

$$\begin{aligned}
m^*(t,y^{(j)};M_n) &= \int K_h(t,y^{(j)})dM_n \\
&\to \int K_h(t,y^{(j)})dM_0 = m^*(t,y^{(j)};M_0).
\end{aligned}$$

from the kernel assumption (K1) and $\hat{M}_n \xrightarrow{v} M_0$. The second and third limits almost surely converges to $P\left(Y = y^{(j)}\right)$ and $\int K_h(t,x)dM_\tau(x|y=y^{(j)})$ from SLLN where $M_\tau(x|y)$ is the true conditional distribution of X given Y. Therefore the last expression in (8.8) almost surely converges to

$$\begin{aligned}
&\sum_{j=1}^{s}\left[\ln m^*(t,y^{(j)};M_0)\right]^- P\left(Y=y^{(j)}\right)\int K_h(t,x)dM_\tau\left(x|y=y^{(j)}\right)\\
&=\iint\left[\ln m^*(t,y;M_0)\right]^- K_h(t,x)dM_\tau(x|y)dM_\tau(y)\\
&=\iint\left[\ln m^*(t,y;M_0)\right]^- K_h(t,x)dM_\tau(x,y)
\end{aligned}$$

where $M_\tau(y)$ is the true marginal distribution of Y. □

LEMMA 8.2. *If* $\hat{M}_n \xrightarrow{v} M_0$,

$$\int \frac{1}{n}\sum_{i=1}^{n}\left[\ln m^*(t,y_i;\hat{M}_n)\right]^+ K_h(t,x_i)dt \xrightarrow{a.s.} \iiint\left[\ln m^*(t,y;M_0)\right]^+ K_h(t,x)dM_\tau(x,y)dt \tag{8.9}$$

Proof :

$$\begin{aligned}
&\lim_n \int \frac{1}{n}\sum_{i=1}^{n}\left[\ln m^*(t,y_i;\hat{M}_n)\right]^+ K_h(t,x_i)dt\\
&=\sum_{j=1}^{s}\lim_n \int \frac{1}{n}\sum_{i=1}^{n}\left[\ln m^*(t,y_i;\hat{M}_n)\right]^+ K_h(t,x_i)I\left(y_i=y^{(j)}\right)dt\\
&=\sum_{j=1}^{s}\lim_n \frac{\sum_{k=1}^{n} I\left(y_k=y^{(j)}\right)}{n}\lim_n \int \sum_{i=1}^{n}\left[\ln m^*(t,y^{(j)};\hat{M}_n)\right]^+ \frac{K_h(t,x_i)I\left(y_i=y^{(j)}\right)}{\sum_{k=1}^{n} I\left(y_k=y^{(j)}\right)}dt
\end{aligned} \tag{8.10}$$

To interchange limit and integral sign in (8.10), we will use the extended version of the dominated convergent theorem. The boundedness of the kernel K_h from (K1) implies that $m^*(t, y; \hat{M}_n)$ is bounded above for a fixed h and y, and so there exists positive number $U(h, y)$ such that $\left[\ln m^*(t, y; \hat{M}_n)\right]^+ < U(h, y)$. Therefore the integrand is bounded by

$$U\left(h, y^{(j)}\right) \frac{\sum_{i=1}^n K_h(t, x_i) I\left(y_i = y^{(j)}\right)}{\sum_{i=1}^n I\left(y_i = y^{(j)}\right)}. \tag{8.11}$$

So we use (8.11) as a dominating sequence. Moreover, we can easily check

$$U\left(h, y^{(j)}\right) \frac{\sum_{i=1}^n K_h(t, x_i) I\left(y_i = y^{(j)}\right)}{\sum_{i=1}^n I\left(y_i = y^{(j)}\right)} \xrightarrow{a.s.} U\left(h, y^{(j)}\right) \int K_h(t, x) dM_\tau(x | y = y^{(j)})$$

and

$$\int U\left(h, y^{(j)}\right) \int K_h(t, x) dM_\tau\left(x | y = y^{(j)}\right) dt = U\left(h, y^{(j)}\right) < \infty.$$

Now, applying the extended version of the dominated convergence theorem, the order of the second limit and integral sign in (8.10) can be interchanged. With the similar argument to the previous lemma, (8.10) almost surely converges to

$$\sum_{j=1}^s P\left(y = y^{(j)}\right) \int \left[\ln m^*(t, y^{(j)}; M_0)\right]^+ K_h(t, x) dM_\tau(x | y = y^{(j)}) dt$$

$$= \int \left[\ln m^*(t, y; M_0)\right]^+ K_h(t, x) dM_\tau(x, y) dt$$

□

LEMMA 8.3.

$$\int \frac{1}{n}\sum_{i=1}^{n}\left[\ln m^{*}(t,y_i;M_\tau)\right]^{-} K_h(t,x_i)dt \xrightarrow{a.s.} \iiint \left[\ln m^{*}(t,y;M_\tau)\right]^{-} K_h(t,x)dM_\tau dt \tag{8.12}$$

Proof : If $\iiint \left[\ln m^{*}(t,y;M_\tau)\right]^{-} K_h(t,x)dM_\tau dt = \infty$, equation (8.12) holds from (Chung, 1974, Theorem 5.4.2). Otherwise, (8.12) holds from Fubini theorem and SLLN. □

THEOREM 8.1. *Let $\mathcal{M} = \{M_\theta(x,y)\}$ be a class of model distributions indexed by θ. Suppose that $\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$ is a random sample from the true distribution $M(x,y;\theta_\tau) \in \mathcal{M}$. If Y has a finite number of support points, then the minimizer $\hat{M}_n$ of the objective function Q weakly converges to M_τ on a set of probability one.*

Proof : Since $\hat{M}_n = \hat{M}_n^{\omega}$ is a sequence of distributions on $\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$, for any subsequence $\{m\} \subset \{n\}$ by Helly's selection principle we can always select a further subsequence $\{k\} \subset \{m\}$ such that $\hat{M}_k$ vaguely convergent to subprobability measure M_0. If we can show that $M_0 = M_\tau$, then we are done by the method of subsequences (Chung, 1974, Theorem 4.3.4). We can then justify the following sequence of inequalities.

$$\begin{aligned} 0 &\geq \liminf_k \sum_i \int \ln\left(\frac{m^{*}(t,y_i;M_{\theta_\tau})}{\hat{m}_n^{*}(t,y_i;M_{\theta_k})}\right) f^{*}(t,y_i)dt \\ &= \liminf_k \int \frac{1}{k}\sum_{i=1}^{k} \ln\left(\frac{m^{*}(t,y_i;M_\tau)}{\hat{m}(t,y_i;\hat{M}_k)}\right) K_h(t,x_i)dt \end{aligned}$$

$$\geq \liminf_k \int \frac{1}{k} \sum_{i=1}^{k} \left[\ln m^*(t, y_i; M_\tau)\right]^+ K_h(t, x_i) dt$$

$$+ \liminf_k \int \frac{1}{k} \sum_{i=1}^{k} \left[\ln m^*(t, y_i; \hat{M}_k)\right]^- K_h(t, x_i) dt$$

$$- \limsup_k \int \frac{1}{k} \sum_{i=1}^{k} \left[\ln m^*(t, y_i; M_\tau)\right]^- K_h(t, x_i) dt$$

$$- \limsup_k \int \frac{1}{k} \sum_{i=1}^{k} \left[\ln m^*(t, y_i; \hat{M}_k)\right]^+ K_h(t, x_i) dt$$

$$\geq \liminf_k \frac{1}{k} \sum_{i=1}^{k} \int \left[\ln m^*(t, y_i; M_\tau)\right]^+ K_h(t, x_i) dt$$

$$+ \int \liminf_k \frac{1}{k} \sum_{i=1}^{k} \left[\ln m^*(t, y_i; \hat{M}_k)\right]^- K_h(t, x_i) dt$$

$$- \limsup_k \int \frac{1}{k} \sum_{i=1}^{k} \left[\ln m^*(t, y_i; M_\tau)\right]^- K_h(t, x_i) dt$$

$$- \limsup_k \int \frac{1}{k} \sum_{i=1}^{k} \left[\ln m^*(t, y_i; \hat{M}_k)\right]^+ K_h(t, x_i) dt$$

From SLLN, the first expression in the last expression converges to

$$\iiint \ln \left[m^*(t, y; M_\tau)\right]^+ K_h(t, x) dt dM_\tau.$$

From the Fubini's theorem for the nonnegative function and the previous lemmas, the last expression converges to

$$\iiint \left(\left[\ln m^*(t,y;M_\tau) \right]^+ + \left[\ln m^*(t,y;M_0) \right]^- \right) K_h(t,x) dt dM_\tau$$

$$- \iiint \left[\ln m^*(t,y;M_\tau) \right]^- K_h(t,x) dt dM_\tau - \iiint \left[\ln m^*(t,y;M_0) \right]^+ K_h(t,x) dt dM_\tau$$

$$= \iiint \ln \left(\frac{m^*(t,y;M_\tau)}{\hat{m}(t,y;\hat{M}_n)} \right) K_h(t,x) dt dM(x,y;\theta_\tau)$$

$$= \iint \ln \left(\frac{m^*(t,y;M_\tau)}{\hat{m}(t,y;\hat{M}_n)} \right) \int K_h(t,x) m(x,y;\theta_\tau) dx dt dy$$

$$= \iint \ln \left(\frac{m^*(t,y;M_\tau)}{\hat{m}(t,y;\hat{M}_n)} \right) m^*(t,y;M_\tau) dt dy \geq 0$$

The same argument in Theorem 3.1 finishes the proof. □

8.3.2 Y is continuous

In this section, we assume the unsmoothed variable Y is continuous. In this case, we need following regularity conditions for the model $m(x,y;\theta)$:

(R1) $m(x,y;\theta)$ is continuous in (y,θ) for each x with a suitable metric.

(R2) There exists a measurable function $m(x)$ such that $m(x,y;\theta) \leq m(x)$ where $\int m(x)dx < \infty$.

The first condition is similar to the model continuity condition (M2) in section 3.3 and it is easy to verify this condition for a given model. The second condition is slightly stronger than the usual regularity conditions in the consistency study. However, this can be also easily verified in many models of interest. As a simple example, if the conditional density Y given X is normal and the marginal density for X is $f(x)$, then the joint density for

(X, Y) is

$$f(x,y) = f(y|x)f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\theta x)^2}{2\sigma^2}\right\} f(x) \leq \frac{1}{\sqrt{2\pi\sigma^2}} f(x)$$

Hence $f(x,y)$ satisfies (R2).

LEMMA 8.4. *Assuming (R1) and (R2), if* $y_n \to y$ *and* $\theta_n \to \theta_0$, *then*

$$\liminf_n \iint \left(\left[\ln m^*(t,y;M_\tau)\right]^+ + \left[\ln m^*(t,y;\hat{M}_n)\right]^- \right) K_h(t,x) d\hat{F}_n(x,y) \tag{8.13}$$

$$\geq \iint \left(\left[\ln m^*(t,y;M_\tau)\right]^+ + \left[\ln m^*(t,y;M_0)\right]^- \right) K_h(t,x) dM(x,y;\theta_\tau) \tag{8.14}$$

Proof : From the given two conditions and DCT, we can easily prove for each t,

$$m^*(t,y_n;M_n) = \int K_h(t,x) m(x,y_n;\theta_n) dx \to \int K_h(t,x) m(x,y,\theta_0) dx = m^*(t,y;M_0). \tag{8.15}$$

Using Skorohod construction theorem, we can build a probability space $(\Omega, \mathfrak{B}, \mathcal{P})$ on which $(X_n, Y_n) \to (X, Y)$ with probability one. Thus, (8.13) can be expressed as

$$\liminf_n \iint \left(\left[\ln m^*(t,Y_n;M_\tau)\right]^+ + \left[\ln m^*(t,Y_n;\hat{M}_n)\right]^- \right) K_h(t,X_n) dP \tag{8.16}$$

$$\geq \iint \liminf_n \left(\left[\ln m^*(t,Y_n;M_\tau)\right]^+ + \left[\ln m^*(t,Y_n;\hat{M}_n)\right]^- \right) K_h(t,X_n) dP \tag{8.17}$$

$$= \iint \left(\left[\ln m^*(t,Y;M_\tau)\right]^+ + \left[\ln m^*(t,Y;M_0)\right]^- \right) K_h(t,X) dP \tag{8.18}$$

$$= \iint \left(\left[\ln m^*(t,y;M_\tau)\right]^+ + \left[\ln m^*(t,y;M_0)\right]^- \right) K_h(t,x) dM(x,y;\theta_\tau) \tag{8.19}$$

We have the inequality in (8.17) from Fatou's Lemma. From (R1) and (K1), the equation (8.18) holds. □

LEMMA 8.5.

$$\limsup_n \iiint \left[\ln m^*(t,y;M_\tau)\right]^- K_h(t,x)dtd\hat{F}(x,y)$$
$$= \iiint \left[\ln m^*(t,y;M_\tau)\right]^- K_h(t,x)dtdM(x,y;\theta_\tau) \quad a.s.$$

Proof : Similar to the proof of Lemma 8.3. □

LEMMA 8.6. *Assuming (R1) and (R2), if* $\theta_n \longrightarrow \theta_0$,

$$\limsup_n \iiint \left[\ln m^*(t,y;\hat{M}_n)\right]^+ K_h(t,x)dtd\hat{F}_n(x,y)$$
$$= \iiint \left[\ln m^*(t,y;M_0)\right]^+ K_h(t,x)dtdM(x,y;\theta_\tau) \quad a.s.$$

Proof : Using Skorohod construction theorem again like previous lemma,

$$\lim_n \iiint \left[\ln m^*(t,y;\hat{M}_n)\right]^+ K_h(t,x)dtd\hat{F}_n(x,y)$$
$$= \lim_n \iiint \left[\ln m^*(t,Y_n;\hat{M}_n)\right]^+ K_h(t,X_n)dtdP \tag{8.20}$$

Now, we use dominated convergence theorem. From the kernel assumption, $K_h(t,X_n) \leq U_h$ where U_h is a positive real number. Because $m^*(t,Y;\hat{M}_n) = \int K_h(t,x)m(x,y;\theta)dt \leq \int U_h m(x)dx < \infty$, $\left[\ln m^*(t,Y_n;\hat{M}_n)\right]^+$ is bounded above by $\left[\ln \int U_h m(x)dx\right]^+ = R$. So integrand in (8.20) is bounded by $RK_h(t,X_n)$ and $\int RK_h(t,X_n)dt = R < \infty$. Using the

extended version of DCT, (8.20) converges to

$$\begin{aligned}
&\iiint \lim_n \left[\ln m^*(t, Y_n; \hat{M}_n)\right]^+ K_h(t, X_n) dt dP \\
&= \iiint \left[\ln m^*(t, Y; \hat{M}_0)\right]^+ K_h(t, X) dt dP \\
&= \iiint \left[\ln m^*(t, y; \hat{M}_0)\right]^+ K_h(t, x) dt dM(x, y; \theta_\tau)
\end{aligned}$$

□

THEOREM 8.2. *Let $\mathcal{M} = \{M_\theta(x, y)\}$ be a class of model distributions indexed by θ satisfying (R1) and (R2). Suppose that $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ is a random sample from true distribution $M_\tau \in \mathcal{M}$ and Y is continuous. Then the minimizer $\hat{M}_n$ of the objective function Q weakly converges to M_τ on a set of probability one.*

Proof : Applying the same method of subsequences used in Theorem 3.1 and 8.1, we can then justify the following inequalities:

$$\begin{aligned}
0 &\geq \liminf_k \frac{1}{k} \sum_{i=1}^{k} \int \ln \left(\frac{m^*(t, y_i; M_{\theta_\tau})}{\hat{m}^*(t, y_i; M_{\theta_k})} \right) \hat{f}_k^*(t, y_i) dt \\
&= \liminf_k \frac{1}{k} \sum_{i=1}^{k} \int \ln \left(\frac{m^*(t, y_i; M_\tau)}{\hat{m}(t, y_i; \hat{M}_k)} \right) K_h(t, x_i) dt \\
&= \liminf_k \iiint \ln \left(\frac{m^*(t, y; M_\tau)}{\hat{m}(t, y; \hat{M}_k)} \right) K_h(t, x) dt d\hat{F}_k(x, y) \\
&\geq \liminf_k \iiint \left(\left[\ln m^*(t, y; M_\tau)\right]^+ + \left[\ln m^*(t, y; \hat{M}_k)\right]^- \right) K_h(t, x) dt d\hat{F}_k(x, y) \\
&\quad - \limsup_k \iiint \left[\ln m^*(t, y; M_\tau)\right]^- K_h(t, x) dt d\hat{F}_k(x, y) \\
&\quad - \limsup_k \iiint \left[\ln m^*(t, y; \hat{M}_k)\right]^+ K_h(t, x) dt d\hat{F}_k(x, y) \qquad (8.21)
\end{aligned}$$

Using Lemma 8.4, 8.5, and 8.6 and the same argument of Theorem 8.1, the last expression in (8.21) converges to

$$\iint \ln\left(\frac{m^*(t,y;M_\tau)}{\hat{m}(t,y;\hat{M}_n)}\right) m^*(t,y;M_\tau)dtdy \geq 0$$

The same argument in Theorem 3.1 will apply to finish the proof. □

Chapter 9

Computation

Finding the maximizer in equation (8.7) involves some computational difficulties because we need to maximize (8.7) with respect to the parametric components and the nonparametric component simultaneously. Moreover, due to kernel smoothing, the numerical integration is also required. This chapter builds on the algorithm discussed in chapter 5. We will first describe how to estimate the covariate distribution and then the complete estimation for both the parameters and the covariate distribution.

9.1 Estimation of covariate distribution

Let us first consider the nonparametric estimation of the covariate distribution G in section 8.2.1. The doubly-smoothed log-likelihood that we maximize can be written as

$$\iint \ln\left[m^*(t_1,t_2;G)\right]\hat{f}_n^*(t_1,t_2)dt_1dt_2 = \frac{1}{n}\iint \ln\left[m^*(t_1,t_2;G)\right]K_h(t_1-w_i,t_2-z_i)dt_1dt_2. \tag{9.1}$$

For numerical integration, we suggest the Monte-Carlo integration. By this means, we can use some well-known mixture algorithms to estimate the covariate distribution G nonparametrically.

To apply the Monte-Carlo integration, for each i, generate b Monte-Carlo sample $\left(T_{1i1}, T_{2i1}\right), \cdots, \left(T_{1ib}, T_{2ib}\right)$ from $K_h(t_1 - w_i, t_2 - z_i)$. Then (9.1) can be approximated by

$$\frac{1}{nb} \sum_{i=1}^{n} \sum_{j=1}^{b} \ln \left[m^*(t_{1ij}, t_{2ij}; G) \right]. \tag{9.2}$$

Because the smoothed model density $m^*(t_1, t_2; G) = \iint K_h(t_1 - w, t_2 - z')f(w|x)I(z' = z)dG(x, z')$ can be viewed as a mixture density with the atomic density $K_h(t_1 - x, t_2 - z)f(w|x)I(z' = z)$ and mixing distribution G, if we think the simulated Monte-Carlo samples $\left(T_{1i1}, T_{2i1}\right), \cdots, \left(T_{1ib}, T_{2ib}\right)$ as random samples from the density $m^*(t_1, t_2; G)$, the initial maximization problem is reduced to estimating the nonparametric mixing or latent distribution G under the mixture density $m^*(t_1, t_2; G)$.

When we estimate the nonparametric distribution of G, we need to choose a set of support points for G. However, because it is not possible to choose support points without extra information, predetermined grid points are required. If the support points are not fixed, the EM algorithm can be also used to estimate both the support points and corresponding weights by determining the number of support points beforehand. However its computational complexity will greatly increase.

As an alternative way to find the nonparametric mixing distribution G, some gradient based methods will be useful such as *VEM* (Bohning, 1985) and *ISDM* (Lesperance and Kalbfleisch, 1992), as discussed in chapter 5.

In our case, if we only have covariates measured with error, either the EM algorithm or a gradient based algorithm can be applied. If there is an additional error free

covariate Z, the EM algorithm will be more useful than gradient based algorithm because we already have the information for the support points of Z. That is, all distinct Z observations would be good candidates for the the support points for Z. If EM-algorithm is used, we need to only estimate support points for X while the support points of Z are fixed at each Z observation. In this case, the weight for each pair of support point for (X, Z) will be fixed with $1/n$.

There are two advantages of this estimating scheme. First, because the program does not have to determine the support points for Z and corresponding weights, the computational burden will be greatly lessened. Second, because we know the NPMLE of the marginal distribution of Z is the empirical distribution based on $Z_1, \cdots, Z_n$, we also hope the marginal distribution of Z induced by the resultant estimator would be the same empirical distribution. This can be achieved without any computational effort because the marginal distribution is always fixed to be the empirical distribution.

However, despite of this computational convenience, we are not sure if the resultant estimator is a real minimizer of the objective function Q. This might be true when X and Z are independent. But if they are not independent, the information from the estimated distribution of X could be helpful to estimate the nature of Z variable. Thus if we fix the support points for Z without considering the information from X, there might be some information loss. We propose to investigate this further in the future.

9.2 Combining algorithms for the parametric and nonparametric components of the model

Now let us consider the full model in section 8.2.2. The basic procedure is same as that of section 9.1 but we have to estimate the parameter θ as well as G. So we need two stage iterative process to estimate θ and G simultaneously. For the estimation of (θ, G), we suggest following algorithm.

1. For the initial nonparametric estimator for $G(x, z)$, choose the empirical distribution based on $\{(w_i, z_i) : i = 1, \cdots, n\}$ and for the initial estimator for θ, use the maximum likelihood estimator ignoring measurement error.

2. For fixed $\hat{\theta}^{Current}$, estimate $\hat{G}(x, z)$ using the algorithm in section 9.1.

3. For fixed $\hat{G}(x, z)$, maximize the approximated objective function (9.2) over θ.

4. Iterate step 2 and 3 until the approximated objective function converges under predetermined stopping rule.

9.3 The choice of tuning parameter h

For the choice of the tuning parameter h, we need to think about the initial objective of this study. Because kernel smoothing was used to repair the failure of the MLE, a reasonable direction would be to choose the tuning parameter that assures one to have an improved estimator. Moreover, because the main failure of the MLE is the wrongly estimated covariate distribution and the estimated marginal distribution of Z is always correct based on the proposed algorithm, it would be desirable to choose the

tuning parameter that enables the estimation to find the correct marginal distribution of X.

To asses the validity of the estimated marginal distribution of X, we need to check if the estimated marginal distribution is close to the true distribution of X. Of course we don't know the true distribution because X is not observed. However we have the information that characterize the relationship between X and W. That is, although we do not know the true marginal distribution of X, it is assumed that the measurement error distribution $W|X$ is known and W is observed. So we can extract some information for the true distribution of X from the known measurement error distribution and observed W_i's. If $\hat{G}$ is close to true G, the information from the known measurement error distribution and the estimated marginal distribution of X must agree. A simple way to extract information could be the moment based information such as the expectation and variance of X. Under this argument, our suggestion is choosing h which makes those information agree.

For example, if an additive normal measurement error is assumed, that is $W = X+U$ where $U \sim N(0, \sigma^2)$ with known σ^2, then $Var(W) = Var(X)+\sigma^2$. So the variance of X can be estimated by $\widehat{Var}_1(X) = \widehat{Var}(W)-\sigma^2$ where $\widehat{Var}(W)$ is the sample variance from $W_1, \cdots, W_n$. Now from the estimated $\hat{G}$, another variance estimator for X can be calculated by $\widehat{Var}_2(X) = Var_{\hat{G}}(X)$. Therefore the reasonable choice for h would be the one that minimize a certain distance between $\widehat{Var}_1(X)$ and $\widehat{Var}_2(X)$. For instance, if X is scalar valued, choose h that minimize

$$\left| \widehat{Var}_1(X) - \widehat{Var}_2(X) \right|. \tag{9.3}$$

To find the tuning parameter h that minimizes the distance between $\widehat{Var_1}(X)$ and $\widehat{Var_2}(X)$, we may need to use a grid search within a predetermined grid. If the range of h is very wide or fine grid points are used, the burden of computation will increase. Even for the small set of grid points for h, computing time could be long because the distance can be calculated only after $\hat{G}_n$ is estimated with fixed h on the grid. In this case, we can narrow the possible range of h using sDOF as discussed in section 4.2. If this range of sDOF too wide, linear or quadratic interpolation method can be used. From our experience, $\widehat{Var_1}(X) - \widehat{Var_2}(X)$ is approximately linear or quadratic in h within the appropriate range of h based on sDOF. Hence the linear or quadratic interpolation can be used to find h that makes (9.3) close to zero. We will show this in the next chapter under some simulated sample.

Chapter 10

Simulation study

This chapter builds on the algorithms discussed in chapter 9 and applies them to a simple measurement error problem. In this simulation study, we consider a linear regression model with an additive measurement error though the proposed method can be easily extended to any measurement error model or any nonlinear regression model.

10.1 Estimation of non-parametric component

For the first simulation study, we focus only on estimating the covariate distribution G in section 8.2.1. For the simulation experiment, we generate (X, W, Z) using following scheme and then we assume that W and Z are only observed, but X is not.

1. Generate u_i, z_i independently from $N(0, \sigma_u^2), N(0, 1)$, respectively.
2. Generate x_i such that $x_i = \begin{cases} -2 & \text{with probability } \frac{e^{z_i}}{1+e^{z_i}} \\ 2 & \text{with probability } \frac{1}{1+e^{z_i}} \end{cases}$
3. Generate $w_i = x_i + u_i$

With this simulation design, the unobserved true covariate X has only two support points, -2 and 2. In this simulation, it is important to check if the estimated marginal covariate distribution of X is close to the estimated one. Because X has only two support points, we can clearly compare the true marginal cumulative distribution of X

with estimated one graphically. Assuming that the distribution of U is completely known to be $N(0, 0.2)$ by setting $\sigma_u^2 = 0.2$, the observed likelihood is then

$$\prod_i f_{W,Z}(w_i, z_i) = \prod_i \int f_{W|X}(w_i|x) I(z_i = \zeta) dG_{X,Z}(\xi, \zeta)$$

In order to apply the proposed methodology, the bivariate normal kernel

$$MVN\left\{ \begin{pmatrix} x \\ z \end{pmatrix}, h\hat{Cov} \begin{pmatrix} W \\ Z \end{pmatrix} \right\}$$

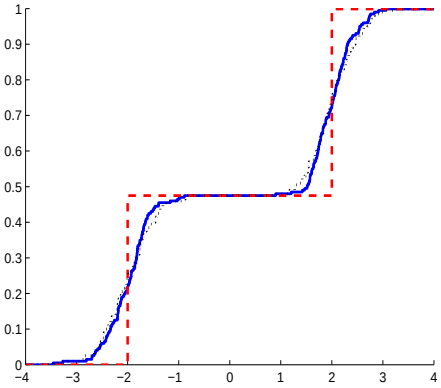
is used for both the model and the data because it can provide us with closed form of the smoothed model density, as discussed in section 4.1. Then the smoothed model density and smoothed kernel density are

$$m_G^*(t_1, t_2) = \iint MVN\left\{ \begin{pmatrix} t_1 \\ t_2 \end{pmatrix}; \begin{pmatrix} x \\ z \end{pmatrix}, \left[h\hat{Cov} \begin{pmatrix} W \\ Z \end{pmatrix} + \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & 0 \end{pmatrix} \right] \right\} dG(x, z)$$

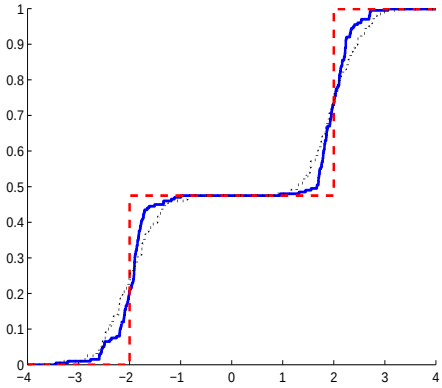
and

$$\hat{f}_n^*(y, t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} MVN\left\{ \begin{pmatrix} t_1 \\ t_2 \end{pmatrix}; \begin{pmatrix} w_i \\ z_i \end{pmatrix}, h\hat{Cov} \begin{pmatrix} W \\ Z \end{pmatrix} \right\} I(y = y_i)$$

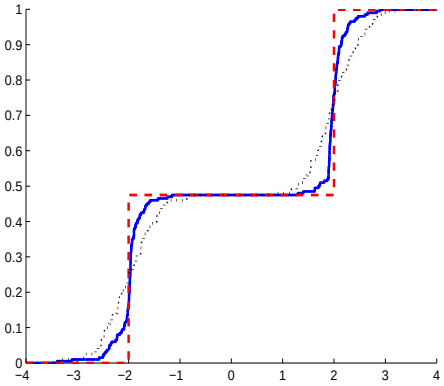
The estimated marginal cumulative distribution of X is shown in Figure 10.1 over different tuning parameter values h=0.0001, 0.001, 0.01, 0.1, 1, and 10 based on one sample with $n = 200$. In each plot, dotted line, dashed line and solid line represent



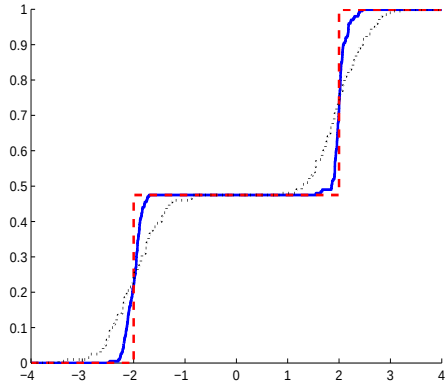
. $h = 0.0001$



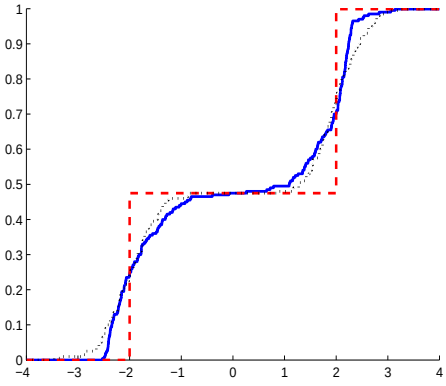
$h = 0.001$



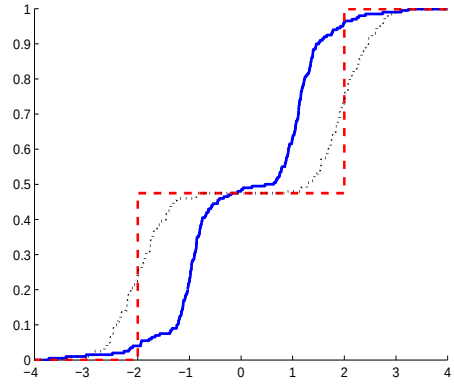
. $h = 0.01$



$h = 0.1$



. $h = 1$



$h = 10$

Fig. 10.1. Estimated marginal cumulative distribution of X

Table 10.1. The choice of tuning parameter h

h	0.0001	0.001	0.01	0.1	1	10
sDOF	67.88	41.60	17.23	5.11	1.46	0.73
$\widehat{Var_1}(X) - \widehat{Var_2}(X)$	-0.1238	-0.0737	-0.0253	0.026	0.2504	2.3993

empirical distribution of W, the empirical distribution of X and estimated distribution of X, respectively. As h decreases, we can see that the estimated distribution of X becomes close to the empirical distribution of W, obviously wrong, as we showed in Theorem 8.1. When h is very large, it also shows an inappropriate estimator for the distribution of X due to the large amount of smoothing. Based on Figure 10.1, an appropriate tuning parameter h could be between 0.01 and 0.1.

To choose a tuning parameter h as discussed in section 9.3, we calculate $\widehat{Var_1}(X) - \widehat{Var_2}(X)$ for each h shown in Table 10.1. This table implies that h should be between 0.01 and 0.1. This agrees with the conclusion from Figure 10.1. We can also see that this range agrees with the rule discussed in section 4.2 because the appropriate range for sDOF is 5 to 200/5=40. Moreover, within this range, $\widehat{Var_1}(X) - \widehat{Var_2}(X)$ is almost linear(or quadratic) in h shown in Figure 10.2. This fact enables us to find the optimal h by linear(quadratic) interpolation. Figure 10.2 suggests that $h = 0.04$ is optimal for the choice of the tuning parameter.

10.2 Estimation of both parametric and non-parametric component

Now we consider the full estimation with both the parametric component and nonparametric component. For this simulation study, We assume a linear regression

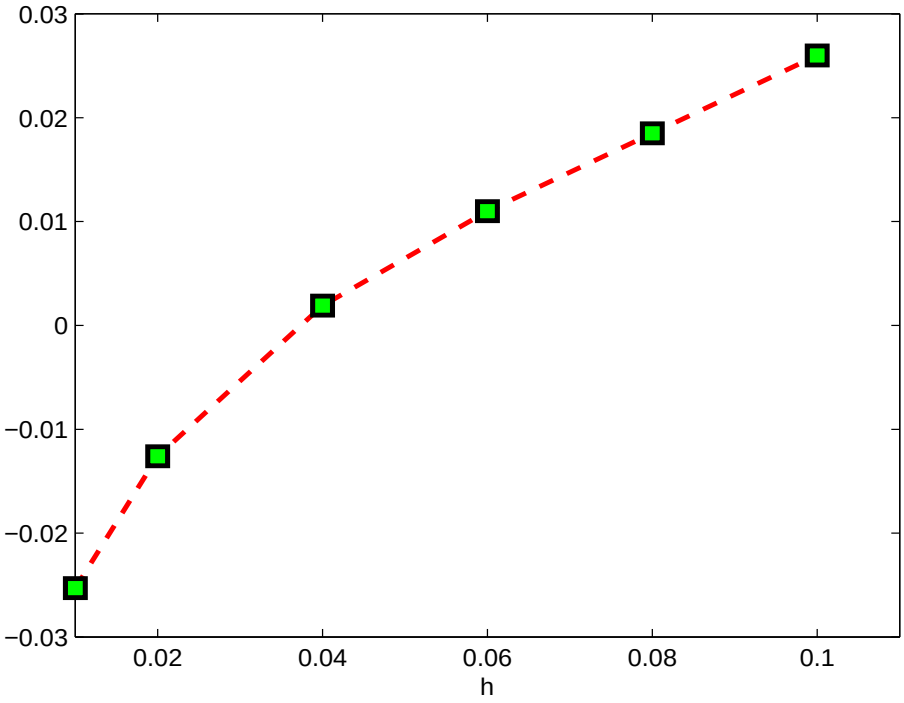


Fig. 10.2. Tuning parameter h versus $\widehat{Var_1(X)} - \widehat{Var_2(X)}$

model

$$Y = \beta_0 + \beta_x X + \beta_z Z + \epsilon,\ \epsilon \sim N(0, \sigma_\epsilon^2) \tag{10.1}$$

where Y is a response variable, X is a covariate measured with error, and Z is an additional covariate measured without error. We assume that X is not observed but instead the surrogate measure W is observed. And we assume additive measurement error $W = X + U$, $U \sim N(0, \sigma_u^2)$ and also assume that σ_u^2 is known.

For the simulation experiment, we fix $\beta_0 = 1$, $\beta_x = 2$, $\beta_z = 4$, $\sigma_\epsilon^2 = 1$, and generate (X, W, Z, Y) using following scheme.

1. Generate ϵ_i, u_i, z_i independently from $N(0,1), N(0, \sigma_u^2), N(0,1)$, respectively.
2. Generate x_i such that $x_i = \begin{cases} -2 & \text{with probability } \frac{e^{z_i}}{1+e^{z_i}} \\ 2 & \text{with probability } \frac{1}{1+e^{z_i}} \end{cases}$
3. Generate $w_i = x_i + u_i$
4. Generate $y_i = 1 + 2x_i + 4z_i + \epsilon_i$

For the proposed method, we define the smoothed model density $m^*(y, t_1, t_2)$ and the smoothed kernel density $\hat{f}_n^*(y, t_1, t_2)$ using bivariate normal kernel as below.

$$m^*_{\beta,G}(y, t_1, t_2) = \iint N(y; X'\beta, \sigma_\epsilon^2) MVN\left\{ \begin{pmatrix} t_1 \\ t_2 \end{pmatrix}; \begin{pmatrix} x \\ z \end{pmatrix}, \left[h\hat{Cov} \begin{pmatrix} W \\ Z \end{pmatrix} + \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & 0 \end{pmatrix} \right] \right\} dG(x, z) \tag{10.2}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_x, \beta_z)'$, $X = (1, x, z)'$ and the smoothed kernel density is

$$\hat{f}_n^*(y, t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} MVN\left\{\begin{pmatrix} t_1 \\ t_2 \end{pmatrix}; \begin{pmatrix} w_i \\ z_i \end{pmatrix}, h\hat{Cov}\begin{pmatrix} W \\ Z \end{pmatrix}\right\} I(y = y_i) \tag{10.3}$$

Now, first we fix $\sigma_u^2 = 0.1$ and generate three different data sets with $n = 50$, $n = 100$, and $n = 200$. For each data set, we used tuning parameters $h = 0.1$, $h = 0.03$, $h = 0.005$ based on the proposed decision rule in section 9.3.

Table 10.2. Estimates for $\boldsymbol{\beta}$

	$n = 50$			$n = 100$			$n = 200$		
	β_0	β_x	β_z	β_0	β_x	β_z	β_0	β_x	β_z
$\boldsymbol{\beta}$	1.0000	2.0000	4.0000	1.0000	2.0000	4.0000	1.0000	2.0000	4.0000
$\widetilde{\boldsymbol{\beta}}_1$	1.0421	2.0154	3.9339	1.0042	1.9759	3.9391	0.9205	2.0254	4.0036
$\widetilde{\boldsymbol{\beta}}_2$	1.1907	1.5970	3.4712	1.1136	1.6262	3.6328	0.9366	1.6848	3.6724
$\widehat{\boldsymbol{\beta}}$	1.0784	2.0896	3.9629	1.0229	1.9548	3.9431	0.8916	2.0044	3.9958

Table 10.2 shows the parameter estimators from each data set with the true parameter value $\boldsymbol{\beta}$. First, using the true covariates and response, $\{Y_i, X_i, Z_i, i = 1, \cdots, 50\}$, we estimated $\widetilde{\boldsymbol{\beta}}_1$ using the linear regression model (10.1). So $\widetilde{\boldsymbol{\beta}}_1$ is our target estimator. Second, ignoring the measurement error, we regress Y on (W, Z), then $\widetilde{\boldsymbol{\beta}}_2$ is an estimator under the model $Y = \beta_0 + \beta_1 W + \beta_2 Z + \epsilon$ hence $\widetilde{\boldsymbol{\beta}}_2$ is the MLE. Next, applying the algorithm in section 9.2, we estimated DSMLE, $\widehat{\boldsymbol{\beta}}$, based on the proposed approach.

The slopes of $\widetilde{\boldsymbol{\beta}}_2$ are biased and attenuated to 0 compared to the target estimator $\widetilde{\boldsymbol{\beta}}_1$, as discussed in section 7.2. Generally speaking, in the linear measurement error model

under additive measurement error, there are well known results for this attenuation with both X and Z (Carroll et al., 2006). That is,

$$\begin{pmatrix} \tilde{\beta}_{2x} \\ \tilde{\beta}_{2z} \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} + \Sigma_{uu} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix}^{-1} \left[\begin{pmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix} \begin{pmatrix} \tilde{\beta}_{1x} \\ \tilde{\beta}_{1z} \end{pmatrix} + \begin{pmatrix} \Sigma_{u\epsilon} \\ 0 \end{pmatrix} \right]$$

where Σ_{ab} means the covariance matrix between random variable $\boldsymbol{A}$ and $\boldsymbol{B}$. Thus the MLE does not work in this result but DSMLE does.

Because the first simulation experiment is restricted to only three data sets, in the second simulation experiment, we replicate the first simulation 50 times for $n = 50$ over two different amounts of measurement error $\sigma_u^2 = 0.1$ and 0.5. Mean and MSE of β are calculated based on 50 repetitions of the simulation. RMSE is the ratio of the MSE of $\tilde{\beta}_1$ to $\hat{\beta}$. These are shown in Table 10.3 and 10.4.

Table 10.3. Comparison between target estimator $\widetilde{\boldsymbol{\beta}}_1$ and $DSMLE$ when $\boldsymbol{\beta} = (1, 2, 4)$, $\sigma_u^2 = 0.1$

	β_0			β_x			β_z		
	Mean	MSE	RMSE	Mean	MSE	RMSE	Mean	MSE	RMSE
$\widetilde{\boldsymbol{\beta}}_{\mathbf{1}}$	0.9786	0.0252	1	1.9981	0.0057	1	4.0358	0.0266	1
$\widehat{\boldsymbol{\beta}}$	0.9536	0.0321	1.27	2.0343	0.0082	1.44	4.0726	0.0378	1.42

10.3 Conclusion and future work

Throughout Part II, we have considered semiparametric mixture methods to estimate both the parameters and joint distribution of the true predictors when we have

Table 10.4. Comparison between target estimator $\widetilde{\boldsymbol{\beta}}_1$ and $DSMLE$ when $\boldsymbol{\beta} = (1, 2, 4)$, $\sigma_u^2 = 0.5$

	β_0			β_x			β_z		
	Mean	MSE	RMSE	Mean	MSE	RMSE	Mean	MSE	RMSE
$\widetilde{\boldsymbol{\beta}}_1$	0.9762	0.0236	1	2.0084	0.0056	1	4.0229	0.0222	1
$\widehat{\boldsymbol{\beta}}$	0.9469	0.0569	2.41	2.0476	0.0191	3.41	4.0700	0.0828	3.73

additional error free covariates. The measurement errors are assumed to have a general parametric distribution under the non-differential error structure. In this case, the usual ML method breaks down when it estimates the covariate distribution due to the non-homogeneity of the accuracy of measurement. We applied the doubly-smoothed maximum likelihood estimation to repair this failure of the MLE and showed the DSMLEs of both parameters and covariate distribution are consistent.

We also discussed a simple algorithm to estimate both parametric and nonparametric components simultaneously. Although it provides us with a simple way to implement the proposed method, the computing time is still problematic. For example, when we estimated only the covariate distribution with $n = 200$, the elapsed computing time varied between 1600 seconds (26 minutes) and 3800 seconds (63 minutes) depending on the choice of the tuning parameter. All routines were coded in MATLAB and run on Pentium 4 CPU 3GHz with 1GB RAM. For the full estimation, if the proposed algorithm is applied, the required computing time will be at least ten times more than that of estimation of covariate distribution only. To resolve this computational difficulty, further research is necessary.

In the simulation study, we have focused only on showing the consistency of the DSMLE because the main objective of this study is to modify the failure of the ML method. However, the advantage of the semiparametric mixture model in the measurement error problem is its complete robustness to the choice of covariate distribution. Therefore, we need to compare DSMLE to other structural modeling approaches in order to see the robustness and efficiency of DSMLE.

Since the proposed estimator involves several non-explicit integrals, studying the efficiency is difficult especially when the model includes nonparametric components. In the measurement error problem, the model index involves a nonparametric distribution as well as vector-valued parameters. For the inference of the parameter of interest, we need to know the asymptotic distribution of DSMLE. So developing asymptotic theory for the DSMLE would be important future work.

For the estimation of the covariate distribution, we suggested using the EM algorithm to estimate X supports while fixing Z supports and corresponding weight. This method appears to be reasonable and makes the estimation simple. However, we need to be careful because we do not know if the resultant estimator is a real maximizer of the doubly-smoothed likelihood. So more computational investigation is required.

Bibliography

Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46:683–705.

Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *Annals of Statistics*, 5:445–463.

Billingsley, P. (1995). *Probability and Measure, 3rd edition.* John wiley & Sons, New York.

Bohning, D. (1985). Numerical estimation of a probability measure. *Journal of statistical planning and inference*, 11:57–69.

Bohning, D. (1986). A vertex-exchange-method in d-optimal design theory. *Metrika*, 33:337–347.

Carroll, R. J., Roeder, K., and Wasserman, L. (1999). Flexible parametric measurement error models. *Biometrics*, 55:44–54.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models. second edition.* Chapman & Hall : London.

Chung, K. L. (1974). *A course in probability theory, 2nd edition.* Academic Press, New York.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of fit tests. *Journal of the Royal Statistical Society, Series B.*, 46:449–464.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, 56:463–474.

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

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

Gustafson, P., Le, N., and Vallée, M. (2002). A beyesian approach to case-control studies with errors in covariables. *Biostatistics*, 3:229–243.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, 13:795–800.

Heckman and Singer (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52:271–320.

Huang, X., Stefanski, L. A., and Davidian, M. (2006). Latent-model robustness in structural measurement error models. *Biometrika*, 93:53–64.

Hunter, D. R. (2003). On the geometry of em algorithms. Technical Report. Penn State University.

Hunter, D. R. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statician*, 58:30–37.

Jones, D. Y., Schatzkin, A., Green, S. B., Block, G., Brinton, L. A., Ziegler, R. G., Hoover, R., and Taylor, P. R. (1987). Dietary fat and breast cancer in the national helath and nutrition survey I: Epidemiologic follow-up study. *Journal of the National Cancer Institute*, 79:465–471.

Kaplan, E. L. and Meier, P. (1958). Non-parametric estimation from incomplete obervations. *Journal of the American Statistical Association*, 53:457–481.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27:886–906.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73:805–811.

Lesperance, M. L. and Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric mle of a mixing distribution. *Journal of the American Statistical Association*, 87:120–126.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum hellinger distance and related methods. *Annals of Statistics*, 22:1081–1114.

Lindsay, B. G. (1995). *Mixture models: theory, geometry, and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5. IMS : US.

Lindsay, B. G., Markatou, M., Ray, S., Yang, K., and Chen, S. (2007). Diffusion kernels and quadratic distances as building blocks for high dimensional inference. To appear in Annals of Statistics.

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

Mcfadden, D. (1989). A method of simulated moments for estimation of discrete response modes without numerical ingegration. *Econometrica*, 57:995–1026.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32.

Ray, S. and Lindsay, B. G. (2007). Model selection in high-dimensions: A quadratic-risk based approach. Manuscript.

Richardson, S., Leblond, L., Jaussent, I., and Green, P. J. (2002). Mixture models in measurement error problems with reference to epidemiological studies. *Journal of the Royal Statistical Society, Series A*, 165:549–566.

Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariates. *Journal of the American Statistical Association*, 91:722–732.

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

Schafer, D. W. (2001). Semiparametric maximum likelihood for measurement error model regression. *Biometrics*, 57:53–61.

Simpson, D. G. (1987). Minimum hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association*, 82:802–807.

Simpson, D. G. (1989). Hellinger deviance tests: efficiency, breakdown points, and examples. *Journal of the American Statistical Association*, 84:107–113.

Tamura, R. N. and Boos, D. D. (1989). Minimum hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81:223–229.

Tan, X., Chen, J., and Zhang, R. (2006). Consistency of the constrained maximum likelihood estimator in finite normal mixture models. Manuscript.

van der Laan, M. J. (1996). Efficient estimation in the bivaraite censoring model and repairing npmle. *Annals of Statistics*, 24:596–627.

Wang, L. (2004). Estimation of nonlinear models with berkson measurement errors. *Annals of Statistics*, 32:2559–2579.

Yang, K. (2004). *Using the Poisson kernel in model building and selection.* PhD thesis, Penn State.

Vita

Byungtae Seo

Byungtae Seo was born in Seoul, Korea 1975. He received his B.S. degree in Statistics from Seoul National University in 2001. He enrolled in the Ph.D. program in Statistics at The Pennsylvania State University in 2002. His research interest involves mixture models, minimum distance estimation, measurement error models and nonparametric statistics.