The Pennsylvania State University

The Graduate School

LIKELIHOOD-TUNED DENSITY ESTIMATOR AND ITS APPLICATION TO CLUSTERING

A Dissertation in

Statistics

by

Yeojin Chung

© 2010 Yeojin Chung

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

August 2010

The thesis of Yeojin Chung was reviewed and approved* by the following:

Bruce G. Lindsay
Willaman Professor of Statistics and Department Head
Dissertation Co-advisor, Co-chair of Committee

Jia Li
Associate Professor of Statistics
Dissertation Co-advisor, Co-chair of Committee

Bing Li
Professor of Statistics

David Hunter
Associate Professor of Statistics

Jesse Barlow
Professor of Computer Science and Engineering

*Signatures are on file in the Graduate School.

Abstract

Nonparametric density estimation is widely used for investigating underlying features of data. We introduce a likelihood enhanced nonparametric density estimator which arises from treating the kernel density estimator as an element of the model that consists of all mixtures of the kernel, continuous or discrete. One can obtain the kernel density estimator with "likelihood-tuning" by using the uniform density as the starting value in an EM algorithm. We prove algorithmic convergence of this EM algorithm to the nonparametric mixture maximum likelihood estimator. The second tuning step leads to a fitted density with higher likelihood than the kernel density estimator. This twice tuned density estimator reduces the bias of the kernel density estimator while the order of variance stays the same. Our simulation study shows that the second-tuned estimator performed robustly against the type of densities, but this feature tended to weaken relative to a competing estimator as the data dimension grew.

Along with the type of density estimators, the bandwidth selection problem is very crucial in the nonparametric density estimation, particularly in higher dimensions. We introduce a new bandwidth selection method using the spectral degrees of freedom introduced in Lindsay et al. (2008). Investigating the theoretical sDOF and simulation results, we found that the bandwidth need to increase proportionally to the square root of dimension if we are to achieve adequate smoothing in higher dimensions.

We also develop a penalized version of the likelihood-tuning procedure that allows the mixture model to adapt to local shape and scale features. This model gives the kernel density estimator with the t-kernel with the first tuning. The second penalized-tuning leads to a density estimator with local shape adaptation in the t-kernel function. We compare the performance of the new density estimators with unpenalized likelihood tuned density estimators.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I owe my deepest gratitude to my advisor, Dr. Bruce G. Lindsay for his guidance, support and patience. He not only gave invaluable advice through this dissertation but also showed me how fun and exciting research is. Whenever I was stuck, he always encouraged me to go ahead and enlightened the direction of my research. I would like to thank my co-advisor, Dr. Jia Li, for motivating me to explore my research topic. She gave me a wonderful support in my Ph. D. study.

This is a great opportunity to express my respect to my parents. Though they have been far apart, their heartfelt support and trust are always my biggest fuel of my life. This dissertation would not have been possible unless my mother-in-law helped me in the last few months. Her assistance and sacrifice enabled me to continue to work on this dissertation with my little girl. I especially thank my love, Kion. As the husband, as the best friend, and sometimes as a colleague, he has been giving me endless encouragement and being a pillar of my life. Lastly, this dissertation is dedicated to all my love, Gina.

Chapter 1

Introduction

Nonparametric density estimation has been widely used for investigating underlying features in a set of data in many areas. Of sepcial interest to this thesis, many authors have focused on applying nonparametric density estimation to cluster analysis and bump hunting. (Li et al. (2007), Tran et al. (2006), Good and Gaskins (1980), Azzalini and Torelli (2007)) Among them, Li et al. (2007) proposed a clustering algorithm based on finding the modes of the kernel density estimator, given by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x, X_i). \tag{1.1}$$

Their method has the flexibility to have characteristics between K-means clustering and linkage clustering, and showed good performance on various shapes of data. In addition, it has gained attention in various applications. (Sarro et al. (2009), Yao and Lindsay (2009))

Motivated by the modal-based clustering problem, this dissertation is partly devoted to develop a new nonparametric density estimation method focused on application to clustering and to investigate its properties. Our main tool to improve the kernel density

estimator is based on a nonparametric mixture model

$$f(x) = \int K(x;\phi)d\Pi(\phi) \tag{1.2}$$

and the nonparametric likelihood function with a random sample $X_1, \ldots, X_n$ following $f(x)$, given by

$$L(\Pi) = \sum_i \log \int K(X_i;\phi)d\Pi(\phi).$$

Our goal is to obtain a density estimator from (1.2), $\hat{f}(x) = \int K(x;\phi)d\hat{\Pi}(\phi)$ by fitting $\hat{\Pi}$ with higher likelihood.

In Chapter 2, we propose the *likelihood-tuning procedure* to update Π to increase the likelihood by employing the continuous EM algorithm found in Vardi and Lee (1993). One can obtain the kernel density estimator with "likelihood-tuning" by using the uniform density as the starting value in an EM algorithm. The second tuning step leads to a fitted density with higher likelihood than the kernel density estimator. The two step EM estimator can be written explicitly when K is a Gaussian kernel, and its bias is one order of magnitude smaller than the kernel estimator. In addition, the order of magnitude of the variance stays of the same order, so that the asymptotic mean squared error can be reduced significantly. Compared with other modified density estimators, our simulation results show that the two-step likelihood-tuned density estimator performs robustly against different types of true density in one dimension.

Chapter 3 extends the likelihood-tuned density estimator to higher dimensions. The asymptotic bias is also reduced from the bias of the kernel density estimator while the order of variance stays the same. In a simulation study, however, compared to other improved density estimators, the advantage that the likelihood-tuned density estimator gained for non-Gaussian densities in the univariate simulation tended to be weaken as dimension grew. However, the simulation to observe the performance on mode identification showed that the likelihood-tuned density estimator was somewhat superior to the other density estimators. In this section, we also provide proofs for the algorithmic

convergence of the continuous EM algorithm to the nonparametric MLE, a new result.

Along with the problem of finding a mode-sensitive density estimator, the problem of bandwidth selection is crucial in kernel-type density estimation. In keeping with our approach, it will be influenced by the purpose of our investigation, as noted in Silverman (1986). In higher dimensions, there are important new difficulties in selecting an appropriate bandwidth because of sparsity of data. In Chapter 4, we examine a new bandwidth selection method, focused on clustering in higher dimensions. Instead of standard bandwidth selection methods to optimize some selection criteria, we employ the spectral degrees of freedom (sDOF) introduced in Lindsay et al. (2008). The sDOF gives a measure of smoothing that does not depend on dimension. In addition, keeping the sDOF approximately constant, we would control variance of the model to ensure we are capturing key features of the true density in higher dimensions. When we examine theoretical sDOF and simulation results, we found that in both venues the bandwidth would need to increase proportionally to the square root of dimension in order to produce good clustering results.

In Chapter 5, we consider a more relaxed model than the model in Chapter 2 and 3. With a Gaussian kernel $K_h(x;\phi) = h^{-1}\varphi(h^{-1}(x-\phi))$, we regarded the mixing variable ϕ in (1.2) to be the location parameter. We next extend ϕ to include both the location and the scale parameters. However, this fully unrestricted model has unbounded likelihood, so we consider a penalized likelihood function, given by

$$L(\Pi; X_i) = \int K(X_i;\mu,\sigma)P(\sigma)d\Pi(\mu,\sigma). \tag{1.3}$$

With inverse-Gamma or inverse-Wishart for $P(\sigma)$ and with starting from an initial non-informative $\hat{\Pi}$, the first likelihood-tuning gives the kernel density estimator with t-kernel and the second likelihood-tuning yields the density estimator with t-kernel including a local scale and shape adaptation term. Although the simulation results did not show immediate improvement from the unpenalized likelihood-tuned density estimator, there are

a number of possible extensions to investigate including reparameterization and bandwidth selection.

Chapter 2

A Likelihood-tuned Density Estimator via a Nonparametric Mixture Model

HMAC is based on the kernel density estimator, which has been modified to reduce the bias. We might expect that, if HMAC is applied to an improved density estimator, it would show better performance in clustering.

In this chapter, we consider an improved density estimator which arises from treating the kernel density estimator as an element of the model that consists of all mixtures of the kernel, continuous or discrete. One can obtain the kernel density estimator with "likelihood-tuning" by using the uniform density as the starting value in an EM algorithm. The second tuning leads to a fitted density with higher likelihood than the kernel density estimator. The two step EM estimator can be written explicitly with a Gaussian kernel, and its bias is one order of magnitude smaller than the kernel estimator. In addition, the order of magnitude of the variance stays of the same order, so that the asymptotic mean squared error can be reduced significantly. Compared with other modified density estimators, the simulation result shows that the two-step likelihood-tuned

density estimator performs robustly against different types of true density.

This chapter of the thesis has been submitted for publication in a festschrift in honor of Tom Hettmansperger in the Lecture Notes - World Scientific.

2.1 Introduction

The kernel density estimator is a widely used nonparametric density estimator. Let $X_1, \cdots, X_n$ be a random sample from $f(x)$. Then the kernel density estimator of $f(x)$, denoted $\hat{f}_{\mathrm{KER}}(x)$ in this paper, is defined by

$$\hat{f}_{\mathrm{KER}}(x) = n^{-1} \sum_{i=1}^{n} K_h(x, X_i), \tag{2.1}$$

where $K_h(\cdot, \cdot)$ is a known kernel function with a bandwidth h. Many authors have modified it to reduce its bias, a key element of its mean squared error. In this paper, we propose a new method that is based on applying one step of the EM algorithm to $\hat{f}_{\mathrm{KER}}$ in a class of mixture models. This also reduces bias, and it has an advantage over other methods in that it is based on a likelihood device that can easily be generalized to other smoothing problems.

Breiman et al. (1977) proposed to replace h in (2.1) by a variable bandwidth $h(X_i)$ depending on an observation. Abramson (1982) suggested taking $h(X_i)$ proportional to $f(X_i)^{-1/2}$ by showing that this reduces the bias to $O(h^4)$. Although this procedure requires one to know the true $f(x)$, Silverman (1986) mentioned that the basic kernel estimator in (2.1) works well as a pilot estimator of $f(X_i)$ for the variable bandwidth. In addition, he defined *the adaptive bandwidth density estimator*, denoted by $\hat{f}_{\mathrm{ABW}}$ throughout this paper, as

$$\hat{f}_{\mathrm{ABW}}(x) = n^{-1} \sum_{i=1}^{n} K_{h(\hat{f}_{\mathrm{KER}}(X_i)/g)^{-1/2}}(x, X_i), \tag{2.2}$$

where g is the geometric mean of the $\hat{f}_{\mathrm{KER}}(X_1), \cdots, \hat{f}_{\mathrm{KER}}(X_n)$.

Jones et al. (1995) considered a multiplicative bias correction that results in the density estimator given by

$$\hat{f}_{\mathrm{MBC}}(x) = \hat{f}_{\mathrm{KER}}(x) \cdot n^{-1} \sum_{i=1}^{n} \hat{f}_{\mathrm{KER}}(X_i)^{-1} K_h(x, X_i). \tag{2.3}$$

The leading bias term of $\hat{f}_{\mathrm{KER}}$ in the numerator is canceled by the leading bias of $\hat{f}_{\mathrm{KER}}$ in the denominator so that $\hat{f}_{\mathrm{MBC}}$ attains a bias of order $O(h^4)$ (Jones and Signorini, 1997). The estimator in (2.3) is referred as the *multiplicative bias correction density estimator* in this paper, following Jones and Signorini (1997).

Recently, DiMarzio and Taylor (2004) applied boosting to kernel density estimation. They adopted a goodness-of-fit measure that compares the kernel density estimator with the leave-one-out estimator (Silverman, 1986). Based on this measure, the boosting step updates the weight for each $K(X_i, \cdot)$ and fits the weighted kernel density estimator. Starting from the uniform initial weight, n^{-1}, the first boosting step provides $\hat{f}_{\mathrm{MBC}}$. While a further boosting step reduces the average ISE for the Gaussian distribution, the simulation result shows that it does not clearly improve the performance in the case of non-Gaussian distributions.

In this paper, we will compare our new method with $\hat{f}_{\mathrm{KER}}$, $\hat{f}_{\mathrm{ABW}}$ and $\hat{f}_{\mathrm{MBC}}$. We have chosen $\hat{f}_{\mathrm{ABW}}$ because this seems to be the most commonly used modified density estimator. In addition, Jones et al. (1995) compared six higher-order bias density estimators, including $\hat{f}_{\mathrm{ABW}}$ and $\hat{f}_{\mathrm{MBC}}$, and found that these two estimators were the most competitive. Adopting another tuning parameter of the bandwidth for $\hat{f}_{\mathrm{KER}}$ in (2.2) or in (2.3) can give better performance for some densities. However, for a fair comparison, we will use the common bandwidth for $\hat{f}_{\mathrm{KER}}$ and the kernel function in (2.2) and (2.3) because it is most natural to compare our estimator with other estimators in their basic form.

This article proposes a new density estimator that reduces the bias. To create it, we look at the kernel density estimator as an estimator via the nonparametric mixture

model and use the EM algorithm to improve its likelihood. In Section 2.1, we describe the connection between the nonparametric mixture model and nonparametric density estimation, two areas that have been mostly treated in separate literatures to date. In Section 2.2, the likelihood-tuning procedure and the resulting density estimators are introduced. In Section 3, we investigate the asymptotic properties of the new density estimator and compare them with existing estimators such as the basic kernel, adaptive bandwidth and multiplicative bias correction density estimator. The simulation comparisons are given in Section 4.

2.2 Methodology

2.2.1 Background

Consider the nonparametric mixture model with a mixing (latent) distribution $\Pi(\phi)$, given by

$$f(x;\Pi) = \int K(x,\phi)d\Pi(\phi), \tag{2.4}$$

where $K(\cdot,\cdot)$ is a known density function, called a component density, and ϕ represents a component parameter. Here the distribution function $\Pi(\phi)$ is allowed to be discrete, continuous or from any specific family of distributions. If Π is discrete with point masses π_i at $\phi_i,\ i = 1,\cdots,m$, then (3.5) becomes the m-component finite mixture model, written as

$$f(x;\phi_1,\cdots\phi_m,\pi_1,\cdots,\pi_m) = \sum_{i=1}^{m}\pi_i K(x,\phi_i). \tag{2.5}$$

Let $X_1,\cdots,X_n$ be a random sample from $f(x;\Pi)$. If Π is estimated by $\hat{\Pi}(\phi) = n^{-1}$ at $\phi = X_i,\ i = 1,\cdots,n$ and 0 elsewhere, (3.5) can be written as

$$f(x;\hat{\Pi}) = n^{-1}\sum_{i=1}^{n}K(x,X_i),$$

and this is the same as the basic kernel density estimator in (2.1). This view of the

kernel density estimator in the context of nonparametric mixture models leads to the idea that an improved estimator of Π would provide a good estimator of $f(x)$.

There is an extensive literature concerning the nonparametric maximum likelihood estimator of the mixing distribution Π. Consider the likelihood function of an observation, given by

$$L_i(\Pi) = \int K(X_i, \phi) d\Pi(\phi),$$

and the log-likelihood function with multiple observations $n(i)$ of a single L_i, given by

$$l(\Pi) = \sum_{i=1}^{D} n(i) \ln(L_i(\Pi)), \tag{2.6}$$

where $\sum_{i=1}^{D} n(i) = n$. Here we do not specify any parametric form of Π, allowing it to be either discrete or continuous. The estimator $\hat{\Pi}$ that maximizes (3.7) is called the nonparametric maximum likelihood estimator (NPMLE) of Π (Lindsay, 1995). Lindsay (1995) proves that if the L_i's are all distinct and all $n(i) > 0$, then there exists an NPMLE, $\hat{\Pi}$, that is a discrete distribution with no more than D distinct points of support.

Since the NPMLE is discrete on finitely many support points, we can apply the EM algorithm by restricting $\Pi(\phi)$ to be a discrete distribution on a large number of support points (Laird, 1978; Vardi et al., 1985). On the other hand, we could consider the generalized EM algorithm that allows the initial $\Pi(\phi)$ to be continuous with a density function $\pi(\phi)$, even though we know it converges to the discrete NPMLE. Vardi and Lee (1993) and Lindsay (1995) described *the continuous EM algorithm for NPMLE*, given by

$$\hat{\pi}_{(k+1)}(\phi) = \hat{\pi}_{(k)}(\phi) \cdot n^{-1} \sum_{i=1}^{n} \frac{K(X_i, \phi)}{L_i(\hat{\Pi}_{(k)})}. \tag{2.7}$$

Notice that this is a generalization of the standard EM algorithm for updating the mixing proportion π_j in the discrete mixture model in (3.6), written as

$$\pi_{j,(k+1)} = \pi_{j,(k)} \cdot n^{-1} \sum_{i=1}^{n} \frac{K(X_i, \phi_j)}{f(x_i)}.$$

In this paper, we consider the consequences of using a continuous uniform density, $\pi(\phi) = 1$, as the initial estimate of $\hat{\pi}$. Each update by (3.9) will produce a spikier estimate of π that assigns more weights to the support points of the discrete NPMLE. A single update by the EM gives the basic kernel density estimator and the second update generates our new and improved density estimator. Although the second-step estimator of π is still not the NPMLE, it does provide a density estimator that is asymptotically superior to the kernel density estimator in many cases. It also shows better overall performance in our simulation study than other modified kernel density estimators.

Similar ideas have appeared in the mixture literature, but the focus has been on the estimator of Π, not on the resulting estimator for the density of x. Laird and Louis (1991) viewed $\Pi(\phi)$ as a prior distribution of a parameter ϕ and x_i's as the realizations from $x_i|\phi \sim N(\phi, 1)$. They employed a uniform prior and updated Π with an empirical Bayes estimate. This procedure of updating a prior distribution Π is identical to the continuous EM algorithm of updating the mixing distribution Π in (3.9).

There have been other attempts to smooth the NPMLE for Π that do not depend on the EM. Goutis (1997)'s method is motivated by the kernel density estimator of the unobserved data ϕ_i's in the incomplete data (X_i, ϕ_i) problem. Since we cannot observe the ϕ_i's, he proposed to iteratively estimate the conditional expectation of the kernel density estimate of ϕ_i. With the Gaussian kernel with a bandwidth h, this procedure turns out to be the same as the generalized EM in (3.9) except that the update of π is generated by kernels with a larger bandwidth than the update of π by the EM. Silverman et al. (1990) proposed the EMS algorithm, which adds a smoothing step to the EM for smoothing the estimate in each iteration. These authors have focused on reducing the spiky feature of the NPMLE of $\hat{\Pi}$. On the contrary, we are utilizing our methodology to improve the kernel density estimator for x.

2.2.2 Likelihood-tuning Procedure

Prior to introducing the likelihood-tuning procedure, we define a gradient function of the log-likelihood function $l(\Pi)$ in (3.7) at Π_0 toward the direction of ϕ by

$$D_{\Pi_0}(\phi) = \sum_{i=1}^{D} n(i) \left(\frac{K(X_i, \phi)}{L_i(\Pi_0)} - 1 \right)$$

(Lindsay, 1995). This is derived from the directional derivative of $l(\Pi)$ at Π_0 along the path between Π_0 and Π_1, that is,

$$D_{\Pi_0}(\Pi_1) = \frac{\partial}{\partial \alpha} l(\alpha \Pi_0 + (1-\alpha)\Pi_1)|_{\alpha=0}.$$

When Π_1 is a degenerate distribution at ϕ, $D_{\Pi_0}(\Pi_1)$ becomes $D_{\Pi_0}(\phi)$. Thus if $D_{\Pi_0}(\phi)$ is positive, $l(\Pi_0)$ will be increased by adding more probability at ϕ in the mixing distribution Π_0. On the contrary, if $D_{\Pi_0}(\phi)$ is negative for ϕ, $l(\Pi_0)$ will be increased by shrinking the mass at ϕ in Π_0.

Using the gradient function, the continuous EM algorithm in (3.9) can be written as

$$\hat{\pi}_{(k+1)}(\phi) = \hat{\pi}_{(k)}(\phi) \left(1 + n^{-1} D_{\hat{\Pi}_{(k)}}(\phi) \right).$$

This equation implies that the EM algorithm increases the density at ϕ where the gradient function is positive and reduces the density where the gradient function is negative. We note that $\hat{\pi}_{(k+1)}(\phi)$ integrates to 1 if $\hat{\pi}_{(k)}(\phi)$ does.

The likelihood-tuning procedure that updates a density estimator of x includes two steps: updating the mixing density π and updating the density estimator of x. Given an initial estimate $\pi_0(\phi)$, the *likelihood-tuning procedure* is as follows.

1. Update the estimator of the mixing density π by

$$\hat{\pi}_{(k+1)}(\phi) = \hat{\pi}_{(k)}(\phi) \Delta_{(k)}(\phi),$$

where $\Delta_{(k)}(\phi) = 1 + n^{-1} D_{\Pi_{(k)}}(\phi)$.

2. Update the density estimator of x by

$$\hat{f}_{(k+1)}(x) = \int K(x, \phi) \hat{\pi}_{(k+1)}(\phi) \, d\phi.$$

Notice that it is desirable to avoid numerical integration in step 2. We will take the uniform density $\pi_0(\phi) = 1$ to be the initial estimate of π given a lack of any prior information about the location of peaks. There are sound theoretical reasons for this choice, as it gives a gradient function that tends to identify the needed points of support in the NPMLE. In each step thereafter, $\Delta_{(k)}(\phi)$ indicates the deviation of the latent density $\pi_{(k+1)}(\phi)$ from $\pi_{(k)}(\phi)$, which therefore increases mass at the highest gradient values. In fact, one can show that repeated application of the continuous EM algorithm converges to the NPMLE without the parameter space searches ("gradient checks") that are required when one uses discrete Π estimators (Lindsay, 1995).

Let $K_h(\cdot, \cdot)$ be a kernel function with a bandwidth h. If we consider a diffusion kernel function, which satisfies the diffusion equation

$$\int K_h(x, \phi) K_h(\phi, y) d\phi = K_{\sqrt{2}h}(x, y),$$

then the estimators of $\pi(\phi)$ and $f(x)$ by the likelihood-tuning procedure are reduced to explicit forms. The Gaussian kernel function, which is such a diffusion kernel, is defined by $K_h(x, \phi) = h^{-1} \varphi\left\{h^{-1}(x - \phi)\right\}$, where φ is the standard normal density function. Then the first likelihood-tuning iteration provides

$$\hat{\pi}_{(1)}(\phi) = n^{-1} \sum_{i=1}^{n} K_h(\phi, X_i)$$

and

$$\begin{aligned}\hat{f}_{(1)}(x) &= \int K_h(x,\phi)\cdot n^{-1}\sum_{i=1}^{n} K_h(\phi, X_i)\; d\phi \\ &= n^{-1}\sum_{i=1}^{n} K_{h\sqrt{2}}(x, X_i). \end{aligned} \tag{2.8}$$

Notice that the estimate $\hat{\pi}_{(1)}(\phi)$ from a single likelihood-tuning step turns out to be the basic kernel density estimator. The resulting fitted density for x, $\hat{f}_{(1)}$, is also the basic kernel density estimator but with a wider bandwidth than the bandwidth in $\hat{\pi}_{(1)}$.

If we apply the likelihood-tuning procedure once more, we obtain

$$\hat{\pi}_{(2)}(\phi) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} K_{h/\sqrt{2}}(\phi, \bar{X}_{ij}),$$

where $w_{ij} = K_{h\sqrt{2}}(X_i, X_j)/\hat{f}_1(X_i)$ and $\bar{X}_{ij} = (X_i + X_j)/2$. The second step density estimator is given by

$$\begin{aligned}\hat{f}_{(2)}(x) &= n^{-2}\sum_{i}\sum_{j} w_{ij}\int K_h(x,\phi) K_{h/\sqrt{2}}(\phi, \bar{X}_{ij}) d\phi \\ &= n^{-2}\sum_{i}\sum_{j} w_{ij} K_{h\sqrt{\frac{3}{2}}}(x, \bar{X}_{ij}). \end{aligned} \tag{2.9}$$

We define this second step density estimator in (3.13) as the *two-step likelihood-tuned density estimator* and denote it by $\hat{f}_{\mathrm{EM2}}(x)$.

Proceeding in the same manner, one can take more EM steps to move $\hat{\pi}$ even closer to the NPMLE than $\hat{\pi}_{(2)}$. For any positive integer k, $\pi_{(k)}$ is generalized by

$$\hat{\pi}_{(k)}(\phi) = \frac{1}{n^k}\sum_{i_1=1}^{n}\cdots\sum_{i_k=1}^{n} w_{i_1,\cdots,i_{k-1}} K_{h/\sqrt{k}}(\phi, \bar{X}_{i_1,\cdots,i_k}),$$

where

$$w_{i_1,\cdots,i_{k-1}} = c(h) \cdot \frac{K_{h\sqrt{k}}(X_{i_1}, X_{i_2}) \cdots K_{h\sqrt{k}}(X_{i_{k-1}}, X_{i_k})}{\hat{f}_{(1)}(X_{i_1}) \cdots \hat{f}_{(k-1)}(X_{i_{k-1}})}$$

with known constant $c(h)$ and $\bar{X}_{i_1,\cdots,i_k}$ is the mean of $(X_{i_1}, \cdots, X_{i_k})$. The updated density estimator can be expressed in the generalized form

$$\hat{f}_{(k)}(x) = \frac{1}{n^k} \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} w_{i_1,\cdots,i_{k-1}} K_{h\sqrt{\frac{k+1}{k}}}(x, \bar{X}_{i_1,\cdots,i_k}). \tag{2.10}$$

Although the density estimator in (3.15) with $k \geq 3$ might have more desirable asymptotic properties than the second step estimator, it requires intensive computations. In fact, when one moves from the $(k-1)$th step estimator to the kth step, the number of summands is increased from n^{k-1} to n^k. Without proceeding further, therefore, this paper focuses on the second step estimator, $\hat{f}_{\mathrm{EM2}}(x)$, and will show that it has desirable asymptotic properties and promising simulation results in Sections 3 and 4.

2.3 Asymptotic Properties

In this section, we show the asymptotic properties of the likelihood-tuned density estimator and compare them with other improved density estimators.

Theorem 2.1. *Suppose that $f(x)$ is four times continuously differentiable and $K_h(x, \phi) = h^{-1}\varphi\left(h^{-1}(x-\phi)\right)$. Then, when $h \to 0$ and $nh \to \infty$,*

$$\begin{aligned} E\left[\hat{f}_{\mathrm{EM2}}(x)\right] &= f(x)\left(-\frac{f^{(4)}(x)}{f(x)} + \frac{f^{(3)}(x)f'(x) + f''(x)^2}{f(x)^2} - \frac{f''(x)f'(x)^2}{f(x)^3}\right)h^4 \\ &\quad + f(x) + o(h^4) \end{aligned}$$

and

$$Var\left[\hat{f}_{\mathrm{EM2}}(x)\right] = (nh)^{-1} f(x) \frac{1}{\sqrt{\pi}}\left(\frac{1}{4} + \sqrt{2} - \frac{2}{\sqrt{3}}\right) + o\left((nh)^{-1}\right).$$

See the Appendix for an outline proof.

Theorem 2.1 reveals that the likelihood-tuned density estimator has bias of order $O(h^4)$ and variance of order $O((nh)^{-1})$, as do the other modified density estimators. Before comparing it with other density estimators, one should note that we had a bandwidth of $h\sqrt{2}$ instead of h for the one-step kernel density estimator in Section 2.2. Thus it is reasonable to compare the one-step and two-step estimators to see the effect of likelihood tuning. When we make comparisons with the adaptive bandwidth and the multiplicative bias correction estimator, we also need to account for bandwidth effects.

Jones et al. (1995) provided bias and variance formulas for the adaptive bandwidth density estimator in (2.2). With the Gaussian kernel defined in Theorem 1, the asymptotic bias reduces to

$$\left(-\frac{3f^{(4)}(x)}{2f(x)}+\frac{6f^{(3)}(x)f'(x)+4f''(x)^2}{f(x)^2} \right. \\ \left. -20\frac{f''(x)f'(x)^2}{f(x)^3}+12\frac{f'(x)^4}{f(x)^4}\right)h^4+o(h^4)$$

and the asymptotic variance to

$$(nh)^{-1}f(x)\frac{1}{\sqrt{\pi}}\left(\frac{1}{2\sqrt{2}}+\frac{1}{2\sqrt{3}}+\frac{1}{16}\left(1+\frac{x^2}{4}\right)\right).$$

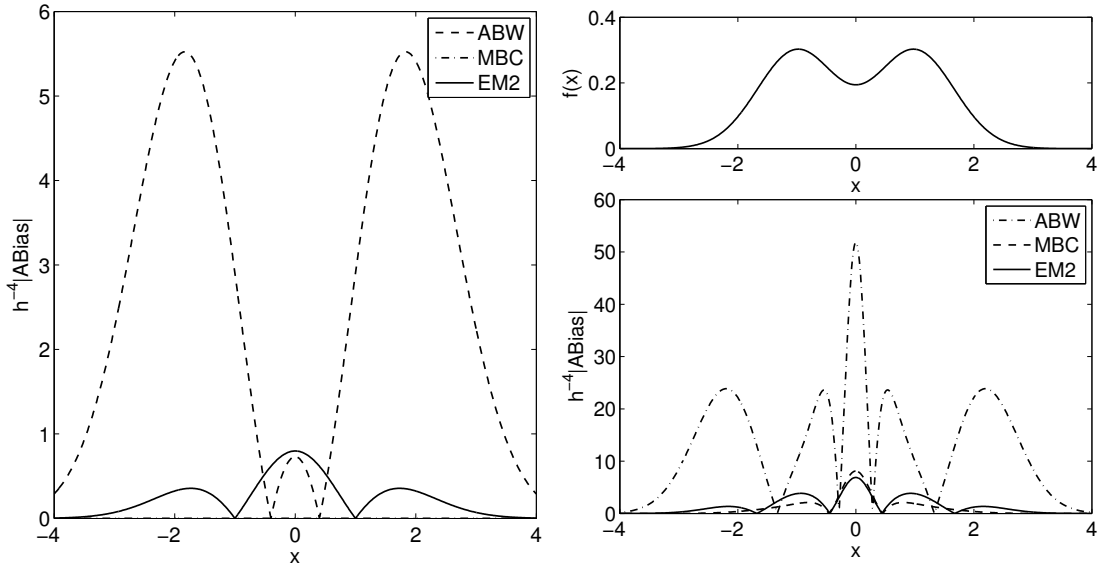
Since $\hat{f}_{\mathrm{MBC}}(x)$ given in (2.3) does not integrate to 1, Jones et al. (1995) suggested rescaling it by dividing $\hat{f}_{\mathrm{MBC}}(x)$ by its integral. They provided the bias of the rescaled version. With the Gaussian kernel function, the rescaled bias becomes

$$f(x)\left(-\frac{f^{(4)}(x)}{f(x)}+\frac{2f^{(3)}(x)f'(x)+f''(x)^2}{f(x)^2} \right. \\ \left. -\frac{2f''(x)f'(x)^2}{f(x)^3}+\int\frac{f''(z)^2}{f(z)}dz\right)h^4+o(h^4).$$

The rescaled $\hat{f}_{\mathrm{MBC}}(x)$ has exactly the same asymptotic variance as the likelihood-tuned density estimator.

To compare the theoretical MSE of $\hat{f}_{\mathrm{ABW}}$, $\hat{f}_{\mathrm{MBC}}$, and $\hat{f}_{\mathrm{EM2}}$, we replace h of $\hat{f}_{\mathrm{ABW}}$

by $h \cdot AVar(\hat{f}_{\mathrm{ABW}})/AVar(\hat{f}_{\mathrm{EM2}})$, as Jones et al. (1995) proposed. Then $\hat{f}_{\mathrm{ABW}}$ also has the same asymptotic variance as $\hat{f}_{\mathrm{EM2}}$, so that MSE comparison can be based only on the rescaled bias.



(a) Normal(0,1)

(b) Bimodal (with density shown at top)

Figure 2.1: Asymptotic biases after rescaling for $\hat{f}_{\mathrm{ABW}}$.

Figure 2.1(a) displays the absolute values of asymptotic biases after rescaling the bias of $\hat{f}_{\mathrm{ABW}}$ for the standard normal distribution (after removing h^4). Since $\hat{f}_{\mathrm{MBC}}$ has zero bias for $N(0,1)$ with order $o(h^4)$, the dash-dotted line is flat on the horizontal axis. Around the mode, $\hat{f}_{\mathrm{ABW}}$ is superior to $\hat{f}_{\mathrm{EM2}}$. However, in most regions, $\hat{f}_{\mathrm{ABW}}$ has much larger bias than $\hat{f}_{\mathrm{EM2}}$ and so $\hat{f}_{\mathrm{EM2}}$ beats $\hat{f}_{\mathrm{ABW}}$ in terms of the asymptotic MSE.

Figure 2.1(b) illustrates the absolute values of asymptotic biases for the bimodal distribution defined in Marron and Wand (1992), also after rescaling the bias of $\hat{f}_{\mathrm{ABW}}$. The upper plot in Figure 2.1 (b) is the true density $f(x)$. The adaptive bandwidth density estimator, $\hat{f}_{\mathrm{ABW}}$, again has larger bias than the others except for a small region around the shoulders of both modes. The likelihood-tuned estimator, $\hat{f}_{\mathrm{EM2}}$, beats both $\hat{f}_{\mathrm{ABW}}$ and $\hat{f}_{\mathrm{MBC}}$ around the "valley" of f, while it is more biased around the two modes and the tails.

In both densities, the adaptive bandwidth density estimator has larger MSE than the others in most regions. Although the likelihood-tuned estimator is asymptotically worse than the multiplicative bias correction estimator in some places, it is known that asymptotic results may differ significantly from actual finite sample performance (Bowman and Foster, 1993) and so we need to study the methods further by simulation.

2.4 Simulation Comparison

In this section, we use simulated data to compare the performances of $\hat{f}_{\mathrm{KER}}(x)$, $\hat{f}_{\mathrm{ABW}}(x)$, $\hat{f}_{\mathrm{MBC}}(x)$ and $\hat{f}_{\mathrm{EM2}}(x)$. We consider the first eight Marron-Wand distributions (Marron and Wand, 1992), which are mixtures of Gaussian densities with various shapes, named 'Gaussian,' 'skewed unimodal,' 'strongly skewed,' 'kurtotic unimodal,' 'outlier,' 'bimodal,' 'separated bimodal,' and 'skewed bimodal.' In addition, four non-Gaussian distributions, namely Gamma$(2,1)$, Beta$(2,5)$, Beta$(2,2)$ and Beta$(1,3)$, were considered. Each density estimate was obtained on a grid of 301 points on [-3,3] for Gaussian mixtures, 350 points on [0,7] for Gamma$(2,5)$, and 100 points on [0,1] for beta distributions. For each distribution and each estimator, we calculated an *optimal bandwidth* by minimizing the average of integrated square errors (ISE), given by

$$ISE(\hat{f}) = \int \left\{\hat{f}(x) - f(x)\right\}^2 dx.$$

The ISE was calculated by numerical integration except that the ISE values for Gaussian mixture distributions with $n = 100$ were analytically calculated.

To investigate local performance, 500 random samples of size 500 were generated and the density was estimated by the four density estimators using the corresponding optimal bandwidths. Then, at each grid point of x, we found the mean of the squared errors (SEs) over 500 replicates. Figure 2.2 displays the bimodal density and the square root of mean SEs for this density. Here the mean SE of $\hat{f}_{\mathrm{EM2}}$ (thick solid line) was uniformly smaller than that of $\hat{f}_{\mathrm{KER}}$ (thin solid line). Although $\hat{f}_{\mathrm{ABW}}$ (dashed line)

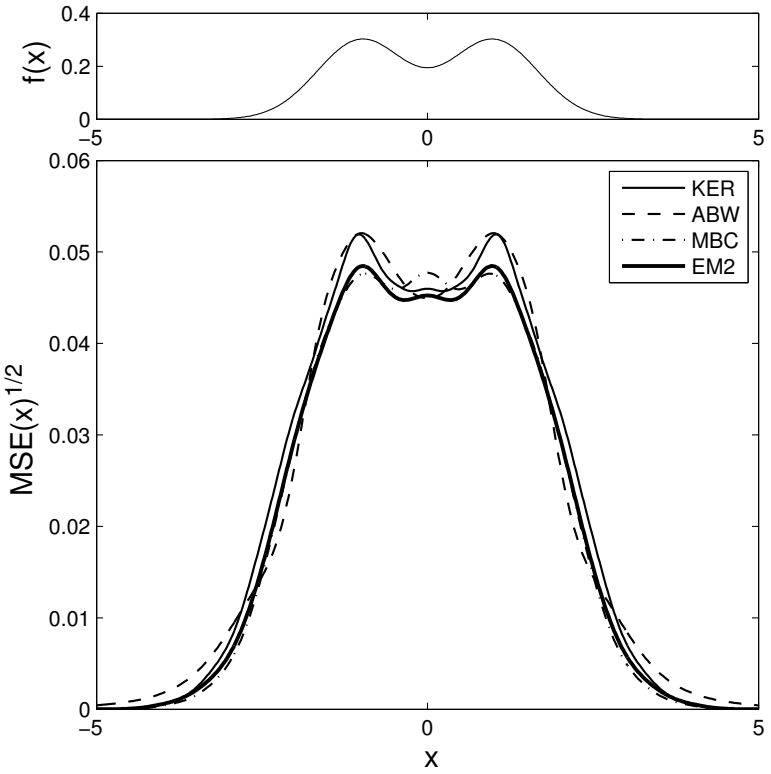


Figure 2.2: Bimodal density (top) and symmetrized mean squared errors for Bimodal with n=500 and replicates=500 (bottom).

behaved slightly better in the left and right shoulder, it clearly had higher mean SEs in both the peaks and tails. The most competitive estimator in this case was $\hat{f}_{\mathrm{MBC}}$ (dash-dotted line), which performed better in the tails and peaks. However, we can see that $\hat{f}_{\mathrm{EM2}}$ was still superior in the valley, where it also beat $\hat{f}_{\mathrm{MBC}}$ in asymptotic bias, as shown in Section 3.

Figure 2.3 shows the square root of the mean SE for Beta$(2, 5)$, one of the non-Gaussian distributions. The squared errors were calculated on $[-0.5, 1.5]$ to observe the behavior beyond the support of the true density, $[0, 1]$. In this case, $\hat{f}_{\mathrm{KER}}$ worked much better than in the bimodal case. In Figure 2.3, the thin solid line for $\hat{f}_{\mathrm{KER}}$ is found significantly below the others around the left shoulder. Though $\hat{f}_{\mathrm{EM2}}$ is located slightly above $\hat{f}_{\mathrm{KER}}$ in the left shoulder, it beats all of three competitors in the peak. In addition, $\hat{f}_{\mathrm{EM2}}$ was superior to $\hat{f}_{\mathrm{ABW}}$ and $\hat{f}_{\mathrm{MBC}}$ in most areas except where x was

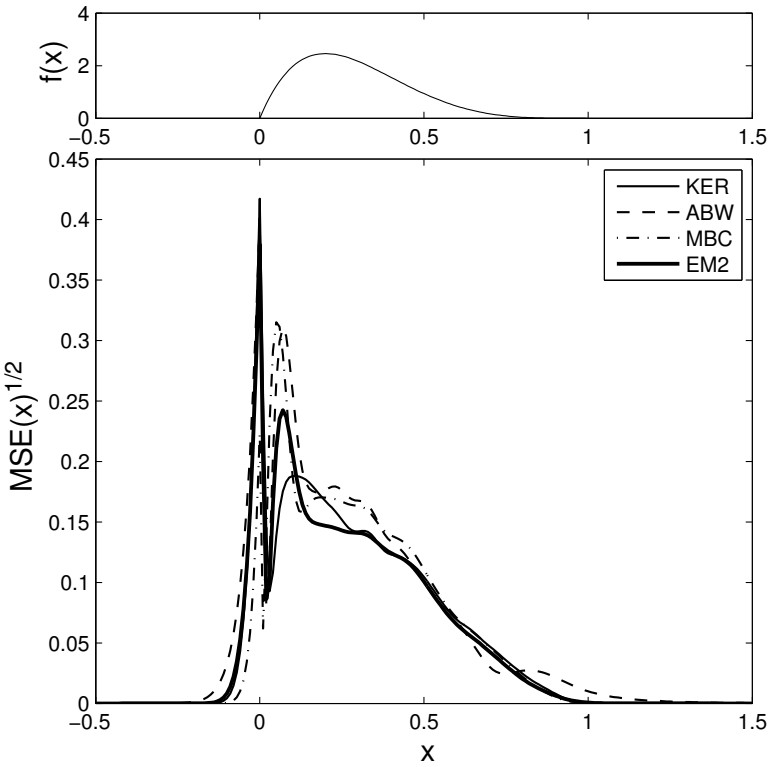


Figure 2.3: Beta$(2, 5)$ density (top) and mean squared errors for Beta$(2, 5)$ with $n = 500$ and 500 replicates (bottom).

negative. Outside of the support of f, $\hat{f}_{\mathrm{MBC}}$ had the smallest mean SE. However, it clearly lost to other estimators inside the support. For both the bimodal and Beta$(2, 5)$, we cannot conclude that the likelihood-tuned estimator was the best. However, it still performed very well for both, while the other estimators tended to work well for one or the other.

Table 2.1 provides a summary of simulation results over all our trial densities. For the comparison of global performances, 1000 random samples were generated by each f and we calculated density estimates by $\hat{f}_{\mathrm{KER}}$, $\hat{f}_{\mathrm{ABW}}$, $\hat{f}_{\mathrm{MBC}}$, and $\hat{f}_{\mathrm{EM2}}$ with corresponding optimal bandwidths. The ISE was calculated exactly or approximately as in the bandwidth selection, then we found the mean ISE for each density esitmator. The simulation was conducted with sample sizes $n = 100$ and $n = 500$, separately. Table 2.1 contains the ratio of the mean ISE for the basic kernel, the adaptive bandwidth and the

multiplicative bias correction estimator versus the likelihood-tuned estimator. Thus, a ratio greater than 1 implies that the likelihood-tuned density estimator was superior to the corresponding density estimator.

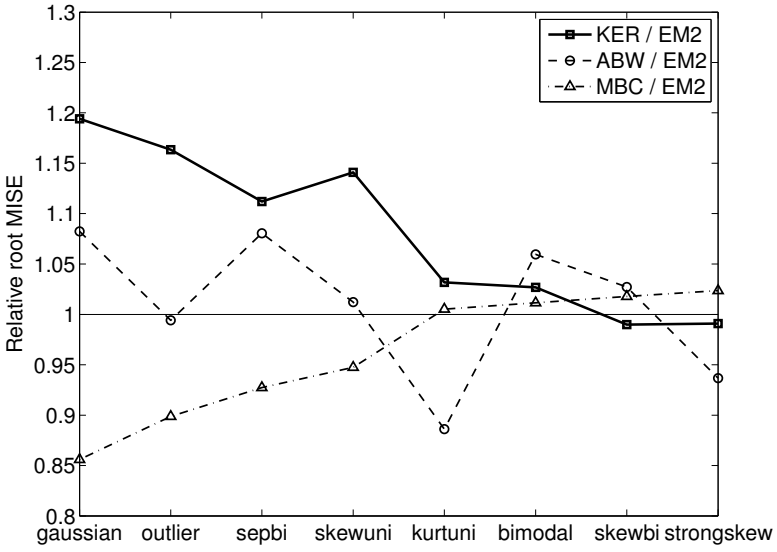
When $n = 100$, $\hat{f}_{\mathrm{EM2}}$ beat $\hat{f}_{\mathrm{KER}}$ for eight out of 12 densities, it beat $\hat{f}_{\mathrm{ABW}}$ for nine densities, and it beat $\hat{f}_{\mathrm{MBC}}$ for seven densities. When $n = 500$, for 10 of 12 densities, $\hat{f}_{\mathrm{EM2}}$ had the smaller mean ISE than $\hat{f}_{\mathrm{KER}}$ and $\hat{f}_{\mathrm{ABW}}$. However, $\hat{f}_{\mathrm{MBC}}$ outperformed $\hat{f}_{\mathrm{EM2}}$ for six out of 12 densities.

Figure 2.4 shows the relative root mean ISE of $\hat{f}_{\mathrm{KER}}$, $\hat{f}_{\mathrm{ABW}}$ and $\hat{f}_{\mathrm{MBC}}$ compared to $\hat{f}_{\mathrm{EM2}}$ for eight Gaussian mixture distributions at two sample sizes, $n = 100$ in plot (a) and $n = 500$ in plot (b). Since the mean ISE of $\hat{f}_{\mathrm{EM2}}$ is in the denominator, a relative root mean ISE greater than one implies that $\hat{f}_{\mathrm{EM2}}$ was superior to the competitor for the corresponding distribution in terms of the mean ISE. The thin solid line represents the ratio one. The horizontal axis represents eight Gaussian mixture distributions ordered so that the relative root mean ISE of $\hat{f}_{\mathrm{MBC}}$ was increasing, which turned out to be the most competitive density estimator for Gaussian mixture distributions.

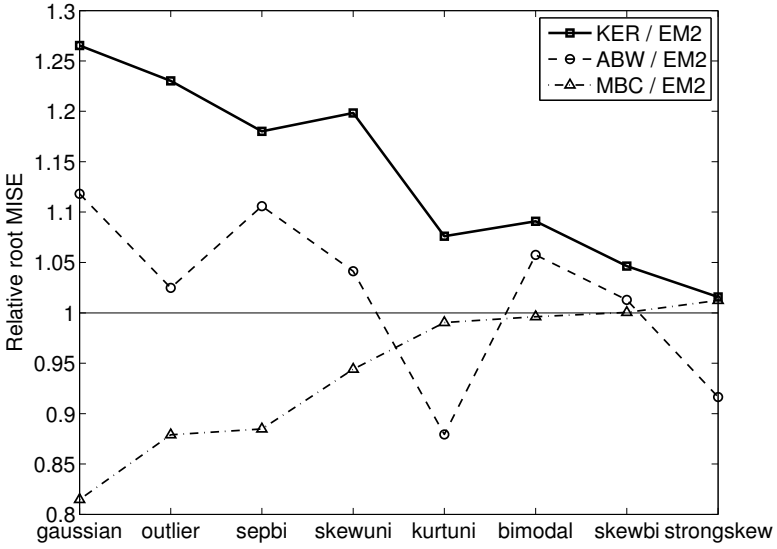
For the Gaussian mixture densities, $\hat{f}_{\mathrm{EM2}}$ was superior to $\hat{f}_{\mathrm{KER}}$ at both sample sizes

	f_{KER} vs. f_{EM2}		f_{ABW} vs. f_{EM2}		f_{MBC} vs. f_{EM2}	
	n=100	n=500	n=100	n=500	n=100	n=500
Gaussian	1.4259	1.6010	1.1716	1.2504	0.7327	0.6638
skewed unimodal	1.3019	1.4359	1.0245	1.0843	0.8976	0.8913
strongly skewed	0.9819	1.0316	0.8776	0.8397	1.0476	1.0247
kurtotic unimodal	1.0644	1.1581	0.7853	0.7731	1.0105	0.9806
outlier	1.3536	1.5135	0.9885	1.0501	0.8076	0.7724
bimodal	1.0543	1.1900	1.1227	1.1181	1.0231	0.9924
separated bimodal	1.2366	1.3925	1.1675	1.2230	0.8600	0.7827
skewed bimodal	0.9799	1.0952	1.0551	1.0258	1.0358	1.0008
$Gamma(2, 1)$	0.9836	0.9575	1.0944	1.1963	1.1966	1.2704
$Beta(2, 5)$	1.0686	1.0210	1.2214	1.3511	1.1548	1.2159
$Beta(2, 2)$	1.1649	1.0685	1.4961	1.6988	0.9035	1.3418
$Beta(1, 3)$	0.8735	0.8515	1.0057	1.0073	1.2134	1.2404

Table 2.1: Ratio of the mean ISE of $\hat{f}_{KER}$, $\hat{f}_{ABW}$ and $\hat{f}_{EM2}$ against $\hat{f}_{EM2}$ for sample sizes $n = 100$ and $n = 500$ from eight Gaussian mixture densities and four non-Gaussian densities over 1000 simulations.



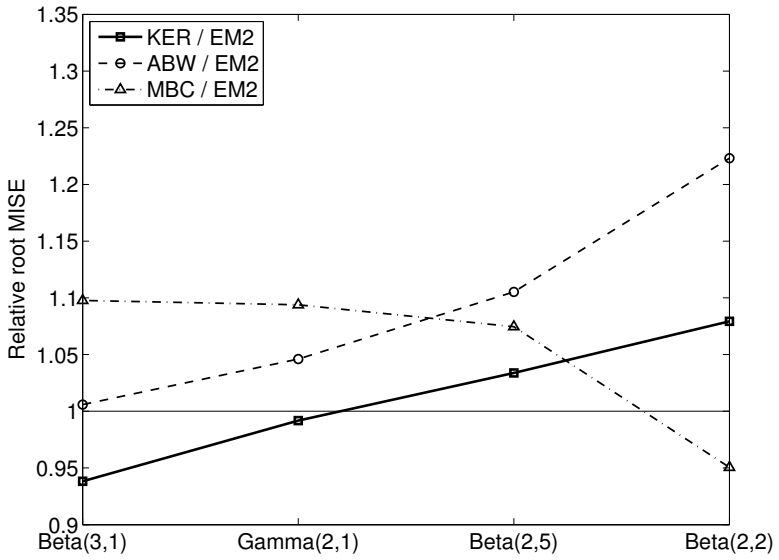
(a) $n = 100$



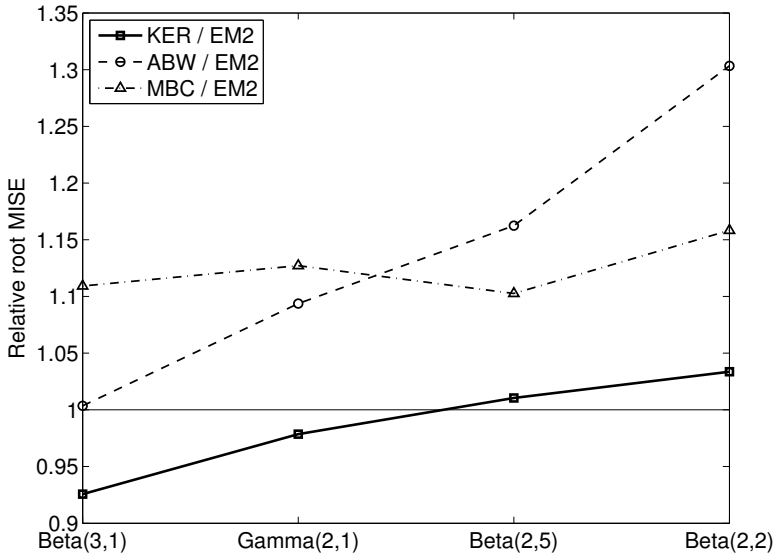
(b) $n = 500$

Figure 2.4: Relative root mean ISE for $\hat{f}_{\mathrm{KER}}$, $\hat{f}_{\mathrm{ABW}}$ and $\hat{f}_{\mathrm{MBC}}$ vs. $\hat{f}_{\mathrm{EM2}}$ for Gaussian mixture distributions. Dotted line represents the ratio 1.

while $\hat{f}_{\mathrm{MBC}}$ performed somewhat better than $\hat{f}_{\mathrm{EM2}}$ at both. When $n = 100$ in Figure 2.4 (a), one can observe that six out of eight points of $\hat{f}_{\mathrm{KER}}$ are above the thin solid line. In the case of $n = 500$ in Figure 2.4 (b), all points of $\hat{f}_{\mathrm{KER}}$ are greater than one. On the other hand, a half of relative root mean ISEs of $\hat{f}_{\mathrm{MBC}}$ were less than one when $n = 100$ and six out of eight were less than one when $n = 500$. Thus we can conclude that $\hat{f}_{\mathrm{MBC}}$



(a) $n = 100$



(b) $n = 500$

Figure 2.5: Relative root mean ISE for $\hat{f}_{\mathrm{KER}}$, $\hat{f}_{\mathrm{ABW}}$ and $\hat{f}_{\mathrm{MBC}}$ vs. $\hat{f}_{\mathrm{EM2}}$ for non-Gaussian distributions. Dotted line represents the ratio 1.

worked better than $\hat{f}_{\mathrm{EM2}}$ for Gaussian mixture distributions. Based on the trend, we expect that, as the sample size increases, the performance of $\hat{f}_{\mathrm{EM2}}$ and $\hat{f}_{\mathrm{MBC}}$ would be enhanced relative to $\hat{f}_{\mathrm{KER}}$ and $\hat{f}_{\mathrm{EM2}}$, respectively. The performance of $\hat{f}_{\mathrm{ABW}}$ relative to $\hat{f}_{\mathrm{EM2}}$ was mixed, but $\hat{f}_{\mathrm{EM2}}$ was superior in more cases.

Figure 2.5 includes relative root mean ISE for non-Gaussian distributions. In these

plots, the horizontal axis is in increasing order of the relative root mean ISE of $\hat{f}_{\mathrm{KER}}$, which performed best for the non-Gaussian distributions. For the non-Gaussian distributions, the result was opposite to the one for Gaussian mixture distributions. $\hat{f}_{\mathrm{KER}}$ performed better than $\hat{f}_{\mathrm{EM2}}$ while $\hat{f}_{\mathrm{MBC}}$ performed worse than $\hat{f}_{\mathrm{EM2}}$, and this pattern is clearest when the sample size is large. In Figures 2.5 (a) and (b), a half of relative root mean ISEs of $\hat{f}_{\mathrm{KER}}$ are below the thin solid line whereas all points of $\hat{f}_{\mathrm{MBC}}$ are above the line. For non-Gaussian distributions, $\hat{f}_{\mathrm{ABW}}$ was inferior to $\hat{f}_{\mathrm{EM2}}$. In Figure 2.5, all points are greater than one regardless of the sample size.

In conclusion, the two-step likelihood-tuned density estimator was not the best for the Gaussian mixture densities or for the non-Gaussian densities, but it was more robust than the other estimators. The adaptive bandwidth density estimator was worse than the likelihood-tuned for both the Gaussian mixtures and the non-Gaussians. The kernel density estimator performed well for the non-Gaussians while it worked poorly for the Gaussian mixtures. On the other hand, the multiplicative bias correction estimator beat all the others for the Gaussian mixtures while it did not perform well for the non-Gaussians. However, the likelihood-tuned density estimator showed a robust performance for both types of density.

2.5 Discussion

The likelihood-tuned density estimator is a two-step EM estimator, taking the uniform density as a prior for the mixing density $\pi(\phi)$. However, this procedure can be modified by applying a different prior density or proceeding with further tuning steps. By experiment, we found that using $N(\bar{X}, S^2 - h^2)$ as the initial Π_0 results in an estimator that is more robust against outliers than the two-step likelihood-tuned density estimator proposed here. We also suspect it would further debias the estimator in the Gaussian case. As mentioned in Section 2.2, further tuning steps could possibly improve asymptotic properties, but at a heavy computational cost.

The likelihood-tuning procedure and the resulting density estimator can be naturally extended to the multi-dimensional case. Compared to the adaptive bandwidth density estimator, the performance of the likelihood-tuned density estimator is expected to be even better than in the univariate case. This is because the adaptive bandwidth density estimator has a disadvantage in sparse areas, a situation made worse in multiple dimensions. Compared to the multiplicative bias correction estimator, we expect the likelihood-tuned estimator to be still more robust against the type of the true density.

Chapter 3

Multivariate Likelihood-tuned Density Estimator and Modal Inference

In this chapter, we extend the likelihood-tuned density estimator introduced in Chapter 2 to multivariate versions. We prove algorithmic convergence of the continuous EM algorithm, which the likelihood-tuning procedure is based on. The asymptotic and finite sample properties of the likelihood-tuned density estimators are compared to the multivariate version of the other density estimators that were compared in Chapter 2. In addition, we also examine their performance on mode identification.

3.1 Introduction

Let $\boldsymbol{X}_i \in \mathbb{R}^d,\ i = 1, \cdots, n$ be a random sample from $f(\boldsymbol{x})$. Then the multivariate kernel density estimator, denoted $\hat{f}_{KER}(\boldsymbol{x})$, is defined by

$$\hat{f}_{KER}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} K_H(\boldsymbol{x}, \boldsymbol{X}_i) \tag{3.1}$$

where $K_H(\boldsymbol{x}, \boldsymbol{y})$ can often be expressed as $|H|^{-1}K\left(H^{-1}(\boldsymbol{x}-\boldsymbol{y})\right)$ with a known density function $K(\cdot)$. When $H = hI$, $\hat{f}_{KER}(\boldsymbol{x})$ has bias of order $O\left(h^2\right)$ and variance of order $O\left(n^{-1}h^{-d}\right)$ under the true density $f(\boldsymbol{x})$. (Wand and Jones (1995)) These results require certain regularity conditions on f. Many authors have modified $\hat{f}_{KER}$ to reduce its bias, a key element of its mean squared error.

For univariate density estimation, Jones and Signorini (1997) compared the asymptotic and finite sample properties of six density estimators, all of which improve bias from $O(h^2)$ to $O(h^4)$. Among those six estimators, the adaptive bandwidth estimator of Abramson (1982) and the multiplicative bias correction of Jones et al. (1995) showed the best performance in terms of MISE.

Recently, Chung and Lindsay (2010) proposed a new nonparametric density estimator based on nonparametric maximum likelihood, which they called a likelihood-tuned density estimator. In univariate case, they viewed at the standard kernel density estimator as an estimator within a nonparametric mixture model and they used the EM algorithm to improve its likelihood. They compared the proposed density estimator with $\hat{f}_{KER}$ and the two methods found to be best in Jones and Signorini (1997) in the univariate case. The likelihood-tuned density estimator had the same order of bias and variance with the two improved density estimators. Simulation study showed that the likelihood-tuned density estimator behaved more robustly in efficiency against different types of distribution. This section extends the likelihood-tuned density estimator to the multivariate case and compares it with the multivariate version of the same bias-reduced density estimators that were used in Chung and Lindsay (2010).

The adaptive bandwidth (ABW) estimator of Abramson (1982) can be extended to the multivariate case in a natural way by letting H be $H_i = H(\boldsymbol{x}_i)$, a local bandwidth matrix. Breiman, Meisel, and Purcell (1977) proposed to replace H in (3.1) by $h_i I = h \cdot d_{i,k} I$ where $d_{i,k}$ is the distance from the point $\boldsymbol{x}_i$ to its kth nearest neighbor. Abramson

(1982) suggested taking $h(\boldsymbol{X}_i)$ proportional to $f(\boldsymbol{X}_i)^{-1/2}$ as follows

$$\tilde{f}_{ABW}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} K_{hf(\boldsymbol{X}_i)^{-1/2}I}(\boldsymbol{x}, \boldsymbol{X}_i), \tag{3.2}$$

and showed that it reduced the bias to $O(h^4)$. Since this estimator depends on the unknown f, Hall and Marron (1988) suggested to replace $f(\boldsymbol{X}_i)$ with a pilot estimator $\hat{f}_{KER}(\boldsymbol{X}_i)$ constructed using a bandwidth matrix $h_0 I$, say $\hat{f}_{KER,h_0}$, thereby introducing an additional parameter h_0. They showed that the adaptive bandwidth estimator with a bandwidth matrix $h \cdot \hat{f}_{KER,h_0}(\boldsymbol{X}_i)^{-\frac{1}{2}} I$ had the same rate of convergence to the ideal estimator as $\hat{f}_{ABW}$ (3.2) but the constant coefficient was worse than (3.2). They also verified that $\tilde{f}_{ABW}(\boldsymbol{x})$ reduced bias to $O(h^4)$ and had variance of order $O(n^{-1}h^{-d})$.

In this paper, we compare our new density estimator with the density estimator

$$\hat{f}_{ABW}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} K_{h\hat{f}_{KER,h}(\boldsymbol{X}_i)I}(\boldsymbol{x}, \boldsymbol{X}_i), \tag{3.3}$$

which replaces $f(\boldsymbol{X}_i)$ with $\hat{f}_{KER,h}(\boldsymbol{X}_i)$ in (3.3). That is, we restrict the bandwidth for the pilot estimator, h_0, to be the same as the bandwidth for the density estimator itself, h. Taking an additional bandwidth parameter causes a problem of choosing one more tuning parameter. In addition, Jones and Signorini (1997) mentioned that this choice of bandwidth was reasonable and simulation results showed that its performance was as good as the two-bandwidth version in the case where the latter worked well.

Jones, Linton, and Nielsen (1995) also proposed a multivariate version of their multiplicative bias correction (MBC) estimator, which was the other winner in Jones and Signorini (1997), given by

$$\hat{f}_{MBC}(\boldsymbol{x}) = \hat{f}_{KER}(\boldsymbol{x}) \frac{1}{n} \sum_{i=1}^{n} \frac{K_H(\boldsymbol{x}, \boldsymbol{X}_i)}{\hat{f}_{KER}(\boldsymbol{X}_i)}. \tag{3.4}$$

They showed that it had a bias of order $O(h^4)$ and a variance of order $O(n^{-1}h^{-d})$ when $H = hI$.

In this chapter, we propose a multivariate version of the likelihood-tuned density estimator. In section 3.2, we extend the likelihood-tuning procedure in Chung and Lindsay (2010) to the multivariate case, and so the resulting density estimators are introduced. Section 3.3 includes theorems and proofs of the EM convergence. In Section 3.4, the asymptotic properties for the multivariate likelihood-tuned density estimator and the optimal bandwidth are investigated. Simulation comparisons for the mean integrated squared errors are in Section 3.5. In Section 3.6, we compare the performance of the density estimators on identifying modes of two-component mixture models.

3.2 Methodology

3.2.1 Background

Consider the nonparametric mixture model with a mixing (latent) distribution $\Pi(\phi)$, given by

$$f(\boldsymbol{x};\Pi) = \int K(\boldsymbol{x},\boldsymbol{\phi})d\Pi(\boldsymbol{\phi}) \tag{3.5}$$

where $K(\cdot,\cdot)$ is a known density function, called a component density and $\boldsymbol{\phi}$ represents a component parameter. Here the distribution function $\Pi(\boldsymbol{\phi})$ is allowed to be discrete, continuous or any specific family of distributions. If Π is discrete with point masses π_i at $\phi_i,\ i = 1,\cdots,m$, then (3.5) becomes the m-component finite mixture model, written as

$$f(x;\boldsymbol{\phi}_1,\cdots\boldsymbol{\phi}_m,\pi_1,\cdots,\pi_m) = \sum_{i=1}^{m}\pi_i K(\boldsymbol{x},\boldsymbol{\phi}_i). \tag{3.6}$$

Let $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n$ be a random sample from $f(\boldsymbol{x};\Pi)$. If Π is estimated by $\hat{\Pi}(\boldsymbol{\phi}) = n^{-1}$ at $\boldsymbol{\phi} = \boldsymbol{X}_i,\ i = 1,\cdots,n$ and 0 elsewhere, (3.5) can be written as

$$f(\boldsymbol{x};\hat{\Pi}) = n^{-1}\sum_{i=1}^{n}K(\boldsymbol{x},\boldsymbol{X}_i)$$

and this is the same as the basic kernel density estimator in (3.1). This view of the

kernel density estimator in the context of the nonparametric mixture models leads to the idea that an improved estimator of Π could provide a better estimator of $f(\boldsymbol{x})$.

Consider the likelihood function of an observation, given by

$$L_i(\Pi) = \int K(\boldsymbol{X}_i, \boldsymbol{\phi}) d\Pi(\boldsymbol{\phi})$$

and the log-likelihood function, given by

$$l(\Pi) = \sum_{i=1}^{n} \ln(L_i(\Pi)). \tag{3.7}$$

For simplicity, we assume that all individual likelihoods L_i's are distinct. In case there exist identical L_i's, the log-likelihood (3.7) can be expressed in a more general form which Lindsay (1995) used, given by

$$l(\Pi) = \sum_{i=1}^{D} n(i) \ln(L_i(\Pi)), \tag{3.8}$$

where $\sum_{i=1}^{D} n(i) = n$.

Here we do not specify any parametric form of Π, allowing it to be either discrete or continuous. The estimator $\hat{\Pi}$ that maximizes (3.7) is called the nonparametric maximum likelihood estimator (NPMLE) of Π (Lindsay (1995)). Lindsay (1995) proved that if the L_i's are bounded and all distinct, then there exists an NPMLE, $\hat{\Pi}$, that is a discrete distribution with no more than n distinct points of support. Although $\hat{\pi}$ is not necessarily unique, the fitted likelihood vector $\hat{\boldsymbol{L}} = (L_1(\hat{\pi}), \ldots, \hat{L}_D(\hat{\pi}))^T$ is unique.

Since the NPMLE $\hat{\Pi}$ is discrete on finitely many support points, one algorithmic approach to finding $\hat{\Pi}$ is to apply the EM algorithm with $\Pi(\boldsymbol{\phi})$ restricted to be a discrete distribution on a large number of support points (Laird (1978), Vardi et al. (1985)).

On the other hand, we will here consider the generalized EM algorithm that allows the initial $\Pi(\boldsymbol{\phi})$ to be continuous with a density function $\pi(\boldsymbol{\phi})$. Ideally, this EM would still converge to the discrete NPMLE. Vardi and Lee (1993) and Lindsay (1995) described

the continuous EM algorithm for the NPMLE, given by

$$\hat{\pi}_{(k+1)}(\boldsymbol{\phi}) = \hat{\pi}_{(k)}(\boldsymbol{\phi}) \cdot n^{-1} \sum_{i=1}^{n} \frac{K(\boldsymbol{X}_i, \boldsymbol{\phi})}{L_i(\hat{\Pi}_{(k)})}. \tag{3.9}$$

In this paper, we consider the consequences of using a continuous uniform density, specifically $\hat{\pi}_{(0)}(\boldsymbol{\phi}) = 1$ as the initial estimate of $\hat{\pi}$. Each update by (3.9) will produce a spikier estimate of π that assigns more weights to the support points of the discrete NPMLE. The first update by the EM gives the basic kernel density estimator and the second update generates our new density estimator.

Although the second-step estimator of π is still not the NPMLE, it does provide a density estimator that is asymptotically superior to the kernel density estimator in many cases. It also shows better overall performance in our simulation study than other modified kernel density estimators.

3.2.2 Likelihood-tuning Procedure

Prior to introducing the likelihood-tuning procedure, we define a gradient function of the log-likelihood function $l(\Pi)$ in (3.7) at Π_0 toward the direction of ϕ by

$$D_{\Pi_0}(\boldsymbol{\phi}) = \sum_{i=1}^{D} \left(\frac{K(\boldsymbol{X}_i, \boldsymbol{\phi})}{L_i(\Pi_0)} - 1 \right)$$

(Lindsay (1995)). This is derived from the directional derivative of $l(\Pi)$ at Π_0 along the path between Π_0 and Π_1, that is

$$D_{\Pi_0}(\Pi_1) = \frac{\partial}{\partial \alpha} l(\alpha\Pi_0 + (1-\alpha)\Pi_1)|_{\alpha=0}.$$

When Π_1 is a degenerate distribution at ϕ, $D_{\Pi_0}(\Pi_1)$ becomes $D_{\Pi_0}(\boldsymbol{\phi})$. Thus if $D_{\Pi_0}(\boldsymbol{\phi})$ is positive, $l(\Pi_0)$ will be increased by adding some small mass at ϕ in the mixing distribution Π_0. On the contrary, if $D_{\Pi_0}(\boldsymbol{\phi})$ is negative for $\boldsymbol{\phi}$, $l(\Pi_0)$ will be increased by shrinking the mass at $\boldsymbol{\phi}$ in Π_0, unless the mass is already zero.

Using the gradient function, the continuous EM algorithm in (3.9) can be written as

$$\hat{\pi}_{(k+1)}(\boldsymbol{\phi}) = \hat{\pi}_{(k)}(\boldsymbol{\phi}) \left(1 + n^{-1} D_{\hat{\Pi}_{(k)}}(\boldsymbol{\phi})\right). \tag{3.10}$$

This equation implies that the EM algorithm increases the density at $\boldsymbol{\phi}$ where the gradient function is positive and reduces the density where the gradient function is negative. We note that $\hat{\pi}_{(k+1)}(\boldsymbol{\phi})$ integrates to 1 if $\hat{\pi}_{(k)}(\boldsymbol{\phi})$ does.

The likelihood-tuning procedure that updates a density estimator of $\boldsymbol{x}$ includes two steps: updating the mixing density π and updating the density estimator of $\boldsymbol{x}$. Given an initial estimate $\pi_0(\boldsymbol{\phi})$, the *likelihood-tuning procedure* is as follow.

1. Update the estimator of the mixing density π by

$$\hat{\pi}_{(k+1)}(\boldsymbol{\phi}) = \hat{\pi}_{(k)}(\boldsymbol{\phi}) \Delta_{(k)}(\boldsymbol{\phi}),$$

   where $\Delta_{(k)}(\boldsymbol{\phi}) = 1 + n^{-1} D_{\hat{\Pi}_{(k)}}(\boldsymbol{\phi})$.

2. Update the density estimator of x by

$$\hat{f}_{(k+1)}(\boldsymbol{x}) = \int K(x, \boldsymbol{\phi}) \hat{\pi}_{(k+1)}(\boldsymbol{\phi}) \, d\boldsymbol{\phi}.$$

Notice that it is desirable to avoid numerical integration in step 2.

We will take the continuous uniform density $\hat{\pi}_0(\boldsymbol{\phi}) = 1$ to be the initial estimate of π given a lack of any prior information about the location of peaks. Starting from a flat density, a gradient function tends to identify the needed points of support in the NPMLE. In each step thereafter, $\Delta_{(k)}(\boldsymbol{\phi})$ indicates the deviation of the latent density $\hat{\pi}_{(k+1)}(\boldsymbol{\phi})$ from $\hat{\pi}_{(k)}(\boldsymbol{\phi})$, which therefore increases mass at the highest gradient values.

In fact, one can show that repeated application of the continuous EM algorithm converges to NPMLE without the parameter space searches ("gradient checks") that are required when one use discrete Π estimators (Lindsay (1995)). (See Section 3.3.)

3.2.3 Application to Diffusion Kernels

If the two steps in the likelihood-tuning procedure can be calculated analytically, one can express these steps in closed forms and obtain general formula for $\hat{f}_{(k)}$ and $\hat{\pi}_{(k)}$. In this section, we apply the likelihood-tuning steps to the nonparametric mixture model in a specific family of kernels.

Consider a family of kernels $K_t(\cdot,\cdot)$ with a bandwidth matrix H, that satisfy

$$\int K_{t_1}(x,y)K_{t_2}(y,z)dz = K_{t_1+t_2}(x,z). \tag{3.11}$$

Here we call a kernel with the above property as a *Markov diffusion kernel*. This is related to a density of particles in Brownian motion. Starting from a location x, a density of a location z in time t_1+t_2, $K_{t_1+t_2}(x,z)$, can be expressed as an integral of the product of two transition densities; $K_{t_1}(x,y)$ and $K_{t_2}(y,z)$. Notice that this equation is satisfied by the normal (Gaussian) kernel, one of the most popular kernels in nonparametric density estimation. See Chen and Lindsay (2006) and Yang (2004) for two other examples of Markov diffusion kernels.

When the kernel is a Markov diffusion kernel, the estimator of $\pi(\boldsymbol{\phi})$ and $f(\boldsymbol{x})$ by the likelihood-tuning procedure are reduced into explicit forms. The Gaussian kernel function is defined by $K_h(\boldsymbol{x},\boldsymbol{\phi}) = |H|^{-1}\varphi\left\{H^{-1}(\boldsymbol{x}-\boldsymbol{\phi})\right\}$, where φ is the standard normal density function. If one uses the initial $\pi_0(\boldsymbol{\phi}) = 1$, an improper prior, then the first likelihood-tuning iteration provides

$$\hat{\pi}_{(1)}(\boldsymbol{\phi}) = n^{-1}\sum_{i=1}^{n} K_H(\boldsymbol{\phi}, X_i)$$

and

$$\hat{f}_{(1)}(\boldsymbol{x}) \;=\; \int K_H(\boldsymbol{x},\boldsymbol{\phi})\cdot n^{-1}\sum_{i=1}^{n} K_H(\boldsymbol{\phi},\boldsymbol{X}_i)\; d\boldsymbol{\phi}$$

$$= \quad n^{-1}\sum_{i=1}^{n} K_{\sqrt{2}H}(\boldsymbol{x}, \boldsymbol{X}_i). \tag{3.12}$$

Notice that the estimate $\hat{\pi}_{(1)}(\boldsymbol{\phi})$ from a single likelihood-tuning step turns out to be the basic kernel density estimator. The resulting fitted density for $\boldsymbol{x}$, $\hat{f}_{(1)}$, is also the basic kernel density estimator but with a wider bandwidth than the bandwidth in $\hat{\pi}_{(1)}$. Here this kernel density estimator obtained by the first likelihood-tuning step in (3.12) is denoted by $\hat{f}_{EM1}$.

If we apply the likelihood-tuning procedure once more, we obtain

$$\hat{\pi}_{(2)}(\boldsymbol{\phi}) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} K_{\frac{1}{\sqrt{2}}H}(\boldsymbol{\phi}, \bar{\boldsymbol{X}}_{ij})$$

where $w_{ij} = K_{\sqrt{2}H}(\boldsymbol{X}_i, \boldsymbol{X}_j)/\hat{f}_1(\boldsymbol{X}_i)$ and $\bar{\boldsymbol{X}}_{ij} = (\boldsymbol{X}_i + \boldsymbol{X}_j)/2$. The second step density estimator is given by

$$\begin{aligned} \hat{f}_{(2)}(\boldsymbol{x}) &= n^{-2}\sum_i\sum_j w_{ij}\int K_H(\boldsymbol{x}, \phi) K_{\frac{1}{\sqrt{2}}H}(\boldsymbol{\phi}, \bar{\boldsymbol{X}}_{ij}) d\boldsymbol{\phi} \\ &= n^{-2}\sum_i\sum_j w_{ij} K_{\sqrt{\frac{3}{2}}H}(\boldsymbol{x}, \bar{\boldsymbol{X}}_{ij}). \end{aligned} \tag{3.13}$$

We define this second step density estimator in (3.13) as *the (two step) likelihood-tuned density estimator*, and denote it by $\hat{f}_{EM2}(\boldsymbol{x})$.

Applying the likelihood-tuning procedure repeatedly, one can take more EM steps to move $\hat{\pi}$ even closer to the NPMLE than $\hat{\pi}_{(2)}$. With $\pi^{(0)}(\phi) = 1$, for any positive integer k, $\pi_{(k)}$ is generalized by

$$\hat{\pi}_{(k)}(\phi) = \frac{1}{n^k}\sum_{i_1=1}^{n}\cdots\sum_{i_k=1}^{n} w_{i_1,\cdots,i_{k-1}} K_{\frac{1}{\sqrt{k}}H}(\phi, \bar{X}_{i_1,\cdots,i_k}) \tag{3.14}$$

where $w_{i_1,\cdots,i_{k-1}} = c(h)\cdot\frac{K_{\sqrt{k}H}(X_{i_1},X_{i_2})\cdots K_{\sqrt{k}H}(X_{i_{k-1}},X_{i_k})}{\hat{f}_{(1)}(X_{i_1})\cdots\hat{f}_{(k-1)}(X_{i_{k-1}})}$ with known constant $c(h)$ and $\bar{X}_{i_1,\cdots,i_k}$ is the mean of $(X_{i_1},\cdots,X_{i_k})$.

The updated density estimator can be expressed in the generalized form

$$\hat{f}_{(k)}(x) = \frac{1}{n^k} \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} w_{i_1,\cdots,i_{k-1}} K_{\sqrt{\frac{k+1}{k}}H}(x, \bar{X}_{i_1,\cdots,i_k}). \tag{3.15}$$

Although the density estimator in (3.15) with $k \geq 3$ might have more desirable asymptotic properties than the second step estimator, it requires intensive computations. In fact, when one move from the $(k-1)$-th step estimator to the k-th step, the number of summands is increased from n^{k-1} to n^k. Without proceeding further, therefore, this paper focuses on the second step estimator, $\hat{f}_{EM2}(x)$, and will show it has desirable asymptotic properties and promising simulation results in Section 3.4 and Section 3.5.

Remark 3.1. If one started from another $\hat{\pi}_{(0)}(\phi)$, for example, Gaussian, one can obtain different likelihood-tuned estimators. In univariate case, let $\hat{\pi}_{(0)}$ be a Gaussian density with mean $\bar{X}$ and variance $S^2 - h^2$, where $\bar{X}$ is the sample mean and S^2 is the sample variance multiplied by $(n-1)/n$. With this initial $\hat{\pi}_{(0)}$, an initial density estimator $\hat{f}_{(0)}(x)$ becomes the probability density function of $N(\bar{X}, S^2)$, which attains the maximum likelihood in the Gaussian model. Thus this initial $\hat{\pi}_{(0)}$ let us start from the ideal case for Gaussian distribution.

With $\hat{\pi}_{(0)}(\phi) = (S^2 - h^2)^{-1/2} \varphi\left\{(S^2 - h^2)^{-1/2}(\phi - \bar{X})\right\}$, the first likelihood-tuning step gives

$$\tilde{f}_{(1)}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{h\sqrt{2-\frac{h^2}{S^2}}} \left(x, \frac{h^2}{S^2}\bar{X} + \left(1 - \frac{h^2}{S^2}\right) X_i \right) \tag{3.16}$$

and the second likelihood-tuning step gives

$$\tilde{f}_{(2)}(x) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} K_{h\sqrt{\frac{3-2h^2/S^2}{2-h^2/S^2}}} \left(x, \frac{\frac{h^2}{S^2}\bar{X} + 2\left(1 - \frac{h^2}{S^2}\right)\bar{X}_{ij}}{2 - \frac{h^2}{S^2}} \right) \tag{3.17}$$

where $w_{ij} = K_{h\sqrt{2-h^2/S^2}} \left(\frac{h^2}{S^2}\bar{X} + \left(1 - \frac{h^2}{S^2}\right) X_i, X_j \right) / f_1(X_j)$.

When $h = 0$, $\tilde{f}_{(1)}$ and $\tilde{f}_{(2)}$ become the empirical distribution. Thus a random variable following either of them will have a mean $\bar{X}$ and variance S^2. On the other hand, when

$h = S^2$, both of them turn out to be the density function of $N(\bar{X}, S^2)$. Therefore, there does not exist any variance inflation due to the bandwidth h. Note that a random variable following the kernel density estimator with the bandwidth h has variance S^2+h^2 (see Jones (1991).)

In addition, the first likelihood-tuned estimator (3.16) can be viewed as a kernel density estimator with moving data points toward the sample mean. When h is large, the estimators in (3.16) and (3.17) tend to shrink more toward the sample mean $\bar{X}$, so this would work more robustly again outliers. This is a similar idea to Samiuddin and El-Sayyad (1990) while they adjusted data points in the direction of positive slope to tighten peaks.

3.3 Algorithmic Convergence

In Section 3.2.3, we obtained a general form of $\hat{\pi}_{(k)}$ in (3.14), which was calculated from the continuous EM algorithm in (3.10). In this section, we demonstrate that $\hat{\pi}^{(k)}$ converges to the NPMLE in a suitable sense.

Convergence of the EM algorithm was proved in Wu (1983) in a standard complete-incomplete data structure when the parameter space is in $\mathbb{R}^p$. Under some regularity conditions, he proved an EM sequence of likelihood values converges to some stationary points. Then with more strict conditions, he showed that an EM sequence of parameter updates converges to a fixed point.

Wu's proof is for the EM algorithm on p-dimensional parameter vector. On the other hand, the continuous EM algorithm is on a functional parameter space consisting of mixing densities $\pi(\cdot)$. Here we prove the continuous counterparts of the theorems in Wu (1983). At the beginning, we simplify the problem by considering discrete mixtures on fixed grid of points. Then the results will be extended to a continuous π.

Let ϕ be in a finite or infinite grid of ϕ-values $\{\phi_1, \phi_2, \ldots\}$ and consider a model

$$f(x) = \sum_j \pi_j K(x; \phi_j), \tag{3.18}$$

where $\sum_j \pi_j = 1$ and random sample $X_i, \ldots, X_n$ are from f. Assume that the individual kernels of the likelihood $K(x_i, \phi)$ are both nonnegative and bounded as functions of ϕ. With this model, the likelihood can be written by $L(\boldsymbol{\pi}) = \prod_{i=1}^n L_i(\boldsymbol{\pi})$, where $L_i(\boldsymbol{\pi}) = \sum_j \pi_j K(x_i, \phi_j)$ and $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)^T$. In this model, a maximum likelihood estimator of $\boldsymbol{\pi}$ is a finite or infinite length vector $\hat{\boldsymbol{\pi}}$ that satisfies $\hat{\boldsymbol{\pi}} = \arg\max_{\boldsymbol{\pi}} L(\boldsymbol{\pi})$. Note that this $\hat{\boldsymbol{\pi}}$ does not have to be unique, although once again $\hat{\boldsymbol{L}}$ is so. The continuous EM algorithm in (3.10) can be applied here for discrete π when the support points $\{\phi_1, \phi_2, \ldots\}$ are predetermined, which is given by, for each j,

$$\pi_j^{(k+1)} \quad = \quad \pi_j^{(k)} \left(\frac{1}{n} D_{\pi^{(k)}}(\phi_j) + 1 \right). \tag{3.19}$$

Note that we can write this equation as $\pi^{(k+1)} = G(\pi^{(k)})$ for a function G. In the following, we will need the following definition. Say that π^* is a *fixed point* of the algorithm if application of the algorithm in (3.19) to $\pi^{(k)} = \pi^*$ yields $\pi^{(k+1)} = \pi^*$. That is, $G(\pi^*) = \pi^*$. Note that for each $\pi^{(k)}$, we can define $\boldsymbol{L}^{(k)} = (L_1(\pi^{(k)}), \ldots, L_D(\pi^{(k)}))^T$ and that each iteration of the algorithm corresponds to an EM map $H(\boldsymbol{L}^{(k)}) = \boldsymbol{L}^{(k+1)}$. We will focus on the fixed points of this map, with $H(\boldsymbol{L}^*) = \boldsymbol{L}^*$.

Lemma 3.2. *For each j, if $\pi_j^{(0)} > 0$, then $\pi_j^{(k)} > 0$ for all iterations.*

Proof. For each j, suppose $\pi_j^{(0)} > 0$ and $\pi_j^{(k)} > 0$. Then the $(k+1)$-th EM update of π_j is

$$\pi_j^{(k+1)} \quad = \quad \pi_j^{(k)} \cdot \frac{1}{n} \sum_{i=1}^n \frac{K(x_i; \phi_j)}{L_i(\pi^{(k)})}.$$

Since $K(x_i; \phi_j)$'s are assumed to be positive, with the induction assumption, we have

$\pi_j^{(k+1)} > 0$. By mathematical induction, we conclude that $\pi_j^{(k)} > 0$ for each j and for all iterations. □

Lemma 3.2 guarantees that the weights $\pi_j^{(k)}$'s can never be exactly zero when we start them at nonzero values. This is important in the convergence theory because it is clear from (3.19) that if one starts with $\pi_j^{(0)} = 0$ for some j, then $\pi_j^{(k)} = 0$ for all k, making it clear that the algorithm could not possibly converge to the NPMLE if the NPMLE has $\hat{\pi}_j > 0$. In fact, the EM algorithm has many fixed points in the model, corresponding to the NPMLEs for the various sub models in which some of the π's are constrained to be zero.

Lemma 3.3. *The convergence of $L_i(\boldsymbol{\pi}^{(k)})$ to L_i^* means the gradient functions converge. That is,*

$$D^{(k)}(\phi) \to D^*(\phi) = \sum_{i=1}^{n} \left(\frac{K(x_i;\phi)}{L_i^*} - 1 \right)$$

Proof. For $i = 1, \ldots, n$, suppose $L_i(\boldsymbol{\pi}^{(k)})$ converges to L_i^* for each i. Then $L_i(\boldsymbol{\pi}^{(k)})^{-1}$ also converges to $(L_i^*)^{-1}$ pointwisely.

$$\begin{aligned} \left| D^{(k)}(\phi) - D^*(\phi) \right| &= \sum_{i=1}^{n} K(x_i;\phi) \left| L_i(\boldsymbol{\pi}^{(k)})^{-1} - (L_i^*)^{-1} \right| \\ &\leq n\, M \max_i \left| L_i(\boldsymbol{\pi}^{(k)})^{-1} - (L_i^*)^{-1} \right| \to 0 \end{aligned}$$

where M is a bound for $K(x;\phi)$. □

Theorem 3.4. *Suppose that the initial value of $\boldsymbol{\pi}$, $\boldsymbol{\pi}^{(0)} = (\pi_1^{(0)}, \pi_2^{(0)}, \ldots)^T$, satisfies $\pi_j^{(0)} > 0$ and that the sequence of iterates $\{L_i(\boldsymbol{\pi}^{(k)})\}$ are convergent with limit points L_i^*. Then, for each i, $L_i^* = L_i(\hat{\boldsymbol{\pi}})$, so the EM algorithm maximizes the likelihood.*

Proof. The gradient inequality says that L_i^* is the maximal likelihood vector if and only if $D^*(\phi) \leq 0$ (Lindsay (1995).) Suppose the gradient inequality does not hold for L_i^*, so $D^*(\phi_r) > 0$ for some ϕ_r. Then Lemma 3.3 implies $D^{(k)}(\phi_r) \to D^*(\phi_r) > 0$ and it follows

that there exists $K > 0$ such that, for all $k > K$,

$$D^{(k)}(\phi_r) \geq \frac{1}{2} D^*(\phi_r) \equiv \delta > 0.$$

Then, as $m \to \infty$, we have

$$\begin{aligned} \pi_r^{(K+m+1)} &= \pi_r^{(K)} \left(1 + D^{(K+1)}(\phi_r)\right) \left(1 + D^{(K+2)}(\phi_r)\right) \ldots \left(1 + D^{(K+m)}(\phi_r)\right) \\ &\geq \pi_r^{(K)} (1+\delta)^m. \end{aligned} \tag{3.20}$$

Since Lemma 3.2 implies that $\pi_r^{(K)} > 0$, the right-hand side of (3.20) goes to ∞. Thus we have a contradiction to $\sum_j \pi_j = 1$. □

Theorem 3.4 shows that, for each i, if the likelihoods convey $\{L_i(\boldsymbol{\pi}^{(k)})\}$ to a limit L_i^*, the limit point must yield the maximum of the likelihood function $L_i(\boldsymbol{\pi})$. However, how do we know that $\{L_i(\boldsymbol{\pi}^{(k)})\}$ converges to a limit? We here borrow from Wu (1983) a simple sufficient condition to this to occur.

Remark 3.5. Note that this convergence is stronger than that of Wu (1983) in the sense that we are guaranteed to converge to the global maximum.

Theorem 3.6. *Suppose that the EM algorithm has a finite set of fixed points, each with a different likelihood. Then the EM algorithm converges to exactly one of the fixed points. (That is, it cannot oscillate between fixed points.)*

Proof. Suppose that a subsequence $\{L_i(\boldsymbol{\pi}^{(k')})\}$ of iterates $\{L_i(\boldsymbol{\pi}^{(k)})\}$ converges to a fixed point L_i^*, with likelihood $L^* = \prod_i L_i^*$. Since the EM algorithm has a property $L(\boldsymbol{\pi}^{(k+1)}) > L(\boldsymbol{\pi}^{(k)})$, the entire sequence of likelihood $\{L(\boldsymbol{\pi}^{(k)})\}$ converges to L^*. It follows that no subsequence of iterates can converge to a different fixed point, as they would then have different limiting likelihood. □

We note here the possible challenge one faces when there are two fixed points with the same likelihoods. One could, in theory, have a situation where the odd steps of the

algorithm $\boldsymbol{\pi}^{(2k+1)}$ converge to one fixed point $\boldsymbol{L}^*$ and the even steps $\boldsymbol{\pi}^{(2k)}$ converge to the other $\boldsymbol{L}^{**}$, all without violating likelihood monotonicity.

One can construct more general convergence proofs, but they take a bit more work. One method is to show that $G(\boldsymbol{\pi})$ (or $H(\boldsymbol{L})$) is a contraction operator in the neighborhood of its fixed points $\boldsymbol{\pi}^*$. Here is the argument: a Taylor expansion of G around $\boldsymbol{\pi}^*$ gives, for $\dot{G} = \nabla G(\boldsymbol{\pi}^*)$,

$$G(\boldsymbol{\pi}) = G(\boldsymbol{\pi}^*) + \dot{G}(\boldsymbol{\pi} - \boldsymbol{\pi}^*),$$

so that locally

$$\boldsymbol{\pi}^{(k+1)} \doteq \boldsymbol{\pi}^* + \dot{G}(\boldsymbol{\pi}^{(k)} - \boldsymbol{\pi}^*).$$

It follows that

$$\|\boldsymbol{\pi}^{(k)} - \boldsymbol{\pi}^*\|^2 \doteq (\boldsymbol{\pi}^{(k)} - \boldsymbol{\pi}^*)\dot{G}^T\dot{G}(\boldsymbol{\pi}^{(k)} - \boldsymbol{\pi}^*).$$

If all the eigenvalues of $\dot{G}^T\dot{G}$ are smaller than one, then

$$\|\boldsymbol{\pi}^{(k+1)} - \boldsymbol{\pi}^*\|^2 < \|\boldsymbol{\pi}^{(k)} - \boldsymbol{\pi}^*\|^2.$$

We plan to investigate this approach in future work.

The preceding proofs focus on the convergence of the likelihood $L_i(\boldsymbol{\pi}^{(k)})$. This is because the theory of mixture likelihoods shows that $L_i(\hat{\boldsymbol{\pi}})$ is unique, even if $\hat{\boldsymbol{\pi}}$ itself is not unique. If we assume $\hat{\boldsymbol{\pi}}$ is unique, then we have the following result.

Theorem 3.7. *Suppose that $K(x_i, \phi)$ is continuous in ϕ and converges to zero when $|\phi| \to \infty$. If the NPMLE, $\hat{\boldsymbol{\pi}}$, is unique and has positive likelihood, and if the likelihood vectors $\boldsymbol{L}(\boldsymbol{\pi}^{(k)})$ form a convergent sequence, then $\boldsymbol{\pi}^{(k)} \xrightarrow{w} \hat{\boldsymbol{\pi}}$ in the sense of weak convergence of measures.*

Proof. We use the method of subsequences. That is, we need to show that for any subsequence $\{k'\}$ of $\{k\}$, there exist a further subsequence $\{k''\}$ such that $\boldsymbol{\pi}^{(k'')} \xrightarrow{w} \hat{\boldsymbol{\pi}}$.

For the given subsequence $\{k'\}$, we can choose a further subsequence $\{k''\}$ so that $\boldsymbol{\pi}^{(k'')} \xrightarrow{v} \boldsymbol{\alpha}$ by Corollary 5.6 in Bhattacharya and Waymire (2007). Here $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots)^T$ ∎

is a subprobability mass vector with $\sum_j \alpha_j \leq 1$ and $\xrightarrow{v}$ means vague convergence. Then by the definition of vague convergence in Bhattacharya and Waymire (2007), we have

$$L_i(\boldsymbol{\pi}^{(k'')}) \;\; = \;\; \int K(x_i;\phi)\,\boldsymbol{\pi}^{(k'')}(\phi)d\phi \to \int K(x_i;\phi)\,\boldsymbol{\alpha}(\phi)d\phi = L_i(\boldsymbol{\alpha})$$

since $K(x_i,\phi)$ is continuous and vanishes at $\pm\infty$. However, by Theorem 3.4, $L_i(\boldsymbol{\pi}^{(k'')}) \to L_i(\hat{\boldsymbol{\pi}})$. Hence, $L_i(\boldsymbol{\alpha}) = L_i(\hat{\boldsymbol{\pi}})$, and so $\boldsymbol{\alpha}$ maximizes the likelihood. Note that $\boldsymbol{\alpha}$ is a probability measure with $\sum_j \alpha_j = 1$, as otherwise $L_i(\boldsymbol{\alpha}) = \sum_j \alpha_j K(x_i;\phi_j)$ can be increased for each i by normalizing $\boldsymbol{\alpha}$, then we have a contradiction on $\hat{\boldsymbol{\pi}}$ is the NPMLE. Thus, by uniqueness of the NPMLE, $\boldsymbol{\alpha} = \hat{\boldsymbol{\pi}}$, and we have $\boldsymbol{\pi}^{(k'')} \xrightarrow{w} \hat{\boldsymbol{\pi}}$. It follows $\hat{\boldsymbol{\pi}}^{(k)} \xrightarrow{w} \hat{\boldsymbol{\pi}}$. □

Now we extend the previous theorems to the continuous case. Again let $K(x;\boldsymbol{\phi})$ be bounded and nonnegative. Moreover, let $K(x;\boldsymbol{\phi})$ be continuous in $\boldsymbol{\phi}$. Consider a model $f(x) = \int K(x;\boldsymbol{\phi})\pi(\boldsymbol{\phi})d\boldsymbol{\phi}$ where π is a density function of $\boldsymbol{\phi} \in \mathbb{R}^p$. Let $\hat{\Pi}(\boldsymbol{\phi})$ be the discrete NPMLE that maximizes the log-likelihood in (3.7). The continuous EM algorithm in (3.10) updates $\pi(\boldsymbol{\phi})$, but convergence to discrete $\hat{\Pi}$ must be in a suitable metric that includes both discrete and continuous distributions.

Lemma 3.8. *Suppose $\pi^{(0)}(\boldsymbol{\phi})$ is a continuous probability density function that satisfies $\pi^{(0)}(\boldsymbol{\phi}) > 0$ for all $\boldsymbol{\phi} \in \mathbb{R}^p$. Then $\pi^{(k)}(\boldsymbol{\phi}) > 0$ for all $\boldsymbol{\phi}$ for all iterations. In addition, $\Pi^{(k)}(\boldsymbol{\phi})$ is also continuous.*

Proof. The gradient function $D^{(k)}(\boldsymbol{\phi}) = \sum_{i=1}^n \left(\frac{K(x_i;\boldsymbol{\phi})}{L_i(\pi^{(k)})} - 1\right)$ is continuous in $\boldsymbol{\phi}$ since $K(\cdot,\boldsymbol{\phi})$ is. Suppose $\pi^{(k)}(\boldsymbol{\phi})$ is continuous and positive for all $\boldsymbol{\phi} \in \mathbb{R}^p$. The EM updates $\pi^{(k)}$ by

$$\pi^{(k+1)}(\boldsymbol{\phi}) = \pi^{(k)}(\boldsymbol{\phi})\left(1 + n^{-1}D^{(k)}(\boldsymbol{\phi})\right). \tag{3.21}$$

Since $\pi^{(k)}$ and $D^{(k)}$ are continuous, $\pi^{(k+1)}$ is also continuous. Also, both factors in the right-hand side of (3.21) are positive and it follows $\pi^{(k+1)}$ is also continuous and positive.

In addition,

$$\begin{aligned}\int \pi^{(k+1)}(\boldsymbol{\phi})\, d\boldsymbol{\phi} &= \int \pi^{(k)}(\boldsymbol{\phi}) \frac{1}{n} \sum_{i=1}^{n} \frac{K(x_i;\boldsymbol{\phi})}{L_i(\pi^{(k)})} d\boldsymbol{\phi} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\int \pi^{(k)}(\boldsymbol{\phi}) K(x_i;\boldsymbol{\phi}) d\boldsymbol{\phi}}{L_i(\pi^{(k)})} = 1.\end{aligned}$$

Thus by mathematical induction we have the conclusion. □

Theorem 3.9. *Suppose that $\pi^{(0)}(\boldsymbol{\phi})$ is continuous and satisfies $\pi^{(0)}(\boldsymbol{\phi}) > 0$ for all $\boldsymbol{\phi} \in \mathbb{R}^p$ and that the sequence of iterates $L_i(\pi^{(k)})$ are convergent with limit points L_i^*. Then, for each i, $L_i^* = L_i(\hat{\Pi})$, so the EM algorithm maximizes the likelihood.*

Proof. Similar to the proof of Theorem 3.4, suppose the gradient inequality does not hold for L_i^*, so that $D^*(\boldsymbol{\phi}_0) > 0$ for some $\boldsymbol{\phi}_0$. For any ϕ with $D(\phi) > 0$, we can choose a large enough K that, for every $k > K$, $D^{(k)}(\boldsymbol{\phi}) > \delta > 0$. Similar argument in Theorem 3.4 gives

$$\pi^{(K+m)}(\boldsymbol{\phi}) \geq \pi^{(K)}(\boldsymbol{\phi})\,(1+\delta)^m \to \infty \quad \text{as } m \to \infty.$$

That is, for every $\boldsymbol{\phi}$ with $D(\boldsymbol{\phi}) > 0$, we have

$$\lim_{k\to\infty} \pi^{(k)}(\boldsymbol{\phi}) = \infty.$$

Moreover, the continuity of $D(\boldsymbol{\phi})$ implies that if $D(\boldsymbol{\phi}_0) > 0$, then there exists a nonempty ϵ-neighborhood of ϕ_0, say $N(\epsilon, \boldsymbol{\phi}_0)$, where $D(\boldsymbol{\phi}) > 0$. On this neighborhood, we have

$$\liminf_{k\to\infty} \Pi^{(k)}\left(N(\epsilon,\phi_0)\right) = \liminf_{k\to\infty} \int_{N(\epsilon,\phi_0)} \pi^{(k)}(\boldsymbol{\phi}) d\boldsymbol{\phi} \geq \int_{N(\epsilon,\phi_0)} \liminf_{k\to\infty} \pi^{(k)}(\boldsymbol{\phi}) d\boldsymbol{\phi} = \infty$$

where the last inequality follows from Fatou's lemma. □

Theorem 3.10. *Suppose that the continuous EM algorithm has a finite set of fixed points, each with a different likelihood. Then the continuous EM algorithm converges to exactly one of the fixed points.*

Proof. Similar to the proof of Theorem 3.6. □

Theorem 3.11. *Suppose that $K(x_i, \phi)$ is continuous in ϕ and converges to zero when $|\phi| \to \infty$. If the NPMLE, $\hat{\Pi}$, is unique and has positive likelihood, and if the likelihood vectors $\boldsymbol{L}(\Pi^{(k)})$ form a convergent sequence, then $\Pi^{(k)} \xrightarrow{w} \hat{\Pi}$ in the sense of weak convergence of measures.*

Proof. As in Theorem 3.7, we use the method of subsequences. For any given subsequence $\{k'\} \subset \{k\}$, we can choose a further subsequence $\{k''\}$ so that $\Pi^{(k'')} \xrightarrow{v} \Pi^*$. Then again by the property of vague convergence, we have

$$L_i(\Pi^{(k'')}) \;=\; \int K(x_i;\phi)\, d\Pi^{(k'')}(\phi) \to \int K(x_i;\phi)\, d\Pi^*(\phi) = L_i(\Pi^*)$$

since $K(x_i, \phi)$ is continuous and vanishes at $\pm\infty$. However, by Theorem 3.9, $L_i(\Pi^{(k)}) \to L_i(\hat{\Pi})$. Hence, $L_i(\Pi^*) = L_i(\hat{\Pi})$ and it implies $\Pi^*(\mathbb{R}^p) = 1$ as otherwise we can increase the likelihood by normalizing it. By uniqueness, $\Pi^* = \hat{\Pi}$ and, therefore, $\Pi^{(k'')} \xrightarrow{w} \hat{\Pi}$ and it implies $\Pi^{(k)} \xrightarrow{w} \hat{\Pi}$. □

3.4 Asymptotic Properties

3.4.1 Asymptotic Bias and Variance

Chung and Lindsay (2010) compared the theoretical MSE of $\hat{f}_{ABW}$, $\hat{f}_{MBC}$ and $\hat{f}_{EM2}$ for the univariate case. In the standard normal and the bimodal density, the adaptive bandwidth density estimator had larger MSE than the others in most regions. Although the likelihood-tuned estimator was asymptotically worse than the multiplicative bias correction estimator in some places, it is known that asymptotic results may differ significantly from actual finite sample performance (Bowman and Foster (1993)). In fact, the simulation result showed that $\hat{f}_{MBC}$ performed better than the others for the Gaussian mixture densities but worse for the non-Gaussian densities.

In this section, we investigate the asymptotic properties of the multivariate likelihood-tuned density estimator. In addition, let $\boldsymbol{\varphi}(\cdot)$ be the probability density function of $N(0, I_d)$.

Theorem 3.12. *Let the bandwidth matrix H be hA where h is a scalar bandwidth and $A \in \mathbb{R}^{d\times d}$ is a positive-definite symmetric matrix which satisfies $|A| = 1$ and whose i-th column is denoted by $\boldsymbol{a}_i$. Suppose that all fourth order partial derivatives of f exist and are continuous in a neighborhood of $\boldsymbol{x}$, and $K_H(\boldsymbol{x}, \boldsymbol{\phi}) = H^{-1}\boldsymbol{\varphi}\left(H^{-1}(\boldsymbol{x} - \boldsymbol{\phi})\right)$. Then, when $h \to 0$ and $nh^d \to \infty$,*

$$E\left[\hat{f}_{EM2}(\boldsymbol{x})\right] = f(\boldsymbol{x}) \quad + \quad \sum_{i,j,k,l} f(\boldsymbol{x}) \left[-\frac{f^{(i,j,k,l)}(\boldsymbol{x})}{f(\boldsymbol{x})} + \frac{f^{(i,j,k)}(\boldsymbol{x})f^{(l)}(\boldsymbol{x}) + f^{(i,j)}(\boldsymbol{x})f^{(k,l)}(\boldsymbol{x})}{f^2(\boldsymbol{x})} \right.$$
$$\left. - \frac{f^{(i,j)}(\boldsymbol{x})f^{(k)}(\boldsymbol{x})f^{(l)}(\boldsymbol{x})}{f^3(\boldsymbol{x})} \right] (\boldsymbol{a}_i^T\boldsymbol{a}_j)(\boldsymbol{a}_k^T\boldsymbol{a}_l)h^4 + o(h^4) \tag{3.22}$$

and

$$Var\left[\hat{f}_{EM2}(\boldsymbol{x})\right] = n^{-1}h^{-d}f(x)|A|^{-1}\pi^{-\frac{d}{2}}(2^{-\frac{3}{2}d+2} + 2^{-2d} - 2^{-d+2}3^{-\frac{d}{2}}) + o(n^{-1}h^{-d}). \tag{3.23}$$

where $f^{(i)}(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial x_i}$, $f^{(i,j)}(\boldsymbol{x}) = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}$, and so on.

Proof. Similar to the proof of the univariate case in Appendix A.1. □

Theorem 3.12 reveals that the likelihood-tuned density estimator has pointwise bias of order $O(h^4)$ and variance of order $O(n^{-1}h^{-d})$ as do the other modified density estimators. Before comparing it with other density estimators, one should note that we used a bandwidth of $h\sqrt{2}$ instead of h for the one step kernel density estimator in Section 2.2. Thus it is reasonable to compare the one step and two step estimators to see the effect of likelihood tuning. When we make comparisons with the adaptive bandwidth and the multiplicative bias correction estimator, we also need to account for bandwidth effects.

Since $\hat{f}_{MBC}(\boldsymbol{x})$ given in (3.4) does not integrate to 1, Jones et al. (1995) suggested rescaling it by dividing $\hat{f}_{MBC}(\boldsymbol{x})$ by its integral. They provided the bias of the rescaled version. As in the univariate case, with the Gaussian kernel function, the rescaled

$\hat{f}_{MBC}(\boldsymbol{x})$ has exactly the same asymptotic variance as $\hat{f}_{EM2}(\boldsymbol{x})$. In addition, the rescaled $\hat{f}_{MBC}$ has zero bias for $N(\mathbf{0}, I)$ at the order $O(h^4)$. This could be a reason that $\hat{f}_{MBC}$ had good simulation results in Gaussian distribution in Section 3.5.

3.4.2 Optimal bandwidths and Mean Integrated Squared Errors in the normal case

In this section, we investigate the optimal bandwidth of the kernel density estimator and the likelihood-tuned density estimator as dimension increases. Consider the standard normal case, $N(0, I_d)$, and let the bandwidth matrix H be hI. Scott (1992) obtained that the bandwidth minimizing the Asymptotic Mean Integrated Squared error (AMISE) of the kernel density estimator at the standard normal is given as

$$h_{opt,KER} = \left(\frac{4}{(d+2)}\right)^{\frac{1}{d+4}} n^{\frac{1}{d+4}}. \tag{3.24}$$

Similarly, the AMISE of the likelihood-tuned density estimator, $\hat{f}_{EM2}$, can be calculated by integrating the square of the asymptotic bias and the asymptotic variance in Theorem 3.12 over $\boldsymbol{x}$.

Lemma 3.13. *Suppose $f(\boldsymbol{x})$ is a standard normal density and let A be the identity matrix I. Then the bandwidth that minimizes the AMISE of $\hat{f}_{EM2}$ is*

$$h_{opt,EM2} = \left[\frac{2^{-d}(3\pi)^{-\frac{d}{2}}\left\{-2^{2+d}\pi^{\frac{d}{2}} + (3\pi)^{\frac{d}{2}} + 2^{2+\frac{d}{2}}(3\pi)^{\frac{d}{2}}\right\}}{(16+8d)}\right]^{\frac{1}{d+8}} n^{\frac{1}{d+8}}.$$

Proof. Theorem 3.12 includes the asymptotic bias and variance of $\hat{f}_{EM2}$. When $A = I$, the asymptotic bias in (3.22) is reduced to

$$\begin{aligned} ABias(\boldsymbol{x}) &= \sum_{i=1}^{d}\sum_{j=1}^{d}\left[-f^{(i,j)}(\boldsymbol{x}) + \frac{f^{(i,i,j)}(\boldsymbol{x})f^{(j)}(\boldsymbol{x}) + f^{(i,i)}(\boldsymbol{x})f^{(j,j)}(\boldsymbol{x})}{f(\boldsymbol{x})}\right. \\ &\quad \left. - \frac{f^{(i,i)}(\boldsymbol{x})f^{(j)}(\boldsymbol{x})f^{(j)}(\boldsymbol{x})}{f^2(\boldsymbol{x})}\right] h^4. \end{aligned} \tag{3.25}$$

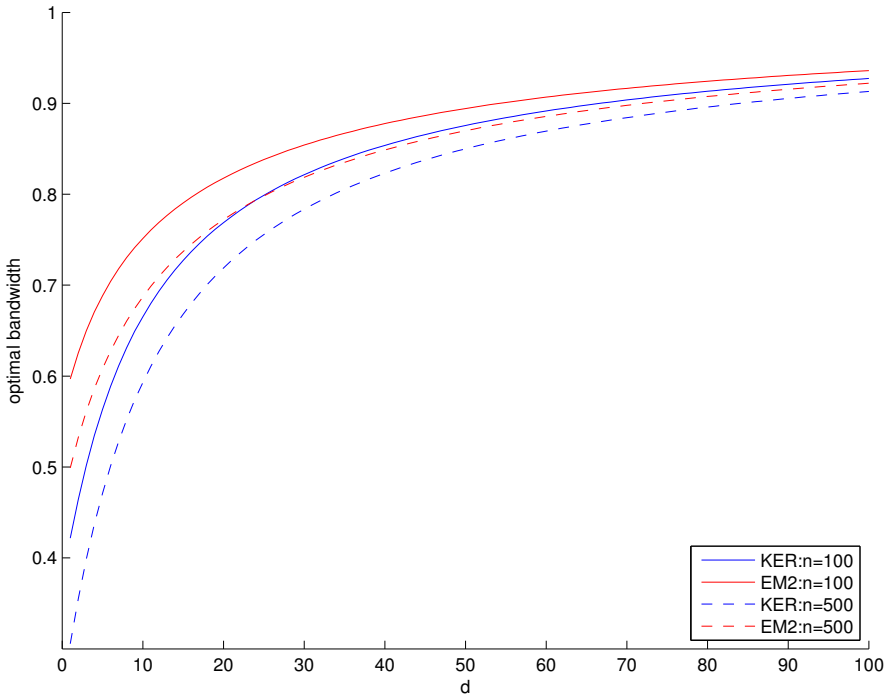


Figure 3.1: Optimal bandwidths for $\hat{f}_{KER}$ and $\hat{f}_{EM2}$

When $f(\boldsymbol{x}) = N(0, I)$, each term in (3.25) can be written in a simple polynomial of x_i and x_j. Then integrating the square of $ABias(\boldsymbol{x})$ is given by

$$IASB(h,d) = h^8 \left(2^{1-d} d \pi^{-2+\frac{4-d}{2}} + 2^{-d} d^2 \pi^{-2+\frac{4-d}{2}}\right).$$

Again, integrating the asymptotic variance in (3.23) becomes

$$IAV(h,d,n) = \frac{\left(2^{2-\frac{3d}{2}} + 2^{-2d} - 2^{2-d} 3^{-d/2}\right) h^{-d} \pi^{-d/2}}{n}.$$

Solving the equation

$$\frac{\partial \left(IASB(h,d) + IAV(h,d,n)\right)}{\partial h} = 0,$$

we have the result above. □

Figure 3.1 shows the optimal bandwidths of $\hat{f}_{KER}$ and $\hat{f}_{EM2}$ for $n = 100$ and 500

as the dimension increases. One should note that we need a bandwidth $\sqrt{2}H$ instead of H for the one step kernel density estimator in Section 3.2.3. Thus it is reasonable to compare the one step and two step estimators to see the effect of likelihood tuning. When we make comparisons with the adaptive bandwidth and the multiplicative bias correction estimator, we also need to account for bandwidth effects.

For both $\hat{f}_{KER}$ and $\hat{f}_{EM2}$, a larger sample size corresponds to a smaller optimal bandwidth. When the sample size is small, one needs a relatively large bandwidth in order to avoid the bias generated by outlying bumps.

The optimal bandwidths for $\hat{f}_{KER}$ and $\hat{f}_{EM2}$ converge to 1 as $d \to \infty$ regardless of the sample size. However, for fixed d, the optimal bandwidth for $\hat{f}_{EM2}$ is greater than the bandwidth for $\hat{f}_{KER}$. For the standard normal density, the kernel density estimator tends to underestimate the peak and, thus it has the largest bias at the peak. Compared with the kernel density estimator, $\hat{f}_{EM2}$ is more flexible and so able to fit the mode more accurately at the same bandwidth.

3.5 Simulation Comparisons

In this section, we use simulated data to compare the performances of $\hat{f}_{KER}(x)$, $\hat{f}_{ABW}(x)$, $\hat{f}_{MBC}(x)$ and $\hat{f}_{EM2}(x)$. For the univariate case, Chung and Lindsay (2010) considered two kinds of distributions; Gaussian mixture densities and non-Gaussian densities. For each distribution and each estimator, they calculated an *optimal bandwidth* by minimizing the average of integrated square errors (ISE), given by

$$ISE(\hat{f}) = \int \left\{ \hat{f}(x) - f(x) \right\}^2 dx.$$

The result showed that, in terms of MISE, $\hat{f}_{MBC}(x)$ was superior to the other density estimators for the Gaussian mixture densities while it performed worst for the non-Gaussian densities. On the other hand, $\hat{f}_{KER}$ was the worst for the Gaussian mixture densities but it performed well for the non-Gaussian densities. $\hat{f}_{ABW}$ was worse than

$\hat{f}_{EM2}$ for both the Gaussian mixture and non-Gaussian densities. In conclusion, even though $\hat{f}_{EM2}$ was not the best estimator for both kinds of densities, it was robust in efficiency compared to the other estimators.

In this section, we did similar simulation study to Chung and Lindsay (2010) in dimensions one to four. We considered Gaussian distributions with $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = I$, and with $\Sigma = 0.7 \cdot \mathbf{1}_{d\times d} + 0.3 \cdot I$. In addition, we also examined independent coordinate $Beta$ distributions, $Beta(1, 3)$ and $Beta(2, 5)$. The sample size is $n = 100$ and the simulation replication is $R = 100$. Again, we used the optimal bandwidth that minimized MISE for each density estimator.

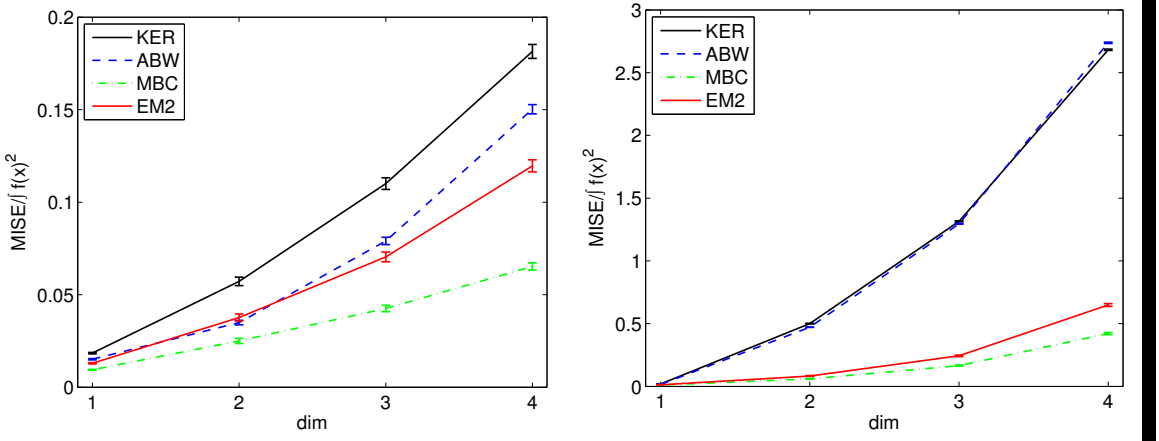
In comparison over dimensions, squared errors can mislead the result. In higher dimensions, since the $f(x)$ and $\hat{f}(x)$ have relatively low heights, squared errors will be also smaller. Therefore, if we compare MISE between low and high dimensions, MISE in the higher dimension would be more likely to be smaller regardless of accuracy of the estimates. Scott (1991) discussed several dimensionless quantities for density estimates. Among them, the relative MISE, defined by

$$\frac{MISE(\hat{f})}{\int_{\mathbb{R}^d} f(x)^2 dx},$$

was used in the following simulation study.

Figures 3.2 contain results for two Gaussian distributions. For both the standard normal and correlated normal densities, $\hat{f}_{MBC}$ has the smallest relative MISE in all four dimensions. As mentioned in Section 3.4, this could be caused by the fact that $\hat{f}_{MBC}$ has a zero bias for $N(\mathbf{0}, I)$. As in the univariate simulations, even though $\hat{f}_{EM2}$ had larger MISE than $\hat{f}_{MBC}$, it performed still better than the other two density estimators. As dimension increases, this tendency looks more obvious.

When the true density was correlated in Figure 3.2 (b), the advantage of $\hat{f}_{MBC}$ and $\hat{f}_{EM2}$ against $\hat{f}_{KER}$ and $\hat{f}_{ABW}$ was more clear compared to the standard normal case. While $\hat{f}_{MBC}$ had the smallest MISE and $\hat{f}_{EM2}$ had the next smallest, the other two

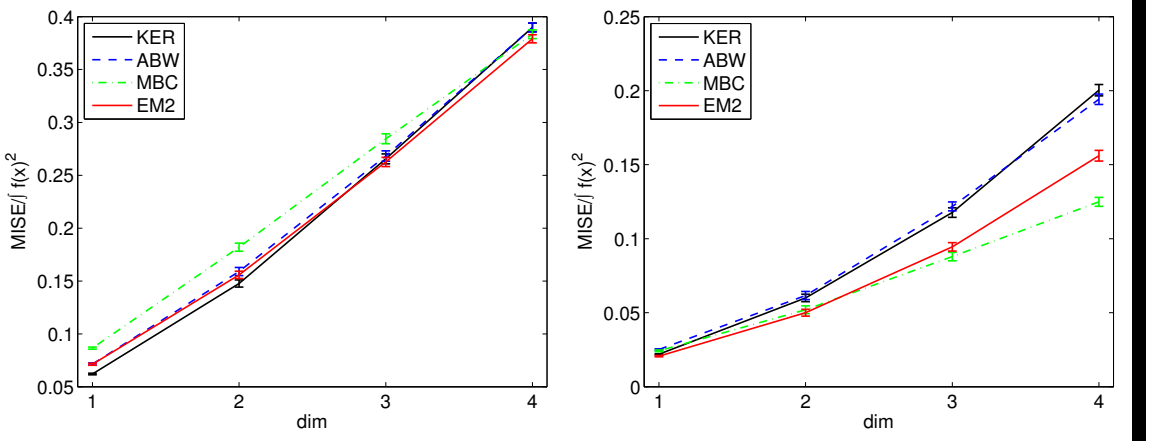


(a) Standard Normal

(b) Normal distribution with a covariance matrix $\Sigma = 0.7 \cdot \mathbf{1}_{d\times d} + 0.3 \cdot I$

Figure 3.2: $MISE/\int f(x)^2 dx$ for Gaussian distributions

estimators had obviously larger MISE and they were not distinguishable each other. In dimension four, $\hat{f}_{ABW}$ was even worse than $\hat{f}_{KER}$.



(a) Independent Beta(1,3)

(b) Independent Beta(2,5)

Figure 3.3: $MISE/\int f(x)^2 dx$ for Beta distributions

Figure 3.3 (a) shows the results for $Beta(1,3)$ distribution. In lower dimensions, $\hat{f}_{KER}$ had the smallest MISE while $\hat{f}_{MBC}$ had the largest. As dimension increased, the performance of $\hat{f}_{KER}$ became worse and, on the other hand, the performance of $\hat{f}_{MBC}$ started to be better. In dimension one and two, $\hat{f}_{EM2}$ was the second best estimator.

However, in the higher dimensions, $\hat{f}_{EM2}$ beat $\hat{f}_{KER}$, so it performed the best.

Although $\hat{f}_{EM2}$ was favorable for $Beta(1,3)$ as dimension increased, this pattern did not stay the same for $Beta(2,5)$. While $\hat{f}_{EM2}$ had the smallest MISE in one dimension, $\hat{f}_{MBC}$ started to overcome $\hat{f}_{EM2}$ as dimension increased, so it had the smallest MISE in the four dimension case. As for $Beta(1,3)$, the performance of $\hat{f}_{KER}$ was getting worse in higher dimensions.

It is clear that the univariate result does not hold up in higher dimensions. In the univariate case, $\hat{f}_{KER}$ performed well for $Beta$ distributions while it was poor for Gaussian distributions. However, as dimension increased, the performance of $\hat{f}_{KER}$ became worse for $Beta$ distributions while it remained worse for Gaussian distributions. On the other hand, $\hat{f}_{MBC}$ was improving for both Gaussian and $Beta$ distributions as the dimension increased, and, in fact, it seems to work better than $\hat{f}_{EM2}$.

In spite of the fact that the most competitive estimator, $\hat{f}_{MBC}$, gained an advantage in higher dimensions, $\hat{f}_{EM2}$ performed reasonably well. For Gaussian distributions, $\hat{f}_{EM2}$ kept working as the second best. For $Beta$ distributions, $\hat{f}_{EM2}$ tends to be better than $\hat{f}_{KER}$ even though it seems to be beaten by $\hat{f}_{MBC}$ as dimension increased.

One question remaining in our minds is the extent to which this deficiency in relative performance is due in part to our initial value $\pi^{(0)}(\phi)$, which was uniform. If we had chosen $\pi^{(0)}(\phi)$ to be Gaussian as described briefly in Section 3.2.3, say then would certainly expect to improve performance of $\hat{f}_{EM2}$ when the true density is Gaussian. We leave the role of starting value to be a subject of future work.

3.6 Mode Identification

This section investigates the performance of density estimators on mode identification. In cluster analysis, the location and the number of modes in the underlying density are closely related to the number of clusters and the process to group observations. Especially, a model-based clustering algorithm, proposed by Li, Ray, and Lindsay (2007)

is based on the modes of density estimates. In this section, I will explain the model-based clustering algorithm by Li et al. (2007) and investigate how well the density estimators identify underlying modes in two-component Gaussian mixture model.

3.6.1 Mode Association Clustering by Li, Ray, and Lindsay (2007)

Recently, Li et al. (2007) proposed a new model-based hierarchical agglomerative clustering method. They did not impose a parametric assumption on each cluster and, instead, employed a nonparametric kernel density estimator. They developed the Modal EM (MEM) algorithm, which iteratively searched a local maximum of a mixture density

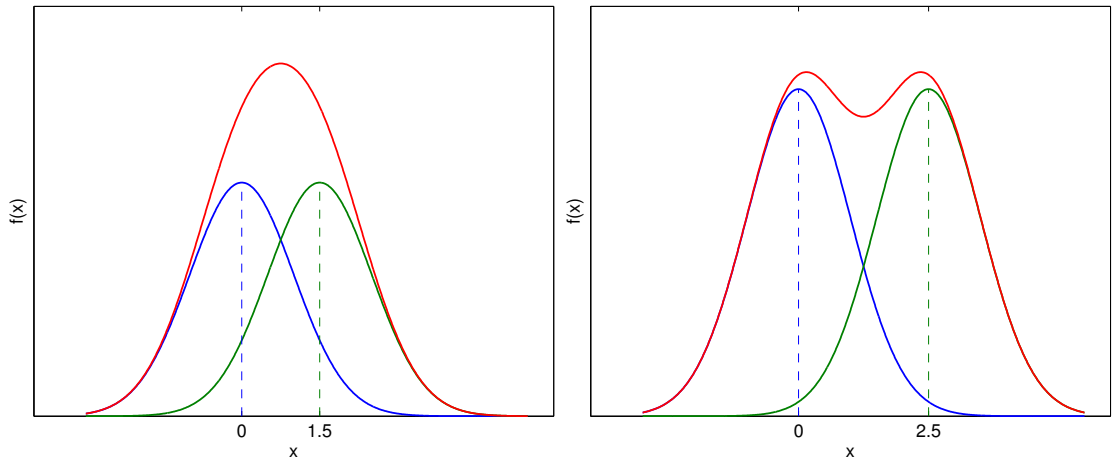
$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x), \tag{3.26}$$

where π_k was a mixing proportion and $f_k(x)$ was a component density. Given any initial value of x, MEM repeats the following steps until a stopping criterion met.

1. Fit
$$p_k = \frac{\pi_k f_k(x^{(r)})}{f(x^{(r)})}, k = 1, \cdots, K$$

2. Updates
$$x^{(r+1)} = \arg\max_x \sum_{k=1}^{K} p_k \log f_k(x). \tag{3.27}$$

As a special case of (3.26), they considered the kernel density estimator with Gaussian kernel, $f(x) = n^{-1} \sum_{i=1}^{n} h^{-1} \varphi(\frac{x_i - x}{h})$. Starting from each observation, MEM finds a mode of the kernel density estimate. If some observations climbed up to the same mode, they are merged into a cluster and the mode is named to be the cluster representative of the corresponding cluster. Since the kernel density estimate becomes smooth as h is increased, the resulting density estimate tend to have the fewer modes. In the next stage, the modes are identified starting from the cluster representatives instead of individual observations so that we obtain the hierarchical clustering by merging clusters whose



(a) $|\mu_1 - \mu_2| = 1.5 < 2\sigma$ (b) $|\mu_1 - \mu_2| = 2.5 > 2\sigma$

Figure 3.4: Two-component mixture density with equal weight and $\sigma = 1$

cluster representative reached the same mode. They called this the hierarchical mode association clustering (HMAC).

This clustering method had the characteristics of both mixture model clustering and linkage clustering by tuning h. In addition, this method was robust in high dimensions and with non-Gaussian data. Since this method was derived from modal estimates based on the kernel density estimator, a density estimator with more accurate modal estimates than the kernel density estimator might give a better clustering result. With this motivation, the following section investigates the performance on mode identification of the density estimators that have been examined in this chapter.

3.6.2 Mode Identification in Two-component Gaussian Mixture Density

In this section, we carry out an experiment to test each density estimator to detect the modes of the true density. Consider a univariate two-component Gaussian mixture model, written as

$$f(x) = 0.5 \cdot \varphi_1(x; \mu_1, \sigma^2) + 0.5 \cdot \varphi_2(x; \mu_2, \sigma^2). \tag{3.28}$$

Here we restrict the two components to have the same weight and variance. It is known that that $f(x)$ is bimodal when $|\mu_1 - \mu_2| > 2\sigma$ (Ray and Lindsay (2005).) In Figure 3.4 (a), $f(x)$ is unimodal when $|\mu_1 - \mu_2|$ is less than 2σ. On the other hand, Figure 3.4 (b) shows that $f(x)$ is bimodal when $|\mu_1 - \mu_2|$ is greater than 2σ.

Even though the true density $f(x)$ is bimodal when μ_1 is far enough from μ_2, the density estimate based on a data set could be still unimodal. Let the bandwidth fixed to be h_0. Then the expectation of $\hat{f}_{KER}$ becomes

$$\begin{aligned} E\hat{f}_{KER}(x) &= \int K_{h_0}(y;x) f(x) dx \\ &= \int \varphi(y;x,h_0^2) \left[0.5\varphi(y;\mu_1,\sigma^2) + 0.5\varphi(y;\mu_2,\sigma^2)\right] dy \\ &= 0.5\varphi(x;\mu_1,\sigma^2+h_0^2) + 0.5\varphi(x;\mu_2,\sigma^2+h_0^2). \end{aligned}$$

Thus the expected value of $\hat{f}_{KER}(x)$ is bimodal when $|\mu_1 - \mu_2| > 2\sqrt{\sigma^2 + h_0^2}$. That is, when $2\sigma < |\mu_1 - \mu_2| < 2\sqrt{\sigma^2 + h_0^2}$, $\hat{f}_{KER}$ is unlikely to detect two modes even though the true density is bimodal. This is the price of kernel smoothing.

In the following simulation, we compared the performance of the four density estimators in mode identification. We set σ to be one and the bandwidth of density estimators was fixed. Given a fixed distance between μ_1 and μ_2, for each density estimator we took a pointwise mean of the density estimates from R =1000 samples of size n =100 and located the modes of this expected curve. If, at the given separation $|\mu_1 - \mu_2|$, there were two modes, we called them $mode1$ and $mode2$. In higher dimensions, we kept the same distance between μ_1 and μ_2 the first coordinate, then let the other means be zero for both components.

We did this analysis with h fixed at the normal reference rule in (3.24). In the one dimensional case with the sample size 100, the formula in (3.24) equals 0.4217. The results are shown in Figure 3.5. Since the true density $f(x)$ starts to be bimodal only when $|\mu_1 - \mu_2|$ is two, the black dashed line which shows the relative distance between modes, $|mode1 - mode2|/|\mu_1 - \mu_2|$, is zero at $|\mu_1 - \mu_2| = 2$. As the separation increases,

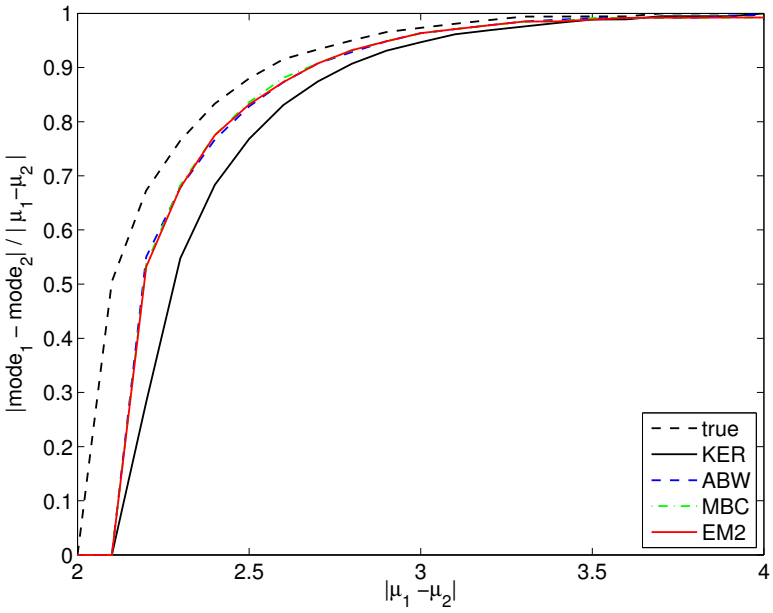


Figure 3.5: Mode Identification result when h = normal reference rule.

the distance between the two modes moves way from zero. It eventually converges to one as $|\mu_1 - \mu_2|$ increases, being nearly so at $|\mu_1 - \mu_2| = 3.5$. At this separation, the modes and means are nearly the same.

Until $|\mu_1 - \mu_2|$ is about 2.2, none of the four density estimators separated two modes. They started to distinguish the modes all together when $|\mu_1 - \mu_2|$ is about 2.2. However, as the separation increased, $\hat{f}_{KER}$ clearly underestimated the mode difference compared to the other three improved density estimators. Thus we can conclude that $\hat{f}_{ABW}$. $\hat{f}_{MBC}$ and $\hat{f}_{EM2}$ are more sensitive to detect modes than $\hat{f}_{KER}$ even at the optimal bandwidth for $\hat{f}_{KER}$.

For comparison, we considered another bandwidth $h = 1$ because we wish to understand model discrimination in higher dimensions. Recall that in Section 3.4.2, we investigated the bandwidths for $\hat{f}_{KER}$ and $\hat{f}_{EM2}$ that minimize the asymptotic MISE when the true density is $N(0, I)$. It was shown that the optimal bandwidths for both estimators converged to one when dimension increased to the infinity. Thus the bandwidth $h = 1$ is meaningful in higher dimensions. Our results are shown in Figure 3.6.

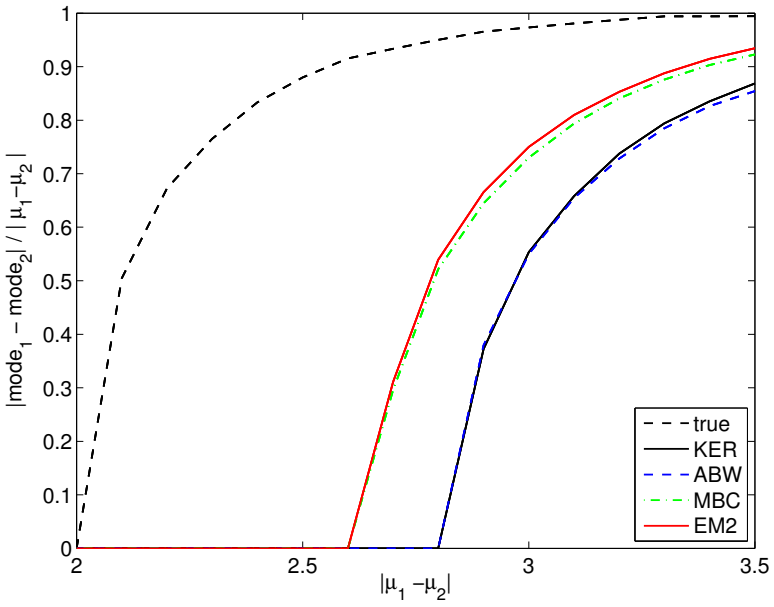


Figure 3.6: Mode Identification result when $h = 1$.

Until $|\mu_1 - \mu_2|$ is about 2.6, none of these four density estimators separate two modes. $\hat{f}_{EM2}$ and $\hat{f}_{MBC}$ start to distinguish the modes when $|\mu_1 - \mu_2|$ is about 2.6, and $\hat{f}_{ABW}$ and $\hat{f}_{KER}$ do when $|\mu_1 - \mu_2|$ is around 2.8. Thus we can conclude that $\hat{f}_{EM2}$ and $\hat{f}_{MBC}$ are more sensitive to detect the modes than the other two density estimators at $h = 1$. Even though $\hat{f}_{EM2}$ does not look significantly different from $\hat{f}_{MBC}$, the red solid line is still a little above the green dash-dotted line. Therefore, in this evaluation, $\hat{f}_{EM2}$ was a little more sensitive than $\hat{f}_{MBC}$.

3.7 Future Work

Based on the simulation study, even though the likelihood-tuned density estimator had a good performance in mode identification and its MISE were reasonable in most cases, there seem to exist some room to improve our method. As mentioned at the end of Section 3.5, we expect that Gaussian initial π would bring some advantage especially when the true is a Gaussian mixture distribution.

In addition, regarding the sparsity of data in higher dimensions, we definitely need to

examine larger sample size in simulation. Increasing the sample size in higher dimensions, we can get an idea where the advantage of $\hat{f}_{MBC}$ (and $\hat{f}_{EM2}$ for $Beta(1,3)$) in higher dimensions came from.

Chapter 4

Spectral Degrees of Freedom and Bandwidth Selection

4.1 Introduction

In Chapter 3, we investigated the multivariate likelihood-tuned density estimator and compared with other density estimators. Our simulation comparison was on the optimal bandwidth that minimized MISE for each density estimator, which is a widely used model selection criterion. However, the problem of bandwidth selection is crucial in density estimation and should always be influenced by the purpose of investigator as noted in Silverman (1986). Especially in higher dimensions, it is more difficult to select an appropriate bandwidth because of sparsity of data. In addition, it seems unlikely that one wants to give equal weight to bias and variance, as done by the MISE criterion. In this chapter, we examine a bandwidth selection method, focused on clustering in higher dimensions. Instead of standard bandwidth selection methods, we employ the spectral degrees of freedom introduced in Lindsay et al. (2008).

4.1.1 Bandwidth Selection in Kernel Density Estimation

Most of the standard bandwidth selection methods are based on the integrated square error (ISE), given by

$$ISE(\hat{f}) = \int (\hat{f}(x) - f(x))^2 dx. \tag{4.1}$$

For the standard kernel density estimator, many authors suggested methods to estimate the asymptotic MISE (AMISE). Here a difficulty caused by the term $\int f''(x)^2 dx$ in AMISE of $\hat{f}_{KER}$. Silverman (1986) includes a rule of thumb method based on the normal distribution. Sheather and Jones (1991) proposed a plug-in selector to replace $\int f''(x)^2 dx$ with a kernel based estimate in the optimal bandwidth representation that minimizes AMISE. The biased cross-validation (BCV) method by Scott and Terrell (1987) used a cross-validatory kernel estimator in the AMISE formula. However, all these methods are based on the AMISE for the kernel density estimator. Therefore, it is hard to apply them other improved density estimators, which have more complex form of AMISE, especially in higher dimensions.

One of the most widely studied bandwidth selection method is least squares cross-validation (LSCV), proposed by Rudemo (1982) and Bowman (1984), which attempts to minimize MISE instead of AMISE. In the expansion of MISE, $\int \hat{f}(x)^2 - 2\hat{f}(x)f(x) + f(x)^2 dx$, the cross term $\int \hat{f}(x)f(x)dx$ is estimated by leave-one-out cross-validation. Since this does not depend on the form of the density estimator $\hat{f}$, it can be applied to any density estimators. However, studies have shown that the performance of LSCV can be disappointing (Hall and Marron (1987).) (For a contrary point of view, see Loader (1999).)

Differing from the attempts to minimize integrated square errors by MISE or AMISE, Geisser (1975) used likelihood cross-validation, which aimed to maximize the likelihood. Instead of minimizing L_2 error in (4.1), it intends to minimize the Kullback-Leibler

deviance between $\hat{f}$ and f, given by

$$KL(\hat{f}) = \int f(x) \log \left\{ f(x)/\hat{f}(x) \right\} dx. \tag{4.2}$$

This method is generally applicable, not only for density estimators, but Scott and Factor (1981) noted that corresponding cross-validation method's performance is sensitive to outliers.

In the following sections, we investigate the relationship of bandwidth selection to two quantities, spectral degrees of freedom and coalescence level in higher dimensions. Unlike the standard methods mentioned above, we are looking to find a good bandwidth for clustering in higher dimensions where sparsity of data causes a trouble in density estimation and clustering. We do not consider minimizing or maximizing a model selection criterion as in (4.1) and (4.2). Instead, we examine a general distance measure between two distributions and the associated degrees of freedom, which give an idea about tightness of data points.

4.1.2 Spectral Degree of Freedom and Kernel Density Estimation

Lindsay et al. (2008) built a unified structure of model assessment. They defined a quadratic distance between two distribution F and G, given by

$$d_K(F, G) = \int \int K_G(s, t) d(F - G)(s) d(F - G)(t). \tag{4.3}$$

Some conventional distances can be written as a quadratic distance. In the discrete space, $K(s, t) = \mathbb{I}[s = t]$ results in the ordinary L_2 distance, $\sum (f(i) - g(i))^2$. When $K_G(s, t) = \sum \mathbb{I}[s \in A_i]\mathbb{I}[t \in A_i]/G(A_i)$, the equation (4.3) becomes a Pearson's chi-square statistics.

They mentioned that the kernel density estimation problem can be expressed as a quadratic distance problem; the form is quite simple when the kernel is a Markov diffusion kernel as defined in Section 3.2.3. Let $f^*(x) = \int K_{h^2}(x, z) dF(z)$ and $K(\cdot, \cdot)$ be

a Markov diffusion kernel. Then the kernel density estimator can be expressed as $\hat{f}^*(x) = \int K_{h^2}(x,z)d\hat{F}(z)$ where $\hat{F}$ is the empirical distribution. The L_2 distance between $\hat{f}^*$ and the smoothed true density, f^*, can be written by

$$\begin{aligned} L_2(\hat{f}^*, f^*) &= \int (\hat{f}^*(x) - f^*(x))^2 dx \qquad (4.4) \\ &= \int \left\{ \int K_{h^2}(x,y) d(\hat{F} - F)(y) \right\} \left\{ \int K_{h^2}(x,z) d(\hat{F} - F)(z) \right\} dx \\ &= \int \left\{ \int K_{h^2}(x,y) K_{h^2}(x,z) dx \right\} d(\hat{F} - F)(y) d(\hat{F} - F)(z) \\ &= \int K_{2h^2}(x,y) d(\hat{F} - F)(y) d(\hat{F} - F)(z) = d(\hat{F}, F) \end{aligned}$$

If K_h^2 was not a Markov kernel, K_{2h^2} would be replaced by the convolution kernel defined by $K^*(x,z) = \int K_{h^2}(x,y) K_{h^2}(y,z) dy$.‘ In formula (4.3), Lindsay et al. (2008) considered G to be a null distribution and F to be the true distribution. When F was estimated by the empirical distribution $\hat{F}$, then the empirical quadratic distance $d_K(\hat{F}, G)$ could be viewed as a test statistic for testing $H_0 : F = G$. Then they showed that, under H_0, $d_K(\hat{F}, G)$ converges to a distribution that is approximately a scale multiple of a chi-squared distribution with its degrees of freedom given by the spectral degrees of freedom (sDOF),

$$sDOF_G(K) = \frac{trace_G(K)^2}{trace_G(K^2)}. \qquad (4.5)$$

Here $trace_G(K) = \int K(x,x) dG(x)$ and $trace_G(K^2) = \int \int K(x,y)^2 dG(x) dG(y)$. In the Pearson's chi-square example in the previous paragraph, the sDOF corresponds to the usual degrees of freedom of the Pearson's chi-squared statistics. That is, dDOF is $(\#\ cells - 1)$.

In higher dimensions, data tend to be sparse and it makes nonparametric density estimation very difficult. As we will see later in this section, for a fixed value of n and for data from multivariate normal, the bandwidth needed to provide coalescence of the kernel density estimator to a unimodal one is proportional to $\sqrt{d}$, the square root

of dimension. Thus it is clear that the bandwidth should be substantially increased as dimension grows if we wish to get the number of modes right. For this reason, we investigate in this chapter whether the sDOF could provide information about the role of bandwidth in nonparametric density estimates.

We note here that the spectral degrees of freedom of a kernel density estimator is a measure of the variance of it. That is, in (4.4), $\hat{f}^*$ is an unbiased estimator of f^*, and so the expected quadratic distance

$$E\left[d(\hat{F}, F)\right] = E\left[L_2(\hat{f}^*, f^*)\right] = \int var(\hat{f}^*(x))dx$$

is an integrated variance measure. However, such an expected distance measure is by itself not easy to interpret over different scales of measurement and different numbers of dimensions. The spectral degrees of freedom is a measure that eliminates scale factors and provides a measure of smoothing whose interpretation is not dependent on dimension.

We might contrast these considerations with MISE, written by

$$\begin{aligned} MISE(\hat{f}^*) &= E\int(\hat{f}^*(x) - f(x))^2 dx \\ &= E\left[d(\hat{F}, F)\right] + E\int(f^*(x) - f(x))^2 dx. \end{aligned} \tag{4.6}$$

MISE focuses equally on both bias and variance, and so minimizing it in higher dimensions forces one to let the variances become very large. That is, if one minimizes MISE, one might not even estimate f^* well. Our point of view is that the integrated bias in (4.6) could be quite large even though f^* has the key features of the shape of $f(x)$, and we should at least do a good job of estimating the shape of f^* by controlling the overall variance.

To help understand the spectral degrees of freedom better, consider the kernel func-

tion for d-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, defined by

$$K_h(\boldsymbol{x}, \boldsymbol{y}) = \left(\frac{1}{2h}\right)^d \prod_{j=1}^{d} \mathbb{I}(|x_j - y_j| \leq h). \tag{4.7}$$

The kernel density estimator with the above kernel is

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n(2h)^d} NN_h(\boldsymbol{x}), \tag{4.8}$$

where $NN_h(\boldsymbol{x})$ is the number of data neighbors of x; that is, it is the number of observations in the cube that is centered at x, each edge having length $2h$. Although this estimator is cruder than the kernel density estimator with the Gaussian kernel, we use it to better understand the relationship between nonparametric density estimators and the corresponding spectral degrees of freedom.

With the kernel in (4.7), the numerator of the sDOF in (4.5) is estimated by

$$trace_{\hat{F}}(K)^2 = \int \left(\frac{1}{2h}\right)^d \prod_{j=1}^{d} \mathbb{I}(|x_j - x_j| \leq h) d\hat{F}(\boldsymbol{x})^2 = \frac{1}{(2h)^{2d}},$$

and the denominator can be estimated by

$$\begin{aligned} trace_{\hat{F}}(K^2) &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K^2(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= \frac{1}{(2h)^{2d}} \frac{1}{n^2} \sum_{i=1}^{n} NN_h(\boldsymbol{x}_i). \end{aligned}$$

Thus the estimated sDOF with the kernel in (4.7) becomes

$$\frac{n}{\frac{1}{n}\sum_{i=1}^{n} NN_h(\boldsymbol{x}_i)}. \tag{4.9}$$

The denominator of (4.9) equals the average number of neighbors of each data point. If h is very large, then $NN_h(\boldsymbol{x}_i) = n$ for all i, so the sDOF will be one. On the other hand, if h is very small, each individual $\boldsymbol{x}_i$ has no neighbors other than itself, $NN_h(x_i) = 1$,

and the sDOF will be n.

We might compare this formula with the chi-squared one, $(\#\ cells - 1)$ as follows. Suppose we have a chi-squared partition with K cells and n observations. Then the average number of observations per cell, $\bar{n}$ is n/K, and the degrees of freedom is

$$sDOF = K - 1 = \frac{n}{\bar{n}} - 1.$$

That is, we see that $\frac{1}{n}\sum_i NN_h(\boldsymbol{x}_i)$ is approximately playing the role of $\bar{n}$, the average number of observations per cell. Thus we might cautiously interpret $n/sDOF$ as the average number of observations per cell using the chi-squared analogy or as the average number of neighbors of each $\boldsymbol{x}_i$, using the uniform analogy. Intuitively, then, we control the spectral degrees of freedom by ensuring that enough neighbors are being averaged together.

4.2 Theoretical sDOF for Gaussian distribution

In this section, we investigate the role of sDOF in bandwidth selection for the kernel density estimator.

Lemma 4.1. *Consider the Gaussian kernel $K_{h^2}(x, y) = \varphi(x; y, h^2)$. Assume the true distribution F is $N(0, I)$. Then $sDOF_F(K)$ is*

$$\left(\frac{1}{1 + \frac{2}{h^2}}\right)^{\frac{d}{2}}.$$

In addition, when h^2/d is fixed to be α, the sDOF converges to the constant, e^α, as d goes to ∞.

Remark 4.2. Notice that the bandwidth h must grow proportionally to $\sqrt{d}$ in order for the degrees of freedom to be approximately constant over change in dimension.

Proof. With $K_{h^2}(x,y) = \varphi(x; y, h^2)$, the sDOF can be written as

$$\frac{K_{h^2}(0,0)^2}{\int\int K_{h^2}(x,y)^2 dF(x)dF(y)} = \frac{1}{\int\int e^{-\frac{1}{h^2}(x-y)^T(x-y)} dF(x)dF(y)}.$$

Let $Z = (X-Y)^T(X-Y) \sim \chi^2(d)$. Then the denominator of the right-hand side can be expressed by

$$\begin{aligned}\int\int e^{-\frac{1}{h^2}(x-y)^T(x-y)} dF(x)dF(y) &= \int e^{-\frac{1}{h^2}z} dG(z) \\ &= \left(\frac{1}{1+\frac{2}{h^2}}\right)^{\frac{d}{2}}.\end{aligned}$$

Let $h^2 = \frac{d}{\alpha}$ in the last term above. Now we obtain

$$\begin{aligned}\int\int e^{-\frac{1}{h^2}(x-y)^T(x-y)} dF(x)dF(y) &= \left(\frac{1}{1+\frac{2\alpha}{d}}\right)^{\frac{d}{2}} \\ &= \left(1+\frac{-\alpha}{-d/2}\right)^{-\frac{d}{2}} \to e^{-\alpha} \quad \text{as} \quad d \to \infty.\end{aligned}$$

□

With simple calculation with $n = 2$ in the standard normal case, we can investigate the relationship between h^2 and d in Remark 4.2. Let $\{\boldsymbol{X}_1, \boldsymbol{X}_2\}$ be a random sample with size $n = 2$ from $N(0, I_d)$. Then the kernel density estimator is given by

$$\hat{f}(\boldsymbol{x}) = \frac{1}{2}(K_{h^2}(\boldsymbol{x}, \boldsymbol{X}_1) + K_{h^2}(\boldsymbol{x}, \boldsymbol{X}_2)). \tag{4.10}$$

By Corollary 4 in Ray and Lindsay (2005), this density estimator is unimodal when

$$(\boldsymbol{X}_2 - \boldsymbol{X}_1)^T(\boldsymbol{X}_2 - \boldsymbol{X}_1) \leq 4h^2.$$

Let h_0 be the smallest bandwidth to make the density estimator in (4.10) to be unimodal.

Then, since $\boldsymbol{X}_2 - \boldsymbol{X}_1 \sim N(0, 2I_d)$, it follows

$$2{h_0}^2 = \frac{(\boldsymbol{X}_2 - \boldsymbol{X}_1)^T(\boldsymbol{X}_2 - \boldsymbol{X}_1)}{2} \sim \chi^2_d.$$

By Fisher's approximation to the χ^2 distribution, we have

$$\sqrt{2 \cdot 2{h_0}^2} - \sqrt{2d-1} \to N(0,1) \quad \text{as } d \to \infty,$$

which can be written as

$$h_0 - \frac{1}{2}\sqrt{2d-1} \to N(0, \frac{1}{4}).$$

This approximation implies that

$$\Rightarrow P\left[h_0^2 < \frac{1}{2}\left(d - \frac{1}{2}\right)\right] \to 0.5$$

and so (the median of $h_0^2)/\frac{1}{2}d$ converges to 1. That is, the median of the smallest bandwidth that results in the unimodality of (4.10) converges to $\frac{1}{2}(d - \frac{1}{2})$ as $d \to \infty$, which is linear in d as $d \to \infty$.

This calculation can be generalized to other X-distributions by using the asymptotic normality of the sum of squares $(\boldsymbol{X}_1 - \boldsymbol{X}_2)^T(\boldsymbol{X}_1 - \boldsymbol{X}_2)$. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a random vector from any arbitrary distribution and the component of $\boldsymbol{X}$ are independent and identically distributed. Again consider a random sample with size $n = 2$, $\{\boldsymbol{X}_1, \boldsymbol{X}_2\}$. Then h_0 is again written as

$$h_0^2 = \frac{\boldsymbol{Y}^T\boldsymbol{Y}}{4}$$

where $\boldsymbol{Y} = \boldsymbol{X}_1 - \boldsymbol{X}_2$. Since each component of $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^T$, is independent and identically distributed, we have

$$\frac{\frac{1}{d}h_0^2 - \mu}{\sigma/\sqrt{d}} = \frac{h_0^2 - d\mu}{\sqrt{d}\sigma} \to N(0,1)$$

where $\mu = E(Y_1^2/4)$ and $\sigma^2 = Var(Y_1^2/4)$. In addition, μ can be written as

$$\mu = \frac{E(Y_1^2)}{4} = \frac{2\,Var(X_1)}{4} = \frac{1}{2}Var(X_1).$$

Thus the asymptotic median of h_0^2 is $d\,Var(X_1)/2$. As the previous example with the standard normal random sample, this confirms that the median of the smallest bandwidth for obtaining unimodality of the kernel density estimator is approximately linear in d.

In the previous section, it was noted that sDOF can be interpreted approximately as the average number of neighbors of each data point. With the random sample with $n = 2$, we calculated analytically the smallest bandwidth for attaining the unimodality of the kernel density estimator so that these two observations are considered as a neighbor each other. Then it showed that the median of the squared bandwidth h_0^2 is linear in d. This confirms that h_0^2 needs to grow proportionally in d to obtain the same amount of smoothing as d grows, which is consistent with Remark 4.2.

4.3 Simulation Study

In this section, we investigate the behavior of sDOF and the mode association clustering (MAC) algorithm explained in Chapter 3. We generated $R = 500$ random samples with size $n = 100$ and $n = 500$ from $N(0, I)$ for dimension 1,5, 10, 15, 20 and 25. With a grid of bandwidth h, we performed the hierarchical mode association clustering based on the kernel density estimator. Then we observed the smallest bandwidths that merged the data to one modal. That is, we found the smallest bandwidth that coalesced the kernel density estimator to a unimodal density estimate.

Table 4.1 includes the mean and standard error of the bandwidths to provide coalescence to a single mode for the sample sizes $n = 100$ and $n = 500$. Note that the sample size is fixed over the various dimensions so that any asymptotic results would here depend on $d \to \infty$. When $n = 100$, we observed that $\sqrt{d/6}$ formula matches closely with the average bandwidth merging all points into one group. Similarly, when $n = 500$,

	n=100				n=500			
d	mean(h)	se(h)	$\sqrt{d/6}$	asym. sDOF	mean(h)	se(h)	$\sqrt{d/7}$	asym. sDOF
1	0.3588	0.0030	0.4082	2363.1436	0.2867	0.0033	0.3780	192126.7162
5	0.8824	0.0038	0.9129	614.9477	0.7932	0.0031	0.8452	2827.1524
10	1.2738	0.0034	1.2910	474.8865	1.1602	0.0029	1.1952	1684.2355
15	1.5600	0.0032	1.5811	475.1868	1.4264	0.0028	1.4639	1591.4408
20	1.7908	0.0030	1.8257	511.0292	1.6391	0.0025	1.6903	1709.9480
25	2.0031	0.0029	2.0412	508.0958	1.8208	0.0023	1.8898	1883.2739

Table 4.1: Bandwidth and asymptotic sDOF for 100% coalescence.

$\sqrt{d/7}$ is close to the average bandwidth. This confirms Remark 4.2 that mentioned that the bandwidth needs to be grow proportionally to $\sqrt{d}$ if we are to obtain the same amount of smoothing. However, this result also shows that the rate of increment varies over the sample size. For $n = 100$, the slope was roughly $1/\sqrt{6}$ while it was $1/\sqrt{7}$ for $n = 500$.

For both sample sizes, sDOF calculated by the formula in Lemma 4.1 is also shown in Table 4.1. Though sDOF in dimension one is quite larger than the other dimensions, sDOF tends to be stable as dimension grows. That is, the bandwidths for obtaining the same amount of smoothing has approximately constant sDOF, which implies controlling sDOF could be a method to control the amount of smoothing in higher dimensions.

In Figure 4.1, one can find the scatter plot of h over $\sqrt{d}$ in Table 4.1, and the regression lines based on these points. For both sample sizes, the points are located very close to the indicated line. Compared to $n = 500$, the line is slightly steeper when $n = 100$. It means that we need to increase h faster as dimension grow when the sample size is smaller. This is consistent with the MISE optimal bandwidth result in Section 3.4.2. The partial derivative of the logarithm of the optimal bandwidth for in (3.24) with respect to d becomes

$$\frac{\partial \log h_{opt,KER}(d,n)}{\partial d} = C - \frac{\log n}{(d+4)^2}.$$

This implies that the optimal bandwidth for $\hat{f}_{KER}$ to minimize the AMISE increases faster with d when the sample size is smaller.

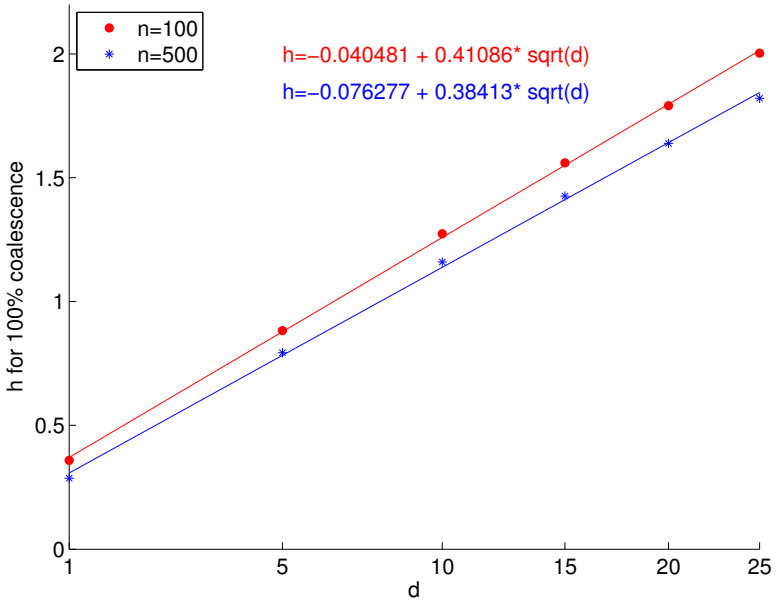


Figure 4.1: Relationship between bandwidths to attain 100% coalescence and dimensions

The increment of h also depends on the coalescence level. Table 4.2 contains the simulation result for 50% coalescence level. That is, the average of the smallest bandwidths that have merged at least 50% of the data into one group are presented. When $n = 100$, h increased approximately with slope $1/\sqrt{9}$ while it increased with slope $1/\sqrt{12}$ when $n = 500$. Again, the smaller sample size required the faster increment in h. In addition, compared to the result in Table 4.1, the 50% coalescence had the smaller slope than 100% coalescence. For the 50% coalescence level, sDOF does not look stable as clear as for the 100% coalescence level. This could be affected by the fact that the exponent in the convergent formula of sDOF in Lemma 4.1, $e^{h^2/d}$, exaggerates the difference in h^2/d. If one calculates h^2/d instead of the exponent of it, it is more clear that this value tends to be constant as dimension grows.

In Figure 4.2, the scatter plot is not quite as well aligned with the regression line compared to Figure 4.1, being slightly convex. Note though that the plot goes up to $d = 25$, so near linearity holds over a long range.

Remark 4.3. The preceding results are quite amazing in their near perfect linearity.

	n=100				n=500			
d	mean(h)	se(h)	$\sqrt{d/9}$	asym. sDOF	mean(h)	se(h)	$\sqrt{d/12}$	asym. sDOF
1	0.2106	0.0022	0.3333	6.1932E+09	0.1574	0.0013	0.2887	3.3862E+17
5	0.6117	0.0023	0.7454	6.3582E+05	0.4791	0.0015	0.6455	2.8856E+09
10	0.9579	0.0016	1.0541	5.4086E+04	0.7842	0.0009	0.9129	1.1536E+07
15	1.2365	0.0012	1.2910	1.8229E+04	1.0490	0.0006	1.1180	8.3184E+05
20	1.4803	0.0013	1.4907	9.2009E+03	1.2656	0.0005	1.2910	2.6471E+05
25	1.6864	0.0013	1.6667	6.5722E+03	1.4583	0.0005	1.4434	1.2747E+05

Table 4.2: Bandwidth and asymptotic sDOF for 50% coalescence.

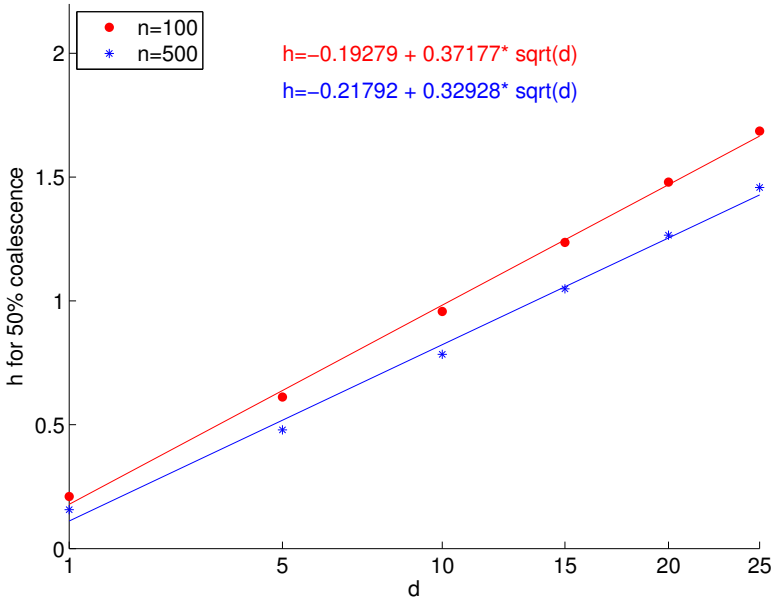


Figure 4.2: Relationship between bandwidths to attain 50% coalescence and dimensions

They suggest that a careful asymptotic analysis might prove a fruitful way to better understand the coalescence property of the kernel density estimator, and give results with high accuracy across many dimensions.

4.4 Modal Estimate in Higher Dimensions

In the previous section, we considered how the bandwidth should grow when dimension grows. This section applies this result to compare the performance of the density estimators in Chapter 3 on mode estimation problems. We generated $R = 500$ samples with

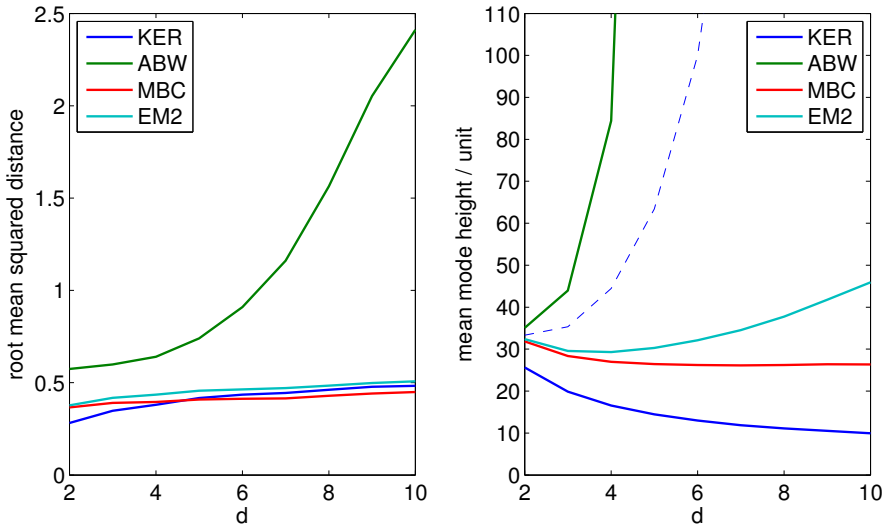


Figure 4.3: Mode Estimates for $N(0, I)$.

the sample size $n = 100$ from $N(0, I)$. Considering dimension 2 to 10, we increased the bandwidth as dimension grows with $h = \sqrt{d/6}$ found in the previous section. Then we fit the density estimates from the four density estimators used in Chapter 3 and found the mode of them using the modal EM algorithm starting from the true mode. The left plot in Figure 4.3 includes the average of $\|\hat{m} - m\|$ where $\hat{m}$ is the estimated mode from the density estimate and $m = 0$ is the true mode of the normal. Except for $\hat{f}_{ABW}$, the other three estimators are not significantly different in estimating the location of the mode although $\hat{f}_{EM2}$ had slightly larger error than $\hat{f}_{KER}$ and $\hat{f}_{MBC}$.

We define $\frac{1}{n}K(\hat{m}, \hat{m})$ to be one unit. This is the approximate height of the kernel density estimator at an outlier observation. With this measure, we count height of modes of $\hat{f}$ in units, thereby representing how high they are relatively to the density of modal outliers. Since the density f has a lower height in higher dimensions, the value $|\hat{f}(x) - f(x)|$ does not give an appropriate measure to compare the performance of the density estimators over various dimensions. Thus by measuring density heights in units, we might also have a fairer comparison on the mode estimates over dimensions.

The right plot in Figure 4.3 shows the mean of $\hat{f}(\hat{m})$ divided by the unit $\frac{1}{n}K(\hat{m}, \hat{m})$. Since the dashed line represents the true mode height $f(0)$ in units, the estimator with

the closest line has the better performance at estimating the height of the mode. Note that $h = \sqrt{d/6}$ ranged between 0.58 and 1.29 in dimensions 2 to 10, and in higher dimensions, the number of units in the true mode height rises rapidly. In this plot, we can observe that $\hat{f}_{EM2}$ performed better than the other density estimators at capturing the true modal height. In addition, $\hat{f}_{EM2}$ tends to improve its performance as dimension grows.

We think that this analysis is suggestive, but still very preliminary. Our case study of the multivariate normal has the unfortunate feature that the optimal bandwidth for estimating the mode is $h \approx \infty$, which, for the kernel density estimator, gives a modal estimator of $\bar{x}$, the normal MLE, which is fully efficient. Our observed comparison of estimators then could be affected by the fact we used the same bandwidth for all the estimators. However, it is clear that the $\hat{f}_{EM2}$ has less bias at $x = 0$. Thus, to be comparable to the other estimators in bias, it would require a larger bandwidth in bias, which would make it more accurate at estimating $m = 0$.

4.5 Future Work

The simulation results in Section 4.4 leave many rooms to be examined. The strange behavior of $\hat{f}_{ABW}$ should be more carefully examined. We are simply seeing bandwidth effectsm not actually sensitivity That is, here we used the same bandwidth for all the density estimators. Considering that $\hat{f}_{MBC}$ had the largest optimal bandwidth in the simulation in Chapter 3, using the same bandwidth as the others might give it a disadvantage. However, the bias results suggest otherwise. It is clear that we need to investigate different densities starting from a bimodal density, so that the optimal bandwidth for modal estimation is not $h \approx \infty$.

Chapter 5

Penalized Likelihood-tuned Density Estimator

5.1 Introduction

In Chapter 2 and 3, we discussed the nonparametric mixture model

$$f(x;\Pi) = \int K(\boldsymbol{x};\boldsymbol{\phi})d\Pi(\boldsymbol{\phi}). \tag{5.1}$$

In Section 3.2.3, we specified the kernel to be Gaussian, $K_H(\cdot,\boldsymbol{\mu}) = \varphi(\cdot;\boldsymbol{\mu},H)$, which is one of the Markov kernel. In this model, we regarded the location parameter $\boldsymbol{\mu}$ as a mixing variable $\boldsymbol{\phi}$ in (5.1), so the model can be written as

$$f(x;\Pi) = \int K_H(\boldsymbol{x};\mu)d\Pi(\boldsymbol{\mu}). \tag{5.2}$$

In this way, we put flexibility on the location of the mixing components $K_H(\cdot;\boldsymbol{\mu})$ and investigated where these locations should be placed in a support of Π by EM iterations. In the meantime, the shape and scale of the kernel to be fixed by a predetermined bandwidth matrix H. In this section, we allow mixtures not only of locations but also shape and scale parameters so that components in the nonparametric mixture model can

potentially capture local variance and shape features. This extends (5.2) to

$$f(x;\Pi) = \int K(x;\boldsymbol{\mu},\Sigma)d\Pi(\boldsymbol{\mu},\Sigma). \tag{5.3}$$

Here $(\boldsymbol{\mu},\Sigma)$ corresponds to a mixing variable $\boldsymbol{\phi}$ in (5.1).

A problem arises from the fact that the model in (5.3) is fully unrestricted for both $\boldsymbol{\mu}$ and Σ, which causes the likelihood

$$L(\Pi;X_i) = \int K(X_i;\boldsymbol{\mu},\Sigma)d\Pi(\boldsymbol{\mu},\Sigma) \tag{5.4}$$

to be unbounded. (This is easy to see in a finite mixture model when a component density is degenerate on an observation. See Kiefer and Wolfowitz (1956).) Therefore, there does not exist a maximum likelihood estimator of Π. To address this problem, we consider adding a "penalty-like" term $P(\Sigma)$ in the individual likelihood function, so that

$$L(\Pi;X_i) = \int K(X_i;\boldsymbol{\mu},\Sigma)P(\Sigma)d\Pi(\boldsymbol{\mu},\Sigma). \tag{5.5}$$

Here $P(\cdot)$ is referred to a "penalty-like" term instead of a penalty term since it does not exactly penalize small σ^2 but instead allows us to control the value of σ^2 in the fitting procedure. One of the key goals, of course, is to choose $P(\Sigma)$ so that the penalized likelihood generated by (5.5) is bounded. Then one can directly apply the NPMLE theory and the EM algorithm of our preceding work. Note that if $P(\Sigma)$ is very nearly a spike density at $\Sigma_0 = h^2 I$, then this problem corresponds to our fixed bandwidth in Chapter 3. The less spiky $P(\Sigma)$ is, the more it will allow an adaptation to Σ in a neighborhood of each $\boldsymbol{x}_i$.

As an example of a "penalty-like" term, one can consider $Gamma(1,\theta)$ on σ^{-2} for the univariate case, given by

$$P(\sigma^{-2}) = \frac{1}{\theta}\exp\left(-\frac{1}{\beta\sigma^2}\right).$$

Since the $Gamma(\sigma^{-2}; 1, \theta)$ density is monotone increasing in σ^2, the penalized likelihood function in (5.5) tends to prefer large σ^2. However, a *Gamma* distribution is not guaranteed to be monotone increasing with different parameters. In this chapter, we do not restrict *Gamma* to be monotone but allow it to have two free parameters, and investigate their role in density estimation. Thus, strictly speaking, $P(\cdot)$ does not penalize on small variance, but instead gives a likelihood "boost" to those values of Σ that have the highest density $P(\Sigma)$.

Several authors have addressed the likelihood degeneracy problem in the context of finite mixture models. Hathaway (1985) proposed constrained maximum likelihood estimation for normal mixture distributions. In a mixture of m univariate normal densities, give by

$$f(x) = \sum_{j=1}^{m} \pi_j \varphi(x; \mu_j, \sigma_j),$$

he restricted $\min_{i,j}(\sigma_i/\sigma_j) \geq c > 0$. Then the existence of a constrained global maximizer and its consistency were proved.

Ciuperca et al. (2003) addressed the likelihood degeneracy problem with a penalized likelihood. They considered a penalized likelihood, given as

$$L(\boldsymbol{\mu}, \boldsymbol{\sigma}; X_1, \ldots, X_n) \prod_{j=1}^{m} g(\sigma_j)$$

where $L(\boldsymbol{\mu}, \boldsymbol{\sigma}^2; X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i)$ and g is a penalty on individual σ_j with some assumptions. They showed consistency and asymptotic efficiency of the penalized MLE. In numerical examples, they compared Hathaway (1985)'s constrained method with their penalized likelihood taking the inverse-gamma distribution for $g(\boldsymbol{\sigma}^2)$. It was shown that Hathaway's method had a restriction when the true component variances did not belong to the constrained parameter space, while their method gave reasonable point estimates.

The likelihood degeneracy problem for the mixture model in higher dimensions has not been well investigated yet. Ingrassia and Rocci (2007) extended Hathaway (1985)'s

constrained method to

$$a \leq \lambda_i(\Sigma_j) \leq b.$$

Then they provided an EM algorithm to find the maximum likelihood estimator in constrained parameter space. Warton (2008) investigated the multivariate penalized likelihood in the context of the multivariate regression with a response random vector $\boldsymbol{Y}$. He showed that $\hat{\Sigma}$ maximizing the penalized likelihood, given by

$$\log L(\boldsymbol{\mu}, \Sigma; \boldsymbol{Y}) = \sum_{i=1}^{n} \varphi(\boldsymbol{Y}; \boldsymbol{\mu}, \Sigma) - \frac{c}{2} tr(\Sigma^{-1})$$

was the ridge estimator $\hat{\Sigma}_\lambda = \hat{\Sigma} + \lambda \boldsymbol{I}$, where $\hat{\Sigma}$ is the maximum likelihood estimator. Then he proposed to estimate λ by K-fold cross-validation.

One should note that the existing penalized likelihood methods described above have a penalty term subtracted from a log-likelihood function. Therefore they need to adjust the penalty function so that the penalized likelihood is bounded as in Ridolfi and Idier (2001). On the other hand, our method includes a penalty-like term inside of individual likelihoods $L(X_i)$ as shown in (5.5). Thus, as far as the individual likelihood (5.5) is bounded, it guarantees the existence of the global maximum. In addition, since this penalized likelihood keeps the general form of the nonparametric likelihood function, written as

$$L(\Pi; X_i) = \int K(X_i, \phi) d\Pi(\phi),$$

the theory in Lindsay (1995) for the nonparametric likelihood methods could be applied for this problem.

In Section 5.2, we describe the details of nonparametric density estimation based on the penalized likelihood-tuning in both univariate and multivariate cases. In Section 5.3, we investigate a connection of the penalized likelihood-tuning to the (regular) likelihood-tuning in Chapter 2 and 3. Section 5.4 shows simulation examples to compare their performance.

5.2 Methodology

For the univariate case, given a fixed bandwidth h^2, let $P(\delta)$ be $Gamma(\alpha, \frac{1}{\alpha h^2})$, given by

$$P(\delta) = \frac{(\alpha h^2)^\alpha}{\Gamma(\alpha)} \delta^{\alpha-1} \exp\left(-\alpha h^2 \delta\right)$$

where $\delta = \sigma^{-2}$. Here α will be a second tuning parameter. That is, we consider a penalized likelihood function for X_i, given by

$$L(\Pi; X_i) = \int K(X_i; \mu, \delta^{-1}) Gamma(\delta; \alpha, \frac{1}{\alpha h^2})\, d\Pi(\mu, \delta). \tag{5.6}$$

Here we employ the parameterization $Gamma(\alpha, \frac{1}{\alpha h^2})$, instead of a conventional parameters $Gamma(\alpha, \beta)$. This parameterization gives $E(\delta) = \frac{1}{h^2}$ and $Var(\delta) = \frac{1}{\alpha h^4}$. That is, we set $\frac{1}{\sigma^2}$ to be, on average, $\frac{1}{h^2}$ and allow variability with a variance $\frac{1}{\alpha h^4}$. Therefore, when $\alpha \to \infty$ with h^2 fixed, $Gamma(\alpha, \frac{1}{\alpha h^2})$ degenerates to h^{-2}, and the penalized likelihood in (5.6) becomes the likelihood in Chapter 2, given by

$$L(\Pi; X_i) = \int K_h(X_i; \mu) d\Pi(\mu),$$

where the component variance is fixed to be a bandwidth h. The advantage of the *Gamma* distribution is that it is the Bayesian conjugate prior for $\delta = \sigma^{-2}$ in $N(\mu, \sigma^2)$. Thus with this choice, we can calculate integrals in the likelihood-tuning steps analytically and the resulting estimators are in closed form.

Now we apply the likelihood-tuning procedure, introduced in Chapter 2, to the penalized likelihood function in (5.6). Here we restate the likelihood-tuning procedure in terms of the penalized likelihood for the Gaussian kernel and *Gamma* penalty-like term. Let the kernel function be $K(x; \mu, \delta^{-1})$ the probability density function for $N(\mu, \delta^{-1})$. Given an initial estimate $\pi_0(\mu, \delta)$,

1. Update the estimator of the mixing density π by

$$\hat{\pi}_{(k+1)}(\mu, \delta) = \hat{\pi}_{(k)}(\mu, \delta)\Delta_{(k)}(\mu, \delta),$$

where

$$\Delta_{(k)}(\mu, \delta) = \sum_{i=1}^{n} \frac{K(X_i; \mu, \delta^{-1})Gamma(\delta; \alpha, \frac{1}{\alpha h^2})}{L_i(\hat{\Pi}_{(k)})}$$

and

$$L_i(\hat{\Pi}_{(k)}) = \int K(X_i; \mu, \delta^{-1})Gamma(\delta; \alpha, \frac{1}{\alpha h^2})\, d\mu\, d\delta$$

2. Update the density estimator of x by

$$\hat{f}_{(k+1)}(x) = \int K(x; \mu, \delta^{-1})\hat{\pi}_{(k+1)}(\mu, \delta)\, d\mu\, d\delta.$$

As an initial estimate for $\pi(\mu, \delta)$, we consider $\pi_0(\mu, \delta) = 1 \cdot \delta$. Here, '1' is an initial mixing distribution for μ and 'δ' is an initial for δ, both of which are Jeffeys' non-informative prior. As in Chapter 2, it is assumed that we do not have any prior information about the mixing distribution for μ and δ. Then the likelihood-tuning procedure will indicate where the mass for (μ, δ) should be located to be closer to the discrete NPMLE.

With the Gaussian kernel and the *Gamma* penalty-like term, integrals in the likelihood-tuning steps can be expressed in closed forms. The first tuning step gives

$$\hat{\pi}_{(1)}(\mu, \delta) = \frac{1}{n}\sum_{i} \delta h^2 K(x_i; \mu, \frac{1}{\delta})Gamma(\delta; \alpha, \frac{1}{\alpha h^2})$$

and

$$\hat{f}_{(1)}(x) = \frac{1}{n}\sum_{i} \sqrt{\frac{(\alpha+1)}{2\alpha h^2}}\, t\left(\sqrt{\frac{(\alpha+1)}{2\alpha h^2}}(x - x_i); 2\alpha + 2\right).$$

As in Chapter 2, the first penalized likelihood-tuning step gives a standard kernel density estimator. While $\hat{f}_{EM1}$ in Chapter 2 was the kernel density estimator with the Gaussian

kernel, here we obtain $\hat{f}_{(1)}$ that is the kernel density estimator with t-kernel. We call it as the first penalized likelihood-tuned density estimator (PEM1), denoted by $\hat{f}_{PEM1}(x)$.

The second-tuning gives

$$\hat{\pi}_{(2)}(\mu,\delta) \quad = \quad \frac{1}{n^2}\sum_i\sum_j \frac{\delta h^2}{L_j(\pi_1)} K(x_i;\mu,\frac{1}{\delta})K(x_j;\mu,\frac{1}{\delta})Gamma(\delta;\alpha,\frac{1}{\alpha h^2})^2$$

and

$$\hat{f}_{(2)}(x) = \frac{1}{n^2}\sum_i\sum_j w_{ij}\sqrt{\frac{4(4\alpha+1)}{3((x_i-x_j)^2+8\alpha h^2)}}t\left(\sqrt{\frac{4(4\alpha+1)}{3((x_i-x_j)^2+8\alpha h^2)}}(x-\bar{x}_{ij});4\alpha+1\right)$$

where

$$w_{ij} = \frac{\sqrt{\frac{1}{2h^2}}t\left(\sqrt{\frac{1}{2h^2}}(x_i-x_j);4\alpha\right)}{\frac{1}{n}\sum_{k=1}^n \sqrt{\frac{1}{2h^2}}t\left(\sqrt{\frac{1}{2h^2}}(x_k-x_j);4\alpha\right)}.$$

The second penalized likelihood-tuned estimator $\hat{f}_{(2)}(x)$ is denoted by $\hat{f}_{PEM2}(x)$.

The penalized likelihood-tuning procedure can be easily extended to the multivariate case. We can replace the *Gamma* penalty-like term with a Wishart distribution on Σ^{-1}. With a scalar m and a matrix V, the $Wishart(m,V)$ density is defined on a matrix U with a probability density function

$$f(U) = \frac{|U|^{(n-p-1)/2}\exp\left[-\frac{1}{2}trace(V^{-1}U)\right]}{2^{np/2}|V|^{n/2}\Gamma_d(\frac{m}{2})}.$$

When dimension $d=1$, $Wishart(m,V)$ is reduced to $Gamma(\frac{m}{2},2V)$.

Define U to be Σ^{-1}. We consider $P(U) = Wishart(U;m,\frac{1}{mh^2}I_d)$ as a penalty-like term in (5.5), which leads to the penalized likelihood, written as

$$L(X_i;\Pi) = \int K(X_i;\boldsymbol{\mu},U^{-1})Wishart(U;m,\frac{1}{mh^2}I_d)d\Pi(\boldsymbol{\mu},U). \tag{5.7}$$

Since $Wishart(U;m,\frac{1}{mh^2}I)$ has $E(U)=\frac{1}{h^2}I$ and $Var(U_{ij})=\frac{1}{mh^4}$, this parameterization

implies that Σ^{-1} is expected to be $\frac{1}{h^2}I$ on average and each (i,j)-th element in Σ^{-1} has a variability $\frac{1}{mh^4}$ around the mean. As the univariate penalized likelihood does, the penalized likelihood in (5.7) converges to the nonparametric likelihood function in Chapter 3, given as

$$L(X_i;\Pi) = \int K_{h^2 I}(X_i;\boldsymbol{\mu})dP(\boldsymbol{\mu})$$

when $Wishart(U;m,\frac{1}{mh^2}I)$ degenerates to $\frac{1}{h^2}I$ by letting $m \to \infty$.

For an initial $\hat{\pi}_{(0)}$, again we use Jeffreys' non-informative priors for $\boldsymbol{\mu}$ and U, given by

$$\pi_0(\boldsymbol{\mu},U) = 1 \cdot |U|^{\frac{d+1}{2}}.$$

The first tuning step gives

$$\hat{\pi}_{(1)}(\boldsymbol{\mu},U) = \frac{1}{n}\frac{1}{L_i(\pi_0)}\sum_i |U|^{\frac{d+1}{2}} K(x_i;\boldsymbol{\mu},U^{-1})Wishart(U;m,\frac{1}{mh^2}I)$$

and

$$\hat{f}_{PEM1}(\boldsymbol{x}) = \frac{1}{n}\sum_i T\left(\boldsymbol{x}-\boldsymbol{x}_i; m+2, \frac{2mh^2}{m+2}I\right).$$

With degrees of freedom n and covariance matrix Σ, let $T(\cdot;n,\Sigma)$ represents the probability density function of the multivariate t-distribution, defined by the random variable $\boldsymbol{y}/\sqrt{u/n}$, where $\boldsymbol{y} \sim N(\boldsymbol{0},\Sigma)$, $u \sim \chi^2_n$, and $\boldsymbol{y}$ and u are independent.

The second tuning step gives

$$\hat{\pi}_{(2)}(\boldsymbol{\mu},U) = \frac{1}{n^2}\frac{1}{L_i(\pi_0)L_j(\pi_1)}\sum_i\sum_j |U|^{\frac{d+1}{2}} K(x_i;\boldsymbol{\mu},U^{-1})K(x_j;\boldsymbol{\mu},U^{-1})Wishart(U;m,\frac{1}{mh^2}I)^2$$

and

$$\hat{f}_{PEM2}(\boldsymbol{x}) = \frac{1}{n^2}\sum_i\sum_j w_{ij}\cdot T\left(\boldsymbol{x}-\bar{\boldsymbol{x}}_{ij}; 2m-d+2, \frac{3}{4(2m-d+2)}[(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T + 4mh^2 I]\right) \tag{5.8}$$

where

$$w_{ij} = \frac{t(\boldsymbol{x}_i - \boldsymbol{x}_j; \frac{4mh^2}{2m-d+1} I, 2m-d+1)}{\frac{1}{n}\sum_k t(\boldsymbol{x}_k - \boldsymbol{x}_j; \frac{4mh^2}{2m-d+1}, 2m-d+1)}.$$

The covariance parameter of t-kernel in (5.8) includes the term $(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T$, which corresponds to the local shape and scale from a pair of data points. Therefore, we expect that the second penalized likelihood-tuning would take local shapes and scales into account in the fitting process.

The following lemma shows the relationship between the unpenalized likelihood-tuned estimators in Chapter 3, and the penalized likelihood-tuned estimators above. As mentioned earlier, when the penalty-like term $Wishart(U; m, \frac{1}{mh^2} I)$ degenerates to its mean $\frac{1}{h^2} I$, the penalized likelihood converges to the unpenalized likelihood.

Lemma 5.1. *Let h to be fixed and $m \to \infty$. Then $\hat{f}_{PEM1}(x)$ converges to $\hat{f}_{EM1}(x)$ and $\hat{f}_{PEM2}(x)$ converges to $\hat{f}_{EM2}(x)$ for each x.*

Proof. It is straightforward by the approximation of a t-distribution to a normal. □

5.3 Simulation

In this section, we compare $\hat{f}_{PEM2}$ to $\hat{f}_{EM2}$ in a simple simulation study. We considered a two-component Gaussian mixture density, $f(x) = 0.5N(0, 0.1) + 0.5N(5, 1)$. As shown in Figure 5.1, each component is centered at 0 and 5 with extremely different scale. We examine whether $\hat{f}_{PEM2}$ could have an advantage in estimating this density f by somehow adjusting to the local scale in the model.

We generated $R = 100$ random samples from $f(x)$ with the sample size $n = 300$ and fit the density with four density estimator in Chapter 3. The bandwidth was fixed to be the optimal bandwidth for $\hat{f}_{KER}$ in (3.24) and the degree of freedom for $\hat{f}_{PEM2}$ was chosen to be 5, 10 and 15. Then the mean squared errors (MSE) were estimated at both component mean, $x = 0$ and $x = 5$ by averaging $(\hat{f}(x) - f(x))^2$ over replicates.

Figure 5.2 contains MSE of at $x = 0$ with standard error bars. At the spikier mode

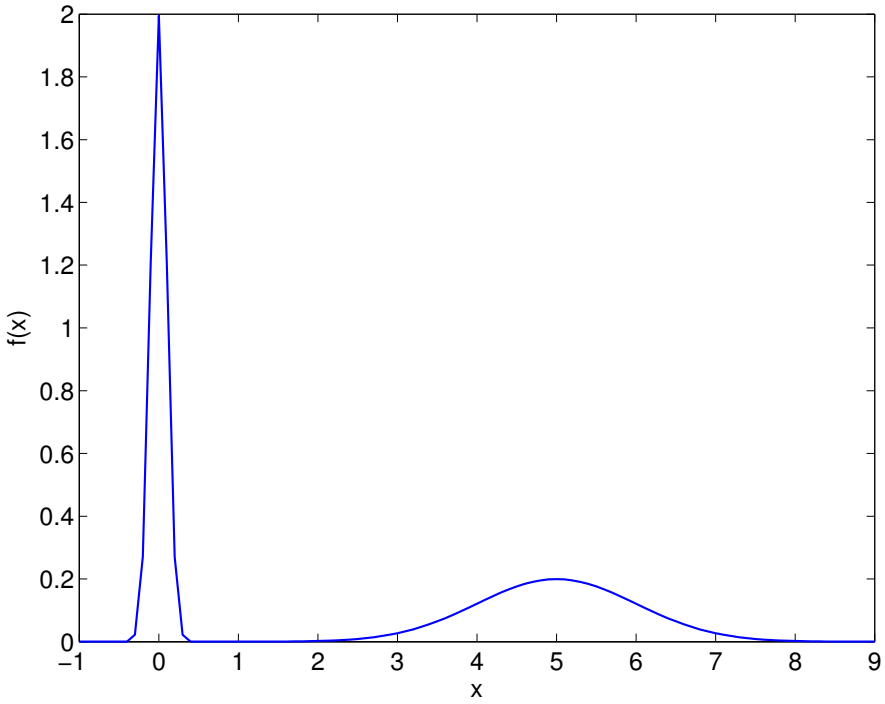


Figure 5.1: Two-component Gaussian mixture density with unequal variances.

$x = 0$, $\hat{f}_{PEM2}$ and $\hat{f}_{EM2}$ had obviously smaller MSE than $\hat{f}_{KER}$. When the degree of freedom is 5 for $\hat{f}_{PEM2}$, it had the smallest MSE. As the degree of freedom of $\hat{f}_{PEM2}$ increases, the MSE approached to $\hat{f}_{EM2}$, which is a consistent result with Lemma 5.1.

At $x = 5$, Figure 5.3 had a similar pattern to Figure 5.2. However, since all the differences between the density estimators were within two standard errors, we cannot conclude that $\hat{f}_{PEM2}$ or $\hat{f}_{EM2}$ was significantly better at estimating the flatter mode.

Though $\hat{f}_{PEM2}$ performed better in estimating the sharp mode, the fixed bandwidth could have an large influence on this result. Therefore it is more reasonable to consider flexible bandwidths for each density estimator in the simulation study. Now we simulated $R = 100$ random samples with size $n = 100$ from $0.5N(0, 0.1^2) + 0.5N(5, 1)$, and then calculated MISE over a grid of h. Again three different degrees of freedom (5, 10, 15), were considered for $\hat{f}_{PEM2}$.

In Figure 5.4, there are two clusters of lines; one for $\hat{f}_{EM1}$ and $\hat{f}_{PEM1}$'s, and the other for $\hat{f}_{EM2}$ and $\hat{f}_{PEM2}$'s. It is clear that the first-tuning estimators had larger MISE than

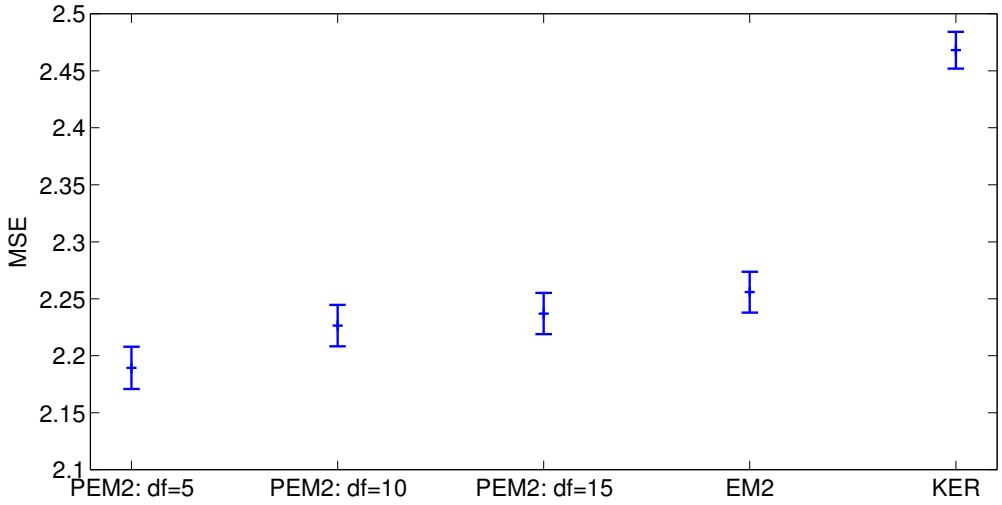


Figure 5.2: MSE at $x = 0$

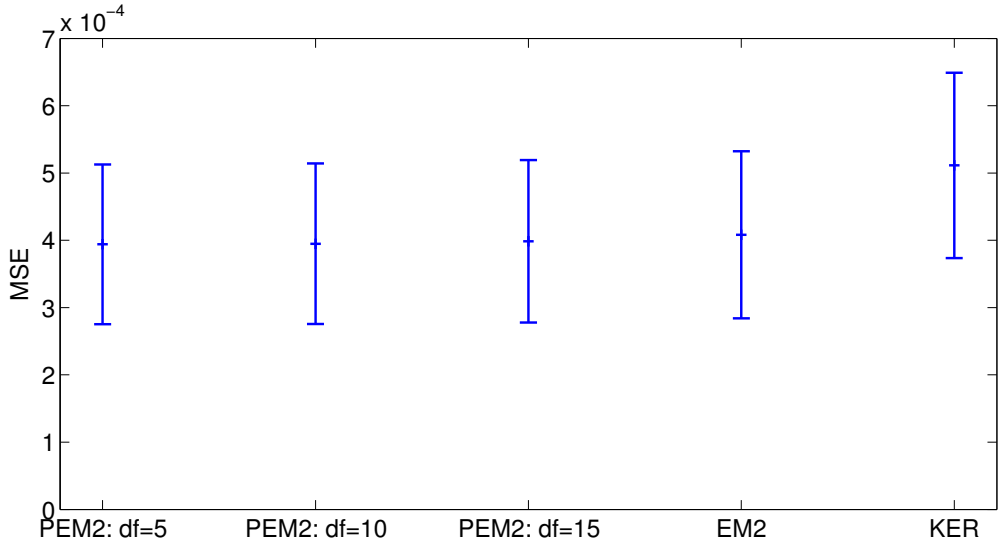


Figure 5.3: MSE at $x = 5$

the second-tuning estimators, regardless of penalized-tuning or unpenalized-tuning. This implies that the likelihood-tuning procedure reduced the bias as the asymptotic result for $\hat{f}_{EM2}$ had shown while the increment of variance were comparably small. In addition, the minimum MISEs were obtained with smaller bandwidths for the first-tuning estimators.

Among the first-tuning estimators, $\hat{f}_{EM1}$ had the smallest MISE and $\hat{f}_{PEM1}$ with $df = 15$ had the largest MISE. Similar to Figure 5.2 and 5.3, the MISE of $\hat{f}_{PEM1}$ moved

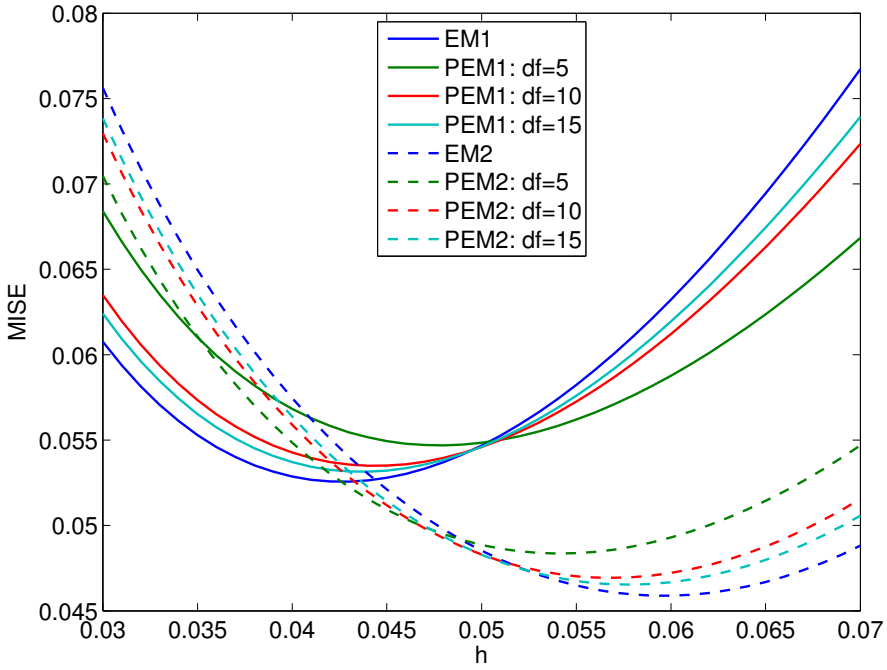


Figure 5.4: MISE over a grid of h for $0.5N(0, 0.1^2) + 0.5N(5, 1)$.

toward the MISE of $\hat{f}_{EM1}$ as the degrees of freedom increased. The second-tuning estimators also behaved in the same pattern. This is contradictory to our expectation that the penalized likelihood-tuned estimators might adjust the different scales in the data and, so would reduce the MISE. We infer that the penalized likelihood-tuning procedure overfit the model, so the model complexity increased variance more than it reduced bias.

5.4 Future Works

We note that our first preliminary test of the penalized method was unsuccessful. It is possible that in higher dimensions the flexibility given by local shape adaptation could increase the usefulness of this method. This remains to be investigated.

We can consider an alternate formulation that would be to specify a two stage mixing

distribution, so that

$$\begin{aligned} X|\mu,\Sigma_1 &\sim K(x;\mu,\Sigma_1) \\ \Sigma_1|\Sigma &\sim \textit{Inverse Wishart} \text{ with mean } \Sigma \\ \mu &\sim \Pi(\mu) \end{aligned}$$

where $\Sigma = h^2 A$ is a fixed bandwidth matrix. This formulation would have some similarities to the our present treatment (5.5), with $K(x_i;\mu,\Sigma)P(\Sigma)$ in (5.5) replaced with $\int K(x_i;\mu,\Sigma_1)P(\Sigma_1|\Sigma)d\Sigma_1$. The possible advantage of this treatment is that we would be working with a true likelihood. However, the likelihood-tuning procedure would not yield a new class of density estimators.

The penalized likelihood idea in this chapter can be also applied to the finite mixture model. Considering a penalty term $P(\sigma_k^{-1}) = Gamma(1, 1/\lambda)$ in the individual likelihood function, it can be written as

$$L(\boldsymbol{\mu},\boldsymbol{\sigma},\lambda;x_i) = \sum_{k=1}^{p} \pi_k \varphi(x_i;\mu_k,\sigma_k) e^{-\lambda/\sigma_k^2}.$$

The EM algorithm to find the maximum likelihood estimator in the above likelihood function is the follow.

1.
$$\hat{\gamma}_{ik} = \frac{\pi_k \varphi(x_i;\mu_k,\sigma_k)}{\sum_{j=1}^{p} \pi_j \varphi(x_i;\mu_j,\sigma_j)}$$

2.
$$\hat{\mu}_k = \left(\sum_{i=1}^{n} \hat{\gamma}_{ik}\right)^{-1} \sum_{i=1}^{n} \hat{\gamma}_{ik} x_i$$

$$\hat{\sigma}_k = \left(\sum_{i=1}^{n} \hat{\gamma}_{ik}\right)^{-1} \left(\sum_{i=1}^{n} \gamma_{ik}(x_i - \hat{\mu}_k)^2\right) + 2\lambda$$

$$\pi_k = \sum_{i=1}^{n} \gamma_{ik}/N$$

In the update of σ_k, 2λ was added to the regular EM update. If we consider the penalty term in the form of Ciuperca et al. (2003), given by

$$L(\boldsymbol{\mu}, \boldsymbol{\sigma}, \lambda) = \prod_{i=1}^{n} \left[\sum_{k} \pi_k \varphi(x_i; \mu_k, \sigma_k) - \lambda \sum_{k} \frac{1}{\sigma_k} \right],$$

the EM update for $\hat{\sigma}_k$ becomes $\hat{\sigma}_k = \left(\sum_{i=1}^{n} \hat{\gamma}_{ik}\right)^{-1} \left(\sum_{i=1}^{n} \gamma_{ik}(x_i - \hat{\mu}_k)^2 + 2\lambda\right)$. We plan to investigate the properties of the penalized likelihood on finite mixture models.

Chapter 6

Summary and Future Work

This thesis proposed a new nonparametric density estimation method based on the nonparametric maximum likelihood. The proposed density estimator reduces the bias of the standard kernel density estimator and its performance was reasonable for both Gaussian mixture densities and non-Gaussian densities in the univariate case. Although its properties were similar in the multivariate cases, its advantage over the other improved density estimators tends to fade slightly away as dimension grows. As note in Remark 3.1, the likelihood-tuning starting from a normal density instead of the uniform is expected to enhance the performance for Gaussian mixture densities. In addition, since the estimator twice tuned from the normal does not seem to inflate the variance, its asymptotic bias might be lower than the estimator tuned from the uniform.

The algorithmic convergence of the continuous EM algorithm to the NPMLE in Chapter 3 extends the previous proof by Wu (1983) to a functional parameter space of mixing densities. In addition, our proof is stronger than Wu's results in the sense that it proves convergence to the global maximum. In the proof that the EM algorithm converges, the regularity condition that each fixed point has a different likelihood could be relaxed under the condition that the EM algorithm is a contraction operator.

Bandwidth choice is a difficult problem in nonparametric density estimation, particularly in higher dimensions where data are comparably sparse. In Chapter 4, we

investigated the spectral degrees of freedom as a way to measure the amount of smoothing approximately. Theoretical work and simulation results showed that the bandwidth needs to increase proportionally to the square root of dimension for spectral degrees of freedom and coalescence. Noting that the spectral degrees of freedom is a measure related only to the integrated variance of the kernel density estimator and MISE gives equal weight to the integrated bias and the integrated variance, as one area of future work, we could explore a different weight between these two factors. We also think that there are other measures, such as the Kullback-Leibler divergence, that might work better in higher dimensions than squared error loss. It is invariant under scaling change and so should be more homogeneous in interpretation our dimension.

The penalized likelihood-tuning method in Chapter 5 was proposed as a more flexible procedure than the unpenalized likelihood-tuning in Chapter 2 and 3. We obtained the density estimator that includes a local shape and scale adaptation term in the t-kernel function. Instead of penalizing the likelihood function, we could consider in the future the alternate formulation that is described in Section 5.4. We could obtain a posterior of the kernel function with a prior density of the variance parameter, then apply the likelihood-tuning procedure on the location parameter only. In this setting, we would investigate how the local scale and shape is adapted in the likelihood-tuning procedure compared to the penalized likelihood-tuning.

Appendix A

A.1 Outline proof of Theorem 1

The likelihood-tuned density estimator in (3.13) can be expressed as

$$\hat{f}_{\mathrm{EM2}}(x) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{f}_{(1)}(X_j)^{-1} \int K_h(X_i, y) K_h(X_j, y) K_h(y, x) dy$$

if we do not expand terms in the second tuning step. Then the expectation of $\hat{f}_{\mathrm{EM2}}(x)$ is

$$\int \int \frac{\int K_h(x_i, y) K_h(x_j, y) K_h(y, x) dy}{\hat{f}_{(1)}(x_j)} dF(x_i) dF(x_j), \tag{A.1}$$

where F is a distribution function for the true density f. The numerator of the integrand in (A.1) is regarded as a conditional density of (x_i, x_j) given y. This conditional density can be factored into two terms:

$$K(x_i | x_j, x) \;=\; N\left(\frac{x + x_j}{2}, \frac{3}{2}h^2\right) \quad \text{and} \quad K(x_j | x) \;=\; N(x, 2h^2).$$

Using these terms, we rewrite (A.1) as

$$\int \left\{ \int K(x_i | x_j, x) f(x_i) dx_i \right\} \frac{K(x_j | x) f(x_j) dx_j}{\int K_{h\sqrt{2}}(x_k, x_j) f(x_k) dx_k}, \tag{A.2}$$

where $K_h(x, y) = h^{-1}\varphi(h^{-1}(x - y))$ with φ denoting the standard normal density. Define $B_1(x)$, $B_2(x)$ and $B_3(x)$ as two numerators and a denominator as follows and change

variables x_i, x_j, and x_k into z_1, z_2, and z_3 that follow the standard normal density $\varphi(\cdot)$:

$$\begin{aligned} B_1(x) &= \int K(x_i|x_j,x)f(x_i)dx_i = \int \varphi(z_1)\, f\left(\frac{x+x_j}{2} + z_1 h\sqrt{\frac{3}{2}}\right) dz_1; \\ B_2(x) &= K(x_j|x)f(x_j)dx_j = \varphi(z_2)f(x+z_2h\sqrt{2})dz_2; \\ B_3(x) &= \int K_{h\sqrt{2}}(x_k,x_j)f(x_k)dx_k = \int \varphi(z_3)f(x_j+z_3h\sqrt{2})dz_3. \end{aligned}$$

After we can expand $f(\cdot)$ in B_1, B_2 and B_3 around $h=0$, $B_1(x)$ and $B_3(x)$ are integrated over z_1 and z_3, respectively. Now there remains only one random variable, z_2, in three terms. By multiplying $B_2(x)/B_3(x)$ and $B_1(x)$ and integrating over z_2, the density estimator (A.2) is written as

$$\int \varphi(z_2)\{f(x) + a_1(z_2)h + a_2(z_2)h^2 + a_3(z_2)h^3 + a_4(z_2)h^4 + O(h^5)\}dz_3$$

for appropriate coefficients $a_1(z_3), \cdots, a_4(z_3)$. With some calculation, we find that

$$\int a_1(z_2)dz_2 = \int a_2(z_2)dz_2 = \int a_3(z_2)dz_2 = 0$$

and

$$\int a_4(z_2)dz_2 = -f^{(4)}(x) + \frac{1}{f(x)}\left(f'(x)f^{(3)}(x) + f''^2(x)\right) - \frac{1}{f^2(x)}f''(x)f'^2(x).$$

Thus we attain

$$\begin{aligned} E\left[\hat{f}_{\mathrm{EM2}}(x)\right] &= f(x)\left(-\frac{f^{(4)}(x)}{f(x)} + \frac{f^{(3)}(x)f'(x) + f''(x)^2}{f(x)^2} - \frac{f''(x)f'(x)^2}{f(x)^3}\right)h^4 \\ &\quad + f(x) + o(h^4). \end{aligned}$$

For the asymptotic variance, rewrite (A.2) as a functional on the distribution F as

$$T(F) = \int\int K_{h\sqrt{3/2}}\left(\frac{x+x_j}{2}, x_i\right) dF(x_i) \frac{K_{h\sqrt{2}}(x_j,x)dF(x_j)}{\int K_{h\sqrt{2}}(x_k,x_j)dF(x_k)}.$$

We will find the first von Mises derivative $T'(y)$, and then use the result that

$$T(\hat{F}) - T(F) \approx \int T'(y) d(\hat{F} - F)$$

so that the asymptotic variance for $T(\hat{F})$ is

$$\text{asyvar}(T(\hat{F})) = \text{Var}_F(T'(y))/n.$$

The first von Mises derivative $T'(y)$ is a sum of three terms:

$$\begin{aligned}
&\int \frac{\int K_{h\sqrt{3/2}}(\frac{x+x_j}{2}, x_i) dF(x_i)}{\int K_{h\sqrt{2}}(x_k, x_j) dF(x_k)} K_{h\sqrt{2}}(x_j, x) d\Delta(x_j), \\
&\int \frac{\int K_{h\sqrt{3/2}}(\frac{x+x_j}{2}, x_i) d\Delta(x_i)}{\int K_{h\sqrt{2}}(x_k, x_j) dF(x_k)} K_{h\sqrt{2}}(x_j, x) dF(x_j), \text{ and} \\
&-\int \frac{\int K_{h\sqrt{3/2}}(\frac{x+x_j}{2}, x_i) dF(x_i) \int K_{h\sqrt{2}}(x_k, x_j) d\Delta(x_k)}{\left(\int K_{h\sqrt{2}}(x_k, x_j) dF(x_k)\right)^2} K_{h\sqrt{2}}(x_j, x) dF(x_j),
\end{aligned}$$

where the measure $d\Delta(x) = d\delta_y(x) - dF(x)$ and δ_y is the distribution degenerate at y. We can rewrite by letting

$$\begin{aligned}
C_1(y) &= \frac{\int K_{h\sqrt{3/2}}(\frac{x+y}{2}, x_i) dF(x_i)}{\int K_{h\sqrt{2}}(x_k, y) dF(x_k)} K_{h\sqrt{2}}(y, x), \\
C_2(y) &= \int \frac{K_{h\sqrt{3/2}}(\frac{x+x_j}{2}, y)}{\int K_{h\sqrt{2}}(x_k, x_j) dF(x_k)} K_{h\sqrt{2}}(x_j, x) dF(x_j), \\
C_3(y) &= -\int \frac{\int K_{h\sqrt{3/2}}(\frac{x+x_j}{2}, x_i) dF(x_i) \cdot K_{h\sqrt{2}}(y, x_j)}{\left(\int K_{h\sqrt{2}}(x_k, x_j) dF(x_k)\right)^2} K_{h\sqrt{2}}(x_j, x) dF(x_j),
\end{aligned}$$

and noting that $T'(y) = C_1(y) + C_2(y) + C_3(y) - E(C_1(y) + C_2(y) + C_3(y))$.

Further, $E(C_1(y) + C_2(y) + C_3(y))$ is just the asymptotic mean calculated above. Thus we seek $E(C_1(y) + C_2(y) + C_3(y))^2$. By expanding and integrating as in calculating

the asymptotic mean, we have the following limiting results:

$$\begin{aligned}
h\int C_1(y)^2 f(y)dy &\to f(x)\frac{1}{\sqrt{2}}\int \varphi(z)^2 dz,\\
h\int C_2(y)^2 f(y)dy &\to f(x)\frac{1}{\sqrt{2}}\int \varphi^2(z)dz,\\
h\int C_3(y)^2 f(y)dy &\to f(x)\frac{1}{2}\int \varphi^2(z)dz,\\
h\int C_1(y)C_2(y) f(y)dy &\to f(x)\frac{1}{\sqrt{2}}\int \varphi^2(z)dz,\\
h\int C_1(y)C_3(y) f(y)dy &\to -f(x)\frac{1}{\sqrt{3}}\int \varphi^2(z)dz, \text{ and}\\
h\int C_2(y)C_3(y) f(y)dy &\to -f(x)\frac{1}{\sqrt{3}}\int \varphi^2(z)dz.
\end{aligned}$$

By adding three square terms and three cross terms twice, we have the asymptotic variance,

$$\mathrm{Var}\left[\hat{f}_{\mathrm{EM2}}(x)\right] = (nh)^{-1} f(x)\left(\frac{1}{2}+\frac{4}{\sqrt{2}}-\frac{4}{\sqrt{3}}\right)\int \varphi^2(z)dz + o\left((nh)^{-1}\right).$$

Bibliography

Abramson, I. "On bandwidth variation in kernel estimates-a square root law." *The Annals of Statistics*, 10(4):1217–1223 (1982).

Azzalini, A. and Torelli, N. "Clustering via nonparametric density estimation." *Statistics and Computing*, 17(1):71–80 (2007).

Bhattacharya, R. and Waymire, E. *A basic course in probability theory*. Springer Verlag (2007).

Bowman, A. "An alternative method of cross-validation for the smoothing of density estimates." *Biometrika*, 71(2):353 (1984).

Bowman, A. and Foster, P. "Adaptive smoothing and density-based tests of multivariate normality." *Journal of the American Statistical Association*, 88(422):529–537 (1993).

Breiman, L., Meisel, W., and Purcell, E. "Variable kernel estimates of multivariate densities." *Technometrics*, 19(2):135–144 (1977).

Chen, S. and Lindsay, B. "Building mixture trees from binary sequence data." *Biometrika*, 93(4):843 (2006).

Chung, Y. and Lindsay, B. "A Likelihood-tuned Density Estimator via a Nonparametric Mixture Model." *Festschrift in honor of Tom Hettmansperger, World Scientific* (2010).

Ciuperca, G., Ridolfi, A., and Idier, J. "Penalized Maximum Likelihood Estimator for Normal Mixtures." *Scandinavian Journal of Statistics*, 30(1):45–59 (2003).

DiMarzio, M. and Taylor, C. "Boosting kernel density estimates: A bias reduction technique?" *Biometrika* (2004).

Geisser, S. "The predictive sample reuse method with applications." *Journal of the American Statistical Association*, 70(350):320–328 (1975).

Good, I. and Gaskins, R. "Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data." *Journal of the American Statistical Association*, 75(369):42–56 (1980).

Goutis, C. "Nonparametric Estimation of a Mixing Density Via the Kernel Method." *Journal of the American Statistical Association*, 92(440) (1997).

Hall, P. and Marron, J. "Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation." *Probability Theory and Related Fields*, 74(4):567–581 (1987).

—. "Variable window width kernel estimates of probability densities." *Probability Theory and Related Fields*, 80(1):37–49 (1988).

Hathaway, R. "A constrained formulation of maximum-likelihood estimation for normal mixture distributions." *The Annals of Statistics*, 13(2):795–800 (1985).

Ingrassia, S. and Rocci, R. "Constrained monotone EM algorithms for finite mixture of multivariate Gaussians." *Computational Statistics and Data Analysis*, 51(11):5339–5351 (2007).

Jones, M. "On correcting for variance inflation in kernel density estimation." *Computational Statistics and Data Analysis*, 11(1):3–15 (1991).

Jones, M., Linton, O., and Nielsen, J. "A simple bias reduction method for density estimation." *Biometrika*, 82(2):327–338 (1995).

Jones, M. and Signorini, D. "A Comparison of Higher-Order Bias Kernel Density Estimators." *Journal of the American Statistical Association*, 92(439) (1997).

Kiefer, J. and Wolfowitz, J. "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters." *The Annals of Mathematical Statistics*, 27(4):887–906 (1956).

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

Laird, N. and Louis, T. "Smoothing the non-parametric estimate of a prior distribution by roughening: A computational study." *Computational Statistics and Data Analysis*, 12:27–37 (1991).

Li, J., Ray, S., and Lindsay, B. "A Nonparametric Statistical Approach to Clustering via Mode Identification." *The Journal of Machine Learning Research*, 8:1687–1723 (2007).

Lindsay, B. *Mixture Models: Theory, Geometry, and Applications*. Ims (1995).

Lindsay, B., Markatou, M., Ray, S., Yang, K., and Chen, S. "Quadratic distances on probabilities: A unified foundation." *The Annals of Statistics*, 36:983–1006 (2008).

Loader, C. "Bandwidth selection: classical or plug-in?" *The Annals of Statistics*, 27(2):415–438 (1999).

Marron, J. and Wand, M. "Exact mean integrated squared error." *The Annals of Statistics*, 20(2):712–736 (1992).

Ray, S. and Lindsay, B. G. "The topography of multivariate normal mixtures." *The Annals of Statistics*, 33(5):2042–2065 (2005).

Ridolfi, A. and Idier, J. "Penalized maximum likelihood estimation for univariate normal mixture distributions." In *AIP Conference Proceedings*, 229–240. Citeseer (2001).

Rudemo, M. "Empirical choice of histograms and kernel density estimators." *Scandinavian Journal of Statistics*, 65–78 (1982).

Samiuddin, M. and El-Sayyad, G. "On nonparametric kernel density estimates." *Biometrika*, 77(4):865 (1990).

Sarro, L., Debosscher, J., Aerts, C., and López, M. "Comparative clustering analysis of variable stars in the Hipparcos, OGLE Large Magellanic Cloud, and CoRoT exoplanet databases." *Astronomy and Astrophysics*, 506(1):535–568 (2009).

Scott, D. "Feasibility of multivariate density estimates." *Biometrika*, 78(1):197 (1991).

—. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley-Interscience (1992).

Scott, D. and Factor, L. "Monte Carlo study of three data-based nonparametric probability density estimators." *Journal of the American Statistical Association*, 76(373):9–15 (1981).

Scott, D. and Terrell, G. "Biased and unbiased cross-validation in density estimation." *Journal of the American Statistical Association*, 82(400):1131–1146 (1987).

Sheather, S. and Jones, M. "A reliable data-based bandwidth selection method for kernel density estimation." *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 683–690 (1991).

Silverman, B. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC (1986).

Silverman, B., Jones, M., Wilson, J., and Nychka, D. "A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography." *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(2):271–324 (1990).

Tran, T., Wehrens, R., and Buydens, L. "KNN-kernel density-based clustering for high-dimensional multivariate data." *Computational Statistics and Data Analysis*, 51(2):513–525 (2006).

Vardi, Y. and Lee, D. "From image deblurring to optimal investments: Maximum likelihood solutions for positive linear inverse problems." *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(3):569–612 (1993).

Vardi, Y., Shepp, L., and Kaufman, L. "A statistical model for positron emission tomography (with discussion)." *Journal of the American Statistical Association*, 80:8–37 (1985).

Wand, M. and Jones, M. *Kernel Smoothing*. Chapman & Hall/CRC (1995).

Warton, D. "Penalized Normal Likelihood and Ridge Regularization of Correlation and Covariance Matrices." *Journal of the American Statistical Association*, 103(481):340–349 (2008).

Wu, C. "On the convergence properties of the EM algorithm." *The Annals of Statistics*, 11(1):95–103 (1983).

Yang, K. "Using the Poisson kernel in model building and selection." Ph.D. thesis, The Pennsylvania State University, University Park, PA, USA (2004).

Yao, W. and Lindsay, B. "Bayesian Mixture Labeling by Highest Posterior Density." *Journal of the American Statistical Association*, 104(486):758–767 (2009).

Vita

Yeojin Chung

Yeojin Chung was born in Seoul, South Korea, 1980. She received her Bachelor degree in Economics and Applied Statistics from Yonsei University in 2003 and her Master degree in Applied Statistics from Yonsei University in 2005. She enrolled in the Ph. D. program in Statistics at the Pennsylvania State University in 2005.