Supplemental materials for this article are available through the JCGS web page at http://www.amstat.org/publications.

A Functional EM Algorithm for Mixing Density Estimation via Nonparametric Penalized Likelihood Maximization

Lei LIU, Michael LEVINE, and Yu ZHU

When the true mixing density is known to be continuous, the maximum likelihood estimate of the mixing density does not provide a satisfying answer due to its degeneracy. Estimation of mixing densities is a well-known ill-posed indirect problem. In this article, we propose to estimate the mixing density by maximizing a penalized likelihood and call the resulting estimate the nonparametric maximum penalized likelihood estimate (NPMPLE). Using theory and methods from the calculus of variations and differential equations, a new functional EM algorithm is derived for computing the NPMPLE of the mixing density. In the algorithm, maximizers in M-steps are found by solving an ordinary differential equation with boundary conditions numerically. Simulation studies show the algorithm outperforms other existing methods such as the popular EMS algorithm. Some theoretical properties of the NPMPLE and the algorithm are also discussed. Computer code used in this article is available online.

Key Words: Mixture model; Nonparametric maximum penalized likelihood estimate; Ordinary differential equation with boundary conditions.

1. INTRODUCTION

Suppose $y_1, y_2, \ldots, y_n$ are independent and identically distributed with a mixture density

$$h(y|G) = \int f(y|x)\, dG(x), \tag{1.1}$$

where $f(y|x)$ is a known component density function indexed by x and $G(x)$ is a mixing distribution. Laird (1978) showed that, under some mild conditions on f, the nonparametric maximum likelihood estimate (NPMLE) of G, denoted by $\hat{G}$, is a step function with at most n jumps. Laird (1978) also proposed an EM algorithm to find $\hat{G}$. Lindsay (1983a, 1983b) proved the existence and uniqueness of $\hat{G}$ and obtained other important properties of $\hat{G}$. When the true distribution G is known to have a continuous density $g(x)$, which is

Lei Liu is Statistical Analyst, HealthCare Interactive, Inc., Minneapolis, MN 55435 (E-mail: *liulnk@gmail.com*). Michael Levine is Assistant Professor (E-mail: *mlevins@stat.purdue.edu*) and Yu Zhu is Associate Professor (E-mail: *yuzhu@stat.purdue.edu*), Department of Statistics, Purdue University, West Lafayette, IN 47907-2066.

Journal of Computational and Graphical Statistics, Volume 18, Number 2, Pages 481–504
DOI: 10.1198/jcgs.2009.07111

referred to as a mixing density, and g is the target of statistical inference, the NPMLE of G becomes improper because of its degeneracy. In this article, we propose a new nonparametric method that uses penalized maximum likelihood to estimate g. When the density g exists, the model (1.1) can be rewritten as

$$h(y|g) = \int_{\mathcal{X}} f(y|x)g(x)\,dx, \tag{1.2}$$

where $\mathcal{X}$ is the support of $g(x)$. The support $\mathcal{X}$ is assumed to be a known compact interval throughout this article. In what follows, we first give a brief review of existing methods for estimating mixing densities, then discuss the ideas behind the new algorithm we develop in this article. The layout of the article is given at the end of this section.

1.1 EXISTING METHODS FOR ESTIMATING MIXING DENSITIES

The existing methods for estimating mixing densities in the literature can be roughly divided into three categories: EM-based algorithms, kernel methods, and methods based on orthogonal series expansion.

As mentioned earlier, an EM algorithm was originally proposed by Laird to compute $\hat{G}$, the NPMLE of G. Observing that the EM algorithm produces smooth estimates before it converges to $\hat{G}$, Vardi, Shepp, and Kaufman (1985) recommended to start the EM algorithm from a uniform distribution and let it run for a limited number of iterations. The resulting estimate is then used as a continuous estimate of g, whose likelihood can be fairly close to the maximum when the number of iterations is properly specified. The smoothing-by-roughening method proposed by Laird and Louis (1991) uses a similar strategy of stopping the EM algorithm early, with the suggested number of iterations proportional to $\log n$ where n is the sample size. A common drawback of the above two methods is that both lack a formal stopping rule to terminate the EM algorithm. Silverman et al. (1990) proposed the Smoothed EM (EMS) algorithm, which adds a smoothing step to each expectation-maximization iteration. Empirically, this algorithm was found to converge quickly to an estimate close to the true mixing density. There are two drawbacks of the EMS algorithm. First, it does not preserve the monotonicity property of the original EM algorithm due to the added smoothing steps. Second, the estimate obtained by the EMS algorithm is hard to interpret because there does not exist an apparent objective function it optimizes. To overcome the second drawback, Eggermont and LaRiccia (1995, 1997) incorporated a smoothing operator into the likelihood function and proposed the Nonlinearly Smoothed EM (NEMS) algorithm. They showed that the NEMS algorithm performs similarly to the EMS algorithm; in addition, the estimate given by the NEMS is the maximizer of the smoothed likelihood function. Other EM-based algorithms include an EM algorithm with stepwise knot deletion and model selection (Koo and Park 1996), the One Step Late (OSL) procedure (Green 1990), and the Doubly Smoothed EM (EMDS) algorithm (Szkutnik 2003). The last algorithm was specifically designed and optimized to deal with grouped data.

When the component density function $f(y|x)$ can be written as $\phi(y-x)$ where x is a location parameter, estimating the mixing density function $g(x)$ is referred to as deconvolution in the literature. Using the Fourier transform, a kernel-type estimate can be derived for

g; see Stefanski and Carroll (1990), Zhang (1990), and Fan (1991). Fan (1991) showed that this kernel estimate can achieve the optimal convergence rate in a certain sense. Unfortunately, this approach is limited to the deconvolution problem only. Goutis (1997) proposed a general kernel-type procedure for estimating g without assuming that x is a location parameter of $f(y|x)$. The resulting estimate is conceptually similar to a kernel estimate having the form of $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x-x_i}{h})$ where $K(\cdot)$ is a kernel function and $h > 0$ is the bandwidth. The method of Mixture-of-Gaussians proposed by Magder and Zeger (1996) can essentially be classified as a kernel-type method; similar ideas were discussed in Lindsay (1995) as well.

The third group of existing methods includes those based on orthogonal series expansion. Let K be an integral operator: $g \to Kg = \int f(y|x)g(x)\,dx$. Johnstone and Silverman (1990) and Jones and Silverman (1989) proposed to expand and estimate g using the orthogonal basis in the singular value decomposition (SVD) of the operator K. Smoothing is enforced through cutting off the infinite expansion of $g(x)$ or, more generally, through tapering it using a sequence of weights w_ν satisfying $w_\nu \to 0$ as $\nu \to \infty$; see Silverman (1986) and Izenman (1991) for more details. Koo and Chung (1998) proposed to approximate and estimate $\log g(x)$ using a finite linear combination of the eigenfunctions of K; the corresponding estimate is called the Maximum Indirect Likelihood Estimator (MILE). For the deconvolution problem with $f(y|x) = \phi(y - x)$, estimators based on wavelet expansion and coefficients' thresholding have been proposed, and their convergence behavior has been studied; see Pensky and Vidakovic (1999) and Fan and Koo (2002).

1.2 MAXIMUM PENALIZED LIKELIHOOD METHOD

Another well-known way to generate continuous density estimates is to penalize the roughness of a density function. One of the most popular is the maximum penalized likelihood method. Consider direct density estimation first whereby the density is estimated based on observations directly sampled from it. The penalized log-likelihood functional for an arbitrary density f, denoted by $l_p(f)$, has the form

$$l_p(f) = l(f) - \lambda J(f), \tag{1.3}$$

where $l(f)$ is the usual log-likelihood function, $J(f)$ is a roughness penalty term, and λ is a smoothing parameter. The maximum penalized likelihood estimate (MPLE) is defined as the maximizer of $l_p(f)$ over a collection of density functions.

The penalized likelihood method for direct density estimation was pioneered by Good and Gaskins (1971). De Montricher, Tapia, and Thompson (1975) and Klonias (1982) proved the existence and consistency of the MPLE defined by Good and Gaskins (1971). For a comprehensive introduction to this method, see Tapia and Thompson (1978); for a more recent account, see Eggermont and LaRiccia (2001). To better accommodate positivity and unity constraints of a density function, Leonard (1978) and Silverman (1982) proposed to estimate the log-density function $\eta = \log f$ using the penalized likelihood method. Gu and Qiu (1993) and Gu (1993) further studied this problem using smoothing splines, and developed an algorithm that can be used to estimate multivariate density functions.

The application of MPLE to estimate mixing densities is a natural idea. Silverman et al. (1990) and Green (1990) discussed the possibility of using this approach for indirect density estimation or, equivalently, mixing density estimation. Both considered this approach reasonable, but the computational difficulties in M-steps kept them from implementing the MPLE for mixing density estimation directly. Instead, Silverman et al. (1990) proposed the EMS algorithm by adding a smoothing step after each EM iteration, and Green (1990) proposed the One Step Late (OSL) procedure, a pseudo-EM algorithm, to circumvent the computational difficulties. Both methods were discussed in the previous section.

In this article, we aim at fully implementing the maximum penalized likelihood method for mixing densities estimation. A functional EM algorithm is proposed to compute the maximum penalized likelihood estimate of a mixing density over a function space. During each M-step of this EM algorithm, maximization is conducted over the same function space and the maximizer is characterized by a nonlinear ordinary differential equation with boundary conditions, which is solved by a numeric procedure called the collocation method.

1.3 Organization of the Article

The rest of the article is organized as follows. Section 2 defines the nonparametric maximum penalized likelihood estimate (NPMPLE). We derive the new functional EM algorithm in Section 3. Some theoretical results supporting the definition of the new estimator and the new algorithm are included in these two sections as well. Section 4 discusses the numeric solution to the nonlinear ordinary differential equations generated in M-steps of the algorithm. Section 5 focuses on the selection of smoothing parameter λ. Section 6 compares the new algorithm with the EMS algorithm through a simulation study. Some concluding remarks are given in the last section. Due to space limitation, the proofs of the propositions, theorems, and corollaries are included in a separate supplementary document that is available on the *JCGS* website. An R code of the proposed algorithm can also be found on the same website.

2. NONPARAMETRIC MAXIMUM PENALIZED LIKELIHOOD ESTIMATE

Let g_0 be the true mixing density in model (1.2). Assume that the support of g_0 is a finite interval $\mathcal{X} = [a, b]$, and g_0 is bounded above and below away from 0. In other words, there exist positive constants M_0 and M_1 such that $M_0 \leq g_0(x) \leq M_1$ for all $x \in [a, b]$. These assumptions are collectively labeled as Assumption (A0), which is assumed to hold throughout this article. The assumption that g_0 has a compact support is not uncommon in density estimation. The results developed in this article can be extended to estimating mixing densities with unbounded support. Because the extension requires different techniques, it is not pursued in this article. Nevertheless, in practice, when facing mixing densities with unbounded support, the methods developed in this article can still be applied to generate reasonably good estimates by using intervals wider than the data range. The second part of Assumption (A0) is for convenience, due to the fact that the log mixing density instead of the mixing density itself is directly estimated. For more discussions of the assumptions, see Silverman (1982).

A density g satisfying Assumption (A0) can be represented as

$$g(x) = \frac{e^{\eta(x)}}{\int_a^b e^{\eta(t)}\,dt}, \quad \text{where } \eta(x) = \log g(x) + \text{const}, \tag{2.1}$$

and the mixture density of $h(y|G)$ becomes

$$h(y|\eta) = \frac{\int_a^b f(y|x)e^{\eta(x)}\,dx}{\int_a^b e^{\eta(x)}\,dx}. \tag{2.2}$$

Given a random sample $y_1, y_2, \ldots, y_n$ from the above density (2.2), the log-likelihood functional of η is

$$l(\eta) = \frac{1}{n}\sum_{i=1}^{n} \log \int_a^b f(y_i|x)e^{\eta(x)}\,dx - \log \int_a^b e^{\eta(x)}\,dx. \tag{2.3}$$

As discussed in the Introduction, we want to penalize the roughness of η using a penalty term. In this article, we choose the penalty

$$J(\eta) = \int_a^b [\eta''(x)]^2\,dx, \tag{2.4}$$

which was originally proposed by Leonard (1978) for (direct) density estimation. Combining $l(\eta)$ and $J(\eta)$ gives a penalized likelihood functional

$$\begin{aligned} l_p(\eta) &= l(\eta) - \lambda J(\eta) \\ &= \frac{1}{n}\sum_{i=1}^{n} \log \int_a^b f(y_i|x)e^{\eta(x)}\,dx - \log \int_a^b e^{\eta(x)}\,dx - \lambda \int_a^b [\eta''(x)]^2\,dx, \end{aligned} \tag{2.5}$$

where $\lambda > 0$ is a smoothing parameter.

To obtain a proper estimate of η by maximizing $l_p(\eta)$, we need to specify a proper function space, denoted by $\mathcal{H}$, as the "parameter" space. Given the penalty that we use, a natural choice is to assume that $\eta(x) \in \mathcal{H} = W^{2,2}[a,b]$ where $W^{2,2}[a,b]$ is the second-order Sobolev space based on L_2-norm; see, for example, Adams (1975) for formal definitions. It is known that for any $\eta \in \mathcal{H}$, both the function itself and its first derivative are absolutely continuous (Wahba 1990). Hence, the functions in $\mathcal{H}$ are smooth enough for our purpose. The nonparametric maximum penalized likelihood estimates (NPMPLEs) of η_0 and g_0 are defined, respectively, as

$$\hat{\eta} = \arg\max_{\eta\in\mathcal{H}} l_p(\eta) \qquad \text{and} \qquad \hat{g} = \frac{e^{\hat{\eta}(x)}}{\int_a^b e^{\hat{\eta}(t)}\,dt}. \tag{2.6}$$

Note that if $\hat{\eta}$ is a maximizer of $l_p(\eta)$, then clearly $\hat{\eta} + C$ is also a maximizer, where C is an arbitrary constant. Both $\hat{\eta}$ and $\hat{\eta} + C$, however, give the same $\hat{g}$. Therefore the difference between $\hat{\eta}$ and $\hat{\eta} + C$ will not cause confusion for our purpose and we consider $\hat{\eta}$ well-defined up to a constant shift.

Let $\mathcal{N}_J = \{\eta \in \mathcal{H} : J(\eta) = 0\} = \{cx + d : c, d \in R\}$, which is the null space of the penalty functional $J(\eta) = \int_a^b [\eta''(x)]^2\,dx$. Let $\mathcal{Y} = \bigcup_{x\in\mathcal{X}}\{y : f(y|x) > 0\}$. In addition to Assumption (A0) about g_0 stated at the beginning of this section, one more assumption needs to be imposed to ensure the existence of the NPMPLE $\hat{\eta}(x)$, which is: (A1) For any given $y \in \mathcal{Y}$, $f(y|x)$ is a continuous function of x in $[a, b]$. Based on Assumption (A1), it can be shown that there exists a positive number $M > 0$ such that $0 < \int_a^b f(y|x)\,dx < M$ for any given $y \in \mathcal{Y}$. Assumption (A1) is a regularity condition imposed on $f(y|x)$ as a function of x for any given y. Under Assumptions (A0) and (A1), the true mixture density $h(y|g_0)$ has an upper bound. Popularly used component densities usually satisfy Assumption (A1). Together with Assumption (A0), Assumption (A1) is assumed to hold throughout this article; and they are not restated in the theorems and propositions below. The following two results establish the existence of $\hat{\eta}$.

Theorem 1. *If there exists an $\eta^*(x) = c^*x + d^* \in \mathcal{N}_J$ such that*

$$l(\eta^*) > \max\left\{\frac{1}{n}\sum_{i=1}^{n} \log f(y_i|a), \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|b)\right\}, \tag{2.7}$$

then there exists $\hat{\eta} \in \mathcal{H}$ such that $l_p(\hat{\eta}) \geq l_p(\eta)$ for all $\eta \in \mathcal{H}$.

Corollary 1. *If the uniform distribution $U(a, b)$ has a higher likelihood than the point mass distribution on a or on b, that is,*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} &\log\left(\frac{1}{b-a}\int_a^b f(y_i|x)\,dx\right) \\ &> \max\left\{\frac{1}{n}\sum_{i=1}^{n} \log f(y_i|a), \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|b)\right\},\end{aligned} \tag{2.8}$$

then there exists $\hat{\eta} \in \mathcal{H}$ such that $l_p(\hat{\eta}) \geq l_p(\eta)$ for all $\eta \in \mathcal{H}$.

Theorem 1 indicates that, if there exists a density $g^*(x) = \exp\{c^*x + d^*\}$ that gives a better explanation to the sample $\{y_i\}$ in terms of likelihood than the one-point mass distributions at a and b, then the maximizer of $l_p(\eta)$ over $\mathcal{H}$ exists. Intuitively, this condition should be satisfied except in some extremely rare situations where the sample is concentrated around a or b as if it was drawn from the density $f(y|a)$ or $f(y|b)$. Corollary 1 gives a convenient sufficient condition for the existence of $\hat{\eta}$, which is easy to verify and should always be checked first.

Finding the maximizer of a functional over a function space is a typical problem in the calculus of variations. Usual techniques used to deal with finite-dimensional parameters, such as those used to solve a system of likelihood score equations, are not directly applicable to finding the maximizer $\hat{\eta}$ of $l_p(\eta)$. In this article, we resort to concepts and techniques from the calculus of variations and differential equations instead. In the next section, we first present some properties of $\hat{\eta}$, then propose and develop a functional EM algorithm for computing the NPMPLE $\hat{\eta}$.

3. FUNCTIONAL EM ALGORITHM FOR COMPUTING $\hat{\eta}$

Because the likelihood part of $l_p(\eta)$ involves logarithms of mixture densities $h(y_i|\eta)$ that are conditional on η inside an integral, its direct maximization is usually difficult even in the situation where η depends on a finite-dimensional parameter and no penalty exists. One popular way to circumvent this difficulty is to use the EM algorithm. We adopt this approach to develop a Functional EM algorithm (FEM) to compute the NPMPLE $\hat{\eta}$. This algorithm is effectively nonparametric because it attempts to find optimal $\eta \in \mathcal{H}$.

3.1 DERIVATION OF FEM AND THE E-STEP

It is well known that the random sample $\{y_i\}_{1\le i\le n}$ from the mixture density $h(y|\eta_0)$ can be generated using the following two-step procedure: first a random sample denoted $\{x_i\}_{1\le i\le n}$ is drawn from the mixing density $g_0(x)$, then y_i is randomly drawn from the component density $f(y|x_i)$. Because x_i's are not observable, they are referred to as missing or latent values. $\{(y_i, x_i)\}_{1\le i\le n}$ forms a random sample from the joint density $f(y|x)g_0(x)$ and is referred to as the complete data. Given this complete data, a complete penalized log-likelihood functional of η can be defined as

$$l_{cp}(\eta) = \frac{1}{n}\sum_{i=1}^{n}\{\log f(y_i|x_i) + \eta(x_i)\} - \log\int_a^b e^{\eta(x)}\,dx - \lambda\int_a^b [\eta''(x)]^2\,dx. \tag{3.1}$$

If η were a function depending on a finite-dimensional parameter, with or without the penalty term, the classical EM algorithm (Dempster, Laird, and Rubin 1977) would have started with an expectation step (E-step) involving the complete likelihood $l_{cp}(\eta)$, then proceeded on to the maximization step (M-step) to calculate $\hat{\eta}$, and then repeated the two steps iteratively, beginning with some initial value of η. Here we attempt to develop a similar iterative process in the functional space $\mathcal{H}$. The details are described below.

In the E-step, we compute the expectation of $l_{cp}(\eta)$ given the current estimate of η and the data. Let $\vec{y} = (y_1, y_2, \ldots, y_n)$ be the (observable) data, η_{cur} denote the current estimate of η, and $Q(\eta|\eta_{\text{cur}}) = E[l_{cp}(\eta)|\vec{y}, \eta_{\text{cur}}]$. Because y_i's are independent, the expectation of the complete likelihood can be simplified to

$$\begin{aligned} Q(\eta|\eta_{\text{cur}}) &= E[l_{cp}(\eta)|\vec{y}, \eta_{\text{cur}}] \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_a^b \{\log f(y_i|x_i) + \eta(x_i)\}\varphi(x_i|y_i, \eta_{\text{cur}})\,dx_i \\ &\quad - \log\int_a^b e^{\eta(x)}\,dx - \lambda\int_a^b [\eta''(x)]^2\,dx, \end{aligned} \tag{3.2}$$

where

$$\varphi(x|y_i, \eta_{\text{cur}}) = \frac{f(y_i|x)e^{\eta_{\text{cur}}(x)}}{\int_a^b f(y_i|t)e^{\eta_{\text{cur}}(t)}\,dt} \tag{3.3}$$

is the conditional density of x_i given data y_i and the current estimate η_{cur} of η. Effectively, $\varphi(x|y_i, \eta_{\text{cur}})$ can be seen as a posterior density and its computation process can be viewed as a Bayesian updating scheme.

In the M-step, we compute the maximizer of $Q(\eta|\eta_{\text{cur}})$, which is denoted by η_{new} and used as the current estimate for the next iteration. The E-step and M-step are thus iterated until the estimate $\hat{\eta}$ converges. Although the algorithm defined above is not a classical EM algorithm ($l_{cp}(\eta)$ is a penalized likelihood functional over the function space $\mathcal{H}$), it still retains the monotonicity property of a classical EM algorithm as stated in the next proposition.

Proposition 1. *After each iteration of the E-step and M-step above,* $l_p(\eta_{\text{new}}) \geq l_p(\eta_{\text{cur}})$.

Proposition 1 implies that the FEM algorithm converges to a maximum of $l_p(\eta)$. However, this maximum may not be global, because $l_p(\eta)$ is not necessarily concave in η and may have many local maxima. Although the FEM algorithm may be trapped in a local maximum, our simulation study shows that the problem is not severe. The E-steps of FEM are straightforward; the M-steps involve maximizing a new functional of η [i.e., $Q(\eta|\eta_{\text{cur}})$] and thus are not trivial. Though $Q(\eta|\eta_{\text{cur}})$ is simpler than $l_p(\eta)$, it is not straightforward to compute its maximizer directly. This is also where Silverman et al. (1990) and Green (1990) stopped implementing the EM algorithm fully.

3.2 M-Step: Maximization of $Q(\eta|\eta_{\text{cur}})$

For convenience, (3.2) can be rewritten as

$$Q(\eta|\eta_{\text{cur}}) = \frac{1}{n}\sum_{i=1}^{n} E[\log f(y_i|x_i)|\vec{y}, \eta_{\text{cur}}] + \int_a^b \eta(x)\psi(x|\vec{y}, \eta_{\text{cur}})\,dx - \log\int_a^b e^{\eta(x)}\,dx - \lambda\int_a^b [\eta''(x)]^2\,dx, \qquad (3.4)$$

where

$$\psi(x|\vec{y}, \eta_{\text{cur}}) = \frac{1}{n}\sum_i \varphi(x|y_i, \eta_{\text{cur}}). \qquad (3.5)$$

Removing the term that does not depend on η and using a similar method by Silverman (1982), we define a new functional

$$\tilde{Q}(\eta|\eta_{\text{cur}}) = \int_a^b \eta(x)\psi(x|\vec{y}, \eta_{\text{cur}})\,dx - \int_a^b e^{\eta(x)}\,dx - \lambda\int_a^b [\eta''(x)]^2\,dx. \qquad (3.6)$$

$\tilde{Q}$ can be used as a surrogate of Q because both functionals share the same maximizer. This property is summarized in the following proposition.

Proposition 2. *Maximizing* $\tilde{Q}(\eta|\eta_{\text{cur}})$ *is equivalent to maximizing* $Q(\eta|\eta_{\text{cur}})$. *If the maximizer of* $\tilde{Q}$ *exists, which is denoted as* $\hat{\eta}$, *it must satisfy* $\int_a^b \exp(\hat{\eta}(x))\,dx = 1$.

The following theorems state that the maximizer of $\tilde{Q}(\eta|\eta_{\text{cur}})$ exists, is unique, and satisfies an ordinary differential equation with some boundary conditions.

Theorem 2. *The maximizer of* $\tilde{Q}(\eta|\eta_{\text{cur}})$ *in* $\mathcal{H}$ *exists and is unique.*

Theorem 3. *If the maximizer of* $\tilde{Q}(\eta|\eta_{\text{cur}})$ *exists and is in* $C^4[a,b]$, *it must satisfy the ordinary differential equation (ODE)*

$$\psi(x|\vec{y}, \eta_{\text{cur}}) - e^{\eta(x)} - 2\lambda\eta^{(4)}(x) = 0 \qquad (3.7)$$

with boundary conditions

$$\eta''(a) = \eta'''(a) = 0, \qquad \eta''(b) = \eta'''(b) = 0. \tag{3.8}$$

The next theorem (Theorem 4) concludes that if a solution of the ODE (3.7) with boundary conditions (3.8) exists, such a solution must be the maximizer of $\tilde{Q}$.

Theorem 4. *If $\eta_*(x) \in \mathcal{H}$ is the solution of the ODE (3.7) with boundary conditions (3.8), then $\tilde{Q}(\eta_*|\eta_{\text{cur}}) \geq \tilde{Q}(\eta|\eta_{\text{cur}})$ for any $\eta \in \mathcal{H}$. Furthermore, the solution of the ODE (3.7) with boundary conditions (3.8) is unique, provided it exists.*

Theorem 2 asserts the existence of the maximizer η_{new} of $\tilde{Q}(\eta|\eta_{\text{cur}})$ in $\mathcal{H}$, which only guarantees that η_{new} and η'_{new} are absolutely continuous on $[a, b]$ and $\eta''_{\text{new}} \in L_2(a, b)$. But this is not enough to derive (3.7) and (3.8), which include a fourth-order differential equation with boundary conditions. To use the above equations, η_{new} needs to be smoother, for example, $\eta_{\text{new}} \in C^4[a, b]$. The smoothness property of η_{new} is referred to as the regularity of the maximizer in the calculus of variations. In fact, the regularity of η_{new} (i.e., the existence of up to fourth derivatives) can be established by applying the results developed in Clarke and Vinter (1990). The smoothness of η_{new} depends on the smoothness of $\psi(\cdot|\vec{y}, \eta_{\text{cur}})$. If η_{cur} is smooth enough, then ψ is smooth enough to guarantee that the maximizer of $\tilde{Q}(\eta|\eta_{\text{cur}})$ has the required smoothness of (3.7) and (3.8). In theory, if we start the algorithm with a smooth function η such as the uniform distribution, then (3.7) and (3.8) can be used to compute η_{new} in all the subsequent M-steps of the FEM algorithm. Readers are referred to Liu (2005) for more technical details. The numerical solution to the nonlinear ordinary differential equation (3.7) with the boundary conditions (3.8) will be discussed in detail in Section 4.

3.3 THE FEM ALGORITHM

Based on the results above, the steps of the FEM algorithm are summarized as follows.

Algorithm 1:

(a) Specify λ.

(b) Set $k = 0$, and select an initial guess of η. Usually we use $\eta_0(x) \equiv \log \frac{1}{b-a}$ for $x \in [a, b]$.

(c) Compute $\psi(x|\vec{y}, \eta_k)$. Numerically solve the ODE (3.7) with boundary conditions (3.8), and denote the solution as η_{k+1}. Normalize the solution before proceeding to the next step as $\eta_{k+1}(x) \leftarrow \eta_{k+1}(x) - \log \int_a^b \exp\{\eta_{k+1}(x)\}\, dx$.

(d) $k \leftarrow k + 1$. Run Step (c) until η_k converges.

Because of Assumption (A1), the penalized likelihood of the uniform distribution is finite. The uniform density usually serves as a good initial guess of the true mixing density. When there is a concern that the FEM algorithm may get trapped by local maxima, other initializations should also be tried. In each M-step, we need to use numerical methods to solve the ordinary differential equation (3.7) with boundary conditions (3.8), which will be the subject of the next section. Notice that we have added a normalization step in Step (c) above. This step is necessary because in theory the solution η_k of the ODE (3.7) already satisfies $\int e^{\eta_k} = 1$ (see Proposition 2), but the numerical solution we actually obtain is only an approximation. The normalization in Step (c) not only makes e^{η_k} a density so that the computed marginal density in the next iteration will be legitimate, but also ensures that $l_p(\text{modified } \eta_k) \geq l_p(\eta_k)$; see the proof of Proposition 2 in the supplementary document.

In Step (d) of the FEM algorithm (i.e., Algorithm 1), the convergence of η_k is assessed by the change in the penalized likelihood $l_p(\eta_k)$. In general, the algorithm is terminated when $l_p(\eta_{k+1}) - l_p(\eta_k)$ is less than a prespecified threshold. In our simulation study, we specify the threshold to be 10^{-15}. The convergence of EM algorithms is generally known to be slow, even when implemented to calculate maximum likelihood estimates in finite-dimensional parametric models. One may question if the convergence of the FEM algorithm is also slow. It turns out that it is not a big concern, because the FEM algorithm usually does not need many iterations to converge. The same kind of quick convergence can also be observed in the EMS algorithm. One possible explanation is that it is the effect caused by the penalty term in the penalized likelihood function l_p in the FEM algorithm and the smoothing step added in the EMS algorithm. In the FEM algorithm, maximizing the penalized likelihood function is equivalent to maximizing the original likelihood function subject to the constraint that the penalty term $J(\eta)$ is less than a prespecified constant, which leads to a restricted thus smaller "parameter" space. Therefore, intuitively, the FEM algorithm does not need to explore the whole parameter space as most traditional EM algorithms usually have to do. The computational burden of M-steps is not severe either because of the high efficiency of the collocation method; we will give a more detailed discussion in the next section.

4. NUMERICAL SOLUTION FOR M-STEPS

Recall that in each M-step of the FEM algorithm, the maximizer is a function satisfying the ordinary differential equation (3.7) with boundary conditions (3.8). When implementing FEM, we need to choose a numerical method to solve (3.7)–(3.8). The collocation method is an efficient and stable method for numerical solution of ordinary differential equations with boundary conditions. In the following, we describe how to apply the collocation method to solving (3.7)–(3.8); more information about this method can be found in Ascher, Mattheij, and Russell (1988).

For convenience, we restate the equations (3.7)–(3.8) as $L[\eta](x) = 0$ and $B[\eta] = 0$ where

$$L[\eta](x) = \psi(x) - e^{\eta(x)} - 2\lambda\eta^{(4)}(x) \tag{4.1}$$

and

$$B[\eta] = \big(\eta''(a), \eta'''(a), \eta''(b), \eta'''(b)\big). \tag{4.2}$$

Here L and B can be viewed as linear operators on $\mathcal{H}$ and $\psi(x)$ is an abbreviation of $\psi(x|\vec{y}, \eta_{\text{cur}})$.

4.1 COLLOCATION METHOD

The collocation method approximates the exact solution of (3.7)–(3.8) with a linear combination of basis functions $\{\phi_d(x)\}_{d=1}^D$:

$$u(x) = \sum_{d=1}^{D} \alpha_d \phi_d(x), \tag{4.3}$$

where $\{\phi_d(x)\}$ satisfy (3.7)–(3.8) at a number of interior points of $[a, b]$. In this article, B-spline functions are used as the basis functions.

Recall that (3.7) is an ODE of order $m = 4$. Let N and k be two positive integers and

$$\pi_0 : a = x_0 < x_1 < \cdots < x_{N-1} < x_N = b$$

be an equally spaced partition of $[a, b]$. We use $\{a_1, a_2, \ldots, a_{k+m}\} \cup \{c_{ij} : 1 \le i \le N - 1, 1 \le j \le k\} \cup \{b_1, b_2, \ldots, b_{k+m}\}$ as the knot vector to construct B-spline functions. Here $a_1 < a_2 < \cdots < a_{k+m} \le a$, $b \le b_1 < b_2 < \cdots < b_{k+m}$, and $c_{ij} = x_i$, $1 \le i \le N - 1$, $1 \le j \le k$. These functions form a basis $\{\phi_d(x)\}_{d=1}^D$ with $D = (N-1)k + 2(k+m) - (k+m) = Nk + m$, which is the length of the knot vector minus the order of basis functions. These functions are the nonuniform B-spline basis functions of order $k + m$. By the standard property of nonuniform B-splines, $u(x) \in C^{(m-1)}[a, b]$, and in each subinterval (x_{i-1}, x_i) $u(x)$ is a polynomial function of degree $k + m - 1$.

Next, we need to determine the interior points of $[a, b]$ where $u(x)$ satisfies $L[u](x) = 0$. The number of interior points required by the collocation method is $D - m = Nk$. The set of points we choose is

$$\pi = \{x_{ij} = x_{i-1} + \rho_j (x_i - x_{i-1}) : 1 \le i \le N, 1 \le j \le k\},$$

where $0 < \rho_1 < \rho_2 < \cdots < \rho_k < 1$ are the abscissas or canonical points for Gaussian quadrature of order k over [0, 1].

The collocation method requires that $u(x)$ should satisfy the following system of D equations with D unknown coefficients:

$$\begin{aligned} L[u](x_{ij}) &= 0, \qquad i = 1, 2, \ldots, N, j = 1, 2, \ldots, k; \\ B[u] &= 0. \end{aligned} \tag{4.4}$$

In the system above, the coefficients are α_d, $1 \le d \le D$. Because the ODE (3.7) is nonlinear, the system (4.4) is also nonlinear. In the next subsection, we describe a quasilinearization method for solving the system (4.4).

4.2 QUASILINEARIZATION

Suppose $u = \sum_d \alpha_d^u \phi_d(x)$ is an initial guess of the solution of (4.4). Using the Gâteaux derivative, we derive the following approximations:

$$\begin{cases} L[u + z](x) \approx L[u](x) - e^{u(x)} z(x) - 2\lambda z^{(4)}(x), & \text{for } x \in \pi \\ B[u + z] \approx B[u] + \big(z''(a), z'''(a), z''(b), z'''(b)\big). \end{cases} \tag{4.5}$$

Based on the approximation (4.5), we use the following iterative procedure to solve the system (4.4):

(a) Solve the linear system with respect to z with u given,

$$\begin{cases} L[u](x) - e^{u(x)}z(x) - 2\lambda z^{(4)}(x) = 0, & \text{for } x \in \pi \\ B[u] + \big(z''(a), z'''(a), z''(b), z'''(b)\big) = 0, & \end{cases}$$

where it is assumed that $z = \sum_d \alpha_d^z \phi_d(x)$; in terms of α_d^z (that are unknown), the system is

$$\begin{cases} \sum_{d=0}^{D}\big(e^{u(x)}\phi_d(x) + 2\lambda\phi_d^{(4)}(x)\big)\alpha_d^z = L[u](x), \quad \text{for } x \in \pi \\ \big(\sum_{d=0}^{D}\phi_d''(a)\alpha_d^z, \sum_{d=0}^{D}\phi_d'''(a)\alpha_d^z, \sum_{d=0}^{D}\phi_d''(b)\alpha_d^z, \sum_{d=0}^{D}\phi_d'''(b)\alpha_d^z\big) = -B[u]. \end{cases}$$

(b) Update $u(x)$ by

$$u(x) \leftarrow u(x) + z(x) = \sum_{d=1}^{D}(\alpha_d^u + \alpha_d^z)\phi_d(x).$$

(c) Repeat Steps (a) and (b) until $\sup|z|$ is below some prespecified threshold.

The prespecified threshold in Step (c) above is chosen to be 10^{-4} in the simulation study that will be discussed in Section 6. Because the algorithm is essentially a combination of solving a system of linear equations and some Newton–Raphson type of iterations, it usually converges fairly quickly and does not give much computational burden to the FEM algorithm. This will be demonstrated in Section 6.

5. DATA-DRIVEN SELECTION OF λ

It is well known that the choice of smoothing parameter is one of the most important steps of the penalized likelihood method in direct density estimation. We expect it to be the same for indirect density estimation. In this section, we begin with briefly reviewing the cross-validation (CV) method as used for selecting λ in direct density estimation. Then, we extend it to select the smoothing parameter λ when estimating the mixing density.

5.1 CV FOR DIRECT DENSITY ESTIMATION

Suppose a sample $x_1, x_2, \ldots, x_n$ is randomly drawn from a density $g_0(x)$. The nonparametric maximum penalized likelihood estimate of g_0 is defined as the maximizer of

$$l_p(g) = \frac{1}{|V|}\sum_{i \in V}\log(g(x_i)) - \lambda J(g),$$

where g is a density, $V = \{1, 2, \ldots, n\}$ is the index set of the sample, and $|V|$ is the cardinality of V. The K-fold CV is a popular method for selecting λ. The data $\{x_i\}_{1 \le i \le n}$ are divided into K disjoint subsets of approximately the same size. Let V_k be the index set of the kth subset, $k = 1, \ldots, K$, $\hat{g}_\lambda(x)$ be the density estimate based on the entire data set, and $\hat{g}_{\lambda,-k}(x)$ be the density estimate based on all data points except those in the kth subset. Two popular CV-type scores, the least squares CV score $\mathrm{LS}(\lambda)$ and the likelihood CV score $\mathrm{KL}(\lambda)$, are routinely used in practice (see Izenman 1991). They are defined as

$$\mathrm{LS}(\lambda) = \int \hat{g}_\lambda(x)^2\,dx - \frac{2}{K}\sum_{k=1}^{K}\frac{1}{|V_k|}\sum_{i \in V_k}\hat{g}_{\lambda,-k}(X_i),$$

$$\mathrm{KL}(\lambda) = -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|V_k|}\sum_{i\in V_k}\log(\hat{g}_{\lambda,-k}(X_i)),$$

respectively. The smoothing parameter is then chosen as the minimizer of either $\mathrm{LS}(\lambda)$ or $\mathrm{KL}(\lambda)$.

5.2 CV FOR INDIRECT DENSITY ESTIMATION

In indirect density estimation, the observed data $\{y_i\}$ are drawn from the mixture density $h(y|g_0)$ instead of the mixing density g_0. Hence, the CV scores $\mathrm{LS}(\lambda)$ and $\mathrm{KL}(\lambda)$ cannot be computed directly. Recall that $\{y_i\}$ can be considered to have been generated from the two-step procedure discussed at the beginning of Section 3.1 whereof a direct sample $\{x_i\}$ from the targeted mixing density is postulated. Although the sample $\{x_i\}$ is latent and thus not available, we can consider the conditional density of x_i given y_i and g_0, $\varphi(x|y_i, g_0) = f(y_i|x)g_0(x)/\int_a^b f(y_i|t)g_0(t)\,dt$. Based on $\varphi(x|y_i, g_0)$, we propose the following two pseudo-CV scores:

$$\mathrm{pLS}(\lambda|g_0) = \int \hat{g}_\lambda(x)^2\,dx - \frac{2}{K}\sum_{k=1}^{K}\frac{1}{|V_k|}\sum_{i\in V_k}\int \hat{g}_{\lambda,-k}(x)\varphi(x|y_i, g_0)\,dx, \qquad (5.1)$$

$$\mathrm{pKL}(\lambda|g_0) = -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|V_k|}\sum_{i\in V_k}\int \log(\hat{g}_{\lambda,-k}(x))\varphi(x|y_i, g_0)\,dx, \qquad (5.2)$$

which correspond to $\mathrm{LS}(\lambda)$ and $\mathrm{KL}(\lambda)$ above, respectively. The following proposition justifies using $\mathrm{pLS}(\lambda|g_0)$ and $\mathrm{pKL}(\lambda|g_0)$ as the cross-validation scores for selecting λ in indirect density estimation.

Proposition 3. *If g_0 is the true mixing density, then*

$$E[\mathrm{pLS}(\lambda|g_0)] = E[\mathrm{LS}(\lambda)] \qquad \textit{and} \qquad E[\mathrm{pKL}(\lambda|g_0)] = E[\mathrm{KL}(\lambda)].$$

Proposition 3 indicates that the expectation of $\mathrm{pLS}(\lambda|g_0)$ [or $\mathrm{pKL}(\lambda|g_0)$] is exactly the same as that of $\mathrm{LS}(\lambda)$ [or $\mathrm{KL}(\lambda)$] based on a sample drawn directly from the true density g_0. Thus, these pseudo-CV scores are analogous to the true CV scores based on observations from the mixing density g_0. However, another difficulty arises when trying to use these scores directly. Note that the true density g_0 is in fact not known and the scores are not computable. Next, we propose an implementable procedure to determine the smoothing parameter λ, treating $\mathrm{pLS}(\lambda|g)$ and $\mathrm{pKL}(\lambda|g)$ as two score functions for any given density g.

Let Λ be a collection of λ values to be considered. For each $\lambda \in \Lambda$, a NPMPLE estimate can be computed by the FEM algorithm and is denoted by $\hat{g}_\lambda$. Our goal is to select the best smoothing parameter from Λ, or equivalently, the best density estimate from $\{\hat{g}_\lambda, \lambda \in \Lambda\}$. Instead of minimizing $\mathrm{pLS}(\lambda|g_0)$ [or $\mathrm{pKL}(\lambda|g_0)$], which is infeasible as pointed out previously, we take a different approach following the self-voting principle proposed by Gu (1992). For any pair of values λ_1 and λ_2 from Λ, define

$$\mathrm{pCV}(\lambda_2|\lambda_1) = \mathrm{pLS}(\lambda_2|\hat{g}_{\lambda_1}) \qquad \text{or} \qquad \mathrm{pCV}(\lambda_2|\lambda_1) = \mathrm{pKL}(\lambda_2|\hat{g}_{\lambda_1}),$$

depending on which pseudo-CV score is used. $\mathrm{pCV}(\lambda_2|\lambda_1)$ can be viewed as the voting score from λ_2 to λ_1. Gu's self-voting principle in our setting states that the optimal smoothing parameter must satisfy

$$\mathrm{pCV}(\lambda^*|\lambda^*) \leq \mathrm{pCV}(\lambda|\lambda^*) \quad \text{for any } \lambda \in \Lambda. \tag{5.3}$$

In other words, the optimal λ^* or the corresponding density estimate $\hat{g}_{\lambda^*}$ must vote for itself. In general, the smoothing parameter satisfying the self-voting principle is not unique. In particular, the principle tends to be satisfied by small λ values. Hence, the self-voting principle is not enough for determining the optimal smoothing parameter uniquely. We suggest using a version of this principle supplemented by the maximum smoothing principle to choose the optimal λ. Because the larger λ is, the smoother the density estimate $\hat{g}_\lambda$ is, our maximum smoothing principle states that the largest λ satisfying (5.3) should be selected. Jones, Marron, and Sheather (1996) commented that the largest local minimizer of $\mathrm{CV}(h)$ in the kernel density estimation setting usually gives better estimates than the global minimizer of $\mathrm{CV}(h)$. This is analogous to our maximum smoothing principle. Hall and Marron (1991) observed that spurious local minima of the cross-validation function $\mathrm{CV}(h)$ are more likely to occur when the bandwidth values used are very small rather than very large. We combine the self-voting principle and the maximum smoothing principle in the following algorithm to obtain the optimal density estimate.

Algorithm 2:

```
(a) Specify $\Lambda$ and divide data randomly into $K$ subsets of approximately the same size.
(b) For each $\lambda \in \Lambda$, use Algorithm 1 to compute $\hat{g}_\lambda$, and $\hat{g}_{\lambda,-k}$ for $1 \leq k \leq K$.
(c) Find $w(\lambda) = \arg\min_{\lambda_1 \in \Lambda} \mathrm{pCV}(\lambda_1|\lambda)$ for each $\lambda \in \Lambda$.
(d) Find $\lambda^* = \arg\max\{\lambda : \lambda = w(\lambda)\}$; then output $\hat{g}_{\lambda^*}$.
```

6. SIMULATIONS

We have conducted various simulation studies to compare the performance of the FEM algorithm and the EMS algorithm proposed by Silverman et al. (1990). The EMS algorithm has been shown to be an effective algorithm for estimating mixing densities, and it usually outperforms the kernel method proposed by Stefanski and Carroll (1990) and Fan (1991) in the case of deconvolution; see Eggermont and LaRiccia (1997). In this section, we report simulation results for two deconvolution problems and one general mixture problem; and the computing costs of the two algorithms are also compared and discussed. The effectiveness of our smoothing parameter selection procedure is also demonstrated by a simulation example.

6.1 FEM VERSUS EMS IN DECONVOLUTION

In this simulation study, we compare FEM and EMS in deconvolution only, in which random samples generated from

$$h(y) = \int_0^1 \phi(y-x)g(x)\,dx \tag{6.1}$$

are used to estimate the mixing density $g(x)$. Let us denote φ the density of the standard normal distribution $N(0,1)$ and $\beta(x;2,4)$ the density of the beta distribution Beta$(2,4)$. Six different mixing densities denoted by $\{g_i\}_{i=1}^6$ and two different component densities denoted by $\{\phi_j\}_{j=1}^2$ are considered; they are

$$g_1(x) \propto 1 + \beta(x;2,4), \qquad x \in [0,1];$$

$$g_2(x) \propto \frac{1}{3}\varphi\left(\frac{x-0.3}{0.1}\right) + \frac{2}{3}\varphi\left(\frac{x-0.7}{0.1}\right), \qquad x \in [0,1];$$

$$g_3(x) \propto \frac{3}{10}\varphi\left(\frac{x-0.1}{0.1}\right) + \frac{4}{10}\varphi\left(\frac{x-0.5}{0.1}\right) + \frac{3}{10}\varphi\left(\frac{x-0.85}{0.1}\right), \qquad x \in [0,1];$$

$$g_4(x) \propto \exp(-5x), \qquad x \in [0,1];$$

$$g_5(x) \propto \exp(x^2 - 1.2x), \qquad x \in [0,1];$$

$$g_6(x) \propto \exp(x^4 - 1.2x) - 0.5, \qquad x \in [0,1];$$

$$\phi_1(x) = \varphi(x/0.05) \qquad \text{and} \qquad \phi_2(x) = 10\sqrt{2}\exp(-20\sqrt{2}|x|).$$

In the above, $\phi_1(x)$ and $\phi_2(x)$ are the densities of the normal distribution and the double exponential distribution with mean 0 and standard deviation 0.05, respectively. All of the mixing densities considered (i.e., g_1 to g_6) have $[0,1]$ as their support. Following (6.1), each combination of g_i $(1 \le i \le 6)$ and ϕ_j $(j = 1,2)$ generates a mixture density, denoted by h_{ij}. In total, twelve mixture densities are used in the simulation study.

Three different distance measures are used to calculate the distance between the density estimate $\hat{g}$ and the true density g. They are the integrated squared error distance $\mathrm{ISE}(g,\hat{g}) = \int_a^b [g(x) - \hat{g}(x)]^2\,dx$, the integrated absolute error distance $\mathrm{IAE}(g,\hat{g}) = \int_a^b |g(x) - \hat{g}(x)|\,dx$, and the Kullback–Leibler distance $\mathrm{KLD}(g,\hat{g}) = \int_a^b \log(g(x)/\hat{g}(x)) \times g(x)\,dx$. To compare FEM and EMS directly and eliminate the impact of smoothing parameter selection on their performances, we adopt the strategy of comparing the estimates based on oracle choices of smoothing parameters. For both FEM and EMS, the oracle choice of smoothing parameter is the one that minimizes the average distance between the true density and the corresponding density estimate. For FEM, the best smoothing parameter λ is chosen from $S_\lambda = \{10^{-8} \times 2^{k/2}\}_{k=0}^{40}$, whereas for EMS, the best smoothing parameter J is chosen from $S_J = \{4 \times l + 1\}_{l=1}^{36}$. In the simulation study, the EMS algorithm is based on a grid of size 150.

The basic simulation scheme is given below, where L denotes the distance measure that can be either ISE, IAE, or KLD as defined above.

(a) For fixed i and j, generate N independent samples, each of size n, from the mixture density h_{ij}. Denote the kth sample $\{y_{kl}\}_{l=1}^n$ $(1 \le k \le N)$.

(b) For each sample $\{y_{kl}\}_{l=1}^n$, each smoothing parameter $\lambda \in S_\lambda$, and each smoothing parameter $J \in S_J$, use the FEM algorithm (Algorithm 1) and the EMS algorithm, separately, to compute the density estimates, which are denoted by $\hat{g}_{\lambda,k}^{\mathrm{FEM}}$ and $\hat{g}_{J,k}^{\mathrm{EMS}}$, respectively.

(c) For a given distance measure $L(g, \hat{g})$, find

$$\tilde{\lambda} = \arg\min_{\lambda \in S_\lambda} \frac{1}{N} \sum_{k=1}^{N} L(g, \hat{g}_{\lambda,k}^{\text{FEM}}) \qquad \text{and} \qquad \tilde{J} = \arg\min_{J \in S_J} \frac{1}{N} \sum_{k=1}^{N} L(g, \hat{g}_{J,k}^{\text{EMS}}).$$

(d) Compare $\{L(g, \hat{g}_{\tilde{\lambda},k}^{\text{FEM}})\}_{k=1}^{N}$ and $\{L(g, \hat{g}_{\tilde{J},k}^{\text{EMS}})\}_{k=1}^{N}$, using summary statistics such as mean, standard deviation, and side-by-side boxplots.

In Step (c) of the above scheme, $\frac{1}{N}\sum_{k=1}^{N} L(g, \hat{g}_{\lambda,k}^{\text{FEM}}) \approx E[L(g, \hat{g}_{\lambda,j}^{\text{FEM}})]$ is the average distance between a density estimate using the smoothing parameter λ and the true density; thus $\tilde{\lambda}$ is the optimal smoothing parameter in that it minimizes this average distance. The same interpretation applies to $\tilde{J}$. The scheme has been applied to every h_{ij} ($1 \le i \le 6$; $1 \le j \le 2$) with sample sizes $n = 150$ and $n = 400$. In every case, 100 replications are used and the results are recorded for all three distance measures ($L = \text{ISE}, \text{IAE}, \text{ or KLD}$). Therefore, there are in total 72 different scenarios. The simulation results under the scenarios with the normal component density ($\phi = \phi_1$) including means, standard errors, side-by-side boxplots are reported in Table 1 and Figure 1, respectively. To conserve space, only the means and standard errors under the scenarios with the double exponential component density ($\phi = \phi_2$) are reported (Table 2). Note that the results in both Tables 1 and 2 show that the FEM algorithm outperformed the EMS algorithm under most scenarios in average performance as well as performance stability. The small absolute values of the error criteria may make it somewhat hard to appreciate this fact at first. Because of that, the *relative*

Table 1. Deconvolution with the normal component density $\phi = \phi_1$.

		$n = 150$					$n = 400$				
		FEM		EMS			FEM		EMS		
g	Dist.	Mean	St. error	Mean	St. error	RC	Mean	St. error	Mean	St. error	RC
g_1	ISE	0.0206	0.00117	0.0431	0.00192	52.2%	0.0113	0.00065	0.0177	0.00091	36.2%
	IAE	0.1096	0.00348	0.1637	0.00374	33.0%	0.0783	0.00249	0.1022	0.00264	3.4%
	KLD	0.0115	0.00069	0.0233	0.00109	50.6%	0.0061	0.00032	0.0093	0.00043	34.4%
g_2	ISE	0.0474	0.00274	0.0526	0.00321	9.9%	0.0240	0.00125	0.0273	0.00157	12.1%
	IAE	0.1635	0.00457	0.1712	0.00529	4.5%	0.1151	0.00299	0.1235	0.00336	6.8%
	KLD	0.0246	0.00124	0.0286	0.00144	14.0%	0.0127	0.00059	0.0150	0.00071	15.3%
g_3	ISE	0.0617	0.00322	0.0567	0.00278	−8.8%	0.0264	0.00113	0.0268	0.00108	1.5%
	IAE	0.1921	0.00508	0.1881	0.00486	−2.1%	0.1277	0.00293	0.1298	0.00293	1.6%
	KLD	0.0315	0.00162	0.0297	0.00147	−6.1%	0.0135	0.00056	0.0140	0.00055	3.6%
g_4	ISE	0.0116	0.00205	0.0482	0.00416	75.9%	0.0044	0.00060	0.0262	0.00163	83.2%
	IAE	0.0543	0.00427	0.1347	0.00447	68.8%	0.0327	0.00264	0.0947	0.00264	65.5%
	KLD	0.0041	0.00070	0.0284	0.00232	85.6%	0.0015	0.00021	0.0108	0.00057	86.1%
g_5	ISE	0.0121	0.00118	0.0485	0.00223	75.1%	0.0049	0.00049	0.0165	0.00089	70.3%
	IAE	0.0812	0.00430	0.1748	0.00438	53.5%	0.0510	0.00267	0.1013	0.00269	49.7%
	KLD	0.0061	0.00060	0.0249	0.00127	75.5%	0.0024	0.00025	0.0083	0.00045	71.1%
g_6	ISE	0.0391	0.00508	0.0496	0.00381	21.2%	0.0147	0.00114	0.0215	0.00107	31.6%
	IAE	0.1267	0.00530	0.1613	0.00514	21.5%	0.0833	0.00272	0.1056	0.00267	21.1%
	KLD	0.0145	0.00132	0.0250	0.00142	42.0%	0.0062	0.00039	0.0104	0.00046	40.4%

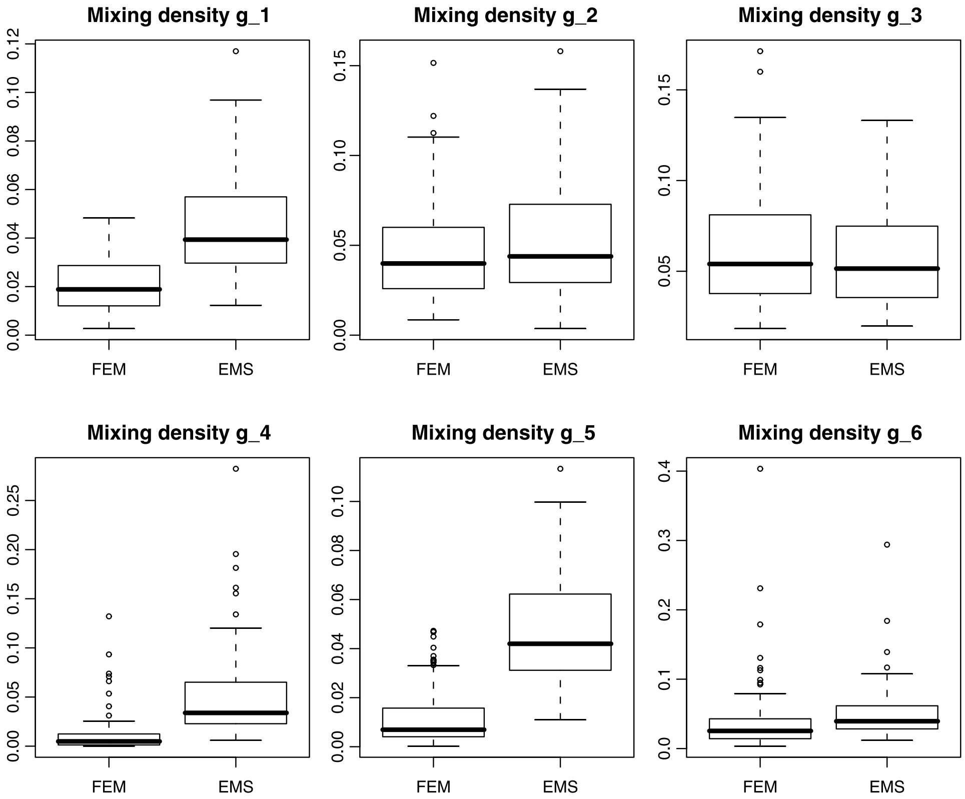

Figure 1. Side-by-side boxplots of the ISE values of the estimates generated by the FEM algorithm and those generated by the EMS algorithm in the case of the normal component density $\phi = \phi_1$. The sample size is 150 and the number of replications is 100.

changes in the respective risk criterion for both sample sizes are given in both Tables 1 and 2 under the heading RC ("Relative Change") to help understand the difference. The side-by-side boxplots (Figure 1) for the cases with the normal component density and the sample size $n = 150$ further demonstrate the superiority of FEM over EMS. The only case in which the EMS algorithm outperformed the FEM algorithm is under the scenario of the mixing density g_3 and $n = 150$; and the improvement is quite small. Under the scenarios of the mixing densities g_1, g_4, and g_5, the relative improvement of FEM over EMS is indeed significant, especially when $n = 150$.

To check the visual effect of the estimates, we have also plotted the density estimates generated by the FEM algorithm and the EMS algorithm. The true density is shown for comparison. The smoothing parameters are oracle ones for all the density estimates plotted. Due to space limitation, only two sets of such plots are included in Figures 2 and 3. The overall impression is that the estimates generated by the FEM algorithm recover the true density better than the estimates generated by the EMS algorithm. The EMS estimates tend to be less smooth than the FEM estimates as demonstrated in Figure 2; they also seem to be worse in capturing the strongly pronounced peaks and the deep valleys (see Figure 3).

Table 2. Deconvolution with the double exponential component density $\phi = \phi_2$.

		$n = 150$					$n = 400$				
		FEM		EMS			FEM		EMS		
g	Dist.	Mean	St. error	Mean	St. error	RC	Mean	St. error	Mean	St. error	RC
g_1	ISE	0.0206	0.00116	0.0446	0.00195	53.8%	0.0113	0.00064	0.0180	0.00089	37.2%
	IAE	0.1095	0.00343	0.1662	0.00365	34.1%	0.0777	0.00247	0.1034	0.00257	24.9%
	KLD	0.0114	0.00069	0.0240	0.00107	52.5%	0.0060	0.00032	0.0095	0.00043	36.8%
g_2	ISE	0.0472	0.00283	0.0533	0.00330	11.4%	0.0233	0.00119	0.0276	0.00150	15.6%
	IAE	0.1626	0.00466	0.1719	0.00531	5.4%	0.1137	0.00294	0.1247	0.00328	8.8%
	KLD	0.0245	0.00125	0.0292	0.00147	16.1%	0.0125	0.00058	0.0154	0.00069	18.8%
g_3	ISE	0.0608	0.00304	0.0571	0.00274	−0.6%	0.0249	0.00106	0.0263	0.00103	5.3%
	IAE	0.1911	0.00495	0.1889	0.00477	−0.1%	0.1246	0.00282	0.1290	0.00275	3.4%
	KLD	0.0309	0.00150	0.0298	0.00144	−0.4%	0.0129	0.00053	0.0137	0.00052	5.8%
g_4	ISE	0.0117	0.00205	0.0477	0.00410	75.5%	0.0043	0.00061	0.0259	0.00165	83.3%
	IAE	0.0535	0.00425	0.1354	0.00431	60.5%	0.0328	0.00263	0.0944	0.00268	65.3%
	KLD	0.0041	0.00070	0.0291	0.00237	85.9%	0.0015	0.00021	0.0109	0.00056	86.2%
g_5	ISE	0.0120	0.00118	0.0499	0.00236	80.0%	0.0049	0.00049	0.0169	0.00089	71%
	IAE	0.0809	0.00432	0.1780	0.00436	54.6%	0.0511	0.00266	0.1025	0.00261	50.1%
	KLD	0.0060	0.00060	0.0256	0.00128	76.6%	0.0024	0.00025	0.0085	0.00044	71.8%
g_6	ISE	0.0389	0.00521	0.0506	0.00393	23.1%	0.0147	0.00115	0.0215	0.00107	31.6%
	IAE	0.1265	0.00510	0.1632	0.00504	22.5%	0.0831	0.00273	0.1056	0.00264	21.3%
	KLD	0.0145	0.00134	0.0252	0.00143	42.5%	0.0062	0.00039	0.0103	0.00046	39.8%

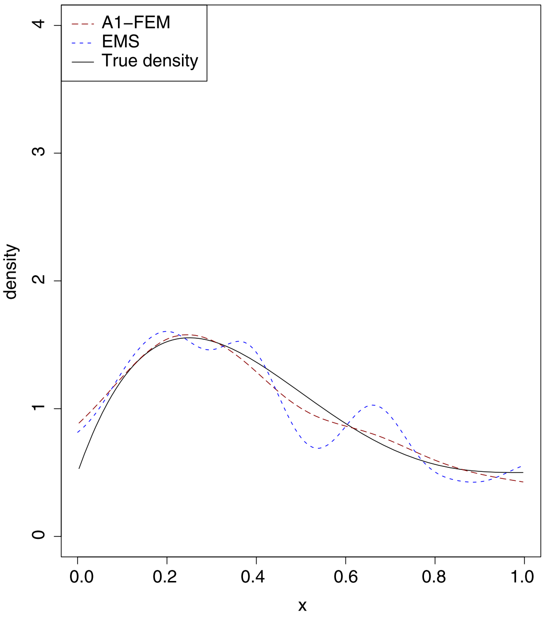

Figure 2. Solid line: the true mixing density (g_1); long-dashed line: the estimate by the FEM algorithm; dashed line: the estimate by the EMS algorithm. All smoothing parameters are oracle ones. The sample size is 150 and the component density is normal.

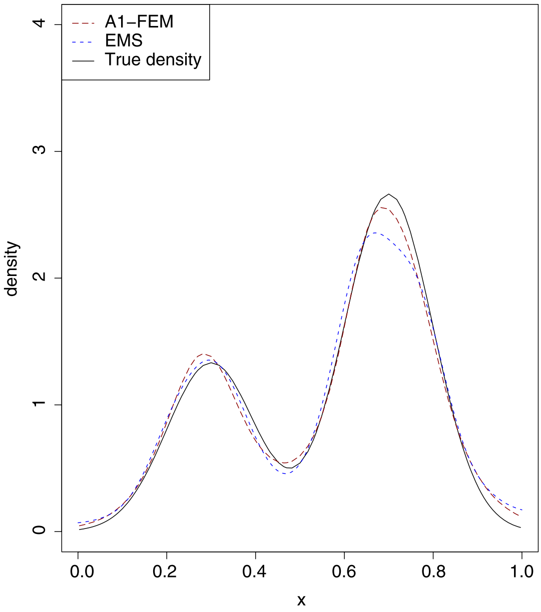

Figure 3. Solid line: the true mixing density (g_2); long-dashed line: the estimate by the FEM algorithm; dashed line: the estimate by the EMS algorithm. All smoothing parameters are oracle ones. The sample size is 150 and the component density is normal.

To compare the computational costs of the two algorithms, we recorded the amount of CPU time required to calculate the estimates under each choice of smoothing parameter considered in the simulation scheme. Note that the required CPU time differs for different smoothing parameters in both the FEM algorithm and the EMS algorithm. Both the algorithms were coded in R (Windows version 2.6.2) and run on a computer with Intel Pentium (R) 4 CPU at 3.0 GHz with 1GB of RAM. The operating system of the machine is Windows XP Professional Service Pack 2. Our general impression is that these two algorithms are comparable in terms of computational cost. To facilitate fair comparison, we report only the time required to calculate the estimates under the optimal smoothing parameter selected by oracle. Table 3 includes the average CPU time used to calculate the estimates by the FEM algorithm and the EMS algorithm, respectively, based on 100 replicated samples. The component density function is $\phi = \phi_1$ and the sample size is 150. The table shows that under the scenarios g_2, g_3, and g_6, the FEM algorithm was slower than the EMS algorithm; but under the scenarios g_1, g_4, and g_5, the FEM algorithm was in fact faster. And under none of the scenarios, one algorithm was slower or faster than the

Table 3. Average CPU time (in sec) for computing the oracle estimates by FEM and EMS.

Algorithm	g_1	g_2	g_3	g_4	g_5	g_6
FEM	0.0218	0.4111	0.5146	0.2152	0.1350	0.2709
EMS	0.0436	0.2122	0.2194	0.2232	0.1972	0.2123

other by more than one and half folds. This is consistent with our discussion at the end of Section 3.3 that the FEM algorithm does not incur too much computational burden, even though it is a fully implemented nonparametric EM procedure.

6.2 FEM versus EMS in Non-Deconvolution

We draw an i.i.d. sample $y_1, y_2, \ldots, y_n$ from the density

$$h(y) = \int_a^b \gamma(y; 25, x/25) g(x)\, dx, \tag{6.2}$$

where $\gamma(y; 25, x/25)$ is the density of the Gamma distribution with a shape parameter $\alpha = 25$ and a scale parameter $\theta = x/25$. Given $x > 0$, the standard deviation of the distribution with density $\gamma(y; 25, x/25)$ is $\sqrt{25(x/25)^2} = x/5$. The same simulation scheme as stated in the previous subsection was used to compare the performances of FEM and EMS in this example. The numerical results are summarized in Table 4. Due to limited space, we only present one plot including the estimates generated by the algorithms in Figure 4 and omit the other plots. Again, the FEM algorithm demonstrated much better performance in terms of mean and variability in all scenarios except the one of density g_3. Visually the FEM algorithm generated smoother estimates than the EMS algorithms. The computing costs of the two algorithms are similar. Therefore, the FEM algorithm outperforms the EMS algorithms in this non-deconvolution problem.

Table 4. Non-deconvolution with the gamma component density γ.

		$n = 150$					$n = 400$				
		FEM		EMS			FEM		EMS		
g	Dist.	Mean	St. error	Mean	St. error	RC	Mean	St. error	Mean	St. error	RC
g_1	ISE	0.0204	0.00140	0.0422	0.00207	51.7%	0.0110	0.00070	0.0168	0.00088	34.5%
	IAE	0.1070	0.00411	0.1634	0.00207	34.5%	0.0792	0.00262	0.1007	0.00262	21.4%
	KLD	0.0114	0.00078	0.0238	0.00142	52.1%	0.0059	0.00033	0.0087	0.00044	32.2%
g_2	ISE	0.1030	0.00865	0.1255	0.00990	17.9%	0.0457	0.00304	0.0567	0.00365	19.4%
	IAE	0.2313	0.00883	0.2611	0.00991	11.4%	0.1565	0.00516	0.1787	0.00580	12.4%
	KLD	0.0497	0.00367	0.0658	0.00413	24.5%	0.0234	0.00140	0.0317	0.00173	26.2%
g_3	ISE	0.1229	0.00456	0.1007	0.00399	−22.0%	0.0790	0.00462	0.0690	0.00239	−14.5%
	IAE	0.2770	0.00684	0.2548	0.00519	−8.7%	0.2112	0.00616	0.2096	0.00402	−0.8%
	KLD	0.0604	0.00209	0.0504	0.00193	−19.8%	0.0388	0.00210	0.0342	0.00111	−13.5%
g_4	ISE	0.0139	0.00241	0.0502	0.00458	72.3%	0.0050	0.00071	0.0221	0.00157	77.4%
	IAE	0.0587	0.00453	0.1372	0.00522	57.2%	0.0349	0.00281	0.0897	0.00302	61.1%
	KLD	0.0048	0.00081	0.0259	0.00205	81.5%	0.0017	0.00024	0.0102	0.00066	83.3%
g_5	ISE	0.0132	0.00130	0.0445	0.00237	70.3%	0.0056	0.00053	0.0194	0.00109	71.1%
	IAE	0.0850	0.00443	0.1648	0.00464	48.4%	0.0553	0.00270	0.1095	0.00322	49.5%
	KLD	0.0066	0.00068	0.0223	0.00122	70.4%	0.0027	0.00026	0.0097	0.00054	72.2%
g_6	ISE	0.0451	0.00452	0.0621	0.00387	27.4%	0.0209	0.00151	0.0308	0.00152	33.0%
	IAE	0.1434	0.00536	0.1817	0.00505	21.1%	0.0997	0.00345	0.1231	0.00295	19.0%
	KLD	0.0198	0.00152	0.0320	0.00203	38.1%	0.0095	0.00073	0.0159	0.00089	40.3%

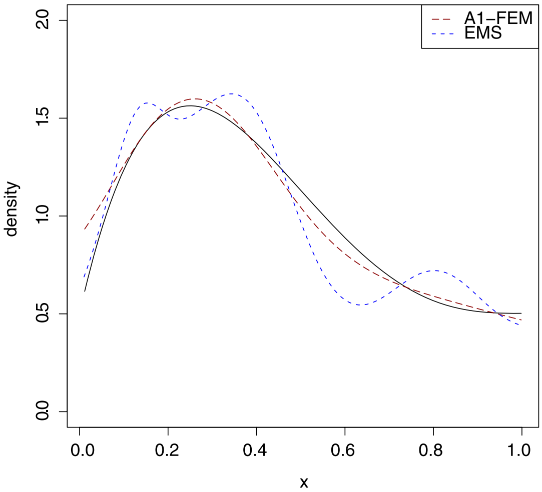

Figure 4. Solid line: true mixing density (g_1); long-dashed line: estimate by Algorithm 1; dashed line: estimate by the EMS algorithm. Smoothing parameters are oracle ones. The sample size is 150 and the component density is gamma.

6.3 EFFECTIVENESS OF SMOOTHING PARAMETER SELECTION

Recall that the self-voting principle and the maximum smoothing principle are used to select λ. In this subsection, we show the effectiveness of this approach by comparing $\min_\lambda \text{ISE}(\hat{g}_\lambda, g)$ with $\text{ISE}(\hat{g}_{\lambda_*^{\text{LS}}}, g)$ and $\min_\lambda \text{KLD}(\hat{g}_\lambda, g)$ with $\text{KLD}(\hat{g}_{\lambda_*^{\text{KL}}}, g)$, where λ_*^{LS} and λ_*^{KL} are the values selected by the pLS CV score and the pKL CV score, respectively. The deconvolution examples from Section 6.1 are used in the comparison. Recall that $g \in \{g_i\}_{i=1}^6$, $\phi \in \{\phi_1, \phi_2\}$, and $S_\lambda = \{10^{-8} \times 2^{k/2}\}_{k=0}^{40}$. Let $n = 400$, $N = 100$, and $K = 10$. We randomly partition $\{1, 2, \ldots, n\}$ into ten folds of roughly the same size, which are denoted as $V_1, V_2, \ldots, V_K$. The basic comparison procedure is given below. Note that the pseudo-CV score in the procedure can be pLS or pKL.

(a) Generate N samples of size n. Denote the jth sample as $\{y_{ij}\}_{i=1}^n$ where $j = 1, 2, \ldots, N$.

(b) For any $1 \le j \le N$ and $\lambda \in S_\lambda$, use Algorithm 1 to compute the density estimate based on $\{y_{ij}\}_{i=1}^n$ and denote the resulting estimate as $\hat{g}_{\lambda,j}$; for any $1 \le k \le K$, compute the estimate based on $\{y_{ij}\}_{i \notin V_k}$ and denote the resulting estimate as $\hat{g}_{\lambda,j}^{[-k]}$, $k = 1, 2, \ldots, K$.

(c) Compute the pseudo-CV score $\text{pCV}_j(\lambda'|\lambda)$ for any $\lambda, \lambda' \in S_\lambda$, where the subscript j indicates that the pseudo-CV score is based on the jth sample.

(d) Apply the self-voting and maximum smoothing principles to select λ, that is, to find the largest $\lambda \in S_\lambda$ that satisfies $\text{pCV}_j(\lambda|\lambda) = \max_{\lambda' \in S_\lambda} \text{pCV}_j(\lambda'|\lambda)$. Denote the result by λ_j^{LS} or λ_j^{KL} depending on whether pLS or pKL is used as the pCV score.

(e) Generate the scatterplots of $\text{ISE}(\hat{g}_{\lambda_j^{\text{LS}},j}, g)$ versus $\min_{\lambda \in S_\lambda} \text{ISE}(\hat{g}_{\lambda,j}, g)$ and $\text{KLD}(\hat{g}_{\lambda_j^{\text{KL}},j}, g)$ versus $\min_{\lambda \in S_\lambda} \text{KLD}(\hat{g}_{\lambda,j}, g)$, separately, where $1 \le j \le N$.

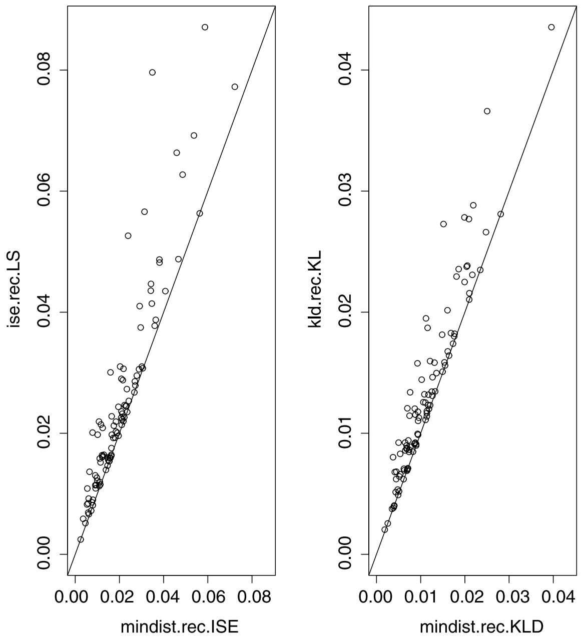

Figure 5. The left plot: $\mathrm{ISE}(\hat{g}_{\lambda_*^{\mathrm{LS}}}, g)$ versus $\min_\lambda \mathrm{ISE}(\hat{g}_\lambda, g)$; the right plot: $\mathrm{KLD}(\hat{g}_{\lambda_*^{\mathrm{KL}}}, g)$ versus $\min_\lambda \mathrm{KLD}(\hat{g}_\lambda, g)$. λ_*^{LS} and λ_*^{KL} are data-driven selected smoothing parameters based on pLS and pKL, respectively. The comparisons are based on g_2.

In essence, the above procedure is to compare the density estimates based on λ selected by oracle and by Algorithm 2 in various deconvolution problems. A representative pair of plots generated from the procedure are shown in Figure 5. In the left plot, the vertical axis represents $\mathrm{ISE}(\hat{g}_{\lambda_j^{\mathrm{LS}},j}, g)$ whereas the horizontal axis represents $\min_{\lambda \in S_\lambda} \mathrm{ISE}(\hat{g}_{\lambda,j}, g)$. Notice that the majority of the points are close to the straight line $y = x$. This indicates the performances of the oracle estimate and the estimate based on the λ selected by Algorithm 2 are similar to each other. The right plot is for $\mathrm{KLD}(\hat{g}_{\lambda_j^{\mathrm{KL}},j}, g)$ and $\min_{\lambda \in S_\lambda} \mathrm{KLD}(\hat{g}_{\lambda,j}, g)$, and it demonstrates the same pattern as the left plot. Both plots suggest that Algorithm 2 is an acceptable smoothing parameter selection procedure.

7. CONCLUDING REMARKS

In this article, we have proposed the FEM algorithm to compute the mixing density in a mixture model. The algorithm can be considered an extension of the maximum penalized likelihood approach for direct density estimation to indirect density estimation. Simulation studies have shown that the new algorithm outperforms many existing methods such as the EMS algorithm and kernel methods. We have proposed to use Gu's self-voting principle and the maximum smoothing principle to select the smoothing parameter. Though it performs well in general, the optimal selection of the smoothing parameter for the FEM algorithm is still an open problem and invites further study. An important characteristic of our work is the use of methods from the calculus of variations and differential equations.

As a matter of fact, theories and methods in the calculus of variations and differential equations are developed to study functions that possess certain optimality over various function spaces. They are naturally related to many nonparametric function estimation problems in statistics. We believe that their use in statistics deserves further exploration.

SUPPLEMENTAL MATERIALS

Computer Code: See the supplemental files `!Read Me.pdf` and `Codereadme.txt` for details. Both files are in the archive. (Computer-code.tar, tar archive)

Appendix: This file contains the proofs of the major propositions, theorems and corollaries in the paper. (appendix.pdf, pdf file)

[Received September 2007. Revised April 2009.]

REFERENCES

Adams, R. (1975), *Sobolev Spaces*, New York: Academic Press.

Ascher, U. M., Mattheij, R. M., and Russell, R. D. (1988), *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Englewood Cliffs, NJ: Prentice Hall.

Clarke, F. H., and Vinter, R. B. (1990), "A Regularity Theory for Variational Problems With Higher Order Derivatives," *Transactions of the American Mathematical Society*, 320, 227–251.

De Montricher, G. M., Tapia, R. A., and Thompson, J. R. (1975), "Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods," *The Annals of Statistics*, 3, 1329–1348.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Eggermont, P., and LaRiccia, V. (2001), *Maximum Penalized Likelihood: Volume I, Density Estimation*, New York: Springer.

Eggermont, P. P. B., and LaRiccia, V. N. (1995), "Maximum Smoothed Likelihood Density Estimation for Inverse Problems," *The Annals of Statistics*, 23, 199–220.

——— (1997), "Nonlinearly Smoothed EM Density Estimation With Automated Smoothing Parameter Selection for Nonparametric Deconvolution Problems," *Journal of the American Statistical Association*, 92, 1451–1458.

Fan, J. (1991), "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *The Annals of Statistics*, 19, 1257–1272.

Fan, J., and Koo, J.-Y. (2002), "Wavelet Deconvolution," *IEEE Transactions on Information Theory*, 48, 734–747.

Good, I. J., and Gaskins, R. A. (1971), "Nonparametric Roughness Penalties for Probability Densities," *Biometrika*, 58, 255–277.

Goutis, C. (1997), "Nonparametric Estimation of a Mixing Density via the Kernel Method," *Journal of the American Statistical Association*, 92, 1445–1450.

Green, P. J. (1990), "On the Use of the EM Algorithm for Penalized Likelihood Estimation," *Journal of the Royal Statistical Society*, Ser. B, 52, 443–452.

Gu, C. (1992), "Cross-Validating Non-Gaussian Data," *Journal of Computational and Graphical Statistics*, 1, 169–179.

——— (1993), "Smoothing Spline Density Estimation: A Dimensionless Automatic Algorithm," *Journal of the American Statistical Association*, 88, 495–504.

Gu, C., and Qiu, C. (1993), "Smoothing Spline Density Estimation: Theory," *The Annals of Statistics*, 21, 217–234.

Hall, P., and Marron, J. S. (1991), "Local Minima in Cross-Validation Functions," *Journal of the Royal Statistical Society*, Ser. B, 53, 245–252.

Izenman, A. J. (1991), "Recent Developments in Nonparametric Density Estimation," *Journal of the American Statistical Association*, 86, 205–224.

Johnstone, I. M., and Silverman, B. W. (1990), "Speed of Estimation in Positron Emission Tomography and Related Inverse Problems," *The Annals of Statistics*, 18, 251–280.

Jones, M. C., and Silverman, B. W. (1989), "An Orthogonal Series Density Estimation Approach to Reconstructing Positron Emission Tomography Images," *Journal of Applied Statistics*, 16, 177–191.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996), "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, 91, 401–407.

Klonias, V. K. (1982), "Consistency of a Nonparametric Penalized Likelihood Estimator of the Probability Desity Function," *The Annals of Statistics*, 10, 811–824.

Koo, J.-Y., and Chung, H.-Y. (1998), "Log-Density Estimation in Linear Inverse Problems," *The Annals of Statistics*, 26, 335–362.

Koo, J.-Y., and Park, B. U. (1996), "B-Splines Deconvolution Based on the EM Algorithm," *Journal of Statistical Computation and Simulation*, 54, 275–288.

Laird, N. (1978), "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association*, 73, 805–811.

Laird, N. M., and Louis, T. A. (1991), "Smoothing the Non-Parametric Estimate of a Prior Distribution by Roughening: An Empirical Study," *Computational Statistics and Data Analysis*, 12, 27–38.

Leonard, T. (1978), "Density Estimation, Stochastic Processes and Prior Information," *Journal of the Royal Statistical Society*, Ser. B, 40, 113–132.

Lindsay, B. G. (1983a), "The Geometry of Mixture Likelihoods: A General Theory," *The Annals of Statistics*, 11, 86–94.

——— (1983b), "The Geometry of Mixture Likelihoods, Part II: The Exponential Family," *The Annals of Statistics*, 11, 783–792.

——— (1995), *Mixture Models: Theory, Geometry, and Applications*, Hayward, CA: Institute of Mathematical Statistics/Alexandria, VA: American Statistical Association.

Liu, L. (2005), "On the Estimation of Mixing Distributions: NPMLE and NPMPLE," Ph.D. thesis, Department of Statistics, Purdue University.

Magder, L. S., and Zeger, S. L. (1996), "A Smooth Nonparametric Estimate of a Mixing Distribution Using Mixtures of Gaussians," *Journal of the American Statistical Association*, 91, 1141–1151.

Pensky, M., and Vidakovic, B. (1999), "Adaptive Wavelet Estimator for Nonparametric Density Deconvolution," *The Annals of Statistics*, 27, 2033–2053.

Silverman, B. W. (1982), "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method," *The Annals of Statistics*, 10, 795–810.

——— (1986), *Density Estimation for Statistics and Data Analysis*, London/New York: Chapman & Hall.

Silverman, B. W., Jones, M. C., Wilson, J. D., and Nychka, D. W. (1990), "A Smoothed EM Approach to Indirect Estimation Problems, With Particular Reference to Stereology and Emission Tomography," *Journal of the Royal Statistical Society*, Ser. B, 52, 271–324.

Stefanski, L., and Carroll, R. J. (1990), "Deconvoluting Kernel Density Estimators," *Statistics*, 21, 169–184.

Szkutnik, Z. (2003), "Doubly Smoothed EM Algorithm for Statistical Inverse Problems," *Journal of the American Statistical Association*, 98, 178–190.

Tapia, R. A., and Thompson, J. R. (1978), *Nonparametric Probability Density Estimation*, Baltimore, MD: Johns Hopkins University Press.

Vardi, Y., Shepp, L. A., and Kaufman, L. (1985), "A Statistical Model for Positron Emission Tomography," *Journal of the American Statistical Association*, 80, 8–20.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia, PA: SIAM.

Zhang, C.-H. (1990), "Fourier Methods for Estimating Mixing Densities and Distributions," *The Annals of Statistics*, 18, 806–831.