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

by

L. Liu, M. Levine, and Y. Zhu
Purdue University

Technical Report #07-07

Department of Statistics
Purdue University

October 2007

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

Lei Liu, Michael Levine, Yu Zhu
Department of Statistics, Purdue University

September 11, 2007

Abstract

When the true mixing distribution is known to be continuous, the nonparametric maximum likelihood estimate of the mixing distribution cannot provide a satisfying answer due to its degeneracy. The estimation of mixing densities is an 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 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 and the kernel method. Some theoretical properties of the NPMPLE and the algorithm are also given in the article.

Key words: Mixture model; Mixing density; Nonparametric maximum penalized likelihood estimate; Functional EM algorithm; 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,b) 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 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 et al. (1985) recommended to start the EM algorithm from 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 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. In order 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 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 looks similarly to a kernel estimate having the form of $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)$ 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: $g \rightarrow Kg = \int f(y|x)g(x)\,dx$. 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 $w_\nu \rightarrow 0$ as $\nu \rightarrow \infty$; 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 singular functions 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; 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 et al. (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). In order 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 an 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 paper

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 several existing methods through simulation. Section 7 reports an application of the algorithm to a real problem. Some concluding remarks are given in the last section. The proofs of the propositions, theorems and corollaries are collected in the appendix.

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 $M_0 \le g_0(x) \le M_1$ for all $x \in [a, b]$. These assumptions are collectively labelled as Assumption (A0) and are assumed to hold throughout this article.

Any density g with support $[a, b]$ can be represented as

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

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 $y_1, y_2, \dots, y_n$ 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 section, we want to penalize the roughness of η using a penalty term. In this article, we choose the penalty

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

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

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

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 2nd order Sobolev space based on L_2-norm; see, e.g., 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

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

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 $\mathcal{Y} = \cup_{x \in \mathcal{X}}\{y : f(y|x) > 0\}$. In addition to Assumption

(A0) about g_0 stated at the beginning of this section, two more assumptions need to be imposed to ensure the existence of the NPMLE $\hat{\eta}(x)$, which are: (A1) For any given $y \in \mathcal{Y}$, $f(y|x)$ is a continuous function of x in $[a, b]$; and (A2) 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; Assumption (A2) is equivalent to requiring that the true mixture density $h(y|g_0)$ has an upper bound provided that Assumption (A0) holds. Both Assumptions (A1) and (A2) are commonly satisfied for popular component densities such as the normal density function $N(y-x, \sigma^2)$. Together with Assumption (A0), Assumptions (A1) and (A2) are 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,*

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

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 NPMLE $\hat{\eta}$.

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 since 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 $\{y_i\}_{1\leq i\leq n}$ from the mixture density $h(y|\eta_0)$ can be generated using the following two-step procedure: first a random sample denoted $\{x_i\}_{1\leq i\leq n}$ 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. $\{(y_i, x_i)\}_{1\leq i\leq n}$ 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

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

If η were a function depending on a finite dimensional parameter, with or without the penalty term, the classical EM algorithm (Dempster et al., 1977) would have started with an expectation step (E-step) involving the complete likelihood $l_{cp}(\eta)$, then proceed on to the maximization step (M-step) to calculate $\hat{\eta}$, and then repeat 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 $\vec{y} = (y_1, y_2, \cdots, y_n)$ 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

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

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, $l_p(\eta_{new}) \geq l_p(\eta_{cur})$.*

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

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

where

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

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

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

$\tilde{Q}$ 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 $\tilde{Q}(\eta|\eta_{cur})$ is equivalent to maximizing $Q(\eta|\eta_{cur})$. 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 $\tilde{Q}(\eta|\eta_{\mathrm{cur}})$ exists, is unique and satisfies an ordinary differential equation with some boundary conditions.

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

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

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

with boundary conditions

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

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 $\tilde{Q}(\eta_*|\eta_{cur}) \geq \tilde{Q}(\eta|\eta_{cur})$ for any $\eta \in \mathcal{H}$. Furthormore, 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 $\tilde{Q}(\eta|\eta_{\mathrm{cur}})$ in $\mathcal{H}$, which only guarantees that η_{new} and η'_{new} are absolutely continuous on $[a,b]$ and $\eta''_{\mathrm{new}} \in L_2(a,b)$. But this is not enough to derive equations (3.7) and (3.8), which include a 4th order differential equation with boundary conditions. In order to use the above equations, η_{new} needs to be smoother, for example, $\eta_{\mathrm{new}} \in C^4[a,b]$. 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 applying the results developed in Clarke and Vinter (1990). The smoothness of η_{new} depends on the smoothness of $\psi(\cdot|\vec{y},\eta_{\mathrm{cur}})$. If η_{cur} is smooth enough, then ψ is smooth enough to guarantee that the maximizer of $\tilde{Q}(\eta|\eta_{\mathrm{cur}})$ 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) till η_k converges.

Because of Assumption (A2), 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 (b). 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 (b) not only makes e^{η_k} a density so that the computed marginal density in 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 appendix for details.

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 et al. (1988).

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

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

and

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

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 $\{a_1, a_2, \ldots, a_{k+m}\} \cup \{c_{ij} : 1 \leq i \leq N-1, 1 \leq j \leq k\} \cup \{b_1, b_2, \ldots, b_{k+m}\}$ as the knot vector to construct B-spline functions. Here $a_1 < a_2 < \cdots < a_{k+m} \leq a$, $b \leq b_1 < b_2 < \cdots < b_{k+m}$, and $c_{ij} = x_i, 1 \leq i \leq N-1, 1 \leq j \leq k$. 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 are

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

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

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

In the system above, the coefficients are α_d, $1 \leq d \leq D$. Because the ODE (3.7) is nonlinear, the system (4.4) is also nonlinear. In the next section, 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 Gâteaux derivative, we derive the following approximations

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

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,

$$\left\{ \begin{array}{l} L[u](x) - e^{u(x)}z(x) - 2\lambda z^{(4)}(x) = 0, \quad \text{for } x \in \pi, \\ B[u] + (z''(a), z'''(a), z''(b), z'''(b)) = 0, \end{array} \right.$$

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

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

(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) till $\sup |z|$ is below some pre-specified threshold.

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, $V = \{1, 2, \dots, n\}$ 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 $\{x_i\}_{1 \le i \le n}$ is divided into K disjoint subsets of approximately the same size. Let V_k be the index set of the kth subset, $k = 1, \dots, K$, $\hat{g}_\lambda(x)$ 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*

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

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 $\text{pLS}(\lambda|g)$ and $\text{pKL}(\lambda|g)$ 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 $\text{pLS}(\lambda|g_0)$ (or $\text{pKL}(\lambda|g_0)$), 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

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

depending on which pseudo-CV score is used. $\text{pCV}(\lambda_2|\lambda_1)$ 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

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

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. Since 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 et al. (1996) commented that the largest local minimizer of $\text{CV}(h)$ in kernel density estimation setting usually gives better estimate than the global minimizer of $\text{CV}(h)$. This is analogous to our maximum smoothing principle. Hall and Marron (1991) observed that spurious local minima of the cross-validation function $\text{CV}(h)$ 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 Λ 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 $w(\lambda) = \arg\min_{\lambda_1 \in \Lambda} \text{pCV}(\lambda_1|\lambda)$ 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 performances 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 case of deconvolution; see Eggermont and LaRiccia (1997). In this section, we report simulation results for two deconvolution problems and one general mixture problem. The effectiveness of our smoothing parameter selection procedure is also demonstrated by a simulation example.

6.1 FEM vs. 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 $g(x)$. 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, which are

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

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

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

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

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

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

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

where φ is the density of the standard normal distribution $N(0, 1)$, $\beta(x; 2, 4)$ is the density of the beta distribution Beta$(2, 4)$, and $\phi_1(x)$ and $\phi_2(x)$ are the densities of normal distribution and double exponential distribution with mean 0 and standard deviation 0.05, respectively. All of the 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 ISE$(g, \hat{g}) = \int_a^b [g(x) - \hat{g}(x)]^2\, dx$, the integrated absolute error distance IAE$(g, \hat{g}) = \int_a^b |g(x) - \hat{g}(x)|\, dx$, and the Kullback-Leibler distance KLD$(g, \hat{g}) = \int_a^b \log\left(g(x)/\hat{g}(x)\right) g(x)\, dx$. In order 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 $S_\lambda = \{10^{-8} \times 2^{k/2}\}_{k=0}^{40}$, while 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 $\hat{g}_{\lambda,k}^{\text{FEM}}$ and $\hat{g}_{J,k}^{\text{EMS}}$, respectively.

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

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

(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, first quartile (Q1), median and third quartile (Q3).

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 two different sample sizes $(n = 400$ and $n = 1600)$, 100 replications $(N = 100)$, and all the three distance measures ($L =$ ISE, IAE or KLD). Therefore, there are in total 72 different scenarios. The results under these scenarios are reported in Tables 6.1 and 6.2 including means and standard deviations output from Step (d) of the above scheme. Clearly, the FEM algorithm outperforms the EMS algorithm uniformly across all the scenarios. In some cases, the improvements of FEM over EMS are fairly dramatic.

We have also plotted the density estimates generated by the FEM algorithm, the EMS algorithm as well as the kernel estimates, and compare them visually. Figures 6.1 and 6.2 are two sets of these plots, where solid lines represent the true densities, long-dashed lines represent the FEM estimates, dashed lines the EMS estimates and dotted lines the kernel estimates. The smoothing parameters in the kernel method are also the oracle ones. The overall impression is that the FEM estimates recover the true density much better than the other estimates and demonstrate particularly good behavior at the boundary points. The EMS estimates preserve more local properties of

Table 6.1: Deconvolution example 1: $\phi = \phi_1$ (normal noise).

		$n = 400$				$n = 1600$			
		Algorithm 1		EMS		Algorithm 1		EMS	
g	Dist.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
g_1	ISE	0.0113	0.0065	0.0177	0.0091	0.0049	0.0026	0.0070	0.0033
	IAE	0.0783	0.0249	0.1022	0.0264	0.0500	0.0146	0.0602	0.0158
	KLD	0.0061	0.0032	0.0093	0.0043	0.0026	0.0013	0.0037	0.0016
g_2	ISE	0.0240	0.0125	0.0273	0.0157	0.0083	0.0038	0.0081	0.0041
	IAE	0.1151	0.0299	0.1235	0.0336	0.0678	0.0154	0.0689	0.0164
	KLD	0.0127	0.0059	0.0150	0.0071	0.0044	0.0019	0.0049	0.0020
g_3	ISE	0.0264	0.0113	0.0268	0.0108	0.0096	0.0045	0.0114	0.0048
	IAE	0.1277	0.0293	0.1298	0.0293	0.0775	0.0178	0.0844	0.0187
	KLD	0.0135	0.0056	0.0140	0.0055	0.0051	0.0022	0.0061	0.0025
g_4	ISE	0.0044	0.0060	0.0262	0.0163	0.0010	0.0015	0.0121	0.0069
	IAE	0.0327	0.0264	0.0947	0.0264	0.0160	0.0123	0.0594	0.0139
	KLD	0.0015	0.0021	0.0108	0.0057	0.0003	0.0005	0.0038	0.0015
g_5	ISE	0.0049	0.0049	0.0165	0.0089	0.0014	0.0013	0.0040	0.0021
	IAE	0.0510	0.0267	0.1013	0.0269	0.0273	0.0135	0.0494	0.0136
	KLD	0.0024	0.0025	0.0083	0.0045	0.0007	0.0006	0.0020	0.0010
g_6	ISE	0.0147	0.0114	0.0215	0.0107	0.0042	0.0026	0.0083	0.0041
	IAE	0.0833	0.0272	0.1056	0.0267	0.0469	0.0143	0.0619	0.0139
	KLD	0.0062	0.0039	0.0104	0.0046	0.0020	0.0011	0.0037	0.0017

the likelihood function, and appear to be less smooth than the FEM estimates, especially around the boundary points. The kernel estimates are much bumpier, which is an indication of their susceptibility to noise.

6.2 FEM vs. EMS in non-deconvolution

We draw i.i.d. sample $Y_1, Y_2, \dots, Y_n$ 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 6.3 and the estimates generated by the algorithms were plotted in Figure 6.3. Both the numerical and visual comparisons show that the FEM algorithm outperforms the EMS algorithm.

Table 6.2: Deconvolution example 2: $\phi = \phi_2$ (double exponential noise).

		$n = 400$				$n = 1600$			
		Algorithm 1		EMS		Algorithm 1		EMS	
g	Dist.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
g_1	ISE	0.0111	0.0064	0.0180	0.0089	0.0047	0.0025	0.0069	0.0032
	IAE	0.0777	0.0247	0.1034	0.0257	0.0497	0.0148	0.0605	0.0157
	KLD	0.0060	0.0032	0.0095	0.0043	0.0026	0.0013	0.0037	0.0016
g_2	ISE	0.0233	0.0119	0.0276	0.0150	0.0076	0.0036	0.0085	0.0041
	IAE	0.1137	0.0294	0.1247	0.0328	0.0653	0.0152	0.0700	0.0159
	KLD	0.0125	0.0058	0.0154	0.0069	0.0043	0.0019	0.0052	0.0020
g_3	ISE	0.0249	0.0106	0.0263	0.0103	0.0091	0.0043	0.0110	0.0047
	IAE	0.1246	0.0282	0.1290	0.0275	0.0753	0.0177	0.0825	0.0183
	KLD	0.0129	0.0053	0.0137	0.0052	0.0048	0.0022	0.0058	0.0024
g_4	ISE	0.0043	0.0061	0.0259	0.0165	0.0010	0.0015	0.0115	0.0067
	IAE	0.0328	0.0263	0.0944	0.0268	0.0161	0.0123	0.0589	0.0136
	KLD	0.0015	0.0021	0.0109	0.0056	0.0003	0.0005	0.0037	0.0014
g_5	ISE	0.0049	0.0049	0.0169	0.0089	0.0014	0.0013	0.0041	0.0021
	IAE	0.0511	0.0266	0.1025	0.0261	0.0272	0.0134	0.0502	0.0130
	KLD	0.0024	0.0025	0.0085	0.0044	0.0007	0.0006	0.0020	0.0010
g_6	ISE	0.0147	0.0115	0.0215	0.0107	0.0043	0.0027	0.0083	0.0041
	IAE	0.0831	0.0273	0.1056	0.0264	0.0471	0.0141	0.0619	0.0138
	KLD	0.0062	0.0039	0.0103	0.0046	0.0020	0.0011	0.0037	0.0016

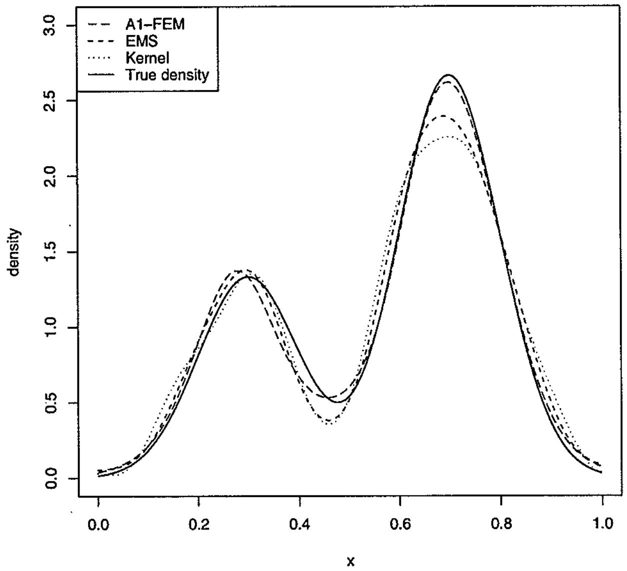

Figure 6.1: Solid line: true mixing density (g_2); long-dashed line: estimate by Algorithm 1; dashed line: estimate by the EMS algorithm; dotted line: estimate by the kernel method. All smoothing parameters are oracle ones.

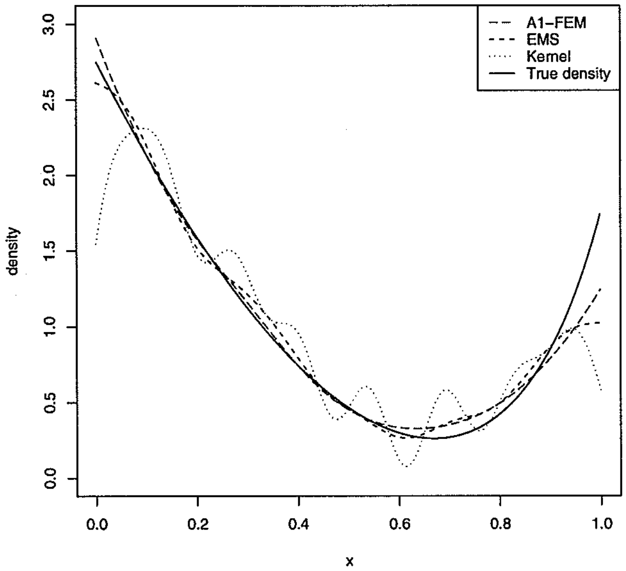

Figure 6.2: Solid line: true mixing density (g_6); long-dashed line: estimate by Algorithm 1; dashed line: estimate by the EMS algorithm; dotted line: estimate by the kernel method. All smoothing parameters are oracle ones.

Table 6.3: Non-deconvolution example: A gamma density as the component density.

		$n = 400$				$n = 1600$			
		Algorithm 1		EMS		Algorithm 1		EMS	
g	Dist.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
g_1	ISE	0.0110	0.0070	0.0168	0.0088	0.0044	0.0026	0.0056	0.0032
	IAE	0.0792	0.0262	0.1007	0.0262	0.0499	0.0156	0.0566	0.0173
	KLD	0.0059	0.0033	0.0087	0.0044	0.0025	0.0014	0.0029	0.0016
g_2	ISE	0.0457	0.0304	0.0567	0.0365	0.0191	0.0094	0.0230	0.0171
	IAE	0.1565	0.0516	0.1787	0.0580	0.1028	0.0258	0.1114	0.0405
	KLD	0.0234	0.0140	0.0317	0.0173	0.0102	0.0046	0.0130	0.0078
g_3	ISE	0.0790	0.0462	0.0690	0.0239	0.0337	0.0227	0.0411	0.0151
	IAE	0.2112	0.0616	0.2096	0.0402	0.1362	0.0434	0.1576	0.0293
	KLD	0.0388	0.0210	0.0342	0.0111	0.0172	0.0115	0.0210	0.0074
g_4	ISE	0.0050	0.0071	0.0221	0.0157	0.0010	0.0014	0.0096	0.0050
	IAE	0.0349	0.0281	0.0897	0.0302	0.0161	0.0127	0.0540	0.0144
	KLD	0.0017	0.0024	0.0102	0.0066	0.0004	0.0005	0.0034	0.0015
g_5	ISE	0.0056	0.0053	0.0194	0.0109	0.0016	0.0015	0.0042	0.0024
	IAE	0.0553	0.0270	0.1095	0.0322	0.0288	0.0142	0.0510	0.0146
	KLD	0.0027	0.0026	0.0097	0.0054	0.0008	0.0008	0.0021	0.0012
g_6	ISE	0.0209	0.0151	0.0308	0.0152	0.0064	0.0044	0.0154	0.0057
	IAE	0.0997	0.0345	0.1231	0.0295	0.0565	0.0164	0.0825	0.0172
	KLD	0.0095	0.0073	0.0159	0.0089	0.0031	0.0020	0.0077	0.0032

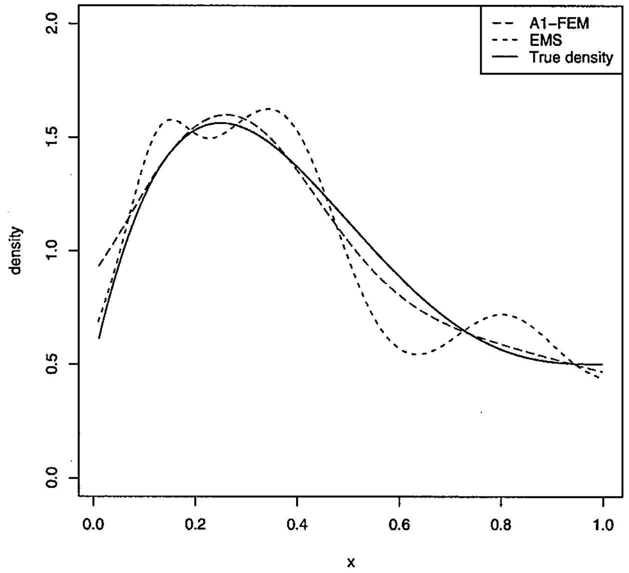

Figure 6.3: Solid line: true mixing density (g_1); long-dashed line: estimate by Algorithm 1; dashed line: estimate by the EMS algorithm. Smoothing parameters are oracle ones.

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 $\min_\lambda \mathrm{ISE}(\hat{g}_\lambda, g)$ with $\mathrm{ISE}(\hat{g}_{\lambda_*^{\mathrm{LS}}}, g)$ and $\min_\lambda \mathrm{KLD}(\hat{g}_\lambda, g)$ with $\mathrm{KLD}(\hat{g}_{\lambda_*^{\mathrm{KL}}}, g)$, 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 $\mathrm{pCV}_j(\lambda'|\lambda)$ 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 $\mathrm{pCV}_j(\lambda|\lambda) = \max_{\lambda' \in S_\lambda} \mathrm{pCV}_j(\lambda'|\lambda)$. Denote the result by λ_j^{LS} or λ_j^{KL} depending on whether pLS or pLS is used as the pCV score.

(e) Generate the scatter plots of $\mathrm{ISE}(\hat{g}_{\lambda_j^{\mathrm{LS}},j}, g)$ versus $\min_{\lambda \in S_\lambda} \mathrm{ISE}(\hat{g}_{\lambda,j}, g)$ and $\mathrm{KLD}(\hat{g}_{\lambda_j^{\mathrm{KL}},j}, g)$

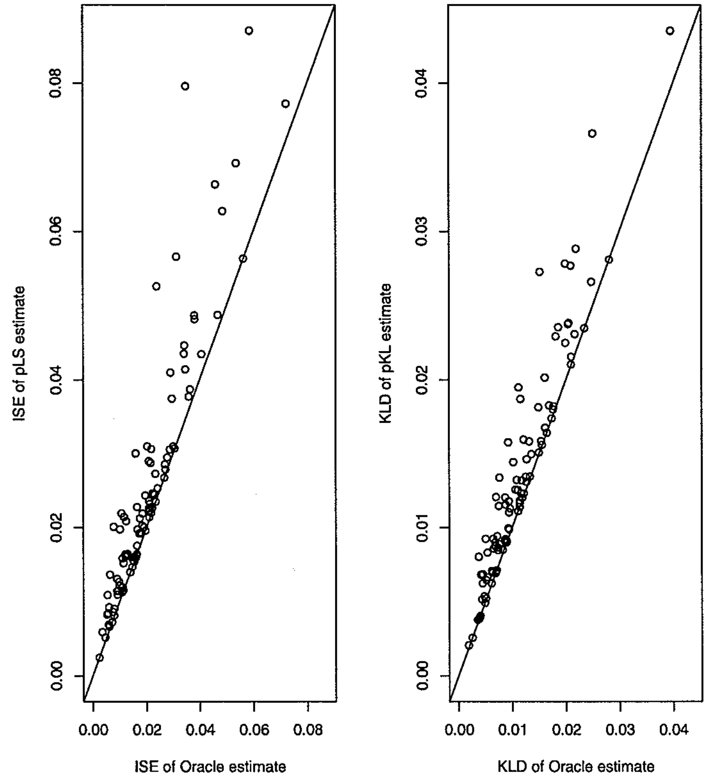

Figure 6.4: The left plot: ISE$(\hat{g}_{\lambda_*^{\text{LS}}}, g)$ vs. $\min_\lambda$ ISE$(\hat{g}_\lambda, g)$; the right plot: KLD$(\hat{g}_{\lambda_*^{\text{KL}}}, g)$ vs. $\min_\lambda$ KLD$(\hat{g}_\lambda, g)$. λ_*^{LS} and λ_*^{KL} are data-driven selected smoothing parameters based on pLS and pKL, respectively. The comparisons are based on g_2.

versus $\min_{\lambda \in S_\lambda}$ KLD$(\hat{g}_{\lambda,j}, g)$, separately, where $1 \leq j \leq N$.

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 Figures 6.4. In the left plot, the vertical axis represents ISE$(\hat{g}_{\lambda_j^{\text{LS}},j}, g)$ whereas the horizontal axis represents $\min_{\lambda \in S_\lambda}$ 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 KLD$(\hat{g}_{\lambda_j^{\text{KL}},j}, g)$ and $\min_{\lambda \in S_\lambda}$ 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 An Application in Stereology

We apply the FEM algorithm to solve the well-known "corpuscle problem" first discussed in Wicksell (1925). Suppose there is a three-dimensional specimen consisting of many small spheres embedded in a medium and we are interested in finding the distribution density of the radii of these spheres. However, we cannot measure the radii directly; instead, a very thin slice is taken through the specimen in a random direction. Examining the thin slice, we can observe a number of circles and measure their radii. The statistical problem here is to estimate the distribution density of radii of the three-dimensional spheres from observations generated by the density of the radii of their two-dimensional projections. Let r represent the radius of a sphere and y be the radius of a circle observed in the slice. Due to practical considerations, we usually set $\epsilon \leq r, y \leq R$ where $\epsilon \geq 0$ and R are known. The relationship between the density $h(y)$ of y and the density $g(r)$ of r is

$$h(y) = \frac{1}{\mu_\epsilon} \int_y^R \frac{y}{\sqrt{r^2 - y^2}} g(r)\, dr = \frac{1}{\mu_\epsilon} \int_\epsilon^R 1_{(y,R)}(r) \frac{y}{\sqrt{r^2 - y^2}} g(r)\, dr, \ \epsilon \leq y \leq R, \tag{7.1}$$

where $\mu_\epsilon = \int_\epsilon^R \sqrt{r^2 - \epsilon^2}\, g(r)\, dr$ and $1_{(y,R)}(r)$ is an indicator function. Given an i.i.d. random sample $y_1, y_2, \cdots, y_n$ from $h(y)$, we aim to estimate $g(r)$, the density of the radii. For a more detailed account of this problem, see Nychka et al. (1984), Wilson (1989), Silverman et al. (1990) and references therein.

In order to apply the FEM algorithm, we rewrite (7.1) as

$$h(y) = \int_\epsilon^R \left(\frac{y\, 1_{(\epsilon,r)}(y)}{\sqrt{r^2 - \epsilon^2}\sqrt{r^2 - y^2}} \right) \left(\frac{\sqrt{r^2 - \epsilon^2}\, g(r)}{\mu_\epsilon} \right) dr, \ \epsilon < y < R.$$

Let

$$f^*(y|r) = \frac{y\, 1_{(\epsilon,r)}(y)}{\sqrt{r^2 - \epsilon^2}\sqrt{r^2 - y^2}}, \quad g^*(r) = \frac{\sqrt{r^2 - \epsilon^2}\, g(r)}{\mu_\epsilon}.$$

Then it can be verified that $f^*(y|r)$ for any given $\epsilon < r < R$ and $g^*(r)$ are probability densities on the interval $[\epsilon, R]$. We treat f^* and g^* as the component and mixing densities of $h(y)$, respectively. Estimating $g^*(r)$ is equivalent to estimating $g(r)$ because there exists a one-to-one correspondence between them. Theoretically, if we can get an estimate $\hat{g}^*$ of g^*, then the estimate of g is

$$\hat{g}(r) = \frac{\hat{g}^*(r)/\sqrt{r^2 - \epsilon^2}}{\int_\epsilon^R \hat{g}^*(s)/\sqrt{s^2 - \epsilon^2}\, ds}. \tag{7.2}$$

However, as $r \to \epsilon+$, $1/\sqrt{r^2 - \epsilon^2} \to \infty$. Therefore, a small amount of error in estimating g^* near ϵ will result in a large amount of error in the estimation of g and the estimate will be numerically unstable near the left end point ϵ. For this reason, we adapt the FEM algorithm to estimate g directly, instead of estimating g^*.

Suppose $g(r) \propto e^{\eta(r)}$. We define a penalized likelihood functional of η,

$$\begin{aligned} l_p(\eta) = &\frac{1}{n} \sum_{i=1}^n \log \int_\epsilon^R f^*(y_i|r) \sqrt{r^2 - \epsilon^2}\, e^{\eta(r)}\, dr \\ &- \log \int_\epsilon^R \sqrt{r^2 - \epsilon^2}\, e^{\eta(r)}\, dr - \lambda \int_\epsilon^R [\eta''(r)]^2\, dr. \end{aligned} \tag{7.3}$$

The NPMPLE of η is $\arg\max_{\eta\in\mathcal{H}} l_p(\eta)$. The same procedure to derive the FEM algorithm can be used to derive an algorithm to compute the NPMLE of η, which is a variant of Algorithm 1. In the M-steps of the new algorithm, we need to find the maximizer of the following functional

$$\tilde{Q}(\eta|\eta_{\text{cur}}) = \int_{\epsilon}^{R} \eta(r)\psi(r|\vec{y},\eta_{\text{cur}})dr - \int_{\epsilon}^{R} \sqrt{r^2-\epsilon^2}\, e^{\eta(r)}dr - \lambda \int_{\epsilon}^{R} \left[\eta''(r)\right]^2 dr \tag{7.4}$$

where

$$\psi(r|\vec{y},\eta_{\text{cur}}) = \frac{1}{n}\sum_{i=1}^{n} \frac{f^*(y_i|r)\sqrt{r^2-\epsilon^2}\, e^{\eta(r)}}{\int_{\epsilon}^{R} f^*(y_i|s)\sqrt{s^2-\epsilon^2}\, e^{\eta(s)}\, ds} \tag{7.5}$$

It can be shown that the maximizer of $\tilde{Q}$ exists, is unique, and statisfies $\int_{\epsilon}^{R}\sqrt{r^2-\epsilon^2}\, e^{\hat{\eta}(r)}\, dr = 1$. Furthermore, the maximizer of $\tilde{Q}$ satisfies the ODE

$$\psi(r|\vec{y},\eta_{\text{cur}}) - \sqrt{r^2-\epsilon^2}\, e^{\eta(r)} - 2\lambda\, \eta^{(4)}(r) = 0, \tag{7.6}$$

with boundary conditions

$$\eta''(\epsilon) = \eta'''(\epsilon) = 0,\ \eta''(R) = \eta'''(R) = 0. \tag{7.7}$$

We summarize the steps of the new algorithm to compute the maximizer of (7.3) as follows and refer to the algorithm as Algorithm 1'.

Algorithm 1$'$

(a) Specify λ.

(b) Set $k = 0$ and $\eta_0(r) = \log \frac{1}{R-\epsilon}, r \in [\epsilon, R]$.

(c) Compute $\psi(r|\vec{y},\eta_k)$ as in (7.5). Numerically solve the ODE (7.6) with boundary conditions (7.7), and denote the solution η_{k+1}. Normalize the solution before proceeding to the next step,

$$\eta_{k+1}(r) \leftarrow \eta_{k+1}(r) - \log \int_{\epsilon}^{R} \sqrt{r^2-\epsilon^2}\, e^{\eta_{k+1}(r)}\, dr.$$

(d) $k \longrightarrow k+1$. Run step (c) until η_k converges. Let η_* be the final estimate of η. Then $\hat{g} = e^{\eta_*} / \int e^{\eta_*}$.

Silverman et al. (1990) and Wilson (1989) proposed to use the EMS algorithm to compute the density $g(r)$ of the sphere radii. We have conducted a simulation study to compare the performances of Algorithm 1$'$ and EMS. The results show that Algorithm 1$'$ outperforms EMS numerically and visually. In particular, the new algorithm works much better than EMS on the left boundary point. Due to limited space, the simulation results are not reported here.

8 Concluding Remarks

In this article, we have proposed the FEM algorithm to compute the 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 outperform many existing methods such as the EMS algorithm and the 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 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.

Appendix

The following lemma is needed in the proof of Theorem 1.

Lemma 1. *For any constants $C, B > 0$, define $S^{C,B} = \{\eta \in H : \eta(a) = 0, |\eta'(a)| \leq C, J(\eta) \leq B\}$. Then there exists an $\hat{\eta} \in S^{C,B}$ such that $l_p(\hat{\eta}) \geq l_p(\eta)$ for any $\eta \in S^{C,B}$.*

Proof. For any $\eta \in S^{C,B}$, because $\eta' \in W^{1,2}$, by Theorem A.6 of Braides (2002), without loss of generality, we can assume $\eta' \in C[a,b]$ and $\eta'(x) - \eta'(y) = \int_x^y \eta''(t)\,dt$ for all $x, y \in [a,b]$. Hence we have

$$|\eta'(x) - \eta'(y)| \leq \left(|x-y| \int_{\min(x,y)}^{\max(x,y)} [\eta''(t)]^2\,dt\right)^{1/2} \quad \text{for all} x, y \in [a,b] \tag{a-1}$$

by Cauchy inequality and

$$|\eta'(x)| \leq |\eta'(a)| + |\eta'(x) - \eta'(a)| \leq C + (b-a)^{1/2} B^{1/2} \quad \text{forall} x, y \in [a,b]. \tag{a-2}$$

Since $\eta(a) = 0, \eta(x) = \int_a^x \eta'(t)\,dt$, we have

$$|\eta(x)| \leq (C + (b-a)^{1/2} B^{1/2})(x-a) \leq (C + (b-a)^{1/2} B^{1/2})(b-a).$$

Therefore

$$\frac{e^{\eta(x)}}{\int_a^b e^{\eta(t)}\,dt} \leq \frac{e^{(C+(b-a)^{1/2}B^{1/2})(b-a)}}{e^{-(C+(b-a)^{1/2}B^{1/2})(b-a)}(b-a)} (= U(C,B)). \tag{a-3}$$

For all $\eta \in S^{C,B}$, we have $l_p(\eta) \leq l(\eta) \leq \log(MU(C,B)) < \infty$, hence $\sup_{\eta \in S^{C,B}} l_p(\eta) < \infty$. Therefore, there exist a sequence $\{\eta_k\} \in S^{C,B}$ such that $l_p(\eta_k) \to D = \sup_{\eta \in S^{C,B}} l_p(\eta)$ as $k \to \infty$. Without loss of generality, assume $\eta'_k(x) \in C[a,b]$ for all$k$. Since $\eta'_k(x)$'s satisfy the condition (a-2), $\{\eta'_k\}$'s are *equicontinuous* and *equibounded.* By the Arzelà-Ascoli theorem, there exists a subsequence $\{\eta'_{k_m}\}$ and a continuous function ξ_0 on $[a,b]$ such that $\eta'_{k_m}(x) \to \xi_0(x)$ uniformly as $m \to \infty$. Then $\eta_{k_m}(x) = \int_a^x \eta'_{k_m}(t)dt \to \int_a^x \xi_0(t)dt = \eta_0(x)$ uniformly. For convenience, we still use $\{\eta'_k\}$ to denote $\{\eta'_{k_m}\}$.

Because $\int_a^b \eta''_k(x)^2 dx \leq B$, that is, the L_2-norms of η''_k are *equibounded*, and $L_2(a,b)$ is a reflexive space, there exists a subsequence $\{\eta'_{k_m}\}$, such that η''_{k_m} *converges weakly in* $L_2(a,b)$ *to some function* $\nu \in L_2(a,b)$. Again for convenience, we still denote the subsequence by $\{\eta'_k\}$. The weak convergence of η'_k implies that

$$\int_a^b \eta''_k \phi dx \to \int_a^b \nu \phi dx \tag{a-4}$$

for all $\phi \in L_2(a,b)$. In particular, for any ϕ which is smooth and has compact support in (a,b), that is, $\phi \in C_c^\infty(a,b)$, we have

$$\int_a^b \eta''_k \phi dx = -\int_a^b \eta'_k \phi' dx.$$

Hence

$$\int_a^b \xi_0 \phi' dx = \lim_{k\to\infty} \int_a^b \eta'_k \phi' dx = -\lim_{k\to\infty} \int_a^b \eta''_k \phi dx = -\int_a^b \nu \phi dx, \quad \text{for all } \phi \in C_c^\infty(a,b).$$

The above equation can be interpreted as saying that ν is the *weak derivative* of $\xi_0 = \eta'_0$.

Since $b^2 - a^2 \geq 2a(b-a)$, we infer that

$$\liminf_{k\to\infty} \int_a^b [\eta''_k(x)]^2 dx - \int_a^b [\nu(x)]^2 dx \geq \liminf_{k\to\infty} \int_a^b 2\nu(\eta''_k - \nu) dx = 0$$

The last equality is due to equation (a-4).

Therefore

$$\begin{aligned}
\limsup_{k\to\infty} l_p(\eta_k) &= \int_a^b f(y_i|x) e^{\eta_0(x)} dx - \int_a^b e^{\eta_0(x)} dx - \liminf_{k\to\infty} \int_a^b [\eta''_k(x)]^2 dx \\
&\leq \int_a^b f(y_i|x) e^{\eta_0(x)} dx - \int_a^b e^{\eta_0(x)} dx - \int_a^b [\nu(x)]^2 dx \\
&= \int_a^b f(y_i|x) e^{\eta_0(x)} dx - \int_a^b e^{\eta_0(x)} dx - \int_a^b [\eta''_0(x)]^2 dx.
\end{aligned}$$

Let $\hat{\eta} = \eta_0$. Since $\eta_0(a) = 0, |\eta'_0(a)| = |\lim_k \eta'_k(a)| \leq C$ and $\int_a^b [\eta''_0(x)]^2\, dx \leq \liminf_{k\to\infty} \int_a^b [\eta''_k(x)]^2 dx$ $\leq B$, one has $\hat{\eta} = \eta_0 \in S^{C,B}$. Because $\lim_{k\to\infty} l_p(\eta_k) = \sup_{\eta \in S^{C,B}} l_p(\eta)$, one can conclude that

$l_p(\hat{\eta}) = \sup_{\eta \in S^{C,B}} l_p(\eta)$. Note that the proof follows the techniques in Buttazzo et al. (1998). □

Proof of Theorem 1

Proof. By Assumption A2, let $\eta \equiv 0$, we know that

$$l_p(\eta) = l(\eta) = \frac{1}{n} \sum_i \log \left(\frac{1}{b-a} \int_a^b f(y_i|x)\, dx \right)$$

exists. Therefore $\sup_{\eta \in \mathcal{H}} l_p(\eta) > -\infty$.

Let $\mathcal{H}_0 = \{\eta : \eta \in \mathcal{H}, \eta(a) = 0\}$. It is easy to see that $\sup_{\eta \in \mathcal{H}_0} l(\eta) = \sup_{\eta \in \mathcal{H}} l(\eta)$ and the existence of one side implies the existence of the other side. Let $\{\eta_k\} \subset \mathcal{H}_0$ be a sequence such that $l_p(\eta_k) \to \sup_{\eta \in \mathcal{H}_0} l_p(\eta)$. Let $C_k = |\eta_k'(a)|, B_k = \int_a^b [\eta'' k(x)]^2\, dx$.

If both C_k's and B_k's are bounded above, that is, there exist constants C and B such that $C_k \leq C$ and $B_k \leq B$, then by Lemma 1, there exists $\hat{\eta} \in S^{C,B}$ such that $l_p(\hat{\eta}) \geq l_p(\eta)$ for all $\eta \in S^{C,B}$. Since $\eta_k \in S^{C,B}$, $\sup_{\eta \in \mathcal{H}_0} l_p(\eta) = \lim_k l_p(\eta_k) \leq l_p(\hat{\eta})$, so $l_p(\hat{\eta}) = \sup_{\eta \in \mathcal{H}_0} l_p(\eta)$ and the theorem is proved. Therefore, it is sufficient to show that C_k and B_k are indeed bounded above.

We will prove the theorem in three steps. Step 1 shows that $C_k - (b-a)^{1/2} {B_k}^{1/2}$'s are bounded above; Step 2 shows that B_k's are bounded above; and Step 3 shows that C_k's are bounded above as well.

Step 1:

By (a-1), we have

$$\eta_k'(a) - (b-a)^{1/2} {B_k}^{1/2} \leq \eta_k'(x) \leq \eta_k'(a) + (b-a)^{1/2} {B_k}^{1/2}. \tag{a-5}$$

Assume that $\{C_k - (b-a)^{1/2} {B_k}^{1/2}\}$'s are not bounded above. Then we can find a subsequence of $\{\eta_k\}$, which is still denoted as $\{\eta_k\}$, such that one of the following statements is true: (a) $C_k - (b-a)^{1/2} {B_k}^{1/2} \to \infty$, which implies $C_k \to \infty$, and $\eta_k'(a) = C_k$; or (b) $C_k - (b-a)^{1/2} {B_k}^{1/2} \to \infty$ and $\eta_k'(a) = -C_k$.

Without loss of generality, we assume the statement (a) is true. Then

$$\eta_k'(x) \geq C_k - (b-a)^{1/2} {B_k}^{1/2} > 0 \text{ for large } k. \tag{a-6}$$

For any $\gamma > 0$, we can find $\Delta > 0$ such that $f(y_i|x) < f(y_i|b) + \gamma$, for all $x \in (b - \Delta, b)$, and $i =$

$1, 2, \ldots, n$. Since $\eta_k(x)$ is a increasing function for large enough k, for $x \in [a, b-\Delta]$,

$$\begin{aligned} \frac{e^{\eta_k(x)}}{\int_a^b e^{\eta_k(t)}\,dt} &\le \frac{e^{\eta_k(b-\Delta)}}{\int_{b-\Delta/2}^b e^{\eta_k(x)}\,dx} \le \frac{e^{\eta_k(b-\Delta)}}{e^{\eta_k(b-\Delta/2)}\Delta/2} \\ &= e^{\eta_k(b-\Delta)-\eta_k(b-\Delta/2)}\frac{1}{\Delta/2} = e^{-\eta_k'(b-\theta\Delta)\Delta/2}\frac{1}{\Delta/2} \quad \text{where } 1/2 < \theta < 1 \\ &\le e^{-\Delta/2(C_k-(b-a)^{1/2}{B_k}^{1/2})}\frac{1}{\Delta/2} \to 0 \text{ as } k \to \infty. \end{aligned}$$

Therefore, for large enough k and $x \in [a, b-\Delta]$,

$$\frac{e^{\eta_k(x)}}{\int_a^b e^{\eta_k(t)}\,dt} < \gamma.$$

Hence

$$\begin{aligned} \frac{\int_a^b f(y_i|x)e^{\eta_k(x)}dx}{\int_a^b e^{\eta_k(x)}dx} &= \frac{\int_a^{b-\Delta} f(y_i|x)e^{\eta_k(x)}dx + \int_{b-\Delta}^b f(y_i|x)e^{\eta_k(x)}dx}{\int_a^b e^{\eta_k(x)}dx} \\ &\le \gamma\int_a^{b-\Delta} f(y_i|x)\,dx + (f(y_i|b)+\gamma)\frac{\int_{b-\Delta}^b e^{\eta_k(x)}dx}{\int_a^b e^{\eta_k(x)}dx} \\ &\le M\gamma + f(y_i|b) + \gamma \end{aligned}$$

for $i = 1, 2, \ldots, n$ and

$$\limsup_{k\to\infty} l(\eta_k) \le \frac{1}{n}\sum_{i=1}^n \log\{f(y_i|b) + (1+M)\gamma\}.$$

0 Since γ can be any positive number, we have

$$\limsup_{k\to\infty} l(\eta_k) \le \frac{1}{n}\sum_{i=1}^n \log f(y_i|b).$$

Since η^* satisfies (2.7), we have $\limsup_k l(\eta_k) < l(\eta^*) = l_p(\eta^*)$, which implies

$$\sup_{\eta\in\mathcal{H}} l_p(\eta) = \sup_{\eta\in\mathcal{H}_0} l_p(\eta) = \lim_k l_p(\eta_k) \le \limsup_k l(\eta_k) < l_p(\eta^*).$$

Noting that $\eta^* \in N_J \subseteq \mathcal{H}$, the foregoing inequality contradicts the the definition of $\sup_{\eta\in\mathcal{H}} l_p(\eta)$. Thus the initial assumption must be false, and $C_k - (b-a)^{1/2}{B_k}^{1/2}$'s are bounded above.

Step 2:

By Assumption A2 and (a-3), $l(\eta_k) \le \log(MU(C_k, B_k)) = O(C_k + (b-a)^{1/2}{B_k}^{1/2})$. In Step 1 we have shown that $C_k - (b-a)^{1/2}{B_k}^{1/2}$'s are bounded above. Therefore, $O(C_k + (b-a)^{1/2}{B_k}^{1/2}) = O({B_k}^{1/2})$. If B_k's are not bounded above, then $l_p(\eta_k) = l(\eta_k) - \lambda J(\eta) \le O({B_k}^{1/2}) - \lambda B_k \to -\infty$

and $\sup_{\eta\in\mathcal{H}} l_p(\eta) = \lim l_p(\eta_k) = -\infty$. This is clearly a contradiction. Hence B_k's are bounded above.

Step 3:

Because both $C_k - (b-a)^{1/2}B_k{}^{1/2}$'s and B_k's are bounded above, C_k's are also bounded above.

We have shown that B_k's and C_k's are bounded above. Therefore, the theorem follows. □

Proof of Proposition 1

Proof. Let $g(x|\eta) = e^{\eta(x)}/\int_a^b e^{\eta(t)}\,dt$. Then the complete penalized log-likelihood functional $l_{cp}(\eta)$ can be rewritten as

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

Again let $\varphi(x|y,\eta) = f(y|x)e^{\eta(x)}/\int_a^b f(y|t)e^{\eta(t)}\,dt$. Because

$$f(y|x)g(x|\eta) = h(y|\eta)\varphi(x|y,\eta) = p(x,y|\eta)$$

where $p(x,y|\eta)$ is the joint probability density of (x,y) given η, we have

$$\begin{aligned} l_{cp}(\eta) &= \frac{1}{n}\sum_i \{\log h(y_i|\eta) + \log \varphi(x_i|y_i,\eta)\} - \lambda \int_a^b [\eta''(x)]^2\,dx \\ &= l_p(\eta) + \frac{1}{n}\sum_i \{\log \varphi(x_i|y_i,\eta)\}. \end{aligned}$$

Therefore,

$$Q(\eta|\eta_{\text{cur}}) = l_p(\eta) + \frac{1}{n}\sum_{i=1}^n \int_a^b \{\log \varphi(x_i|y_i,\eta)\}\varphi(x_i|y_i,\eta_{\text{cur}})\,dx_i = l_p(\eta) + H(\eta|\eta_{\text{cur}}).$$

By the definition of η_{new} and the Jason's inequality, we have $Q(\eta_{\text{new}}|\eta_{\text{cur}}) = l_p(\eta_{\text{new}}) + H(\eta_{\text{new}}|\eta_{\text{cur}}) \geq Q(\eta_{\text{cur}}|\eta_{\text{cur}}) = l_p(\eta_{\text{cur}}) + H(\eta_{\text{cur}}|\eta_{\text{cur}})$ and $H(\eta_{\text{new}}|\eta_{\text{cur}}) \leq H(\eta_{\text{cur}}|\eta_{\text{cur}})$. Therefore $l_p(\eta_{\text{new}}) \geq l_p(\eta_{\text{cur}})$. □

Proof of Proposition 2

Proof. Let $S = \{\eta \in \mathcal{H} : \int e^\eta = 1\}$. Let $\eta^* = \eta - \log\int e^\eta$. Then $\int e^{\eta^*} = 1$ and $\eta^* \in S$. It is clear that $\max_{\eta\in S}\tilde{Q}(\eta|\eta_{\text{cur}}) = \max_{\eta\in S} Q(\eta|\eta_{\text{cur}})$. Because $Q(\eta|\eta_{\text{cur}}) = Q(\eta^*|\eta_{\text{cur}})$ and $\eta^* \in S$, $\max_{\eta\in\mathcal{H}} Q(\eta|\eta_{\text{cur}}) = \max_{\eta\in S} Q(\eta|\eta_{\text{cur}})$. Because $\tilde{Q}(\eta|\eta_{\text{cur}}) = \tilde{Q}(\eta^*|\eta_{\text{cur}}) + 1 - \int e^\eta + \log(\int e^\eta) \leq \tilde{Q}(\eta^*|\eta_{\text{cur}})$ and the equality holds if and only if $\int e^\eta = 1$, that is, $\eta \in S$, we can obtain that

$\max_{\eta\in\mathcal{H}} \tilde{Q}(\eta|\eta_{\text{cur}}) = \max_{\eta\in S} \tilde{Q}(\eta|\eta_{\text{cur}})$. Therefore,

$$\max_{\eta\in\mathcal{H}} Q(\eta|\eta_{\text{cur}}) = \max_{\eta\in\mathcal{H}} \tilde{Q}(\eta|\eta_{\text{cur}}) \quad \text{and} \quad \arg\max_{\eta\in\mathcal{H}} \tilde{Q}(\eta|\eta_{\text{cur}}) \in S.$$

□

Sketch proof of Theorem 2

Proof. Theorem 4.1 of Gu and Qiu (1993) states: *Suppose $A(\eta)$ is a continuous and strictly concave functional in a Hilbert space $\mathcal{H} = \mathcal{H}_J \oplus \mathcal{N}_J$, where $\mathcal{H}_J$ has a square norm $J(\eta)$ and $\mathcal{N}_J$ is the null space of $J(\eta)$ of finite dimensions. If $A(\eta)$ has a maximizer in $\mathcal{N}_J$, then $A(\eta) - \lambda J(\eta)$ has a unique maximizer in $\mathcal{H}$ for any $\lambda > 0$.*

Let $A(\eta) = \int_a^b \eta(x)\psi(x|\vec{y},\eta_{\text{cur}})dx - \int_a^b e^{\eta(x)}dx$, $J(\eta) = \int_a^b [\eta''(x)]^2\,dx$, $\mathcal{H} = W^{2,2}(a,b)$, and $\mathcal{N}_J = \{cx + d : c, d \in R,\}$. We can prove that $A(\eta)$ is a continuous and strictly concave functional and $A(\eta)$ has a unique maximizer in $\mathcal{N}_J$. By Theorem 4.1 of Gu and Qiu (1993), Theorem 2 holds. □

Proof of Theorem 3

Proof. Define

$$\begin{aligned} A(t) &= \tilde{Q}(\eta(x) + t\epsilon(x)|\eta_{\text{cur}}) && \text{(a-7)} \\ &= \int_a^b [\eta(x) + t\epsilon(x)]\psi(x|\vec{y},\eta_{\text{cur}})\,dx - \int_a^b e^{\eta(x)+t\epsilon(x)}\,dx - \lambda \int_a^b [\eta''(x) + t\epsilon''(x)]^2\,dx. && \text{(a-8)} \end{aligned}$$

Under some mild conditions, we have

$$\begin{aligned} A'(t) = &\int_a^b \epsilon(x)\psi(x|\vec{y},\eta_{\text{cur}})\,dx \\ &- \int_a^b \epsilon(x)e^{\eta(x)+t\epsilon(x)}\,dx - \lambda \int_a^b 2\epsilon''(x)[\eta''(x) + t\epsilon''(x)]\,dx. && \text{(a-9)} \end{aligned}$$

and

$$A'(0) = \int_a^b \epsilon(x)\psi(x|\vec{y},\eta_{\text{cur}})\,dx - \int_a^b \epsilon(x)e^{\eta(x)}\,dx - 2\lambda \int_a^b \epsilon''(x)\eta''(x)\,dx. \qquad \text{(a-10)}$$

A necessary condition of η maximizing $\tilde{Q}(\eta|\eta_{\text{cur}})$ is $A'(0) = 0$ for any function ϵ.

Using intergration by parts, we have $\int_a^b \epsilon''\eta''\,dx = \epsilon'(b)\eta''(b) - \epsilon'(a)\eta''(a) - \epsilon(b)\eta'''(b) + \epsilon(a)\eta'''(a) + \int_a^b \epsilon\eta^{(4)}\,dx$. If ϵ is chosen to be a function such that $\epsilon(a) = \epsilon(b) = \epsilon'(a) = \epsilon'(b) = 0$, then

$$A'(0) = \int_a^b \epsilon(x)\left\{\psi(x|\vec{y},\eta_{\text{cur}}) - e^{\eta(x)} - 2\lambda\eta^{(4)}(x)\right\}\,dx. \qquad \text{(a-11)}$$

If η is a maximizer of $Q(\eta|\eta_{\text{cur}})$, then $A'(0) = 0$, which further implies

$$\psi(x|\vec{y}, \eta_{\text{cur}}) - e^{\eta(x)} - 2\lambda\eta^{(4)}(x) = 0, \text{for } x \in (a, b). \tag{a-12}$$

Under the above equation (a-12), we have $A'(0) = -2\lambda\{\epsilon'(b)\eta''(b) - \epsilon'(a)\eta''(a) - \epsilon(b)\eta'''(b) + \epsilon(a)\eta'''(a)\}$ for any ϵ. So $A'(0) = 0$ for allϵ implies

$$\eta''(a) = \eta'''(a) = 0,\ \eta''(b) = \eta'''(b) = 0 \tag{a-13}$$

In conclusion, if η maximizes $Q(\eta|\eta_{\text{cur}})$, it must satisfy (a-12) and (a-13), which are exactly (3.7) and (3.8). The theorem is proved. □

The proofs of Theorem 4 and Proposition 3 are omitted due to limited space; readers are referred to Liu (2005) for more details.

References

Abramovich, F. and B. W. Silverman (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika 85*, 115–129.

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

Ascher, U. M., R. M. Mattheij, and R. D. Russell (1988). *Numerical solution of boundary value problems for ordinary differential equations.* Englewood Cliffs, N.J.: Prentice Hall.

Braides, A. (2002). *Gamma-convergence for beginnersn.* New York : Oxford University Press.

Buttazzo, G., M. Giaquinta, and S. Hildebrandt (1998). *One-dimensional variational problems : an introduction.* Oxford : Clarendon Press.

Clarke, F. H. and R. B. Vinter (1990). A regularity theory for variational problems with higher order derivatives. *Transactions of the American Mathematical Society 320*, 227–251.

Cox, D. and F. O'Sullivan (1990). Asymptotic analysis of penalized likelihood and related estimators. *The Annals of Statistics 18*, 1676–1695.

De Montricher, G. M., R. A. Tapia, and J. R. Thompson (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *The Annals of Statistics 3*, 1329–1348.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Donoho, D. L. (1995). Nolinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harm. Anal. 2*, 101–126.

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

Eggermont, P. P. B. and V. N. LaRiccia (1995). Maximum smoothed likelihood density estimation for inverse problems. *The Annals of Statistics 23*, 199–220.

Eggermont, P. P. B. and V. N. LaRiccia (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 J.-Y. Koo (2002). Wavelet deconvolution. *IEEE Transactions on Information Theory 48*, 734–747.

Fan, J. and Y. K. Truong (1993). Nonparametric regression with errors in variables. *The Annals of Statistics 21*, 1900–1925.

G., L. B. and K. Roeder (1993). Uniqueness and identifiability in nonparametric mixtures. *Canadian Journal of Statistics 21*, 139–147.

Good, I. J. and R. A. Gaskins (1971). Nonparametric roughness penalties for probability densities. *Biometrika 58*, 255–277.

Good, I. J. and R. A. Gaskins (1980). Density estimation and bump hunting by penalized likelihood method exemplified by scattering and meteorite data (with discussions and rejoinder). *Journal of the American Statistical Association 75*, 42–73.

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, Series B 52*, 443–452.

Gu, C. (1992). Cross-validating non-gaussian data. *Journal of Computational and Graphical Statistics 1*, 169–179.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm. *Journal of the American Statistical Association 88*, 495–504.

Gu, C. and C. Qiu (1993). Smoothing spline density estimation: Theory. *The Annals of Statistics 21*, 217–234.

Hall, P. and J. S. Marron (1991). Local minima in cross-validation functions. *Journal of the Royal Statistical Society, Series 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 B. W. Silverman (1990). Speed of estimation in positron emission tomography and related inverse problems. *The Annals of Statistics 18*, 251–280.

Jones, M. C., J. S. Marron, and S. J. Sheather (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association 91*, 401–407.

Jones, M. C. and B. W. Silverman (1989). An orthogonal series density estimation approach to reconstructing positron emission tomography images. *Journal of Applied Statistics 16*, 177–191.

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 H.-Y. Chung (1998). Log-density estimation in linear inverse problems. *The Annals of Statistics 26*, 335–362.

Koo, J.-Y. and B. U. Park (1996). B-splines deconvolution based on the EM algorithm. *J. Statist. Comput. Simul. 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 T. A. Louis (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, Series B 40*, 113–132.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods: a general theory. *The Annals of Statistics 11*, 86–94.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods, part II: The exponential family. *The Annals of Statistics 11*, 783–792.

Lindsay, B. G. (1995). *Mixture models : theory, geometry, and applications.* Hayward, Calif. : Institute of Mathematical Statistics; Alexandria, Va. : American Statistical Association.

Lindsay, B. G. and M. L. Lesperance (1995). A review of semiparametric mixture models. *Journal of statistical planning and inference 47*, 29–39.

Liu, L. (2005). *On the estimation of mixing distributions: NPMLE and NPMPLE.* Ph. D. thesis, Department of Statistics, Purdue University.

Magder, L. S. and S. L. Zeger (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association 91*, 1141–1151.

Nychka, D., G. Wahba, T. D. Pugh, and S. Goldfarb (1984). Cross-validated spline methods for the estimation of three-dimensional tumor size distributions from observations on two-dimensional cross sections. *Journal of the American Statistical Association 79*, 832–846.

Pensky, M. and B. Vidakovic (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.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* London ; New York : Chapman and Hall.

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

Stefanski, L. and R. J. Carroll (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 J. R. Thompson (1978). *Nonparametric probability density estimation.* Johns Hopkins University Press, Baltimore, Maryland.

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

Vardi, Y., L. A. Shepp, and L. Kaufman (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.* Cambridge University Press.

Wand, M. P. (1998). Finite sample performance of deconvolving density estimators. *Statistics & Probability Letters 37*, 131–139.

Wicksell, S. D. (1925). The corpuscle problem: A mathematical study of a biometric problem. *Biometrika 17*, 84–99.

Wilson, J. D. (1989). A smoothed EM algorithm for the solution of Wicksell's corpuscle problem. *Journal of Statistical Computation and Simulation 31*, 195–221.

Zhang, C.-H. (1990). Fourier methods for estimating mixing densities and distributions. *The Annals of Statistics 18*, 806–831.