To appear in the *Journal of Nonparametric Statistics*
Vol. 00, No. 00, Month 20XX, 1–20

Theoretical Grounding for Estimation in Conditional Independence Multivariate Finite Mixture Models

Xiaotian Zhu[a*] and David R. Hunter[b†]

[a]*Data Science and Statistics, AbbVie, North Chicago, IL*; [b]*Statistics Department, Pennsylvania State University, University Park, PA*

(*Received 00 Month 20XX; accepted 00 Month 20XX*)

For the nonparametric estimation of multivariate finite mixture models with the conditional independence assumption, we propose a new formulation of the objective function in terms of penalized smoothed Kullback-Leibler distance. The nonlinearly smoothed majorization-minimization (NSMM) algorithm is derived from this perspective. An elegant representation of the NSMM algorithm is obtained using a novel projection-multiplication operator, a more precise monotonicity property of the algorithm is discovered, and the existence of a solution to the main optimization problem is proved for the first time.

Keywords: mixture model; penalized smoothed likelihood; majorization-minimization.

AMS Subject Classification: 62G05; 62H30

1. Introduction

In recent years, several studies have advanced the development of estimation algorithms, based on expectation-maximization (EM) and its generalization called majorization-minimization (MM), for nonparametric estimation for conditional independence multivariate finite mixture models. The idea for these algorithms had its genesis in the stochastic EM algorithm of Bordes et al. (2007) and was later extended to a deterministic algorithm by Benaglia et al. (2009) and Benaglia et al. (2011). These algorithms were placed on a more stable theoretical foundation due to the ascent property established by Levine et al. (2011). A detailed account of these algorithms, along with the related theory of parameter identifiability, is presented in the survey article by Chauveau et al. (2015). This paper follows up on this line of research, extending the theoretical foundations of this method and deriving novel results while also simplifying their formulation.

Conditional independence multivariate finite mixture models have fundamental importance in both statistical theory and applications; for example, as Chauveau et al. (2015) point out, these models are related to the random-effects models of Laird and Ware (1982). The basic setup assumes that r-dimensional vectors $\mathbf{X}_i = (X_{i,1}, X_{i,2}, ..., X_{i,r})^\top$, $1 \le i \le n$, are simple random samples from a finite mixture of $m > 1$ components with positive mixing proportions $\lambda_1, \lambda_2, ..., \lambda_m$ that sum to 1, and density functions $f_1, f_2, ..., f_m$. Here, we assume m is known. For recent work that addresses the esti-

* Email: xiaotian.zhu@abbvie.com

† Email: dhunter@stat.psu.edu

mation of m, along with a different approach to the estimation of the model parameters than the one outlined here, see Bonhomme et al. (2014) and Kasahara and Shimotsu (2014).

The conditional independence assumption, which arises naturally in analysis of data with repeated measurements, says each f_j, $1 \le j \le m$, is equivalent to the product of its marginal densities $f_{j,1}, f_{j,2}, ..., f_{j,r}$. Thus, the mixture density is

$$g(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j f_j(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{j,k}(x_k) \tag{1}$$

for any $\mathbf{x} = (x_1, ...x_r)^\top \in \mathbb{R}^r$. This is often regarded as a semi-parametric model with $\lambda_1, \ldots, \lambda_m$ being the Euclidean parameters and $f_{j,k}$, $1 \le j \le m$, $1 \le k \le r$ being the functional parameters. Let θ denote all of these parameters.

The identifiability of the parameters in the model (1) was not clear until the breakthrough in Hall and Zhou (2003) which established the identifiability when $m = 2$ and $r \ge 3$. Some follow-up work appeared, for example, Hall et al. (2005) and Kasahara and Shimotsu (2009), until the fundamental result that established generic identifiability of (1) for $r \ge 3$ was obtained (Allman et al. 2009) based on an algebraic result of Kruskal (1976, 1977).

Bordes et al. (2007) proposed a stochastic nonparametric EM algorithm (npEM) estimation algorithm for the estimation of semiparametric mixture models. Benaglia et al. (2009) and Benaglia et al. (2011) proposed a deterministic version of the algorithm for the estimation of (1) and studied bandwidth slection related to it. However, all these algorithms lack an objective function as well as the descent property which chracterizes any traditional EM algorithm (Dempster et al. 1977). A significant improvement comes from Levine et al. (2011), which proposes a smoothed likelihood as the objective function and leads to a smoothed version of the npEM that does possess the desired descent property. The authors point out the similarities between their approach and the one in Eggermont (1999) for non-mixtures. However, the constraints imposed by the condition that each f_{jk} must integrate to one lead to tricky optimization issues and necessitate a slightly awkward normalization step to satisfy these constraints. In reformulating the parameter space, the current paper removes the constraints and provides a rigorous justification for the algorithm, proving the existence of a solution to the main optimization problem for the first time. In addition, this paper sharpens the descent property by deriving a positive lower bound on the size of the decrease in the objective function at each iteration.

2. Reframing the Estimation Problem

In the following, we first consider an ideal setting where the target density is known (i.e., the sample size is infinity). Then we replace the target density by its empirical version and obtain the discrete algorithm.

2.1. *Setup and Notation*

Let $\mathbf{x} = (x_1, x_2, \cdots, x_r)^\top \in \mathbb{R}^r$ and let g denote a target density on $\mathbb{R}^r$, with support in the interior of Ω, where Ω is a compact and convex set in $\mathbb{R}^r$. Without loss of generality, assume Ω is the closed r-dimensional cube $[a, b]^r$. We are interested in the case when g

is a finite mixture of products of fully unspecified univariate measures, with unknown mixing parameters.

We make the following assumptions:

(i) Let the number of mixing components in g be fixed and denoted by m. There exist non-negative functions $e_j(\mathbf{x})$, $1 \le j \le m$, such that

$$g(\mathbf{x}) = \sum_{j=1}^{m} e_j(\mathbf{x}). \tag{2}$$

(ii) For each $1 \le j \le m$,

$$e_j(\mathbf{x}) = \theta_j \prod_{k=1}^{r} e_{j,k}(x_k), \tag{3}$$

where $\theta_j > 0$ and for each k, $1 \le k \le r$, $e_{j,k} \in L^1(\mathbb{R})$ is positive with support in $[a, b]$. Hence each $e_j(\mathbf{x})$ is in $L^1(\mathbb{R}^r)$, positive, and with support in Ω.

Given a bandwidth $h \in \mathbb{R}$, let $s_h(\cdot, \cdot) \in L^1(\mathbb{R} \times \mathbb{R})$ be nonnegative and with support in $[a, b] \times [a, b]$, such that

(iii) For $v, z \in \mathbb{R}$,

$$\int s_h(v, u)\, \mathrm{d}u = \int s_h(u, z)\, \mathrm{d}u = 1. \tag{4}$$

(iv) There exist positive numbers $M_1(h)$ and M_2 such that for any $v, z \in [a, b]$,

$$M_1(h) \le s_h(v, z) \le M_2. \tag{5}$$

(v) The function s_h has continuous first-order partial derivatives on $(a, b) \times (a, b)$ and there exists a constant B such that for any $u, x \in (a, b)$,

$$\left| \frac{\partial}{\partial v} s_h(v, z)|_{v=u} \right| \le B \qquad \text{and} \qquad \left| \frac{\partial}{\partial z} s_h(v, z)|_{z=x} \right| \le B. \tag{6}$$

(vi) If we define $f_j(\mathbf{x}) = e_j(\mathbf{x}) / \int e_j(\mathbf{z}) d\mathbf{z}$, then

$$f_j(\mathbf{x}) \ge (M_1(h))^r, \tag{7}$$

for all $\mathbf{x} \in \Omega$ and for each $j \in \{1, 2, ..., m\}$.

Before stating the optimization problem, we define the smoothing operators S_h, S_h^*, and $\mathcal{N}_h$, as follows.

For any $f \in L^1(\mathbb{R}^r)$, let

$$(S_h f)(\mathbf{x}) = \int \tilde{s}_h(\mathbf{x}, \mathbf{u}) f(\mathbf{u})\, \mathrm{d}\mathbf{u} \quad \text{and} \quad (S_h^* f)(\mathbf{x}) = \int \tilde{s}_h(\mathbf{u}, \mathbf{x}) f(\mathbf{u})\, \mathrm{d}\mathbf{u}, \tag{8}$$

where

$$\tilde{s}_h(\mathbf{x},\mathbf{u}) = \prod_{k=1}^{r} s_h(x_k,u_k) \qquad \text{for } \mathbf{x},\mathbf{u} \in \mathbb{R}^r. \tag{9}$$

Furthermore, let

$$(\mathcal{N}_h f)(\mathbf{x}) = \begin{cases} \exp[(S_h^* \log f)(\mathbf{x})] & \text{for } \mathbf{x} \in \Omega, \\ 0 & \text{elsewhere.} \end{cases} \tag{10}$$

These smoothing operators are well-known and have many desirable properties (Eggermont 1999). For instance, Lemma 1.1 of Eggermont (1999) states that for any nonnegative functions g_1 and g_2 in $L^1(\mathbb{R}^r)$,

$$KL\left(S_h g_1, S_h g_2\right) \leq KL\left(g_1, g_2\right), \tag{11}$$

where KL is the Kullback-Leibler divergence defined by

$$KL\left(g_1, g_2\right) = \int \left[g_1 \log \frac{g_1}{g_2} + g_2 - g_1 \right]. \tag{12}$$

2.2. *Main Optimization Problem*

Now, we assume conditions (i) through (vi) and propose to estimate $\mathbf{e}$ by minimizing the function

$$l(\mathbf{e}) = \int g(\mathbf{x}) \log \left[g(\mathbf{x}) / \sum_{j=1}^{m} (\mathcal{N}_h e_j)(\mathbf{x}) \right] \mathrm{d}\mathbf{x} + \int \left[\sum_{j=1}^{m} e_j(\mathbf{x}) \right] \mathrm{d}\mathbf{x} \tag{13}$$

subject to these conditions. In fact the only assumptions that impose any constraints on $\mathbf{e}$ are (ii) and (vi). Minimization of $l(\mathbf{e})$ can be written equivalently as minimization of the penalized smoothed Kullback-Leibler divergence

$$KL\left(g, \sum_{j=1}^{m} (\mathcal{N}_h e_j) \right) + \int \left[\sum_{j=1}^{m} e_j - \sum_{j=1}^{m} (\mathcal{N}_h e_j) \right] (\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{14}$$

where in (14) the second term acts like a roughness penalty.

The discrete version of the optimization problem replaces $g(\mathbf{x})\,\mathrm{d}\mathbf{x}$ by $\mathrm{d}G_n(\mathbf{x})$, where G_n is the empirical distribution function of a random sample of size n, and in this case we minimize

$$l_{\text{discrete}}(\mathbf{e}) = -\frac{1}{n} \sum_{i=1}^{n} \log \sum_{j=1}^{m} (\mathcal{N}_h e_j)(\mathbf{x}_i) + \int \sum_{j=1}^{m} e_j(\mathbf{x}). \tag{15}$$

Although we do not constrain $\mathbf{e}$ to require that the sum of all e_i is a density as required by Equation (2), this property is guaranteed by the main optimization:

THEOREM 2.1 *Any solution* $\tilde{\mathbf{e}}$ *to (13) or (15) satisfies*

$$\int \sum_{j=1}^{m} \tilde{e}_j(\mathbf{x}) = 1. \tag{16}$$

Proof. For any fixed $\mathbf{e}$, differentiation shows that the function $l(\alpha \mathbf{e})$ is minimized at the unique value

$$\hat{\alpha} = 1 \Big/ \int \sum_{j=1}^{m} e_j(\mathbf{x}) \, . \tag{17}$$

Thus if $\mathbf{e}$ is a minimizer then Equation (16) must hold. ■

From (16), we see that for each $1 \leq j \leq m$, $\int \tilde{e}_j$ can be interpreted as the mixing weight corresponding to the jth mixture component.

3. The NSMM Algorithm

In this section, we derive an iterative algorithm, using majorization-minimization (Hunter and Lange 2004), to minimize Equation (13). The algorithm, which we refer to as the nonlinearly smoothed majorization-minorization (NSMM) algorithm, coincides with that of Levine et al. (2011), despite the different derivation.

3.1. *An MM Algorithm*

Given the current estimate $\mathbf{e}^{(0)}$ satisfying assumptions (ii) and (vi), let us define

$$w_j^{(0)}(\mathbf{x}) = \frac{(\mathcal{N}_h e_j^{(0)})(\mathbf{x})}{\sum_{j'=1}^{m} (\mathcal{N}_h e_{j'}^{(0)})(\mathbf{x})} \tag{18}$$

for $1 \leq j \leq m$, noting that $\sum_j w_j^{(0)}(\mathbf{x}) = 1$. The concavity of the logarithm function gives

$$\begin{aligned} & l(\mathbf{e}) - l(\mathbf{e}^{(0)}) \\ & \quad = -\int g(\mathbf{x}) \log \sum_{j=1}^{m} \frac{(\mathcal{N}_h e_j^{(0)})(\mathbf{x})}{\sum_{j'=1}^{m} (\mathcal{N}_h e_{j'}^{(0)})\mathbf{x}} \cdot \frac{(\mathcal{N}_h e_j)(\mathbf{x})}{(\mathcal{N}_h e_j^{(0)})(\mathbf{x})} \, \mathrm{d}\mathbf{x} + \int \left(\sum_{j=1}^{m} e_j - \sum_{j=1}^{m} e_j^{(0)} \right) \\ & \quad \leq -\int g(\mathbf{x}) \sum_{j=1}^{m} \frac{(\mathcal{N}_h e_j^{(0)})(\mathbf{x})}{\sum_{j'=1}^{m} (\mathcal{N}_h e_{j'}^{(0)})(\mathbf{x})} \cdot \log \frac{(\mathcal{N}_h e_j)(\mathbf{x})}{(\mathcal{N}_h e_j^{(0)})(\mathbf{x})} \, \mathrm{d}\mathbf{x} + \int \left(\sum_{j=1}^{m} e_j - \sum_{j=1}^{m} e_j^{(0)} \right). \end{aligned} \tag{19}$$

So if we let

$$b^{(0)}(\mathbf{e}) = -\int g(\mathbf{x}) \sum_{j=1}^{m} w_j^{(0)}(\mathbf{x}) \cdot \log\left(\mathcal{N}_h e_j\right)(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int \left(\sum_{j=1}^{m} e_j\right), \tag{20}$$

we obtain

$$l(\mathbf{e}) - l(\mathbf{e}^{(0)}) \le b^{(0)}(\mathbf{e}) - b^{(0)}(\mathbf{e}^{(0)}). \tag{21}$$

Using the MM algorithm terminology of Hunter and Lange (2004), Inequality (21) means that $b^{(0)}$ may be said to majorize l at $\mathbf{e}^{(0)}$, up to an additive constant. Minimizing $b^{(0)}$ therefore yields a function $\mathbf{e}^{(1)}$ satisfying

$$l(\mathbf{e}^{(1)}) \le l(\mathbf{e}^{(0)}). \tag{22}$$

Thus, we now consider how to minimize $b^{(0)}(\mathbf{e})$, subject to the assumptions on $\mathbf{e}$ that were stated at the beginning. This is to be done component-wise. That is, for each j, we wish to minimize

$$\begin{aligned} b_j^{(0)}(\mathbf{e}) &= -\int g(\mathbf{x}) w_j^{(0)}(\mathbf{x}) \cdot \int \tilde{s}_h(\mathbf{u},\mathbf{x}) \log e_j(\mathbf{u})\, \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{x} + \int e_j \\ &= -\iint g(\mathbf{x}) w_j^{(0)}(\mathbf{x}) \cdot \tilde{s}_h(\mathbf{u},\mathbf{x}) \left[\sum_{k=1}^{r} \log e_{j,k}(u_k) + \log\theta_j\right] \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{x} \\ &\qquad + \int \theta_j \prod_{k=1}^{r} e_{j,k}(u_k)\, \mathrm{d}\mathbf{u}. \end{aligned} \tag{23}$$

Up to an additive term that does not involve any $e_{j,k}$, Expression (23) is

$$-\sum_{k=1}^{r} \iint g(\mathbf{x}) w_j^{(0)}(\mathbf{x}) \cdot s_h(u_k, x_k) \log e_{j,k}(u_k)\, \mathrm{d}u_k\ \mathrm{d}\mathbf{x} + \int \theta_j \prod_{k=1}^{r} e_{j,k}(u_k)\, \mathrm{d}\mathbf{u}. \tag{24}$$

For any k in $1,\ldots,r$, we can view Expression (24) as an integral with respect to $\mathrm{d}u_k$. Differentiating the integrand with respect to $e_{j,k}(u_k)$ and equating the result to zero, Fubini's Theorem gives

$$\hat{e}_{j,k}(u_k) \propto \int g(\mathbf{x}) w_j^{(0)}(\mathbf{x}) \cdot s_h(u_k, x_k)\ \mathrm{d}\mathbf{x}. \tag{25}$$

This tells us, according to (3), that

$$\hat{e}_j(\mathbf{u}) = \alpha_j \prod_{k=1}^{r} \int g(\mathbf{x}) w_j^{(0)}(\mathbf{x}) \cdot s_h(u_k, x_k)\ \mathrm{d}\mathbf{x} \tag{26}$$

for some constant α_j. To find α_j, we plug (26) into (23) and differentiate with respect to α_j, which gives as a final result

$$\hat{e}_j(\mathbf{u}) = \frac{\prod_{k=1}^{r} \int g(\mathbf{x}) w_j^{(0)}(\mathbf{x}) \cdot s_h(u_k, x_k)\ \mathrm{d}\mathbf{x}}{\left[\int g(\mathbf{x}) w_j^{(0)}(\mathbf{x})\ \mathrm{d}\mathbf{x}\right]^{r-1}}. \tag{27}$$

To summarize, our NSMM algorithm starts with some initial estimate $e^{(0)}$ satisfying assumptions (ii) and (vi), then iterates according to

$$e^{(p+1)}(\mathbf{u}) = G(e^{(p)})(\mathbf{u}), \tag{28}$$

where $G(\cdot)$ performs the one-step update of Equation (27). In practical terms, NSMM is identical to the non-parametric maximum smoothed likelihood algorithm proposed in Levine et al. (2011). However, our derivation uses a simpler parameter space and the normalization involved in each step of the algorithm is now a result of optimization. We have thus rigorously derived the NSMM algorithm as a special case of the majorization-minimization method.

In the discrete case, we replace the density $g(\cdot)$ by the empirical distribution defined by the sample; thus, the algorithm iterates according to the following until convergence, assuming $e^{(p)}$ is the current step estimate:

Majorization Step: For $1 \leq i \leq n$, $1 \leq j \leq m$, compute

$$w_j^{(p)}(\mathbf{x}_i) = \frac{(\mathcal{N}_h e_j^{(p)})(\mathbf{x}_i)}{\sum_{j=1}^{m} (\mathcal{N}_h e_j^{(p)})(\mathbf{x}_i)}. \tag{29}$$

Minimization Step: Let

$$e_j^{(p+1)}(\mathbf{u}) = \frac{\prod_{k=1}^{r} \sum_{i=1}^{n} \frac{1}{n} w_j^{(p)}(\mathbf{x}_i) s_h(u_k, x_{ik})}{\left(\sum_{i=1}^{n} \frac{1}{n} w_j^{(p)}(\mathbf{x}_i)\right)^{r-1}}. \tag{30}$$

3.2. *The Projection-Multiplication Operator*

The NSMM algorithm of Section 3.1 can be summarized in an elegant way using the projection-multiplication operator, defined as follows. For any nonnegative function f on $\mathbb{R}^r$ such that $\int f > 0$, and $x = (x_1, x_2, \cdots, x_r)^\top \in \mathbb{R}^r$, let the operator P, which factorizes f as a product of marginal functions on $\mathbb{R}^r$, be defined by

$$(Pf)(\mathbf{x}) = \frac{\left[\prod_{k=1}^{r} \int_{R^{r-1}} f(\mathbf{x})\, \mathrm{d}x_1\, \mathrm{d}x_2 \cdots \mathrm{d}x_{k-1}\, \mathrm{d}x_{k+1} \cdots \mathrm{d}x_r\right]}{\left[\int f\right]^{(r-1)}}. \tag{31}$$

When f is a density on $\mathbb{R}^r$, the right side of (31) simplifies because the denominator is 1. As the next lemma points out, the P operator commutes with the S_h Operator.

Lemma 3.1 *Assume f is an integrable nonnegative function on $\mathbb{R}^r$ with support in a compact set Ω. We have*

$$(P \circ S_h)f = (S_h \circ P)f. \tag{32}$$

Proof. See Appendix (A.1). ∎

Lemma 3.1 implies that $G(\cdot)$, which performs the one-step update of the NSMM algorithm, can be expressed concisely as

$$\left(G(e^{(p)})\right)_j (\mathbf{u}) = \left[P \circ S_h(g \cdot w_j^{(p)})\right](\mathbf{u}) \tag{33}$$

for $1 \leq j \leq m$. In the discrete or finite-sample case, $g(\cdot)$ places weight $1/n$ at each sampled point. Equation (33) therefore suggests a geometric intuition of $G(\cdot)$ in the discrete case, which is illustrated in Figure 1.

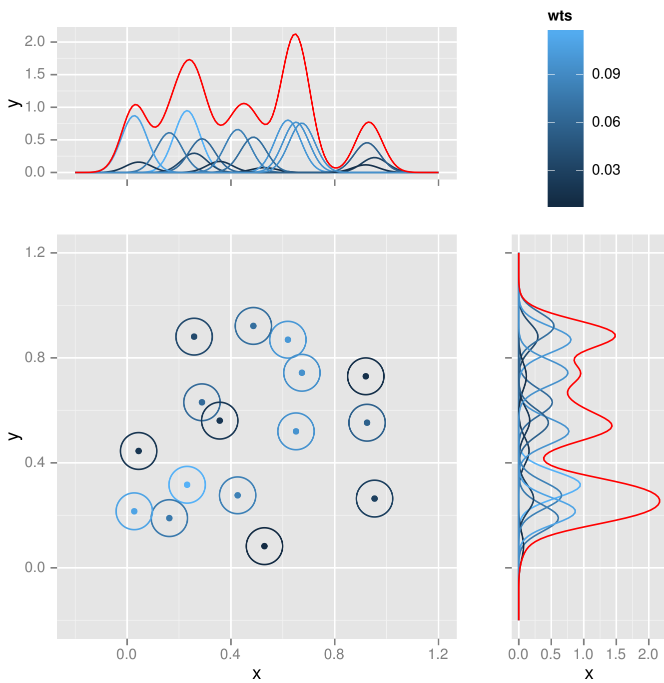

Figure 1. Illustration of the $G(\cdot)$ operator for a finite ($n = 16$) sample in the case $r = 2$: The operator first smoothes the weighted dataset and then applies the P operator to it, yielding the product of the smoothed marginals, shown here in red, as the density estimator at the next iteration.

3.3. *Sharpened Monotonicity*

For any MM algorithm, including any EM algorithm (Dempster et al. 1977), the well-known monotonicity property of Inequality (22) says that the value of the objective function moves, at each iteration, toward the direction of being optimized (Hunter and Lange 2004). For the NSMM algorithm, this descent property was first proved in Levine et al. (2011). In Proposition 3.2, we present a novel result that strengthens Inequality (22) by giving an explicit formula for the nonnegative value $l(\mathbf{e}^{(p)}) - l(\mathbf{e}^{(p+1)})$.

Proposition 3.2 *In the continuous (infinite-sample) version of the NSMM algorithm, at any step p, we have*

$$l(\mathbf{e}^{(p)}) - l(\mathbf{e}^{(p+1)}) = \sum_{j=1}^{m} KL(e_j^{(p+1)}, e_j^{(p)}) + \sum_{j=1}^{m} KL(g \cdot w_j^{(p)}, g \cdot w_j^{(p+1)}). \tag{34}$$

Proof. See Appendix (A.2). ■

Remark 1 The discrete version of Proposition 3.2 is

$$l(\mathbf{e}^{(p)}) - l(\mathbf{e}^{(p+1)}) = \sum_{j=1}^{m} KL(e_j^{(p+1)}, e_j^{(p)}) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} w_j^{(p)}(\mathbf{x}_i) \log \frac{w_j^{(p)}(\mathbf{x}_i)}{w_j^{(p+1)}(\mathbf{x}_i)}. \tag{35}$$

Proposition 3.2 implies the following corollary:

Corollary 3.3 *In the NSMM algorithm, at any step p, we have*

$$l(\mathbf{e}^{(p)}) - l(\mathbf{e}^{(p+1)}) \geq \sum_{j=1}^{m} KL(e_j^{(p+1)}, e_j^{(p)}). \tag{36}$$

Inequality (36) may be established directly, using Jensen's Inequality, and we include this proof separately as an appendix because it is interesting in its own right.

Proof. Direct proof of Corollary 3.3 can be found in Appendix (A.3). ■

Corollary 3.3 implies the following two novel results. First, Corollary 3.4 guarantees that we only need to search among fixed point(s) of the NSMM algorithm for a solution to the minimization problem. This gives a theoretical basis for using the NSMM algorithm for this estimation problem.

Corollary 3.4 *Any minimizer* $\mathbf{e}$ *of* $l(\mathbf{e})$ *or* $l_{discrete}(\mathbf{e})$ *is a fixed point of the corresponding NSMM algorithm.*

Proof. Since the right side of (36) is strictly positive when $e_j^{(p+1)} \neq e_j^{(p)}$ for any j, a necessary condition for $\mathbf{e}^{(p)}$ to minimize $l(\mathbf{e})$ is that $\mathbf{e}^{(p+1)} = \mathbf{e}^{(p)}$, i.e., that $\mathbf{e}^{(p)}$ is a fixed point of the algorithm. The same reasoning works for $l_{\text{discrete}}(\mathbf{e})$. ■

Second, Corollary 3.5 ensures among other things that the L^1 distance between estimates of adjacent steps from an NSMM sequence will tend to zero, a result that is used in the next section.

Corollary 3.5 *In the NSMM algorithm, at any step p, we have*

$$l(\mathbf{e}^{(p)}) - l(\mathbf{e}^{(p+1)}) \geq \sum_{j=1}^{m} \frac{1}{4} \left\| e_j^{(p+1)} - e_j^{(p)} \right\|_1^2, \tag{37}$$

where $\|\cdot\|_1$ *denotes the* L^1 *norm.*

Proof. The result follows from Inequality (3.21) in Eggermont and LaRiccia (2001), which states that

$$KL(g_1, g_2) \geq \frac{1}{4}\|g_1, g_2\|_1^2 \tag{38}$$

for functions g_1 and g_2. ■

4. Existence of a Solution to the Maximization Problem

In this section, we verify the existence of at least one solution to the main optimization problem of Section 2.2, a novel result as far as we are aware.

LEMMA 4.1 *Given* $\mathbf{e}$ *satisfying assumption (ii), we have* $l(\mathbf{e}) \geq 1$*. In the discrete case, we have* $l_{discrete}(\mathbf{e}) \geq -\log M_2$*.*

Proof. See Appendix (A.4). ■

Together, Lemma 3.3 and Lemma 4.1 imply the following corollary.

COROLLARY 4.2 *In the NSMM algorithm,* $l(\mathbf{e}^{(p)})$ *will tend to a finite limit as* p *goes to infinity. This result also holds in the discrete case.*

We now establish some technical results that lead to the main conclusion of this section, namely, the existence of a minimizer of both $l(\mathbf{e})$ and $l_{\text{discrete}}(\mathbf{e})$.

LEMMA 4.3 *Assume conditions (i) through (vi). For each* j*,* $1 \leq j \leq m$*, any NSMM sequence* $\{e_j^{(p)}\}_{1\leq p<\infty}$ *is uniformly bounded and equicontinuous on* Ω*. This result also holds in the discrete case.*

Proof. See Appendix (A.5). ■

More generally, Lemma 4.3 implies the following result:

LEMMA 4.4 *For* $\mathbf{e}$ *satisfying assumptions (i) through (vi), in either the discrete or the continuous case, for* $1 \leq j \leq m$ *and* $\mathbf{u}$*,* $\mathbf{v} \in \Omega$*, we have*

$$(G(\mathbf{e}))_j(\mathbf{u}) \leq M_2^r, \tag{39}$$

$$|(G(\mathbf{e}))_j(\mathbf{u}) - (G(\mathbf{e}))_j(\mathbf{v})| \leq [B \cdot M_2^{r-1}] \cdot \|\mathbf{u} - \mathbf{v}\|_1. \tag{40}$$

The following lemma establishes a sort of lower semi-continuity of the functional $l(\cdot)$, which will be needed in proving existence of at least one solution to the main optimization problem.

LEMMA 4.5 *Let* $\gamma_j^{(p)} \in L^1(\mathbb{R}^r)$ *be nonnegative and with support in* Ω *for each* p *and* j*, where* $0 \leq p \leq \infty$ *and* $1 \leq j \leq m$*. Assume each* $\gamma_j^{(p)}$ *uniformly converges to* $\gamma_j^{(\infty)}$ *in* $L^1(\mathbb{R}^r)$ *and that all* $\gamma_j^{(p)}$ *are bounded from above by a constant* $Q > 1$*. Let* $\boldsymbol{\gamma}^{(p)}$ *and* $\boldsymbol{\gamma}^{(\infty)}$ *represent* $(\gamma_1^{(p)}, \cdots, \gamma_m^{(p)})$ *and* $(\gamma_1^{(\infty)}, \cdots, \gamma_m^{(\infty)})$*, respectively. Then we have*

$$l(\boldsymbol{\gamma}^{(\infty)}) \leq \liminf_{p\to\infty} l(\boldsymbol{\gamma}^{(p)}). \tag{41}$$

This is also true for the discrete case.

Proof. See Appendix (A.6). ∎

THEOREM 4.6 *Under assumptions (i) through (vi), there exists at least one solution to the main optimization problem (13). This is also true in the discrete case.*

Proof. See Appendix (A.7). ∎

To conclude this section, we discuss the rationale behind assumption (vi) and related issues such as why the $\mathcal{N}_h$ operator is well-defined as we applied it.

LEMMA 4.7 *In an NSMM sequence* $\{\mathbf{e}^{(p)}\}_{0\le p\le\infty}$, *the* $e_j^{(p)}$ *are all strictly positive for all* j. *Moreover, if we let* $\lambda_j^{(p)} = \int e_j^{(p)}$ *and* $f_j^{(p)}(\mathbf{u}) = e_j^{(p)}(\mathbf{u})/\lambda_j^{(p)}$, *then*

$$(M_1(h))^r \le f_j^{(p)}(\mathbf{u}) \le M_2^r \tag{42}$$

for all $\mathbf{u}\in\Omega$, $p>0$, *and* $1\le j\le m$.

Proof. See Appendix (A.8). ∎

Lemma (4.7) shows why in assumption (vi) we require the marginal densities of each mixture component to be bounded below by $(M_1(h))^r$ and guarantees that dividing by zero never occurs in any NSMM sequence.

5. Discussion

Starting from the conditional independence finite multivariate mixture model as set forth in the work of Benaglia et al. (2009) and Levine et al. (2011), this manuscript proposes an equivalent but simplified parameterization. This reformulation leads to a novel and mathematically coherent version of the penalized Kullback-Leibler divergence as the main optimization criterion for the estimation of the parameters.

In this new framework, certain constraints that were previously imposed on the parameter space may be eliminated, and the solutions obtained may be shown to follow these constraints naturally. These contributions help to rigorously justify the nonparametric maximum smoothed likelihood (npMSL) estimation algorithm established by Levine et al. (2011).

As part of our investigation, we have discovered several new results, including a sharper monotonicity property of the NSMM algorithm that could ultimately contribute to future investigations of the true convergence rate or other asymptotic properties of the algorithm. We also prove, for the first time, the existence of at least one solution for the estimation problem of this model.

Because of the elegant simplicity and mathematical tractability associated with this framework, we believe the results herein will serve as the basis for future research on this useful nonparametric model.

Appendix A. Mathematical Proofs

A.1. *Proof of Lemma 3.1*

Proof. Since $(P \circ S_h)$ is linear, we only need to consider the case where f is a density function. By Fubini's Theorem and Equation (31),

$$
\begin{aligned}
&[(P \circ S_h)f](\mathbf{x}) \\
&= \prod_{k=1}^{r} \int_{R^{r-1}} \left(\int_{\mathbb{R}^r} \tilde{s}_h(\mathbf{x}, \mathbf{u}) f(\mathbf{u}) \, \mathrm{d}\mathbf{u} \right) \mathrm{d}x_1 \, \mathrm{d}x_2 \cdots \mathrm{d}x_{k-1} \, \mathrm{d}x_{k+1} \cdots \mathrm{d}x_r \\
&= \prod_{k=1}^{r} \int_{R^r} \left(\int_{R^{r-1}} \tilde{s}_h(\mathbf{x}, \mathbf{u}) f(\mathbf{u}) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \cdots \mathrm{d}x_{k-1} \, \mathrm{d}x_{k+1} \cdots \mathrm{d}x_r \right) \mathrm{d}\mathbf{u} \\
&= \prod_{k=1}^{r} \int_{R} s_h(x_k, u_k) \left(\int_{R^{r-1}} f(\mathbf{u}) \, \mathrm{d}u_1 \, \mathrm{d}u_2 \cdots \mathrm{d}u_{k-1} \, \mathrm{d}u_{k+1} \cdots \mathrm{d}u_r \right) \mathrm{d}u_k \\
&= \int_{R^r} \left(\prod_{k=1}^{r} s_h(x_k, u_k) \right) \cdot \prod_{k=1}^{r} \left(\int_{R^{r-1}} f(\mathbf{u}) \, \mathrm{d}u_1 \, \mathrm{d}u_2 \cdots \mathrm{d}u_{k-1} \, \mathrm{d}u_{k+1} \cdots \mathrm{d}u_r \right) \mathrm{d}\mathbf{u} \\
&= [(S_h \circ P)f](\mathbf{x}).
\end{aligned}
$$

■

A.2. *Proof of Propositon 3.2*

Proof. Direct evaluation and the definition of Kullback-Leibler divergence in Equation (12) give

$$
\begin{aligned}
&l(\mathbf{e}^{(p)}) - l(\mathbf{e}^{(p+1)}) \\
&= \int g(\mathbf{x}) \log \frac{\sum_{c=1}^{m} (\mathcal{N}_h e_c^{(p+1)})(\mathbf{x})}{\sum_{d=1}^{m} (\mathcal{N}_h e_d^{(p)})(\mathbf{x})} \, \mathrm{d}\mathbf{x} = \sum_{j=1}^{m} \int g(\mathbf{x}) w_j^{(p)}(\mathbf{x}) \log \frac{\sum_{c=1}^{m} (\mathcal{N}_h e_c^{(p+1)})(\mathbf{x})}{\sum_{d=1}^{m} (\mathcal{N}_h e_d^{(p)})(\mathbf{x})} \, \mathrm{d}\mathbf{x} \\
&= \sum_{j=1}^{m} \int g(\mathbf{x}) w_j^{(p)}(\mathbf{x}) \log \frac{w_j^{(p)}(\mathbf{x})}{w_j^{(p+1)}(\mathbf{x})} \cdot \frac{(\mathcal{N}_h e_j^{(p+1)})(\mathbf{x})}{(\mathcal{N}_h e_j^{(p)})(\mathbf{x})} \, \mathrm{d}\mathbf{x} \\
&= \sum_{j=1}^{m} \int g(\mathbf{x}) w_j^{(p)}(\mathbf{x}) \log \frac{(\mathcal{N}_h e_j^{(p+1)})(\mathbf{x})}{(\mathcal{N}_h e_j^{(p)})(\mathbf{x})} \, \mathrm{d}\mathbf{x} + \sum_{j=1}^{m} \int g(\mathbf{x}) w_j^{(p)}(\mathbf{x}) \log \frac{w_j^{(p)}(\mathbf{x})}{w_j^{(p+1)}(\mathbf{x})} \, \mathrm{d}\mathbf{x} \\
&= \sum_{j=1}^{m} KL(e_j^{(p+1)}, e_j^{(p)}) \\
&\quad + \sum_{j=1}^{m} \int \left[g(\mathbf{x}) w_j^{(p)}(\mathbf{x}) \log \frac{g(\mathbf{x}) w_j^{(p)}(\mathbf{x})}{g(\mathbf{x}) w_j^{(p+1)}(\mathbf{x})} + g(\mathbf{x}) w_j^{(p+1)}(\mathbf{x}) - g(\mathbf{x}) w_j^{(p)}(\mathbf{x}) \right] \mathrm{d}\mathbf{x}
\end{aligned}
$$

$$= \sum_{j=1}^{m} KL(e_j^{(p+1)}, e_j^{(p)}) + \sum_{j=1}^{m} KL(g \cdot w_j^{(p)}, g \cdot w_j^{(p+1)}). \tag{A1}$$

■

A.3. *Direct Proof of Corollary (3.3)*

Proof. If we define $\lambda_j = \int e_j(\mathbf{x})\,\mathrm{d}\mathbf{x}$ and $f_j(\mathbf{x}) = e_j(\mathbf{x})/\lambda_j$, then Jensen's inequality together with some simplification give

$$
\begin{aligned}
&l(\mathbf{e}^{(p)}) - l(\mathbf{e}^{(p+1)}) \\
&= \int g(\mathbf{x}) \log \frac{\sum_{j=1}^{m} (\mathcal{N}_h e_j^{(p+1)})(\mathbf{x})}{\sum_{j'=1}^{m} (\mathcal{N}_h e_{j'}^{(p)})(\mathbf{x})} \,\mathrm{d}\mathbf{x} = \int g(\mathbf{x}) \log \sum_{j=1}^{m} \frac{(\mathcal{N}_h e_j^{(p)})(\mathbf{x})}{\sum_{j'=1}^{m} (\mathcal{N}_h e_{j'}^{(p)})(\mathbf{x})} \cdot \frac{(\mathcal{N}_h e_j^{(p+1)})(\mathbf{x})}{(\mathcal{N}_h e_j^{(p)})(\mathbf{x})} \,\mathrm{d}\mathbf{x} \\
&\geq \int g(\mathbf{x}) \sum_{j=1}^{m} \frac{(\mathcal{N}_h e_j^{(p)})(\mathbf{x})}{\sum_{j'=1}^{m} (\mathcal{N}_h e_{j'}^{(p)})(\mathbf{x})} \cdot \log \frac{(\mathcal{N}_h e_j^{(p+1)})(\mathbf{x})}{(\mathcal{N}_h e_j^{(p)})(\mathbf{x})} \,\mathrm{d}\mathbf{x} \\
&= \sum_{j=1}^{m} \int g(\mathbf{x}) w_j^{(p)}(\mathbf{x}) \log \frac{\lambda_j^{(p+1)} \prod_{k=1}^{r} \mathcal{N}_h f_{j,k}^{(p+1)}(x_k)}{\lambda_j^{(p)} \prod_{k=1}^{r} \mathcal{N}_h f_{j,k}^{(p)}(x_k)} \,\mathrm{d}\mathbf{x} \\
&= \sum_{j=1}^{m} \int g(\mathbf{x}) w_j^{(p)}(\mathbf{x}) \left[\log \frac{\lambda_j^{(p+1)}}{\lambda_j^{(p)}} + \sum_{k=1}^{r} \int s_h(u_k, x_k) \log \frac{f_{j,k}^{(p+1)}(u_k)}{f_{j,k}^{(p)}(u_k)} \,\mathrm{d}u_k \right] \mathrm{d}\mathbf{x} \\
&= \sum_{j=1}^{m} \lambda_j^{(p+1)} \log \frac{\lambda_j^{(p+1)}}{\lambda_j^{(p)}} + \sum_{j=1}^{m} \sum_{k=1}^{r} \int \left(\int g(\mathbf{x}) w_j^{(p)}(\mathbf{x}) s_h(u_k, x_k) \,\mathrm{d}\mathbf{x} \right) \log \frac{f_{j,k}^{(p+1)}(u_k)}{f_{j,k}^{(p)}(u_k)} \,\mathrm{d}u_k \\
&= \sum_{j=1}^{m} \lambda_j^{(p+1)} \log \frac{\lambda_j^{(p+1)}}{\lambda_j^{(p)}} + \sum_{j=1}^{m} \sum_{k=1}^{r} \int \lambda_j^{(p+1)} f_{j,k}^{(p+1)}(u_k) \log \frac{f_{j,k}^{(p+1)}(u_k)}{f_{j,k}^{(p)}(u_k)} \,\mathrm{d}u_k \\
&= \sum_{j=1}^{m} \lambda_j^{(p+1)} \log \frac{\lambda_j^{(p+1)}}{\lambda_j^{(p)}} + \sum_{j=1}^{m} \int \lambda_j^{(p+1)} \left(\prod_{k=1}^{r} f_{j,k}^{(p+1)}(u_k) \right) \log \frac{\prod_{k=1}^{r} f_{j,k}^{(p+1)}(u_k)}{\prod_{k=1}^{r} f_{j,k}^{(p)}(u_k)} \,\mathrm{d}\mathbf{u} \\
&= \sum_{j=1}^{m} \int \lambda_j^{(p+1)} \left(\prod_{k=1}^{r} f_{j,k}^{(p+1)}(u_k) \right) \log \frac{\lambda_j^{(p+1)} \prod_{k=1}^{r} f_{j,k}^{(p+1)}(u_k)}{\lambda_j^{(p)} \prod_{k=1}^{r} f_{j,k}^{(p)}(u_k)} \,\mathrm{d}\mathbf{u} \\
&= \sum_{j=1}^{m} \int e_j^{(p+1)}(\mathbf{u}) \log \frac{e_j^{(p+1)}(\mathbf{u})}{e_j^{(p)}(\mathbf{u})} \,\mathrm{d}\mathbf{u} \\
&= \sum_{j=1}^{m} \int \left(e_j^{(p+1)}(\mathbf{u}) \log \frac{e_j^{(p+1)}(\mathbf{u})}{e_j^{(p)}(\mathbf{u})} + e_j^{(p)}(\mathbf{u}) - e_j^{(p+1)}(\mathbf{u}) \right) \mathrm{d}\mathbf{u}
\end{aligned}
$$

$$= \sum_{j=1}^{m} KL(e_j^{(p+1)}, e_j^{(p)}).$$

■

A.4. *Proof of Lemma 4.1*

Proof. For each j, $1 \le j \le m$, and $\mathbf{x} \in \Omega$, Jensen's Inequality gives

$$\begin{aligned}(\mathcal{N}_h e_j)(\mathbf{x}) &= \exp\left\{\int \tilde{s}_h(\mathbf{u},\mathbf{x}) \log e_j(\mathbf{u})\, \mathrm{d}\mathbf{u}\right\} \\ &\le \int \tilde{s}_h(\mathbf{u},\mathbf{x}) \exp\left[\log e_j(\mathbf{u})\right] \mathrm{d}\mathbf{u} \\ &= \int \tilde{s}_h(\mathbf{u},\mathbf{x}) e_j(\mathbf{u})\, \mathrm{d}\mathbf{u}. \end{aligned} \tag{A2}$$

Integrate both sides with respect to $\mathbf{x}$, then use Fubini's Theorem to obtain

$$\int (\mathcal{N}_h e_j)(\mathbf{x})\, \mathrm{d}\mathbf{x} \le \int e_j(\mathbf{u})\, \mathrm{d}\mathbf{u}. \tag{A3}$$

Summing over j, we get

$$\int \sum_{j=1}^{m} (\mathcal{N}_h e_j) \le \int \sum_{j=1}^{m} e_j. \tag{A4}$$

Therefore, all three terms on the right hand side of

$$l(\mathbf{e}) = KL\left(g, \sum_{j=1}^{m} (\mathcal{N}_h e_j)\right) + \int g + \left[\int \sum_{j=1}^{m} e_j - \int \sum_{j=1}^{m} (\mathcal{N}_h e_j)\right] \tag{A5}$$

are nonnegative and the middle term is 1, which implies that $l(\cdot)$ is always bounded below by 1.

For discrete case, Jensen's Inequality gives

$$\begin{aligned} l_{\text{discrete}}(\mathbf{e}) &= -\frac{1}{n}\sum_{i=1}^{n} \log \sum_{j=1}^{m} (\mathcal{N}_h e_j)(\mathbf{x}_i) + \int \sum_{j=1}^{m} e_j(\mathbf{x}_i) \\ &\ge -\frac{1}{n}\sum_{i=1}^{n} \log \sum_{j=1}^{m} (\mathcal{N}_h e_j)(\mathbf{x}_i) \\ &\ge -\frac{1}{n}\sum_{i=1}^{n} \log \sum_{j=1}^{m} (\mathcal{S}_h^* e_j)(\mathbf{x}_i) \ge -\frac{1}{n}\sum_{i=1}^{n} \log M_2 = -\log M_2. \end{aligned} \tag{A6}$$

■

A.5. *Proof of Lemma 4.3*

Proof. In the continuous case, for $p \geq 1$ and $\mathbf{u} \in \Omega$,

$$e_j^{(p)}(\mathbf{u}) = \frac{\prod_{k=1}^{r} \int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x}) \cdot s_h(u_k, x_k)\ \mathrm{d}\mathbf{x}}{\left[\int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x})\ \mathrm{d}\mathbf{x}\right]^{r-1}} \leq M_2^r \cdot \left[\int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x})\ \mathrm{d}\mathbf{x}\right] \leq M_2^r.$$

Thus $\{e_j^{(p)}\}_{1 \leq p < \infty}$ is uniformly bounded. Also, for any $\mathbf{u}$ in the interior of Ω,

$$\begin{aligned}\left|\frac{\partial}{\partial u_l} e_j^{(p)}(\mathbf{u})\right| &= \left| \frac{\left[\prod_{k \neq l}^{r} \int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x}) s_h(u_k, x_k)\ \mathrm{d}\mathbf{x}\right] \cdot \int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x}) \frac{\partial}{\partial u_l} s_h(u_l, x_l)\ \mathrm{d}\mathbf{x}}{\left[\int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x})\ \mathrm{d}\mathbf{x}\right]^{r-1}} \right| \\ &\leq B \cdot M_2^{r-1} \cdot \left[\int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x})\ \mathrm{d}\mathbf{x}\right] \\ &\leq B \cdot M_2^{r-1}. \end{aligned} \tag{A7}$$

By the Dominated Convergence Theorem, the above differentiation under the integral is allowed because the term $|g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x}) \frac{\partial}{\partial u_l} s_h(u_l, x_l)|$ is uniformly bounded by the integrable function $B \cdot g(\mathbf{x})$.

Now by the Mean Value Theorem for functions of several variables, for any $\mathbf{u}, \mathbf{v} \in \Omega$, there is some $d \in (0, 1)$ such that

$$e_j^{(p)}(\mathbf{u}) - e_j^{(p)}(\mathbf{v}) = \nabla e_j^{(p)}[(1-d)\mathbf{v} + d\mathbf{u}] \cdot (\mathbf{u} - \mathbf{v}). \tag{A8}$$

So

$$\left| e_j^{(p)}(\mathbf{u}) - e_j^{(p)}(\mathbf{v}) \right| \leq [B \cdot M_2^{r-1}] \cdot \|\mathbf{u} - \mathbf{v}\|_1, \tag{A9}$$

which shows that $\{e_j^{(p)}\}_{1 \leq p < \infty}$ is equicontinuous on Ω in the L^1 norm.

This proof can be readily adapted to the discrete case by replacing the integrals by summations. ■

A.6. *Proof of Lemma 4.5*

Proof. We first consider the continuous case. In the following, Fatou's Lemma will be applied twice to get the desired result. First, we show by Jensen's Inequality that all $\mathcal{N}_h \gamma_j^{(p)}$ are bounded from above by Q:

$$\begin{aligned}\left(\mathcal{N}_h \gamma_j^{(p)}\right)(\mathbf{x}) &= \exp\left\{\int \tilde{s}_h(\mathbf{u}, \mathbf{x}) \log \gamma_j^{(p)}(\mathbf{u})\, \mathrm{d}\mathbf{u}\right\} \\ &\leq \exp\left\{\int \tilde{s}_h(\mathbf{u}, \mathbf{x}) \log Q\, \mathrm{d}\mathbf{u}\right\} = Q. \end{aligned} \tag{A10}$$

Now, for any fixed value of $\mathbf{x}$, the nonnegative measurable function $\tilde{s}_h(\cdot,\mathbf{x})[Q - \log\gamma_j^{(p)}(\cdot)]$ converges to $\tilde{s}_h(\cdot,\mathbf{x})[Q - \log\gamma_j^{(\infty)}(\cdot)]$ pointwise in $L^1(\mathbb{R}^r)$. These functions are allowed to attain the value $+\infty$. By Fatou's Lemma, we have

$$\liminf_{p\to\infty}\int \tilde{s}_h(\mathbf{u},\mathbf{x})[Q-\log\gamma_j^{(p)}(\mathbf{u})]\,\mathrm{d}\mathbf{u} \geq \int \tilde{s}_h(\mathbf{u},\mathbf{x})[Q-\log\gamma_j^{(\infty)}(\mathbf{u})]\,\mathrm{d}\mathbf{u}. \tag{A11}$$

Exponentiating, this implies

$$\limsup_{p\to\infty}\exp\left\{\int \tilde{s}_h(\mathbf{u},\mathbf{x})\log\gamma_j^{(p)}(\mathbf{u})\,\mathrm{d}\mathbf{u}\right\} \leq \exp\left\{\int \tilde{s}_h(\mathbf{u},\mathbf{x})\log\gamma_j^{(\infty)}(\mathbf{u})\,\mathrm{d}\mathbf{u}\right\}. \tag{A12}$$

That is,

$$\limsup_{p\to\infty}\left(\mathcal{N}_h\gamma_j^{(p)}\right)(\mathbf{x}) \leq \left(\mathcal{N}_h\gamma_j^{(\infty)}\right)(\mathbf{x}), \tag{A13}$$

which implies that

$$\liminf_{p\to\infty} g(\mathbf{x})\log\frac{g(\mathbf{x})}{\sum_{j=1}^{m}\left(\mathcal{N}_h\gamma_j^{(p)}\right)(\mathbf{x})} \geq g(\mathbf{x})\log\frac{g(\mathbf{x})}{\sum_{j=1}^{m}\left(\mathcal{N}_h\gamma_j^{(\infty)}\right)(\mathbf{x})}. \tag{A14}$$

Since $a\log(a/b)+b-a$ is nonnegative for all $a,b\geq 0$, we have

$$g(\mathbf{x})\log\frac{g(\mathbf{x})}{\sum_{j=1}^{m}\left(\mathcal{N}_h\gamma_j^{(p)}\right)(\mathbf{x})} \geq g(\mathbf{x}) - \sum_{j=1}^{m}\left(\mathcal{N}_h\gamma_j^{(p)}\right)(\mathbf{x}) \geq -m\cdot Q. \tag{A15}$$

Thus, we can rewrite (A14) as

$$\liminf_{p\to\infty}\left[g(\mathbf{x})\log\frac{g(\mathbf{x})}{\sum_{j=1}^{m}\left(\mathcal{N}_h\gamma_j^{(p)}\right)(\mathbf{x})} + m\cdot Q\right] \geq g(\mathbf{x})\log\frac{g(\mathbf{x})}{\sum_{j=1}^{m}\left(\mathcal{N}_h\gamma_j^{(\infty)}\right)(\mathbf{x})} + m\cdot Q, \tag{A16}$$

so that both sides are nonnegative.

Now apply Fatou's Lemma again to obtain

$$\begin{aligned}\liminf_{p\to\infty}\int&\left[g(\mathbf{x})\log\frac{g(\mathbf{x})}{\sum_{j=1}^{m}\left(\mathcal{N}_h\gamma_j^{(p)}\right)(\mathbf{x})} + m\cdot Q\right]\mathrm{d}\mathbf{x}\\ &\geq \int\left[\liminf_{p\to\infty} g(\mathbf{x})\log\frac{g(\mathbf{x})}{\sum_{j=1}^{m}\left(\mathcal{N}_h\gamma_j^{(p)}\right)(\mathbf{x})} + m\cdot Q\right]\mathrm{d}\mathbf{x}\end{aligned}$$

$$\geq \int \left[g(\mathbf{x}) \log \frac{g(\mathbf{x})}{\sum_{j=1}^{m} \left(\mathcal{N}_h \gamma_j^{(\infty)} \right)(\mathbf{x})} + m \cdot Q \right] \mathrm{d}\mathbf{x}. \tag{A17}$$

We conclude that

$$\liminf_{p\to\infty} \int g(\mathbf{x}) \log \frac{g(\mathbf{x})}{\sum_{j=1}^{m} \left(\mathcal{N}_h \gamma_j^{(p)} \right)(\mathbf{x})} \mathrm{d}\mathbf{x} \geq \int g(\mathbf{x}) \log \frac{g(\mathbf{x})}{\sum_{j=1}^{m} \left(\mathcal{N}_h \gamma_j^{(\infty)} \right)(\mathbf{x})} \mathrm{d}\mathbf{x}. \tag{A18}$$

The uniform convergence of $\gamma_j^{(p)}$ to $\gamma_j^{(\infty)}$ for each j, together with (A18), imply

$$\liminf_{p\to\infty} \left[\int g(\mathbf{x}) \log \frac{g(\mathbf{x})}{\sum_{j=1}^{m} \left(\mathcal{N}_h \gamma_j^{(p)} \right)(\mathbf{x})} \mathrm{d}\mathbf{x} + \int \sum_{j=1}^{m} \gamma_j^{(p)}(\mathbf{x})\, \mathrm{d}\mathbf{x} \right]$$
$$\geq \int g(\mathbf{x}) \log \frac{g(\mathbf{x})}{\sum_{j=1}^{m} \left(\mathcal{N}_h \gamma_j^{(\infty)} \right)(\mathbf{x})} \mathrm{d}\mathbf{x} + \int \sum_{j=1}^{m} \gamma_j^{(\infty)}(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{A19}$$

That is,

$$\liminf_{p\to\infty} l(\gamma^{(p)}) \geq l(\gamma^{(\infty)}), \tag{A20}$$

which establishes the desired lower semi-continuity.

The proof can be adpated to the discrete case by replacing the integrals with summations. ■

A.7. *Proof of Theorem 4.6*

Proof. By Lemma 4.1, $\tau := \inf\{l(\mathbf{e}) | \mathbf{e}$ satisfies assumptions (ii) and (vi)$\}$ is a finite constant. So there exists a sequence $\{\boldsymbol{\psi}^{(p)}\}_{0\leq p\leq\infty}$ satisfying assumptions (ii) and (vi) such that

$$\lim_{p\to\infty} l(\boldsymbol{\psi}^{(p)}) = \tau. \tag{A21}$$

By Lemma 4.4, for each j, $1 \leq j \leq m$, the sequence $\{(G(\boldsymbol{\psi}^{(p)}))_j\}_{0\leq p\leq\infty}$ is bounded and equicontinuous.

By the Arzelà-Ascoli theorem, we know that $\{(G(\boldsymbol{\psi}^{(p)}))_j\}_{0\leq p\leq\infty}$ has a uniformly convergent subsequence. Applying this theorem m times to $\{(G(\boldsymbol{\psi}^{(p)}))\}_{0\leq p\leq\infty}$ we can extract a subsequence that converges uniformly in every component. This subsequence also satisfies (ii) and (vi).

That is, there exists a sequence $\{(G(\boldsymbol{\psi}^{(p_k)}))\}_{0\leq k\leq\infty}$, such that, for each j, $1 \leq j \leq m$, $\{(G(\boldsymbol{\psi}^{(p_k)}))_j\}_{0\leq k\leq\infty}$ converges uniformly to a limit function in $L^1(\mathbb{R}^r)$. Denote this limit function by $\tilde{\psi}_j$. As usual, let $\tilde{\boldsymbol{\psi}}$ denote the m-tuples $(\tilde{\psi}_1, \cdots, \tilde{\psi}_m)$. If all components of

$\tilde{\boldsymbol{\psi}}$ are nonzero, then $\tilde{\boldsymbol{\psi}}$ satisfies (iii). If not, we can split up some nonzero components of $\tilde{\boldsymbol{\psi}}$ so that all components become nonzero, which does not change the value of $l(\tilde{\boldsymbol{\psi}})$. In a word, we can assume that $\tilde{\boldsymbol{\psi}}$ satisfies (vi).

Now, by Lemma 4.5 and the fact that G does not increase the value of l (see the proof of Lemma 3.3), we have

$$\tau \leq l(\tilde{\boldsymbol{\psi}}) \leq \lim_{k\to\infty} l(G(\boldsymbol{\psi}^{(p_k)})) \leq \lim_{k\to\infty} l(\boldsymbol{\psi}^{(p_k)}) = \lim_{p\to\infty} l(\boldsymbol{\psi}^{(p)}) = \tau, \tag{A22}$$

so that $l(\tilde{\boldsymbol{\psi}}) = \tau$. Apply the operator G to $\tilde{\boldsymbol{\psi}}$. By Lemma 3.3 and the fact that $l(\tilde{\boldsymbol{\psi}})$ has already attained the infimum value in this setting, we have

$$0 \geq l(\tilde{\boldsymbol{\psi}}) - l(G(\tilde{\boldsymbol{\psi}})) \geq \sum_{j=1}^{m} KL((G(\tilde{\boldsymbol{\psi}}))_j, \tilde{\psi}_j) \geq 0. \tag{A23}$$

So for each j, $1 \leq j \leq m$, $G(\tilde{\boldsymbol{\psi}})_j = \tilde{\psi}_j$ in $L^1(\mathbb{R}^r)$. Thus in particular, by (33), $\tilde{\boldsymbol{\psi}}$ also satisfies assumption (ii). We have proved the existence of a solution, $\tilde{\boldsymbol{\psi}}$, to the main optimization problem (13).

As above, the proof can readily be adapted to the discrete case. ■

A.8. *Proof of Lemma 4.7*

Proof. First, by assumption (vi), each $e_j^{(0)}$ is strictly positive on Ω. So given any $\mathbf{x} \in \Omega$,

$$\left(\mathcal{N}_h e_j^{(0)}\right)(\mathbf{x}) = \exp\left[\left(S_h^* \log e_j^{(0)}\right)(\mathbf{x})\right] > 0. \tag{A24}$$

Thus,

$$w_j^{(0)}(\mathbf{x}) = \frac{\left(\mathcal{N}_h e_j^{(0)}\right)(\mathbf{x})}{\sum\limits_{j=1}^{m}\left(\mathcal{N}_h e_j^{(0)}\right)(\mathbf{x})} > 0, \tag{A25}$$

which implies

$$\int g(\mathbf{x}) w_j^{(0)}(\mathbf{x})\ \mathrm{d}\mathbf{x} > 0. \tag{A26}$$

Now, we use induction. Assume

$$\int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x})\ \mathrm{d}\mathbf{x} > 0. \tag{A27}$$

We have

$$f_j^{(p)}(u) = \frac{\prod\limits_{k=1}^{r} \int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x}) \cdot s_h(u_k, x_k)\ \mathrm{d}\mathbf{x}}{\left[\int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x})\ \mathrm{d}\mathbf{x}\right]^r} \leq M_2^r. \tag{A28}$$

Similarly,

$$f_j^{(p)}(u) \quad = \frac{\prod_{k=1}^{r} \int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x}) \cdot s_h(u_k, x_k) \, \mathrm{d}\mathbf{x}}{\left[\int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x}) \, \mathrm{d}\mathbf{x}\right]^r} \geq (M_1(h))^r. \tag{A29}$$

Therefore,

$$\begin{aligned} \left(\mathcal{N}_h e_j^{(p)}\right)(\mathbf{x}) &= \int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x}) \, \mathrm{d}\mathbf{x} \cdot \exp\left[\left(S_h^* \log f_j^{(p)}\right)(\mathbf{x})\right] \\ &\geq \int g(\mathbf{x}) w_j^{(p-1)}(\mathbf{x}) \, \mathrm{d}\mathbf{x} \cdot (M_1(h))^r \\ &> 0. \end{aligned} \tag{A30}$$

We conclude that

$$w_j^{(p)}(\mathbf{x}) = \frac{\left(\mathcal{N}_h e_j^{(p)}\right)(\mathbf{x})}{\sum_{j=1}^{m} \left(\mathcal{N}_h e_j^{(p)}\right)(\mathbf{x})} > 0, \tag{A31}$$

which gives

$$\int g(\mathbf{x}) w_j^{(p)}(\mathbf{x}) \, \mathrm{d}\mathbf{x} > 0. \tag{A32}$$

The next step of the induction follows in the same way, and the result is established. ■

References

Allman, E.S., Matias, C., and Rhodes, J.A. (2009), "Identifiability of parameters in latent structure models with many observed variables," *The Annals of Statistics*, 37, 3099–3132.

Benaglia, T., Chauveau, D., and Hunter, D.R. (2009), "An EM-like algorithm for semi-and nonparametric estimation in multivariate mixtures," *Journal of Computational and Graphical Statistics*, 18, 505–526.

Benaglia, T., Chauveau, D., Hunter, D.R. et al. (2011), "Bandwidth selection in an EM-like algorithm for nonparametric multivariate mixtures," *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger*, pp. 15–27.

Bonhomme, S., Jochmans, K., and Robin, J.M. (2014), "Nonparametric estimation of finite mixtures," Centre for Microdata Methods and Practice (cemmap) Working Paper CWP11/14, London.

Bordes, L., Chauveau, D., and Vandekerkhove, P. (2007), "A stochastic EM algorithm for a semiparametric mixture model," *Computational Statistics & Data Analysis*, 51, 5429–5443.

Chauveau, D., Hunter, D.R., and Levine, M. (2015), "Semi-Parametric Estimation for Conditional Independence Multivariate Finite Mixture Models," *Statistics Surveys*, 9, 1–31.

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

Eggermont, P., and LaRiccia, V., *Maximum Penalized Likelihood Estimation: Volume I: Density Estimation*, Vol. 1, Springer (2001).

Eggermont, P. (1999), “Nonlinear smoothing and the EM algorithm for positive integral equations of the first kind,” *Applied mathematics & optimization*, 39, 75–91.
Hall, P., Neeman, A., Pakyari, R., and Elmore, R. (2005), “Nonparametric inference in multivariate mixtures,” *Biometrika*, 92, 667–678.
Hall, P., and Zhou, X.H. (2003), “Nonparametric estimation of component distributions in a multivariate mixture,” *Annals of Statistics*, pp. 201–224.
Hunter, D.R., and Lange, K. (2004), “A tutorial on MM algorithms,” *The American Statistician*, 58, 30–37.
Kasahara, H., and Shimotsu, K. (2009), “Nonparametric identification of finite mixture models of dynamic discrete choices,” *Econometrica*, 77, 135–175.
Kasahara, H., and Shimotsu, K. (2014), “Non-parametric identification and estimation of the number of components in multivariate mixtures,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 97–111.
Kruskal, J.B. (1976), “More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling,” *Psychometrika*, 41, 281–293.
Kruskal, J.B. (1977), “Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics,” *Linear algebra and its applications*, 18, 95–138.
Laird, N.M., and Ware, J.H. (1982), “Random-effects models for longitudinal data,” *Biometrics*, pp. 963–974.
Levine, M., Hunter, D., and Chauveau, D. (2011), “Maximum smoothed likelihood for multivariate mixtures,” *Biometrika*, 98, 403–416.