A. Numerical Implementation of the Sieve Estimation

In this section, we discuss the numerical implementation of obtaining our proposed sieve estimators. The estimation procedure is based on the EM algorithm. In particular, we describe the special considerations regarding efficient optimization of the constrained maximization problem in each maximization step.

Let’s start from the model (3.4) in Section 3.2 where the target variable $X$ is approximated by replacing its continuous component $X_c$ with $\tilde{X}_c$. To simplify the notation, let $u = (a_1, \ldots, a_r, x_1, \ldots, x_r)^T$ and let $p = (\pi_1, \ldots, \pi_r, \pi_{r+1}\theta^T)^T$. Note that $p$ and $(\pi, \theta)$ has 1-1 correspondence, so the estimation of $p$ is equivalent to the estimation of $(\pi, \theta)$. Hence the log-likelihood (3.5) in Section 3.2 can be re-formulated as

$$\ell(p) = \sum_{i=1}^{N} \log f_Y(Y_i \mid p) = \sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{r+\nu} f_Y(Y_i \mid X = u_k)P(X = u_k \mid p) \right\}$$

$$= \sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{r+\nu} p(X = u_k \mid Y_i, p_\nu) \frac{f_Y(Y_i \mid X = u_k)P(X = u_k \mid p)}{p(X = u_k \mid Y_i, p_\nu)} \right\}$$

$$\geq \sum_{i=1}^{N} \sum_{k=1}^{r+\nu} p(X = u_k \mid Y_i, p_\nu) \log \left\{ \frac{f_Y(Y_i \mid X = u_k)P(X = u_k \mid p)}{p(X = u_k \mid Y_i, p_\nu)} \right\}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{r+\nu} p(X = u_k \mid Y_i, p_\nu) \log \left\{ f_Y(Y_i \mid X = u_k)P(X = u_k \mid p) \right\}$$

$$- \sum_{i=1}^{N} \sum_{k=1}^{r+\nu} p(X = u_k \mid Y_i, p_\nu) \log p(X = u_k \mid Y_i, p_\nu)$$

$$= \ell_1(p \mid p_\nu) - \ell_2(p_\nu)$$

for any $p_\nu$, by Jensen’s inequality. Due to the nonnegativity conditions on $\pi$ and $\theta$, the maximization of $\ell(p)$ does not have a closed form.
Instead, we apply the EM algorithm which maximizes $\ell_1(p \mid p_\star)$ in each maximization step. More precisely, we set the initial value $p^{(0)}$. In $n$th iteration,

\[ E\text{-step}: \text{determine the conditional probability } p(X = u_k \mid Y_i, p^{(n-1)}); \]
\[ M\text{-step}: p^{(n)} = \arg\max_p \ell_1(p \mid p^{(n-1)}). \]

### A.1. E-step

Under our model,

\[
\begin{align*}
  f_Y(Y_i \mid X = u_k, p) &= f_Z(Y_i - u_k \mid p) = f_Z(Y_i - a_k)I(k \leq \nu) + f_Z(Y_i - x_{k-\nu} \mid p)I(k \geq \nu + 1) \\
  P(X = u_k \mid p) &= p_k = \pi_k I(k \leq \nu) + \pi_{\nu+1} \theta_{k-\nu} I(k \geq \nu + 1) \\
  P(X = u_k \mid Y_i, p) &= \frac{P(Y_i \mid X = u_k, p)P(X = u_k \mid p)}{\sum_{k=1}^{\nu+\nu} P(Y_i \mid X = u_k, p)P(X = u_k \mid p)} \\
  &= \frac{f_Z(Y_i - a_k)I(k \leq \nu) + f_Z(Y_i - x_{k-\nu})I(k \geq \nu + 1)}{\sum_{l=1}^{\nu} \pi_l f_Z(Y_i - a_l) + \pi_{\nu+1} \sum_{j=1}^{\nu} \theta_j f_Z(Y_i - x_j)}.
\end{align*}
\]

By combining those three, we get

\[
\ell_1(p \mid p^{(n-1)}) = \sum_{i=1}^{N} \frac{\sum_{l=1}^{\nu} f_Z(Y_i - a_l) \log \{ \pi_l f_Z(Y_i - u_l) \} + \sum_{j=1}^{r} f_Z(Y_i - x_j) \log \{ \pi_{\nu+1} \theta_j f_Z(Y_i - x_j) \}}{\sum_{i=1}^{N} \sum_{l=1}^{\nu} \pi_l^{(n-1)} f_Z(Y_i - a_l) + \pi_{\nu+1}^{(n-1)} \sum_{j=1}^{\nu} \theta_j^{(n-1)} f_Z(Y_i - x_j)}
\]

\[
= \sum_{i=1}^{N} \frac{\sum_{l=1}^{\nu} f_Z(Y_i - a_l) \log \pi_l + \sum_{j=1}^{r} f_Z(Y_i - x_j) \log \{ \pi_{\nu+1} \theta_j f_Z(Y_i - x_j) \}}{\sum_{i=1}^{N} \sum_{l=1}^{\nu} \pi_l^{(n-1)} f_Z(Y_i - a_l) + \pi_{\nu+1}^{(n-1)} \sum_{j=1}^{\nu} \theta_j^{(n-1)} f_Z(Y_i - x_j)}
\]

\[
+ \sum_{l=1}^{\nu} f_Z(Y_i - a_l) \log f_Z(Y_i - u_l) + \sum_{j=1}^{r} f_Z(Y_i - x_j) \log f_Z(Y_i - x_j)
\]

\[
= \sum_{i=1}^{N} \frac{\sum_{l=1}^{\nu} \pi_l^{(n-1)} f_Z(Y_i - a_l) + \sum_{j=1}^{r} \theta_j^{(n-1)} f_Z(Y_i - x_j)}{\sum_{i=1}^{N} \sum_{l=1}^{\nu} \pi_l^{(n-1)} f_Z(Y_i - a_l) + \sum_{j=1}^{r} \theta_j^{(n-1)} f_Z(Y_i - x_j)}.
\]

### A.2. M-step

In the $n$th maximization step, $(\pi^{(n)}, \theta^{(n)})$ is obtained by maximizing $\ell_1(p \mid p^{(n-1)})$, i.e.

\[
(\pi^{(n)}, \theta^{(n)}) = \arg\max_{\pi, \theta} \ell_1(p \mid p^{(n-1)})
\]

\[
= \arg\min_{\pi, \theta} \left\{ \sum_{i=1}^{N} \left( \sum_{l=1}^{\nu} P_{i,l} \log \pi_l + \sum_{j=1}^{r} Q_{i,j} \log \pi_{\nu+1} + \log \theta_j \right) \right\}
\]  

subject to $\pi_j \leq 0$, $\theta_j \leq 0$, $\sum_{l=1}^{\nu+1} \pi_l = \sum_{j=1}^{r} \theta_j = 1$, where

\[
P_{i,l} = \frac{f_Z(Y_i - a_l)}{\sum_{l=1}^{\nu} \pi_l^{(n-1)} f_Z(Y_i - a_l) + \pi_{\nu+1}^{(n-1)} \sum_{j=1}^{r} \theta_j^{(n-1)} f_Z(Y_i - x_j)}
\]
\[
Q_{i,j} = \frac{f_Z(Y_i - x_j)}{\sum_{l=1}^{\nu} \pi_l^{(n-1)} f_Z(Y_i - a_l) + \pi_{\nu+1}^{(n-1)} \sum_{j=1}^{r} \theta_j^{(n-1)} f_Z(Y_i - x_j)}
\]
for $1 \leq i \leq N$, $1 \leq l \leq \nu$ and $1 \leq j \leq r$.

To solve the nonlinear constrained optimization, the most popular method uses Lagrange multipliers. Suppose that we want to maximize (or minimize) some function defined on the $r$-dimensional space with $k$ constraints. By introducing a new unknown variable, called the Lagrange multiplier, for each constraint, this method replaces the problem by one of maximizing (or minimizing) an unconstrained function, the Lagrangian, in $r + k$ variables.

The Karush-Kuhn-Tucker (KKT) condition is simply the generalization of the Lagrange multiplier theorem. Compared to the fact that the Lagrange multiplier theorem contains only the equality constraints, the KKT condition provides the analogous conditions in the optimization problem with both the equality and inequality constraints.

Consider the following problem:

$$\text{minimize } \{ f(\theta) : \theta \in \mathbb{R}^n \} \text{ subject to } h(\theta) = 0, \text{ and } g(\theta) \leq 0 \quad (A.2)$$

with a strictly convex function $f(\cdot)$, a $p$-dimensional vector $h(\cdot)$ and an $m$-dimensional vector $g(\cdot)$. From the KKT condition, the necessary and sufficient conditions for a unique minimum point, say $\theta^*$, are that there exist multipliers $\lambda^* \in \mathbb{R}^p$ and $s^* \in \mathbb{R}^m$ such that

$$\nabla f(\theta^*) + \sum_{j=1}^{p} \lambda^*_j \nabla h_j(\theta^*) + \sum_{j=1}^{m} s^*_j \nabla g_j(\theta^*) = 0,$$

where

$$h(\theta^*) = (0, 0, \ldots, 0)^T, \quad g_j(\theta^*)s^*_j = 0 \text{ for any } j,$$

$$g(\theta^*) \leq (0, 0, \ldots, 0)^T \quad \text{and} \quad s^* \geq 0.$$

Here, $\nabla$ represents the derivative of a function.

The standard approach for solving $(A.2)$ is to consider the following penalized problem:

$$\text{minimize } \left\{ f(\theta) - \mu \sum_{j=1}^{r} \log \theta_j : \theta \in \mathbb{R}^r \right\} \text{ subject to } h(\theta) = 0, \text{ and } g(\theta) \leq 0,$$

for small positive $\mu$. Note that $\theta_j$ should be greater than zero for this problem to make sense. If the solutions are $\theta(\mu)$ and $\lambda(\mu)$, then by letting $\mu \rightarrow 0$, we can obtain

$$\theta^* = \lim_{\mu \rightarrow 0} \theta(\mu), \quad \text{and} \quad \lambda^* = \lim_{\mu \rightarrow 0} \lambda(\mu).$$

This $\theta^*$ corresponds to our final parameter estimate.
When this method is applied to (A.1), \((\pi^{(n)}, \theta^{(n)})\) is obtained by minimizing

\[
g(\pi, \theta, \lambda) = -\sum_{i=1}^{N} \left\{ \sum_{l=1}^{\nu} P_{i,l} \log \pi_l + \sum_{j=1}^{r} Q_{i,j} (\log \pi_{l+1} + \log \theta_j) \right\} \\
-\mu \left( \sum_{l=1}^{\nu+1} \log \pi_l + \sum_{j=1}^{r} \log \theta_j \right) + \lambda_1 \left( \sum_{j=1}^{r} \theta_j - 1 \right) + \lambda_2 \left( \sum_{l=1}^{\nu+1} \pi_l - 1 \right) \quad (A.3)
\]

where \(\mu \to 0\).

In particular, when \(\nu = 1\), the above minimization has a closed form solution. For notational simplicity, let \(\pi_1 = \pi, \pi_2 = 1 - \pi, P_{i,1} = P_i\) and \(\lambda_1 = \lambda\). (\(\lambda_2\) need not be considered anymore). Then, (A.3) is simplified as

\[
g(\pi, \theta, \lambda) = \sum_{i=1}^{N} \left\{ P_i \log \pi + \sum_{j=1}^{r} Q_{i,j} \log(1 - \pi) + \sum_{j=1}^{r} Q_{i,j} \log \theta_j \right\} \\
-\mu \left( \log \pi + \log(1 - \pi) + \sum_{j=1}^{r} \log \theta_j \right) + \lambda \left( \sum_{j=1}^{r} \theta_j - 1 \right).
\]

To get the minimizer of \(g\), we solve the following equations:

\[
\frac{\partial g}{\partial \pi} = -\frac{1}{\pi} \left( \sum_{i=1}^{N} P_i + \mu \right) + \frac{1}{1 - \pi} \left( \sum_{i=1}^{N} \sum_{j=1}^{r} Q_{ij} + \mu \right) = 0;
\]

\[
\frac{\partial g}{\partial \theta_j} = -\frac{1}{\theta_j} \sum_{i=1}^{N} Q_{ij} - \frac{\mu}{\theta_j} + \lambda = 0, \quad j = 1, \ldots, r;
\]

\[
\frac{\partial g}{\partial \lambda} = \sum_{j=1}^{r} \theta_j - 1 = 0.
\]

By letting \(\mu \to 0\) after that, we finally get the parameter estimates

\[
\pi^{(n)} = \frac{\sum_{i=1}^{N} P_i}{N}, \quad \theta_j^{(n)} = \left( \sum_{i=1}^{N} Q_{ij} \right) / \left( \sum_{i=1}^{N} \sum_{j=1}^{r} Q_{ij} \right) = \left( \sum_{i=1}^{N} Q_{ij} \right) / \left( N - \sum_{i=1}^{N} Q_{i0} \right).
\]

In general, closed form solutions do not exist, and optimizers are obtained using Newton’s iteration algorithm.

**A.3. M-step When Considering a Roughness Penalty**

In the penalized method discussed in Section 3.3, we want to maximize the penalized log-likelihood \(\ell_\gamma(\pi, \theta)\) instead of \(\ell(\pi, \theta)\). Then, each M-step, we need to minimize \(g_\gamma(\pi, \theta, \lambda) = g(\pi, \theta, \lambda) + \gamma P(\theta)\) instead of \(g(\pi, \theta, \lambda)\).

Unlike the unpenalized case, we do not have a closed form solution even when \(\nu = 1\). The
estimation of $\pi$ does not change, but $\theta$ is obtained by Newton’s iteration algorithm.

$$\frac{\partial g}{\partial \theta_j} = -\frac{1}{\theta_j} \sum_{i=1}^N Q_{ij} - \frac{\mu}{\theta_j} + \lambda + \gamma \frac{\partial}{\partial \theta_j} P(\theta) \quad \Rightarrow \quad \frac{\partial g}{\partial \theta} = -(Q + \mu)1_r/\theta + \lambda 1_r + \gamma \dot{P}(\theta) = 0_r$$

$$\frac{\partial g}{\partial \lambda} = \sum_{j=1}^r \theta_j - 1 = 1_r^t \theta - 1 = 0$$

$$\frac{\partial^2 g}{\partial \theta^2} = \text{Diag} \left( \frac{\sum_{i=1}^N Q_{ij} + \mu}{\theta_j^2} \right)_{1 \leq j \leq r} + \gamma \ddot{P}(\theta), \quad \frac{\partial^2 g}{\partial \theta \partial \lambda} = 1_r, \quad \frac{\partial^2 g}{\partial \lambda^2} = 0$$

where $Q = [Q_{ij}]_{1 \leq i \leq N, 1 \leq j \leq r}, \ 1_r$ and $0_r$ are $r \times 1$ vectors of entries 1s and 0s, and the symbol “/” represents the entry-wise division.

Let

$$\dot{\theta} = \left[ \frac{\partial g}{\partial \theta} \right]_{|\pi = \pi^{(0)}, \ \theta = \theta^{(0)}} \quad \text{and} \quad \ddot{\theta} = \left[ \frac{\partial^2 g}{\partial \theta^2} \frac{\partial^2 g}{\partial \theta \partial \lambda} \right]_{|\pi = \pi^{(0)}, \ \theta = \theta^{(0)}}$$

Then,

$$\begin{bmatrix} \theta^{(1)} \\ \lambda^{(1)} \end{bmatrix} = \begin{bmatrix} \theta^{(0)} \\ \lambda^{(0)} \end{bmatrix} - \alpha^* (\dot{\theta})^{-1} \ddot{\theta}$$

where $\alpha^*$ is the maximum constant between 0 and 1 which makes the $\theta_j^{(1)} \geq 0$ for all $j$. Set $(\lambda^{(0)}, \theta^{(0)}) = (\lambda^{(1)}, \theta^{(1)})$ and repeat the above updating until converges. When $\gamma = 0$, the penalized method is exactly the same as the non-penalized one, so this procedure can be used for the standard sieve estimation as well.

**B. Proofs of The Consistency Theorems**

In this section, we provide detailed proofs of the two theorems that state the consistency of our proposed sieve estimators.

**Proof of Theorem 1.**

Let $m = m(n) \geq 2$ be an integer. We shall choose $m$ to diverge to infinity as $n$ increases, and such that

$$\frac{r}{m} \to 0. \quad (B.1)$$

Details will be given below (B.6). Consider a lattice of values of $q_1, \ldots, q_r$, where each $q_j$ is expressed as $m_j/m$, each $m_j$ is a nonnegative integer, and $m_1 + \cdots + m_r = m$. And suppose that each $q_j \leq C_1 h,$
where \( C_1 \) is as in (R4). Then \( q_j \leq C_2/r \) for some constant \( C_2 \) from (R2). Therefore, these \( q_j \)s can assume at most \( C_3 m/r \) different values for some \( C_3 \geq 1 \).

In addition, consider \( \phi = (\phi_1, \ldots, \phi_{\nu+1})^T \), where each \( \phi_l = n_l/[m/r] \), where \([x]\) denotes the integer part of \( x \), \( n_l \) is a nonnegative integer, and \( n_1 + \cdots + n_{\nu+1} = [m/r] \). Then each \( \phi_l \) can have at most \( m/r \) different values. Write \( Q \) for the class of all \( q \)'s which have the form \( q = (\phi_1, \ldots, \phi_{\nu+1} q^T) \) arising in this way. Then

\[
\text{card}(Q) \leq \left( \frac{C_3 m}{r} \right)^{r+\nu-1}.
\] (B.2)

Define \( d(y|q) = \log \left\{ \sum_{i=1}^{\nu} \phi_i f_Z(y - a_i) + \phi_{\nu+1} \sum_{j=1}^{\nu} q_j f_Z(y - x_j) \right\} \), and let \( g \) be as defined in (9). Note that if \( x_0 \in [c, d] \) then, for all \( y \),

\[
\frac{1}{g(y)} \leq \frac{\exp[d(y|q)]}{f_Z(y - x_0)} \leq g(y).
\]

Therefore, \( |d(y|q)| \leq 2 \max\{\log g(y), |\log f_Z(y - x_0)|\} \), from which it follows that

\[
|d(y|q)|^* \leq 2^n \left\{ \log g(y)|^* + |\log f_Z(y - x_0)|^* \right\}.
\] (B.3)

This result, and the assumptions (10) and (11), imply that there exists a constant \( B_1 > 0 \) such that

\[
\text{for all integers } s \geq 3, \quad E\left\{ |d(Y|q)|^s \right\} \leq \frac{s!}{2} \text{Var}\{d(Y|q)\}^2 B_1^{s-2},
\] (B.4)

where \( B_1 \) does not depend on \( q \) or \( s \). In view of (B.4), Bernstein-type bounds imply that if \( \tau^2 \) denotes an upper bound (Theorem 1.1 of De La Peña (1999)) to \( \text{Var}\{d(Y|q)\} \) for all \( q \). If \( 0 < \eta < \tau^2/B_1 \), then

\[
P\left( \left| \frac{1}{n} \sum_{i=1}^{n} [d(Y_i|q) - E\{d(Y_i|q)\}] \right| > \eta \right) \leq 2 \exp \left( - \frac{n \eta^2}{2 \left[ \text{Var}\{d(Y|q)\} + B_1 \eta \right]} \right)
\]

\[
\leq 2 \exp \left( - \frac{n \eta^2}{4 \tau^2} \right).
\] (B.5)

Note that Condition (B.3) implies that the upper bound \( \tau^2 \) exists and is finite.

Defining \( B_2 = (4\tau^2)^{-1} \) and

\[
\delta(q) = \frac{1}{n} \sum_{i=1}^{n} [d(Y_i|q) - E\{d(Y_i|q)\}],
\]

we deduce from (B.2) and (B.5) that

\[
P\left( \sup_{q \in Q} |\delta(q)| > \eta \right) \leq \sum_{q \in Q} P\{|\delta(q)| > \eta \}
\]

\[
\leq \text{card}(Q) \cdot \sup_{q \in Q} P\{|\delta(q)| > \eta \}
\]

\[
\leq 2 \left( \frac{C_3 m}{r} \right)^{r+\nu-1} \exp \left( - B_2 n \eta^2 \right).
\] (B.6)
Since \( r = o(n) \) by (R5) and \( \nu \) is finite, we may write \( \xi = (n) = n/(r + \nu - 1) \), which diverges to infinity as \( n \) goes to infinity. It suffices to treat the case where \( r \) is relatively large, and indeed we may assume without loss of generality that \( \xi \leq \log n \). Let \( m \) equal the integer part of \( rC_3^{-1}\exp(B_2 \xi^{1/2}/2) \) and \( \eta = \xi^{-1/4} \). Then \( r/m \to 0 \) and

\[
\log \left\{ \left( \frac{C_3 m}{r} \right)^{r+\nu-1} \exp \left( -B_2 n \eta^2 \right) \right\} \leq \frac{r+\nu-1}{2} B_2 \xi^{1/2} - B_2 n \xi^{-1/2} = -\frac{B_2 n}{2} \xi^{-1/2}.
\]

Hence, by (B.6),

\[
P \left\{ \sup_{q \in \mathcal{Q}} |\delta(q)| > \eta \right\} \leq 2 \exp \left( -\frac{B_2 n}{2} \xi^{-1/2} \right),
\]

which converges to zero faster than any power of \( n^{-1} \). Hence, by the Borel-Cantelli Lemma,

\[
\sup_{q \in \mathcal{Q}} |\delta(q)| \to 0 \quad \text{with probability } 1.
\]

Let \( \mathcal{P} \) be the class of all \( p = (\pi_1, \ldots, \pi_\nu, \pi_{\nu+1} \mathbf{\theta}^T) \), where each \( \pi_i \geq 0 \), \( \theta_j \geq 0 \), and \( \sum_i \pi_i = \sum_j \theta_j = 1 \). In fact, \( \mathcal{P} \) is the class of all \((r + \nu)\)-variate discrete distributions. In addition, each \( \theta_j \leq C_1 h, \) and given \( p \in \mathcal{P} \), let \( q = q(p) \) be the best approximation of \( q \) in the sense that it minimizes the \( L_1 \) distance between \( p \) and \( q \), i.e., \( ||p - q||_1 = \sum_{i=1}^\nu |\pi_i - \phi_i| + \sum_{j=1}^r |\pi_{\nu+1} \theta_j - \phi_{\nu+1} q_j| \) over all \( q \in \mathcal{Q} \). By construction of \( \mathcal{Q} \),

\[
||p - q||_1 \leq \frac{r(\nu + 2)}{m}, \tag{B.8}
\]

for sufficient large \( n \) from (B.1). Define \( D(y|q) = \sum_{i=1}^\nu \phi_i f_Z(y - a_i) + \phi_{\nu+1} \sum_j q_j f_Z(y - x_j) \) and define \( D(y|p) \) in the same way; and put

\[
D(y|q, p) = \frac{|D(y|q) - D(y|p)|}{\min\{D(y|q), D(y|p)\}} \geq 0.
\]

Property (B.8) implies that

\[
|D(y|q) - D(y|p)| \leq \frac{r(\nu + 2)}{m} \sup_{c \leq u \leq d} f_Z(y - u),
\]

from which it follows that \( D(y|q, p) \leq r(\nu + 2)g(y)/m \). Hence, if \( \delta \in (0, 1] \) and \( r < m/(\nu + 2) \),

\[
| \log D(y|q) - \log D(y|p) | = \log \{ 1 + D(y|q, p) \}
\]

\[
\leq \log \left\{ 1 + \frac{r(\nu + 2)g(y)}{m} \right\}
\]

\[
\leq \delta + \log \{ 1 + g(y) \} I \left\{ \frac{r(\nu + 2)g(y)}{m} > \delta \right\}
\]

\[
\leq \delta + \{ \log 2 + \log g(y) \} I \left\{ \frac{r(\nu + 2)g(y)}{m} > \delta \right\},
\]

7
where the last inequality follows from the fact that \( g(\cdot) \geq 1 \). Therefore,

\[
\sup_{p \in P} \left| \frac{1}{n} \sum_{i=1}^{n} \{ d(Y_i|q) - d(Y_i|p) \} \right| = \left| \frac{1}{n} \sum_{i=1}^{n} \{ \log D(Y_i|q) - \log D(Y_i|p) \} \right|
\leq \delta + \frac{1}{n} \sum_{i=1}^{n} \left\{ \log 2 + \log g(Y_i) \right\} I \left\{ g(Y_i) > \frac{m\delta}{r(\nu + 2)} \right\}.
\]

(B.9)

In view of (B.1), the right-hand side of (B.9) converges to \( \delta \), with probability 1, as \( n \to \infty \). Since this is true for each \( \delta > 0 \) then the left-hand side converges to zero:

\[
\sup_{p \in P} \left| \frac{1}{n} \sum_{i=1}^{n} \{ d(Y_i|q) - d(Y_i|p) \} \right| \to 0 \quad \text{with probability 1}.
\]

(B.10)

A similar argument shows that, as \( n \to \infty \),

\[
\sup_{p \in P} \left| E \left\{ d(Y_i|q) - d(Y_i|p) \right\} \right| \to 0.
\]

(B.11)

Results (B.10), (B.11) and (B.7) imply that

\[
\sup_{p \in P} \left| \delta(p) \right| \to 0 \quad \text{with probability 1}.
\]

(B.12)

Let \( F_X \) denote the true distribution of \( X \). For any \( p \in P \), let \( f_Y(y|p) = \sum_{i=1}^{\nu} \pi_i f_Z(y - a_i) + \pi_{\nu+1} \sum_{j=1}^{r} \theta_j f_Z(y - x_j) \) denote the approximation to the density \( f_Y \) of \( Y \), which is obtained by approximating \( F_X \) by the distribution with atoms \( a_1, \ldots, a_\nu, x_1, \ldots, x_r \) having probability \( \pi_1, \ldots, \pi_\nu, \pi_{\nu+1} \theta_1, \ldots, \pi_{\nu+1} \theta_r \), respectively. Define the negative entropy, \( e_Y \), of the distribution of \( Y \) to be \( e_Y = \int (\log f_Y) f_Y \). Now,

\[
E\{d(Y|p)\} = \int \log \{ f_Y(y|p) \} f_Y(y) \, dy,
\]

and the Kullback-Leibler divergence of the distribution with density \( f_Y(\cdot|p) \) from that with density \( f_Y \) is given by

\[
d\{f_Y, f_Y(\cdot|p)\} = \int \log \left\{ \frac{f_Y}{f_Y(\cdot|p)} \right\} f_Y \geq 0.
\]

Hence, \( e_Y - E\{d(Y|p)\} = d\{f_Y, f_Y(\cdot|p)\} \), and so by (B.12), and with probability 1,

\[
e_Y - \frac{1}{n} \sum_{i=1}^{n} d(Y_i|p) = d\{f_Y, f_Y(\cdot|p)\} + O(1), \quad \text{uniformly in} \quad p \in P.
\]

(B.13)

Let \( \hat{p} \) denote a random element of \( P \) which gives a global maximum of \( \sum_i d(Y_i|p) \) over \( p \in P \). Knowing the distribution \( f_X \) it is straightforward to construct a vector

\[
\overline{p} = (\pi_1, \ldots, \pi_\nu, \pi_{\nu+1} \theta_1, \ldots, \pi_{\nu+1} \theta_r)^T \in P
\]

for which

\[
d\{f_Y, f_Y(\cdot|\overline{p})\} \to d(f_Y, f_Y) = 0.
\]

(B.14)
One can simply choose $\pi_l = P(X = a_l)$ for $1 \leq l \leq \nu$, $\pi_{\nu+1} = 1 - \sum_{l=1}^{\nu} \pi_l$, and $\bar{\theta}_j$ to equal the probability that $X_c \in [(j-1)h, jh)$ for $j \geq 1$. By definition of $\hat{\theta}$, $\sum_i d(Y_i|\hat{\theta}) \geq \sum_i d(Y_i|\bar{\theta})$. Therefore, by (B.13) and (B.14), and with probability 1,

$$d\{f_Y, f_Y(\cdot|\bar{\theta})\} + o(1) = e_Y - \frac{1}{n} \sum_{i=1}^{n} d(Y_i|\bar{\theta}) \leq e_Y - \frac{1}{n} \sum_{i=1}^{n} d(Y_i|\hat{\theta}) = d\{f_Y, f_Y(\cdot|\hat{\theta})\} + o(1) \to 0.$$  \hspace{1cm} (B.15)

However $d\{f_Y, f_Y(\cdot|\hat{\theta})\}$, being a Kullback-Leibler divergence, is nonnegative, and therefore (B.15) implies that $d\{f_Y, f_Y(\cdot|\hat{\theta})\} \to 0$ with probability 1. Hence, with probability 1, the distribution with density $f_Y(\cdot|\hat{\theta})$ converges, almost surely, to the distribution with density $f_Y$. The theorem then follows from this result and the assumption (R6).

**Proof of Theorem 2.**

The proof is very similar with the proof of Theorem 1. Define $\hat{\theta}^{PS}$ as a global maximum of the penalized log-likelihood $\sum_i d(Y_i|\theta) - \lambda P(\theta)$ over $\theta \in \mathcal{P}$. By the definition of $\hat{\theta}^{PS}$, it is clear that

$$\sum_{i=1}^{n} d(Y_i|\hat{\theta}^{PS}) \geq \sum_{i=1}^{n} d(Y_i|\bar{\theta}) + \lambda \left\{ P(\hat{\theta}^{PS}) - P(\bar{\theta}) \right\},$$

where $\bar{\theta}$ is chosen in the same way as in the previous proof. Then, from (B.13) and (B.14), and with probability 1,

$$d\{f_Y, f_Y(\cdot|\hat{\theta}^{PS})\} + o(1) = e_Y - \frac{1}{n} \sum_{i=1}^{n} d(Y_i|\hat{\theta}^{PS}) \leq e_Y - \frac{1}{n} \sum_{i=1}^{n} d(Y_i|\bar{\theta}) - \frac{\lambda}{n} \left\{ P(\hat{\theta}^{PS}) - P(\bar{\theta}) \right\} = d\{f_Y, f_Y(\cdot|\bar{\theta})\} + o(1) + O\left(\frac{\lambda}{n}\right) \to 0.$$  \hspace{1cm} (B.16)

since $P(\cdot)$ is asymptotically bounded and $\lambda = o(n)$. Because $d\{f_Y, f_Y(\cdot|\hat{\theta}^{PS})\}$ is nonnegative, it converges to 0 with probability 1 by (B.16). Hence the distribution with density $f_Y(\cdot|\hat{\theta}^{PS})$ converges to the true distribution $f_Y$ almost everywhere, with probability 1. Hence, from the fact (R6), we conclude that the distribution estimator of $X$ characterized by $\hat{\theta}^{PS}$ converges to the true distribution with probability 1.
References