ON ESTIMATING MIXING DENSITIES IN EXPONENTIAL FAMILY MODELS FOR DISCRETE VARIABLES II

by

Wei-Liem Loh and Cun-Hui Zhang
Purdue University Rutgers University

Technical Report #95-1

Department of Statistics
Purdue University

January 1995

ON ESTIMATING MIXING DENSITIES IN EXPONENTIAL FAMILY MODELS FOR DISCRETE VARIABLES II

BY WEI-LIEM LOH[1] AND CUN-HUI ZHANG[2]

Purdue University and Rutgers University

This paper is concerned with estimating a mixing density g using a random sample from the mixture distribution $f(x) = \int f(x|\theta)g(\theta)d\theta$ where $f(.|\theta)$ is a known discrete exponential family of density functions. Recently two techniques for estimating g have been proposed. The first uses Fourier analysis and the method of kernels and the second uses orthogonal polynomials. It is known that the first technique is capable of yielding estimators that achieve (or almost achieve) the minimax convergence rate. We show that this is true for the technique based on orthogonal polynomials as well. The practical implementation of these estimators is also addressed. Computer experiments indicate that the kernel estimators give somewhat disappointing finite sample performances. However the orthogonal polynomial estimators appear to do much better. To improve on the finite sample performance of the orthogonal polynomial estimators, a way of estimating the optimal truncation parameter is proposed. The resultant estimators retain the convergence rates of the previous estimators and a Monte Carlo finite sample study reveals that they perform well relative to the ones based on the optimal truncation parameter.

Key words: Discrete exponential family, mixing density, orthogonal polynomials, rate of convergence.

1 Introduction

Let $X_1, \ldots, X_n$ be independent observations from a mixture distribution with probability law

$$f(x; g) = \int_0^{\theta^*} f(x|\theta)g(\theta)d\theta, \tag{1}$$

where g is a mixing probability density function on $(0, \theta^*)$ and $f(.|\theta)$ is a known parametric family of probability density functions with respect to a σ-finite measure ν. In particular we assume that

$$f(x|\theta) = C(\theta)q(x)\theta^x, \quad \forall x = 0, 1, 2, \cdots, \tag{2}$$

where $0 < \theta < \theta^* \leq \infty$, $q(x) > 0$ whenever $x = 0, 1, 2, \ldots$ and ν is the counting measure on the set of nonnegative integers. In this paper we are concerned with the estimation of g using the random sample $X_1, \ldots, X_n$.

Over the last few years, there has been a great deal of interest in the above problem and other related mixture problems. Important advances have been made on the deconvolution problem by Devroye and Wise (1979), Carroll and Hall (1988), Zhang (1990), Fan (1991) and many others using Fourier techniques. In particular kernel estimators have been obtained which achieve the minimax convergence rate.

In the context of mixtures of discrete exponential families, Tucker (1963) considered the estimation of the mixing distribution of a Poisson mixture via the method of moments and Simar (1976)

[1]Research supported in part by NSA Grant MDA 904-93-3011.
[2]Research supported in part by NSF Grant 89-16180 and ARO Grant DAAL03-91-G-0045.

approached the same problem using maximum likelihood. Rolph (1968), Meeden (1972) and Datta (1991) used Bayesian methods to construct consistent estimators for the mixing distribution.

Quite recently, two techniques for the estimation of the mixing density g, as given in (1), have been proposed. The first was proposed by Zhang (1992) which uses Fourier analysis and the method of kernels. The second was proposed by Walter and Hamedani (1989), (1991) which uses orthogonal polynomials. It has been shown by Zhang (1992) and Loh and Zhang (1994) that the first technique is capable of yielding estimators that achieve (or almost achieve) the minimax convergence rate with respect to integrated mean squared error over various smoothness classes of mixing density functions.

The rest of this paper is organized as follows. We shall first very briefly review the kernel mixing density estimators and their properties in Section 2. In Section 3 we shall show that the technique based on orthogonal polynomials is also capable of yielding mixing density estimators that achieve (or almost achieve) the minimax convergence rate with respect to integrated mean squared error over various nonparametric classes of mixing density functions. However even with this property, the minimax convergence rates of these estimators are logarithmic (not polynomial). This leaves us with the important question as to how well can these estimators actually perform in practice.

Section 4 addresses the issue of the finite sample performances as well as the practical implementation of these estimators. Computer experiments indicate that the kernel mixing density estimators (for the particular kernel used here) give somewhat disappointing finite sample performances. On the other hand, the orthogonal polynomial mixing density estimators appear to do much better. To improve upon the finite sample performance of the orthogonal polynomial mixing density estimators further, a way of estimating the optimal truncation parameter is proposed in Section 5. The resultant estimators retain the convergence rates of the previous estimators and a Monte Carlo finite sample study reveals that they perform well relative to the ones based on the optimal truncation parameter.

All proofs in this paper have been deferred to the Appendix. Finally we shall denote by $P = P_g$ and $E = E_g$ the probability and expectation corresponding to g respectively, by $h^{(j)}$ the jth derivative (if it exists) of any function h with $h^{(0)} = h$, and the weighted L^p-norm of any measurable function h by $||h||_{w,p} = (\int |h(y)|^p w(y) dy)^{1/p}$, $\forall 1 \leq p < \infty$. If $w(y) \equiv 1$, we denote $||.||_{w,p}$ by $||.||_p$.

2 Kernel mixing density estimators

This section treats the case $\theta^* < \infty$ and, for completeness, gives a brief review of the kernel mixing density estimators that we are concerned with here. We refer the reader to Loh and Zhang (1994) for the proofs and a more detailed discussion of these estimators.

Let $k : R \to R$ be a symmetric function satisfying

$$\int_{-\infty}^{\infty} k(y)dy = 1, \qquad k^*(t) = 0, \quad \forall |t| > 1,$$

(3)
$$\int_{-\infty}^{\infty} y^j k(y)dy = 0, \quad \forall 1 \leq j < \alpha_0,$$

and

(4)
$$\int_{-\infty}^{\infty} |y^{\alpha_0} k(y)| \, dy < \infty,$$

for some positive number α_0, where k^* denotes the Fourier transform of k, that is $k^*(t) =$

$\int_{-\infty}^{\infty} e^{ity}k(y)dy$. Define

(5) $$K_n(x,\theta) = \frac{I\{0 \le x \le d_n\}}{2\pi q(x)x!}\int_{-c_n}^{c_n} \mathcal{R}\{(it)^x e^{-it\theta}\}k^*(t/c_n)dt,$$

where c_n and d_n are positive constants tending to ∞, $I\{.\}$ denotes the indicator function and $\mathcal{R}(z)$ is the real part of the complex number z. Observing that

(6) $$E_g K_n(X_1,\theta) - C(\theta)g(\theta) \to 0, \quad \forall -\infty < \theta < \infty,$$

as $(c_n, d_n) \to (\infty,\infty)$ along a suitable path, Loh and Zhang (1994) proposed estimating $g(\theta)$ by the kernel mixing density estimator

(7) $$\hat{g}_{K,n}(\theta) = n^{-1}\sum_{j=1}^{n}\{K_n(X_j,\theta)/C(\theta)\}I\{0 < \theta < a_n\}, \quad \forall 0 < \theta < \theta^*.$$

where a_n, c_n, and d_n are constants satisfying

(8) $$c_n + \max_{1\le x\le d_n}\log(1/q(x)) = \beta_0\log n, \qquad c_n = (\theta^* e)^{-1}(d_n - \beta_1\log c_n),$$

and

(9) $$a_n = \begin{cases} \theta^* & \text{if } C(\theta^*) > 0, \\ \theta^* - a^*/c_n & \text{if } C(\theta^*) = 0, \end{cases}$$

with absolute (independent of n) constants $0 < \beta_0 < 1/2$, $\beta_1 > 0$, and $0 < a^* < \infty$. The performance of these estimators are investigated with respect to the following smoothness classes of mixing density functions. Let w be a measurable function on $(0,\theta^*)$ with $||w||_1$ finite. For $\alpha > 0$ we define $\mathcal{G}_{\alpha,\theta^*}(w,M)$ to be the set of all probability density functions g on $(0,\theta^*)$ such that

(10) $$||g^{(\alpha')}(.) - g^{(\alpha')}(.+\delta)||_{w,2} < M|\delta|^{\alpha''}, \quad \forall\delta,$$

where α' is the integer with $0 < \alpha'' = \alpha - \alpha' \le 1$, and M is a constant such that $\mathcal{G}_{\alpha,\theta^*}(w,M)$ is nonempty.

We further assume that there exist constants $\gamma \ge 0$, C_1^*, C_2^*, and C_3^* such that

(11) $$\sup_{0<\theta<\theta^*}(\theta^*-\theta)^\gamma/C(\theta) < C_1^*,$$

(12) $$\sup_{0<\theta<\theta^*}(\theta^*-\theta)^j|C^{(j)}(\theta)|/\{C(\theta)j!\} < C_2^*, \quad \forall 0 \le j \le \rho',$$

and

(13) $$|C^{(\rho')}(\theta+\delta) - C^{(\rho')}(\theta)| < C_3^*\delta^{\rho''}, \ 0 < \theta < \theta+\delta < \theta^*,$$

where ρ' is a nonnegative integer with $0 < \rho'' = \rho - \rho' \le 1$.

Theorem 1 below shows that the kernel mixing density estimators $\hat{g}_{K,n}$ achieve (or almost achieve) the minimax convergence rate with respect to $\mathcal{G}_{\alpha,\theta^*}(w,M)$ under reasonably mild conditions.

Theorem 1 *Suppose $\alpha > 0$ and that (11)-(13) hold with $\gamma \ge 0$ and $\rho = \alpha + \gamma$. Let $\hat{g}_{K,n}$ be given by (7) with the kernel $K_n(x,\theta)$ in (5) such that $\alpha_0 \ge \alpha+\gamma$ in (4). Let (8) and (9) hold with $\beta_1 \ge \alpha + \gamma$. Then if*

$$q(x)\gamma_0\gamma_1^x(x!)^\beta \ge 1, \quad \forall x \ge 0,$$

for some constants γ_0, γ_1, and β, we have

$$\sup_{g\in\mathcal{G}_{\alpha,\theta^*}(w,M)} E_g||\hat{g}_{K,n} - g||_{w,2} = \begin{cases} O(1)(1/\log n)^{\alpha} & \text{if } \beta = 0, \\ O(1)(\log\log n/\log n)^{\alpha} & \text{if } 0 < \beta < \infty. \end{cases}$$

Furthermore

$$\liminf_{n\to\infty}(\log n)^{\alpha}\inf_{\hat{g}_n}\ \sup\{E_g||\hat{g}_n - g||_2 : g \in \mathcal{G}_{\alpha,\theta^*}(1, M)\} > 0,$$

where the infimum runs over all possible estimators $\hat{g}_n$ based on $X_1, \ldots, X_n$ and $\mathcal{G}_{\alpha,\theta^*}(1, M)$ is given by (10) with $w(\theta) = I\{0 < \theta < \theta^*\}$.

3 Orthogonal polynomial mixing density estimators

In this section we introduce the class of orthogonal polynomial mixing density estimators that we are concerned with and also establish upper and lower bounds for their convergence rates with respect to various nonparametric classes of mixing density functions. Let $C : (0, \theta^*) \to R^+$ be as in (2) and $w : (0, \theta^*) \to R^+$ be a measurable function such that $||C^2/w||_1 < \infty$. Let $\{p_{w_0,j}\}_{j=0}^{\infty}$ be a sequence of orthogonal polynomials on $(0, \theta^*)$ with weight function

$$w_0(\theta) = C^2(\theta)/w(\theta). \tag{14}$$

In particular, we assume that these polynomials are normalized so that

$$p_{w_0,j}(\theta) = \sum_{x=0}^{j} k_{w_0,j,x}\theta^x, \tag{15}$$

with $k_{w_0,j,j} > 0$ for all $j \geq 0$, and $\int_0^{\theta^*} p_{w_0,i}(\theta)p_{w_0,j}(\theta)w_0(\theta)d\theta = \delta_{ij}$, where δ_{ij} denotes the Kronecker delta. We further assume that $\{p_{w_0,j}\}_{j=0}^{\infty}$ is complete with respect to $||.||_{w_0,2}$. Note that this is always true if $\theta^* < \infty$ [see for example Szegö (1975) page 40]. Next define

$$\lambda_{w_0,j}(x) = \begin{cases} k_{w_0,j,x}/q(x) & \text{if } 0 \leq x \leq j, \\ 0 & \text{otherwise.} \end{cases}$$

We write

$$h(\theta) = w(\theta)g(\theta)/C(\theta), \quad \forall 0 < \theta < \theta^*, \tag{16}$$

and assume that the mixing density g satisfies $||g||_{w,2} = ||h||_{w_0,2} < \infty$. Then h has the formal orthogonal polynomial series expansion $h(\theta) \sim \sum_{j=0}^{\infty} h_{w_0,j}p_{w_0,j}(\theta)$, where

$$h_{w_0,j} = \int_0^{\theta^*} h(\theta)p_{w_0,j}(\theta)w_0(\theta)d\theta, \quad \forall j = 0, 1, 2, \cdots. \tag{17}$$

Observing that

$$E_g\lambda_{w_0,j}(X_1) = \sum_{x=0}^{\infty} f(x; g)\lambda_{w_0,j}(x) = h_{w_0,j}, \quad \forall j = 0, 1, 2, \cdots,$$

we estimate $h_{w_0,j}$ by $\hat{h}_{w_0,j} = n^{-1}\sum_{i=1}^{n}\lambda_{w_0,j}(X_i)$ and $g(\theta)$ by the orthogonal polynomial mixing density estimator

$$\hat{g}_{OP,n}(\theta) = [C(\theta)/w(\theta)]\sum_{j=0}^{m_n}\hat{h}_{w_0,j}p_{w_0,j}(\theta), \quad \forall 0 < \theta < \theta^*, \tag{18}$$

where m_n is a positive constant (truncation parameter) which tends to ∞ as $n \to \infty$. The following proposition gives an upper bound on the convergence rate of $\hat{g}_{OP,n}$.

Proposition 1 *Suppose* $||C^2/w||_1 < \infty$ *and* $||g||_{w,2} < \infty$. *Let* $\hat{g}_{OP,n}$ *be as in (18). Then*

$$E_g||\hat{g}_{OP,n} - g||_{w,2} \le \{n^{-1} \sum_{j=0}^{m_n} \max_{0\le x\le j} [k_{w_0,j,x}/q(x)]^2 + \sum_{j=m_n+1}^{\infty} h^2_{w_0,j}\}^{1/2},$$

with $k_{w_0,j,x}$ *and* $h_{w_0,j}$ *as in (15) and (17) respectively.*

REMARK. The motivation for (18) originates from Walter and Hamedani (1989) who proposed a similar class of estimators. They also obtained a result analogous to Proposition 1.

We now study the performance of the estimators $\hat{g}_{OP,n}$ with respect to the following nonparametric classes of mixing density functions. For positive constants α, M and $m = 1, 2, \ldots$, we define $\mathcal{G}(\alpha, m, M, w_0)$ to be the set of all probability density functions g on $(0, \theta^*)$ such that $||g||_{w,2} < \infty$ and $\sum_{j=m}^{\infty} j^{2\alpha} h^2_{w_0,j} < M$ with $h_{w_0,j}$ as in (17). We note that this class implicitly depends on the discrete exponential family of interest, in particular on $C(\theta)$. This ellipsoidal class is chosen mainly for reasons of mathematical tractability. However ellipsoid conditions can amount to the imposition of smoothness and integrability requirements, see for example Johnstone and Silverman (1990) page 258. In our case, we have the following characterization.

Proposition 2 *Let* $m \ge 1$ *and* $\{p_{w_0,j}\}_{j=0}^{\infty}$ *be as in (15). Suppose there exist constants* $\nu_{j,m}$, $j \ge m$ *and another sequence of (normalized) complete orthogonal polynomials* $\{p_{w_1,j}\}_{j=0}^{\infty}$ *with weight function* w_1 *such that*

$$[p_{w_1,j}(\theta)w_1(\theta)]^{(m)} = (-1)^m \nu_{j+m,m} p_{w_0,j+m}(\theta) w_0(\theta), \quad \forall j \ge 0, \tag{19}$$

and

$$\alpha_1 < \inf_{j\ge m} |\nu_{j,m}|/j^{\alpha} \le \sup_{j\ge m} |\nu_{j,m}|/j^{\alpha} < \alpha_2, \tag{20}$$

where α, α_1 *and* α_2 *are positive constants. Then if* h *is a measurable function on* $(0, \theta^*)$ *such that* $h^{(m)}$ *exists,*

$$0 = \lim_{\theta\to 0+} h^{(m-i)}(\theta)[p_{w_1,j}(\theta)w_1(\theta)]^{(i-1)} = \lim_{\theta\to\theta^*-} h^{(m-i)}(\theta)[p_{w_1,j}(\theta)w_1(\theta)]^{(i-1)} \tag{21}$$

whenever $0 < i \le m$, $j \ge 0$, *and* $||h^{(m)}||_{w_1,2} < \infty$, *we have*

$$\alpha_1(\sum_{j=m}^{\infty} j^{2\alpha} h^2_{w_0,j})^{1/2} \le ||h^{(m)}||_{w_1,2} \le \alpha_2(\sum_{j=m}^{\infty} j^{2\alpha} h^2_{w_0,j})^{1/2}, \tag{22}$$

where $h_{w_0,j}$ *is defined as in (17). (19) and (20) are satisfied by the classical orthogonal polynomials of Laguerre and Jacobi with* $\alpha = m/2$, m *respectively.*

For the rest of this section, we shall assume that M is sufficiently large so that $\mathcal{G}(\alpha, m, M, w_0)$ is nonempty. The next two theorems and their corollaries establish upper bounds on the convergence rate of $\hat{g}_{OP,n}$ over the class of mixing densities $\mathcal{G}(\alpha, m, M, w_0)$.

Theorem 2 *Suppose* $||C^2/w||_1 < \infty$. *Let* $\hat{g}_{OP,n}$ *be as in (18) and*

$$\max_{0\le x\le j\le m_n} \log(|k_{w_0,j,x}|/q(x)) \le \beta_0 \log n, \tag{23}$$

for some constant $0 < \beta_0 < 1/2$. *Then*

$$\sup\{E_g||\hat{g}_{OP,n} - g||_{w,2} : g \in \mathcal{G}(\alpha, m, M, w_0)\} = O(1)(m_n^{-\alpha} + m_n^{1/2} n^{(2\beta_0-1)/2}).$$

Corollary 1 *Suppose* $\theta^* = \infty$, $w(\theta) = \theta^{-\beta}C^2(\theta)e^{\theta}$ *and* $w_0(\theta) = \theta^{\beta}e^{-\theta}$ *with* $\beta > -1$. *Let* $\{p_{w_0,j}\}_{j=0}^{\infty}$ *be the sequence of (normalized) Laguerre polynomials on* $(0, \infty)$ *with weight function* w_0, $\hat{g}_{OP,n}$ *as in (18) and*

$$q(x)\gamma_0\gamma_1^x(x!) > 1, \quad \forall x \ge 0,$$

for constants γ_0 *and* $\gamma_1 \ge 1$. *Then by choosing* $m_n = \delta \log n$ *with* $0 < \delta \le \beta_0/\log(2\gamma_1)$ *and* $0 < \beta_0 < 1/2$, *we have*

$$\sup\{E_g||\hat{g}_{OP,n} - g||_{w,2} : g \in \mathcal{G}(\alpha, m, M, w_0)\} = O(1)(1/\log n)^{\alpha}.$$

Theorem 3 is a specialization of Theorem 2 which proves to be useful when $\theta^* < \infty$.

Theorem 3 *Let* $\hat{g}_{OP,n}$ *be as in (18) and that for some constant* $\zeta > 1$,

$$\max_{0\le x\le j} k^2_{w_0,j,x} < \zeta^{2j}, \quad \forall j \ge 0. \tag{24}$$

Suppose further that

$$\max_{0\le x\le m_n} \log(1/q(x)) + m_n \log\zeta \le \beta_0 \log n, \tag{25}$$

with constant $0 < \beta_0 < 1/2$. *Then*

$$\sup\{E_g||\hat{g}_{OP,n} - g||_{w,2} : g \in \mathcal{G}(\alpha, m, M, w_0)\} = O(m_n^{-\alpha}).$$

Corollary 2 *Let* $\hat{g}_{OP,n}$ *be as in (18) and that (24) holds for some constant* $\zeta > 1$. *Suppose*

$$q(x)\gamma_0\gamma_1^x(x!)^{\gamma} > 1, \quad \forall x \ge 0, \tag{26}$$

for constants γ_0, $\gamma_1 \ge 1$ *and* γ. *Then*

(a) if $\gamma = 0$, *by choosing* $m_n = \delta\log n$ *with* $0 < \delta \le \beta_0/\log(\gamma_1\zeta)$ *and* $0 < \beta_0 < 1/2$, *we have*

$$\sup\{E_g||\hat{g}_{OP,n} - g||_{w,2} : g \in \mathcal{G}(\alpha, m, M, w_0)\} = O(1)(1/\log n)^{\alpha},$$

(b) if $0 < \gamma < \infty$, *by choosing* $m_n = \delta \log n/\log\log n$ *with* $0 < \delta \le \beta_0/\gamma$ *and* $0 < \beta_0 < 1/2$, *we have*

$$\sup\{E_g||\hat{g}_{OP,n} - g||_{w,2} : g \in \mathcal{G}(\alpha, m, M, w_0)\} = O(1)(\log\log n/\log n)^{\alpha}.$$

REMARK 1. The negative binomial and Poisson mixtures satisfy (26) with $\gamma = 0$ and 1 respectively.

REMARK 2. The classical orthogonal polynomials of Jacobi satisfy (24).

The next theorem complements the above results by establishing lower bounds on the minimax convergence rate over the class of mixing densities $\mathcal{G}(\alpha, m, M, w_0)$ under the condition that (19) and (20) hold.

Theorem 4 *Let $w : (0,\theta^*) \to R^+$ be a measurable function such that $||w||_1 < \infty$ and $||w_0||_1 < \infty$ with w_0 as in (14) and $\{p_{w_0,j}\}_{j=0}^{\infty}$ be a sequence of (normalized) orthogonal polynomials with weight function w_0 such that (19) and (20) are satisfied. Suppose there exists an open interval such that w is strictly positive and m times continuously differentiable. Then for sufficiently large M,*

$$\lim_{n\to\infty} (\log n)^m \inf_{\hat{g}_n} \sup\{E_g||\hat{g}_n - g||_{w,2} : g \in \mathcal{G}(\alpha, m, M, w_0)\} > 0,$$

where the infimum runs over all possible estimators $\hat{g}_n$ based on $X_1, \ldots, X_n$.

We close this section with the following consequence of Corollary 2, Remark 2 and Theorem 4. Suppose $\theta^* < \infty$ and that there exist constants $\beta_1 > -1$, $\beta_2 > -1$, $\gamma_0 > 0$, $\gamma_1 \geq 1$ and $\gamma \geq 0$ such that

$$w(\theta) = C^2(\theta)\theta^{-\beta_1}(\theta^* - \theta)^{-\beta_2}, \quad \forall 0 < \theta < \theta^*,$$

and $q(x)\gamma_0\gamma_1^x(x!)^\gamma > 1$, for all $x \geq 0$. Then

(a) if $\gamma = 0$, the minimax convergence rate with respect to $||.||_{w,2}$ loss is $(1/\log n)^m$ for mixing densities g in the class $\mathcal{G}(m, m, M, w_0)$ where w_0 is as in (14). This rate is attained by the mixing density estimators $\hat{g}_{OP,n}$ of Corollary 2.

(b) if $0 < \gamma < \infty$, the convergence rate [namely $(\log\log n/\log n)^m$] of the estimators of Corollary 2 almost achieve the lower bound of $(1/\log n)^m$ obtained in Theorem 4 for mixing densities within the class $\mathcal{G}(m, m, M, w_0)$.

4 Finite sample performance

A key consequence of the results of Sections 2 and 3 is that both the kernel and orthogonal polynomial mixing density estimators, i.e. $\hat{g}_{K,n}$ and $\hat{g}_{OP,n}$ respectively, are capable of achieving (or almost achieving) the minimax rate of convergence. However even with this property, the minimax convergence rate of these estimators is logarithmic (not polynomial), This leads us to the following problem. Typically how large must a sample be in order that the desired asymptotics of these estimators (as described in the previous two sections) can take effect.

4.1 Kernel mixing density estimators

In order to gauge typically how well the kernel mixing density estimators perform in practice, we focus on the problem of estimating the mixing density g of a negative binomial mixture with $\theta^* = 1$ and $C(\theta) = 1 - \theta$ with respect to integrated squared error, that is $||\hat{g}_n - g||_2^2$. To construct the kernel mixing density estimator $\hat{g}_{K,n}$, we take

$$k(y) = \frac{6}{\pi}|\frac{2}{y}\sin(\frac{y}{4})|^4, \quad \forall -\infty < y < \infty.$$

Our motivation for such a choice of k is its relative simplicity and that (3) and (4) hold with $\alpha_0 = 2$. We observe from (6) and (7) that an upper bound on the finite sample performance of $\hat{g}_{K,n}$ can be obtained by investigating how close

$$(27) \qquad ||E_g[K_n(X_1,\theta)/C(\theta)]I\{0 < \theta < a_n\} - g(\theta)||_2^2$$

is to 0. In this case we take $g(\theta) = I\{0 < \theta < 1\}$ and use $ERR_n = (1/10)\sum_{i=1}^{10}\{E_g[K_n(X_1, 0.1i - 0.05)/C(0.1i - 0.05)] - 1\}^2$ as an approximation to (27).

REMARK. The reason for such a choice of g is that we feel that the uniform distribution is arguably one of the distributions that any reasonable estimation procedure should be able to estimate adequately well.

Computations show that in order to have $ERR_n \approx 0.1$, we need $c_n \approx 17$. Since $c_n \leq (1/2)\log n$, this implies that the sample size n must be astronomically large and is quite impossible to obtain in practice.

This presents a disappointing setback for the practical implementation of $\hat{g}_{K,n}$. However it should be noted that this can be due to a possibly inappropriate choice of the kernel k and that it does not eliminate the possibilty that there exist other kernels which give dramatically better results.

4.2 Orthogonal polynomial mixing density estimators

We observe that the integrated mean squared error of the orthogonal polynomial mixing density estimators has a simple closed form expression. In particular, we observe as in (31) that

$$\begin{aligned} & E_g \int_0^{\theta^*} [\hat{g}_{OP,n}(\theta) - g(\theta)]^2 w(\theta) d\theta \\ (28) \qquad &= \int_0^{\theta^*} g^2(\theta) w(\theta) d\theta + n^{-1} \sum_{j=0}^{m_n} \{E_g \lambda^2_{w_0,j}(X_1) - (n+1)[E_g \lambda_{w_0,j}(X_1)]^2\}. \end{aligned}$$

The right hand side of (28) enables us to compute the integrated mean squared error of $\hat{g}_{OP,n}$ in any given situation. We illustrate this below with two examples.

EXAMPLE 3. This example deals with the problem of estimating a mixing density g of a negative binomial mixture with $\theta^* = 1$ and $C(\theta) = 1 - \theta$ using integrated squared error loss. In this case the orthogonal polynomial mixing density estimators are given as in (18) where $\{p_{w_0,j}\}_{j=0}^{\infty}$ corresponds to the Jacobi polynomials with weight function $w_0(\theta) = (1-\theta)^2$, $\forall 0 < \theta < 1$.

Tables 1, 2 and 3 give the integrated mean squared error of $\hat{g}_{OP,n}$ for sample sizes $n = 1000$, 10000 and 100000 as well as for truncation parameters $0 \leq m_n \leq 4$.

TABLE 1. $g(\theta) = 1$					
	truncation parameter m_n				
sample size n	0	1	2	3	4
1000	0.251	0.128	0.330	5.186	110.752
10000	0.250	0.113	0.089	0.555	11.100
100000	0.250	0.111	0.065	0.091	1.135
TABLE 2. $g(\theta) = (\pi/2)\sin(\pi\theta)$					
	truncation parameter m_n				
sample size n	0	1	2	3	4
1000	0.48445	0.02041	0.32504	6.16375	128.20208
10000	0.48378	0.00333	0.03352	0.61638	12.82021
100000	0.48371	0.00162	0.00437	0.06164	1.28202
TABLE 3. $g(\theta) = \exp(\theta)/(e-1)$					
	truncation parameter m_n				
sample size n	0	1	2	3	4
1000	0.558	0.291	0.431	5.701	126.007
10000	0.558	0.277	0.184	0.660	12.663
100000	0.558	0.275	0.159	0.156	1.329

EXAMPLE 4. This example deals with the estimation of the mixing density g of a Poisson mixture with $\theta^* = \infty$ using integrated weighted squared error loss $\|\hat{g}_n - g\|_{w,2}^2$ where $w(\theta) = e^{-\theta}$, $\forall\theta > 0$. In this case $\{p_{w_0,j}\}_{j=0}^{\infty}$ corresponds to the Laguerre polynomials with weight function $w_0(\theta) = e^{-\theta}$. Table 4 gives the integrated mean squared error of the estimator $\hat{g}_{OP,n}$ when $g(\theta) = e^{-\theta}$, $\forall\theta > 0$ for sample sizes $n = 1000$, 10000 and 100000 as well as for $0 \le m_n \le 6$.

TABLE 4. $g(\theta) = \exp(-\theta)$

	truncation parameter m_n						
sample size n	0	1	2	3	4	5	6
500	0.08383	0.02271	0.01030	0.01426	0.03335	0.08570	0.22609
1000	0.08358	0.02177	0.00775	0.00778	0.01684	0.04289	0.11305
10000	0.08336	0.02093	0.00546	0.00195	0.00198	0.00436	0.01132
100000	0.08334	0.02084	0.00523	0.00137	0.00049	0.00051	0.00115

REMARK. Examples 3 and 4 (plus other unreported ones) indicate that $\hat{g}_{OP,n}$ perform well for sample sizes $n \ge 1000$ as long as h, defined as in (16), can be reasonably approximated by a low degree polynomial and that the optimal truncation parameter is used.

5 Estimating the optimal truncation parameter

In this section, a way is proposed to estimate the optimal truncation parameter m_n^* for the orthogonal polynomial mixing density estimator $\hat{g}_{OP,n}$, given as in (18), where m_n^* is defined to be the value of the truncation parameter m_n which minimizes $E_g\|\hat{g}_{OP,n} - g\|_{w,2}$. We write

$$t_{n,j} = n^{-1}\{E_g\lambda_{w_0,j}^2(X_1) - (n+1)[E_g\lambda_{w_0,j}(X_1)]^2\}. \tag{29}$$

We observe from (28) that $\sum_{j=0}^{m_n^*} t_{n,j} \le \sum_{j=0}^{m} t_{n,j}$, for all $m \ge 0$. This implies that m_n^* can be determined if the sign of $\sum_{j=a}^{b} t_{n,j}$ is known for each $a \le b$. Let $\hat{t}_{n,j}$ be the unbiased estimator of $t_{n,j}$ based on $X_1, \ldots, X_n$, $\hat{t}_{n,i,j} = \sum_{l=i}^{j} \hat{t}_{n,l}$, $\forall 0 \le i \le j$ and $\hat{\sigma}^2(\hat{t}_{n,i,j})$ be the unbiased estimator of the variance of $\hat{t}_{n,i,j}$. Let $0 < \alpha^* < 1$ and B_n be the largest possible constant satisfying the inequalities

$$\max_{0 \le x \le j \le B_n} \log(|k_{w_0,j,x}|/q(x)) \le \beta_0 \log n, \qquad B_n \le \beta_1 \log n, \tag{30}$$

for positive constants $\beta_0 < 1/2$ and β_1. Our algorithm for estimating m_n^* is as follows:

1. Set $\hat{m}_n^* = 0$ and $n_1 = n_2 = 1$.
2. Compute $Y = \hat{t}_{n,n_1,n_2} + z_{\alpha^*}\hat{\sigma}(\hat{t}_{n,n_1,n_2})$, where $\Phi(z_{\alpha^*}) = 1 - \alpha^*$ and Φ denotes the distribution function of the standard normal distribution.
3. CASE 1. If $Y < 0$ and $n_2 \le B_n$, set $\hat{m}_n^* = n_2$, $n_1 = n_2 + 1$ and then set $n_2 = n_1$. Let $Y = \hat{t}_{n,n_1,n_2} + z_{\alpha^*}\hat{\sigma}(\hat{t}_{n,n_1,n_2})$ and return to the start of Step 3.

   CASE 2. If $Y \ge 0$ and $n_2 \le B_n$, increase n_2 by 1, compute $Y = \hat{t}_{n,n_1,n_2} + z_{\alpha^*}\hat{\sigma}(\hat{t}_{n,n_1,n_2})$ and return to the beginning of Step 3.

   CASE 3. If $n_2 > B_n$, the estimate of the optimal truncation parameter m_n^* is given by $\hat{m}_n^*$.

REMARK 1. The above algorithm can be thought of as a successive sequence of hypotheses tests each at level α^* where the null hypothesis always has fewer terms than the alternative.

REMARK 2. The constant B_n can be chosen in the following manner. Under the conditions of Corollary 1, take $B_n = \beta_0 \log n/\log(2\gamma_1)$. Under the conditions of Corollary 2(a) and (b), we take $B_n = \beta_0 \log n/\log(\gamma_1\zeta)$ and $(\beta_0/\gamma)\log n/\log\log n$ respectively.

REMARK 3. The closer α^* is chosen to 0, the more likely it is that $\hat{m}_n^*$ will underestimate m_n^*. The previous section (see Tables 1 to 4) indicates that the risk of $\hat{g}_{OP,n}$ is asymmetrical about m_n^* and that there is a distinct possibility that the risk increases very dramatically with overestimation. As such we recommend that α^* be chosen to be 0.01, 0.05 or 0.10, which are in line with the usual values of α^* for classical hypotheses testing.

Let $\hat{g}^*_{OP,n}$ be as in (18) with m_n replaced by $\hat{m}_n^*$. The following theorem gives an upper bound to the convergence rate of $\hat{g}^*_{OP,n}$.

Theorem 5 *Let $\|C^2/w\|_1 < \infty$ and B_n be the largest possible constant satisfying (30). Then*

$$\sup\{E_g\|\hat{g}^*_{OP,n} - g\|_{w,2} : g \in \mathcal{G}(\alpha, m, M, w_0)\} = O(B_n^{-\alpha}).$$

REMARK. By choosing $B_n = m_n$ in Corollaries 1 and 2, we observe that the estimators $\hat{g}^*_{OP,n}$ essentially retain the convergence rates of $\hat{g}_{OP,n}$.

TABLE 5. $g(\theta) = 1$

		truncation parameter m_n				
sample size n	IMSE	0	1	2	3	4
1000	0.231	0.84	0.16	0.00	0.00	0.00
10000	0.110	0.00	0.98	0.02	0.00	0.00
100000	0.072	0.00	0.29	0.71	0.00	0.00

TABLE 6. $g(\theta) = (\pi/2)\sin(\pi\theta)$

		truncation parameter m_n				
sample size n	IMSE	0	1	2	3	4
1000	0.0552	0.08	0.92	0.00	0.00	0.00
10000	0.00320	0.00	1.00	0.00	0.00	0.00
100000	0.00158	0.00	1.00	0.00	0.00	0.00

TABLE 7. $g(\theta) = \exp(\theta)/(e-1)$

		truncation parameter m_n				
sample size n	IMSE	0	1	2	3	4
1000	0.360	0.30	0.70	0.00	0.00	0.00
10000	0.263	0.00	0.94	0.06	0.00	0.00
100000	0.142	0.00	0.00	1.00	0.00	0.00

EXAMPLE 3 CONTD. Here we have applied the above algorithm to Example 3. In particular the algorithm is used to determine $\hat{m}_n^*$ using $\alpha^* = 0.05$ and $B_n = \lfloor (1/2)\log n \rfloor$. For convenience we use 50 bootstrap replications to approximate each $\hat{\sigma}(\hat{t}_{n,n_1,n_2})$. The second column of Tables 5, 6 and 7 give the average value of

$$(1/10)\sum_{i=1}^{10}[\hat{g}^*_{OP,n}(0.1i - 0.05) - g(0.1i - 0.05)]^2,$$

for 100 independent replications of $X_1, \ldots, X_n$. These values approximate the integrated mean squared error (IMSE) of the mixing density estimator

$$\hat{g}^*_{OP,n}(\theta) = [C(\theta)/w(\theta)] \sum_{j=0}^{\hat{m}^*_n} \hat{h}_{w_0,j} p_{w_0,j}(\theta), \qquad \forall 0 < \theta < \theta^*.$$

We recall that in this case, we have $\theta^* = 1$, $w(\theta) = 1$ and $C(\theta) = (1-\theta)$. The remaining 5 columns of Tables 5, 6 and 7 give the proportion of the time $\hat{m}^*_n$ takes the values 0 to 4.

EXAMPLE 4 CONTD. The above algorithm is now applied to Example 4 with $\alpha^* = 0.05$, $B_n = 0.7 \log n$ and 50 bootstrap replications to approximate each $\hat{\sigma}(\hat{t}_{n,n_1,n_2})$. As in Example 3, the second column of Table 8 gives the average value of

$$(1/10) \sum_{i=1}^{500} e^{-(0.1i-0.05)} [\hat{g}^*_{OP,n}(0.1i - 0.05) - g(0.1i - 0.05)]^2,$$

for 100 independent replications of $X_1, \ldots, X_n$. These values approximate the integrated weighted mean squared error (IMSE) of the orthogonal polynomial mixing density estimator $\hat{g}^*_{OP,n}$, namely $E_g \|\hat{g}^*_{OP,n} - g\|^2_{w,2}$ with $w(\theta) = e^{-\theta}$, $\forall \theta > 0$. The remaining 5 columns of Table 8 give the proportion of the time $\hat{m}^*_n$ takes the values 0 to 4.

TABLE 8. $g(\theta) = \exp(-\theta)$

		truncation parameter m_n				
sample size n	IMSE	0	1	2	3	4
1000	0.0190	0.00	0.75	0.24	0.01	0.00
10000	0.00455	0.00	0.00	0.71	0.29	0.00
100000	0.00111	0.00	0.00	0.00	0.66	0.34

Both of the above Monte Carlo studies indicate that the risks of the orthogonal polynomial mixing density estimators $\hat{g}^*_{OP,n}$ compare well to the ones based on the optimal truncation parameter.

We conclude with the remark that in general the following two conditions do not hold: $\hat{g}^*_{OP,n}(\theta) \geq 0$, $\forall 0 < \theta < \theta^*$ and $\int_0^{\theta^*} \hat{g}^*_{OP,n}(\theta) d\theta = 1$. As such the accuracy of estimate $\hat{g}^*_{OP,n}$ can be further gauged by how close the above two conditions are to being satisfied.

6 Appendix

PROOF OF PROPOSITION 1. We observe that

$$\begin{aligned} E_g \{ \int_0^{\theta^*} [\hat{g}_{OP,n}(\theta) - g(\theta)]^2 w(\theta) d\theta \}^{1/2} &= E_g \{ \int_0^{\theta^*} [\sum_{j=0}^{m_n} \hat{h}_{w_0,j} p_{w_0,j}(\theta) - h(\theta)]^2 w_0(\theta) d\theta \}^{1/2} \\ &\leq \{ \sum_{j=0}^{m_n} E_g (\hat{h}_{w_0,j} - h_{w_0,j})^2 + \sum_{j=m_n+1}^{\infty} h^2_{w_0,j} \}^{1/2}. \qquad (31) \end{aligned}$$

The last inequality follows from Jensen's inequality and the completeness of $\{p_{w_0,j}\}_{j=0}^{\infty}$. Since $\hat{h}_{w_0,j} = n^{-1} \sum_{i=1}^{n} \lambda_{w_0,j}(X_i)$, the r.h.s. of (31) is bounded by

$$\{ n^{-1} \sum_{j=0}^{m_n} E_g [\lambda^2_{w_0,j}(X_1)] + \sum_{j=m_n+1}^{\infty} h^2_{w_0,j} \}^{1/2} \leq \{ n^{-1} \sum_{j=0}^{m_n} \max_{0 \leq x \leq j} [k_{w_0,j,x}/q(x)]^2 + \sum_{j=m_n+1}^{\infty} h^2_{w_0,j} \}^{1/2}.$$

This proves the proposition. □

PROOF OF PROPOSITION 2. We observe from (19), (21) and repeated integration by parts that

$$\begin{aligned}\int_0^{\theta^*} h^{(m)}(\theta)p_{w_1,j}(\theta)w_1(\theta)d\theta &= (-1)^m\int_0^{\theta^*} h(\theta)[p_{w_1,j}(\theta)w_1(\theta)]^{(m)}d\theta \\ &= \nu_{j+m,m}\int_0^{\theta^*} h(\theta)p_{w_0,j+m}(\theta)w_0(\theta)d\theta \\ &= \nu_{j+m,m}h_{w_0,j+m}, \quad \forall j\geq 0.\end{aligned}$$

From the completeness of $\{p_{w_1,j}\}_{j=0}^{\infty}$, we get $||h^{(m)}||_{w_1,2}^2 = \sum_{j=m}^{\infty}\nu_{j,m}^2 h_{w_0,j}^2$. Now (22) follows immediately from (20). To prove the second statement of Proposition 2, we argue as follows.

LAGUERRE POLYNOMIALS. Suppose $w_0(\theta) = \theta^{\beta}e^{-\theta}$, with $\theta > 0$ and $\beta > -1$, is the weight function of the normalized Laguerre polynomials

$$p_{w_0,j}(\theta) = [\Gamma(\beta+1)\binom{j+\beta}{j}]^{-1/2}\sum_{x=0}^{j}\binom{j+\beta}{j-x}\frac{(-\theta)^x}{x!}, \quad \forall j\geq 0.$$

For $j\geq 0$ and $m\geq 1$, we write $w_1(\theta) = \theta^{\beta+m}e^{-\theta}$,

$$p_{w_1,j}(\theta) = [\Gamma(\beta+m+1)\binom{j+\beta+m}{j}]^{-1/2}\sum_{x=0}^{j}\binom{j+\beta+m}{j-x}\frac{(-\theta)^x}{x!},$$

and

$$\nu_{j+m,m} = (-1)^m\frac{(j+m)!}{j!}[\Gamma(\beta+1)\binom{j+\beta+m}{j+m}]^{1/2}[\Gamma(\beta+m+1)\binom{j+\beta+m}{j}]^{-1/2}.$$

Then (19) follows from the Rodrigues' formula for Laguerre polynomials and (20) holds for $\alpha = m/2$.

JACOBI POLYNOMIALS. Suppose $w_0(\theta) = \theta^{\beta_1}(\theta^*-\theta)^{\beta_2}$, with $\beta_1 > -1$, $\beta_2 > -1$ and $0 < \theta < \theta^* < \infty$. Then the orthogonal polynomials with w_0 as the weight function correspond to the normalized Jacobi polynomials

$$p_{w_0,j}(\theta) = C_{j,\beta_1,\beta_2}\binom{j+\beta_2}{j}(\theta^*)^{-j}\sum_{x=0}^{j}\frac{j(j-1)\cdots(j-x+1)}{(\beta_2+1)(\beta_2+2)\cdots(\beta_2+x)}\binom{j+\beta_1}{x}\theta^{j-x}(\theta-\theta^*)^x,$$

where

$$C_{j,\beta_1,\beta_2} = [\frac{(2j+\beta_1+\beta_2+1)\Gamma(j+1)\Gamma(j+\beta_1+\beta_2+1)}{(\theta^*)^{\beta_1+\beta_2+1}\Gamma(j+\beta_1+1)\Gamma(j+\beta_2+1)}]^{1/2} \quad \text{if } j\geq 1,$$

and is equal to

$$[\frac{\Gamma(\beta_1+\beta_2+2)}{(\theta^*)^{\beta_1+\beta_2+1}\Gamma(\beta_1+1)\Gamma(\beta_2+1)}]^{1/2} \quad \text{if } j=0.$$

For $m\geq 1$, let $p_{w_1,j}$, $j\geq 0$, denote the set of normalized Jacobi polynomials with weight function

$$w_1(\theta) = \theta^{\beta_1+m}(\theta^*-\theta)^{\beta_2+m}, \quad \forall 0<\theta<\theta^*,$$

and

$$\nu_{j+m,m} = (\theta^*)^m(j+m)!C_{j,\beta_1+m,\beta_2+m}/[j!C_{j+m,\beta_1,\beta_2}].$$

Then (19) follows from the Rodrigues' formula for Jacobi polynomials and (20) holds for $\alpha = m$. □

PROOF OF THEOREM 2. We first observe from (23) that

$$n^{-1}\sum_{j=0}^{m_n}\max_{0\le x\le j}[k_{w_0,j,x}/q(x)]^2 = O(m_n n^{2\beta_0-1}). \tag{32}$$

We also observe that

$$\sup\{\sum_{j=m_n+1}^{\infty} h^2_{w_0,j} : g\in\mathcal{G}(\alpha,m,M,w_0)\} = O(m_n^{-2\alpha}). \tag{33}$$

Now the theorem follows from (32), (33) and Proposition 1. □

PROOF OF COROLLARY 1. From the properties of Laguerre polynomials, we have

$$\begin{aligned}
|k_{w_0,j,x}/q(x)| &\le \gamma_0\gamma_1^x(x!)^{\gamma-1}\binom{j+\beta}{j-x}[\Gamma(\beta+1)\binom{j+\beta}{j}]^{-1/2}\\
&= \gamma_0\gamma_1^x(x!)^{\gamma-1}\binom{j}{x}[\prod_{i=x+1}^{j}(1+\beta i^{-1})]^{1/2}[\Gamma(\beta+1)\prod_{i=1}^{x}(1+\beta i^{-1})]^{-1/2}\\
&\le \gamma_0\gamma_1^j 2^j[\prod_{i=x+1}^{j}(1+\beta i^{-1})]^{1/2}[\Gamma(\beta+1)\prod_{i=1}^{x}(1+\beta i^{-1})]^{-1/2}. \qquad (34)
\end{aligned}$$

Here we follow the convention that $\prod_{i=x_1}^{x_2}(1+\beta i^{-1})=1$ if $x_1>x_2$. We further observe that there exist positive constants c_1^* and c_2^* such that

$$c_1^* j^{-1}\le\prod_{i=1}^{j}(1+\beta i^{-1})\le c_2^* j,\qquad \forall j\ge 1.$$

Thus it follows from (34) that

$$\max_{0\le x\le j\le m_n}\log(|k_{w_0,j,x}|/q(x)) = m_n(1+o(1))\log(2\gamma_1)\le\beta_0(1+o(1))\log n.$$

This proves (23) and the corollary follows from Theorem 2. □

PROOF OF COROLLARY 2. If $\gamma=0$, we observe that

$$\max_{0\le x\le m_n}\log(1/q(x)) + m_n\log\zeta\le m_n(1+o(1))\log(\gamma_1\zeta)\le\beta_0(1+o(1))\log n.$$

This proves (25) and (a) follows from Theorem 3. The case of $0<\gamma<\infty$ is similar and is omitted. □

PROOF OF THEOREM 4. Let a, θ_0, θ_1, θ_2 and θ_3 be fixed constants satisfying $0<\theta_0<\theta_1<a<\theta_2<\theta_3<\theta^*$ such that w is strictly positive and m times continuously differentiable on $[\theta_0,\theta_3]$. Define $h_{u,v}(\theta)=v^u\theta^{u-1}e^{-v\theta}/\Gamma(u)$ and

$$g_{u,v}(\theta)=\begin{cases}0 & \text{if } 0<\theta<\theta_0,\\ l_{1,u,v}(\theta)/C(\theta) & \text{if } \theta_0\le\theta<\theta_1,\\ h_{u,v}(\theta)/C(\theta) & \text{if } \theta_1\le\theta\le\theta_2,\\ l_{2,u,v}/C(\theta) & \text{if } \theta_2<\theta\le\theta_3,\\ 0 & \text{if } \theta_3<\theta<\theta^*,\end{cases}$$

where $l_{j,u,v}$, $j = 1,2$, are $(2m+1)$th degree polynomials such that $g_{u,v}$ is m times continuously differentiable. Let g_0 be a probability density in $\mathcal{G}(\alpha, m, M-\varepsilon_1, w_0)$ for some small positive constant ε_1 and define

$$g_{0n}(\theta) = g_0(\theta) + \frac{3\varepsilon}{u_n^{1/4}}\left(\frac{\theta_2}{u_n}\right)^m \{g_{u_n,v_n}(\theta) - w_{0n}g_0(\theta)\}$$

$$g_{1n}(\theta) = g_{0n}(\theta) + \frac{\varepsilon}{u_n^{1/4}}\left(\frac{\theta_2}{u_n}\right)^m \left[\sin\left(u_n\frac{\theta-a}{\theta_2}\right) - \frac{w_{1n}}{w_{0n}}\right] g_{u_n,v_n}(\theta),$$

$$g_{2n}(\theta) = g_{0n}(\theta) + \frac{\varepsilon}{u_n^{1/4}}\left(\frac{\theta_2}{u_n}\right)^m \left[\cos\left(u_n\frac{\theta-a}{\theta_2}\right) - \frac{w_{2n}}{w_{0n}}\right] g_{u_n,v_n}(\theta),$$

where the constants w_{jn} are given by $\int_0^{\theta^*} g_{jn}(\theta)d\theta = 1$, ε is a small positive constant, $u_n = \delta_0 \log n$, and $v_n = u_n/a$, with

$$\delta_0 = \max\{\frac{\theta_2/(\theta_3-\theta_2)}{\log(\theta_3/\theta_2)}, \frac{2}{\log(1+a^2/\theta_2^2)}, \frac{1}{\theta_1/a - 1 - \log(\theta_1/a)}, \frac{1}{\theta_2/a - 1 - \log(\theta_2/a)}\}.$$

The rest of the proof is almost identical to Steps 1 to 3 of the proof of Theorem 3 of Loh and Zhang (1994). As such it suffices only to verify that $g_{jn} \in \mathcal{G}(\alpha, m, M, w_0)$ for $j = 0,1,2$. Define for $0 < \theta < \theta^*$,

$$h(\theta) = \frac{3\varepsilon}{u_n^{1/4}}(\frac{\theta_2}{u_n})^m w(\theta) g_{u,v}(\theta)/C(\theta).$$

Then using Leibniz rule we have $||h^{(m)}||_{w_1,2} = \varepsilon O(1)$, where the $O(1)$ term does not depend on ε. Since (19) and (20) hold, we observe from Proposition 2 that $(\sum_{j=m}^\infty j^{2\alpha} h_{w_0,j}^2)^{1/2} = \varepsilon O(1)$, where $h_{w_0,j} = \int_0^{\theta^*} h(\theta)p_{w_0,j}(\theta)w_0(\theta)d\theta$. Writing

$$g_{0n,w_0,j} = \int_0^{\theta^*} C(\theta)g_{0n}(\theta)p_{w_0,j}(\theta)d\theta, \quad \forall j \geq m,$$

it follows from Minkowski's inequality that $(\sum_{j=m}^\infty j^{2\alpha} g_{0n,w_0,j}^2)^{1/2} \leq M - \varepsilon_1 + \varepsilon O(1)$. Thus we conclude that $g_{0n} \in \mathcal{G}(\alpha, m, M, w_0)$ for sufficiently small ε. Likewise we have $g_{jn} \in \mathcal{G}(\alpha, m, M, w_0)$, $j = 1,2$. □

PROOF OF THEOREM 5. Let $g \in \mathcal{G}(\alpha, m, M, w_0)$ and $h_{w_0,j}$ be as in (17). Define for each $\beta > 0$,

$$j_n^*(\beta) = \begin{cases} \max\{j : 0 \leq j \leq B_n, h_{w_0,j}^2 > (\log n)^{-\beta}\}, & \text{if } \{j : 0 \leq j \leq B_n, h_{w_0,j}^2 > (\log n)^{-\beta}\} \neq \phi, \\ 0, & \text{otherwise.} \end{cases}$$

We shall first show that

$$\sup\{P_g[\hat{m}_n^* < j_n^*(\beta)] : g \in \mathcal{G}(\alpha, m, M, w_0)\} = O(1)(\log n)^{2(1+\beta)} n^{2\beta_0 - 1}. \tag{35}$$

Since (35) clearly holds when $j_n^*(\beta) = 0$, it suffices to assume that $j_n^*(\beta) \geq 1$. Let $t_{n,i,j} = \sum_{l=i}^j t_{n,l}$ and $\sigma(\hat{t}_{n,i,j})$ be the standard deviation of $\hat{t}_{n,i,j}$. We observe from (29) and the definition of $\lambda_{w_0,j}$ that $\sup\{t_{n,j,j_n^*(\beta)} : g \in \mathcal{G}(\alpha, m, M, w_0), 0 \leq j \leq j_n^*(\beta)\} \leq -(\log n)^{-\beta}/2$ for sufficiently large n. Also

$$\begin{aligned} & P_g[\hat{m}_n^* < j_n^*(\beta)] \\ \leq & \sum_{j=0}^{j_n^*(\beta)-1} P_g[\hat{t}_{n,j+1,j_n^*(\beta)} + z_{\alpha^*}\hat{\sigma}(\hat{t}_{n,j+1,j_n^*(\beta)}) \geq 0] \\ = & \sum_{j=0}^{j_n^*(\beta)-1} P_g[\frac{\hat{t}_{n,j+1,j_n^*(\beta)} - t_{n,j+1,j_n^*(\beta)}}{\sigma(\hat{t}_{n,j+1,j_n^*(\beta)})} + z_{\alpha^*}(\frac{\hat{\sigma}(\hat{t}_{n,j+1,j_n^*(\beta)})}{\sigma(\hat{t}_{n,j+1,j_n^*(\beta)})} - 1) \geq -z_{\alpha^*} - \frac{t_{n,j+1,j_n^*(\beta)}}{\sigma(\hat{t}_{n,j+1,j_n^*(\beta)})}] \\ \leq & 8(1+o(1))B_n(1+4z_{\alpha^*}^2)(\log n)^{2\beta}\sup\{\sigma^2(\hat{t}_{n,j+1,j_n^*(\beta)}) : 0 \leq j < j_n^*(\beta)\}, \end{aligned} \tag{36}$$

uniformly over $g \in \mathcal{G}(\alpha, m, M, w_0)$. (35) now follows from (36) and the observation that

$$\sup\{\sigma^2(\hat{t}_{n,j+1,j_n^*(\beta)}) : g \in \mathcal{G}(\alpha, m, M, w_0), 0 \leq j < j_n^*(\beta)\} = O((\log n)^2 n^{2\beta_0 - 1}).$$

In a similar manner, we have

$$\sup\{\sum_{j=1}^{m-1} h_{w_0,j}^2 P_g(\hat{m}_n^* < j) : g \in \mathcal{G}(\alpha, m, M, w_0)\} = o(B_n^{-2\alpha}). \quad (37)$$

Next as in (31), we observe that

$$E_g \int_0^{\theta^*} [\hat{g}_{OP,n}^*(\theta) - g(\theta)]^2 w(\theta) d\theta$$

$$\leq E_g\{n^{-1} \sum_{j=0}^{B_n} \max_{0 \leq x \leq j} [k_{w_0,j,x}/q(x)]^2 + \sum_{j=B_n+1}^{\infty} h_{w_0,j}^2 + \sum_{j=(\hat{m}_n^*+1)\vee m}^{B_n} h_{w_0,j}^2 + \sum_{j=1}^{m-1} h_{w_0,j}^2 I\{\hat{m}_n^* < j\}\}. \quad (38)$$

Conditioning on whether or not $\hat{m}_n^* \geq j_n^*(\beta)$, we observe using (35) that for sufficiently large β, the third term on the r.h.s. of (38) is bounded by

$$MP_g[\hat{m}_n^* < j_n^*(\beta)] + B_n(\log n)^{-\beta} = o(B_n^{-2\alpha}), \quad (39)$$

uniformly over $g \in \mathcal{G}(\alpha, m, M, w_0)$ as $n \to \infty$. The theorem now follows from (37), (38) and (39). □

References

[1] Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83** 1184-1186.

[2] Datta, S. (1991). On the consistency of posterior mixtures and its applications. *Ann. Statist.* **19** 338-353.

[3] Devroye, L. P. and Wise, G. L. (1979). On the recovery of discrete probability densities from imperfect measurements. *J. Franklin Inst.* **307** 1-20.

[4] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257-1272.

[5] Johnstone, I. M. and Silverman, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18** 251-280.

[6] Loh, W. L. and Zhang, C.-H. (1994). Global properties of kernel estimators for mixing densities in exponential family models for discrete variables. Revised for *Statist. Sinica.*

[7] Meeden, G. (1972). Bayes estimation of the mixing distribution, the discrete case. *Ann. Math. Statist.* **43** 1993-1999.

[8] Rolph, J. E. (1968). Bayesian estimation of mixing distributions. *Ann. Math. Statist.* **39** 1289-1302.

[9] SIMAR, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.* 4 1200-1209.

[10] SZEGÖ, G. (1975). *Orthogonal Polynomials.* Amer. Math. Soc., Providence, Rhode Island.

[11] TUCKER, H. G. (1963). An estimate of the compounding distribution of a compound Poisson distribution. *Theor. Probab. Appl.* **8** 195-200.

[12] WALTER, G. G. and HAMEDANI, G. G. (1989). Bayes empirical Bayes estimation for discrete exponential families. *Ann. Inst. Statist. Math.* **41** 101-119.

[13] WALTER, G. G. and HAMEDANI, G. G. (1991). Bayes empirical Bayes estimation for natural exponential families with quadratic variance functions. *Ann. Statist.* **19** 1191-1224.

[14] ZHANG, C.-H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.* **18** 806-831.

[15] ZHANG, C.-H. (1992). On estimating mixing densities in exponential family models for discrete variables. Tech. Rep. 055-92, Mathematical Sciences Research Institute, Berkeley, California.