The Annals of Statistics
1997, Vol. 25, No. 3, 917–928

ADAPTIVE DEMIXING IN POISSON MIXTURE MODELS

BY NICOLAS W. HENGARTNER

Yale University

Let $X_1, X_2, \dots, X_n$ be an i.i.d. sample from the Poisson mixture distribution $p(x) = (1/x!)\int_0^\infty s^x e^{-s} f(s)\, ds$. Rates of convergence in mean integrated squared error (MISE) of orthogonal series estimators for the mixing density f supported on $[a, b]$ are studied. For the Hölder class of densities whose rth derivative is Lipschitz α, the MISE converges at the rate $(\log n/\log\log n)^{-2(r+\alpha)}$. For Sobolev classes of densities whose rth derivative is square integrable, the MISE converges at the rate $(\log n/\log\log n)^{-2r}$. The estimator is adaptive over both these classes.

For the Sobolev class, a lower bound on the minimax rate of convergence is $(\log n/\log\log n)^{-2r}$, and so the orthogonal polynomial estimator is rate optimal.

1. Introduction. Consider the following problem: given an i.i.d. sample from a Poisson mixture distribution

$$P[X = k] = \pi_k(f) = \frac{1}{k!}\int_0^\infty y^k e^{-y} f(y)\, dy, \tag{1}$$

estimate the mixing density f. As shown in Teicher (1961), the mixing distribution $F(t) = \int_0^t f(s)\, ds$ in (1) is identifiable, a modern proof of which is found in Lindsay and Roeder (1993). Many estimators of the mixing density have been proposed. Tucker (1963) smoothed the empirical frequencies $k!\#\{X_j = k\}/n$ to solve a moment problem for $e^{-t}f(t)$. Simar (1976) and Lambert and Tierney (1984) studied the non-parametric maximum likelihood estimator. Walter (1985) and Walter and Hamedani (1991) considered estimators based on Laguerre polynomials. All these estimators were shown to be consistent without analyzing the rate of convergence.

More recently, Zhang (1992) considered smoothing kernel estimators for demixing mixtures of exponential family for discrete variables. These include, as a special case, the Poisson mixture distribution. Subsequently with Loh [Loh and Zhang (1993)], he derived upper and lower bounds for the integrated mean squared error (MISE), for densities f whose rth derivative is Lipschitz α, of order $(\log n/\log\log n)^{-2(r+\alpha)}$ and $(\log n)^{-2(r+\alpha)}$, respectively. Note that these bounds do not match.

In this paper we consider estimators of the mixing density obtained by estimating the first m coefficients of the expansion of the density into a series of orthonormal polynomials. We show that by estimating the first $m_n =$

Received September 1994; revised August 1996.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.

Key words and phrases. Poisson mixtures, demixing, optimal rates of convergence, orthonormal polynomial estimator, adaptive estimation.

$\log n/\log\log n$ coefficients, the estimator converges in MISE at rate $(\log n/\log\log n)^{-2r}$, assuming the mixing density f has r square integrable derivatives and converges at rate $(\log n/\log\log n)^{-2(r+\alpha)}$ when the true mixing density f belongs to the Hölder classes of functions whose rth derivatives are Lipschitz α. Since the number of estimated coefficients does not depend on the assumed smoothness, we say that our estimator is automatically adaptive, or universal, over the above function classes.

The orthogonal polynomial and the kernel estimator achieving the same rate is suggestive that Zhang's lower bound may be too small. Applying the technique of Farrell (1967, 1972) and Bickel and Ritov (1988), of finding pairs of mixing densities far apart in the $\mathscr{L}_2$ distance and whose Poisson mixture distributions are contiguous, a lower bound of order $(\log n/\log\log n)^{-2r}$ for the minimax MISE over a Sobolev ball is obtained. Hence for this class, the orthonormal polynomial and kernel density estimators achieve the optimal rate. For Lipschitz classes of densities, it is conjectured that Zhang's lower bound can be improved to show rate optimality of kernel and orthonormal polynomial demixing estimators.

2. Orthonormal function estimator. Throughout this paper, assume the probability density f to be square integrable and supported on the interval $[a, b]$ with known end points $0 < a < b < \infty$. Then, for any complete orthonormal basis $\{q_j\}_{j=0}^{\infty}$ of $\mathscr{L}_2[a, b]$, f can be represented by the Fourier series $f(t) = \sum_{k=0}^{\infty} \alpha_k q_k(t)$. Cencov (1962) first considered density estimators obtained by estimating the coefficients of the approximation

$$f_m(t) = \sum_{j=0}^{m} \alpha_j q_j(t)$$

of the density f. Slowly increasing the number of terms m, with the sample size n, makes the estimator consistent for the density f in the $\mathscr{L}_2$ sense.

2.1. *Estimating the Fourier coefficients.* For the Poisson demixing problem, one needs to estimate the Fourier coefficients of the mixing density from Poisson mixture observations. For this, assume that the basis functions q_k are analytic (with radius of convergence at least b) and define the kernel $K_m(t, s) = \sum_{k=0}^{m} q_k(s)q_k(t)$. The approximation $f_m(t)$ is then written as

$$f_m(t) = \int K_m(t, s) f(s)\, ds. \tag{2}$$

For each fixed t, the latter is a linear functional of the mixing density f. Following Hengartner (1994), if there exists a linear functional of the form

$$f_m(t) = \sum_{k=0}^{\infty} \xi_m(t, k)\pi_k(f) \tag{3}$$

of the Poisson mixture probabilities that agrees with (2) for every f, then formally

$$
\begin{aligned}
f_m(t) &= \sum_{k=0}^{\infty} \xi_m(t,k) \int_0^\infty \frac{s^k}{k!} e^{-s} f(s)\, ds \\
&= \int_0^\infty \left[\sum_{k=1}^{\infty} \frac{\xi_m(t,k)}{k!} s_k \right] e^{-s} f(s)\, ds \\
&= \int_0^\infty e^s K_m(t,s) e^{-s} f(s)\, ds.
\end{aligned}
$$

A power series expansion of $e^s K_m(t,s)$ identifies $\xi_m(t,k)$ as

$$
(4) \qquad \xi_m(t,k) = \frac{\partial^k}{\partial s^k} e^s K_m(t,s)\bigg|_{s=0} = \sum_{j=0}^{m} \left\{ \frac{d^k}{ds^k} e^s q_j(s)\bigg|_{s=0} \right\} q_j(t).
$$

Substituting the empirical frequencies $\hat{\pi}_k = (1/n)\sum_{j=1}^{n} I(X_j = k)$ for the Poisson mixture probabilities in (3) yields the estimator

$$
\begin{aligned}
\hat{f}_{m,n}(t) &= \sum_{k=0}^{\infty} \xi_m(t,k)\, \hat{\pi}_k \\
&= \sum_{k=0}^{\infty} \sum_{j=0}^{m} \left\{ \frac{d^k}{ds^k} e^s q_j(s)\bigg|_{s=0} \right\} q_j(t)\, \hat{\pi}_k \\
&= \sum_{j=0}^{m} \left\{ \sum_{k=1}^{\infty} \frac{d^k}{ds^k} e^s q_j(s)\bigg|_{s=0} \hat{\pi}_k \right\} q_j(t).
\end{aligned}
$$

Thus $\hat{f}_{m,n}$ is an orthonormal function demixing estimator with

$$
(5) \qquad \hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d^{X_i}}{ds^{X_i}} e^s q_k(s)\bigg|_{s=0} \right\}
$$

estimating the kth Fourier coefficient from Poisson mixture observations. By the law of large numbers, if $E[|\xi_m(t,X)|] < \infty$ for all m, then $\hat{f}_{m,n}$ is consistent for f, provided m increases slowly enough. Not every orthonormal basis satisfies this requirement, and even fewer will lead to estimators of α_k having finite variance. However, if q_k is a polynomial, then the estimator (5) has finite moments of all order. This motivates our interest in orthonormal polynomial bases.

Let $q_k(t) = \sum_{j=0}^{k} c_{k,j} t^j$ be the orthonormal polynomial of degree k on $[a,b]$. For this basis, the coefficient α_k is estimated by

$$
\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{d^{X_i}}{ds^{X_i}} e^s \left(\sum_{j=0}^{k} c_{k,j} s^j \right) \bigg|_{s=0} \right\}
$$

$$= c_{k,0} + \sum_{j=1}^{k} c_{k,j} \frac{1}{n} \sum_{i=1}^{n} X_i(X_i - 1) \cdots (X_i - j + 1)$$

$$= c_{k,0} + \sum_{j=1}^{k} c_{k,j} \hat{\mu}_j,$$

where $\hat{\mu}_k = (1/n)\sum_{i \leq n} X_i(X_i - 1) \cdots (X_i - k + 1)$ estimates the kth moment of the mixing distribution. By Lemma A.1 in the Appendix, $\hat{\mu}_k$ is unbiased and has finite variance for compactly supported mixing distributions. It can further be shown that it has finite moments of all orders.

2.2. *Order of convergence in mean integrated squared error*. The difficulty of estimating the mixing density using the orthonormal polynomial estimator is assessed via the rate the integrated mean squared error (MISE) converges to zero. By the usual argument, the MISE is decomposed into integrated variance and integrated squared bias terms. The former is bounded by a function of the sample size n and the number of terms m, while the latter only depends on how well f_m approximates f, which depends on the smoothness class to which the density f belongs. Typical smoothness classes are balls in Sobolev spaces

$$\mathscr{G}_{\nu, M} := \left\{ f \colon \int \left| f^{(\nu)}(t) \right|^2 dt \leq M \right\}, \tag{6}$$

with integer ν, and balls in Hölder spaces

$$\mathscr{F}_{\nu, M} := \left\{ f \colon \left| f^{(r)}(t) - f^{(r)}(s) \right| \leq M |t - s|^{\alpha} \right\}, \tag{7}$$

where $\nu = r + \alpha$, r integer and $0 < \alpha \leq 1$. Here, upper bounds on the rate of convergence of the maximal MISE over (6) and (7) are considered, while lower bounds on the minimax MISE over Sobolev balls will be derived in Section 3.

The following theorem, whose proof is relegated to the Appendix, upper-bounds the integrated variance.

THEOREM 2.1. *The integrated variance of the orthonormal polynomial estimator of the mixing density is bounded by*

$$n \int_a^b \mathrm{Var}_f\left[\hat{f}_{n,m}(t) \right] ds \leq \frac{(2m + 1)m^2}{b - a} \left[(m + b)\rho b \right]^m, \tag{8}$$

where

$$\rho = \left[2\left(\frac{b + a}{b - a} + \frac{b - a}{b + a} \right)\left(1 + \frac{2}{b + a} \right) \right]^2. \tag{9}$$

Jackson-type theorems from approximation theory are used to bound the integrated squared bias. For example, specializing Theorem 6.3 of DeVore and Lorentz (1993) gives the following bounds.

THEOREM 2.2. *Let $\mathscr{P}_m$ be the collection of polynomials of degree at most m on $[a, b]$. The maximal integrated squared bias is bounded by*

$$\sup_{f \in \mathscr{G}_{\nu, M}} \inf_{f_m \in \mathscr{P}_m} \int_a^b (f_m(t) - f(t))^2 \, dt \le c_1 m^{-2\nu}$$

and

$$\sup_{f \in \mathscr{F}_{\nu, M}} \inf_{f_m \in \mathscr{P}_m} \int_a^b (f_m(t) - f(t))^2 \, dt \le c_2 m^{-2\nu},$$

where c_1 and c_2 are positive constants depending on M and ν.

Combining these theorems results in the following theorem.

THEOREM 2.3. *For mixing densities $f \in \mathscr{F}_{\nu, M}$ or $f \in \mathscr{G}_{r, M}$, the orthonormal demixing estimator with $m(n) = \log n/\log\log n$, has MISE converging to zero at rate $(\log n/\log\log n)^{-2\nu}$.*

REMARKS.

1. For densities $f \in \mathscr{F}_{\nu, M}$, these rates are the same as those of the kernel demixing estimator in Loh and Zhang (1993), and thus the orthogonal polynomial demixing estimator is *almost* rate optimal. These two different demixing methods' converging at the same rate suggests that it may be possible to improve upon the $(\log n)^{-2\nu}$ lower bound.
2. In the next section, MISE minimax lower bounds for densities in $\mathscr{G}_{\nu, M}$ are shown to be of order $(\log n/\log\log n)^{-2\nu}$, and therefore, the orthonormal polynomial demixing estimator is rate optimal.
3. The number of terms $m(n)$ does not depend on the smoothness class to which the density f belongs. Because the estimator, with a single choice for $m = m(n)$, achieves (or almost achieves) the optimal rate for the considered smoothness class of densities justifies naming these orthonormal polynomial demixing estimators as automatically adaptive or universal.
4. Estimating only $\kappa \log n/\log\log n$ coefficients, with $\kappa < 1$, does not affect the rate of convergence but may change the constant. The choice of the κ leading to the best constant is expected to depend on the assumed class of densities.

PROOF. Combining Theorems 2.1 and 2.2 yields for either $f \in \mathscr{F}_{\nu, M}$ or $f \in \mathscr{G}_{\nu, M}$

$$\begin{aligned}\int_a^b E_f\Big[\big(\hat{f}_{n,m}(t) - f(t)\big)^2\Big] &= \int_a^b \mathrm{Var}_f\big(\hat{f}_{n,m}(t)\big)\, dt + \int_a^b (f_m(t) - f(t))^2 \, dt \\ &\le C_1 \frac{m^m}{n} + C_2 m^{-2\nu},\end{aligned}$$

for positive constants C_1 and C_2. The optimal rate is obtained by choosing $m = m(n)$ to equate the integrated variance to the integrated squared bias. Taking $m(n) = \kappa \log n/\log\log n$, with $0 < \kappa \le 1$, the MISE converges at the rate $(\log n/\log\log n)^{-2\nu}$. □

3. Optimal rate of convergence. For densities lying in the Sobolev ball $\mathscr{G}_{\nu, M}$, the following theorem holds.

THEOREM 3.1. *Assume that* $\nu \ge 1$. *Then for every sequence of estimators* $\tilde{f}_n(t)$, *based on an i.i.d. sample* $X_1, X_2, \dots, X_n$ *from a Poisson mixtures distribution*,

$$\liminf_{n\to\infty} \sup_{f\in\mathscr{G}_{\nu,M}} P_f\left[\int_a^b \left(\tilde{f}_n(t) - f(t)\right)^2 dt > c\left(\frac{\log n}{\log\log n}\right)^{-2\nu}\right] > 0 \tag{10}$$

holds for some constant $c > 0$.

REMARKS.

1. The orthogonal polynomial estimator achieves this rate, and hence $(\log n/\log\log n)^{-2\nu}$ is the optimal rate of convergence as defined by Stone (1980).
2. Recall that in the definition of $\mathscr{G}_{\nu, M}$, ν is an integer.

Theorem 3.1 is proved by finding pairs of mixing densities $f_{n,1}, f_{n,2} \in \mathscr{G}_{\nu,M}$ with large $\mathscr{L}_2$ distance and contiguous Poisson mixture distributions. This idea is implicit in Farrell (1967, 1972) and Bickel and Ritov (1988) and is exploited in Zhang (1995), whose Lemma 1 is key.

LEMMA 3.2. *Let* $\mathscr{F}$ *be a collection of probability distributions and let* $\theta(F)$ *be a mapping from* $\mathscr{F}$ *to a metric space with metric* $d(\cdot,\cdot)$. *Let* $f_{n,1}$ *and* $f_{n,2}$ *be the joint density of* $X_1, X_2, \dots, X_n$ *under* $F_{n,j} \in \mathscr{F}$, $j = 1, 2$, *respectively. Set* $d_n = d(\theta(F_{n,1}), \theta(F_{n,2}))/2$. *If for a* $\lambda > 0$,

$$\liminf_{n\to\infty} P_{F_{n,1}}[f_{n,1} \le \lambda f_{n,2}] \ge \rho, \tag{11}$$

then any estimator of T_n *based on* $X_1, X_2, \dots, X_n$ *satisfies*

$$\liminf_{n\to\infty} \sup_{F\in\mathscr{F}} P_F[d(T_n, \theta(F)) > d_n] > \frac{\rho}{1+\lambda}. \tag{12}$$

PROOF OF THEOREM 3.1. Let $f_{n,1}$ be the uniform density on $[a, b]$ and $f_{n,2} = f_{n,1} + p_n$ a perturbation thereof. In particular, let $p_n(t) = \gamma_m T_m(t) = \gamma_m \sum_{j=0}^m a_{m,j} \cos(jt)$ be an appropriately scaled trigonometric polynomial of degree $m = m(n)$, whose coefficients $\alpha_{m,j}$ are uniquely determined by the requirements

$$\int_a^b T_m(t)\,dt = 0 \tag{13}$$

and

$$\int_a^b T_m(t)t^k e^{-t}\,dt = 0 \quad \text{for } k = 0, 1, \dots, m-1. \tag{14}$$

By condition (13), $f_{n,2}$ integrates to 1 and condition (14) makes $\pi_k(f_{n,1}) = \pi_k(f_{n,2})$ for $k = 0, 1, \dots, m-1$. Thus choosing

$$m = m(n) \approx \max_{j \le n} X_j \approx c \log n / \log\log n$$

ensures contiguity of the Poisson mixture distribution, regardless of the size of the scaling factor γ_m, which is determined by the requirements that $f_{n,2} \in \mathscr{G}_{\nu,M}$. This choice of perturbation, in the spirit of Donoho and Liu (1991), disentangles the stochastic and analytic arguments.

To determine the size γ_m putting $f_{n,2} \in \mathscr{G}_{\nu,M}$, apply Bernstein's theorem [see Rahman (1982) or page 98 of DeVore and Lorentz (1993)] to obtain

$$\|T_m^{(\nu)}\|_2 \le \left(\frac{2\pi}{b-a}\right)^{\nu} m^{\nu} \|T_m\|_2,$$

and using Nikolskii's bound [see page 102 of DeVore and Lorentz (1993)] yields

$$\|T_m\|_\infty \le \sqrt{\frac{(b-a)(2m+1)}{2\pi}}\,\|T_m\|_2.$$

Setting

$$\gamma_m = \frac{C}{m^{\nu}\|T_m\|_2},$$

we first have that $f_{n,2}(t) \ge 0$, and hence a probability density, provided $\nu \ge 1/2$ and C is small enough. Furthermore, $\|f_{n,2}^{(\nu)}\|_2 \le C$, so that for $C \le M^2$, $f_{n,2} \in \mathscr{G}_{\nu,M}$.

The index $m = m(n)$ will be chosen such that the Poisson mixture distribution under $f_{n,1}$ and $f_{n,2}$ satisfy (11) of Lemma 3.2. For this, note that by construction $\pi_k(f_{n,1}) = \pi_k(f_{n,2})$ for $k = 0, 1, \dots, m-1$, so that for $\lambda > 1$,

$$P_{f_{n,1}}\left[\prod_{j \le n} \frac{\pi_{X_j}(f_{n,1})}{\pi_{X_j}(f_{n,2})} \le \lambda\right] = 1 - P_{f_{n,1}}\left[\prod_{j \le n} \frac{\pi_{X_j}(f_{n,1})}{\pi_{X_j}(f_{n,2})} > \lambda\right] \tag{15}$$

$$\ge 1 - P_{f_{n,1}}\left[\max_{1 \le j \le n} X_j \ge m\right] \tag{16}$$

$$\ge 1 - n\frac{b_m}{m!}. \tag{17}$$

Taking $m(n) = \kappa \log n / \log\log n$, the latter converges to one for any $\kappa > 1$. In fact, convergence to one still occurs if $\kappa = 1 + c \log\log\log n / \log\log n$, for any $c > 1$.

With these choices, apply Lemma 3.2 and note that

$$\begin{aligned}\|f_{n,1} - f_{n,2}\|_2 &= \gamma_{m(n)}\|T_{m(n)}\|_2 \\ &= \frac{C}{\|T_{m(n)}\|_2} m(n)^{-\nu}\|T_{m(n)}\|_2 \\ &= \frac{C}{\kappa^{2\nu}}\left(\frac{\log n}{\log\log n}\right)^{-2\nu}. \qquad \square\end{aligned}$$

APPENDIX

Proof of Theorem 2.1. The following lemmas are needed in proving Theorem 2.1.

LEMMA A.1. *Let X have a Poisson mixture distribution, with mixing density f supported on $[a, b]$. Define $(X)_k = X(X-1)\cdots(X-k+1)$, then*

$$E_f[(X)_k] = \int_a^b t^k f(t)\,dt \equiv \mu_k(f) \tag{18}$$

and

$$E_f[(X)_k(X)_m] = \sum_{l=0}^{\min(k,m)} \binom{m}{l}\binom{k}{l} l!\,\mu_{k+m-l}. \tag{19}$$

Furthermore, the bound

$$E_f\left[((X)_m)^2\right] \le (b+m)^m \mu_m \tag{20}$$

holds for all $m = 0, 1, \dots$.

PROOF. Equation (18) is found in Walter (1985). Without loss of generality, assume $k \le m$, so that

$$E[(X)_m(X)_k] = \sum_{l=m}^{\infty} \frac{l!}{(l-m)!}\,\frac{l!}{(l-k)!}\,\frac{1}{l!}\int_0^\infty t^l e^{-t} f(t)\,dt.$$

All summands being positive, the order of summation and integration can be interchanged, and with the help of the identity

$$\begin{aligned}\sum_{l=m}^{\infty} \frac{l!}{(l-k)!}\,\frac{t^l}{(l-m)!} &= t^k \frac{d^k}{dt^k}[t^m e^t] \\ &= \sum_{l=0}^{\min(k,m)} \binom{m}{l}\binom{k}{l} l!\,t^{k+m-l} e^t,\end{aligned}$$

conclusion (19) follows. Finally,

$$E\left[((X)_m)^2\right] = \sum_{l=0}^{m} \binom{m}{l}^2 l!\mu_{2m-l} \le \sum_{l=0}^{m} \binom{m}{l} \frac{m!}{(m-l)!} b^{m-l}\mu_m$$
$$\le \sum_{l=0}^{m} \binom{m}{l} m^l b^{m-l}\mu_m = (m+b)^m \mu_m. \qquad \square$$

LEMMA A.2. *The coefficients*

$$\mathbf{c}_m = (c_{m,0}, c_{m,1}, c_{m,2}, \dots, c_{m,m})^t$$

of the orthonormal polynomial $q_m(t)$ *of degree* m *with respect to the uniform weight function on the interval* $[a, b]$ *has squared norm bounded by*

$$\|\mathbf{c}_m\|_2^2 \le \frac{2m+1}{2\sigma}\tau^m,$$

where $\mu = (a+b)/2$, $\sigma = (b-a)/2$, *and* $\tau = (2(\mu/\sigma + \sigma/\mu)(1 + 1/\mu))^2$.

PROOF. The orthonormal polynomial $q_m(t)$ of degree m on $[a, b]$ is the translated, rescaled and normalized Legendre polynomial of degree m:

$$q_m(t) = \sum_{k=0}^{m} c_{m,k} t^k = \sqrt{\frac{2m+1}{2\sigma}}\, P_m\left[\frac{t-\mu}{\sigma}\right].$$

Expanding $P_m((t-\mu)/\sigma)$ into powers of t,

$$P_m\left[\frac{t-\mu}{\sigma}\right] = \sum_{k=0}^{m} b_{m,k}\left(\frac{t-\mu}{\sigma}\right)^k$$
$$= \sum_{k=0}^{m}\left\{\sum_{j=0}^{k}\binom{k}{j}\left(-\frac{t}{\mu}\right)^j\right\}\left(-\frac{\mu}{\sigma}\right)^k b_{m,k}$$
$$= \sum_{j=0}^{m}\left\{\sum_{k=j}^{m}\binom{k}{j}\left(-\frac{\mu}{\sigma}\right)^k b_{m,k}\right\}\left(-\frac{t}{\mu}\right)^j,$$

it follows that the kth coefficient of $\mathbf{c}_m$ is

$$c_{m,k} = \sqrt{\frac{2m+1}{2\sigma}}\,\frac{(-1)^k}{\mu^k}\sum_{j=k}^{m}\binom{j}{k}\left(-\frac{\mu}{\sigma}\right)^j b_{m,j}. \tag{21}$$

The coefficients of the Legendre polynomials are [see page 157 of Rainville (1967)]

$$b_{m,m-2j} = \frac{(-1)^j}{2^m}\frac{(2m-2j)!}{(m-j)!(m-2j)!j!}$$

and $b_{m,m-(2j+1)} = 0$. This produces

$$\sqrt{\frac{2\sigma}{2m+1}}\, c_{m,k} = \frac{(-1)^k}{\mu^k} \sum_{2j\le m-k} \binom{m-2j}{k} \left(-\frac{\mu}{\sigma}\right)^{m-2j} b_{m,m-2j}$$

$$= \left(-\frac{\mu}{2\sigma}\right)^m \frac{(-1)^k}{\mu^k}$$

(22)

$$\times \sum_{2j\le m-k} \binom{m-2j}{k} \frac{(2m-2j)!}{(m-j)!\,(m-2j)!\,j!} \left(\frac{\sigma}{\mu}\right)^{2j}$$

$$= \left(-\frac{\mu}{2\sigma}\right)^m \frac{(-1)^k}{\mu^k} \binom{m+k}{k} \sum_{2j\le m-k} \binom{2m-2j}{m+k} \binom{m}{j} \left(\frac{\sigma}{\mu}\right)^{2j}.$$

The sum in (22) is bounded by

$$\left| \sum_{2j\le m-k} \binom{2m-2j}{m+k} \binom{m}{j} \left(\frac{\sigma}{\mu}\right)^{2j} \right| \le \binom{2m}{m+k} \sum_{2j\le m-k} \binom{m}{j} \left(\frac{\sigma}{\mu}\right)^{2j}$$

$$\le \binom{2m}{m+k} \left(1 + \left(\frac{\sigma}{\mu}\right)^2\right)^m,$$

from which it follows that

$$\|\mathbf{c}_m\|^2 \le \frac{2m+1}{2\sigma} \left(\frac{\mu}{2\sigma}\left(1 + \left(\frac{\sigma}{\mu}\right)^2\right)\right)^{2m} \sum_{k=0}^{m} \left(\binom{m+k}{k}\binom{2m}{m+k}\right)^2 \frac{1}{\mu^{2k}}$$

$$= \frac{2m+1}{2\sigma} \left(\frac{1}{2}\left(\frac{\mu}{\sigma} + \frac{\sigma}{\mu}\right)\right)^{2m} \binom{2m}{m} \sum_{k=0}^{m} \binom{m}{k}^2 \frac{1}{\mu^{2k}}.$$

This last sum is less than

$$\sum_{k=0}^{m} \binom{m}{k}^2 \frac{1}{\mu^{2k}} \le \sum_{k=0}^{m} \binom{2m}{2k} \frac{1}{\mu^{2k}} \le \sum_{k=0}^{2m} \binom{2m}{k} \frac{1}{\mu^{k}} = \left(1 + \frac{1}{\mu}\right)^{2m},$$

and therefore

$$\|\mathbf{c}_m\|_2^2 \le \frac{2m+1}{2\sigma} \left(\frac{1}{2}\left(\frac{\mu}{\sigma} + \frac{\sigma}{\mu}\right)\right)^{2m} \left(1 + \frac{1}{\mu}\right)^{2m} \binom{2m}{m}^2$$

$$\le \frac{2m+1}{2\sigma} \left(2\left(\frac{\mu}{\sigma} + \frac{\sigma}{\mu}\right)\left(1 + \frac{1}{\mu}\right)\right)^{2m},$$

which is the desired bound. □

PROOF OF THEOREM 2.1. Let

$$\mathbf{c}_k = \mathbf{c}_k^{(m)} = \Big(c_{k,0}, c_{k,1}, c_{k,2}, \ldots, c_{k,k} \underbrace{0, 0, \ldots, 0}_{m-k}\Big)^t$$

be the vector of coefficients of the orthonormal polynomial $q_k(t)$. By (18) of Lemma A.1, the estimator of the kth Fourier coefficient

$$\hat{\alpha}_k = c_{k,0} + \sum_{j=1}^{k} c_{k,j}\hat{\mu}_j$$

is unbiased for α_k, and thus the integrated variance of $\hat{f}_{n,m}$ is

$$\int_a^b \mathrm{Var}_F\big[\hat{f}_{m,n}(t)\big]\,dt = E_F\left[\sum_{j=0}^{m}\left(\hat{\alpha}_j - \alpha_j\right)^2\right] = \sum_{k=0}^{m}\mathbf{c}_k^t\boldsymbol{\Sigma}\mathbf{c}_k,$$

where Σ denotes the variance-covariance matrix of the vector of moment estimators $(1, \hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_m)$. It is easily verified that every element of $\boldsymbol{\Sigma}$ is nonnegative. The above sum is bounded by

$$\sum_{k=0}^{m}\mathbf{c}_k^t\boldsymbol{\Sigma}\mathbf{c}_k \le \lambda_{\max}(\boldsymbol{\Sigma})\sum_{k=1}^{m}\|\mathbf{c}_k\|_2^2 \le m^2\Big(\max_{k\le m}\|\mathbf{c}_k\|_2^2\Big)\Big(\max_{k\le m}\Sigma_{k,k}\Big).$$

By Lemma A.1, the maximum of the variance-covariance matrix is bounded by

$$\max_{j,i\le m}\Sigma_{i,j} = \Sigma_{m,m} \le \frac{b^m}{n}(b+m)^m, \tag{23}$$

and by Lemma A.2,

$$\max_{k\le m}\|\mathbf{c}_k\|_2^2 \le \frac{2m+1}{b-a}\left(\left(\frac{b+a}{b-a} + \frac{b-a}{b+a}\right)\left(1 + \frac{2}{b+a}\right)\right)^{2m}. \tag{24}$$

Setting

$$\rho = \left[2\left(\frac{b+a}{b-a} + \frac{b-a}{b+a}\right)\left(1 + \frac{2}{b+a}\right)\right]^2$$

yields the upper bound

$$n\int_a^b \mathrm{Var}_f\big[\hat{f}_{n,m}(t)\big]\,ds \le \frac{(2m+1)m^2}{b-a}\big[(m+b)\,\rho b\big]^m. \qquad \square$$

Acknowledgments. This paper originated from my Ph.D. dissertation. I thank my advisor, P. Bickel, for suggesting the problem, helpful discussions and encouragement throughout the course of my Ph.D. at the University of California at Berkeley. This paper also has greatly benefited from constructive suggestions from two referees.

REFERENCES

BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

CENCOV, N. N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.* **3** 1559–1562.

DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation*. Springer, New York.

DONOHO, D. and LIU, R. (1991). Geometrizing rate of convergence, III. *Ann. Statist.* **19** 668–701.
FARRELL, R. H. (1967). On the lack of uniformly consistent sequences of estimators of a density function in certain cases. *Ann. Math. Statist.* **38** 471–474.
FARELL, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of density function at a point. *Ann. Math. Statist.* **43** 170–180.
HENGARTNER, N. W. (1994). Estimating linear functionals of the mixing distribution in Poisson mixture models. Unpublished manuscript.
LAMBERT, D. and TIERNEY, L. (1984). Asymptotic properties of maximum likelihood estimates in the mixed Poisson model. *Ann. Statist.* **12** 1388–1399.
LINDSAY, B. and ROEDER, K. (1993). Uniqueness of estimation and identifiability in mixture models. *Canad. J. Statist.* **21** 139–147.
LOH, W.-L. and ZHANG, C.-H. (1993). Global properties of kernel estimators for mixing densities in exponential family model for discrete variables. Technical Report 93-47, Dept. Statistics, Purdue Univ.
RAHMAN, Q. I. and SCHMEISSER, G. (1982). *Les inégalités de Markoff et de Bernstein*. Presses de l'Univ. Montréal.
RAINVILLE, E. (1967). *Special Functions*. Macmillan, New York.
SIMAR, L. (1976). Maximum likelihood estimations of a compound Poisson process. *Ann. Statist.* **4** 1200–1209.
STONE, C. (1980). Optimal rates of convergence of nonparametric estimators. *Ann. Statist.* **8** 1348–1360.
TEICHER, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32** 244–248.
TUCKER, H. G. (1963). An estimate of the compounding distribution of a compound Poisson distribution. *Theory Probab. Appl.* **8** 195–200.
WALTER, G. (1985). Orthogonal polynomials estimators of the prior distribution of a compound Poisson distribution. *Sankhyā Ser. A* **47** 222–230.
WALTER, G. and HAMEDANI, G. (1991). Bayes empirical Bayes estimation for natural exponential families with quadratic variance function. *Ann. Statist.* **19** 1191–1224.
ZHANG, C.-H. (1995). On estimating mixing densities in discrete exponential family models. *Ann. Statist.* **23** 929–947.

DEPARTMENT OF STATISTICS
YALE UNIVELRSITY
NEW HAVEN, CONNECTICUT 06520-8290
E-MAIL: hengart@stat.yale.edu