Statistica Sinica **19** (2009), 1297-1318

GENERALIZED MAXIMUM LIKELIHOOD ESTIMATION OF NORMAL MIXTURE DENSITIES

Cun-Hui Zhang

Rutgers University

Abstract: We study the generalized maximum likelihood estimator of location and location-scale mixtures of normal densities. A large deviation inequality is obtained which provides the convergence rate $n^{-p/(2+2p)}(\log n)^{\kappa_p}$ in the Hellinger distance for mixture densities when the mixing distributions have bounded finite p-th weak moment, $p > 0$, and the convergence rate $n^{-1/2}(\log n)^{\kappa}$ when the mixing distributions have an exponential tail uniformly. Our results are applicable to the estimation of the true density of independent identically distributed observations from a normal mixture, as well as the estimation of the average marginal densities of independent not identically distributed observations from different normal mixtures. The validity of our results for mixing distributions with p-th weak moment, $0 < p < 2$, and for not identically distributed observations, is of special interest in compound estimation and other problems involving sparse normal means.

Key words and phrases: Convergence rate, Hellinger distance, large deviation, maximum likelihood, mixture density, normal distribution.

1. Introduction.

In this paper we study the generalized maximum likelihood estimator (GMLE) of the average of marginal densities of observations from normal mixtures. Normal mixtures have been used in a broad range of applications and are related to many problems in statistics (Lindsay (1995), Genovese and Wasserman (2000) and Ghosal and van der Vaart (2001, 2007a,b)). A primary motivation of our investigation of the normal mixture is the compound estimation of normal means (Robbins (1951), Stein (1956), James and Stein (1961), Donoho and Johnstone (1994), Abramovich, Benjamini, Donoho and Johnstone (2006), Zhang (1997, 2003, 2005), Johnstone and Silverman (2004), and Jiang and Zhang (2007)), where the oracle Bayes rule can be explicitly expressed in terms of the average of the marginal densities of the observations (Robbins (1956), and Brown (1971)). In this compound estimation context, the average marginal density is a mixture of normal densities but the observations, which are independent not identically distributed (inid), are not generated from the mixture. Although little is known about the GMLE of the average marginal density based on inid observations, it contains the important, better understood special case of the estimation

of a normal mixture density based on independent identically distributed (i.i.d.) observations directly generated from the mixture.

There is a rich literature on the estimation of normal mixtures based on i.i.d. data, including the generalized maximum likelihood method (Kiefer and Wolfowitz (1956)) and its computational algorithms (Dempster, Laird and Rubin (1977) and Vardi and Lee (1993)), Fourier kernel estimators (Zhang (1997)), and more. In a broader context, nonparametric and sieve MLEs of density functions have been considered by many (van de Geer (1993, 1996), Shen and Wong (1994), Wong and Shen (1995), Genovese and Wasserman (2000) and Ghosal and van der Vaart (2001, 2007b)). Among these papers, Genovese and Wasserman (2000), and Ghosal and van der Vaart (2001, 2007b) directly study the estimation of a normal mixture density based on i.i.d. data. For location-scale mixing distributions with bounded support and a known lower bound on the scale, Genovese and Wasserman (2000) provides large deviation inequalities for sieve MLEs of the normal mixture density with convergence rates $n^{-1/6}(\log n)^{(1/6)+}$ or $n^{-1/4}(\log n)^{1/4}$ in the Hellinger distance, depending on the choice of sieves. These rates are improved to $n^{-1/2}(\log n)^{\kappa'}$ in Ghosal and van der Vaart (2001) for the GMLE when the mixing distribution has an exponential tail. For location mixing distributions with finite p-th weak moment, Genovese and Wasserman (2000) provides the convergence rates $n^{-1/4}\sqrt{\log n}$ or $(\log n)^{-p/2}\log\log n$, both for $p > 2$, depending on the choice of sieves, while direct use of the entropy calculations of Ghosal and van der Vaart (2001, 2007b) in the large deviation inequality of Wong and Shen (1995) yields, respectively, the convergence rates $n^{1/p-1/2}(\log n)^{1/2}$ for $p > 2$ and $n^{1/(2p)-1/2}\log n$ for $p > 1$.

In this paper, we establish a large deviation inequality which unifies and improves results in Genovese and Wasserman (2000) and Ghosal and van der Vaart (2001) and establishes the faster convergence rate of $n^{-p/(2+2p)}(\log n)^{\kappa_p}$ when the p-th weak moment is bounded. Our results also improve upon the logarithmic factor of the convergence rates in Ghosal and van der Vaart (2001, 2007b) when the mixing distribution has a heavier-than-normal exponential tail. Moreover, our results are valid for sparse mixing distributions with finite p-weak moment for small $0 < p < 2$ (Donoho and Johnstone (1994)) and in the more general inid setting, both crucial features for applications to the compound estimation and other problems involving sparse normal means.

In order to cover both the inid case and the more standard density estimation problem based on i.i.d. observations, we consider a general model in which the observations are independent and each observation is normally distributed given its latent conditional mean and variance. This includes the inid case of deterministic conditional means and variances and the i.i.d. case where the conditional means and variances are themselves i.i.d. vectors, among other possible data generating models.

We organize the paper as follows. Section 2 states our main theorem for inid observations from possibly different location mixtures with a known common scale. Section 3 contains extensions of the results in Section 2, including location-scale mixtures, deterministic and i.i.d. latent variables, sieve MLE, and more technical discussion of related work. Section 4 discusses the connection to compound estimation. Section 5 provides mathematical proofs.

2. The Main Theorem

In this section we consider inid observations from location mixtures with a known common scale. Our main theorem establishes a large deviation inequality and provides convergence rates of the GMLE to the average marginal densities of the observations. The implication of these results in other settings will be discussed in Section 3. We divide the materials into three subsections to describe the statistical model, the estimator, and the main theorem.

2.1. The inid location-mixture model

Our problem is best formulated in terms of latent location variables or the conditional means as follows. Let (X_i, θ_i) be independent random vectors with the conditional densities

$$X_i|\theta_i \sim \frac{1}{\sigma}\varphi\Big(\frac{x-\theta_i}{\sigma}\Big) \sim N(\theta_i, \sigma^2), \quad i = 1, \ldots, n, \tag{2.1}$$

under a probability measure P_n, where $\boldsymbol{X} \equiv (X_1, \ldots, X_n) \in \mathbb{R}^n$ is observable, $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_n)$ is the unknown vector of the conditional means of the observation $\boldsymbol{X}$, the variance $\sigma^2 > 0$ is known, and $\varphi(x) \equiv (2\pi)^{-1}e^{-x^2/2}$ is the standard normal density. We study the properties of the GMLE for the estimation of the average marginal density of the observations:

$$f_n(x) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{d}{dx}P_n\{X_i \le x\}. \tag{2.2}$$

This includes the cases of i.i.d. θ_i and completely deterministic θ_i under different choices of the probability measure P_n, since a deterministic sequence of constants can be treated as a sequence of degenerate random variables.

The average marginal density (2.2) can be explicitly written as a normal mixture density. Define the standardized normal location-mixture density as

$$h_G(x) \equiv \int \varphi(x-u)dG(u). \tag{2.3}$$

Let G_n be the average of the distribution functions of the standardized (deterministic or random) unknowns $\{\theta_1/\sigma, \ldots, \theta_n/\sigma\}$ under P_n:

$$G_n(u) \equiv \frac{1}{n}\sum_{i=1}^{n} P_n\Big\{\frac{\theta_i}{\sigma} \le u\Big\}. \tag{2.4}$$

Let $\Phi(x) \equiv \int_{-\infty}^{x} \varphi(z)dz$ be the standard normal cumulative distribution function. Since $P_n(X_i \le x) = \int \Phi(x/\sigma - u)dP\{\theta_i/\sigma \le u\}$, we have

$$f_n(x) = \frac{1}{\sigma} h_{G_n}\Big(\frac{x}{\sigma}\Big) = \int \frac{1}{\sigma}\varphi\Big(\frac{x-u}{\sigma}\Big)dG_n\Big(\frac{u}{\sigma}\Big) \tag{2.5}$$

as a mixture of the $N(\theta, \sigma^2)$ densities.

2.2. The GMLE

The GMLE, as suggested in Kiefer and Wolfowitz (1956), is defined as follows. Let $\mathscr{G}$ be the collection of all distributions in the real line $\mathbb{R}$ and define

$$\mathscr{H} \equiv \{h_G : G \in \mathscr{G}\} \tag{2.6}$$

as the family of all location-mixtures (2.3) of normal densities with the unit scale. Given $\boldsymbol{X}$, the GMLE of a normal mixture density with scale σ is

$$\hat{f}_n(x) \equiv \frac{1}{\sigma}\hat{h}_n\Big(\frac{x}{\sigma}\Big), \quad \hat{h}_n \equiv \operatorname*{argmax}_{h\in\mathscr{H}} \prod_{i=1}^{n} h\Big(\frac{X_i}{\sigma}\Big). \tag{2.7}$$

Here $\hat{h}_n$ is the member h in the class $\mathscr{H}$ in (2.6) that maximizes the "likelihood" $\prod_{i=1}^n h(X_i/\sigma)$ of the standardized observation $\boldsymbol{X}/\sigma$, even though $\prod_{i=1}^n h(x_i)$ is not necessarily the joint density of $\boldsymbol{X}/\sigma$ for any $h \in \mathscr{H}$ without the i.i.d. assumption.

Since the family $\mathscr{H}$ is itself indexed by the completely unknown mixing distribution G, given $\boldsymbol{X}$ the GMLE $\hat{f}_n$ in (2.7) is a mixture of normal densities of scale σ:

$$\hat{f}_n(x) = \frac{1}{\sigma} h_{\hat{G}_n}\Big(\frac{x}{\sigma}\Big), \quad \hat{G}_n \equiv \operatorname*{argmax}_{G\in\mathscr{G}} \prod_{i=1}^{n} h_G\Big(\frac{X_i}{\sigma}\Big), \tag{2.8}$$

where $\hat{G}_n$ is the GMLE of the mixing distribution G_n. Although the maximization is usually done over the parameter $G \in \mathscr{G}$ computationally, the formulation in (2.7) is more relevant for studying $\hat{f}_n$ as a density estimator since the properties of $\hat{f}_n$ are largely determined by the entropy of suitable sieves of the parameter space $\mathscr{H}$ as a density family.

The GMLE $\hat{f}_n$ is well defined since its rescaled version $\hat{h}_n$ in (2.7) is a well defined map from $\boldsymbol{X}/\sigma \in \mathbb{R}^n$ into $\mathscr{H}$ (Lindsay (1995), and Ghosal and van der Vaart (2001)). It follows from (2.3) and the definition of $\hat{G}_n$ in (2.8) that the support of $\hat{G}_n$ is always within the range of the standardized data $\{X_i/\sigma, i \le n\}$ due to the monotonicity of $\varphi(x-u)$ in $|x-u|$. Thus, $h_{\hat{G}_n}(X_i/\sigma)$, $i \le n$, are bounded away from zero and infinity for each vector $\boldsymbol{X}/\sigma$. This and the uniform smoothness of $h \in \mathscr{H}$ give the existence of the GMLE $\hat{h}_n$ and the

uniqueness of $\hat{h}_n(X_i/\sigma)$, $i \le n$, due to the convexity of $\mathscr{H}$ and the log-concavity of the likelihood $\prod_{i=1}^n h(X_i/\sigma)$ in $h \in \mathscr{H}$. The computation of the GMLE is typically carried out using iterative algorithms. For example, one may use the EM algorithm to maximize over the subfamily of all discrete distributions G supported on a fine grid in the range of the standardized data.

The GMLE in (2.7) and (2.8) is a sensible estimator for the average mixing density f_n, since the expected log-likelihood

$$E_n \log \prod_{i=1}^n h(X_i/\sigma) = \sum_{i=1}^n \int \log h(x) dP_n\Big\{\frac{X_i}{\sigma} \le x\Big\} = n \int \{\log h(x)\} h_{G_n}(x) dx$$

is uniquely maximized at $h = h_{G_n} \in \mathscr{H}$ and $f_n(x) = \sigma^{-1} h_{G_n}(x/\sigma)$ by (2.2) and (2.5). It is called GMLE in the following senses. First, the maximization in (2.7) is done over an infinite-dimensional parameter space $\mathscr{H}$ and the estimator $\hat{G}_n$ in (2.8) is completely nonparametric in the mixing distribution $G \in \mathscr{G}$. Second, in the general inid model (2.1), the individual X_i are not generated from any member of the family $\mathscr{F}_\sigma \equiv \{\sigma^{-1} h_G(x/\sigma) : G \in \mathscr{G}\}$ unless θ_i themselves are i.i.d. random variables.

2.3. Large deviation inequality and convergence rates

Define

$$d_H(f,g) \equiv \Big(\int \Big(\sqrt{f} - \sqrt{g} \Big)^2 \Big)^{1/2}$$

as the Hellinger distance between two densities f and g. Our main result provides a large deviation inequality for the Hellinger distance $d_H(\hat{f}_n, f_n)$ at a certain convergence rate ϵ_n depending explicitly on the weak moment of the average mixing distribution G_n in (2.4).

The p-th weak moment of a distribution function G is $\{\mu_p^w(G)\}^p$, where

$$\mu_p^w(G) \equiv \Big\{ \sup_{x>0} x^p \int_{|u|>x} G(du) \Big\}^{1/p}. \tag{2.9}$$

Due to the Markov inequality, the p-th weak moment is no greater than the standard p-th absolute moment: $\{\mu_p^w(G)\}^p \le \int |u|^p G(du)$. The convergence rate ϵ_n, as a function of the sample size n, the mixing distribution G, and the power p of the weak moment, is defined as

$$\epsilon(n, G, p) \equiv \max\Big[\sqrt{2\log n}, \Big\{ n^{1/p} \sqrt{\log n} \mu_p^w(G) \Big\}^{p/(2+2p)} \Big] \sqrt{\frac{\log n}{n}}. \tag{2.10}$$

Let $a_n \asymp b_n$ denote $a_n/b_n + b_n/a_n = O(1)$ throughout the paper.

Theorem 1. *Suppose (X_i, θ_i) are independent random vectors with the conditional distribution (2.1) under P_n, with a fixed $\sigma > 0$. Let f_n be the average marginal density in (2.2) and $\hat{f}_n$ be its GMLE in (2.7). Then, there exists a universal constant t_* such that for all $t \ge t_*$ and $\log n \ge 2/p$,*

$$P_n\Big\{d_H(\hat{f}_n, f_n) \ge t\epsilon_n\Big\} \le \exp\Big(-\frac{t^2 n\epsilon_n^2}{2\log n}\Big) \le e^{-t^2\log n}, \tag{2.11}$$

where $\epsilon_n \equiv \epsilon(n, G_n, p)$ is as in (2.10) with the average mixing distribution G_n in (2.4) and $p > 0$. In particular,

$$\epsilon_n \asymp \begin{cases} n^{-p/(2+2p)}(\log n)^{(2+3p)/(4+4p)}, & \mu_p^w(G_n) = O(1) \text{ for a fixed} p \\ n^{-1/2}(\log n)^{3/4}\Big\{M_n^{1/2} \vee (\log n)^{1/4}\Big\}, & G_n([-M_n, M_n]) = 1, p = \infty \\ n^{-1/2}(\log n)^{1/[2(2\wedge\alpha)]+3/4}, & \int e^{|cu|^\alpha} G_n(du) = O(1), p \asymp \log n, \end{cases}$$

where α and c are fixed in the third case.

The strength of the large deviation inequality (2.11) is evident from its uniformity in $\{p, t, \sigma\}$ and the explicit continuous dependence of the convergence rate ϵ_n on p and the weak moment. As a result, the convergence rates in Theorem 1 significantly improve upon the existing results, cf. Section 3.3.

Remark 1. It follows from (2.5) and (2.8) that $f_n(x) = \sigma^{-1}h_{G_n}(x/\sigma)$ and $\hat{f}_n(x) = \sigma^{-1}h_{\hat{G}_n}(x/\sigma)$ are all normal location mixtures, so that

$$d_H(\hat{f}_n, f_n) = d_H(h_{\hat{G}_n}, h_{G_n}), \tag{2.12}$$

due to the scale invariance of the Hellinger distance. It follows that Theorem 1 also implies the consistency of the GMLE $\hat{G}_n$ for the estimation of the average mixing distribution G_n in (2.4) through Fourier inversion. However, this consistency argument does not provide rates for the convergence of $\hat{G}_n$ in distribution, even for i.i.d. data.

Remark 2. According to the proofs in Section 5, the conclusions of Theorem 1 are also valid for any approximate GMLE $\hat{G}_n$ which guarantees

$$\prod_{i=1}^n \left\{\frac{h_{\hat{G}_n}(X_i/\sigma)}{h_{G_n}(X_i/\sigma)}\right\} \ge e^{-2t^2 n\epsilon_n^2/15}.$$

Remark 3. Since the constant t_* in Theorem 1 is universal, we are allowed to optimize over p to obtain

$$P_n\Big\{d_H(\hat{f}_n, f_n) \ge t\epsilon_{n,*}\Big\} \le \exp\Big(-\frac{t^2 n\epsilon_{n,*}^2}{2\log n}\Big) \le e^{-t^2\log n},$$

with $\epsilon_{n,*} \equiv \inf\{\epsilon(n, G_n, p) : \log n \geq 2/p\}$. This optimal p is approximately achieved at $p \asymp \log n$ under the condition $\int e^{|cu|^\alpha} G_n(du) = O(1)$.

Remark 4. Since the large deviation bound is non-asymptotic, Theorem 1 allows $\sigma = \sigma_n$ to depend on n. This is utilized in our discussion of sieve estimators in Section 3.

3. Consequences of the Main Theorem

The general inid formulation (2.1) and the possible dependence of the probability measure P_n on n allow applications of Theorem 1 in many settings, including location-scale mixtures, location mixtures with unknown scale, deterministic latent variables as in the compound estimation theory, and i.i.d. observations from normal mixtures. The large deviation inequality can be also used to study the GMLE (2.8) as a sieve estimator for a general smooth density function. These variations of Theorem 1 and related work are discussed in this section.

3.1. Location-scale mixture or location mixture with unknown scale

As in (2.1), the inid location-scale normal mixture model is best described by a sequence of independent random vectors (X_i, ξ_i, τ_i) with the following conditional densities under P_n:

$$X_i|(\xi_i, \tau_i) \sim \frac{1}{\tau_i}\varphi\Big(\frac{x - \xi_i}{\tau_i}\Big) \sim N(\xi_i, \tau_i^2), \quad \tau_i \geq \sigma, \tag{3.1}$$

where $\sigma > 0$ is a known lower bound for the latent scale variables. This includes the location model with an unknown common scale $\tau_i = \tau$ as long as we have the knowledge of the lower bound σ for the common τ.

As far as the densities of the observations X_i are concerned, the location-scale mixture model (3.1) is identical to the location mixture model (2.1). This can be seen as follows. Since the $N(\xi_i/\sigma, \tau_i^2/\sigma^2)$ density is the convolution of the $N(0, 1)$ and $N(\xi_i/\sigma, \tau_i^2/\sigma^2 - 1)$ densities, (3.1) implies

$$P_n\Big\{\frac{X_i}{\sigma} \leq x\Big\} = E_n\Phi\Big(\frac{x - \xi_i/\sigma}{\tau_i/\sigma}\Big) = E_n \int \varphi(x - u)d\Phi\left(\frac{u - \xi_i/\sigma}{\sqrt{\tau_i^2/\sigma^2 - 1}}\right)$$

with the convention $\Phi((u - \xi)/0) = I\{\xi \leq u\}$. Thus, (2.1) holds with

$$P_n\Big\{\frac{\theta_i}{\sigma} \leq u\Big\} = E_n\Phi\left(\frac{u - \xi_i/\sigma}{\sqrt{\tau_i^2/\sigma^2 - 1}}\right). \tag{3.2}$$

This gives the equivalence between the two models (2.1) and (3.1), since (3.1) is formally more general than (2.1). The equivalence of (2.1) and (3.1) naturally leads to our second theorem.

Theorem 2. *Suppose* (3.1) *holds under certain probability measures* P_n*, instead of* (2.1)*. Then all the conclusions of Theorem 1 hold with* $\epsilon_n \equiv \epsilon(n, G_n, p)$*, where*

$$G_n(u) = \frac{1}{n}\sum_{i=1}^n E_n\Phi\Big(\frac{u-\xi_i/\sigma}{\sqrt{\tau_i^2/\sigma^2-1}}\Big) = \frac{1}{n}\sum_{i=1}^n E_n\Phi\Big(\frac{u\sigma-\xi_i}{\sqrt{\tau_i^2-\sigma^2}}\Big). \tag{3.3}$$

Moreover, (2.11) *provides the convergence rate* $\epsilon_n \asymp n^{-p/(2+2p)}(\log n)^{(2+3p)/(4+4p)}$ *if*

$$\sup_{x>0}\frac{x^p}{n}\sum_{i=1}^n P_n\Big\{|\xi_i|>x\Big\} + \frac{1}{n}\sum_{i=1}^n E_n\tau_i^p = O(1) \tag{3.4}$$

for a fixed $p>0$*, while* $\epsilon_n \asymp n^{-1/2}(\log n)^{1/(2\alpha)+3/4}$ *for* $p \asymp \log n$ *if for some fixed* $0<\alpha\le 2$ *and* $c>0$*,* $n^{-1}\sum_{i=1}^n E_n(e^{|c\xi_i|^\alpha} + e^{(c\tau_i)^{\alpha/(1-\alpha/2)}}) = O(1)$*.*

Remark 5. For $\alpha=2$, we adopt the convention $e^{(c\tau_i)^{\alpha/(1-\alpha/2)}} = 0, =1$, or $=\infty$ when $c\tau_i<1, =1$, or >1, respectively.

Theorem 2 is useful for more explicit calculations of the convergence rates $\epsilon_n \equiv \epsilon(n, G_n, p)$ in terms of the moments of ξ_i and τ_i. It is also helpful for comparisons between our and existing results in Section 3.3, since the convergence rates for models (2.1) and (3.1) are different in Ghosal and van der Vaart (2001).

3.2. Deterministic latent location and scale variables

Since P_n is allowed to depend on n in Theorems 1 and 2, it can be also considered as the probability measure given the means $\boldsymbol{\theta}$ in (2.1) or given the means $\boldsymbol{\xi} \equiv (\xi_1,\dots,\xi_n)$ and standard deviations $\boldsymbol{\tau} \equiv (\tau_1,\dots,\tau_n)$ in (3.1). Here we treat the more general case of deterministic $\boldsymbol{\xi}$ and $\boldsymbol{\tau}$ in (3.1). Define

$$G_{n,\boldsymbol{\xi},\boldsymbol{\tau}}(u) \equiv \frac{1}{n}\sum_{i=1}^n \Phi\Big(\frac{u\sigma-\xi_i}{(\tau_i^2-\sigma^2)^{1/2}}\Big), \tag{3.5}$$

$$f_{n,\boldsymbol{\xi},\boldsymbol{\tau}}(x) \equiv n^{-1}\sum_{i=1}^n \frac{1}{\tau_i}\varphi\Big(\frac{x-\xi_i}{\tau_i}\Big). \tag{3.6}$$

Theorem 3. *Suppose* X_i*,* $i\le n$*, are independent* $N(\xi_i,\tau_i^2)$ *observations under* $P_{n,\boldsymbol{\xi},\boldsymbol{\tau}}$ *with unknown deterministic vectors* $\boldsymbol{\xi}$ *and* $\boldsymbol{\tau}$ *and a known lower bound* $0<\sigma\le\tau_i$*. Let* $\hat f_n$ *be the GMLE in* (2.7) *and* $f_{n,\boldsymbol{\xi},\boldsymbol{\tau}}(x)$ *be the average density of* X_i*,* $i\le n$*, in* (3.6)*. Then, for all* $t\ge t_*$ *and* $\log n \ge 2/p$*,*

$$P_{n,\boldsymbol{\xi},\boldsymbol{\tau}}\Big\{d_H(\hat f_n, f_{n,\boldsymbol{\xi},\boldsymbol{\tau}}) \ge t\epsilon_n\Big\} \le \exp\Big(-\frac{t^2 n\epsilon_n^2}{2\log n}\Big) \le e^{-t^2\log n},$$

where t_* is a universal constant and $\epsilon_n \equiv \epsilon(n, G_{n,\boldsymbol{\xi},\boldsymbol{\tau}}, p)$ is as in (2.10) *with the distribution $G_{n,\boldsymbol{\xi},\boldsymbol{\tau}}$ in* (3.5). *In particular, $\epsilon_n \asymp n^{-p/(2+2p)}(\log n)^{(2+3p)/(4+4p)}$ for a fixed $p > 0$ if $n^{-1}\sum_{i=1}^n (|\xi_i|^p + \tau_i^p) = O(1)$, while $\epsilon_n \asymp n^{-1/2}(\log n)^{1/(2\alpha)+3/4}$ for $p \asymp \log n$ if $n^{-1}\sum_{i=1}^n \exp\big(|c\xi_i|^\alpha + |c\tau_i|^{\alpha/(1-\alpha/2)}\big) = O(1)$ for some fixed $0 < \alpha \le 2$ and $c > 0$.*

Since $\boldsymbol{\xi}$ and $\boldsymbol{\tau}$ are teated as deterministic, Theorem 3 follows immediately from Theorem 2 with $P_n = P_{n,\boldsymbol{\xi},\boldsymbol{\tau}}$. Clearly, Theorem 3 is applicable to the case of deterministic means in (2.1) with $\xi_i = \theta_i$ and $\tau_i = \sigma$.

3.3. The i.i.d. case and related work

Let H denote a bivariate mixing distribution of location and scale. Suppose throughout this subsection that we observe

$$\text{i.i.d.} \quad X_i \sim f_H(x) \equiv \iint \frac{1}{\tau}\varphi\Big(\frac{x-\xi}{\tau}\Big) H(d\xi, d\tau) \tag{3.7}$$

under a probability $P_n = P_H$, and $H(\mathbb{R} \times [\sigma, \infty)) = 1$ with a known lower bound $\sigma > 0$ for the scale.

Since the inid model (3.1) is more general than (3.7), the large deviation inequality (2.11) in Theorems 1 and 2 provides $\epsilon_n \asymp n^{-p/(2+2p)}(\log n)^{(2+3p)/(4+4p)}$ as the convergence rate for the GMLE $\hat{f}_n$ under

$$\sup_{x>0} x^p P_H\{|\xi| > x\} + E_H \tau^p < \infty, \tag{3.8}$$

and $\epsilon_n \asymp n^{-1/2}(\log n)^{1/(2\alpha)+3/4}$ under $E_H \exp(|c\xi|^\alpha + |c\tau|^{\alpha/(1-\alpha/2)}) < \infty$ for some $0 < \alpha \le 2$ and $c > 0$.

For the estimation of f_H under (3.7) and the support bound $H([-M, M] \times [\sigma, \sigma^*]) = 1$ with finite M and σ^*, Genovese and Wasserman (2000) provides large deviation inequalities for certain sieve MLEs with convergence rates $n^{-1/6}(\log n)^{(1/6)+}$ or $n^{-1/4}(\log n)^{1/4}$ in the Hellinger distance, depending on the choice of sieves. For the Kullback-Leibler loss, Li and Barron (1999) obtained the convergence rate $n^{-1/4}(\log n)^{1/4}$. These rates are improved in Ghosal and van der Vaart (2001) to $n^{-1/2}(\log n)^{\kappa'}$ for the GMLE, under the exponential bound $E_H e^{|c\xi|^\alpha} < \infty$ and the support bound $H(\mathbb{R} \times [\sigma, \sigma^*]) = 1$ with finite σ^*, where $\kappa' = 1/2 + 1/(2 \vee \alpha)$ under the independence of location and scale $H(d\xi, d\tau) = H_1(d\xi)H_2(d\tau)$, and $\kappa' = 1/2 + 2/(2 \vee \alpha)$ is slightly worse in general. Under the same conditions, further improved entropy bounds in Ghosal and van der Vaart (2007b) lead to $\kappa' = 1 + 1/\{2(2 \vee \alpha)\}$ via the entropy integral of Wong and Shen (1995). Thus, Theorem 2 improves upon the power of the logarithmic factor of the convergence rates in Ghosal and van der Vaart (2001, 2007b) from κ' to $\kappa = 3/4 + 1/\{2(2 \vee \alpha)\}$, and extends the large deviation inequality to inid

observations and distributions with wider support. The right-hand side of (2.11) is of larger order than the $e^{-c(\log n)^2}$ in Ghosal and van der Vaart (2001), but the difference is minimal since both large deviation inequalities require large t.

Under (3.7) with fixed $\tau = \sigma$ and $(d/dx)P\{|\xi| \le x\} \propto x^{p+1}$ with $p > 2$, Genovese and Wasserman (2000) provides the convergence rates $n^{-1/4}\sqrt{\log n}$ or $(\log n)^{-p/2}\log\log n$ depending on the choice of sieves. Under the weak moment condition $x^p P\{|\xi| > x\} = O(1)$, the entropy bounds in Ghosal and van der Vaart (2001, 2007b) lead respectively to the convergence rates $n^{1/p-1/2}\sqrt{\log n}$ and $n^{1/(2p)-1/2}\log n$ via Wong and Shen (1995). In comparison, the better convergence rate $n^{-p/(2+2p)}(\log n)^{(2+3p)/(4+4p)}$ in Theorem 2 is a significant improvement. In particular, our theorems are applicable for small $0 < p < 2$, a case of particular interest in the compound estimation and other problems for sparse means, cf. Section 4.

3.4. GMLE as a sieve estimator

Let $\mathscr{F}_\sigma \equiv \{\sigma^{-1}h(x/\sigma) : h \in \mathscr{H}\}$ be the family of mixture normal densities with scale σ. For location-scale mixtures (3.7) without a lower bound for the scale, i.e. with $H(\mathbb{R}\times(0,\sigma]) > 0$ for all $\sigma > 0$, Genovese and Wasserman (2000) studies sieve MLEs with sieves inside $\mathscr{F}_{\sigma_n}$ without giving explicit convergence rates to f_H, where $\sigma_n \to 0$ slowly. In Priebe (1994), normal-mixture sieve MLEs exhibit more sparse solutions than standard kernel methods, so that the MLE over normal mixture sieves $\mathscr{F}_{\sigma_n}$ provides an attractive alternative to more conventional smoothing methods for density estimation. Here we provide a direct application of Theorem 1 to the GMLE (2.7) as a sieve estimator for a general dansity f_0, with $\sigma = \sigma_n \to 0$.

For densities $h_G \equiv \int \varphi(x-u)G(du)$ in (2.3) and f_0, define

$$q_0(G,\sigma,t) \equiv P_0\left\{ n^{-1}\sum_{i=1}^n \log\left(\frac{f_0(X_i)}{\sigma^{-1}h_G(X_i/\sigma)}\right) \ge t^2 \right\}.$$

Theorem 4. *Suppose X_i are i.i.d. variables with density f_0 under a probability measure P_0. Suppose $q_0(G_n,\sigma_n,s_n\epsilon_n) \to 0$ for certain distributions G_n and constants σ_n and s_n, where $\epsilon_n \equiv \epsilon(n,G_n,p_n)$ is as in (2.10) for certain $p_n \ge 2/\log n$. Let $f_n(x) = \sigma_n^{-1}h_{G_n}(x/\sigma_n)$, $\hat G_n$ be as in (2.8) with the data $\boldsymbol{X}/\sigma_n$, and $\hat f_n(x) = \sigma_n^{-1}h_{\hat G_n}(x/\sigma_n)$. Then, there exists a universal constant t_* such that for all $t \ge \max\{t_*, 2s_n\sqrt{\log n}\}$,*

$$\begin{aligned} &P_0\Big\{ d_H(\hat f_n, f_0) > t\epsilon_n + d_H(f_n,f_0)\Big\} \\ &\quad\le \exp\Big(-\frac{nt^2\epsilon_n^2}{(4\log n)}\Big) + q_0(G_n,\sigma_n,s_n\epsilon_n) \to 0. \end{aligned} \tag{3.9}$$

Since the observations X_i are i.i.d. random variables from the density f_0 under P_0, $\log(f_0(X_i)/f_n(X_i))$ are i.i.d. variables under P_0. Thus, sufficient conditions for $q_0\big(G_n, \sigma_n, s_n\epsilon_n\big) \to 0$ can be derived from the Weak Law of Large Numbers. The convergence rate in (3.9) is regulated through the tail and smoothness of the density f_0, respectively, in terms of the weak p-th moment of G_n in (2.10) and the bandwidth $\sigma_n > 0$. Since densities with unbounded support are considered, further investigation of sieve MLEs or other smoothing methods is beyond the scope of this paper.

4. Discussion

Although the estimation of normal mixture densities is a problem of important independent interest, a primary motivation of our investigation is the use of the GMLE in the compound estimation of normal means (Zhang (1997, 2003) and Jiang and Zhang (2007)) under the risk function

$$E_{n,\boldsymbol{\theta}}\left\{\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\theta}_i - \theta_i\Big)^2\right\}, \tag{4.1}$$

where $E_{n,\boldsymbol{\theta}}$ is the expectation under which (2.1) holds with deterministic parameter vector $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_n)$ and known $\sigma = 1$. Here we provide a brief technical discussion of this connection, especially the crucial role of the convergence of the GMLE under the p-th weak moment condition for $0 < p < 2$.

Since the normal density is a special case of the Laplace family discussed in Robbins (1956) and Brown (1971)), the oracle Bayes rule, which minimizes the compound risk (4.1) among all separate estimators of the form $\hat{\theta}_i = t(X_i)$, is given by

$$\hat{\theta}_i = t^*_{G_n}(X_i), \quad t^*_G(x) \equiv x + \frac{h'_G(x)}{h_G(x)}, \tag{4.2}$$

where h_G is as in (2.3), $h'_G \equiv (d/dx)h_G$, and G_n in (2.4) becomes the empirical distribution function of the unknown constants $\{\theta_1, \ldots, \theta_n\}$. The basic idea of the general empirical Bayes estimation (Robbins (1956)) is to approximate the oracle (4.2) using a nonparametric estimator of h_{G_n}. For example

$$\hat{\theta}_i = \hat{t}_n(X_i), \quad \hat{t}_n(x) \equiv x + \frac{h'_{\hat{G}_n}(x)}{\max\{h_{\hat{G}_n}(x), n^{-2}\}}, \tag{4.3}$$

provides a general empirical Bayes estimator with the GMLE $h_{\hat{G}_n}$ in (2.8).

An important problem in the compound estimation of normal means is the optimality for sparse means due to its connection to wavelet denoising and other nonparametric estimation problems (Donoho, Johnstone, Kerkyacharian and Picard (1995), Johnstone and Silverman (2005), and Zhang (2005)). It has been

pointed out in Donoho and Johnstone (1994) that a sparse mean vector $\boldsymbol{\theta}$ can be viewed as a member of small ℓ_p balls $\Theta_{n,p,C_n} \equiv \{\boldsymbol{\theta} : n^{-1}\sum_{i=1}^n |\theta_i|^p \le C_n^p\}$ with $0 < p < 2$. Thus, the optimality in the compound estimation of sparse normal means is commonly expressed as the simultaneous asymptotic minimaxity in a broad class of small ℓ_p balls (Donoho and Johnstone (1994), Abramovich, Benjamini, Donoho and Johnstone (2006), Johnstone and Silverman (2004), and Zhang (2005)).

Theorem 1 provides a uniform large deviation inequality for the convergence of the Hellinger distance $d_H(h_{\hat{G}}, h_{G_n})$ in the weak ℓ_p balls. This result plays a crucial role in our investigation of the compound estimation of sparse normal means, since the ℓ_p balls Θ_{n,p,C_n} are subsets of weak ℓ_p balls $\{\boldsymbol{\theta} : \mu_p^w(G_n) \le C_n\}$ by (2.9) and (2.4). In a forthcoming paper (Jiang and Zhang (2007)), the general empirical Bayes estimator (4.3) is demonstrated to provide superior numerical results compared with the extensive simulations in Johnstone and Silverman (2004) and is proved to possess the the simultaneous asymptotic minimaxity in all the weak ℓ_p balls of radii $(\log n)^{4+3/p+p/2}/n \ll C_n^p \ll n^p(\log n)^{4+9p/2}$, $0 < p < 2$, compared with Abramovich, Benjamini, Donoho and Johnstone (2006) and Johnstone and Silverman (2004). Here $a_n \ll b_n$ means $a_n/b_n \to 0$. This simultaneous asymptotic minimaxity is established using Theorem 1 along with an oracle inequality for Bayes rules with misspecified priors and a Gaussian isoperimetric inequality for certain regularized Bayes rules.

5. Proofs

Our proofs are closely related to those in Shen and Wong (1994), Wong and Shen (1995), and Ghosal and van der Vaart (2001, 2007b). Large deviation inequalities for sieve MLEs of a density are typically obtained via an entropy integral (Shen and Wong (1994), and Wong and Shen (1995)). A direct application of this method (Ghosal and van der Vaart (2001)) to the GMLE (2.7) provides convergence rates from the entropy integral for the family

$$\mathscr{H}_M \equiv \Big\{h_G \in \mathscr{H} : G([-M, M]) = 1\Big\} \tag{5.1}$$

at truncation levels $M = M_n$ satisfying $P_n\{\hat{h}_n \in \mathscr{H}_{M_n}\} \approx 1$, i.e., $1 - G_n([-M_n, M_n]) \asymp 1/n$. In the proof of Theorem 1 below, we use much smaller truncation levels M_n by approximating $\hat{h}_n(X_i/\sigma)$ with certain sieves for $|X_i/\sigma| \le M_n$, and bounding $\hat{h}_n(X_i/\sigma)$ with the constant $\varphi(0)$ for $|X_i/\sigma| > M_n$. This is a crucial difference between our and previous proofs, since a smaller truncation level $M = M_n$ leads to a smaller order of the entropy for the family (5.1), which then leads to faster convergence rates. Still, we use a modification of an exponential inequality in Wong and Shen (1995) and an improved version of the entropy bounds in Ghosal and van der Vaart (2001, 2007b).

We need three lemmas. The first one approximates a general normal mixture (2.3) by a certain discrete mixture under the the supreme norm in bounded intervals,

$$\|h\|_{\infty,M} \equiv \sup_{|x|\le M} |h(x)|, \quad M > 0. \tag{5.2}$$

It also considers the L_p norm $\|h\|_p \equiv (\int |h(x)|^p dx)^{1/p}$, $p = 1, \infty$, and the Hellinger distance $d_H(\cdot,\cdot)$ for mixing distributions with bounded support. Let $\lfloor x \rfloor$ denote the greatest integer lower bound and $\lceil x \rceil$ the smallest integer upper bound of x.

Lemma 1. *Let h_G denote the mixture density in* (2.3). *Let $a > 0$, $\eta = \varphi(a)$, and $M > 0$. Given any mixing distribution G, there exists a discrete mixing distribution G_m, with support $[-M - a, M + a]$ and at most $m = (2\lfloor 6a^2 \rfloor + 1)\lceil 2M/a + 2 \rceil + 1$ atoms, such that*

$$\|h_G - h_{G_m}\|_{\infty,M} \le \eta\Big\{1 + \frac{1}{(2\pi)^{1/2}}\Big\}. \tag{5.3}$$

Moreover, if $G([-M, M]) = 1$, then

$$\|h_G - h_{G_m}\|_p \le \eta, p = 1, \infty, \quad d_H(h_G, h_{G_m}) \le \eta, \tag{5.4}$$

for a certain discrete G_m with support $[-M, M]$ and at most m atoms, where $m \le C^* |\log \eta| \max(M/\sqrt{|\log \eta|}, 1)$ for a certain universal constant C^*.

Remark 6. Ghosal and van der Vaart (2007b) proved (5.4) with $m \le C^* |\log \eta| \max(M, 1)$. Compared with their proofs, we partition the interval $(-M-a, M+a]$ into larger subintervals to allow the smaller number of atoms m in (5.4).

Proof. Let $j^* = \lceil 2M/a + 2 \rceil$ and $k^* = \lfloor 6a^2 \rfloor$. Define semiclosed intervals

$$I_j = (-M + (j-2)a, (-M + (j-1)a) \wedge (M + a)\big], \quad j = 1, \ldots, j^*,$$

to form a partition of $(-M - a, M + a]$. It follows from Carathéodory's Theorem that there exists a discrete distribution function G_m, with support $[-M-a, M+a]$ and no more than $m = (2k^* + 1)j^* + 1$ support points, such that

$$\int_{I_j} u^k G(du) = \int_{I_j} u^k G_m(du), \quad k = 0, 1, \ldots, 2k^*, \quad j = 1, \ldots j^*. \tag{5.5}$$

We prove (5.3) for any G_m satisfying (5.5).

For $t^2/2 \le k^* + 2$, the alternating sign of the Taylor expansion of $e^{-t^2/2}$ yields

$$0 \le \mathrm{Rem(t)} \equiv (-1)^{\mathrm{k}^*+1}\Big\{\varphi(\mathrm{t}) - \sum_{\mathrm{k}=0}^{\mathrm{k}^*} \frac{(-\mathrm{t}^2/2)^{\mathrm{k}}}{\mathrm{k}!\sqrt{2\pi}}\Big\} \le \frac{(\mathrm{t}^2/2)^{\mathrm{k}^*+1}}{(\mathrm{k}^*+1)!\sqrt{2\pi}}.$$

Thus, since $k^*+1 \ge 6a^2$, for $x \in I_j \cap [-M,M]$, Stirling's formula yields

$$\begin{aligned}|h_G(x)-h_{G_m}(x)| &\le \left|\int_{(I_{j-1}\cup I_j\cup I_{j+1})^c} \varphi(x-u)\{G(du)-G_m(du)\}\right| \\ &\quad +\left|\int_{I_{j-1}\cup I_j\cup I_{j+1}} \mathrm{Rem(x-u)\{G(du)-G_m(du)\}}\right| \\ &\le \varphi(a)+\frac{\{(2a)^2/2\}^{k^*+1}}{\sqrt{2\pi}(k^*+1)!} \le \eta+\frac{(e/3)^{k^*+1}}{2\pi(k^*+1)^{1/2}}. \end{aligned} \tag{5.6}$$

Since $(e/3)^6 \le e^{-1/2}$ and $k^*+1 \ge 6a^2$, $(e/3)^{k^*+1} \le e^{-a^2/2}$, so that (5.6) implies (5.3).

For (5.4), we define $I_j = (-M+(j-1)a, -M+ja]$ and impose the constraints $\int_{I_j} u^k\{G(du)-G_m(du)\} = 0$ along with $G_m([-M,M]) = 1$ for all intergers $j \le j^*-2$, with $m = (2k^*+1)(j^*-2)+1$ atoms for G_m. The proof of (5.3) then yields $\|h_G - h_{G_m}\|_\infty \le \eta(1+1/\sqrt{2\pi})$. The L_1 bound follows by integrating (5.6) over x, which gives

$$\begin{aligned}\|h_G-h_{G_m}\|_1 &\le \int\left|\int_{|x-u|\ge a}\varphi(x-u)\{G(du)-G_m(du)\}\right|dx \\ &\quad +\eta\Big(1+\frac{1}{\sqrt{2\pi}}\Big)\iint_{|x-u|\le 2a}\{G(du)+G_m(du)\}dx \\ &\le 4\eta+4a\eta\Big(1+\frac{1}{\sqrt{2\pi}}\Big).\end{aligned}$$

Finally, $d_H^2(h_G,h_{G_m}) \le \|h_G-h_{G_m}\|_1$ gives the bound on the Hellinger distance.

Our second lemma provides an entropy bound for the family $\mathscr{H}$ in (2.6) and $\mathscr{H}_M$ in (5.1). For any semi-distance d_0 in $\mathscr{H} = \mathscr{H}_\infty$, the η-covering number of $\mathscr{H}_M$ is the smallest possible cardinality of its η-net: with $\mathrm{Ball}(h_0,\eta,d_0) \equiv \{h : d_0(h,h_0) < \eta\}$,

$$N(\eta,\mathscr{H}_M,d_0) \equiv \inf\Big\{N : \mathscr{H}_M \subseteq \cup_{j=1}^N \mathrm{Ball}(h_j,\eta,d_0)\Big\}.$$

Lemma 2. *There exists a universal constant C^* such that*

$$\log N(\eta,\mathscr{H},\|\cdot\|_{\infty,M}) \le C^*(\log\eta)^2\max\left(\frac{M}{\sqrt{|\log\eta|}},1\right) \tag{5.7}$$

for all $0<\eta\le(2\pi)^{-1/2}$ and $M>0$. Moreover,

$$\log N(\eta,\mathscr{H}_M,d_0) \le C^*(\log\eta)^2\max\left(\frac{M}{\sqrt{|\log\eta|}},1\right), \tag{5.8}$$

where d_0 is the Hellinger, L_1 or L_∞ distances.

Remark 7. Ghosal and van der Vaart (2007b) proved the upper bound

$$\log N(\eta, \mathscr{H}_M, d_0) \le C^* |\log \eta| \max(M, 1) \log\left(\frac{(M+1)}{\eta}\right)$$

with d_0 being the Hellinger, L_1 or L_∞ distances. Compared with their proof, the smaller upper bound (5.8) is a consequence of the smaller m in (5.4) and the cancellation of $\log(M+1)$ in an application of Stirling's formula in the proof below.

Proof. Let $a \equiv a_\eta \equiv \sqrt{2\log(2/\eta)}$ and

$$m = (2\lfloor 6a^2 \rfloor + 1)\lceil \frac{2M}{a} + 2\rceil + 1 \le \frac{C^*}{2} |\log \eta| \max\left(1, \frac{M}{\sqrt{|\log \eta|}}\right). \tag{5.9}$$

It follows from Lemma 1 that there exists a discrete distribution G_m with support $[-M-a, M+a]$ and at most m atoms such that

$$\|h_G - h_{G_m}\|_{\infty, M} \le \frac{\eta}{2}\left\{\frac{1}{(2\pi)^{1/2}} + \frac{1}{2\pi}\right\}. \tag{5.10}$$

The next step is to approximate the h_{G_m} in (5.10) by $h_{G_{m,\eta}}$ where $G_{m,\eta}$ is supported in a lattice and has no more than m atoms. Let θ be a variable with the distribution G_m and $\theta_\eta \equiv \eta, \operatorname{sgn}(\theta)\lfloor |\theta|/\eta \rfloor$, Define $G_{m,\eta}$ as the distribution of θ_η. Since $|\theta - \theta_\eta| \le \eta$,

$$\|h_{G_m} - h_{G_{m,\eta}}\|_\infty \le C_1^* \eta, \quad C_1^* \equiv \sup_x |\varphi'(x)| = (2e\pi)^{-1/2}. \tag{5.11}$$

The support of $G_{m,\eta}$ is in the grid $\Omega_{\eta,M} \equiv \{0, \pm\eta, \pm, 2\eta, \ldots\} \cap [-M - a_\eta, M + a_\eta]$.

The last step is to bound the covering number of the collection of all $h_{G_{m,\eta}}$. Let $\mathscr{P}^m$ be the set of all probability vectors $\boldsymbol{w} \equiv (w_1, \ldots, w_m)$ satisfying $w_j \ge 0$ and $\sum_{j=1}^m w_j = 1$. Let $\mathscr{P}^{m,\eta}$ be an η-net of $\mathscr{P}^m$:

$$\inf_{\boldsymbol{w}^{m,\eta} \in \mathscr{P}^{m,\eta}} \|\boldsymbol{w} - \boldsymbol{w}^{m,\eta}\|_1 \le \eta, \quad \forall \boldsymbol{w} \in \mathscr{P}^m,$$

with $|\mathscr{P}^{m,\eta}| = N(\eta, \mathscr{P}^m, \|\cdot\|_1)$. Let $\{\theta_j, j = 1, \ldots, m\}$ be the support of $G_{m,\eta}$ and $\boldsymbol{w}^{m,\eta}$ be a probability vector in $\mathscr{P}^{m,\eta}$ satisfying $\sum_{j=1}^m |G_{m,\eta}(\{\theta_j\}) - w_j^{m,\eta}| \le \eta$. Then,

$$\left\| h_{G_{m,\eta}} - \sum_{j=1}^m w_j^{m,\eta} \varphi(x - \theta_j) \right\|_\infty \le C_0^* \sum_{j=1}^m \left| G_{m,\eta}(\{\theta_j\}) - w_j^{m,\eta} \right| \le C_0^* \eta,$$

where $C_0^* \equiv \sup_x \varphi(x) = 1/\sqrt{2\pi}$. Thus, by (5.10) and (5.11),

$$\left\| h_G - \sum_{j=1}^m w_j^{m,\eta} \varphi(x - \theta_j) \right\|_{\infty, M} \le \left(\frac{1/2}{\sqrt{2\pi}} + \frac{1/2}{2\pi} + C_1^* + C_0^* \right) \eta \le \eta. \tag{5.12}$$

Counting the number of ways to realize $\boldsymbol{w}^{m,\eta}$ and $\{\theta_j\}$ in (5.12), we find

$$N(\eta, \mathscr{H}, \|\cdot\|_{\infty,M}) \le N(\eta, \mathscr{P}^m, \|\cdot\|_1)\binom{|\Omega_{\eta,M}|}{m}, \tag{5.13}$$

with m satisfying (5.9) and $|\Omega_{\eta,M}| = 1 + 2\lfloor (M+a)/\eta \rfloor$.

Since $\mathscr{P}^m$ is in the ℓ_1 unit sphere of $\mathbb{R}^m$, $N(\eta, \mathscr{P}^m, \|\cdot\|_1)$ is no greater than the maximum number of disjoint $\text{Ball}(\boldsymbol{v}_j, \eta/2, \|\cdot\|, \|\cdot\|_1)$ with $\|\boldsymbol{v}_j\|_1 = 1$. Since all these balls are inside the ℓ_1 $(1+\eta/2)$-sphere, volume comparison yields $N(\eta, \mathscr{P}^m, \|\cdot\|_1) \le (2/\eta+1)^m$. This and (5.13) imply, via Stirling's formula,

$$\begin{aligned} N(\eta, \mathscr{H}, \|\cdot\|_{\infty,M}) &\le \frac{(2/\eta+1)^m |\Omega_{\eta,M}|^m}{m!} \\ &\le \left\{\Big(1+\frac{2}{\eta}\Big)\Big(1+\frac{2(M+a)}{\eta}\Big)\right\}^m \Big\{(m+1)^{m+1/2}e^{-m-1}\sqrt{2\pi}\Big\}^{-1} \\ &\le \left[\frac{(\eta+2)(\eta+2(M+a))e}{(m+1)}\right]^m \eta^{-2m} e\{2\pi(m+1)\}^{-1/2}. \end{aligned} \tag{5.14}$$

For $\eta = o(1)$, $a = \sqrt{2\log(2/\eta)} \to \infty$, so that $m \ge (1+o(1))24a(M+a) \to \infty$ and $(\eta+2)(\eta+2(M+a))e = (1+o(1))4e(M+a) \le m+1$. Thus, $N(\eta, \mathscr{H}, \|\cdot\|_{\infty,M}) \le \eta^{-2m}$ by (5.14). This and (5.9) imply (5.7). The proof of (5.8) is similar to that of (5.4) and is omitted.

Lemma 3. *Let (X_i, θ_i) be independent random vectors with the conditional distributions $X_i|\theta_i \sim N(\theta_i, 1)$ under P_n. Let G_n and $\mu_p^w(G)$ be as in (2.4) and (2.9), respectively, with $\sigma = 1$. Then for all constants $M \ge \sqrt{8\log n}$, $0 < \lambda \le \min(1, p/2)$ and $a > 0$,*

$$E_n\Big\{\prod_{i=1}^n |aX_i|^{I\{|X_i|\ge M\}}\Big\}^\lambda \le \exp\left[2(aM)^\lambda\left\{\frac{2/M}{\sqrt{2\pi}} + n\Big(\frac{2\mu_p^w(G_n)}{M}\Big)^p\right\}\right].$$

Proof. Since $\sum_{i=1}^n Eh(X_i) = n\int h(x)h_{G_n}(x)dx$,

$$\begin{aligned} E_n\Big\{\prod_{i=1}^n |aX_i|^{I\{|X_i|\ge M\}}\Big\}^\lambda &\le \prod_{i=1}^n \Big(1 + a^\lambda E_n|X_i|^\lambda I\{|X_i|\ge M\}\Big) \\ &\le \exp\left\{|a|^\lambda n\int_{|x|\ge M}|x|^\lambda h_{G_n}(x)dx\right\}. \end{aligned} \tag{5.15}$$

Let $Z \sim N(0,1)$ and $\theta \sim G_n$ under P_n. Since $Z+\theta \sim h_{G_n}$ and $\lambda \le 1$,

$$\begin{aligned} \int_{|x|\ge M}|x|^\lambda h_{G_n}(x)dx &= E_n|Z+\theta|^\lambda I\{|Z+\theta|\ge M\} \\ &\le E_n|2Z|^\lambda I\Big\{|Z|\ge \frac{M}{2}\Big\} + E_n|2\theta|^\lambda I\Big\{|\theta|\ge\frac{M}{2}\Big\} \end{aligned}$$

$$\leq 2M^{\lambda-1}E_n|Z|I\Big\{|Z| \geq \frac{M}{2}\Big\} + \int_{|x|\geq \frac{M}{2}} (2|x|)^\lambda G_n(dx). \quad (5.16)$$

Let $C_p \equiv \mu_p^w(G_n)$. It follows from integrating by parts and (2.9) that

$$\begin{aligned}\int_{|x|\geq \frac{M}{2}} (2|x|)^\lambda G_n(dx) &\leq M^\lambda \int_{|x|>\frac{M}{2}} G_n(dx) + 2^\lambda \int_{\frac{M}{2}}^\infty \int_{|u|>x} G_n(du)dx^\lambda \\ &\leq \frac{M^\lambda C_p^p}{(M/2)^p} + 2^\lambda \int_{\frac{M}{2}}^\infty \frac{C_p^p}{x^p}\frac{\lambda}{x^{1-\lambda}}dx = \frac{M^{\lambda-p}2^p C_p^p p}{(p-\lambda)}.\end{aligned}$$

Since $\lambda \leq p/2$ and $M \geq \sqrt{8\log n}$, inserting the above inequality into (5.16) yields

$$\begin{aligned}\int_{|x|\geq M} |x|^\lambda h_{G_n}(x)dx &\leq \frac{4M^\lambda}{M}\int_{\frac{M}{2}}^\infty x\varphi(x)dx + 2M^\lambda \frac{2^p C_p^p}{M^p} \\ &\leq \frac{4M^\lambda}{Mn\sqrt{2\pi}} + 2M^\lambda \frac{2^p C_p^p}{M^p}.\end{aligned}$$

This and (5.15) imply the conclusion.

Proof of Theorem 1. Since the Hellinger distance is scale invariant as in (2.12), we assume $\sigma^2 = 1$ without loss of generality. Since $d_H^2(f,g) \leq 2$, (2.11) automatically holds for $t\epsilon_n \geq t_*\epsilon_n > 2$. Since $\epsilon_n \geq (\log n)\sqrt{2/n}$ by (2.10) and t_* is large, it suffices to consider large n.

Let $C_p \equiv \mu_p^w(G_n)$, $\eta \equiv 1/n^2$, and $M \equiv 2n\epsilon_n^2/(\log n)^{3/2}$. Define

$$h^*(x) \equiv \eta I\{|x| \leq M\} + (\eta M^2)x^{-2}I\{|x| > M\}. \quad (5.17)$$

We first find an upper bound on the likelihood ratio $L_n(h_{\hat{G}_n}, h_{G_n})$ in the event $\{d_H(h_{\hat{G}_n}, h_{G_n}) \geq t\epsilon_n\}$, where for all positive functions h_1 and h_2,

$$L_n(h_1,h_2) = \prod_{i=1}^n \Big\{\frac{h_1(X_i)}{h_2(X_i)}\Big\}.$$

We consider any approximate GMLE that guarantees

$$L_n(h_{\hat{G}_n}, h_{G_n}) = \prod_{i=1}^n \Big\{\frac{h_{\hat{G}_n}(X_i)}{h_{G_n}(X_i)}\Big\} \geq e^{-2t^2 n\epsilon_n^2/15}. \quad (5.18)$$

Let $\{h_j, j \leq N\}$ be an η-net of $\mathscr{H}$ under the seminorm $\|\cdot\|_{\infty,M}$, (5.2), as in Lemma 2, with $N \equiv N(\eta, \mathscr{H}, \|\cdot\|_{\infty,M})$. Let $h_{0,j}$ be densities satisfying

$$h_{0,j} \in \mathscr{H}, \quad d_H(h_{0,j}, h_{G_n}) \geq t\epsilon_n, \quad \|h_{0,j} - h_j\|_{\infty,M} \leq \eta, \quad (5.19)$$

if they exist, and $J \equiv \{j \le N : h_{0,j} \text{ exists } \}$. For any $h \in \mathscr{H}$ with $d_H(h, h_{G_n}) \ge t\epsilon_n$, there exists $j \in J$ such that

$$h(x) \le \begin{cases} h_{0,j}(x) + 2\eta = h_{0,j}(x) + 2h^*(x), & |x| \le M \\ \varphi(0) = \dfrac{1}{\sqrt{2\pi}}, & |x| > M, \end{cases}$$

due to $h^*(x) = \eta$ for $|x| \le M$ and $\sup_{h\in\mathscr{H}} h(x) = \varphi(0)$. It follows that

$$L_n(h_{\hat{G}_n}, h_{G_n}) \le \sup_{j\in J} L_n(h_{0,j} + 2h^*, h_{G_n}) \prod_{|X_i|\ge M} \frac{(2\pi)^{-1/2}}{2h^*(X_i)}$$

in the event $\{d_H(h_{\hat{G}_n}, h_{G_n}) \ge t\epsilon_n\}$. Thus, by (5.18),

$$\begin{aligned} &P_n\Big\{d_H(h_{\hat{G}_n}, h_{G_n}) \ge t\epsilon_n\Big\} \\ &\quad\le P_n\bigg\{ \sup_{j\in J} L_n(h_{0,j} + 2h^*, h_{G_n}) \prod_{|X_i|\ge M} \frac{(2\pi)^{-1/2}}{2h^*(X_i)} \ge e^{-2t^2n\epsilon_n/15}\bigg\} \\ &\quad\le P_n\bigg\{ \bigg[\sup_{j\in J} \prod_{i=1}^n \frac{h_{0,j}(X_i) + 2h^*(X_i)}{h_{G_n}(X_i)}\bigg] \ge e^{-4t^2n\epsilon_n^2/5}\bigg\} \\ &\qquad + P_n\bigg\{ \prod_{|X_i|\ge M} \frac{(2\pi)^{-1/2}}{2h^*(X_i)} \ge e^{2t^2n\epsilon_n^2/3}\bigg\}. \end{aligned} \tag{5.20}$$

Next, we derive large deviation inequalities for the right-hand side of (5.20). For $j \in J$, a modification of the proof of Lemma 1 of Wong and Shen (1995) yields

$$\begin{aligned} &P_n\bigg\{ \prod_{i=1}^n \frac{h_{0,j}(X_i) + h^*(X_i)}{h_{G_n}(X_i)} \ge e^{-4nt^2\epsilon_n^2/5}\bigg\} \\ &\quad\le e^{2nt^2\epsilon_n^2/5} \prod_{i=1}^n E_n \sqrt{\frac{\{h_{0,j}(X_i) + 2h^*(X_i)\}}{h_{G_n}(X_i)}} \\ &\quad\le \exp\bigg\{ \frac{2nt^2\epsilon_n^2}{5} + \sum_{i=1}^n E_n\bigg(\sqrt{\frac{\{h_{0,j}(X_i) + 2h^*(X_i)\}}{h_{G_n}(X_i)}} - 1\bigg)\bigg\} \\ &\quad= \exp\bigg\{ \frac{2nt^2\epsilon_n^2}{5} + n\bigg(\int \sqrt{(h_{0,j} + 2h^*)h_{G_n}} - 1\bigg)\bigg\}. \end{aligned} \tag{5.21}$$

Since $\int h^* = 4\eta M$ by (5.17) and $d_H(h_{0,j}, h_{G_n}) \ge t\epsilon_n$,

$$\int \sqrt{(h_{0,j} + 2h^*)h_{G_n}} - 1 \le -\frac{d_H^2(h_{0,j}, h_{G_n})}{2} + \sqrt{2\int h^*}$$

$$\leq -\frac{(t\epsilon_n)^2}{2} + \sqrt{8\eta M}.$$

Since $|J| \leq N$, the above inequality and (5.21) yield

$$P_n\left\{ \sup_{j\in J} \prod_{i=1}^{n} \frac{h_{0,j}(X_i)+2h^*(X_i)}{h_{G_n}(X_i)} \geq e^{-4nt^2\epsilon_n^2/5} \right\} \leq \exp\left(\log N + n\sqrt{8\eta M} - \frac{nt^2\epsilon_n^2}{10} \right). \tag{5.22}$$

Since $N \equiv N(\eta, \mathscr{H}, \|\cdot\|_{\infty,M})$ with $\eta \equiv 1/n^2$ and $M \equiv 2n\epsilon_n^2/(\log n)^{3/2} \geq 4\sqrt{\log n}$ by (2.10), Lemma 2 provides

$$\begin{aligned} \log N + n\sqrt{8\eta M} &\leq C^*(2\log n)^2 \max\left(\frac{M}{\sqrt{2\log n}}, 1\right) + \sqrt{8M} \\ &\leq \left\{\frac{(t^*)^2}{40}\right\} M(\log n)^{3/2} \leq \left(\frac{t^2}{20}\right) n\epsilon_n^2 \end{aligned}$$

for large n and $t^* \leq t$. Thus, by (5.22)

$$P_n\left\{ \sup_{j\in J} \prod_{i=1}^{n} \frac{h_{0,j}(X_i) + 2h^*(X_i)}{h_{G_n}(X_i)} \geq e^{-4nt^2\epsilon_n^2/5} \right\} \leq e^{-nt^2\epsilon_n^2/20}. \tag{5.23}$$

By (5.17), $1/h^*(x) = x^2/(\eta M^2) = (nx/M)^2$ for $|x| \geq M$, so that

$$P_n\left\{ \prod_{|X_i|\geq M} \frac{(2\pi)^{-1/2}}{2h^*(X_i)} \geq e^{2nt^2\epsilon_n^2/3} \right\} \leq e^{-2nt^2\epsilon_n^2/(3\log n)} E\left\{ \prod_{|X_i|\geq M} \left|\frac{nX_i}{M}\right| \right\}^{1/\log n}. \tag{5.24}$$

Since $M \equiv 2n\epsilon_n^2/(\log n)^{3/2} \geq \sqrt{16\log n}$ by (2.10) and $\log n \geq 2/p$ by assumption, Lemma 3 is applicable with $a \equiv n/M$ and $\lambda \equiv 1/\log n \leq 1$, yielding

$$E\left\{ \prod_{|X_i|\geq M} \left|\frac{nX_i}{M}\right| \right\}^{1/\log n} \leq \exp\left(\frac{e}{\sqrt{2\pi\log n}} + 2en\left(\frac{2C_p}{M}\right)^p \right). \tag{5.25}$$

Since $M/2 = n\epsilon_n^2/(\log n)^{3/2}$ and $C_p \equiv \mu_p^w(G_n)$, (2.10) gives

$$\frac{n\epsilon_n^2/\log n}{n(2C_p/M)^p} = \frac{(\epsilon_n\sqrt{n/\log n})^{2(1+p)}}{n(\sqrt{\log n}C_p)^p} \geq 1.$$

Therefore, (5.24) and (5.25) give

$$P_n\left\{ \prod_{|X_i|\geq M} \frac{(2\pi)^{-1/2}}{2h^*(X_i)} \geq e^{2nt^2\epsilon_n^2/3} \right\} \leq e^{-(2t^2/3-2e)n\epsilon_n^2/\log n + e/\sqrt{2\pi\log n}}.$$

Inserting this inequality and (5.23) into (5.20), we find that (2.11) holds for large n and $t_* \le t$, since $n\epsilon_n^2/\log n \ge 2\log n$ by (2.10).

The rate $\epsilon_n \asymp n^{-p/(2+2p)}(\log n)^{(2+3p)/(4+4p)}$ is clear from (2.10) under $\mu_p^w(G_n) = O(1)$. The rate $\epsilon_n \asymp n^{-1/2}(\log n)^{3/4}\{M_n^{1/2} \vee (\log n)^{1/4}\}$ also follows immediately from (2.10) under $G_n([-M_n, M_n]) = 1$. If $\int e^{|cu|^\alpha} G_n(du) = O(1)$ for certain positive c and $\alpha \le 2$, then

$$\mu_{\alpha k}^w(G_n) \le \frac{1}{c}\biggl\{\int |cu|^{\alpha k} G_n(du)\biggr\}^{1/(\alpha k)} \le \frac{1}{c}\biggl\{k!\int e^{|cu|^\alpha} G_n(du)\biggr\}^{1/(\alpha k)} = O(1)k^{1/\alpha},$$

so that (2.10) with $p = \log n \asymp \alpha k$ yields

$$\sqrt{\frac{n}{\log n}}\epsilon_n \asymp n^{1/(2+2p)}(\log n)^{p/(4+4p)}\Bigl(k^{1/\alpha}\Bigr)^{p/(2+2p)} \asymp (\log n)^{1/4+1/(2\alpha)}.$$

This completes the proof.

Proof of Theorem 2. Due to the equivalence of (2.1) and (3.1), it suffices to translate the rate $\epsilon_n \equiv \epsilon(n, G_n, p)$ into functionals of the moments of ξ_i and τ_i, where G_n is as in (3.3). Again we assume $\sigma = 1$ without loss of generality. By (3.2), we may write $\theta_i|(\xi_i, \tau_i) \sim N(\xi_i, \tau_i^2 - 1)$, so that $\theta_i = \xi_i + Z_i\sqrt{\tau_i^2 - 1}$, where Z_i are i.i.d. $N(0,1)$ random variables independent of (ξ_i, τ_i). For the p-th weak moment, (3.4) implies

$$\begin{aligned}\{\mu_p^w(G_n)\}^p &\le \sup_{x>0}\frac{x^p}{n}\sum_{i=1}^n P_n\{|\theta_i + Z_i\tau_i| > x\}\\ &\le \sup_{x>0}\frac{x^p}{n}\sum_{i=1}^n P_n\Bigl\{|\theta_i| > \frac{x}{2}\Bigr\} + \frac{2^p}{n}\sum_{i=1}^n E_n|Z_i|^p E_n\tau_i^p = O(1).\end{aligned}$$

Moreover, $|(Z_i/2)c\tau_i|^\alpha \le \max\{(Z_i/2)^2, (c\tau_i)^{\alpha/(1-\alpha/2)}\}$ implies

$$E_n e^{|c\theta_i/4|^\alpha} \le E_n\Bigl(e^{|c\xi_i/2|^\alpha} + e^{|c\tau_i|^{\alpha/(1-\alpha/2)}} + e^{Z_i^2/4}\Bigr).$$

Hence, $n^{-1}\sum_{i=1}^n E_n\bigl(e^{|c\xi_i|^\alpha} + e^{(c\tau_i)^{\alpha/(1-\alpha/2)}}\bigr) = O(1)$ implies

$$\int e^{|cu/4|^\alpha} G_n(du) = n^{-1}\sum_{i=1}^n E_n e^{|c\theta_i/4|^\alpha} = O(1),$$

and the conclusions follow from Theorem 1.

Proof of Theorem 4. Define

$$A_n \equiv \biggl\{\prod_{i=1}^n \frac{f_0(X_i)}{\sigma_n^{-1}h_{G_n}(X_i/\sigma_n)} \ge e^{ns_n^2\epsilon_n^2}\biggr\}.$$

Since the Hellinger distance is scale invariant as in (2.12), the left-hand side of (3.9) is no greater than

$$P_0\Big\{d_H(h_{\hat{G}_n}, h_{G_n}) \geq t\epsilon_n, A_n^c\Big\} + P_0\{A_n\}$$
$$\leq e^{ns_n^2\epsilon_n^2} P_n\Big\{d_H(h_{\hat{G}_n}, h_{G_n}) \geq \epsilon_n, A_n^c\Big\} + q_0(G_n, \sigma_n, s_n\epsilon_n),$$

where P_n is the probability under which X_i are i.i.d. variables with common density $\sigma_n^{-1}h_{G_n}(x/\sigma_n)$. Hence, (3.9) follows from Theorem 1.

References

Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584-653.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42**, 855-903.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Donoho, D. L. and Johnstone, I. (1994). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields* **99**, 277-303.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B* **57**, 301-369.

Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28**, 1105-1127.

Ghosal, S. and van der Vaart, A. W. (2001). Entropy and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29**, 1233-1263.

Ghosal, S. and van der Vaart, A. W. (2007a). Convergence rates of posterior distributions for noni.i.d. observations. *Ann. Statist.* **35**, 192-223.

Ghosal, S. and van der Vaart, A. W. (2007b). Posterior convergence rates for Dirichlet mixtures at smooth densities. *Ann. Statist.* **35**, 697-723.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1**, 361-379. Univ. of California Press, Berkeley.

Jiang, W. and Zhang, C.-H. (2007). General maximum likelihood empirical Bayes estimation of normal means. Technical Report 2007-007, Department of Statistics and Biostatistics, Rutgers University.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and hay in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32**, 1594-1649.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33**, 1700-1752.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887-906.

Li, J. and Barron, A. (1999). Mixture density estimation. In *Advances in Neural Information Processings Systems* **12** (Edited by S. A. Solla, T. K. Leen and K.-R. Muller) Morgan Kaufmann Publishers, San Fransisco.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications.* IMS, Hayward, CA.

Priebe, C. (1994). Adaptive mixtures. *J. Amer. Statist. Assoc.* **89**, 796-806.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math. Statist. Probab.* **1**, 131-148. Univ. of California Press, Berkeley.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1**, 157-163. Univ. of California Press, Berkeley.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580-615.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1**, 157-163. Univ. of California Press, Berkeley.

van de Geer, S. (1993). Hellinger consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 14-44.

van de Geer, S. (1996). Rates of convergence for the maximum likelihood estimator in mixture models. *J. Nonparametr. Stat.* **6**, 293-310.

Vardi, Y. and Lee, D. (1993). From image deblurring to optimal investment: Maximum likelihood solutions for positive linear inverse problem (with discussion). *J. Roy. Statist. Soc. Ser. B* **55**, 569-612.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23**, 339-362.

Zhang, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statist. Sinica* **7**, 181-193.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes method. *Ann. Statist.* **33**, 379-390.

Zhang, C.-H. (2005). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.* **33**, 54-100.

Department of Statistics and Biostatistics, Hill Center, Busch Campus, Rutgers University, Piscataway, NJ 08854, U.S.A.

E-mail: czhang@stat.rutgers.edu

(Received March 2008; accepted July 2008)