Записки научных
семинаров ПОМИ
Том 501, 2021 г.

Yibo Wang, N. A. Stepanova

ESTIMATING THE AMOUNT OF SPARSITY IN TWO-POINT MIXTURE MODELS

Abstract. We consider the problem of estimating the fraction of nonzero means in a sparse normal mixture model in the region where variable selection is possible. The focus is on the situation in which the proportion of nonzero means is very small. The proposed estimator is shown to be nearly rate optimal in the asymptotically minimax sense. Using this estimator, one can also consistently estimate the sparsity parameter in sparse normal mixtures, whose knowledge, in particular, is required to carry out the so-called almost full variable selection procedure. The advantage of using the new estimator is illustrated analytically and numerically. The obtained results can be extended to some nonnormal mixtures.

§1. Introduction

In this paper, we consider the problem of estimating the fraction of nonzero means in a two-point normal mixture model when the nonzero means are sparse and only moderately large. The results can be extended to some other mixture models whose tail probabilities are similar to those of the normal mixture model. Our study is partially motivated by recent results on variable selection in sparse mixture models and also by the publications of Meinshausen and Rice (see [17]) and Cai et al. (see [3]) who studied a similar estimation problem, originated from a signal detection problem, that occurred in astrophysics.

Consider the problem of recovering the components of a vector $X = (X_1, \dots, X_n)$ in the Gaussian sequence model

$$X = m + \xi, \tag{1}$$

where $\xi = (\xi_1, \dots, \xi_n) \sim N(0, I_{n \times n})$ and the mean vector $m = (m_1, \dots, m_n)$ is a sparse vector of the form $m = M_n \eta$ with $M_n > 0$ and $\eta = (\eta_1, \dots, \eta_n) \in$

Key words and phrases: sparse normal mixture, fraction of nonzero means, minimax estimation, selection region.

$\mathcal{H}_{n,\beta}$; the set $\mathcal{H}_{n,\beta}$ is defined for some constants $0 < c \leqslant 1 \leqslant C < \infty$ by

$$\mathcal{H}_{n,\beta} = \Big\{\eta = (\eta_1, \dots, \eta_n) : \eta_j \in \{0,1\},\ cn^{1-\beta} \leqslant \sum_{j=1}^{n} \eta_j \leqslant Cn^{1-\beta}\Big\}. \quad (2)$$

When the constants c and C in (2) depend on n and obey the same asymptotics, that is, $c = 1 + o(1)$ and $C = 1 + o(1)$ as $n \to \infty$, then the fraction of nonzero means in model (1) satisfies

$$n^{-1} \sum_{j=1}^{n} \eta_j \sim n^{-\beta}.$$

Therefore, in this case, $n^{-\beta}$ may be viewed as the fraction of nonzero components of vector $m = \mathbf{E}(X)$ in model (1). If, in addition, we assume that the parameter M_n has the form $M_n = \sqrt{2r \ln n}$ for some $r \in (0,4)$, then model (1) entails the two-point normal mixture model

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} (1 - \varepsilon_n) N(0,1) + \varepsilon_n N(\mu_n, 1), \quad (3)$$

where $\varepsilon_n = n^{-\beta}$ for $\beta \in (0,1)$ and $\mu_n = \sqrt{2r \ln n}$ for $r \in (0,4)$. The fact that vector X in (1) contains relatively few nonzero components that are only moderately large makes the problems of identifying nonzero means and estimating their fraction non-trivial. The normal mixture model (3) when $\beta \in (1/2, 1)$ and $r \in (0,1)$ has been studied in a number of publications dealing with high-dimensional inference problems (see, for example, [3, 7, 8, 11]).

The problem of identifying nonzero mean components of a vector X in model (1) (or, equivalently, identifying nonzero components of a vector $\eta \in \mathcal{H}_{n,\beta}$) is typically tackled by providing a suitable measurable function $\widetilde{\eta} = \widetilde{\eta}(X_1, \dots, X_n)$ of $(X_1, \dots, X_n)$ taking its values in $\{0,1\}^n$ and called a selector. A standard way to judge the quality of a given selector $\widetilde{\eta} = (\widetilde{\eta}_j)_{j=1}^n$ is to look at its Hamming risk

$$\mathbf{E}_\eta |\widetilde{\eta} - \eta| := \mathbf{E}_\eta \sum_{j=1}^{n} |\widetilde{\eta}_j - \eta_j|.$$

If the parameter M_n in model (1) satisfies

$$\liminf_{n \to \infty} \frac{M_n}{\sqrt{\ln n}} > \sqrt{2\beta}, \quad (4)$$

then the selector $\widehat{\eta}$ given by

$$\widehat{\eta} = (\widehat{\eta}_j)_{j=1}^n, \quad \widehat{\eta}_j = \mathbb{I}\left(X_j > \sqrt{(2\beta+\Delta)\ln n}\right), \quad j = 1, \dots, n, \tag{5}$$

where $\Delta = \Delta_n > 0$ is such that $\Delta \to 0$ and $\Delta \ln n \to \infty$ as $n \to \infty$, satisfies (see Theorem 4 in [6] and Theorem 9 in [12])

$$\limsup_{n\to\infty} \sup_{\eta \in \mathcal{H}_{n,\beta}} n^{\beta-1} \mathbf{E}_\eta |\widehat{\eta} - \eta| = 0,$$

and thus provides the so-called "almost full" variable selection with respect to the maximum Hamming risk. (A selector is called almost full if its maximum risk is algebraically small as compared to the number of nonzero means.) However, if

$$\limsup_{n\to\infty} \frac{M_n}{\sqrt{\ln n}} < \sqrt{2\beta}, \tag{6}$$

then (see Theorem 5 in [6] and Theorem 10 in [12])

$$\liminf_{n\to\infty} \inf_{\widetilde{\eta}} \sup_{\eta \in \mathcal{H}_{n,\beta}} n^{\beta-1} \mathbf{E}_\eta |\widetilde{\eta} - \eta| > 0,$$

that is, variable selection is impossible. If we assume that $M_n = \sqrt{2r\ln n}$, then inequality (4) reduces to $r > \beta$.

This work is largely motivated by the problem of variable selection in model (3) and also by the findings of [3] and [17] that suggest a reasonable estimator of the fraction ε_n of nonzero means in model (3), in which $\varepsilon_n = n^{-\beta}$ for $\beta \in (1/2, 1)$ and $\mu_n = \sqrt{2r\ln n}$ for $r \in (\rho(\beta), 1)$, where the function $\rho(\beta)$ is given by

$$\rho(\beta) = \begin{cases} \beta - 1/2, & 1/2 < \beta \leqslant 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta < 1. \end{cases} \tag{7}$$

The curve $r = \rho(\beta)$, known in the literature as a *detection boundary*, was found by Ingster (see [14]) who showed that if $r > \rho(\beta)$ then the hypotheses

$$H_0 : X_1, \dots, X_n \overset{\text{iid}}{\sim} N(0,1)$$

and

$$H_{1,n} : X_1, \dots, X_n \overset{\text{iid}}{\sim} (1-\varepsilon_n)N(0,1) + \varepsilon_n N(\mu_n, 1)$$

separate asymptotically, whereas if $r < \rho(\beta)$ then H_0 and $H_{1,n}$ merge asymptotically. The region $\mathcal{D} = \{(\beta, r) \in \mathbb{R}^2 : 1/2 < \beta < 1,\, \rho(\beta) < r < 1\}$ where H_0 and $H_{1,n}$ separate asymptotically is called the detection region. In this paper, we propose an estimator $\widehat{\varepsilon}_n$ of the fraction $\varepsilon_n = n^{-\beta}$ of nonzero means in model (3) that is consistent in the selection region $\mathcal{S} =$

$\{(\beta, r) \in \mathbb{R}^2 : 0 < \beta < 1,\, \beta < r < 4\}$. By the time when the result of [3] on consistent estimation of ε_n inside the detection region $\mathcal{D}$ was published, the existence and shape of the selection region $\mathcal{S}$ have not yet become a common knowledge, hence the focus was on estimating ε_n in $\mathcal{D}$. The new estimator $\widehat{\varepsilon}_n$ proposed in this work is nearly rate optimal and, in the selection region, improves the estimator $\widehat{\varepsilon}^*_{a_n}$ of ε_n introduced by formula (2.8) in [3].

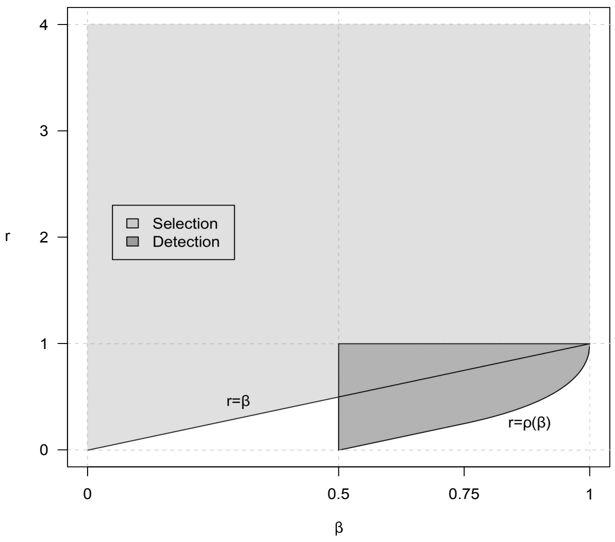

Figure 1. The selection and detection regions, $\mathcal{S}$ and $\mathcal{D}$, as described in Section 1.

Some notations used throughout the paper are as follows. The symbol $\overset{d}{=}$ is used for equality in distribution, and the symbol $\xrightarrow{d}$ denotes convergence in distribution. For an event A, $\mathbb{I}(A)$ is the indicator of the event A. The notation $a_n \sim b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$, whereas the notation $a_n \asymp b_n$ means that $0 < \liminf_{n\to\infty}(a_n/b_n) \leqslant \limsup_{n\to\infty}(a_n/b_n) < \infty$.

§2. Estimation of ε_n in the two-point normal mixture model

As noted in Section 1 of [3], the theory of testing $H_0 : X_1, \dots, X_n \overset{\text{iid}}{\sim} N(0,1)$ versus $H_{1,n} : X_1, \dots, X_n \overset{\text{iid}}{\sim} (1-\varepsilon_n)N(0,1) + \varepsilon_n N(\mu_n, 1)$, where $\varepsilon_n = n^{-\beta}$ for $\beta \in (1/2, 1)$ and $\mu_n = \sqrt{2r \ln n}$ for $r \in (0,1)$, does not automatically yield good estimators of ε_n as the problem of estimating ε_n contains further challenges that are not presented in the above signal detection problem. For a more general testing problem, that includes the signal detection problem in hand as a particular case, Meinshausen and Rice (see [17]) proposed an estimator of ε_n, the proportion of false null hypotheses among a large number of independently tested hypotheses, that estimates ε_n from below with high probability. In continuation of work [17], an estimator of ε_n that is consistent in the detection region $\mathcal{D} = \{(\beta, r) \in \mathbb{R}^2 : 1/2 < \beta < 1,\, \rho(\beta) < r < 1\}$ and, moreover, that is nearly rate optimal in a subregion of $\mathcal{D}$ was constructed in [3]. The proposed estimation procedure, however, is rather complex and is not straightforward to use in applications (see Section 2.3 for details). Also, in order to work as designed, it requires very large n ($n \geqslant 10^7$).

In this work, we modify the proposed in [3] estimator $\widehat{\varepsilon}^*_{a_n}$ of ε_n in such a way that a new estimator, which is applicable in a variable selection framework, is: (i) consistent in the selection region $\mathcal{S}$; (ii) easier implementable; (iii) more efficient (i.e., has a better rate of convergence) in the intersection $\mathcal{D} \cap \mathcal{S} = \{(\beta, r) \in \mathbb{R}^2\,;\, 1/2 < \beta < r < 1\}$ of the detection and selection regions, where the estimator $\widehat{\varepsilon}^*_{a_n}$ introduced by (2.8) in [3] is also defined.

Observe that, in order to be applied, the almost full selector $\widehat{\eta}$ as in (5) requires the knowledge of β. This motivates us for finding an effective estimator for β in the context of variable selection. Clearly, once we get a good estimator $\widetilde{\varepsilon}_n$ for $\varepsilon_n = n^{-\beta}$, we immediately obtain a good estimator $\widetilde{\beta}_n$ for β in the form

$$\widetilde{\beta}_n = \frac{\ln(1/\widetilde{\varepsilon}_n)}{\ln n}.$$

The goal, therefore, is to construct an efficient estimator of ε_n in a two-point normal mixture model (3). This model is obtained from model (1), in which $M_n = \sqrt{2r \ln n}$ and $n^{-1} \sum\limits_{j=1}^{n} \eta_j \sim n^{-\beta}$, by sampling with replacement. The parameters β and r are unknown.

The problem of estimating the fraction ε_n of nonzero means may be viewed as a high-dimensional problem, in which a single vector $X = (X_1, \ldots, X_n)$ of a large dimension n is observed. In the present context, one can only give a non-trivial lower bound for ε_n, and cannot give a useful upper bound for ε_n. Roughly, this can be explained by noting that the possibility that $\varepsilon_n = 1$ can never be ruled out because the nonzero means can be arbitrarily close to zero. For a detailed discussion of this phenomenon, we refer to pp. 2423–2424 of [3] and references therein. Therefore, the estimator $\widehat{\varepsilon}_n$ introduced in this work will *underestimate* the true ε_n with high probability.

In Section 2.2, we shall propose a modification of the estimator from [3] and provide an analytical result pertaining to this estimator (more precisely, a family of estimators), showing its superiority over the original estimator when $(\beta, r) \in \mathcal{D} \cap \mathcal{S}$, which, in its turn, was found to be better than the estimator proposed in [17] (for details, see Sections 6 and 7 of [3]).

2.1. Preliminaries. Return to the two-point normal mixture model (3) and denote by Φ and φ the cdf and pdf of a standard normal distribution, respectively. In terms of a common cdf $F(t) = F_{n,\beta,r}(t)$ of the observations $X_1, \ldots, X_n$, the model can be written as follows:

$$F(t) = (1 - \varepsilon_n)\Phi(t) + \varepsilon_n \Phi(t - \mu_n), \quad t \in \mathbb{R}, \tag{8}$$

where the parameters μ_n and ε_n are as in model (3). The estimation procedure proposed in [3] (that is referred here to as the Cai–Jin–Low (CJL) estimator) first estimates the mean μ_n, and then uses the estimated mean to estimate ε_n, the fraction of nonzero mean observations among $X_1, \ldots, X_n$. The algorithm for constructing the CJL estimator $\widehat{\varepsilon}^*_{a_n}$ of ε_n, as given by (2.8) in [3], resembles the one for $\widehat{\varepsilon}_n$ below, but it has a different step 2 (for details, see Section 2.2 of [3]). A key step in the construction of $\widehat{\varepsilon}^*_{a_n}$ is the choice of an $100(1 - \alpha)\%$ confidence band for the cdf $F(t)$. In [3], the proposed confidence band $[\mathbb{F}^-_{a_n}(t), \mathbb{F}^+_{a_n}(t)]$ on $[0, \sqrt{2 \ln n}]$ is chosen so that $\mathbb{F}^-_{a_n}(t) \leqslant F(t) \leqslant \mathbb{F}^+_{a_n}(t)$ if and only if $\frac{\sqrt{n}|\mathbb{F}_n(t) - F(t)|}{\sqrt{F(t)(1-F(t))}} \leqslant a_n$, where $\mathbb{F}_n(t) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(X_i \leqslant t)$ and a_n is the $(1 - \alpha)$th quantile of $\sup_{t \in [0, \sqrt{2 \ln n}]} \frac{\sqrt{n}|F_n(t) - F(t)|}{\sqrt{F(t)(1-F(t))}}$, see page 2427 of [3]. The lower and upper bounds of this confidence band are obtained by solving (for $F(t)$) the equation

$$\frac{\sqrt{n}|\mathbb{F}_n(t)-F(t)|}{\sqrt{F(t)(1-F(t))}}=a_n,$$

and are given by

$$\mathbb{F}^{\pm}_{a_n}(t)=\frac{2\mathbb{F}_n(t)+a_n^2/n\pm(a_n/\sqrt{n})\sqrt{a_n^2/n+4\left(\mathbb{F}_n(t)-\mathbb{F}_n^2(t)\right)}}{2(1+a_n^2/n)}. \tag{9}$$

By construction, the CJL estimator $\widehat{\varepsilon}^*_{a_n}$ underestimates ε_n whenever $F(t)$ lies inside the confidence band $[\mathbb{F}^-_{a_n}(t),\mathbb{F}^+_{a_n}(t)]$ given by (9).

In this paper, we suggest to modify the CJL estimator $\widehat{\varepsilon}^*_{a_n}$ for ε_n by using a different confidence band in the definition of $\widehat{\varepsilon}^*_{a_n}$. In addition to that, as we consider estimating ε_n in the selection region $\mathcal{S}$, the parameters β and r are allowed to range over the intervals $0<\beta<1$ and $0<r<4$, as specified by model (3).

Instead of using the confidence band with the lower and upper bounds as in (9), we propose to use the *Csörgő–Csörgő–Horváth–Mason* (*CsCsHM*) *confidence band* for $F(t)$ on the interval $[X_{(1)},X_{(n)})$, where $X_{(1)}=\min(X_1,\dots,X_n)$ and $X_{(n)}=\max(X_1,\dots,X_n)$, introduced in [20]. This confidence band is based on Theorem 4.2.3 of [5]. Namely, let $\{B(u),0\leqslant u\leqslant 1\}$ be a Brownian bridge and let the function $q(u)$ be the *Erdős–Feller–Kolmogorov–Petrovski (EFKP) upper-class function* of a Brownian bridge given by

$$q(u)=\sqrt{u(1-u)\ln\ln\left(1/\left(u(1-u)\right)\right)},\quad 0<u<1. \tag{10}$$

Then, by Theorem 4.2.3 of [5], as $n\to\infty$

$$\sup_{0<F(t)<1}\frac{\sqrt{n}|\mathbb{F}_n(t)-F(t)|}{q(F(t))}\xrightarrow{d}\sup_{0<u<1}\frac{|B(u)|}{q(u)}. \tag{11}$$

Now, with c_α denoting the $(1-\alpha)$th quantile of the limit cdf

$$H(t):=\mathbf{P}\left(\sup_{0<u<1}|B(u)|/q(u)\leqslant t\right), \tag{12}$$

an asymptotically correct $100(1-\alpha)\%$ CsCsHM confidence band

$$[\mathbb{F}^-_{n,\alpha}(t),\mathbb{F}^+_{n,\alpha}(t)]$$

for $F(t)$ on the interval $[X_{(1)}, X_{(n)})$ is defined as (see Section 3.2 of [20])

$$\mathbb{F}^{-}_{n,\alpha}(t) = \max\left\{0, \mathbb{F}_n(t) - \frac{c_\alpha}{\sqrt{n}}\, q\left(\mathbb{F}_n(t)\right)\right\},$$

$$\mathbb{F}^{+}_{n,\alpha}(t) = \min\left\{1, \mathbb{F}_n(t) + \frac{c_\alpha}{\sqrt{n}}\, q\left(\mathbb{F}_n(t)\right)\right\}. \tag{13}$$

The limit cdf $H(t)$ is continuous on $(-\infty, \sqrt{2}) \cup (\sqrt{2}, \infty)$ (see Remark 4.2.3 in [5]). Given a small $\alpha \in (0,1)$, the value of c_α may be obtained from Table III in [18] (see also Table 2 in [9]). For instance, $c_{0.05} = 4.57$.

In [20], the asymptotically correct $100(1-\alpha)\%$ confidence band

$$[\mathbb{F}^{-}_{n,\alpha}(t), \mathbb{F}^{+}_{n,\alpha}(t)]$$

was numerically compared to two common confidence bands (at the same level of confidence): one is based on the convergence result for the two-sided Kolmogorov–Smirnov test statistics and the other is based on the 1979 results of Eicker and Jaeschke (see [10] and [15]). Numerical simulations showed that, even for moderate sample sizes, when compared to the Kolmogorov–Smirnov confidence band, the CsCsHM confidence band is of the same length "in the middle" and is shorter on the tails. Also, the CsCsHM confidence band outperforms the Eicker–Jaeschke confidence band "in the middle" and does a similar job on the tails. It is also important for us that the CsCsHM confidence band with the lower and upper bounds given by (13) works better as compared the confidence band in (9) proposed in [3]. As seen from Figure 2, the CsCsHM confidence band $[\mathbb{F}^{-}_{n,\alpha}(t), \mathbb{F}^{+}_{n,\alpha}(t)]$ outperforms the confidence band $[\mathbb{F}^{-}_{a_n}(t), \mathbb{F}^{+}_{a_n}(t)]$ from [3]. An analytical argument in favour of using $[\mathbb{F}^{-}_{n,\alpha}(t), \mathbb{F}^{+}_{n,\alpha}(t)]$ instead of the CJL confidence band $[\mathbb{F}^{-}_{a_n}(t), \mathbb{F}^{+}_{a_n}(t)]$ in our construction below is given in Remark 1 of Section 3.

2.2. A family of estimators and upper bound on the risk. The construction of a new estimator $\widehat{\varepsilon}_n$ of ε_n is implemented by going through steps 1 to 4 of the following algorithm, cf. the procedure on pages 2426–2428 of [3].

Algorithm for the construction of $\widehat{\varepsilon}_n$.

1. Pick two distinct points $0 \leqslant t < t'$ and consider the function

$$D(\mu; t, t') := \frac{\Phi(t) - \Phi(t-\mu)}{\Phi(t') - \Phi(t'-\mu)}. \tag{14}$$

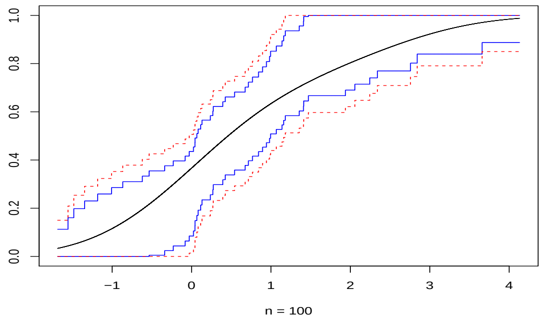

Figure 2. Confidence bands for simulated data. The solid black line is the true cdf of a normal mixture model with $\beta = 1/3$ and $r = 3/4$. The solid blue lines above and below the black line are a 99% CsCsHM confidence band. The red dashed lines are a 99% CJL confidence band.

The function $D(\mu; t, t')$ is a positive and continuous function of μ. Moreover, $D(\mu; t, t')$ is strictly decreasing in $\mu > 0$ for any $t < t'$ (see Lemma 8.1 of [2]). Hence, in view of (8), the parameters ε_n and μ_n are uniquely determined by

$$\varepsilon_n = \frac{\Phi(t) - F(t)}{\Phi(t) - \Phi(t - \mu_n)} \quad \text{and} \quad D(\mu_n; t, t') = \frac{\Phi(t) - F(t)}{\Phi(t') - F(t')}, \tag{15}$$

respectively. Also, it is easy to see that for $t < t'$

$$\inf_{\mu > 0} D(\mu_n; t, t') = \frac{\Phi(t)}{\Phi(t')} < \frac{\Phi(t) - F(t)}{\Phi(t') - F(t')} < \sup_{\mu > 0} D(\mu_n; t, t') = \frac{\varphi(t)}{\varphi(t')}. \tag{16}$$

2. Pick a small $\alpha \in (0, 1)$ and consider $[\mathbb{F}^-_{n,\alpha}(t), \mathbb{F}^+_{n,\alpha}(t)]$, an asymptotically correct $100(1-\alpha)\%$ confidence band for $F(t)$ on $[X_{(1)}, X_{(n)})$ defined by formula (13). Take two points

$$0 < t < t' < \min\left(X_{(n)}, \sqrt{8 \ln n}\right)$$

and define $\widehat{\mu}_n = \widehat{\mu}_{n,\alpha}$ as a solution (if exists) of the equation

$$D(\mu; t, t') = \frac{\Phi(t) - \mathbb{F}^+_{n,\alpha}(t)}{\Phi(t') - \mathbb{F}^-_{n,\alpha}(t')}. \tag{17}$$

With high probability, one has $\widehat{\mu}_n \geqslant \mu_n$ for all large enough n. Then, we set

$$\widehat{\varepsilon}_n = \widehat{\varepsilon}_{n,\alpha} = \frac{\Phi(t) - \mathbb{F}^+_{n,\alpha}(t)}{\Phi(t) - \Phi(t - \widehat{\mu}_n)} \tag{18}$$

if the solution $\widehat{\mu}_n$ exists and $\widehat{\varepsilon}_n = 0$ otherwise. As easily seen, with high probability, one has $\widehat{\varepsilon}_n \leqslant \varepsilon_n$ for all large enough n.

3. Consider the interval $\left[0, \sqrt{8 \ln n}\right]$, pick equidistant grid points

$$t_j = (j-1)/\sqrt{8 \ln n}, \quad 1 \leqslant j \leqslant 8 \ln n + 1,$$

and let the (random) index $1 \leqslant J_0 \leqslant 8 \ln n + 1$ be such that

$$t_{J_0+1} < \min\left(X_{(n)}, \sqrt{8 \ln n}\right) \leqslant t_{J_0+2}.$$

For $j = 1, \ldots, J_0$, apply the procedure in step 2 to each pair of adjacent points $(t, t') = (t_j, t_{j+1})$, and define the estimator $\widehat{\mu}^j_n = \widehat{\mu}^j_{n,\alpha} = \widehat{\mu}_{n,\alpha}(t_j, t_{j+1})$ of μ_n as in step 2, that is, as a solution (if exists) of the equation

$$D(\mu; t_j, t_{j+1}) = \frac{\Phi(t_j) - \mathbb{F}^+_{n,\alpha}(t_j)}{\Phi(t_{j+1}) - \mathbb{F}^-_{n,\alpha}(t_{j+1})}. \tag{19}$$

For each obtained estimator $\widehat{\mu}^j_n$, one needs to check whether or not the following inequalities hold true (see (16) and the comment on p. 2427 of [3]):

$$\frac{\Phi(t_j)}{\Phi(t_{j+1})} \leqslant \frac{\Phi(t_j) - \mathbb{F}^+_{n,\alpha}(t_j)}{\Phi(t_{j+1}) - \mathbb{F}^-_{n,\alpha}(t_{j+1})} \leqslant \frac{\varphi(t_j)}{\varphi(t_{j+1})}. \tag{20}$$

If (20) is not satisfied, then equation (19) does not give a good estimator $\widehat{\mu}^j_n$ for μ_n and in step 4 below we set $\widehat{\varepsilon}^j_n = 0$. If the inequalities in (20) are satisfied, we put

$$\widehat{\varepsilon}^j_n = \widehat{\varepsilon}_{n,\alpha}(t_j, t_{j+1}) = \frac{\Phi(t_j) - \mathbb{F}^+_{n,\alpha}(t_j)}{\Phi(t_j) - \Phi(t_j - \widehat{\mu}^j_n)}. \tag{21}$$

By construction, with high probability we have $\widehat{\varepsilon}^j_n \leqslant \varepsilon_n$ provided n is large enough.

4. Finally, having obtained the estimators $\widehat{\varepsilon}^j_n$, $j = 1, \ldots, J_0$, we define an estimator $\widehat{\varepsilon}_n$ of ε_n by the formula

$$\widehat{\varepsilon}_n = \widehat{\varepsilon}_{n,\alpha} = \max_{1 \leqslant j \leqslant J_0} \widehat{\varepsilon}^j_n. \tag{22}$$

Clearly, for all large enough n, one has $\widehat{\varepsilon}_n \leqslant \varepsilon_n$ with high probability.

We wish to stress that the estimator $\widehat{\varepsilon}_n$ in (22) is dependent on α, where $(1-\alpha)$ is a prescribed confidence level of the CsCsHM confidence band in step 2 of our construction, yielding

$$\lim_{n\to\infty} \mathbf{P}\left(\widehat{\varepsilon}_n \leqslant \varepsilon_n\right) \geqslant 1-\alpha.$$

In the sequel, we shall write $\widehat{\varepsilon}_n$ instead of $\widehat{\varepsilon}_{n,\alpha}$, suppressing for brevity the dependence of $\widehat{\varepsilon}_n$ on the value of α.

Noting that by means of (16) the right inequality in (20) holds true with high probability, we agree with the recommendation on page 2427 of [3] that the left inequality in (20) must also be satisfied as we have seen the importance of this requirement in our empirical study (see Section 4). We also wish to note that, in practice, the left inequality in (20) consistently holds only for very large n such as $n = 10^7$ and greater, and that is the main reason why both the original CJL estimator $\widehat{\varepsilon}^*_{a_n}$ and the new estimator $\widehat{\varepsilon}_n$ are "strongly asymptotic".

In connection with the proposed estimator $\widehat{\varepsilon}_n$, we have the following upper bound on the maximum risk of $\widehat{\varepsilon}_n/\varepsilon_n$.

Theorem 1. *Let $X_1, X_2, \ldots$ be a sequence of iid random variables from a continuous distribution with cdf $F(t)$ as in* (8)*, where $\varepsilon_n = n^{-\beta}$ for $0<\beta<1$ and $\mu_n = \sqrt{2r\ln n}$ for $\beta<r<4$, so that (β, r) falls in the selection region $\mathcal{S}$. Given α, let c_α be such that $H(c_\alpha) = 1-\alpha$, where $H(t)$ is the limit cdf in* (12)*. Assume that α_n is a sequence such that $c_{\alpha_n} = \left(\frac{4\ln n}{\ln\ln 4}\right)^{1/2}$, $n \geqslant 2$, and consider the estimator $\widehat{\varepsilon}_n = \widehat{\varepsilon}_{n,\alpha_n}$ defined in* (22)*. Then for all large enough n*

$$\mathbf{E}\left(\frac{\widehat{\varepsilon}_n}{\varepsilon_n} - 1\right)^2 \leqslant C\left(\beta, r\right)(\ln n)^2(\ln\ln n)n^{-1+\beta}, \tag{23}$$

where $C(\beta, r)$ is a positive constant depending on β and r.

Theorem 1 says that, with α_n tending to zero at a certain rate, for all $(\beta, r) \in \mathcal{S}$ and all sufficiently large n there exists a positive constant $C(\beta, r)$ such that for $0<\beta<1$ and $\beta<r<4$ the estimator $\widehat{\varepsilon}_n = \widehat{\varepsilon}_{n,\alpha_n}$ satisfies

$$\sup_{(\varepsilon_n,\mu_n)\in\mathcal{P}_{n,\beta,r}} \mathbf{E}\left(\frac{\widehat{\varepsilon}_n}{\varepsilon_n} - 1\right)^2 \leqslant C(\beta, r) r_n^2,$$

where for $0 < \beta < 1$ and $0 < r < 4$

$$\mathcal{P}_{n,\beta,r} = \left\{ (\varepsilon_n, \mu_n) : \varepsilon_n = n^{-\beta}, \mu_n = \sqrt{2r \ln n} \right\}, \tag{24}$$

and $r_n = \left(n^{-(1-\beta)}(\ln\ln n)\ln^2 n\right)^{1/2}$ is the rate of convergence of the ratio $\widehat{\varepsilon}_n/\varepsilon_n$ to one. The proof of Theorem 1 will be given in the next section.

The corresponding upper bound on the maximum risk of $\widehat{\varepsilon}^*_{a_n}/\varepsilon_n$ is slightly worse and for the "optimal" choice of $a_n = 4\sqrt{2\pi}(\ln n)^{3/2}$ is as follows (see Theorem 4.1 in [3]): for $1/2 < \beta < 1$ and $\beta < r < 1$

$$\sup_{(\varepsilon_n,\mu_n)\in\mathcal{P}_{n,\beta,r}} \mathbf{E}\left(\frac{\widehat{\varepsilon}^*_{a_n}}{\varepsilon_n} - 1\right)^2 \leqslant c\,(\beta, r)\,(\ln n)^4 n^{-1+\beta} \tag{25}$$

where the set $\mathcal{P}_{n,\beta,r}$ is as in (24) and $c(\beta, r)$ is a generic constant depending on β and r. Comparing the upper bound (23) to the one given in (25), we conclude that, in the intersection $\mathcal{D} \cap \mathcal{S} = \{(\beta, r) \in \mathbb{R}^2 : 1/2 < \beta < r < 1\}$ of the detection and selection regions, the new estimator $\widehat{\varepsilon}_n$ dominates the CJL estimator $\widehat{\varepsilon}^*_{a_n}$; the estimator $\widehat{\varepsilon}_n$ also works well in the region $\mathcal{S} \cap \mathcal{D}^c$ (with $\mathcal{D}^c$ denoting the complement of $\mathcal{D}$ in $(0,1)\times(0,4) \subset \mathbb{R}^2$), where $\widehat{\varepsilon}^*_{a_n}$ has not been originally defined.

Also, in view of Theorem 4.2 in [3], which is initially stated and proved for $1/2 < \beta < 1$ and $0 < r < 1$ but continues to be true for $0 < \beta < 1$ and $0 < r < 4$, the modified CJL estimator $\widehat{\varepsilon}_n$ is nearly *rate optimal* (in the asymptotically minimax sense) over the region Ω_n defined by

$$\begin{aligned}\Omega_n = \{(\varepsilon_n, \mu_n) :\ & B_1 n^{-\beta} \leqslant \varepsilon_n \leqslant B_2 n^{-\beta}, \\ & \sqrt{2r\ln n} - A_1/\ln n \leqslant \mu_n \leqslant \sqrt{2r\ln n} + A_2/\ln n\},\end{aligned}$$

for $0 < r < 4$, $0 < \beta < 1$, $A_1, A_2 > 0$, and $B_2 > B_1 > 0$. More precisely, for all sufficiently large n and some constant $c > 0$

$$\inf_{\widetilde{\varepsilon}_n} \sup_{(\varepsilon_n,\mu_n)\in\Omega_n} \mathbf{E}\left(\frac{\widetilde{\varepsilon}_n}{\varepsilon_n} - 1\right)^2 \geqslant c n^{-1+\beta},$$

where the infimum is over all possible estimators $\widetilde{\varepsilon}_n$ of ε_n based on $X_1, \ldots, X_n$.

Now, the estimator $\widehat{\beta}_n$ of the sparsity parameter β defined by

$$\widehat{\beta}_n = \frac{\ln(1/\widehat{\varepsilon}_n)}{\ln n},$$

which is constructed by taking α in step 2 of the above algorithm depending on n and equal to $\alpha = 1 - H\left(\left(\frac{4\ln n}{\ln\ln 4}\right)^{1/2}\right)$, is a consistent estimator of β, as follows from Theorem 1 and Markov's inequality.

2.3. Discussion. The advantage of the proposed estimator $\widehat{\varepsilon}_n$ over the CJL estimator $\widehat{\varepsilon}^*_{a_n}$ defined by (2.8) in [3], in addition to its higher efficiency, is that $\widehat{\varepsilon}_n$ is easier to implement. Indeed, the choice of the quantity a_n in the definition of $\widehat{\varepsilon}^*_{a_n}$ may vary depending on a purpose (see Section 3.1 of [3] for the discussion on the choice of a_n). For instance, in [3], for the purpose of proving the upper bound (25), a_n is first set to be $a_n = 4\sqrt{2\pi}(\ln n)^{3/2}$ and the value α_n is then chosen to have

$$\mathbf{P}\left(\sup_{t\in[0,\sqrt{2\ln n}]} \frac{\sqrt{n}|F_n(t) - F(t)|}{\sqrt{F(t)(1-F(t))}} \leqslant a_n\right) = 1 - \alpha_n.$$

At the same time, as suggested in [3], in order to compute $\widehat{\varepsilon}^*_{a_n}$ in possible applications, the value of a_n could be found as follows. For a given $\alpha \in (0,1)$, let a_n be such that $\mathbf{P}(Y_n \leqslant a_n) = 1 - \alpha$, where

$$Y_n \overset{d}{=} \max_{F(0)\leqslant t\leqslant F(\sqrt{2\ln n})} \frac{\sqrt{n}|\mathbb{U}_n(t) - t|}{\sqrt{t\,(1-t)}}$$

with $\mathbb{U}_n(t)$ being the edf based on n independent uniform $U(0,1)$ observations $U_1, \ldots, U_n$. Since the underlying cdf $F(t)$ as defined in (8) depends on the parameters β and r that are generally unknown, finding a_n is not straightforward. To find an approximate value of a_n for a given value of α, it was suggested in [3] to use the statistic

$$W_n \overset{d}{=} \max_{1/2\leqslant t\leqslant \Phi(\sqrt{2\ln n})} \frac{\sqrt{n}|\mathbb{U}_n(t) - t|}{\sqrt{t\,(1-t)}},$$

for which, as claimed on p. 2439 of [3],

$$\mathbf{P}(W_n \leqslant a_n) \approx 1 - \alpha.$$

For several choices of α, the (approximate) values of a_n, obtained as the $(1-\alpha)$th quantile of W_n, are given in Table 2 of [3]. The use of $\sup\limits_{0<t<1} \frac{\sqrt{n}|\mathbb{U}_n(t)-t|}{q(t)}$, with q as in (10), in place of

$$\max_{F(0)\leqslant t\leqslant F(\sqrt{2\ln n})} \frac{\sqrt{n}|\mathbb{U}_n(t) - t|}{\sqrt{t\,(1-t)}}$$

resolves the issue with constructing a confidence band for $F(t)$ of a given level, an important ingredient in estimation of the fraction of nonzero means ε_n. This is so because, unlike statistics Y_n and W_n that blow up with probability one as n tends to infinity (see, for example, Chapter 16 in [19]), the statistic $\sup\limits_{0<t<1} \frac{\sqrt{n}|\mathbb{U}_n(t)-t|}{q(t)}$ stays finite with probability one and, in view of (11), it converges in distribution to a non-degenerate random variable whose distribution is tabulated.

At the outset of the study, we anticipated to obtain a new estimator that would be more efficient and also could overcome the property of the CJL estimator $\widehat{\varepsilon}^*_{a_n}$ to be "strongly asymptotic". Unfortunately, in order to work as designed, the new estimator $\widehat{\varepsilon}_n$ continues to require a very large sample size. This is caused by the fact that it is the left requirement in (20) that makes the CJL-type estimators strongly asymptotic, not the choice of a confidence band for $F(t)$.

§3. Proof of Theorem 1

Let $F(t) = F_{n,\beta,r}(t)$ be a common cdf of $X_1, \dots, X_n$ as given by (8), where $\varepsilon_n = n^{-\beta}$ for $\beta \in (0,1)$ and $\mu_n = \sqrt{2r \log n}$ for $r \in (\beta, 4)$. For a given $\alpha \in (0,1)$, consider the event

$$\mathbb{A}_{n,\alpha} = \left\{ \mathbb{F}^-_{n,\alpha}(t) \leqslant F(t) \leqslant \mathbb{F}^+_{n,\alpha}(t),\ \forall\, t \in [0, \mu_n + (2\ln n)^{-1/2}] \right\},$$

and observe that for any $\alpha \in (0,1)$

$$\begin{gathered}\left\{\mathbb{F}^-_{n,\alpha}(t) \leqslant F(t) \leqslant \mathbb{F}^+_{n,\alpha}(t), \forall t \in [X_{(1)}, X_{(n)})\right\} \\ \cap \left\{ X_{(1)} \leqslant 0, X_{(n)} \geqslant \mu_n + (2\ln n)^{-1/2} \right\} \subset \mathbb{A}_{n,\alpha},\end{gathered}$$

where for all large enough n the probability

$$\mathbf{P}\left(X_{(1)} \leqslant 0, X_{(n)} \geqslant \mu_n + (2\ln n)^{-1/2} \right)$$

can be made at least as large as $1-\alpha$. Then, using Bonferroni's inequality (see, for example, Appendix A.2 in [1]), for all large enough n, $\mathbf{P}(\mathbb{A}_{n,\alpha}) \geqslant 1-2\alpha$, and also, for a positive sequence α_n decaying to zero at a polynomial rate and all large enough n, we have

$$\mathbf{P}\left(\mathbb{A}_{n,\alpha_n}\right) \geqslant 1 - 2\alpha_n. \tag{26}$$

By the construction of $\widehat{\varepsilon}_n$, if index $j^* \in \{1, \dots, J_0\}$ is chosen to satisfy

$$t_{j^*} \leqslant \mu_n < t_{j^*+1},$$

then, over the event $\mathbb{A}_{n,\alpha_n}$, we have $\widehat{\varepsilon}_n^{j^*} \leqslant \widehat{\varepsilon}_n \leqslant \varepsilon_n$ and hence

$$(1 - \widehat{\varepsilon}_n/\varepsilon_n)^2 \leqslant \left(1 - \widehat{\varepsilon}_n^{j^*}/\varepsilon_n\right)^2.$$

Therefore, for all large enough n

$$\mathbf{E}\left(\widehat{\varepsilon}_n/\varepsilon_n - 1\right)^2 \leqslant \mathbf{E}\left[\left(\widehat{\varepsilon}_n^{j^*}/\varepsilon_n - 1\right)^2 \mathbb{I}(\mathbb{A}_{n,\alpha_n})\right]$$
$$+\mathbf{E}\left[(\widehat{\varepsilon}_n/\varepsilon_n - 1)^2\, \mathbb{I}(\mathbb{A}^c_{n,\alpha_n})\right] \leqslant \mathbf{E}\left[\left(\widehat{\varepsilon}_n^{j^*}/\varepsilon_n - 1\right)^2 \mathbb{I}(\mathbb{A}_{n,\alpha_n})\right] + \mathbf{P}\left(\mathbb{A}^c_{n,\alpha_n}\right)\varepsilon_n^{-2}, \tag{27}$$

where, in view of (26), $\mathbf{P}(\mathbb{A}^c_{n,\alpha_n}) \leqslant 2\alpha_n$.

Next, since $B(t)/q(t)$ with $q(t)$ as in (10) is a centered Gaussian processes whose trajectories are bounded with probability one, it follows from the Concentration Principe (see, for example, Theorem 6.2 in [16]) that for any $x > 0$

$$\mathbf{P}\left(\sup_{0<t<1} \frac{B(t)}{q(t)} \geqslant x\right) \leqslant 1 - \Phi\left(\frac{x - m}{\sigma}\right),$$

where m is the median of $\sup\limits_{0<1<t} B(t)/q(t)$ and

$$\sigma^2 = \sup_{0<t<1} \frac{\mathbf{E}(B^2(t))}{q^2(t)} = \frac{1}{\ln\ln 4}.$$

Therefore, using

$$1 - \Phi(x) \sim \varphi(x)/x, \quad x \to \infty, \tag{28}$$

we have as $x \to \infty$

$$1 - H(x) \leqslant 2\mathbf{P}\left(\sup_{0<t<1} \frac{B(t)}{q(t)} \geqslant x\right) \leqslant \frac{2\exp\left(-(1/2)(\ln\ln 4)\, x^2(1 + o(1))\right)}{x(1 + o(1))\sqrt{2\pi \ln\ln 4}}.$$

Hence, the choice of the $(1 - \alpha_n)$th quantile c_{α_n} of $H(t)$ in the form $c_{\alpha_n} = \left(\frac{4\ln n}{\ln\ln 4}\right)^{1/2}$ ensures that for some $c > 0$ and all large enough n

$$\alpha_n = 1 - H(c_{\alpha_n}) \leqslant \frac{c}{n^2(\ln n)^{1/2}} = o(n^{-2}). \tag{29}$$

Now, recalling that $\varepsilon_n = n^{-\beta}$, we obtain from (27) and (29) that for all large enough n

$$\mathbf{E}\left(\widehat{\varepsilon}_n/\varepsilon_n - 1\right)^2 \leqslant \mathbf{E}\left[\left(\widehat{\varepsilon}_n^{j^*}/\varepsilon_n - 1\right)^2 \mathbb{I}(\mathbb{A}_{n,\alpha_n})\right] + o(n^{-2+2\beta}). \tag{30}$$

Next, in view of the identity (see (15) and (21))

$$\begin{aligned}\widehat{\varepsilon}_n^{j^*}/\varepsilon_n - 1 &= \frac{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)}{\Phi(t_{j^*}) - \Phi(t_{j^*} - \widehat{\mu}_n^{j^*})} \\ &\quad \times \left(\frac{F(t_{j^*}) - \mathbb{F}^+_{n,\alpha_n}(t_{j^*})}{\Phi(t_{j^*}) - F(t_{j^*})} - \frac{\Phi(t_{j^*} - \mu_n) - \Phi(t_{j^*} - \widehat{\mu}_n^{j^*})}{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)} \right),\end{aligned}$$

the fact that, over $\mathbb{A}_{n,\alpha_n}$,

$$\frac{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)}{\Phi(t_{j^*}) - \Phi(t_{j^*} - \widehat{\mu}_n^{j^*})} \leqslant 1,$$

and the inequality $|a \pm b|^2 \leqslant 2\left(|a|^2 + |b|^2\right)$, we obtain

$$\begin{aligned}\left(\widehat{\varepsilon}_n^{j^*}/\varepsilon_n - 1\right)^2 &\leqslant 2\left(\frac{F(t_{j^*}) - \mathbb{F}^+_{n,\alpha_n}(t_{j^*})}{\Phi(t_{j^*}) - F(t_{j^*})} \right)^2 \\ &\quad + 2\left(\frac{\Phi(t_{j^*} - \mu_n) - \Phi(t_{j^*} - \widehat{\mu}_n^{j^*})}{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)} \right)^2. \qquad (31)\end{aligned}$$

By the definition of $\mathbb{F}^{\pm}_{n,\alpha_n}(t)$ as in (13), we have that over $\mathbb{A}_{n,\alpha_n}$

$$|\mathbb{F}^{\pm}_{n,\alpha_n}(t) - F(t)| \leqslant \frac{2c_{\alpha_n} q(F(t))}{\sqrt{n}}. \qquad (32)$$

Therefore, due to (28), relations (8) and (10), the assumptions that $r > \beta$ and $c_{\alpha_n} = \left(\frac{2p \ln n}{\ln\ln 4}\right)^{1/2}$, the first term on the right side of (31) satisfies (over $\mathbb{A}_{n,\alpha_n}$) as $n \to \infty$

$$\begin{aligned}2\left(\frac{F(t_{j^*}) - \mathbb{F}^+_{n,\alpha}(t_{j^*})}{\Phi(t_{j^*}) - F(t_{j^*})} \right)^2 &\leqslant \frac{8c^2_{\alpha_n} q^2(F(t_{j^*}))}{n\left(\Phi(t_{j^*}) - F(t_{j^*})\right)^2} \\ &= O\left(\frac{c^2_{\alpha_n} q^2(F(\mu_n))}{n\left(\Phi(\mu_n) - F(\mu_n)\right)^2} \right) = O\left(\frac{(\ln\ln n)\ln n}{n^{1-\beta}} \right). \qquad (33)\end{aligned}$$

For the second term on the right side of (31), using the fact that as $n \to \infty$

$$t_{j^*} \to \infty \quad \text{and} \quad t_{j^*} - \mu_n \to \begin{cases} -\infty, & \text{if } t_{j^*} < \mu_n, \\ 0, & \text{if } t_{j^*} = \mu_n, \end{cases}$$

we have, over $\mathbb{A}_{n,\alpha}$, as $n \to \infty$

$$2\left(\frac{\Phi(t_{j^*}-\mu_n)-\Phi(t_{j^*}-\widehat{\mu}_n^{j^*})}{\Phi(t_{j^*})-\Phi(t_{j^*}-\mu_n)}\right)^2 \sim 2\left(\frac{\varphi(t_{j^*}-\mu_n)}{\Phi(t_{j^*})-\Phi(t_{j^*}-\mu_n)}\right)^2 (\mu_n-\widehat{\mu}_n^{j^*})^2$$
$$= \begin{cases} o\left((\mu_n - \widehat{\mu}_n^{j^*})^2\right), & \text{if } t_{j^*} < \mu_n, \\ O\left((\mu_n - \widehat{\mu}_n^{j^*})^2\right), & \text{if } t_{j^*} = \mu_n. \end{cases} \tag{34}$$

Next, it can be seen that for all large enough n

$$\mathbf{E}\left((\mu_n - \widehat{\mu}_n^{j^*})^2 \mathbb{I}(\mathbb{A}_{n,\alpha})\right) \leqslant \frac{C(\ln n)^2 \ln\ln n}{n^{1-\beta}} \tag{35}$$

with some constant $C = C(\beta, r) > 0$. Indeed, by definition, μ_n and $\widehat{\mu}_n^{j^*}$ are solutions of the respective equations

$$D(\mu_n) = \frac{\Phi(t_{j^*}) - F(t_{j^*})}{\Phi(t_{j^*+1}) - F(t_{j^*+1})} \quad \text{and} \quad D(\widehat{\mu}_n^{j^*}) = \frac{\Phi(t_{j^*}) - \mathbb{F}^+_{n,\alpha}(t_{j^*})}{\Phi(t_{j^*+1}) - \mathbb{F}^-_{n,\alpha}(t_{j^*+1})},$$

where the function $D(\mu) = D(\mu; t_{j^*}, t_{j^*+1})$ is as defined in (14). By simple algebra,

$$\frac{D'(\mu_n)}{D(\mu_n)} = \frac{\varphi(t_{j^*} - \mu_n)}{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)} - \frac{\varphi(t_{j^*+1} - \mu_n)}{\Phi(t_{j^*+1}) - \Phi(t_{j^*+1} - \mu_n)}.$$

This can be reduced to the following with more analysis:

$$0 > \frac{D'(\mu_n)}{D(\mu_n)} = O(t_{j^*+1} - t_{j^*}) = O((\ln n)^{-1/2}), \quad n \to \infty,$$

and hence, on the event $\mathbb{A}_{n,\alpha_n}$,

$$\frac{D(\widehat{\mu}_n^{j^*}) - D(\mu_n)}{\widehat{\mu}_n^{j^*} - \mu_n} \frac{1}{D(\mu_n)} = O((\ln n)^{-1/2}), \quad n \to \infty. \tag{36}$$

We now verify that, on the event $\mathbb{A}_{n,\alpha_n}$,

$$\frac{D(\widehat{\mu}_n^{j^*}) - D(\mu_n)}{D(\mu_n)} = O\left(\frac{(\ln\ln n)^{1/2}(\ln n)^{1/2}}{n^{(1-\beta)/2}}\right), \quad n \to \infty, \tag{37}$$

To this end, we set

$$\delta_n = 1 - \frac{D(\mu_n)}{D(\widehat{\mu}_n^{j^*})} = 1 - \frac{(\Phi(t_{j^*}) - F(t_{j^*}))/(\Phi(t_{j^*+1}) - F(t_{j^*+1}))}{(\Phi(t_{j^*}) - \mathbb{F}^+_{n,\alpha}(t_{j^*}))/(\Phi(t_{j^*+1}) - \mathbb{F}^-_{n,\alpha}(t_{j^*+1}))}.$$

Then, using the inequality $\left|\frac{1-a}{1+b}-1\right| \leqslant a+b$ for any $a, b > 0$ and (32), we obtain over $\mathbb{A}_{n,\alpha_n}$

$$\begin{aligned} |\delta_n| &\leqslant \left|\frac{\Phi(t_{j^*}) - \mathbb{F}^+_{n,\alpha_n}(t_{j^*})}{\Phi(t_{j^*}) - F(t_{j^*})} - 1\right| + \left|\frac{\Phi(t_{j^*+1}) - \mathbb{F}^-_{n,\alpha_n}(t_{j^*+1})}{\Phi(t_{j^*+1}) - F(t_{j^*+1})} - 1\right| \\ &= \frac{\left|F(t_{j^*}) - \mathbb{F}^+_{n,\alpha_n}(t_{j^*})\right|}{\Phi(t_{j^*}) - F(t_{j^*})} + \frac{\left|F(t_{j^*+1}) - \mathbb{F}^-_{n,\alpha_n}(t_{j^*+1})\right|}{\Phi(t_{j^*+1}) - F(t_{j^*+1})} \\ &\leqslant \frac{2c_{\alpha_n}}{\sqrt{n}} \left(\frac{q(F(t_{j^*}))}{\Phi(t_{j^*}) - F(t_{j^*})} + \frac{q(F(t_{j^*+1}))}{\Phi(t_{j^*+1}) - F(t_{j^*+1})} \right). \end{aligned}$$

From this, by the choice of c_{α_n} and the points t_{j^*} and t_{j^*+1}, on the event $\mathbb{A}_{n,\alpha_n}$, as $n \to \infty$

$$\delta_n = O\left(\frac{c_{\alpha_n} q(F(\mu_n))}{\sqrt{n}(\Phi(\mu_n) - F(\mu_n))} \right) = O\left(\frac{(\ln\ln n)^{1/2} (\ln n)^{1/2}}{n^{(1-\beta)/2}} \right),$$

and hence, noting that $\delta_n < 0$,

$$\begin{aligned} \left| \frac{D(\widehat{\mu}_n^{j^*}) - D(\mu_n)}{D(\mu_n)} \right| = \left| \frac{D(\widehat{\mu}_n^{j^*})}{D(\mu_n)} - 1 \right| = \left| \frac{\delta_n}{1 - \delta_n} \right| \\ \leqslant |\delta_n| = O\left(\frac{(\ln\ln n)^{1/2} (\ln n)^{1/2}}{n^{(1-\beta)/2}} \right). \end{aligned}$$

Thus, relation (37) is verified, and the combination of (36) and (37) gives, over $\mathbb{A}_{n,\alpha_n}$,

$$\widehat{\mu}_n^{j^*} - \mu_n = O\left(\frac{(\ln\ln n)^{1/2} \ln n}{n^{(1-\beta)/2}} \right), \quad n \to \infty.$$

This leads to (35).

Finally, putting together relations (30), (31), (33), (34), and (35), we obtain for all $(\beta, r) \in \mathcal{S}$ that when $\alpha_n = 1 - H\left(\left(\frac{4 \ln n}{\ln\ln 4}\right)^{1/2} \right)$ and n is sufficiently large

$$\mathbf{E}\left(\widehat{\varepsilon}_n / \varepsilon_n - 1\right)^2 \leqslant \frac{C(\ln n)^2 \ln\ln n}{n^{1-\beta}},$$

where the constant $C = C(\beta, r)$ is as above. The proof is complete. □

Remark 1. The use of the CsCsHM confidence band $[\mathbb{F}^-_{n,\alpha}(t), \mathbb{F}^+_{n,\alpha}(t)]$ in the construction of $\widehat{\varepsilon}_n$ gives us relation (32) with $c_{\alpha_n} = O\left((\ln n)^{1/2}\right)$, whose right side is $\frac{\ln n}{(\ln\ln n)^{1/2}}$ times better than the corresponding bound on

$|\mathbb{F}^{\pm}_{a_n}(t) - F(t)|$ in Lemma 8.2 of Cai et al [3]. As a result, the convergence rate $r_n = \left(n^{-(1-\beta)}(\ln\ln n)\ln^2 n\right)^{1/2}$ of $\widehat{\varepsilon}_n/\varepsilon_n$ to one is $\frac{\ln n}{(\ln\ln n)^{1/2}}$ times faster than that of $\widehat{\varepsilon}^*_{a_n}/\varepsilon_n$, where $\widehat{\varepsilon}^*_{a_n}$ is the CJL estimator of ε_n defined by (2.8) in [3].

Remark 2. The conclusion of Theorem 1 is also true for those non-normal mixtures whose tail behavior is similar to that of the normal mixture (3). For instance, with the properly adjusted algorithm of Section 2.2, the upper bound (23) is valid for the chi-square mixture with the underlying cdf $F(t) = F_{n,\beta,r,\nu}(t)$ of the form

$$F(t) = (1-\varepsilon_n)G_\nu(t;0) + \varepsilon_n G_\nu(t;\mu_n^2), \quad t \in \mathbb{R},$$

where $G_\nu(t;\lambda) = \mathbf{P}(\chi^2_\nu(\lambda) \leqslant t)$ is the cdf of a chi-square random variable $\chi^2_\nu(\lambda)$ with $\nu \geqslant 1$ degrees of freedom and noncentrality parameter $\lambda \geqslant 0$, $\varepsilon_n = n^{-\beta}$ for $\beta \in (0,1)$, and $\mu_n = \sqrt{2r\log n}$ for $r \in (0,4)$. This is so because for any fixed $\nu \geqslant 1$ as $n \to \infty$

$$\begin{aligned}
\mathbf{P}\left(\chi^2_\nu(0) > 2s\ln n\right) &= O\left(n^{-s}\ln^{\nu/2-1} n\right),\ 0 < s < \infty,\\
\mathbf{P}\left(\chi^2_\nu(\mu_n^2) > 2s\ln n\right) &\asymp \mathbf{P}\left(N(\mu_n,1) > \sqrt{2s\ln n}\right)\\
&= O\left(n^{-(\sqrt{s}-\sqrt{r})^2}\ln^{-\frac{1}{2}} n\right),\ 0 < r < s < \infty,\\
\mathbf{P}\left(\chi^2_\nu(\mu_n^2) \leqslant 2s\ln n\right) &\asymp \mathbf{P}\left(N(\mu_n,1) \leqslant \sqrt{2s\ln n}\right)\\
&= O\left(n^{-(\sqrt{s}-\sqrt{r})^2}\ln^{-\frac{1}{2}} n\right),\ 0 < s < r < \infty,
\end{aligned}$$

where the first relation is obvious, the second relation is obtained in [7], an the third relation is a consequence of (1) and (2) in [13].

§4. Numerical study

To compare numerically the quality of the CJL estimator $\widehat{\varepsilon}^*_{a_n}$ introduced by formula (2.8) in [3] and its modification $\widehat{\varepsilon}_n$ as defined in (22), we carry out a small-scale simulation study and compute the estimated mean squared errors of $\widehat{\varepsilon}_n/\varepsilon_n$ and $\widehat{\varepsilon}^*_{a_n}/\varepsilon_n$. In our experiment, we simulate n random samples with a cdf given by formula (8). We pick the same sample size $n = 10^7$ as in [3] so that the number of nonzero means in model (3) is small. To get the observed values of $\widehat{\varepsilon}^*_{a_n}$ and $\widehat{\varepsilon}_n$ and their estimated quadratic risks, we run $M = 100$ independent cycles of simulations and use $\widehat{\mathbf{E}}\left(\widehat{\varepsilon}_n/\varepsilon_n - 1\right)^2 = M^{-1}\sum_{m=1}^{M}\left(\widehat{\varepsilon}_n^{(m)}/\varepsilon_n - 1\right)^2$, where $\widehat{\varepsilon}_n^{(m)}$ is the value of $\widehat{\varepsilon}_n$ obtained in the mth repetition of the experiment, to estimate $\mathbf{E}\left(\widehat{\varepsilon}_n/\varepsilon_n - 1\right)^2$; the estimated risk $\widehat{\mathbf{E}}\left(\widehat{\varepsilon}^*_{a_n}/\varepsilon_n - 1\right)^2$ is defined similarly. Two choices of a point (β, r), namely $(\beta, r) = (1/3, 3/4) \in \mathcal{S}$ and $(\beta, r) = (4/7, 1/2) \in \mathcal{D} \cap \mathcal{S}^c$, and different values of $\alpha \in (0,1)$ are explored in our experiment. The symbol $\mathcal{S}^c$ is used to denote the complement of $\mathcal{S}$ in $(0,1) \times (0,4) \subset \mathbb{R}^2$. The simulation study is run using R language. The numerical results of the study are presented in Tables 1–4 and also displayed in Figure 3 below. Numerically, our estimator $\widehat{\varepsilon}_n$ dominates the CJL estimator $\widehat{\varepsilon}^*_{a_n}$ not only in the selection region but also in the detection region, as seen from the results obtained for $(\beta, r) = (4/7, 1/2) \in \mathcal{D} \cap \mathcal{S}^c$.

The values of $a_n/\sqrt{2\ln\ln n}$ in Tables 2 and 4 were taken from Table 2 on page 2440 of [3]. The values of c_α, the $(1-\alpha)$th quantile of $H(t) = \mathbf{P}\left(\sup_{0<u<1}|B(u)|/q(u) \leqslant t\right)$, appearing in Tables 1 and 3 were found from Table III in [18] (see also Table 2 in [9]). In general, for any $\alpha \in (0,1)$, one can get c_α by using the algorithm in Section 3 of [18] (see also Section 3 of [9]). The estimators $\widehat{\beta}^*_{a_n}$ and $\widehat{\beta}_n$, whose observed values are also given in Tables 1–4, were obtained from $\widehat{\varepsilon}^*_{a_n}$ and $\widehat{\varepsilon}_n$ through the relations (recall that $\varepsilon_n = n^{-\beta}$) $\widehat{\beta}^*_{a_n} = \frac{\ln(1/\widehat{\varepsilon}^*_{a_n})}{\ln n}$ and $\widehat{\beta}_n = \frac{\ln(1/\widehat{\varepsilon}_n)}{\ln n}$. With high probability both estimators $\widehat{\beta}^*_{a_n}$ and $\widehat{\beta}_n$ overestimate the true β.

Table 1. Numerical summary for the new estimator $\widehat{\varepsilon}_n$ in the selection region $\mathcal{S}$

α	c_α	ε_n	$\widehat{\varepsilon}_n$	$\widehat{\mathbf{E}}\left(\widehat{\varepsilon}_n/\varepsilon_n - 1\right)^2$	$\widehat{\beta}_n$
0.01	5.53	0.00464	0.00397	0.02082	0.34300
0.02	5.16	0.00464	0.00400	0.01888	0.34250
0.05	4.57	0.00464	0.00412	0.01254	0.34070
0.1	4.12	0.00464	0.00417	0.01024	0.33995
0.2	3.60	0.00464	0.00422	0.00814	0.33920
0.5	2.80	0.00464	0.00433	0.00436	0.33759

$n = 10^7$, $M = 100$, $\beta = 1/3$, and $r = 3/4$.

As seen from the Tables 1–4, the risk of the estimation procedure based on $\widehat{\varepsilon}_n$ is smaller than that of the CJL procedure. As a result, the observed

Table 2. Numerical summary for the CJL estimator $\widehat{\varepsilon}^*_{a_n}$ in the selection region $\mathcal{S}$

α	$a_n/\sqrt{2\ln\ln n}$	ε_n	$\widehat{\varepsilon}^*_{a_n}$	$\widehat{\mathbf{E}}\left(\widehat{\varepsilon}^*_{a_n}/\varepsilon_n - 1\right)^2$	$\widehat{\beta}^*_{a_n}$
0.005	2.126	0.00464	0.00367	0.04390	0.34792
0.01	1.956	0.00464	0.00386	0.02869	0.34485
0.025	1.699	0.00464	0.00388	0.02689	0.34445
0.05	1.545	0.00464	0.00403	0.01753	0.34215
0.1	1.370	0.00464	0.00408	0.01468	0.34135
0.5	0.940	0.00464	0.00431	0.00507	0.33792

$n = 10^7$, $M = 100$, $\beta = 1/3$, and $r = 3/4$.

values of $\widehat{\varepsilon}_n$ and $\widehat{\beta}_n$ are closer to the true values of ε_n and β than those of $\widehat{\varepsilon}^*_{\alpha_n}$ and $\widehat{\beta}^*_{a_n}$, under different selections of the parameters α, β, and r. These numerical findings are in agreement with the result of Theorem 1 (as compared to the upper bound (25)). In the course of numerical studies, we also noted that the accuracy of estimation does not depend that much on the parameter r alone. As seen from Tables 1 and 3, the quality of estimation is crucially affected by whether the point (β, r) is located inside or outside of the selection region $\mathcal{S}$; in the former case, the performance of $\widehat{\varepsilon}_n$ is much better, especially for very large n.

Figure 3 gives histograms of $\widehat{\varepsilon}_n/\varepsilon_n$ and $\widehat{\varepsilon}^*_{a_n}/\varepsilon_n$ for two choices of (β, r) and four different levels of α. This figure supports the numerical results in Tables 1–4 and nicely illustrates the "thin tail property" of the ratios implying that, with both estimators, when $\alpha \in (0, 1/2]$ the chance of overestimating ε_n is small.

It is also seen from Tables 1–4 that as α gets larger the risk of $\widehat{\varepsilon}_n/\varepsilon_n$ and of $\widehat{\varepsilon}^*_{a_n}/\varepsilon_n$ gets smaller, signifying better performance of both estimators for larger values of α. For the estimator $\widehat{\varepsilon}_n$, this is explained by the fact that a larger α gives a smaller $c_\alpha = H^{-1}(1-\alpha)$ and thus makes the confidence band $[\mathbb{F}^-_{n,\alpha}(t), \mathbb{F}^+_{n,\alpha}(t)]$ used in the construction of $\widehat{\varepsilon}_n$ shorter (see relation (32)). Similar behaviour of the CJL estimator $\widehat{\varepsilon}^*_{a_n}$ has been noticed and commented in Section 7 of [2] as follows: "a larger α_n will not increase much of the chance of overestimation, but it will certainly boost the underestimation and in effect make the whole estimator more accurate". However, as the problem of our interest is to estimate ε_n from below with high probability, both $\widehat{\varepsilon}_n$ and $\widehat{\varepsilon}^*_{a_n}$ will do their job properly only when α is a small number near zero. As α gets larger, the probability of overestimating ε_n by using $\widehat{\varepsilon}_n$ and $\widehat{\varepsilon}^*_{a_n}$ increases and for $\alpha \in (1/2, 1)$, in a series of repeated simulations, one gets estimates that are greater than ε_n.

Table 3. Numerical summary for the new estimator $\widehat{\varepsilon}_n$ in the region $\mathcal{D} \cap \mathcal{S}^c$

α	c_α	ε_n	$\widehat{\varepsilon}_n$	$\widehat{\mathbf{E}}\left(\widehat{\varepsilon}_n/\varepsilon_n - 1\right)^2$	$\widehat{\beta}_n$
0.01	5.53	0.00010	5.62e-05	0.19158	0.60715
0.02	5.16	0.00010	5.87e-05	0.17068	0.60449
0.05	4.57	0.00010	6.06e-05	0.15518	0.60249
0.1	4.17	0.00010	6.57e-05	0.11751	0.59747
0.2	3.60	0.00010	6.63e-05	0.11374	0.59695
0.5	2.80	0.00010	7.27e-05	0.07437	0.59118

$n = 10^7$, $M = 100$, $\beta = 4/7$, and $r = 1/2$.

Table 4. Numerical summary for the CJL estimator $\widehat{\varepsilon}^*_{a_n}$ in the region $\mathcal{D} \cap \mathcal{S}^c$

α	$a_n/\sqrt{2\ln\ln n}$	ε_n	$\widehat{\varepsilon}^*_{a_n}$	$\widehat{\mathbf{E}}\left(\widehat{\varepsilon}^*_{a_n}/\varepsilon_n - 1\right)^2$	$\widehat{\beta}^*_{a_n}$
0.005	2.126	0.00010	4.82e-05	0.26832	0.61671
0.01	1.956	0.00010	4.93e-05	0.25705	0.61531
0.025	1.699	0.00010	5.26e-05	0.22506	0.61133
0.05	1.545	0.00010	5.37e-05	0.21437	0.61000
0.1	1.370	0.00010	5.62e-05	0.19202	0.60720
0.5	0.940	0.00010	5.75e-05	0.18054	0.60575

$n = 10^7$, $M = 100$, $\beta = 4/7$, and $r = 1/2$.

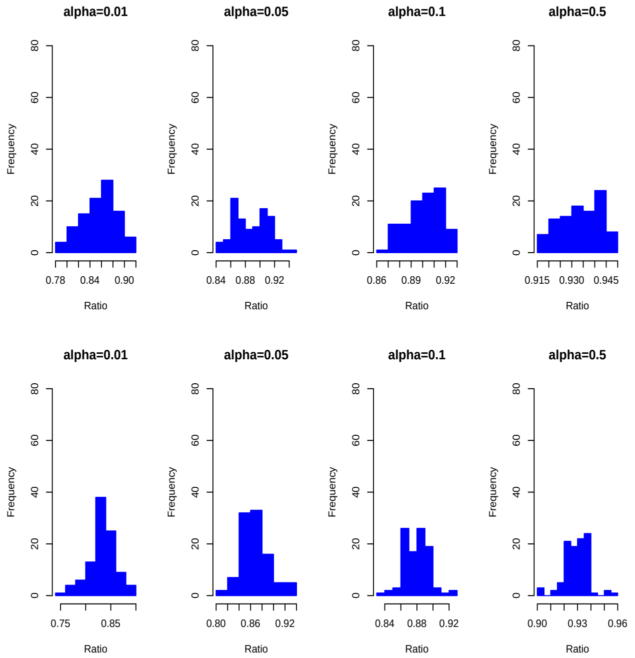

Figure 3. Histograms for $M = 100$ simulated ratios $\widehat{\varepsilon}_n/\varepsilon_n$ (top row) and $\widehat{\varepsilon}^*_{a_n}/\varepsilon_n$ (bottom row) from the normal mixture model (3) with $n = 10^7$ and $(\beta, r) = (1/3, 3/4) \in \mathcal{S}$ for different levels of α: 0.01, 0.05, 0.1 and 0.5.

Acknowledgments. The research of N. A. Stepanova was supported by an NSERC grant. The research of Y. Wang was partially supported by an NSERC grant during the author's graduate studies at Carleton University.

References

1. P. J. Bickel, K. A. Doksum, *Mathematical Statistics. Basic Ideas and Selected Topics. Vol. I*, 2nd ed., Prentice-Hall, New Jersey, 2001.
2. T. T. Cai, J. Jin, M. Low, *Estimation and confidence sets for sparse normal models*, 2006. `http://arxiv.org/abs/math/0612623v1`.
3. T. T. Cai, J. Jin, M. Low, *Estimation and confidence sets for sparse normal models.* — Ann. Statist. **35** (2007), 2421–2449.

4. T. T. Cai, X. J. Jeng, J. Jin, *Optimal detection of heterogeneous and heteroscedastic mixtures.* — J. R. Statist. Soc. B, **73** (2011), 629–662.
5. M. Csörgő, S. Csörgő, L. Horváth, D. Mason, *Weighted empirical and quantile processes.* — Ann. Probab., **14** (1986), 31–85.
6. X. Cui, *Optimal component selection in high dimensions*, MSc Thesis. Carleton University, 2014.
7. D. Donoho, J. Jin, *Higher criticism for detecting sparse heterogeneous mixtures.* — Ann. Statist., **32** (2004), 962–994.
8. D. Donoho, J. Jin, *Feature selection by higher criticism thresholding achieves the optimal phase diagram.* — Phil. Trans. R. Soc. A, **367**, (2009), 4449–4470.
9. B. J. Eastwood, V. R. Eastwood, *Tabulating weighted functionals of Brownian bridges via Monte Carlo simulation*, in: Asymptotic methods in probability and statistics. A volume in honour of Miklós Csörgő, (ed. B. Szyszkowicz), pp. 707–719. Elsivier Science B.V., Amsterdam.
10. F. Eicker, *The asymptotic distribution of the suprema of the standardized empirical processes.* — Ann. Statist., **7** (1979), 116–138.
11. Y. Fan, J. Jin, Z. Yao, *Optimal classification in sparse Gaussian graphic models.* — Ann. Statist., **41**, No. 5 (2013), 2537–2571.
12. C. R. Genovese, J. Jin, L. Wasserman, Z. Yao, *A comparison of the lasso and marginal regression.* — J. Mach. Learn. Res., **13**, (2012), 2107–2143.
13. C.-P. Han, *Some relationships between noncentral chi-squared and normal distributions.* — Biometrika, **62**, No. 1 (1975), 213–214.
14. Yu. I. Ingster, *Some problems of hypothesis testing leading to infinitely divisible distribution.* — Math. Meth. Statist., **6** (1997), 47–69.
15. D. Jaeschke, *The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals.* — Ann. Statist., **7** (1979), 108–115.
16. M. Lifshits, *Lectures on Gaussian Porcesses*, Springer, 2012.
17. N. Meinshausen, J. Rice, *Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses.* — Ann. Statist., **34** (2006), 373–393.
18. M. Orasch, W. Pouliot, *Tabulating weighted sup-norm functionals used in change-point problem.* — J. Stat. Comput. Simul., **74** (2004), 249–276.
19. G. R. Shorack, J. A. Wellner, *Empirical Processes with Applications to Statistics*, Wiley, New York, 1986.
20. N. Stepanova, T. Pavlenko, *Goodness-of-fit tests based on sup-functionals of weighted empirical processes.* — Theory Probab. Appl., **63** (2) (2018), 358–388.

Поступило 28 апреля 2021 г.

Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, AB T6G 2G1, Canada

School of Mathematics and Statistics
Carleton University, Ottawa, ON K1S 5B6, Canada
E-mail: nstep@math.carleton.ca