On sup-functionals of weighted empirical processes with applications to inferential problems in high dimensions

Natalia Stepanova, Carleton University

joint work with Tatjana Pavlenko (KTH Royal Institute of Technology),
Yibo Wang (University of Alberta) and Lee Thompson (Carleton University)

Saint Petersburg
August 31, 2021

Overview

1. Weighted Kolmogorov–Smirnov (KS) type statistics. Choice of the class of weight functions
2. Connection to the higher criticism approach
3. Statistical properties of the properly weighted KS statistics
4. Tabulation of the limit cumulative distribution functions
5. Confidence bands
6. Detection of sparse heterogeneous mixtures
7. Attainment of the optimal detection boundary
8. Estimation of the amount of sparsity in mixture models
9. Feature selection by weighted KS fresholding in sparse classification problems
10. Selected references

Preliminaries

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.'s with a continuous CDF F on $\mathbb{R}$, and let

$$\mathbb{F}_n(t) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(X_i \le t), \quad t \in \mathbb{R},$$

be the EDF based on $X_1, \ldots, X_n$. By the Glivenko-Cantelli theorem,

$$\sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \stackrel{\text{a.s.}}{\to} 0 \text{ as } n \to \infty.$$

Consider testing the **hypothesis of goodness-of-fit**

$$H_0 : F = F_0$$

against either a **two-tailed alternative** $H_1 : F \neq F_0$ and/or an **upper-tailed alternative** $H_1' : F > F_0$. Popular test statistics for testing H_0 are:

$$D_n = \sup_{0<F_0(t)<1} \sqrt{n}|\mathbb{F}_n(t) - F_0(t)|, \quad D_n^+ = \sup_{0<F_0(t)<1} \sqrt{n}\,(\mathbb{F}_n(t) - F_0(t))$$

Preliminaries (cont-d)

For specific types of alternatives, the classical goodness-of-fit test statistics, including the Kolmogorov–Smirnov statistics D_n and D_n^+, may benefit significantly from using proper weights.

Consider the problem of testing H_0 versus two-tailed or upper-tailed alternative by using the **weighted Kolmogorov–Smirnov statistics**

$$D_n(q) = \sup_{0<F_0(t)<1} \frac{\sqrt{n}|\mathbb{F}_n(t) - F_0(t)|}{q(F_0(t))},$$

$$D_n^+(q) = \sup_{0<F_0(t)<1} \frac{\sqrt{n}\,(\mathbb{F}_n(t) - F_0(t))}{q(F_0(t))},$$

where function q belongs to some family of weight functions on $(0, 1)$.

The family of EFKP upper-class functions

Definition 1: Let q be any strictly positive function defined on $(0,1)$ with the property $q(u) = q(1-u)$ for $u \in (0, 1/2)$, which is nondecreasing in a neighborhood of zero and nonincreasing in a neighborhood of one. Such a function is called an **Erdős–Feller–Kolmogorov–Petrovski (EFKP) upper-class function** of a Brownian bridge $\{B(u), 0 \le u \le 1\}$, if there exists a constant $0 \le b < \infty$ such that

$$\limsup_{u \to 0} |B(u)|/q(u) \stackrel{\text{a.s.}}{=} b. \tag{1}$$

An EFKP upper-class function q of a Brownian bridge is called a **Chibisov–O'Reilly function** if $b = 0$ in (1).

Examples of EFKP upper-class functions

An important example of an EFKP upper-class function with $0 < b < \infty$ in (1) is the function

$$q(u) = \sqrt{u(1-u)\log\log(1/(u(1-u)))}, \quad 0 < u < 1. \tag{2}$$

Such a choice of q stems from **Khinchine's local law of the iterated logarithm**, which implies, via the representation of a Brownian bridge in terms of a standard Wiener process, that

$$\limsup_{u\to 0} \frac{|B(u)|}{\sqrt{u(1-u)\log\log(1/u(1-u))}} \stackrel{\text{a.s.}}{=} \sqrt{2}.$$

As examples of Chibisov–O'Reilly weight functions, we may consider the following functions on $(0,1)$:

$$\begin{aligned} q(u) &= \sqrt{u(1-u)}\log\log(1/(u(1-u))), \\ q_\nu(u) &= (u(1-u))^{1/2-\nu}, \quad 0 < \nu < 1/2. \end{aligned}$$

The family of regularly varying functions

Definition 2: Let q be any strictly positive function defined on $(0,1)$ with the property $q(u) = q(1-u)$ for $u \in (0, 1/2)$, which is nondecreasing in a neighborhood of zero and nonincreasing in a neighborhood of one. Such a weight function is called **regularly varying with power** $\tau \in (0, 1/2]$ if for any $b > 0$

$$\lim_{u \to 0} q(bu)/q(u) = b^{\tau}.$$

The so-called **standard deviation proportional (SDP)** weight function

$$q(u) = \sqrt{u(1-u)}, \quad 0 < t < 1,$$

is regularly varying with power $\tau = 1/2$, whereas the Chibisov–O'Reilly function $q_{\nu}(u) = (u(1-u))^{1/2-\nu}$, $\nu \in (0, 1/2)$, is regularly varying with power $\tau = 1/2 - \nu$.

Weighted Kolmogorov-Smirnov statistics

The two-sided statistic $D_n(q)$ with an EFKP upper-class function q appeared for the first time in Csörgő et al. (1986). The following weighted Kolmogorov–Smirnov type statistics are also of interest. For $0 \le a < b \le 1$, let $I = (a, b)$ and define

$$D_n(q, I) = \sup_{a<F_0(t)<b} \frac{\sqrt{n}|\mathbb{F}_n(t) - F_0(t)|}{q(F_0(t))},$$

$$D_n^+(q, I) = \sup_{a<F_0(t)<b} \frac{\sqrt{n}(\mathbb{F}_n(t) - F_0(t))}{q(F_0(t))},$$

which, for each n, have the same null distributions as the statistics

$$\sup_{u\in I} \sqrt{n}|\mathbb{U}_n(u) - u|/q(u) \quad \text{and} \quad \sup_{u\in I} \sqrt{n}(\mathbb{U}_n(u) - u)/q(u),$$

respectively, where $\mathbb{U}_n(u) = n^{-1}\sum_{i=1}^n \mathbb{I}(U_i \le u)$ is the EDF based on i.i.d. uniform $U(0, 1)$ random variables $U_1, \ldots, U_n$.

Connection to the higher criticism approach

The EDF-based tests standardized by the SDP function $q(u) = \sqrt{u(1-u)}$ have been extensively studied in the literature. If under H_0 the i.i.d. observations $U_1, \ldots, U_n$ are $U(0,1)$, a popular statistic of this kind is the **higher criticism statistic**

$$\mathrm{HC}_n = \sup_{0<u<\alpha_0} \frac{\sqrt{n}(\mathbb{U}_n(u) - u)}{\sqrt{u(1-u)}}, \quad 0 < \alpha_0 < 1.$$

It was introduced by Donoho & Jin (2004) as a competitor to the adaptive procedure of Ingster (2002) for certain multiple testing situations. The test based on HC_n rejects H_0 at level $\alpha_n \to 0$ if

$$\mathrm{HC}_n > h(n, \alpha_n),$$

where $h(n, \alpha_n) = \sqrt{2 \log \log n}\,(1 + o(1))$ as $n \to \infty$.

Connection to the higher criticism approach (cont-d)

The test statistic HC_n was derived from the random variable

$$\max_{0<\alpha\le\alpha_0} \frac{\sqrt{n}\,(M_n/n-\alpha)}{\sqrt{\alpha(1-\alpha)}},$$

where M_n is the number of hypotheses among n independently tested hypotheses $H_{0i} : X_i \sim N(0,1)$, $i = 1,\dots,n$, that are rejected at level α. Two modifications of HC_n due to Donoho & Jin (2004) and Jager & Wellner (2007) are:

$$\mathrm{HC}_n^{+} = \sup_{1/n<u<\alpha_0} \frac{\sqrt{n}(\mathbb{U}_n(u)-u)}{\sqrt{u(1-u)}}, \quad \mathrm{HC}_n^{*} = \sup_{U_{(1)}<u<U_{([\alpha_0 n])}} \frac{\sqrt{n}(\mathbb{U}_n(u)-u)}{\sqrt{u(1-u)}},$$

where $U_{(k)}$ is the kth smallest element among $U_1,\dots,U_n$.

Convergence in distribution of the HC statistics

Proposition 1. For any $0 < \alpha_0 < 1$ and any $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbf{P}\left(a_n \sup_{0<u<\alpha_0} \frac{\sqrt{n}|\mathbb{U}_n(u) - u|}{\sqrt{u(1-u)}} - b_n \le x \right) = e^{-e^{-x}},$$

$$\lim_{n\to\infty} \mathbf{P}\left(a_n \sup_{0<u<\alpha_0} \frac{\sqrt{n}(\mathbb{U}_n(u) - u)}{\sqrt{u(1-u)}} - b_n \le x \right) = e^{-\frac{1}{2}e^{-x}},$$

where

$$a_n = \sqrt{2\log\log n}, \quad b_n = 2\log\log n + \frac{1}{2}\log\log\log n - \frac{1}{2}\log(\pi).$$

Thus, regardless of a particular value of $0 < \alpha_0 < 1$, one always has the same **extreme value distribution**. Proposition 1 continues to hold for the modifications HC_n^+ and HC_n^*. It can be used to find the critical values of the asymptotic level α tests based on HC_n and its modifications.

Motivation of the study

In the statistical literature, HC-type statistics of the form

$$\sup_{k<u<1-k} \frac{\sqrt{n}(\mathbb{U}_n(u) - u)}{\sqrt{u(1-u)}},$$

where $0 < k < 1/2$ is either a fixed value of a sequence of values tending to zero as $n \to \infty$, are of interest. See, for example, Donoho & Jin (2009), Fan et al. (2013), Jin & Wang (2016), Ćmiel et al. (2020), etc. In the **sup-norm scenario**, when normalizing $\sqrt{n}(\mathbb{U}_n(u) - u)$ by $\sqrt{u(1-u)}$, one arrives at the situation where **"all the action takes place on the tails but, unfortunately, near infinity"**. This and the fact that, under H_0, the statistics HC_n, HC_n^+, and HC_n^* tend to ∞ in probability (see Prop. 1), as well as almost surely (see Ch. 16 in Shorack & Wellner (1986)), motivated us to search for a better weighed analogue of the HC statistic, for which the **"action is shifted somewhat to the middle, while properly regulated on the tails"** and whose limit distribution depends on α_0.

Motivation of the study (cont-d)

DasGupta (2008), page 611: "It is not clear for what n the asymptotics (for CH_n) start to give reasonably accurate description of the actual finite sample performance and actual finite sample comparison... Simulations would be informative and even necessary. But the range in which n has to be in order that the procedure work well when the distance between the null hypothesis and alternative so small would make the necessary simulations time consuming."

We proposed to use the weighted Kolmogorov–Smirnov test statistics

$$D_n^+(q, I) = \sup_{a<F_0(t)<b} \frac{\sqrt{n}(\mathbb{F}_n(t) - F_0(t))}{q(F_0(t))}, \quad I = (a, b) \subseteq (0, 1),$$

where q is an **EFKP upper-class function of a Brownian bridge**, as competitors to HC_n and its modifications. In order to perform well, the test procedures based on $D_n^+(q, I)$ do not require a very large sample size of $n = 10^6$ and work well even for $n = 10^2$.

Convergence in distribution of $D_n(q, I)$ and $D_n^+(q, I)$

The following extension of Th 4.2.3 in Csörgő et al. (1986) holds true.

Proposition 2. Let q be an EFKP upper-class function of a Brownian bridge $\{B(u), 0 \le u \le 1\}$. Then, under H_0, for any numbers $0 \le a < b \le 1$, as $n \to \infty$,

$$\sup_{a<F_0(t)<b} \frac{\sqrt{n}|\mathbb{F}_n(t) - F_0(t)|}{q(F_0(t))} \xrightarrow{\mathcal{D}} \sup_{a<u<b} \frac{|B(u)|}{q(u)},$$

$$\sup_{a<F_0(t)<b} \frac{\sqrt{n}(\mathbb{F}_n(t) - F_0(t))}{q(F_0(t))} \xrightarrow{\mathcal{D}} \sup_{a<u<b} \frac{B(u)}{q(u)}.$$

In particular, for the competitor of HC_n we have

$$\sup_{0<F_0(t)<\alpha_0} \frac{\sqrt{n}(\mathbb{F}_n(t) - F_0(t))}{q(F_0(t))} \xrightarrow{\mathcal{D}} \sup_{0<u<\alpha_0} \frac{B(u)}{q(u)}.$$

Test procedures based on $D_n(q, I)$ and $D_n^+(q, I)$

The main advantage of using the family of statistics $D_n^+(q, I)$ over the HC statistics is the **identification of the proper limit distribution under the null hypothesis**. This limit distribution is easily tabulated. Proposition 2 suggests the following test procedures of asymptotic level α. Set

$$D(q) := \sup_{0<u<1} |B(u)|/q(u), \quad D^+(q) := \sup_{0<u<1} B(u)/q(u).$$

The CDF of $D(q)$ is continuous on $(-\infty, \sqrt{2}) \cup (\sqrt{2}, \infty)$, and the CDF of $D^+(q)$ is continuous on $\mathbb{R}$. One would reject H_0 in favor of H_1 at level α if $D_n(q) > t_\alpha(q)$, where $P(D(q) \geq t_\alpha(q)) = \alpha$; and one would reject H_0 in favour of H_1' whenever $D_n^+(q) > t_\alpha^+(q)$, where $P(D^+(q) \geq t_\alpha^+(q)) = \alpha$.

The tests based on $D_n(q)$ and $D_n^+(q)$ are **consistent** against the alternatives $H_1 : F \neq F_0$ and $H_1' : F > F_0$, respectively.

Tabulation of the distribution of $\sup_{a<u<b} B(u)/q(u)$

1. Choose a large positive integer n. Generate n independent normal $N(0,1)$ random variables.
2. Choose a large positive integer M. Repeat step 1 M times, and for $m = 1, \ldots, M$, let $X_1^{(m)}, \ldots, X_n^{(m)}$ denote the data obtained on the mth iteration.
3. For each $m = 1, \ldots, M$, calculate the partial sums $S_k^{(m)} = \sum_{i=1}^{k} X_i^{(m)}$, $k = 1, \ldots, n$.
4. For each $m = 1, \ldots, M$, find the value of
$$D_n^{(m)} = \max_{k:k/n\in(a,b)} \frac{S_k^{(m)} - (k/n)S_n^{(m)}}{q(k/n)n^{1/2}}.$$
5. For $x \in \mathbb{R}$, use $G_{n,M}(x) = M^{-1}\sum_{m=1}^{M} \mathbb{I}\left(D_n^{(m)} \le x\right)$ to approximate the limit CDF $G(x) = \mathbf{P}\left(\sup_{a<u<b} B(u)/q(u) \le x\right)$. See Orasch & Pouliot (2004).

x	$G(x)$	x	$G(x)$	x	$G(x)$
0.74	0.01	1.81	0.34	2.57	0.67
0.87	0.02	1.83	0.35	2.60	0.68
0.95	0.03	1.85	0.36	2.63	0.69
1.02	0.04	1.87	0.37	2.66	0.70
1.07	0.05	1.89	0.38	2.69	0.71
1.11	0.06	1.91	0.39	2.72	0.72
1.16	0.07	1.93	0.40	2.76	0.73
1.19	0.08	1.95	0.41	2.79	0.74
1.23	0.09	1.97	0.42	2.83	0.75
1.26	0.10	1.99	0.43	2.87	0.76
1.29	0.11	2.01	0.44	2.91	0.77
1.32	0.12	2.03	0.45	2.95	0.78
1.35	0.13	2.05	0.46	2.99	0.79
1.37	0.14	2.07	0.47	3.03	0.80
1.40	0.15	2.09	0.48	3.08	0.81
1.42	0.16	2.12	0.49	3.13	0.82
1.45	0.17	**2.14**	**0.50**	3.18	0.83
1.47	0.18	2.16	0.51	3.23	0.84
1.49	0.19	2.18	0.52	3.29	0.85
1.51	0.20	2.20	0.53	3.35	0.86
1.54	0.21	2.22	0.54	3.42	0.87
1.56	0.22	2.25	0.55	3.48	0.88
1.58	0.23	2.27	0.56	3.55	0.89
1.60	0.24	2.30	0.57	3.62	0.90
1.63	0.25	2.32	0.58	3.70	0.91
1.65	0.26	2.35	0.59	3.79	0.92
1.67	0.27	2.37	0.60	3.89	0.93
1.69	0.28	2.40	0.61	4.00	0.94
1.71	0.29	2.43	0.62	4.14	0.95
1.73	0.30	2.46	0.63	4.30	0.96
1.75	0.31	2.49	0.64	4.48	0.97
1.77	0.32	2.51	0.65	4.73	0.98
1.79	0.33	2.54	0.66	5.16	0.99

Confidence band based on $D_n(q)$

Proposition 2 continues to hold for the statistics

$$\sup_{a<F_0(t)<b} \frac{\sqrt{n}|\mathbb{F}_n(t) - F_0(t)|}{q(\mathbb{F}_n(t))}, \qquad \sup_{a<F_0(t)<b} \frac{\sqrt{n}(\mathbb{F}_n(t) - F_0(t))}{q(\mathbb{F}_n(t))},$$

where $\sqrt{n}|\mathbb{F}_n(t) - F_0(t)|/q(\mathbb{F}_n(t)) = 0$ for $\mathbb{F}_n(t) \in \{0, 1\}$, and q is a continuous EFKP upper-class function. This result makes it possible to construct an **asymptotically correct** $100(1-\alpha)\%$ **confidence band** $[L_n(t), U_n(t)]$ for $F(t)$ on the interval $t \in [X_{(1)}, X_{(n)})$, where

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

$$U_n(t) = \min\{1, \mathbb{F}_n(t) + \frac{c_\alpha}{\sqrt{n}}\, q(\mathbb{F}_n(t))\},$$

and $c_\alpha = H^{-1}(1-\alpha)$ with $H(t) = \mathbf{P}\left(\sup_{0<u<1} |B(u)|/q(u) \le t\right)$.

Numerical comparison of confidence bands

The graph below depicts confidence bands for simulated data. The solid line is the true CDF. The solid lines above and below the middle line are a 95% confidence band $[L_n(t), U_n(t)]$. The red dashed lines are a 95% **Kolmogorov–Smirnov confidence band**. The blue dotted lines are a 95% **Eicker–Jaeschke confidence band**.

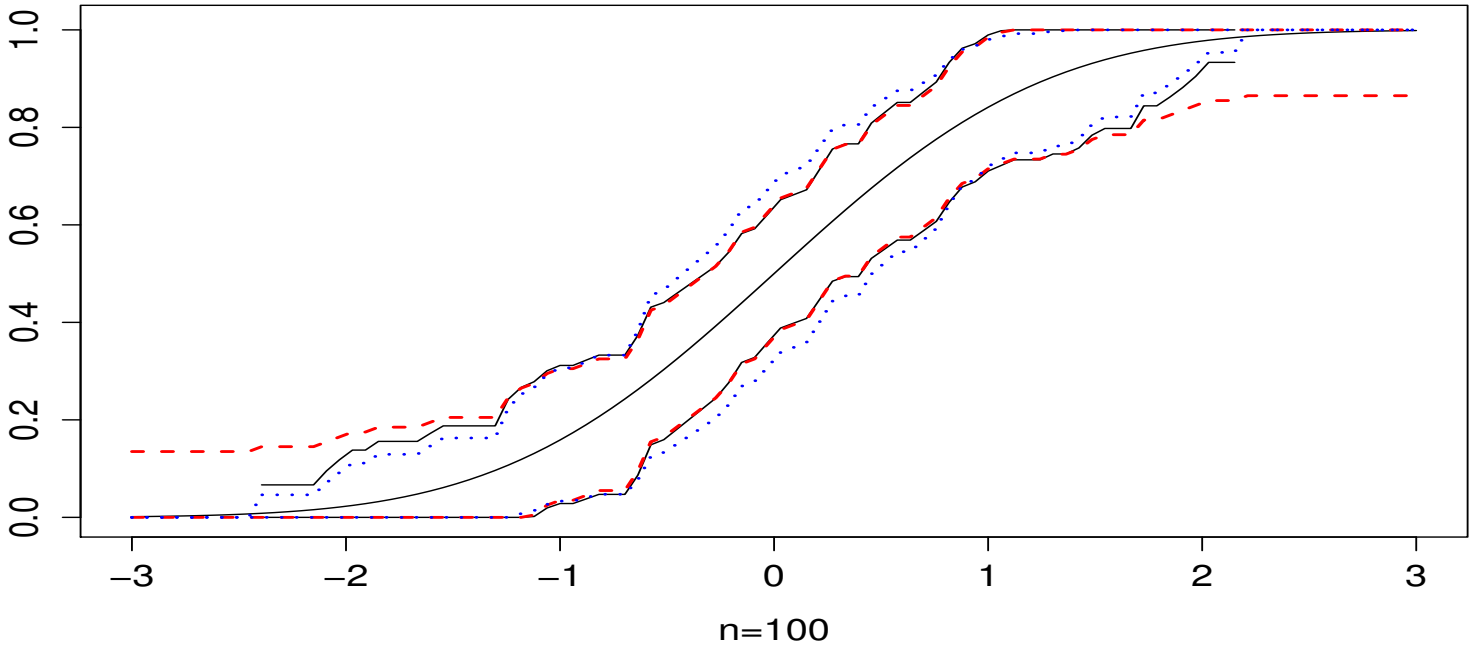
The Kolmogorov–Smirnov confidence band is derived from

$$\lim_{n\to\infty} \mathbf{P}_F\left(\sqrt{n} \sup_{-\infty<t<\infty} |\mathbb{F}_n(t) - F(t)| \le x\right) = K(x),$$

where $K(x) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2}$ is the Kolmogorov CDF. The Eicker–Jaeschke confidence band is obtained from the relation

$$\lim_{n\to\infty} \mathbf{P}_F\left(a_n \sup_{0<F(t)<1} \frac{\sqrt{n}|\mathbb{F}_n(t) - F(t)|}{\sqrt{\mathbb{F}_n(t)(1-\mathbb{F}_n(t))}} - b_n \le x\right) = e^{-2e^{-x}}.$$

Numerical comparison of confidence bands (cont-d)



When compared to the Kolmogorov–Smirnov confidence band, the confidence band $[L_n(t), U_n(t)]$ is of the same length "in the middle" and is shorter on the tails. Also, $[L_n(t), U_n(t)]$ outperforms the Eicker–Jaeschke confidence band "in the middle" and does a similar job on the tails.

Detection of sparse heterogeneous mixtures

An important particular case of a goodness-of-fit testing problem in high dimensions is that of detecting sparse heterogeneous mixtures. The latter problem has been extensively studied after the publication of Ingster (1997). First, consider testing the null hypothesis

$$H_0 : X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(0,1),$$

i.e., the specified CDF F_0 in the hypothesis of goodness-of-fit is the standard normal CDF, against a sequence of alternatives

$$H_{1,n} : X_1, \ldots, X_n \overset{i.i.d.}{\sim} (1-\varepsilon_n)N(0,1) + \varepsilon_n N(\mu_n, 1),$$

where $\varepsilon_n \sim n^{-\beta}$ for some **sparsity index** $\beta \in (1/2, 1)$ and $\mu_n = \sqrt{2r \log n}$ with $0 < r < 1$. The parameter r may be viewed as a **signal strength**. The parameters β and r are assumed unknown, and $n \to \infty$.

Detection of sparse heterogeneous mixtures (cont-d)

The subtlety of this testing problem is seen from the following fact: if $\xi_1, \xi_2, \ldots$ is a sequence of i.i.d. normal $N(0,1)$ random variables, then

$$\mathbf{P}\left(\max_{1\le i\le n}|\xi_i| \ge \sqrt{2\log n}\right) \to 0, \quad n\to\infty.$$

Hence, as $\mu_n < \sqrt{2\log n}$, the nonzero means are, in expectation, smaller than the largest X_i coming from the true component null hypothesis; and the nonzero means cannot have a visible effect in the upper extremes. This makes the problem of distinguishing between H_0 and $H_{1,n}$ very hard but yet solvable.

Detection of sparse heterogeneous mixtures (cont-d)

In order to apply the previously developed theory to the problem of testing H_0 versus $H_{1,n}$, we transform the initial observations. Namely, for $i = 1, \ldots, n$, let $Y_i = 1 - \Phi(X_i)$ and let $\mathcal{G}(u)$ denote a common CDF of the Y_i's taking values in $[0, 1]$. Then the problem of testing H_0 versus $H_{1,n}$ transforms to that of testing

$$\mathcal{H}_0 : \mathcal{G}(u) = F_0(u), \quad \text{the uniform } U(0,1) \text{ CDF}$$

against a sequence of **upper-tailed** alternatives

$$\mathcal{H}_{1,n} : \mathcal{G}(u) = F_0(u) + \varepsilon_n \left((1-u) - \Phi\left(\Phi^{-1}(1-u) - \mu_n\right)\right) > F_0(u).$$

The one-sided weighted Kolmogorov–Smirnov test statistic takes the form

$$D_n^+(q) = \sup_{0<u<1} \sqrt{n}(\mathbb{G}_n(u) - u)/q(u),$$

where $\mathbb{G}_n(u) = n^{-1} \sum_{i=1}^n \mathbb{I}(Y_i \le u)$.

Attainment of the detection boundary (cont-d)

Next theorem shows that if the parameter r is above the **detection boundary** $r = \rho(\beta)$ obtained by Ingster (1997), where

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

then the test procedure based on the one-sided statistic $D_n^+(q)$ distinguishes between $\mathcal{H}_0$ and $\mathcal{H}_{1,n}$. Since $D_n^+(q)$ does not require the knowledge of β and r, following Donoho & Jin (2004), we will call such a test procedure **optimally adaptive.**

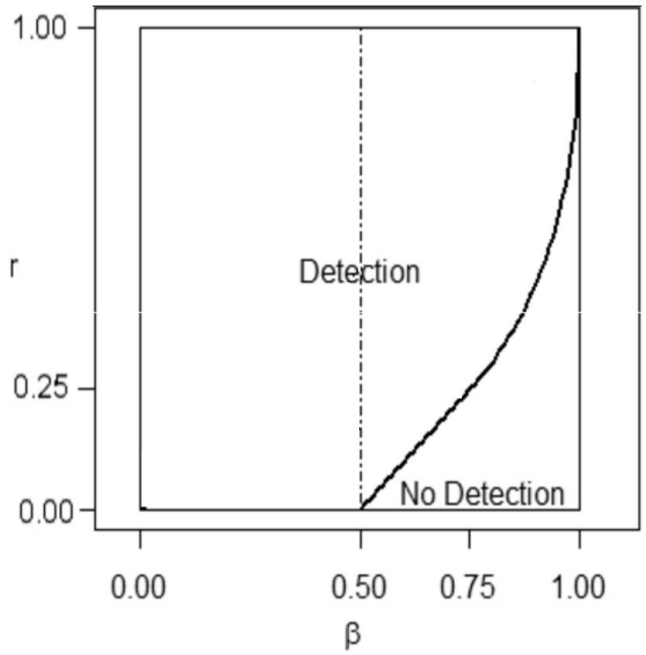


Figure: *Two regions connected to the problem of detecting sparse heterogeneous mixtures*

Attainment of the detection boundary (cont-d)

Theorem 1. *For a weight function q as in (2), consider the test of asymptotic level α that rejects $\mathcal{H}_0$ in favour of $\mathcal{H}_{1,n}$ when*

$$D_n^+(q) \geq t_\alpha^+(q),$$

where the critical value $t_\alpha^+(q)$ is determined by

$$\mathbf{P}\left(\sup_{0<u<1} B(u)/q(u) \geq t_\alpha^+(q)\right) = \alpha.$$

For every alternative $\mathcal{H}_{1,n}$ with $r > \rho(\beta)$, the asymptotic level α test based on $D_n^+(q)$ has a full power, i.e.,

$$\mathbf{P}_{\mathcal{H}_{1,n}}(D_n^+(q) \geq t_\alpha^+(q)) \to 1, \quad n \to \infty.$$

In words, when distinguishing between $\mathcal{H}_0$ and $\mathcal{H}_{1,n}$, the test procedure based on $D_n^+(q)$ **performs optimally adaptively to unknown sparsity and size of non-null effects**. See Stepanova & Pavlenko (2018).

Detection of sparse heterogeneous mixtures (cont-d)

Another model of interest, which was found to be useful in various classification problems has the form:

$$\begin{aligned} H_0' &: X_1, \ldots, X_n \overset{i.i.d.}{\sim} \chi^2_\nu(0), \\ H_{1,n}' &: X_1, \ldots, X_n \overset{i.i.d.}{\sim} (1-\varepsilon_n)\chi^2_\nu(0) + \varepsilon_n \chi^2_\nu(\delta_n), \end{aligned}$$

where $\chi^2_\nu(\delta)$ denotes the noncentral chi-square distribution with ν degrees of freedom and noncentrality parameter δ, $\varepsilon_n \sim n^{-\beta}$ for some $\beta \in (1/2, 1)$, and $\delta_n = 2r \log n$ for some $0 < r < 1$. For $\nu = 2$ this model connects to the problem of detecting covert communications (see Donoho & Jin (2004)). The parameters β and r are assumed unknown, and $n \to \infty$.

The result similar to Theorem 1 holds true: the test procedure based on $D_n^+(q)$ **distinguishes between** the (transformed) hypotheses H_0' and $H_{1,n}'$.

Estimation of the sparsity index

The **problem of estimating the fraction of nonzero means in sparse mixture models** was studies by several authors. Cai et al. (2007) showed that, in the detection region, one can consistently estimate the fraction ε_n of nonzero means in the model

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

where $\varepsilon_n = n^{-\beta}$ for $\beta \in (1/2, 1)$ and $\mu_n = \sqrt{2r \log n}$ for $0 < r < 1$, and obtained an estimator, called the **CJL estimator**, with the nearly optimal rate of convergence. In terms of a common CDF $F(t) = F_{n,\beta,r}(t)$ of the observations $X_1, \dots, X_n$, the normal mixture model (3) takes the form

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

where the parameters μ_n and ε_n are as before. The CJL procedure first estimates the mean μ_n, and then uses the estimated mean to estimate ε_n.

Estimation of the sparsity index (cont-d)

Numerically, the CJL estimator was found to be better as compared to that of Meinshausen & Rice (2006). For a given $\alpha \in (0,1)$, the CJL estimator $\varepsilon^*_{a_n} = \varepsilon^*_{a_n}(X_1, \ldots, X_n)$ satisfies

$$\mathbf{P}\left(\varepsilon^*_{a_n} \leq \varepsilon_n\right) \geq 1 - \alpha.$$

Recalling that $\varepsilon_n = n^{-\beta}$, a natural estimator β^*_n of the **sparsity index** β is then given by

$$\beta^*_n = \frac{\log(1/\varepsilon^*_{a_n})}{\log n}.$$

The quality of the CJL estimator $\varepsilon^*_{a_n}$ depends on a positive parameter a_n whose choice depends on a purpose.

Estimation of the sparsity index (cont-d)

A key step in the construction of $\varepsilon^*_{a_n}$ is the choice of an $100(1-\alpha)\%$ confidence band for the CDF $F(t)$. In Cai et al. (2007), the proposed confidence band $[\mathbb{F}^-_{a_n}(t), \mathbb{F}^+_{a_n}(t)]$ on $[0, \sqrt{2\log n}]$ is chosen so that

$$\mathbb{F}^-_{a_n}(t) \le F(t) \le \mathbb{F}^+_{a_n}(t) \quad \text{if and only if} \quad \frac{\sqrt{n}|\mathbb{F}_n(t) - F(t)|}{\sqrt{F(t)(1-F(t))}} \le a_n,$$

where a_n is the $(1-\alpha)$th quantile of $\sup_{t\in[0,\sqrt{2\log n}]} \frac{\sqrt{n}|F_n(t)-F(t)|}{\sqrt{F(t)(1-F(t))}}$. 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

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

Estimation of the sparsity index (cont-d)

The **selection region** for the two-point mixture model

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

where $\varepsilon_n = n^{-\beta}$ for $\beta \in (0,1)$ and $\mu_n = \sqrt{2r \log n}$ for $r \in (0,4)$, is shown on the Figure. The optimal procedure that provides variable selection with respect to the maximum Hamming risk in the selection region depends on β. It identifies X_j as being a nonzero mean observation if $X_j > \sqrt{(2\beta + \delta)\log n}$ for some positive $\delta = \delta_n$ such that $\delta \to 0$ and $\delta \log n \to \infty$.

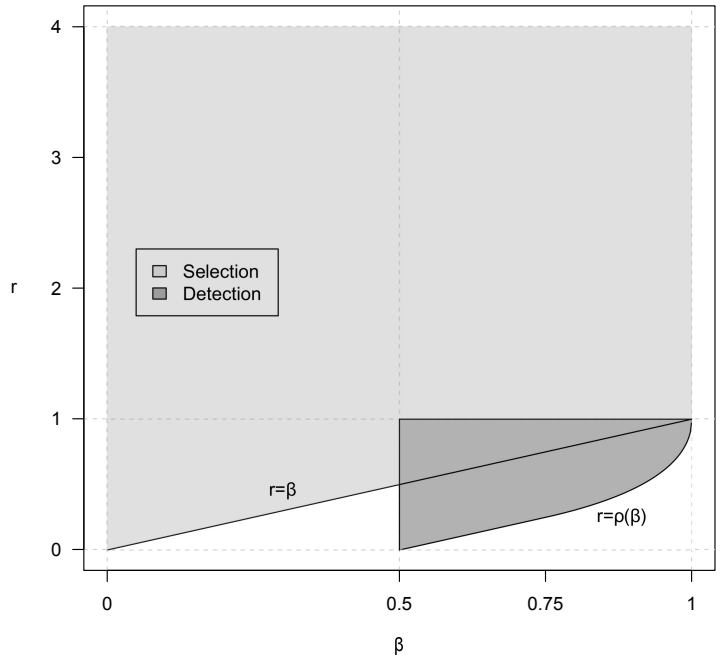


Figure: *The selection and detection regions*

Estimation of the sparsity index (cont-d)

In the context of **variable selection**, dealing with the model

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

where $\varepsilon_n = n^{-\beta}$ for $\beta \in (0,1)$ and $\mu_n = \sqrt{2r \log n}$ for $r \in (0,4)$, we modified the CJL estimator $\varepsilon^*_{a_n}$ by using the confidence band $[L_n(t), U_n(t)]$ in place of $[\mathbb{F}^-_{a_n}(t), \mathbb{F}^+_{a_n}(t)]$. The resulting estimator:

1. is a consistent estimator of $\varepsilon_n = n^{-\beta}$ in the selection region,
2. is easier to compute,
3. has a better rate of convergence.

A faster convergence rate of the new estimator is due to the fact that, at a given level of confidence $1-\alpha$, the confidence band $[L_n(t), U_n(t)]$ is **narrower** than $[\mathbb{F}^-_{a_n}(t), \mathbb{F}^+_{a_n}(t)]$, see Wang & Stepanova (2021+). Consistent estimation of ε_n (and of β) in some other (non-normal) two-point mixture models is also possible (and of interest).

Why $[L_n(t), U_n(t)]$ is better than $[\mathbb{F}^-_{a_n}(t), \mathbb{F}^+_{a_n}(t)]$?

Since $B(u)/q(u)$ with $q(u) = \sqrt{u(1-u)\log\log(1/(u(1-u)))}$ is a centered Gaussian process whose trajectories are **bounded a.s.**, by the Concentration Principe, for any $x > 0$

$$\mathbf{P}\left(\sup_{0<u<1} \frac{B(u)}{q(u)} \geq x\right) \leq 1 - \Phi\left(\frac{x-m}{\sigma}\right),$$

where $m \approx 2.14$ is the median of $\sup_{0<t<1} B(t)/q(t)$ and $\sigma^2 = \sup_{0<u<1} \mathbf{E}(B^2(u))/q^2(u) = 1/\log\log 4$. Hence, using $1 - \Phi(x) \leq \varphi(x)/x$, $x > 0$, we have for all $x > 0$

$$\mathbf{P}\left(\sup_{0<u<1} \frac{|B(u)|}{q(u)} \geq x\right) \leq \frac{\sqrt{2}\exp\left(-\frac{1}{2}\log\log 4\,(x-m)^2\right)}{\sqrt{\pi\log\log 4}\,(x-m)}.$$

This entails that $[L_n(t), U_n(t)]$ is **more accurate** than $[\mathbb{F}^-_{a_n}(t), \mathbb{F}^+_{a_n}(t)]$.

Estimation of the sparsity index (cont-d)

The effect of using $[L_n(t), U_n(t)]$ in place of $[\mathbb{F}^-_{a_n}(t), \mathbb{F}^+_{a_n}(t)]$ is seen from the following upper bounds. Given α, let c_α be such that $H(c_\alpha) = 1 - \alpha$, where $H(t) = \mathbf{P}\left(\sup_{0<u<1} |B(u)|/q(u) \le t\right)$. Assume that α_n is a sequence such that $c_{\alpha_n} = \left(\frac{4 \log n}{\log \log 4}\right)^{1/2}$, $n \ge 2$, and consider the estimator $\hat{\varepsilon}_n = \hat{\varepsilon}_{n,\alpha_n}$, which is defined similar to the CJL estimator $\varepsilon^*_{a_n}$ with $[L_n(t), U_n(t)]$ in place of $[\mathbb{F}^-_{a_n}(t), \mathbb{F}^+_{a_n}(t)]$. Then for all large enough n, uniformly in (β, r) such that $0 < \beta < 1$ and $\beta < r < 4$,

$$\mathbf{E}\left(\frac{\hat{\varepsilon}_n}{\varepsilon_n} - 1\right)^2 \le C(\beta, r)\,(\log n)^2 (\log \log n) n^{-1+\beta},$$

whereas, for the "optimal" choice of $a_n = 4\sqrt{2\pi}(\log n)^{3/2}$,

$$\mathbf{E}\left(\frac{\varepsilon^*_{a_n}}{\varepsilon_n} - 1\right)^2 \le C(\beta, r)\,(\log n)^4 n^{-1+\beta}.$$

Estimation of the sparsity index (cont-d)

Table: Numerical summary for the new estimator $\hat{\varepsilon}_n$ in the selection region

α	ε_n	$\hat{\varepsilon}_n$	$\hat{\mathbf{E}}\left(\hat{\varepsilon}_n/\varepsilon_n - 1\right)^2$	$\hat{\beta}_n$
0.01	0.00464	0.00397	0.02082	0.34300
0.05	0.00464	0.00412	0.01254	0.34070
0.1	0.00464	0.00417	0.01024	0.33995
0.5	0.00464	0.00433	0.00436	0.33759

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

Table: Numerical summary for the CJL estimator $\varepsilon^*_{a_n}$ in the selection region

α	ε_n	$\varepsilon^*_{a_n}$	$\hat{\mathbf{E}}\left(\varepsilon^*_{a_n}/\varepsilon_n - 1\right)^2$	$\beta^*_{a_n}$
0.01	0.00464	0.00386	0.02869	0.34485
0.05	0.00464	0.00403	0.01753	0.34215
0.1	0.00464	0.00408	0.01468	0.34135
0.5	0.00464	0.00431	0.00507	0.33792

Classification in high dimensions

Weighted Kolmogorov–Smirnov statistics can be used to clean the data prior to performing classification in **sparse models**.

Consider a high-dimensional two-class classification problem with equally likely classes. Namely, we assume that $\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_n^{(1)} \overset{\text{iid}}{\sim} N_p(\mathbf{0}, \mathbf{\Sigma}) \equiv \Pi_1$, and $\mathbf{X}_1^{(2)}, \ldots, \mathbf{X}_n^{(2)} \overset{\text{iid}}{\sim} N_p(\boldsymbol{\mu}, \mathbf{\Sigma}) \equiv \Pi_2$, where n is **much smaller** than p. It is assumed that $\boldsymbol{\mu} = (\boldsymbol{\mu}_{[1]}^\top, \ldots, \boldsymbol{\mu}_{[b]}^\top)^\top \neq \mathbf{0}$ and $\mathbf{\Sigma}$ is an unknown nonsingular covariance matrix of the form

$$\mathbf{\Sigma} = \mathrm{BlkDiag}\left(\mathbf{\Sigma}_{[1]}, \ldots, \mathbf{\Sigma}_{[b]}\right),$$

where $bp_0 = p$ for a given number p_0 and each $\mathbf{\Sigma}_{[k]}$ is a $p_0 \times p_0$ (nonsingular) matrix. It is also assumed that the number of blocks b tends to infinity, that

$$n = b^\theta \quad \text{for some } 0 < \theta < 1,$$

and the data are **sparse** in some sense.

Classification in high dimensions (cont-d)

Assume that we observe a p-dimensional vector $\mathbf{X}_0$, which is independent of the training samples $\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_n^{(1)}$ and $\mathbf{X}_1^{(2)}, \ldots, \mathbf{X}_n^{(2)}$, and we know that the distribution of $\mathbf{X}_0$ is either Π_1 or Π_2.

Problem: to classify the new obs. $\mathbf{X}_0$ as coming from either Π_1 or Π_2.

In an **ideal setup**, when $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known and the classes are equally likely the **optimal classifier** that minimizes the risk give by

$$(1/2)\mathbf{P}(\text{misclassifying a } \Pi_1 \text{ observation as } \Pi_2) + (1/2)\mathbf{P}(\text{misclassifying a } \Pi_2 \text{ observation as } \Pi_1)$$

has the form

$$\psi_0(\mathbf{X}_0) = \mathbb{I}\left\{(\mathbf{X}_0 - \boldsymbol{\mu}/2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \leq 0\right\}.$$

It allocates $\mathbf{X}_0$ to Π_1 when $\psi_0(\mathbf{X}_0) = 1$ and to Π_2 otherwise.

Classification boundary

In the **high** β**-sparsity** case when $(1-\theta)/2 < \beta < 1-\theta$, it is known (see Ingster et al. (2009)) that if r falls below the **classification boundary** $r = \rho^*(\beta)$, where

$$\rho^*(\beta) = (1-\theta)\,\rho\left(\frac{\beta}{1-\theta}\right), \quad (1-\theta)/2 < \beta < 1-\theta,$$

and $r = \rho(\beta)$ is the detection boundary, then **classification is impossible** in the sense that

$$\liminf_{b\to\infty} \inf_{\psi} \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in \mathbf{M}_b} \mathcal{R}(\psi) = 1/2,$$

where $\mathcal{R}(\psi)$ is either $\mathcal{R}_B(\psi) = (1/2)\mathbf{E}_{\Pi_2}(\psi) + (1/2)\mathbf{E}_{\Pi_1}(1-\psi)$ or $\mathcal{R}_M(\psi) = \max(\mathbf{E}_{\Pi_2}(\psi), \mathbf{E}_{\Pi_1}(1-\psi))$. We say that **successful classification is possible** when

$$\lim_{b\to\infty} \inf_{\psi} \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in \mathbf{M}_b} \mathcal{R}(\psi) = 0.$$

Related results

- Ingster, Pouet, and Tsybakov (2009): in the case of two normal populations with $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_{p\times p}$, constructed classifiers that provide successful classification for three different regimes when (A) n is fixed and $p \to \infty$; (B) $n \to \infty$ as $p \to \infty$, $\log n = o(\log p)$; (C) $\log n \sim \theta \log p$, $\theta \in (0, 1)$, as $p \to \infty$.
- Donoho and Jin (2009): claimed that, in the case of two normal populations with a sparse mean vector $\boldsymbol{\mu}$ and common $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_{p\times p}$, a **linear classifier with higher criticism thresholding** performs well when $n \sim c(\log p)^\theta$, $\theta \in (0, 1)$, as $p \to \infty$ (i.e., in a special case of scenario (B) above).
- Fan, Jin, and Yao (2013): in the case of two normal populations with a sparse mean vector $\boldsymbol{\mu}$ and a sparse precision matrix $\mathbf{\Sigma}^{-1}$, when $n \sim p^\theta$, $\theta \in (0, 1)$, as $p \to \infty$, proposed a linear classifier with innovative higher criticism thresholding whose classification error tends to zero.

General form of a classification rule

In the region where $(1-\theta)/2 < \beta < 1-\theta$ and $\rho^*(\beta) < r < 4$, we propose to use the classifier $\hat{\psi}_b = \hat{\psi}_b(\mathbf{X}_0; \mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_n^{(1)}; \mathbf{X}_1^{(2)}, \ldots, \mathbf{X}_n^{(2)})$ given by

$$\hat{\psi}_b = \mathbb{I}\left\{ \sum_{k=1\,:\,\hat{\omega}_k=1}^{b} (\mathbf{X}_{0,[k]} - \hat{\boldsymbol{\mu}}_{[k]}/2)^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]} \le 0 \right\},$$

where $\hat{\boldsymbol{\mu}}_{[k]} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{X}_{j,[k]}^{(2)}$ is a natural estimator of $\boldsymbol{\mu}_{[k]}$, $\widehat{\boldsymbol{\Sigma}}_{[k]}$ is the pooled estimator of $\boldsymbol{\Sigma}_{[k]}$, and $\hat{\omega}_k$ is a "good" estimator of

$$\omega_k = \mathbb{I}(k\text{th block is useful})$$

for $k = 1, \ldots, b$. The kth block is useful if $\boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \ge 2r \log b$. There are $s = [b^{1-\beta}] = o(b)$ useful blocks (which are unknown to us) among the b blocks in the data. If $\hat{\psi}_b = 1$, then $\mathbf{X}_0$ is deemed to be a Π_1 observation; if $\hat{\psi}_b = 0$, then $\mathbf{X}_0$ is deemed to be a Π_2 observation.

General form of a classification rule (cont-d)

In the high β-sparsity case, when $\rho^*(\beta) < r \leq \beta$ the classification problem at hand is much harder as compared to the case of $\beta < r < 4$, for which a good solution is available. This is because for $r < \beta$ the problem of deciding which blocks are to be included to the classification procedure is very hard.
The quality of the classifier (see Pavlenko et al. (2021+))

$$\hat{\psi}_b = \mathbb{I}\left\{ \sum_{k=1\,:\,\hat{\omega}_k=1}^{b} (\mathbf{X}_{0,[k]} - \hat{\boldsymbol{\mu}}_{[k]}/2)^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]} \leq 0 \right\},$$

depends strongly on the quality of the "selectors" $\hat{\omega}_k$, $k = 1, \ldots, b$, that decide on which data blocks are useful and are to be retained for classification purposes.

General form of a classification rule (cont-d)

The proposed estimator $\hat{\omega}_k$ of $\omega_k = \mathbb{I}(k\text{th block is useful})$ has the form

$$\hat{\omega}_k = \mathbb{I}(\hat{T}_{k,b} > \hat{t}), \quad k = 1, \ldots, b,$$

where the statistics $\{\hat{T}_{k,b}; k = 1, \ldots b; b = 2, 3, \ldots\}$ are defined as follows:

$$\hat{T}_{k,b} = \frac{(2n - p_0)n}{(2n-1)p_0} \hat{\boldsymbol{\mu}}_{[k]}^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]},$$

and $\hat{t} > 0$ is a random **threshold level** that depends on the training samples. These statistics are independent within each series and

$$\hat{T}_{k,b} \sim F_{p_0, 2n-p_0}(\gamma_{k,b}), \quad k = 1, \ldots, b,$$

where $\gamma_{k,b} = n\hat{\boldsymbol{\mu}}_{[k]}^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]}$.

General form of a classification rule (cont-d)

We convert the statistics $\hat{T}_k = \hat{T}_{k,b}$, $k = 1, \ldots, b$, to the observations taking values on $(0, 1)$:

$$U_k = 1 - F_{p_0, 2n-p_0}(\hat{T}_k; 0), \quad k = 1, \ldots, b$$

where $F_{\nu_1,\nu_2}(x; \gamma) = \mathbf{P}(F_{\nu_1,\nu_2}(\gamma) \le x)$, and put

$$\hat{k}_q = \underset{1 \le k \le [\alpha_0 b]}{\operatorname{argmax}} \frac{\sqrt{b}(k/b - U_{(k)})}{q(k/b)},$$

where $U_{(k)}$ is the kth order statistic. Returning to the F distributed observations $\hat{T}_1, \ldots, \hat{T}_b$, we now define the **weighted KS threshold** $\hat{t}_q$ by

$$\hat{t}_q = F_{p_0}^{-1}(1 - U_{(\hat{k}_q)}; 0) = \hat{T}_{(n+1-\hat{k}_q)}.$$

If $\hat{T}_k > \hat{t}_q$, the kth block is deemed **useful** and hence contributes to $\hat{\psi}_b$.

Choice of threshold

Why the threshold $\hat{t}_q = \hat{T}_{(n+1-\hat{k}_q)}$ is reasonable? In terms of a common CDF $F(u)$ of the U_ks, consider the problem of testing

$$\mathbf{H}_0 : F(u) = F_0(u), \quad \text{the uniform } U(0,1) \text{ CDF}$$

versus a sequence of *upper-tailed* alternatives

$$\mathbf{H}_{1,b}: F(u) = F_0(u) + \varepsilon_b \left((1-u) - F_{p_0,2n-p_0}(F^{-1}_{p_0,2n-p_0}(1-u;0);\gamma) \right) > F_0(u),$$

where $\varepsilon_b = b^{-\beta}$. The choice of the threshold $\hat{t}_q$ is based on the fact that $\mathcal{H}_0$ and $\mathcal{H}_{1,b}$ are **separated** by the weighted KS type test statistic

$$D_b^+(q) = \max_{1 \le k \le [\alpha_0 b]} \frac{\sqrt{b}(k/b - U_{(k)})}{q(k/b)}$$

with an EFKP upper-class function q.

Numerical summary

Estimated classification error $\mathcal{R}(\hat{\psi}_b) = \frac{1}{2}\mathbf{E}_{\Pi_2}(\hat{\psi}_b) + \frac{1}{2}\mathbf{E}_{\Pi_1}(1-\hat{\psi}_b)$

Weight function	$\mathcal{R}_{\text{est}}(\hat{\psi}_b)$
$q_1(u) = \sqrt{u(1-u)}$	0.2786
$q_2(u) = \sqrt{u(1-u)\log\log(1/(u(1-u)))}$	0.2418
$q_3(u) = (u(1-u))^{1/4}$	0.1963
$q_4(u) = \sqrt{u(1-u)\log\log(1/(u(1-u)))}$	0.1721
$q_5(u) \equiv 1$	0.1844
All blocks	**0.2122**
Only informative blocks	0.0018

$$b = 10^4,\ p_0 = 3,\ \beta = 0.375,\ r = 0.25$$

In the region of $\rho^*(\beta) < r < \beta$, where feature selection is impossible, the classifier $\hat{\psi}_b$, for which the selection of useful blocks is done by means of **weighted KS thresholding with EFKP upper-class function** q_4, provides **better classification.**

Graphical representation

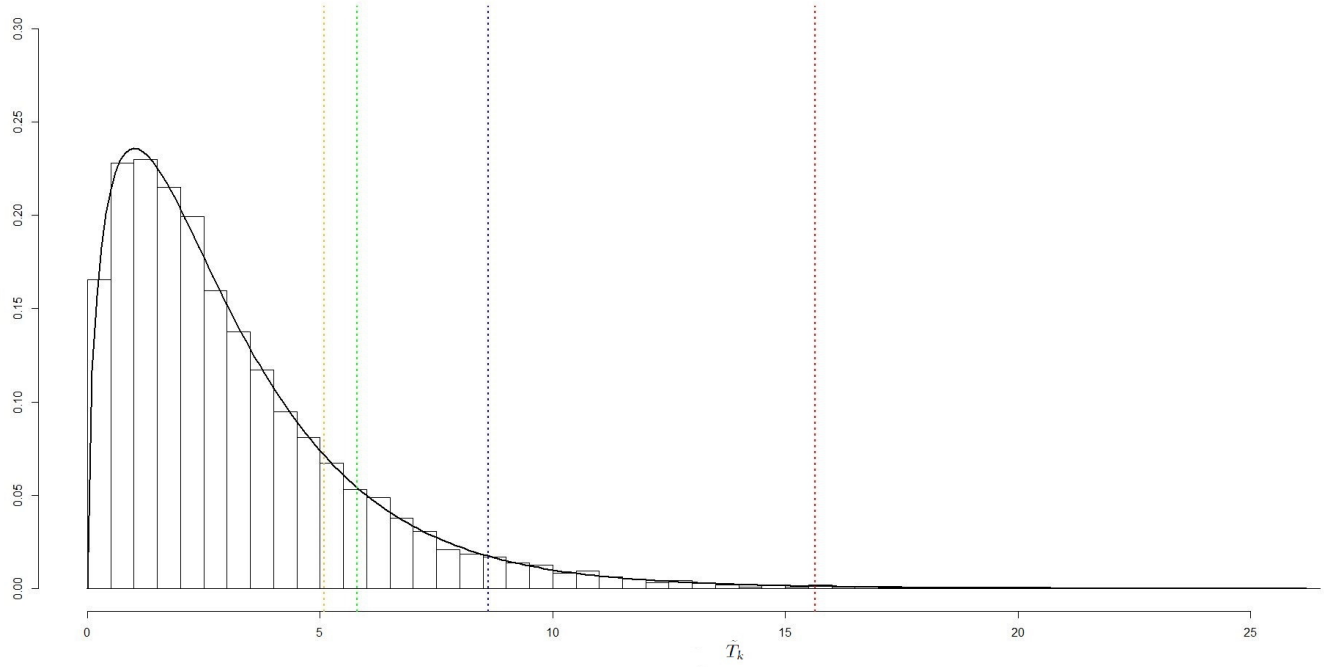


Figure: Threshold $\hat{t}_q = \hat{T}_{(n+1-\hat{k}_q)}$ with four different weight functions q for the F-distributed observations $\{\hat{T}_k : 1 \le k \le b\}$ in the region $\rho^*(\beta) < r < \beta$ with $\theta = 0.5$. The threshold $\hat{t}_q$ is shown with yellow, green, blue, and red vertical lines when q is q_2, q_3, q_4, and q_1, respectively.

Selected references

1. Csörgő, M., Csörgő, S., Horváth, L., & Mason, D. (1986). Weighted empirical and quantile processes. *Ann. Probab.* **14**, 31–85.
2. DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer Science+Business Media, LLC.
3. Donoho, D. & Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 962–994.
4. Donoho, D. & Jin, J. (2009). Feature selection by higher criticism thresholding chieves the optimal phase diagram. *Phil. Trans. R. Soc. A*, **367**, 4449–4470.
5. Eicker, F. (1979). The asymptotic distribution of the suprema of the standardized empirical processes. *Ann. Statist.* **7**, 116–138.

Selected references (cont-d)

6. Ingster, Yu. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distribution. *Math. Meth. Statist.* **6**, 47–69.
7. Ingster, Yu. I. (2002). Adaptive detection of a signal of growing dimension. I, II. *Math. Meth. Statist.* **10**, 395–421; **11**, 37–68.
8. Ingster, Yu. I., Pouet, C. & Tsybakov, A. B. (2009). Classification of sparse high-dimensional vectors. *Phil. Trans. R. Soc. A*, **367**, 4427–4448.
9. Jaeschke, D. (1979). The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *Ann. Statist.*, **7**, 108–115.
10. Orasch, M. & Pouliot, W. (2004). Tabulating weighted sup-norm functionals used in change-point problem. *J. Stat. Comput. Simul.*, **74**, 249–276.

Selected references (cont-d)

11. Pavlenko, T., Stepanova, N., & Thompson, L. (2021+). Adaptive threshold-based classification of sparse high-dimensional data. Submitted.
12. Stepanova, N. & Pavlenko, T. (2018). Goodness-of-fit tests based on weighted empirical processes. *Theory Probab. Its Appl.* **63**, 358–388.
13. Wang, Y. & Stepanova, N. (2021+). Estimating the amount of sparsity in two-point mixture models. Submitted.