Non asymptotic detection of two component mixtures

B. Laurent [1], C. Marteau [2] and Cathy Maugis-Rabusseau[1] .

Colloque ANR MixStatSeq - Toulouse 2016

1. INSA de Toulouse - IMT
2. Université Lyon I - Institut Camille Jordan

Outline

Outline

The mixture model

We have at our disposal a sample $\mathcal{S} = (X_1, \ldots, X_n)$ of i.i.d. random variables ($X_i \in \mathbb{R}^d$), having a common density f.

In an unsupervised classification context, f can be considered of the form

$$f = \sum_{j=1}^{K} \pi_j \phi(. - \mu_j),$$

where ϕ is a known density, $\pi_j \in [0, 1]$, $\mu_j \in \mathbb{R}^d$ and K are unknown parameters.

The mixture model

We have at our disposal a sample $\mathcal{S} = (X_1, \ldots, X_n)$ of i.i.d. random variables ($X_i \in \mathbb{R}^d$), having a common density f.

In an unsupervised classification context, f can be considered of the form

$$f = \sum_{j=1}^{K} \pi_j \phi(. - \mu_j),$$

where ϕ is a known density, $\pi_j \in [0,1]$, $\mu_j \in \mathbb{R}^d$ and K are unknown parameters.

Classical statistical issues

- estimation of the sequences (π_j) and $(\mu_j)_j$ (EM algorithms),
- estimation of the component number K (model selection task).

A testing point of view

In this talk, we want to assess the component number K. Our aim is to test

$$H_0 : \ f \in \mathcal{F}_0 = \left\{ x \in \mathbb{R} \mapsto \phi(x - \mu); \ \mu \in \mathbb{R}^d \right\}.$$

A testing point of view

In this talk, we want to assess the component number K. Our aim is to test

$$H_0: \ f \in \mathcal{F}_0 = \left\{ x \in \mathbb{R} \mapsto \phi(x-\mu); \ \mu \in \mathbb{R}^d \right\}.$$

against

$$\begin{aligned} H_1 \quad : \quad f \in \mathcal{F}_1 = \{ & x \in \mathbb{R} \mapsto (1-\varepsilon)\phi(x-\mu_1) + \varepsilon\phi(x-\mu_2); \\ & \varepsilon \in]0,1[\text{ and } \mu_1, \mu_2 \in \mathbb{R}^d \right\}. \end{aligned}$$

A testing point of view

In this talk, we want to assess the component number K. Our aim is to test

$$H_0 : \; f \in \mathcal{F}_0 = \left\{ x \in \mathbb{R} \mapsto \phi(x-\mu); \; \mu \in \mathbb{R}^d \right\}.$$

against

$$\begin{aligned} H_1 \quad : \quad f \in \mathcal{F}_1 = \{ & x \in \mathbb{R} \mapsto (1-\varepsilon)\phi(x-\mu_1) + \varepsilon\phi(x-\mu_2); \\ & \varepsilon \in]0,1[\text{ and } \mu_1, \mu_2 \in \mathbb{R}^d \Big\}. \end{aligned}$$

In particular, we want to

- construct a test,
- control the first kind error by a fixed level α,
- find condition on (ϵ, μ_1, μ_2) for which the two hypotheses can be separated with a prescribed error.

A testing point of view

This question has already been addressed in the literature

- Test based on the likelihood ratio,
- Seminal contribution by Y. Ingster.
- The Higher-Criticism proposed by Donoho and Jin (2004).
- ...

In all these contributions, it is assumed that $\mu = \mu_1 = 0$ is a known parameter and $d = 1$.

We want to adopt a non-asymptotic point of view.

In this talk, we will focus on the Gaussian case ($\phi = \phi_G$, the density of a standard Gaussian random variable).

References

[1]J-M. Azais, E. Gassiat and C. Mercadier. The likelihood-ratio test for general mixture models with or without structural parameters, *ESAIM Probab. Stat.*, **13**, (2009) 301-327.

[2] T. Cai, X. Jeng and J. Jin. Optimal detection of heterogeneous and heteroscedastic mixture, *J.R. Stat. Soc. Ser. B*, **73**, (2011) 629-662.

[3] D. Donoho and J. Jin. Higher Criticism for detecting sparse heterogeneous mixtures, *Annals of Statistics*, **32**, (2004) 962-994.

[4] M. Fromont et B. Laurent. Adaptive goodness-of-fit tests in a density model. *The Annals of Statistics*, **34** (2006), No 2. p.1-45.

[5] B. Garel. Recent asymptotic results in testing for mixtures, *Comput. Statist. Data. Anal.*, **51** (2007) 5295-5304.

[6] Y. Ingster. Minimax detection of a signal for l^n-balls, *Mathematical Methods of Statistics*, **7** (1999), 401-428.

Outline

Statistical setting

Given $X_1, \ldots, X_n \sim f$, our aim is to test

$$H_0: \ f \in \mathcal{F}_0 = \left\{ x \in \mathbb{R}^d \mapsto \phi(x); \ \mu \in \mathbb{R}^d \right\}.$$

Statistical setting

Given $X_1, \ldots, X_n \sim f$, our aim is to test

$$H_0 : \; f \in \mathcal{F}_0 = \left\{ x \in \mathbb{R}^d \mapsto \phi(x); \; \mu \in \mathbb{R}^d \right\}.$$

against

$$\begin{aligned} H_1 \quad : \quad f \in \mathcal{F}_1 = \Big\{ & x \in \mathbb{R}^d \mapsto (1-\varepsilon)\phi(x) + \varepsilon\phi(x-\mu); \\ & \varepsilon \in]0,1[\text{ and } \mu \in \mathbb{R}^d \Big\}. \end{aligned}$$

Statistical setting

Given $X_1, \dots, X_n \sim f$, our aim is to test

$$H_0: \; f \in \mathcal{F}_0 = \left\{ x \in \mathbb{R}^d \mapsto \phi(x); \; \mu \in \mathbb{R}^d \right\}.$$

against

$$\begin{aligned} H_1 \quad : \quad f \in \mathcal{F}_1 = \Big\{ & x \in \mathbb{R}^d \mapsto (1-\varepsilon)\phi(x) + \varepsilon\phi(x-\mu); \\ & \varepsilon \in]0,1[\text{ and } \mu \in \mathbb{R}^d \Big\}. \end{aligned}$$

In particular, the mean under H_0 is supposed to be known (and is the same in the first component of H_1).

A lower bound

Lemme

Let $\mathcal{F} \subset \mathcal{F}_1$ a subset of alternatives, and π a probability measure on $\mathcal{F}$. Then

$$\inf_{\psi_\alpha} \sup_{f \in \mathcal{F}} \mathbb{P}_f(\psi_\alpha = 0) \geq 1 - \alpha - \frac{1}{2}\left(\mathbb{E}_{H_0}[L_\pi^2(\mathbf{X})] - 1\right)^{1/2},$$

where $L_\pi^2(\mathbf{X})$ the likelihood ratio $d\mathbb{P}_\pi / d\mathbb{P}_0$ and the infimum is taken over all α-level tests.

A lower bound

Lemme

Let $\mathcal{F} \subset \mathcal{F}_1$ a subset of alternatives, and π a probability measure on $\mathcal{F}$. Then

$$\inf_{\psi_\alpha} \sup_{f \in \mathcal{F}} \mathbb{P}_f(\psi_\alpha = 0) \geq 1 - \alpha - \frac{1}{2} \left(\mathbb{E}_{H_0}[L_\pi^2(\mathbf{X})] - 1 \right)^{1/2},$$

where $L_\pi^2(\mathbf{X})$ the likelihood ratio $d\mathbb{P}_\pi / d\mathbb{P}_0$ and the infimum is taken over all α-level tests.

In particular, for some appropriate constant $C(\alpha, \beta)$,

$$\mathbb{E}_{H_0}[L_\pi^2(\mathbf{X})] \leq C(\alpha, \beta) \Rightarrow \inf_{\psi_\alpha} \sup_{f \in \mathcal{F}} \mathbb{P}_f(\psi_\alpha = 0) \geq \beta.$$

See, e.g., Ingster (1995) or Baraud (2002) for more details.

A lower bound

In our setting, we can construct a measure π such that

$$\mathbb{E}_{H_0} L_\pi^2(\mathbf{X}) \leq \mathbb{E}\left(1+\epsilon^2\left(\exp\left[\frac{\|\mu\|^2}{d}\sum_{j=1}^{d} Z_j\right]-1\right)\right)^n,$$

where the Z_j denote i.i.d. Rademacher random variables (with param. $1/2$).

A lower bound

In our setting, we can construct a measure π such that

$$\mathbb{E}_{H_0} L_\pi^2(\mathbf{X}) \leq \mathbb{E}\left(1 + \epsilon^2 \left(\exp\left[\frac{\|\mu\|^2}{d} \sum_{j=1}^{d} Z_j\right] - 1\right)\right)^n,$$

where the Z_j denote i.i.d. Rademacher random variables (with param. $1/2$).

In particular

- If $\epsilon >> 1/\sqrt{n}$, $\|\mu\|$ is allowed to tends to 0 with n.

A lower bound

In our setting, we can construct a measure π such that

$$\mathbb{E}_{H_0} L_\pi^2(\mathbf{X}) \leq \mathbb{E}\left(1+\epsilon^2\left(\exp\left[\frac{\|\mu\|^2}{d}\sum_{j=1}^{d} Z_j\right]-1\right)\right)^n,$$

where the Z_j denote i.i.d. Rademacher random variables (with param. $1/2$).

In particular

- If $\epsilon >> 1/\sqrt{n}$, $\|\mu\|$ is allowed to tends to 0 with n.
- If $\epsilon << 1/\sqrt{n}$, we can only deal with the case where $\|\mu\| \rightarrow +\infty$ with n.

A lower bound

Proposition

Let $\alpha, \beta \in]0,1[$ be fixed. Then,

$$\inf_{\Psi_\alpha} \sup_{f \in \mathcal{F}_1,\ \epsilon\|\mu\| \geq \rho} \mathbb{P}_f(\Psi_\alpha = 0) \geq \beta,$$

for all

$$\rho < \rho^\dagger := c_{\alpha,\beta} \frac{d^{1/4}}{\sqrt{n}}.$$

In some sense, testing is impossible if $\epsilon\|\mu\| \leq c_{\alpha,\beta} d^{1/4} n^{-1/2}$. We recover the asymptotic bound obtained by Cai et al. (2011) for $d = 1$.

Question : Is this bound optimal in dimension d ?

A testing procedure

The sample $\mathbf{X}$ is splited in two different parts $A = (A_1, \ldots, A_{n/2})$ and $Y = (Y_1, \ldots, Y_{n/2})$ (we assume w.l.o.g. n is even and write $n/2 = n$ in the sequel). Set

$$Z_i = \left\langle Y_i, \frac{\bar{A}_n}{\|A_n\|} \right\rangle := \langle Y_i, v_n \rangle \quad \forall i \in \{1, \ldots, n\}.$$

Conditionally to $\mathbf{A}$,

- the Z_i are i.i.d. standard Gaussian random variables under H_0.
- $Z_j \sim (1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(\langle \mu, v_n \rangle, 1)$ under H_1.

Provided v_n is a 'good' approximation of μ, we retrieve the classical uni-dimensional setting investigated in e.g. Cai et al. (2011).

In the following define $Z_{(1)} \leq \cdots \leq Z_{(n)}$ the ordered sample.

A test based on the ordered statistics

Our test statistics is defined as

$$\Psi_\alpha := \sup_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{Z_{(n-k+1)} > q_{\alpha_n, k}} \right\},$$

A test based on the ordered statistics

Assume that $n \geq 2$ and consider the subset $\mathcal{K}_n$ of $\{1, 2, \ldots, n/2\}$ defined as

$$\mathcal{K}_n = \{2^j, 0 \leq j \leq [\log_2(n/2)]\}.$$

Our test statistics is defined as

$$\Psi_\alpha := \sup_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{Z_{(n-k+1)} > q_{\alpha_n, k}} \right\},$$

A test based on the ordered statistics

Assume that $n \geq 2$ and consider the subset $\mathcal{K}_n$ of $\{1, 2, \ldots, n/2\}$ defined as

$$\mathcal{K}_n = \{2^j, 0 \leq j \leq [\log_2(n/2)]\}.$$

Our test statistics is defined as

$$\Psi_\alpha := \sup_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{Z_{(n-k+1)} > q_{\alpha_n, k}} \right\},$$

where, for all $u \in]0, 1[$, $q_{u,k}$ is the $(1-u)$-quantile of $Z_{(n-k+1)}$ under the null hypothesis and

$$\alpha_n = \sup \left\{ u \in]0, 1[, \mathbb{P}_{H_0} \left(\exists k \in \mathcal{K}_n, Z_{(n-k+1)} > q_{u,k} \right) \leq \alpha \right\}.$$

A test based on the ordered statistics

Assume that $n \geq 2$ and consider the subset $\mathcal{K}_n$ of $\{1, 2, \dots, n/2\}$ defined as

$$\mathcal{K}_n = \{2^j, 0 \leq j \leq [\log_2(n/2)]\}.$$

Our test statistics is defined as

$$\Psi_\alpha := \sup_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{Z_{(n-k+1)} > q_{\alpha_n, k}} \right\},$$

where, for all $u \in]0, 1[$, $q_{u,k}$ is the $(1-u)$-quantile of $Z_{(n-k+1)}$ under the null hypothesis and

$$\alpha_n = \sup \left\{ u \in]0, 1[, \mathbb{P}_{H_0} \left(\exists k \in \mathcal{K}_n, Z_{(n-k+1)} > q_{u,k} \right) \leq \alpha \right\}.$$

The terms $q_{\alpha_n,k}$ and α_n can be approximated (via Monte-Carlo simulations for instance) under the assumption that the X_i's have common density ϕ.

Control of the power

The test Ψ_α is in fact an aggregated testing procedure. In particular

Control of the power

The test Ψ_α is in fact an aggregated testing procedure. In particular

$$\mathbb{P}_{H_1}(\Psi_\alpha = 0) \quad = \quad \mathbb{P}_{H_1}\left(\sup_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{Z_{(n-k+1)} > q_{\alpha_n,k}} \right\} = 0\right),$$

Control of the power

The test Ψ_α is in fact an aggregated testing procedure. In particular

$$\begin{aligned}
\mathbb{P}_{H_1}(\Psi_\alpha = 0) &= \mathbb{P}_{H_1}\left(\sup_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{Z_{(n-k+1)} > q_{\alpha_n,k}} \right\} = 0 \right), \\
&= \mathbb{P}_{H_1}\left(\bigcap_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{Z_{(n-k+1)} > q_{\alpha_n,k}} \right\} = 0 \right),
\end{aligned}$$

Control of the power

The test Ψ_α is in fact an aggregated testing procedure. In particular

$$\begin{aligned}
\mathbb{P}_{H_1}(\Psi_\alpha = 0) &= \mathbb{P}_{H_1}\left(\sup_{k\in\mathcal{K}_n}\left\{\mathbf{1}_{Z_{(n-k+1)}>q_{\alpha_n,k}}\right\} = 0\right), \\
&= \mathbb{P}_{H_1}\left(\bigcap_{k\in\mathcal{K}_n}\left\{\mathbf{1}_{Z_{(n-k+1)}>q_{\alpha_n,k}}\right\} = 0\right), \\
&\leq \inf_{k\in\mathcal{K}_n} \mathbb{P}_{H_1}\left(\left\{\mathbf{1}_{Z_{(n-k+1)}>q_{\alpha_n,k}}\right\} = 0\right),
\end{aligned}$$

Control of the power

The test Ψ_α is in fact an aggregated testing procedure. In particular

$$\begin{aligned}
\mathbb{P}_{H_1}(\Psi_\alpha = 0) &= \mathbb{P}_{H_1}\left(\sup_{k\in\mathcal{K}_n}\left\{\mathbf{1}_{Z_{(n-k+1)}>q_{\alpha_n,k}}\right\} = 0\right), \\
&= \mathbb{P}_{H_1}\left(\bigcap_{k\in\mathcal{K}_n}\left\{\mathbf{1}_{Z_{(n-k+1)}>q_{\alpha_n,k}}\right\} = 0\right), \\
&\leq \inf_{k\in\mathcal{K}_n}\mathbb{P}_{H_1}\left(\left\{\mathbf{1}_{Z_{(n-k+1)}>q_{\alpha_n,k}}\right\} = 0\right),
\end{aligned}$$

In some sense, we can 'play' with the spacing k and adapt to the possible values of $\|\mu\|$ (see below).

Control of the power

Proposition

Let $\alpha, \beta \in]0,1[$ be fixed. Then, the testing procedure Ψ_α introduced above is of level α. Moreover, there exists a positive constant $C_{\alpha,\beta}$

$$\sup_{f \in \mathcal{F}_1,\ \epsilon\|\mu\| \geq \rho} \mathbb{P}_f(\Psi_\alpha = 0) \leq \beta,$$

for all $\rho \in \mathbb{R}^+$ such that

$$\rho \geq \rho^\star := C_{\alpha,\beta} \frac{d^{1/4}}{\sqrt{n}} \sqrt{\ln \ln(n)}.$$

We recover the lower bound obtained above up to a logarithmic term.

Outline

Statistical setting

Given $X_1, \ldots, X_n \sim f$, our aim is to test $(d = 1)$

$$H_0 : \; f \in \mathcal{F}_0 = \{x \in \mathbb{R} \mapsto \phi(x-\mu); \; \mu \in \mathbb{R}\} \, .$$

Statistical setting

Given $X_1, \ldots, X_n \sim f$, our aim is to test $(d = 1)$

$$H_0: \; f \in \mathcal{F}_0 = \{x \in \mathbb{R} \mapsto \phi(x - \mu); \; \mu \in \mathbb{R}\}\,.$$

against

$$\begin{aligned} H_1 \quad : \quad f \in \mathcal{F}_1 = \Big\{ & x \in \mathbb{R}^d \mapsto (1-\varepsilon)\phi(x-\mu_1) + \varepsilon\phi(x-\mu_2); \\ & \qquad \varepsilon \in]0,1[\text{ and } \mu_1, \mu_2 \in \mathbb{R}\}\,. \end{aligned}$$

Statistical setting

Given $X_1, \ldots, X_n \sim f$, our aim is to test $(d = 1)$

$$H_0 : \; f \in \mathcal{F}_0 = \{x \in \mathbb{R} \mapsto \phi(x - \mu); \; \mu \in \mathbb{R}\}.$$

against

$$H_1 \quad : \quad f \in \mathcal{F}_1 = \Big\{x \in \mathbb{R}^d \mapsto (1-\varepsilon)\phi(x-\mu_1) + \varepsilon\phi(x-\mu_2); \\ \varepsilon \in]0,1[\text{ and } \mu_1, \mu_2 \in \mathbb{R}\Big\}.$$

The parameters $\epsilon, \mu, \mu_1, \mu_2$ are unknown.

A test based on the ordered statistics

The order statistics are denoted by $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. In particular, remark that

A test based on the ordered statistics

The order statistics are denoted by $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. In particular, remark that

- the spacing of these order statistics are free with respect to the mean under H_0. For some $k < l \in \{1, \ldots, n\}$, the mean value affects the spatial position of a given $X_{(k)}$, but not $X_{(l)} - X_{(k)}$.

A test based on the ordered statistics

The order statistics are denoted by $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. In particular, remark that

- the spacing of these order statistics are free with respect to the mean under H_0. For some $k < l \in \{1, \ldots, n\}$, the mean value affects the spatial position of a given $X_{(k)}$, but not $X_{(l)} - X_{(k)}$.
- the distribution of the variables $X_{(l)} - X_{(k)}$ is known under H_0.

A test based on the ordered statistics

The order statistics are denoted by $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. In particular, remark that

- the spacing of these order statistics are free with respect to the mean under H_0. For some $k < l \in \{1, \ldots, n\}$, the mean value affects the spatial position of a given $X_{(k)}$, but not $X_{(l)} - X_{(k)}$.
- the distribution of the variables $X_{(l)} - X_{(k)}$ is known under H_0.
- it has a different behavior under H_1, provided k and l are well-chosen.

A test based on the ordered statistics

The order statistics are denoted by $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. In particular, remark that

- the spacing of these order statistics are free with respect to the mean under H_0. For some $k < l \in \{1, \ldots, n\}$, the mean value affects the spatial position of a given $X_{(k)}$, but not $X_{(l)} - X_{(k)}$.
- the distribution of the variables $X_{(l)} - X_{(k)}$ is known under H_0.
- it has a different behavior under H_1, provided k and l are well-chosen.

Our testing procedure is based on theses properties.

A test based on the ordered statistics

Our test statistics is defined as

$$\Psi_\alpha := \sup_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{X_{(n-k+1)} - X_{(k)} > q_{\alpha_n, k}} \right\},$$

A test based on the ordered statistics

Assume that $n \geq 2$ and consider the subset $\mathcal{K}_n$ of $\{1, 2, \ldots, n/2\}$ defined as

$$\mathcal{K}_n = \{2^j, 0 \leq j \leq [\log_2(n/2)]\}.$$

Our test statistics is defined as

$$\Psi_\alpha := \sup_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{X_{(n-k+1)} - X_{(k)} > q_{\alpha_n, k}} \right\},$$

A test based on the ordered statistics

Assume that $n \geq 2$ and consider the subset $\mathcal{K}_n$ of $\{1, 2, \ldots, n/2\}$ defined as

$$\mathcal{K}_n = \{2^j, 0 \leq j \leq [\log_2(n/2)]\}.$$

Our test statistics is defined as

$$\Psi_\alpha := \sup_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{X_{(n-k+1)} - X_{(k)} > q_{\alpha_n, k}} \right\},$$

where, for all $u \in]0,1[$, $q_{u,k}$ is the $(1-u)$-quantile of $X_{(n-k+1)} - X_{(k)}$ under the null hypothesis and

$$\alpha_n = \sup \left\{ u \in]0,1[, \mathbb{P}_{H_0} \left(\exists k \in \mathcal{K}_n, X_{(n-k+1)} - X_{(k)} > q_{u,k} \right) \leq \alpha \right\}.$$

A test based on the ordered statistics

Assume that $n \geq 2$ and consider the subset $\mathcal{K}_n$ of $\{1, 2, \ldots, n/2\}$ defined as

$$\mathcal{K}_n = \{2^j, 0 \leq j \leq [\log_2(n/2)]\}.$$

Our test statistics is defined as

$$\Psi_\alpha := \sup_{k \in \mathcal{K}_n} \left\{ \mathbf{1}_{X_{(n-k+1)} - X_{(k)} > q_{\alpha_n, k}} \right\},$$

where, for all $u \in]0, 1[$, $q_{u,k}$ is the $(1-u)$-quantile of $X_{(n-k+1)} - X_{(k)}$ under the null hypothesis and

$$\alpha_n = \sup \left\{ u \in]0, 1[, \mathbb{P}_{H_0} \left(\exists k \in \mathcal{K}_n, X_{(n-k+1)} - X_{(k)} > q_{u,k} \right) \leq \alpha \right\}.$$

The terms $q_{\alpha_n,k}$ and α_n can be approximated (via Monte-Carlo simulations for instance) under the assumption that the X_i's have common density ϕ.

Outline

In the following, we will concentrate our attention on two different schemes :

- The **dense** regime : the term $|\mu_1 - \mu_2|$is supposed to be bounded under H_1. In some sense, it will be impossible to detect mixtures where $\epsilon < 1/\sqrt{n}$.
- The **sparse** regime : the term $|\mu_2 - \mu_1|$ is allow to grow as $\epsilon \to 0$ (asymptotic setting)... which allows to consider smaller values for ϵ.

Main aim : Find *optimal* separation conditions on these parameters.

Outline

Guideline

We suppose in this section that $\mu_2 > \mu_1$ and

$$\mu_2 - \mu_1 \leq M,$$

for some constant M. We will

- establish a lower bound (in the Gaussian case),
- propose a consider upper bound associated to a variance-based test,
- prove that our procedure is optimal (up to a log term).

Lemme

Let $\alpha, \beta \in]0,1[$ be fixed and assume that $|\mu_2 - \mu_1| \leq M$ for some constant $M > 0$. Then, there exists $C = C(\alpha, \beta, M) > 0$ such that

$$\inf_{\psi_\alpha} \sup_{\epsilon(\mu_2-\mu_1)^2 > C/\sqrt{n}} P_f(\psi_\alpha = 0) \geq \beta.$$

Lower bound

Lemme

Let $\alpha, \beta \in]0,1[$ be fixed and assume that $|\mu_2 - \mu_1| \leq M$ for some constant $M > 0$. Then, there exists $C = C(\alpha, \beta, M) > 0$ such that

$$\inf_{\psi_\alpha} \sup_{\epsilon(\mu_2-\mu_1)^2 > C/\sqrt{n}} P_f(\psi_\alpha = 0) \geq \beta.$$

Some remarks

- testing is impossible if $\epsilon(\mu_2 - \mu_1)^2$ is smaller than $C/\sqrt{n}$.
- different result in the case where the mean μ under H_0 is available.

Upper bound (heuristic)

Under H_1, the X_i can be writen as

$$X_i = (\mu_2 - \mu_1)V_i + \eta_i, \ \forall i \in \{1 \dots n\},$$

where $V_i \sim Ber(\epsilon)$ and η_i has density $\phi(. - \mu_1)$.

Upper bound (heuristic)

Under H_1, the X_i can be writen as

$$X_i = (\mu_2 - \mu_1)V_i + \eta_i, \ \forall i \in \{1 \ldots n\},$$

where $V_i \sim Ber(\epsilon)$ and η_i has density $\phi(. - \mu_1)$. In particular

$$\mathrm{Var}(X_i) = \mathrm{Var}(\eta_i) + \epsilon(1-\epsilon)(\mu_2 - \mu_1)^2.$$

Upper bound (heuristic)

Under H_1, the X_i can be writen as

$$X_i = (\mu_2 - \mu_1)V_i + \eta_i, \ \forall i \in \{1 \dots n\},$$

where $V_i \sim Ber(\epsilon)$ and η_i has density $\phi(. - \mu_1)$. In particular

$$\mathrm{Var}(X_i) = \mathrm{Var}(\eta_i) + \epsilon(1-\epsilon)(\mu_2 - \mu_1)^2.$$

Let $\sigma^2 = \mathrm{Var}(\eta_i)$ and $\Psi_{V,\alpha}$ the test defined as

$$\Psi_{V,\alpha} = \mathbf{1}_{\{S_n^2 > \sigma^2 + c_\alpha/\sqrt{n}\}},$$

where c_α is s.t. $\mathbb{P}_{H_0}(S_n^2 - \sigma^2 > c_\alpha/\sqrt{n}) \leq \alpha$.

This test reaches the lower bound presented above (up to a constant).

Upper bound for our procedure

Proposition

There exists $C_{\alpha,\beta}$ s.t.

$$\sup_{\epsilon(\mu_2-\mu_1)^2 > C_{\alpha,\beta}\sqrt{\log\log n}/\sqrt{n}} P_f(\Psi_\alpha = 0) \leq \beta.$$

Remarks

- The proof is based on a control of the deviation of the ordered statistics and associated quantiles.
- The logarithmic loss is due to the adaptation step.
- This results holds for all symmetric and derivable density ϕ.

The asymptotic setting

$$\varepsilon \underset{n\to+\infty}{\sim} n^{-\delta} \text{ and } \mu_2 - \mu_1 \underset{n\to+\infty}{\sim} n^{-r}$$

avec $0 < \delta \leq \frac{1}{2}$ et $0 < r < \frac{1}{2}$.

Proposition

The detection boundary in the *dense* regime is $r^*(\delta) = \frac{1}{4} - \frac{\delta}{2}$

- the detection is possible when $r < r^*(\delta) = \frac{1}{4} - \frac{\delta}{2}$ (for n large enough, the power of our test is greater than $1 - \beta$)
- the detection is impossible if $r > r^*(\delta)$.

The asymptotic setting

$$\varepsilon \underset{n \to +\infty}{\sim} n^{-\delta} \text{ and } \mu_2 - \mu_1 \underset{n \to +\infty}{\sim} n^{-r}$$

avec $0 < \delta \leq \frac{1}{2}$ et $0 < r < \frac{1}{2}$.

Proposition

The detection boundary in the *dense* regime is $r^*(\delta) = \frac{1}{4} - \frac{\delta}{2}$

- the detection is possible when $r < r^*(\delta) = \frac{1}{4} - \frac{\delta}{2}$ (for n large enough, the power of our test is greater than $1 - \beta$)
- the detection is impossible if $r > r^*(\delta)$.

Proof.

$$\epsilon(\mu_2 - \mu_1)^2 > \frac{C}{\sqrt{n}} \quad \Leftrightarrow \quad \frac{1}{n^\delta} \frac{1}{n^{2r}} \gtrsim \frac{1}{\sqrt{n}}.$$

$\square$

Outline

Sparse mixtures : asymptotic setting

In this section, we consider mixtures for which

$$\epsilon << \frac{1}{\sqrt{n}} \text{ quand } n \to +\infty.$$

Proposition (Reminder)

Let $\alpha, \beta \in]0,1[$ be fixed and assume that $\mu_2 - \mu_1 \leq M$ for some given constant $M > 0$. Then there exists $C = C(\alpha, \beta, M) > 0$ such that

$$\inf_{\psi_\alpha} \sup_{\epsilon(\mu_2 - \mu_1)^2 > C/\sqrt{n}} P_f(\psi_\alpha = 0) \geq \beta.$$

According to this result, it is 'necessary to consider situations for which

$$|\mu_1 - \mu_2| \to +\infty \text{ as } n \to +\infty.$$

Gaussian asymptotic setting

Assume that

$$\phi(x) = \phi_G(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \ \forall x \in \mathbb{R}.$$

In the literature, the sparse asymptotic regime is expressed as

$$\varepsilon \underset{n \to +\infty}{\sim} n^{-\delta} \text{ and } \mu_2 - \mu_1 \underset{n \to +\infty}{\sim} \sqrt{2r \log(n)}$$

where $\frac{1}{2} < \delta < 1$ and $0 < r < 1$.

Proposition

Assume that $r > r^*(\delta)$ with

$$r^*(\delta) = \begin{cases} \delta - \frac{1}{2} & \text{if } \frac{1}{2} < \delta < \frac{3}{4} \\ (1 - \sqrt{1-\delta})^2 & \text{if } \frac{3}{4} \leq \delta < 1 \end{cases}.$$

Then, setting $f(.) = (1-\varepsilon)\phi_G(. - \mu_1) + \varepsilon\phi_G(. - \mu_2)$, we have, for n large enough,

$$\mathbb{P}_f(\Psi_\alpha = 0) \leq \beta.$$

In such a case, the separation 'conditions' are the same when the mean μ under H_0 is known (see e.g. Donoho and Jin (2004) for a description of this rate)

The 'adaptive' scheme appears to be necessary in this context.

Outline

Numerical study

Our testing procedure is compared to

- the Kolmogorov-Smirnov test.

Numerical study

Our testing procedure is compared to

- the Kolmogorov-Smirnov test.
- the Higher Criticism Let $\hat{p}_i = \mathbb{P}(Z - \bar{X} > X_i)$ where $Z \sim \mathcal{N}(0,1)$ for all $i \in \{1, \dots, n\}$ and $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \dots \leq \hat{p}_{(n)}$.

Numerical study

Our testing procedure is compared to

- the Kolmogorov-Smirnov test.
- the Higher Criticism Let $\hat{p}_i = \mathbb{P}(Z - \bar{X} > X_i)$ where $Z \sim \mathcal{N}(0,1)$ for all $i \in \{1, \dots, n\}$ and $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \dots \leq \hat{p}_{(n)}$.This test is based on the statistic

$$\widehat{HC} = \max_{1 \leq i \leq n} \frac{\sqrt{n}\left(\frac{i}{n} - \hat{p}_{(i)}\right)}{\sqrt{\hat{p}_{(i)}(1 - \hat{p}_{(i)})}}.$$

Numerical study

Our testing procedure is compared to

- the Kolmogorov-Smirnov test.
- the Higher Criticism Let $\hat{p}_i = \mathbb{P}(Z - \bar{X} > X_i)$ where $Z \sim \mathcal{N}(0,1)$ for all $i \in \{1, \dots, n\}$ and $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \dots \leq \hat{p}_{(n)}$.This test is based on the statistic

$$\widehat{HC} = \max_{1 \leq i \leq n} \frac{\sqrt{n}\left(\frac{i}{n} - \hat{p}_{(i)}\right)}{\sqrt{\hat{p}_{(i)}(1 - \hat{p}_{(i)})}}.$$

Then, define $\hat{\psi}_{HC,\alpha} = \mathbf{1}_{\widehat{HC} > \hat{q}_{HC,\alpha}}$ where $\hat{q}_{HC,\alpha}$ is the $(1-\alpha)$-quantile of $\widehat{HC}$ under H_0.

Numerical study

Our testing procedure is compared to

- the Kolmogorov-Smirnov test.
- the Higher Criticism Let $\hat{p}_i = \mathbb{P}(Z - \bar{X} > X_i)$ where $Z \sim \mathcal{N}(0,1)$ for all $i \in \{1, \ldots, n\}$ and $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \ldots \leq \hat{p}_{(n)}$.This test is based on the statistic

$$\widehat{HC} = \max_{1 \leq i \leq n} \frac{\sqrt{n}\left(\frac{i}{n} - \hat{p}_{(i)}\right)}{\sqrt{\hat{p}_{(i)}(1 - \hat{p}_{(i)})}}.$$

  Then, define $\hat{\psi}_{HC,\alpha} = \mathbf{1}_{\widehat{HC} > \hat{q}_{HC,\alpha}}$ where $\hat{q}_{HC,\alpha}$ is the $(1-\alpha)$-quantile of $\widehat{HC}$ under H_0.

We used $N = 100000$ Monte-Carlo replications of size $n = 100$ for a Gaussian mixture with $\varepsilon \in \{0.05, 0.15, 0.25, 0.35\}$ and $\mu \in [0, 10]$. .

Numerical study (Gaussian case)

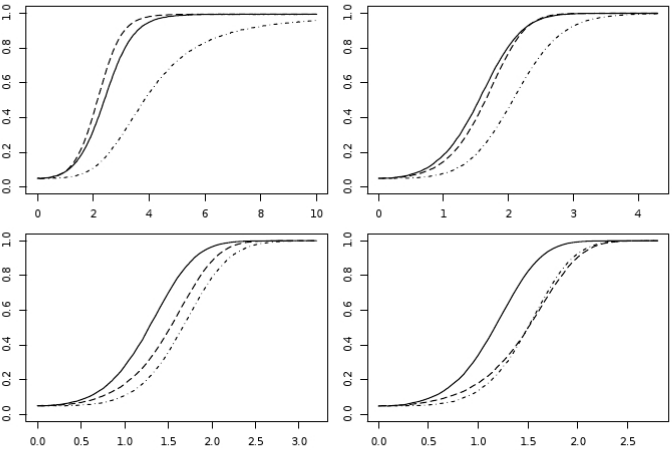

Figure : *Power function of the three considered testing procedures (continuous line for our test* Ψ_α*, dashed line for Higher Criticism and dotted line for the Kolmogorov-Smirnov test) according to* μ*, for* $\varepsilon = 0.05$ *(top-left),* 0.15 *(top right),* 0.25 *(middle left) and* 0.35 *(middle right).*

Numerical study (Laplace case)

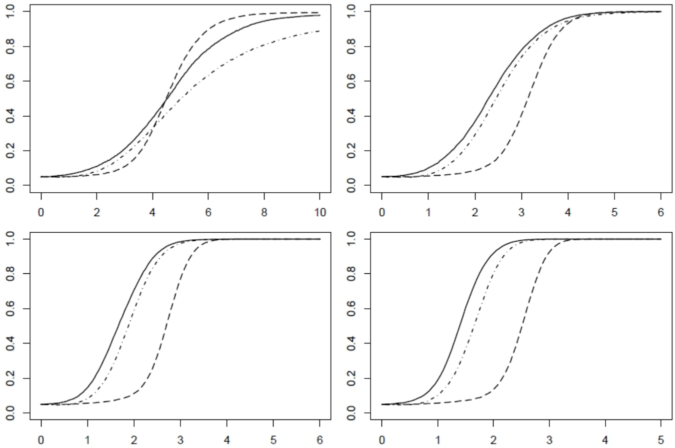

Figure : *Power function of the three considered testing procedures (continuous line for our test Ψ_α, dashed line for Higher Criticism and dotted line for the Kolmogorov-Smirnov test) according to μ, for $\varepsilon = 0.05$ (top-left), 0.15 (top right), 0.25 (middle left) and 0.35 (middle right).*

Conclusion

Possible extensions

- Complete the investigations for the general case $d \neq 1$ (sparse regime and unknown mean under the null).
- generalization to the cases where $K \geq 2$,
- take into account a possible heteroscedasticity,

B. Laurent, C. Marteau and C. Maugis-Rabusseau. Non-asymptotic detection of mixtures with unknown mean. Bernoulli, 22 (2016), pp. 242-274.

B. Laurent, C. Marteau and C. Maugis-Rabusseau. Multidimensional two component Gaussian mixtures detection. *Arxiv :1509.09129*

Non asymptotic detection of two component mixtures

B. Laurent [3], C. Marteau [4] and Cathy Maugis-Rabusseau[1] .

Colloque ANR MixStatSeq - Toulouse 2016

3. INSA de Toulouse - IMT
4. Université Lyon I - Institut Camille Jordan