Classification supervisée et non supervisée: Point de vue des Modèles de mélange Gaussien

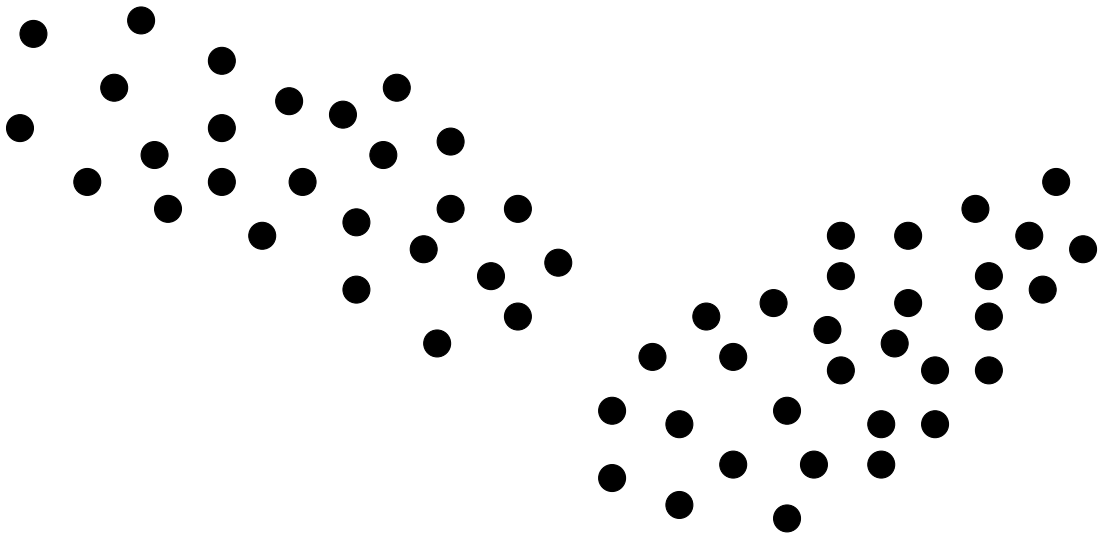
Nicolas Verzelen

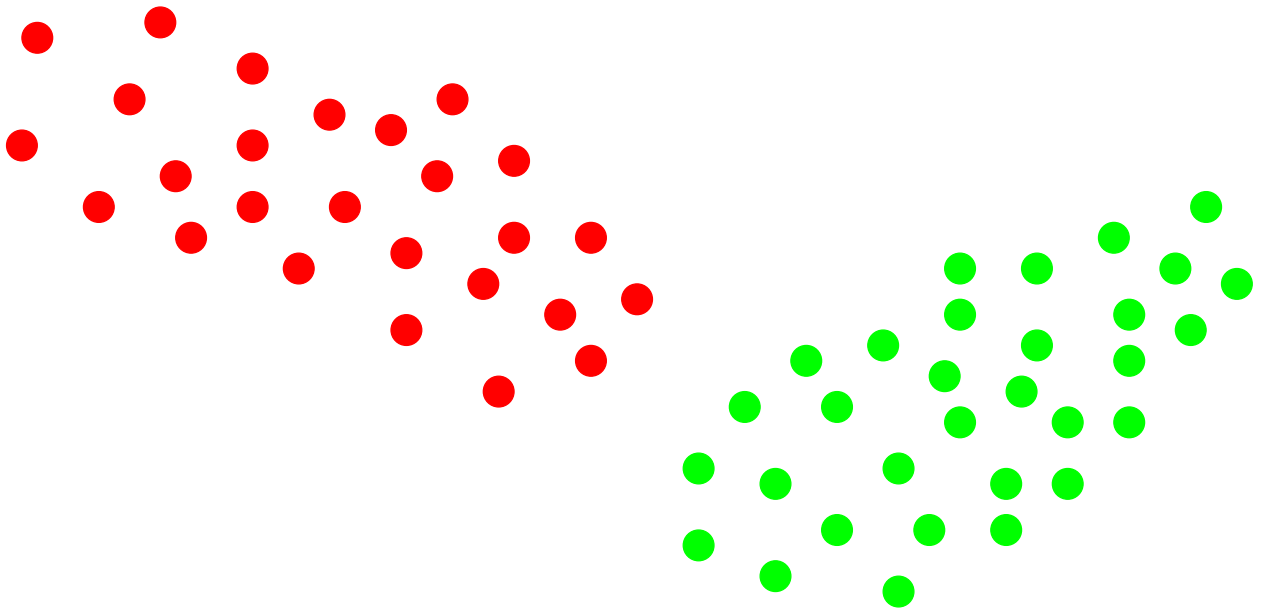
INRA, Montpellier

Travail commun avec

Ery Arias-Castro

UC San Diego





Modèles de Mélange Gaussien

k : nombre de classes
$\nu \in [0,1]^k$ tel que $\sum_{i=1}^{k} \nu_i = 1$: poids du mélange
$\forall i \in \{1, \dots K\}$, $\mu_i \in \mathbb{R}^p$ et $\mathbf{\Sigma}_i \in \mathcal{S}^+(p)$.

Modèles de Mélange Gaussien

k : nombre de classes
$\nu \in [0,1]^k$ tel que $\sum_{i=1}^{k} \nu_i = 1$: poids du mélange
$\forall i \in \{1, \dots K\}$, $\mu_i \in \mathbb{R}^p$ et $\mathbf{\Sigma}_i \in \mathcal{S}^+(p)$.

Loi de $(X, Z) \in \mathbb{R}^p \times \{1, \dots, k\}$:
$Z \sim \mathcal{M}(\nu)$
$X|Z \sim \mathcal{N}(\mu_Z, \mathbf{\Sigma}_Z)$

Loi Marginale : $X \sim f$ avec $f = \sum_{i=1}^{k} \nu_i \phi(\mu_i, \mathbf{\Sigma}_i)$

Modèles de Mélange Gaussien

k : nombre de classes
$\nu \in [0,1]^k$ tel que $\sum_{i=1}^k \nu_i = 1$: poids du mélange
$\forall i \in \{1, \dots K\}$, $\mu_i \in \mathbb{R}^p$ et $\mathbf{\Sigma}_i \in \mathcal{S}^+(p)$.

Loi de $(X, Z) \in \mathbb{R}^p \times \{1, \dots, k\}$:
$Z \sim \mathcal{M}(\nu)$
$X|Z \sim \mathcal{N}(\mu_Z, \mathbf{\Sigma}_Z)$

Loi Marginale : $X \sim f$ avec $f = \sum_{i=1}^k \nu_i \phi(\mu_i, \mathbf{\Sigma}_i)$

Données :

1. Échantillon supervisé $(X_1, Z_1), \dots, (X_n, Z_n)$.
2. Échantillon non supervisé $(X_1, \dots, X_n)$.

Différents objectifs statistiques

Échantillon supervisé :

- Estimation des paramêtres $\mu_i, \mathbf{\Sigma}_i$
- Classification
- Sélection de structure différentiant les classes
- Test d'homogénéité : $\mu_1 = \ldots = \mu_k$, $\mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_k$.

Différents objectifs statistiques

Échantillon supervisé :

- Estimation des paramêtres $\mu_i, \mathbf{\Sigma}_i$
- Classification
- Sélection de structure différentiant les classes
- Test d'homogénéité : $\mu_1 = \ldots = \mu_k$, $\mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_k$.

Échantillon non supervisé :

- Estimation des paramêtres des composantes
- Estimation de densité f.
- Clustering ; Classification
- Sélection de structure différentiant les classes
- Test d'homogénéité

Objectifs de l'exposé

Quelques questions :

- Quelle est la difficulté "intrinsèque" de chacun des problèmes ?

Objectifs de l'exposé

Quelques questions :

- Quelle est la difficulté "intrinsèque" de chacun des problèmes ?
- Quelle différence entre cas supervisé et cas non supervisé ?

Objectifs de l'exposé

Quelques questions :

- Quelle est la difficulté "intrinsèque" de chacun des problèmes ?
- Quelle différence entre cas supervisé et cas non supervisé ?
- Quel est l'effet de la présence de structure ?

Objectifs de l'exposé

Quelques questions :

- Quelle est la difficulté "intrinsèque" de chacun des problèmes ?
- Quelle différence entre cas supervisé et cas non supervisé ?
- Quel est l'effet de la présence de structure ?
- Que peut-on faire en temps polynomial ?

Objectifs de l'exposé

Quelques questions :

- Quelle est la difficulté "intrinsèque" de chacun des problèmes ?
- Quelle différence entre cas supervisé et cas non supervisé ?
- Quel est l'effet de la présence de structure ?
- Que peut-on faire en temps polynomial ?

Cadre statistique

- Cas de petite ($p \ll n$) ou grande dimension ($p \gg n$).
- Possiblement de la structure (parcimonie)

Objectifs de l'exposé

Quelques questions :

- Quelle est la difficulté "intrinsèque" de chacun des problèmes ?
- Quelle différence entre cas supervisé et cas non supervisé ?
- Quel est l'effet de la présence de structure ?
- Que peut-on faire en temps polynomial ?

Cadre statistique

- Cas de petite ($p \ll n$) ou grande dimension ($p \gg n$).
- Possiblement de la structure (parcimonie)

Estimation des composantes

Estimation de la densité

Echantillon supervisé

Estimation des composantes μ_i, $\mathbf{\Sigma}_i$ $\rightsquigarrow$ problème d'estimation paramétrique

Sans hypothèse de structure : Vitesse d'estimation (ex Kullback) en $\frac{p+(p+1)p/2}{\nu_i n}$.

Echantillon supervisé

Estimation des composantes μ_i, $\mathbf{\Sigma}_i$ $\rightsquigarrow$ problème d'estimation paramétrique

Sans hypothèse de structure : Vitesse d'estimation (ex Kullback) en $\frac{p+(p+1)p/2}{\nu_i n}$.

Estimation simultanée de moyennes et covariance avec présence de structure : parcimonie de μ, parcimonie de $\mathbf{\Sigma}^{-1}$,...

pénalisation l_1, Glasso d'Aspremont et al. (2006), Raskutti et al (2011)

Echantillon supervisé

Estimation des composantes μ_i, $\mathbf{\Sigma}_i$ $\rightsquigarrow$ problème d'estimation paramétrique

Sans hypothèse de structure : Vitesse d'estimation (ex Kullback) en $\frac{p+(p+1)p/2}{\nu_i n}$.

Estimation simultanée de moyennes et covariance avec présence de structure : parcimonie de μ, parcimonie de $\mathbf{\Sigma}^{-1}$,...

pénalisation l_1, Glasso d'Aspremont et al. (2006), Raskutti et al (2011)

Estimation de la densité du mélange

Lemme (Genovese et Wasserman (2000))

Soient $f = \sum_{i=1}^{k} a_i f_i$ *et* $g = \sum_{i=1}^{k} b_i g_i$ *deux mélanges. On a*

$$d_H^2(f,g) \leq \sum_{i=1}^{k} d_H^2(a_i f_i, b_i g_i) \ ,$$

Echantillon supervisé

Estimation des composantes μ_i, $\mathbf{\Sigma}_i$ $\rightsquigarrow$ problème d'estimation paramétrique

Sans hypothèse de structure : Vitesse d'estimation (ex Kullback) en $\frac{p+(p+1)p/2}{\nu_i n}$.

Estimation simultanée de moyennes et covariance avec présence de structure : parcimonie de μ, parcimonie de $\mathbf{\Sigma}^{-1}$,...

pénalisation l_1, Glasso d'Aspremont et al. (2006), Raskutti et al (2011)

Estimation de la densité du mélange

Lemme (Genovese et Wasserman (2000))

Soient $f = \sum_{i=1}^k a_i f_i$ et $g = \sum_{i=1}^k b_i g_i$ deux mélanges. On a

$$d_H^2(f,g) \leq \sum_{i=1}^k d_H^2(a_i f_i, b_i g_i) \ ,$$

Première conséquence : controle de l'erreur en estimation de densité.

$$\mathbb{E}[d_H^2(\widehat{f}, f)] \lesssim \sum_{i=1}^k \frac{p^2 \nu_i}{\nu_i n} \lesssim \frac{kp^2}{n}$$

Echantillon non supervisé - Estimation de densité

$$\mathcal{L}(x_1, \ldots, x_n) = \sum_{i=1}^{n} \log \left[\sum_{i=1}^{k} \nu_i \phi(\mu_i, \mathbf{\Sigma}_i) \right]$$

Maximisation de la vraisemblance $\rightsquigarrow$ Algorithme EM.

Echantillon non supervisé - Estimation de densité

$$\mathcal{L}(x_1, \ldots, x_n) = \sum_{i=1}^{n} \log \left[\sum_{i=1}^{k} \nu_i \phi(\mu_i, \mathbf{\Sigma}_i) \right]$$

Maximisation de la vraisemblance $\rightsquigarrow$ Algorithme EM.

$$\text{Rappel : } d_H^2(f, g) \leq \sum_{i=1}^{k} d_H^2(a_i f_i, b_i g_i) \ ,$$

2e conséquence : contrôle de l'entropie d'une classe de mélange

Genovese et Wasserman (2000), Ghosal et Van der Vaart (2001), Maugis et Michel (2011)

$$\mathbb{E}[d_H^2(\widehat{f}, f)] \lesssim \frac{kp^2}{n}$$

+ Inégalités oracle, adaptation à k
+ Pénalisation l_0 : Maugis et Michel (2011)

Echantillon non supervisé - Estimation de densité

$$\mathcal{L}(x_1, \ldots, x_n) = \sum_{i=1}^{n} \log \left[\sum_{i=1}^{k} \nu_i \phi(\mu_i, \mathbf{\Sigma}_i) \right]$$

Maximisation de la vraisemblance $\rightsquigarrow$ Algorithme EM.

$$\text{Rappel : } d_H^2(f, g) \leq \sum_{i=1}^{k} d_H^2(a_i f_i, b_i g_i) \ ,$$

2e conséquence : contrôle de l'entropie d'une classe de mélange

Genovese et Wasserman (2000), Ghosal et Van der Vaart (2001), Maugis et Michel (2011)

$$\mathbb{E}[d_H^2(\widehat{f}, f)] \lesssim \frac{kp^2}{n}$$

+ Inégalités oracle, adaptation à k
+ Pénalisation l_0 : Maugis et Michel (2011)

Prise en compte de structure par pénalisation de type l1 : Städler, Bühlmann, et van de Geer (2010) Xie, Pan et Chen (2008)
$\rightsquigarrow$ Peu de résultats théoriques.

Echantillon non supervisé - Estimation de densité

$$\mathcal{L}(x_1, \ldots, x_n) = \sum_{i=1}^{n} \log \left[\sum_{i=1}^{k} \nu_i \phi(\mu_i, \mathbf{\Sigma}_i) \right]$$

Maximisation de la vraisemblance $\rightsquigarrow$ Algorithme EM.

$$\text{Rappel : } d_H^2(f, g) \leq \sum_{i=1}^{k} d_H^2(a_i f_i, b_i g_i) \ ,$$

2e conséquence : contrôle de l'entropie d'une classe de mélange

Genovese et Wasserman (2000), Ghosal et Van der Vaart (2001), Maugis et Michel (2011)

$$\mathbb{E}[d_H^2(\widehat{f}, f)] \lesssim \frac{kp^2}{n}$$

+ Inégalités oracle, adaptation à k
+ Pénalisation l_0 : Maugis et Michel (2011)

Prise en compte de structure par pénalisation de type l1 : Städler, Bühlmann, et van de Geer (2010) Xie, Pan et Chen (2008)
$\rightsquigarrow$ Peu de résultats théoriques.

Point de vue Computer science : Complexité de la maximisation de vraisemblance ? Aucun algorithme polynomial connu à ce jour.

Estimation non supervisé des composantes ?

Identifiabilité : Teicher (1961) Si $(\mu_i, \mathbf{\Sigma}_i)$ sont 2 à 2 distincts, alors le modèle est identifiable (à permutation près)

Estimation non supervisé des composantes ?

Identifiabilité : Teicher (1961) Si $(\mu_i, \mathbf{\Sigma}_i)$ sont 2 à 2 distincts, alors le modèle est identifiable (à permutation près)

Objectif : Trouver un estimateur $(\nu_i, \hat{\mu}_i, \hat{\mathbf{\Sigma}}_i)$ tel que pour une permutation π

$$\sum_{i=1}^{k} d\left[\phi(\mu_i, \mathbf{\Sigma}_i), \phi(\hat{\mu}_{\pi(i)}, \hat{\mathbf{\Sigma}}_{\pi(i)})\right] \text{ petit (ex : } < \epsilon \text{ fixé).}$$

Identifiabilité robuste : $\kappa(f) := \min(\nu_i) \wedge \min \|\phi(\mu_i, \mathbf{\Sigma}_i) - \phi(\mu_j, \mathbf{\Sigma}_j)\|_{TV} > 0$

Estimation non supervisé des composantes ?

Identifiabilité : Teicher (1961) Si $(\mu_i, \mathbf{\Sigma}_i)$ sont 2 à 2 distincts, alors le modèle est identifiable (à permutation près)

Objectif : Trouver un estimateur $(\nu_i, \hat{\mu}_i, \hat{\mathbf{\Sigma}}_i)$ tel que pour une permutation π

$$\sum_{i=1}^{k} d\left[\phi(\mu_i, \mathbf{\Sigma}_i), \phi(\hat{\mu}_{\pi(i)}, \hat{\mathbf{\Sigma}}_{\pi(i)})\right] \text{ petit (ex : } < \epsilon \text{ fixé).}$$

Identifiabilité robuste : $\kappa(f) := \min(\nu_i) \wedge \min \|\phi(\mu_i, \mathbf{\Sigma}_i) - \phi(\mu_j, \mathbf{\Sigma}_j)\|_{TV} > 0$

Proposition (Une minoration Moitra et Valiant (2010))

Il existe deux modèles de modèle de mélange f, f' avec k composantes :

- *f et f' satisfont $\kappa(f) \geq 1/k$ et $\kappa(f') > 1/k$.*
- $\min_\pi \min_i \|\phi(\mu_i, \mathbf{\Sigma}_i) - \phi(\mu'_{\pi(i)}, \mathbf{\Sigma}'_{\pi(i)})\|_{TV} \geq 1/4$
- $\|f - f'\|_{TV} \leq O(e^{-k/30})$.

Conséquence : lorsque le nombre de composantes k est grand, n doit être exponentiellement grand.

Approches d'estimation des composantes

- Clustering puis estimation : Dasgupta (1999)
  Condition : séparabilité importante $\|\mu_i - \mu_j\|^2 \gtrsim \sqrt{p}\max_s \lambda_{\max}(\mathbf{\Sigma}_s)$

- Méthode spectrale (ie ACP-like) : Vempala, Wang (2004), Achlioptas McSherry (2005), Kannan, Salmasian, Sempala (2008)
  Condition : séparabilité $\|\mu_i - \mu_j\|^2 \gtrsim \left[k + \sqrt{k\log(kp)}\right]\max_s \lambda_{\max}(\mathbf{\Sigma}_s)$
  Taille d'échantillon : $n \gg \frac{k(p+\log(k))}{\kappa(f)}$

Approches d'estimation des composantes

- Clustering puis estimation : Dasgupta (1999)
  Condition : séparabilité importante $\|\mu_i - \mu_j\|^2 \gtrsim \sqrt{p}\max_s \lambda_{\max}(\mathbf{\Sigma}_s)$

- Méthode spectrale (ie ACP-like) : Vempala, Wang (2004), Achlioptas McSherry (2005), Kannan, Salmasian, Sempala (2008)
  Condition : séparabilité $\|\mu_i - \mu_j\|^2 \gtrsim \left[k + \sqrt{k\log(kp)}\right]\max_s \lambda_{\max}(\mathbf{\Sigma}_s)$
  Taille d'échantillon : $n \gg \frac{k(p+\log(k))}{\kappa(f)}$

- Relaxation SDP du critère k-means Mixon, Villar, Ward (2016)
  $\inf_{(A_t)} \sum_{i=1}^k \sum_{j\in A_i} \|x_j - \frac{1}{|A_i|}\sum_{l\in A_i} x_l\|_2^2$.
  Condition clustering approx : séparabilité $\|\mu_i - \mu_j\|^2 \gtrsim k(k\wedge p)\max_s \lambda_{\max}(\mathbf{\Sigma}_s)$
  Taille d'échantillon : $n \gg p$.

Approches d'estimation des composantes

- Clustering puis estimation : Dasgupta (1999)
  Condition : séparabilité importante $\|\mu_i - \mu_j\|^2 \gtrsim \sqrt{p}\max_s \lambda_{\max}(\mathbf{\Sigma}_s)$

- Méthode spectrale (ie ACP-like) : Vempala, Wang (2004), Achlioptas McSherry (2005), Kannan, Salmasian, Sempala (2008)
  Condition : séparabilité $\|\mu_i - \mu_j\|^2 \gtrsim \left[k + \sqrt{k\log(kp)}\right]\max_s \lambda_{\max}(\mathbf{\Sigma}_s)$
  Taille d'échantillon : $n \gg \frac{k(p+\log(k))}{\kappa(f)}$

- Relaxation SDP du critère k-means Mixon, Villar, Ward (2016)
  $\inf_{(A_t)} \sum_{i=1}^k \sum_{j\in A_i} \|x_j - \frac{1}{|A_i|}\sum_{l\in A_i} x_l\|_2^2$.
  Condition clustering approx : séparabilité $\|\mu_i - \mu_j\|^2 \gtrsim k(k\wedge p)\max_s \lambda_{\max}(\mathbf{\Sigma}_s)$
  Taille d'échantillon : $n \gg p$.

- Méthode des moments Kalai, Moitra, Valiant (2010,2010,2014)
  Condition : $n > \left(\frac{p}{\epsilon\kappa(f)}\right)^{c_k}$.
  Basé sur les $4k-2$ moments empiriques de projections de la distribution.

Classification

Test du nombre de composantes

Sélection de variables

Un modèle simplifié

$k = 2$ Classes. $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1$

Un modèle simplifié

$k = 2$ Classes. $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1$

$$X_i = \eta_i \mu_0 + (1 - \eta_i)\mu_1 + \mathbf{\Sigma}^{1/2} Z_i \ ,$$

où $\eta_1, \ldots, \eta_n \overset{\text{iid}}{\sim} \text{Bern}(\nu)$ et indépendant de $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$.

Un modèle simplifié

$k = 2$ Classes. $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1$

$$X_i = \eta_i \mu_0 + (1 - \eta_i)\mu_1 + \mathbf{\Sigma}^{1/2} Z_i \ ,$$

où $\eta_1, \ldots, \eta_n \overset{\text{iid}}{\sim} \text{Bern}(\nu)$ et indépendant de $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$.

$$X \sim \nu \mathcal{N}(\mu_0, \mathbf{\Sigma}) + (1 - \nu)\mathcal{N}(\mu_1, \mathbf{\Sigma}) \ .$$

Modèle proche de celui de l'analyse linéaire discriminante en classification supervisée.

Différence des moyennes : $\Delta\mu := \mu_1 - \mu_0$.

Perte en classification non supervisé

Soit $\psi : \mathbb{R}^p \mapsto \{0, 1\}$ un classifieur

$$
\begin{aligned}
l_{cl}(\psi) &:= \min_{\pi} \{\mathbb{P}_0[\psi(X) = \pi(1)] + \mathbb{P}_1[\psi(X) = \pi(0)]\} \\
l_{\neq}(\psi) &:= 2\,\mathbb{P}_{0,1}[\psi(X) = \psi(\tilde{X})] + \mathbb{P}_{0,0}[\psi(X) \neq \psi(\tilde{X})] \\
&\quad + \mathbb{P}_{1,1}[\psi(X) \neq \psi(\tilde{X})]
\end{aligned}
$$

Perte en classification non supervisé

Soit $\psi : \mathbb{R}^p \mapsto \{0, 1\}$ un classifieur

$$\begin{aligned} l_{cl}(\psi) &:= \min_{\pi} \{\mathbb{P}_0[\psi(X) = \pi(1)] + \mathbb{P}_1[\psi(X) = \pi(0)]\} \\ l_{\neq}(\psi) &:= 2\,\mathbb{P}_{0,1}[\psi(X) = \psi(\tilde{X})] + \mathbb{P}_{0,0}[\psi(X) \neq \psi(\tilde{X})] \\ &\quad + \mathbb{P}_{1,1}[\psi(X) \neq \psi(\tilde{X})] \end{aligned}$$

Lemme

Pour tout classifieur ψ,

$$l_{\neq}(\psi) = 2 l_{cl}(\psi)(2 - l_{cl}(\psi))$$

Comparaison entre classifications

$$l_{sc}(\psi) := \mathbb{P}_0[\psi(X) = 1] + \mathbb{P}_1[\psi(X) = 0] \geq l_{cl}(\psi)$$

Étant donné un échantillon supervisé, le risque d'un classifieur supervisé est donné par

$$R_{sc}(\widehat{\psi}) = \mathbb{E}_{(X_1,Z_1)\ldots,(X_n,Z_n)}[l_{sc}(\widehat{\psi})]$$

Proposition (Comparaison des risques minimax)

$$\inf_{\widehat{\psi}[X_1,\ldots,X_{n-1}]} \sup_{f\in\mathcal{F}} R_{cl}(\widehat{\psi}) \geq \frac{1}{3} \inf_{\widehat{\psi}[(X_1,Z_1)\ldots(X_n,Z_n)]} \sup_{f\in\mathcal{F}} R_{sc}(\widehat{\psi}) \ .$$

Si $\nu = 1/2$, *alors*

$$\inf_{\widehat{\psi}[X_1,\ldots,X_{n-1}]} \sup_{f\in\mathcal{F}} R_{\neq}(\widehat{\psi}) \geq 2 \inf_{\widehat{\psi}[(X_1,Z_1)\ldots(X_n,Z_n)]} \sup_{f\in\mathcal{F}} R_{sc}(\widehat{\psi}) \ .$$

Comparaison entre test d'homogénéité et classification

Soit Π une distribution à priori sur une classe $\mathcal{F}$.

Pour $(i,j) \in \{0,1\}^2$, $\mathbf{P}_{i,j}$ est la loi de $(X_1, \ldots, X_n, X, \tilde{X})$ où :
$(X_1, \ldots, X_n) \sim f$, $X \sim \phi(\mu_i, \mathbf{\Sigma})$, $\tilde{X} \sim \phi(\mu_j, \mathbf{\Sigma})$.

Comparaison entre test d'homogénéité et classification

Soit Π une distribution à priori sur une classe $\mathcal{F}$.

Pour $(i,j) \in \{0,1\}^2$, $\mathbf{P}_{i,j}$ est la loi de $(X_1, \ldots, X_n, X, \tilde{X})$ où :
$(X_1, \ldots, X_n) \sim f$, $X \sim \phi(\mu_i, \mathbf{\Sigma})$, $\tilde{X} \sim \phi(\mu_j, \mathbf{\Sigma})$.

$$\mathbf{P}_{i,j}^{\Pi} := \int \mathbf{P}_{i,j} \Pi(df)$$

$\mathbf{P}^{(0)}$: loi d'un échantillon de $n+2$ gaussiennes standards.

$$\begin{aligned} \sup_{f \in \mathcal{F}} R_{\neq}(\widehat{\psi}) &\geq E_{\Pi}[R_{\neq}(\widehat{\psi})] \\ &\geq 2\left[1 - \| \frac{1}{2}\left(\mathbf{P}_{0,1}^{\Pi} + \mathbf{P}_{1,0}^{\Pi}\right) - \mathbf{P}^{(0)} \|_{TV} - \| \frac{1}{2}\left(\mathbf{P}_{0,0}^{\Pi} + \mathbf{P}_{1,1}^{\Pi}\right) - \mathbf{P}^{(0)} \|_{TV}\right] \end{aligned}$$

Comparaison entre test d'homogénéité et classification

Soit Π une distribution à priori sur une classe $\mathcal{F}$.

Pour $(i,j) \in \{0,1\}^2$, $\mathbf{P}_{i,j}$ est la loi de $(X_1, \ldots, X_n, X, \tilde{X})$ où :
$(X_1, \ldots, X_n) \sim f$, $X \sim \phi(\mu_i, \mathbf{\Sigma})$, $\tilde{X} \sim \phi(\mu_j, \mathbf{\Sigma})$.

$$\mathbf{P}^{\Pi}_{i,j} := \int \mathbf{P}_{i,j} \Pi(df)$$

$\mathbf{P}^{(0)}$: loi d'un échantillon de $n+2$ gaussiennes standards.

$$\begin{aligned} \sup_{f \in \mathcal{F}} R_{\neq}(\widehat{\psi}) &\geq E_{\Pi}[R_{\neq}(\widehat{\psi})] \\ &\geq 2\left[1 - \| \frac{1}{2}\left(\mathbf{P}^{\Pi}_{0,1} + \mathbf{P}^{\Pi}_{1,0}\right) - \mathbf{P}^{(0)} \|_{TV} - \| \frac{1}{2}\left(\mathbf{P}^{\Pi}_{0,0} + \mathbf{P}^{\Pi}_{1,1}\right) - \mathbf{P}^{(0)} \|_{TV}\right] \end{aligned}$$

$\| \mathbf{P}^{\Pi}_{i,j} - \mathbf{P}^{(0)} \|_{TV}$ est relié au risque du test

$$H_0 : (X_1, \ldots, X_n, X, \tilde{X}) \sim \mathbf{P}^{(0)} \quad \text{contre} \quad H_1 : (X_1, \ldots, X_n, X, \tilde{X}) \sim \mathbf{P}^{\Pi}_{i,j}$$

Frontière de détection/classification

Données :

- Classe $\mathcal{F}$ de mélange.
  Ex1 : $\mathbf{\Sigma}$ fixe, μ_0 et μ_1 arbitraires. Ex2 : $\mathbf{\Sigma}$ fixe, $\mu_1 - \mu_0$ s-parcimonieux.
- une distance $d(\mu_0, \mu_1)$ entre les deux composantes.

Résultats asymptotique $(p, n) \to \infty$.

Frontière de détection/classification

Données :

- Classe $\mathcal{F}$ de mélange.
  Ex1 : $\mathbf{\Sigma}$ fixe, μ_0 et μ_1 arbitraires. Ex2 : $\mathbf{\Sigma}$ fixe, $\mu_1 - \mu_0$ s-parcimonieux.
- une distance $d(\mu_0, \mu_1)$ entre les deux composantes.

Résultats asymptotique $(p, n) \to \infty$.

Risque de **détection** de mélange :
$\gamma(T, r) := \sup_{f \in \mathcal{F},\ \mu_0 = \mu_1} \mathbb{E}_f[T = 1] + \sup_{f, d(\mu_0, \mu_1) > r_n} \mathbb{E}_f[T = 0]$

Frontière de détection "molle" r_n^* :

- si $r_n \ll r_n^*$ alors $\inf_{T_n} \gamma(T_n, r_n) \to 1$.
- si $r_n \gg r_n^*$ il existe une suite de tests T_n, $\gamma(T_n, r_n) \to 0$.

Frontière de détection/classification

Données :

- Classe $\mathcal{F}$ de mélange.
  Ex1 : $\mathbf{\Sigma}$ fixe, μ_0 et μ_1 arbitraires. Ex2 : $\mathbf{\Sigma}$ fixe, $\mu_1 - \mu_0$ s-parcimonieux.
- une distance $d(\mu_0, \mu_1)$ entre les deux composantes.

Résultats asymptotique $(p, n) \to \infty$.

Risque de **détection** de mélange :
$\gamma(T, r) := \sup_{f \in \mathcal{F},\ \mu_0 = \mu_1} \mathbb{E}_f[T = 1] + \sup_{f, d(\mu_0, \mu_1) > r_n} \mathbb{E}_f[T = 0]$

Frontière de détection "molle" r_n^* :

- si $r_n \ll r_n^*$ alors $\inf_{T_n} \gamma(T_n, r_n) \to 1$.
- si $r_n \gg r_n^*$ il existe une suite de tests T_n, $\gamma(T_n, r_n) \to 0$.

Définition correspondantes de frontière de **classification non supervisé**, **classification supervisé** :
Si $r_n \ll r_n^*$ alors $\inf_{\widehat{\psi}_n} \sup_{f, d(\mu_0, \mu_1) > r_n} R_{\neq}(\widehat{\psi}_n) \to 2$
si $r_n \gg r_n^*$ alors il existe $\widehat{\psi}$ tel que $\sup_{f, d(\mu_0, \mu_1) > r_n} R_{\neq}(\widehat{\psi}_n) \to 0$

Cas isotropique $\boldsymbol{\Sigma} = \mathbf{I}$

Classification supervisé : Ingster, Pouet, Tsybakov (2009), Donoho et Jin (2009)

Classification non supervisé : Azizyan, Singh et Wasserman (2013), Arias-Castro et V. (2014)

Cas isotropique $\Sigma = \mathbf{I}$

Classification supervisé : Ingster, Pouet, Tsybakov (2009), Donoho et Jin (2009)

Classification non supervisé : Azizyan, Singh et Wasserman (2013), Arias-Castro et V. (2014)

Frontières dans le cas non parcimonieux (en fonction de $\|\Delta\mu\|^2$).

	$p/n \to 0$	$p/n \to \infty$
Test supervisé	$\sqrt{p}/n$	$\sqrt{p}/n$
Test de mélange	$\sqrt{p/n}$	$\sqrt{p/n}$
Classif. supervisée	$\to \infty$	$\sqrt{p/n}$
Classif. non supervisée	$\to \infty$	$\sqrt{p/n}$

Remarques : En petite dimension, le risque de classification est essentiellement dû au risque du classifieur de Bayes.

En grande dimension, à la frontière on a $< \frac{\widehat{\Delta\mu}}{\|\Delta\mu\|}, \frac{\Delta\mu}{\|\Delta\mu\|} > \to 0$ mais $< \frac{\widehat{\Delta\mu}}{\|\Delta\mu\|}, \Delta\mu > \to \infty$

Test non supervisé

Proposition

Tous les tests sont asymptotiquement impuissants si

$$\|\Delta\mu\|^2 \ll \sqrt{p/n} \ .$$

Covariance du vecteur X :

$$\mathrm{Cov}(X) = \nu(1-\nu)\Delta\mu\Delta\mu^\top + \mathbf{I} \ .$$

$$\widehat{\mathbf{\Sigma}} = \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top$$

Proposition

La plus grande valeur propre est asymptotiquement puissante si

$$\nu(1-\nu)\|\Delta\mu\|^2 \geq C\sqrt{p/n} \ .$$

Cas isotropique ($\Sigma = I$) avec parcimonie de $\Delta\mu$

$\Delta\mu$ est s-**parcimonieux** : s composantes au plus de $\Delta\mu$ sont non nulles.
Ingster, Pouet, Tsybakov (2009), Donoho et Jin (2009), Azizyan, Singh et Wasserman (2013), Arias-Castro et V. (2014)

Frontières dans le cas parcimonieux (en fonction de $\|\Delta\mu\|^2$).

	$\frac{s\log(p/s)}{n} \to 0$	$\frac{s\log(p/s)}{n} \to \infty$
Test supervisé	$\frac{s\log(p/s)}{n} \wedge \frac{\sqrt{p}}{n}$	$\frac{s\log(p/s)}{n} \wedge \frac{\sqrt{p}}{n}$
Test de mélange	$\sqrt{\frac{s\log(p/s)}{n}}$	$\frac{s\log(p/s)}{n} \wedge \sqrt{\frac{p}{n}}$
Classif. supervisée	$\to \infty$	$\frac{s\log(p/s)}{n} \wedge \sqrt{\frac{p}{n}}$
Classif. non supervisée	$\to \infty$	$\frac{s\log(p/s)}{n} \wedge \sqrt{\frac{p}{n}}$

Proposition

Supposons que $p/s \to \infty$. Alors aucun test est asymptotiquement puissant si

$$\|\Delta\mu\|^2 \ll \sqrt{p/n} \ .$$

et

$$\limsup \frac{\|\Delta\mu\|^2}{\sqrt{\frac{s\log(p/s)}{n}} \vee \frac{s\log(p/s)}{n}} < 1 \ .$$

$$\mathrm{Cov}(X) = \nu(1-\nu)\Delta\mu\Delta\mu^\top + \mathbf{I} \ .$$

Statistique de la plus grande valeur propre sparse :

$$\hat{\lambda}_s^{\max} := \max_{\|u\|=1,\, \|u\|_0 \le s} u^\top \hat{\mathbf{\Sigma}} u$$

Proposition

Supposons que $p/s \to \infty$.

- Détection. *La plus grande valeur propre parcimonieuse* $\hat{\lambda}_s^{\max}$ *est asymptotiquement puissante si*

$$\limsup \nu(1-\nu) \frac{\|\Delta\mu\|^2}{\sqrt{\frac{s\log(p/s)}{n}} \vee \frac{s\log(p/s)}{n}} > C \ .$$

- Sélection de variable. *Soit* $\hat{u}$ *le vecteur associé à* $\hat{\lambda}_s^{\max}$. *Si*

$$\|\Delta\mu\|^2 \gg \sqrt{\frac{s\log(p/s)}{n}} \vee \frac{s\log(p/s)}{n} \ ,$$

et si le dynamique range de $\Delta\mu$ *est borné alors le support de* $\hat{u}$ *est consistant pour le support de* $\Delta\mu$.

Estimation de vecteur propre en ACP parcimonieuse : Cai et al (2013).

Calcul en temps polynomial

Toutes les méthodes précédentes ont des complexités polynomiales
... sauf $\hat{\lambda}_s^{\max} := \max_{\|u\|=1,\, \|u\|_0 \leq s} u^\top \hat{\mathbf{\Sigma}} u$

Solution naive calculer $\hat{\lambda}_1^{\max} := \max_{\|u\|=1,\, \|u\|_0 \leq 1} u^\top \hat{\mathbf{\Sigma}} u$
Autres solution (SDP) : D'aspremont et al. (2007) Berthet et Rigollet (2013)

Calcul en temps polynomial

Toutes les méthodes précédentes ont des complexités polynomiales
... sauf $\hat{\lambda}_s^{\max} := \max_{\|u\|=1,\, \|u\|_0 \leq s} u^\top \hat{\mathbf{\Sigma}} u$

Solution naive calculer $\hat{\lambda}_1^{\max} := \max_{\|u\|=1,\, \|u\|_0 \leq 1} u^\top \hat{\mathbf{\Sigma}} u$
Autres solution (SDP) : D'aspremont et al. (2007) Berthet et Rigollet (2013)

Condition suffisante en temps polynomial dans le cas parcimonieux.

	$\frac{s\log(p/s)}{n} \to 0$	$\frac{s\log(p/s)}{n} \to \infty$ (et $\log(p/s) = o(n)$)
Test supervisé	$\frac{s\log(p/s)}{n} \wedge \frac{\sqrt{p}}{n}$	$\frac{s\log(p/s)}{n} \wedge \frac{\sqrt{p}}{n}$
Test de mélange	$\sqrt{\frac{s^2\log(p)}{n}} \wedge \sqrt{\frac{p}{n}}$	$\sqrt{\frac{s^2\log(p)}{n}} \wedge \sqrt{\frac{p}{n}}$
Classif. supervisée	$\to \infty$	$\frac{s\log(p/s)}{n} \wedge \sqrt{\frac{p}{n}}$
Classif. non supervisée	$\to \infty$	$\sqrt{\frac{s^2\log(p)}{n}} \wedge \sqrt{\frac{p}{n}}$

Calcul en temps polynomial

Toutes les méthodes précédentes ont des complexités polynomiales
... sauf $\hat{\lambda}_s^{\max} := \max_{\|u\|=1,\, \|u\|_0 \leq s} u^\top \hat{\mathbf{\Sigma}} u$

Solution naive calculer $\hat{\lambda}_1^{\max} := \max_{\|u\|=1,\, \|u\|_0 \leq 1} u^\top \hat{\mathbf{\Sigma}} u$
Autres solution (SDP) : D'aspremont et al. (2007) Berthet et Rigollet (2013)

Condition suffisante en temps polynomial dans le cas parcimonieux.

	$\frac{s\log(p/s)}{n} \to 0$	$\frac{s\log(p/s)}{n} \to \infty$ (et $\log(p/s) = o(n)$)
Test supervisé	$\frac{s\log(p/s)}{n} \wedge \frac{\sqrt{p}}{n}$	$\frac{s\log(p/s)}{n} \wedge \frac{\sqrt{p}}{n}$
Test de mélange	$\sqrt{\frac{s^2\log(p)}{n}} \wedge \sqrt{\frac{p}{n}}$	$\sqrt{\frac{s^2\log(p)}{n}} \wedge \sqrt{\frac{p}{n}}$
Classif. supervisée	$\to \infty$	$\frac{s\log(p/s)}{n} \wedge \sqrt{\frac{p}{n}}$
Classif. non supervisée	$\to \infty$	$\sqrt{\frac{s^2\log(p)}{n}} \wedge \sqrt{\frac{p}{n}}$

Le terme s^2 est vraisemblablement intrinsèque...

Problème de la clique plantée et réduction computationnel

Détection en ACP parcimonieuse : X échantillon de covariance $\mathbf{\Sigma} = \mathbf{I} + \lambda v v^\top$ et v est s-parcimonieux.

$$H_0 : \lambda = 0 \text{ contre } H_1 : \lambda > \lambda_* \ .$$

Théorème (Berthet/Rigollet (2013))

Si le problème de la clique plantée n'est pas résolvable en temps polynomial, alors pour tout $1 < \delta < 2$ aucun test calculable en temps polynomial n'est asymptotiquement puissant si

$$\lambda_* \lesssim \sqrt{\frac{s^\delta \log(p)}{n}}$$

Covariance Σ inconnu.

Dans la suite, $\mathbf{\Sigma}$ est inconnu (mais inversible). Shao et al. (2011) (Classif supervisé)

Frontières (distance de Mahalanobis $\Delta\mu^\top \mathbf{\Sigma}^{-1}\Delta\mu$).

	$p = o(n)$	$p \gg n$
Test supervisé	$\left(\frac{p}{n}\right)^{1/2}$	$e^{O(p/n)}$
Test de mélange	$\left(\frac{p}{n}\right)^{1/4}$	$e^{O(p/n)}$
Classif. supervisée	$\to \infty$	$e^{O(p/n)}$
Classif. non supervisée	$\to \infty$	$e^{O(p/n)}$

Frontières sous s-parcimonie (en fonction de $\frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma}\Delta\mu}$).

	$s\log(p/s) = o(n)$	$s\log(p/s) \gg n$
Test supervisé	$\left(\frac{s\log(p/s)}{n}\right)^{1/2}$	$e^{O(s\log(p/s)/n)}$
Test de mélange	$\left(\frac{s\log(p/s)}{n}\right)^{1/4}$	$e^{O(s\log(p/s)/n)}$
Classif. supervisée	$\to \infty$	$e^{O(s\log(p/s)/n)}$
Classif. non supervisée	$\to \infty$	$e^{O(s\log(p/s)/n)}$

Cas symmétrique ($\nu = 1/2$)

Proposition

Supposons que $\nu = 1/2$.

- Cas non parcimonieux. *Tous les tests sont asymptotiquement impuissants si*

$$\frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu} \ll (p/n)^{1/4} \ .$$

Si $p \gg n$, *alors tous les tests sont asymptotiquement impuissants si*

$$\limsup \frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu} e^{-Cp/n} < 1 \ .$$

- Cas parcimonieux ($p/s \to \infty$). *Tous les tests sont asymp. impuissants si*

$$\limsup \frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu} \left(\frac{s \log(p/s)}{n} \right)^{-1/4} \leq C_1 \ .$$

Si $\liminf \frac{s}{n} \log(p/s) \geq C_2$, *alors tous les tests sont asymptotiquement impuissants lorsque*

$$\limsup \frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu} e^{-C_3 \frac{s \log(p/s)}{n}} \leq 1 \ .$$

Test de moments généralisés

Statistique de kurtosis généralisé :

$$\min_{\|u\|_0 \leq s} \frac{\sum_i \left[u^\top (X_i - \bar{X})\right]^4}{\left(\sum_i \left[u^\top (X_i - \bar{X})\right]^2\right)^2} \ .$$

Test de moments généralisés

Statistique de kurtosis généralisé :

$$\min_{\|u\|_0 \leq s} \frac{\sum_i \left[u^\top (X_i - \bar{X})\right]^4}{\left(\sum_i \left[u^\top (X_i - \bar{X})\right]^2\right)^2} .$$

Statistique de moment d'ordre 1 :

$$\max_{\|u\|_0 \leq s} \frac{\sum_i \left|u^\top (X_i - \bar{X})\right|}{\left(\sum_i \left[u^\top (X_i - \bar{X})\right]^2\right)^{1/2}} .$$

Test de moments généralisés

Statistique de kurtosis généralisé :

$$\min_{\|u\|_0 \leq s} \frac{\sum_i \left[u^\top (X_i - \bar{X})\right]^4}{\left(\sum_i \left[u^\top (X_i - \bar{X})\right]^2\right)^2} .$$

Statistique de moment d'ordre 1 :

$$\max_{\|u\|_0 \leq s} \frac{\sum_i \left|u^\top (X_i - \bar{X})\right|}{\left(\sum_i \left[u^\top (X_i - \bar{X})\right]^2\right)^{1/2}} .$$

Proposition

Supposons que $n \gg s \log(p/s)$. *Le test d'ordre 1 est asymptotiquement puissant si* $|\nu - 1/2| < \frac{1}{6}$ *et si*

$$\frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu} \gg \left(s \frac{\log(p/s)}{n}\right)^{1/4} .$$

Minoration en terme de distance de Mahalanobis

Proposition

Soit $\Delta\mu$ un vecteur 1-parsimonieux fixé. On considère le problème

$$H_0^\ddagger : X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}),$$

contre

$$H_1^\ddagger : X_1, \dots, X_n \overset{\text{iid}}{\sim} \tfrac{1}{2}\mathcal{N}(-\tfrac{1}{2}\Delta\mu, \mathbf{\Sigma}) + \tfrac{1}{2}\mathcal{N}(\tfrac{1}{2}\Delta\mu, \mathbf{\Sigma}),$$
où $\mathbf{\Sigma} - \mathbf{I}$ est de rang 1.

Si $p \gg n$, alors tous les tests sont asymptotiquement impuissants pour

$$\Delta\mu^\top \mathbf{\Sigma}^{-1} \Delta\mu \ll \frac{e^{p/(2n)}}{np} \ .$$

Cas Assymétrique ($\nu \neq 1/2$)

Proposition

- Cas non parcimonieux. *Supposons que* $p \to \infty$ *et que* $p = o(n)$. *Tous les tests sont asymptotiquement impuissants si*

$$\frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu} \ll (p/n)^{1/3} \ .$$

- Cas parcimonieux. *Supposons que* $p/s \to \infty$ *et que* $n \gg s \log(p/s)$. *Tous les tests sont asymptotiquement impuissants si*

$$\limsup \frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu} \leq C \left(\frac{s \log(p/s)}{n} \right)^{1/3} \ .$$

Remarque. Frontière de détection plus grande que pour la covariance connue.

Cas Assymétrique (suite)

Statistique de skewness :

$$\max_{\|u\|_0 \leq s} \frac{\sum_i \left[u^\top (X_i - \bar{X})\right]^3}{\left(\sum_i \left[u^\top (X_i - \bar{X})\right]^2\right)^{3/2}}.$$

Cas Assymétrique (suite)

Statistique de skewness :

$$\max_{\|u\|_0 \leq s} \frac{\sum_i \left[u^\top (X_i - \bar{X})\right]^3}{\left(\sum_i \left[u^\top (X_i - \bar{X})\right]^2\right)^{3/2}}.$$

Statistique signée des moments d'ordre 2 :

$$\max_{\|u\|_0 \leq s} \frac{\sum_i \left[u^\top (X_i - \bar{X})\right]^2 \operatorname{sign}(u^\top (X_i - \bar{X}))}{\sum_i \left[u^\top (X_i - \bar{X})\right]^2} .$$

Proposition

Supposons que $n \gg s \log(p/s)$. *Le test est asymptotiquement puissant si*

$$\liminf \left[\nu(1-\nu)|1-2\nu|\right]^{2/3} \frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu} \left(\frac{s \log(p/s)}{n}\right)^{-1/3} \geq C .$$

Conclusion

- Pour des modèles isotropiques à 2 classes, peu de différences sensibles entre problèmes supervisés et non supervisés...
  .... sauf sous des contraintes de complexité computationnel.

- Pour des modèles de mélange plus complexes (Σ arbitraires, grands k), aucune méthode robuste n'est connue.

- Apports d'un échantillon semi-supervisé ?

Merci pour votre attention !