

Outils statistiques pour la détection de nouveauté

Olivier Cappé
Télécom ParisTech & CNRS

C.E.S. Data Scientist @ Télécom ParisTech
octobre 2014

- Ce cours constitue une *introduction aux méthodes statistiques* utilisées pour la **détection** (de nouveauté, d'anomalie, de changement, de rupture, ...)
- On se placera uniquement dans le cadre simplificateur d'**observations indépendantes dont la distribution peut être modélisée par une loi de probabilité paramétrique**
- A titre d'exemple et du fait de son importance pratique on considérera en particulier la **loi gaussienne multivariée**
- De nombreuses variantes existent : modélisation plus flexible (modèles de mélange, modèles de Markov cachés, ...), méthodes non paramétriques (tests de rang, estimation de densité...), méthodes à noyaux, ...

Test d'adéquation

- Un échantillon X_1, \dots, X_n est-il compatible avec la loi de probabilité P_0 connue qui caractérise le comportement nominal ?

Dans ce contexte, on suppose que P_0 est parfaitement connue ou que les éventuels paramètres de P_0 ont été estimés au préalable avec suffisamment d'observations pour qu'on puisse négliger l'erreur d'estimation

Test à deux échantillons

- Les deux échantillons X_1, \dots, X_{n_x} et Y_1, \dots, Y_{n_y} sont-ils compatibles au sens où il est plausible qu'ils proviennent d'une même loi P_0 , P_0 étant supposée inconnue ?

Cadre proche du précédent mais où n_x et n_y sont comparables, si bien que l'erreur d'estimation ne peut être négligée et que les deux échantillons jouent un rôle symétrique

Détection de changement (ou rupture)

- Etant donné un échantillon X_1, \dots, X_n , peut on trouver une position de changement $\tau \in \{1, \dots, n-1\}$ tel que X_1, \dots, X_τ et $X_{\tau+1}, \dots, X_n$ soient compatibles au sens défini précédemment ?

Cadre souvent rencontré dans les cas où les indices d'observations correspondent à des dates consécutives (on parle alors de *série chronologique*)

Variantes

- Détection séquentielle : n n'est pas fixé a priori et la décision est remise en cause pour chaque valeur de n
- Changement multiples : on cherche à déterminer plusieurs positions de changement τ_1, \dots, τ_K et éventuellement K (le nombre de changements)

- 1 A propos de ce cours
- 2 Éléments de théorie des tests statistiques
- 3 Tests d'adéquation
- 4 Tests à deux échantillons
- 5 Détection de changements

Test Statistique

En statistique, la problématique de détection est liée à celle des *test statistique*

Test statistique

- Une statistique de test $S(X_1, \dots, X_n)$ (à valeur réelle)
- Un seuil t

La région $\mathcal{R} = \{S(X_1, \dots, X_n) > t\}$ définit la **zone de rejet** de l'hypothèse de référence (dite parfois "hypothèse nulle" ou H_0)

- si $S(X_1, \dots, X_n) > t$ on valide l'hypothèse alternative (dite H_1) d'un changement par rapport à la situation de référence

Evaluation de la performance d'un test

Niveau du test

Dit également *taux de faux alarmes*, *taux de faux positifs* ou *probabilité d'erreur de première espèce*

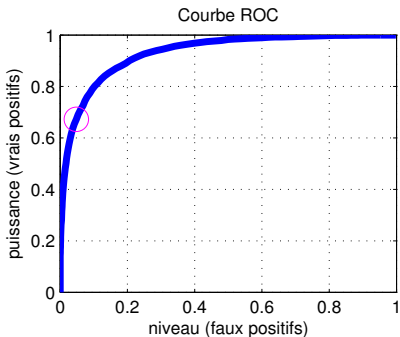
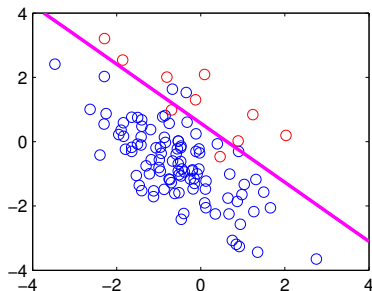
- $P(\mathcal{R})$ sous l'hypothèse de référence

Puissance du test

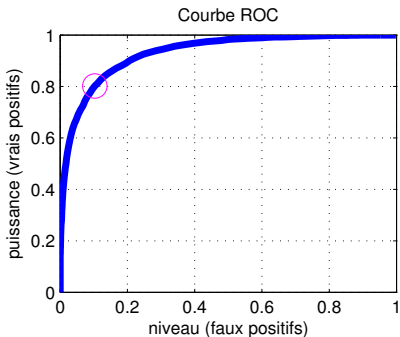
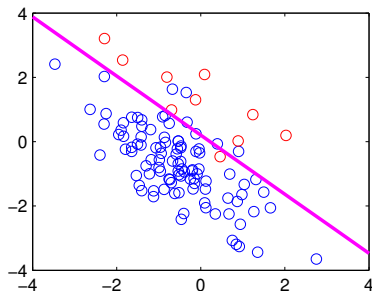
Dit également *taux de vrais positifs*

- $P(\mathcal{R})$ sous l'hypothèse alternative

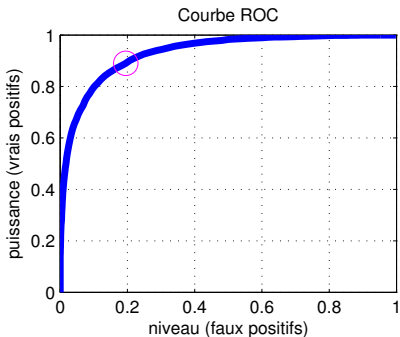
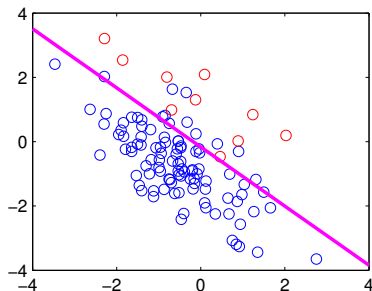
Courbe ROC (Receiver Operating Characteristic)



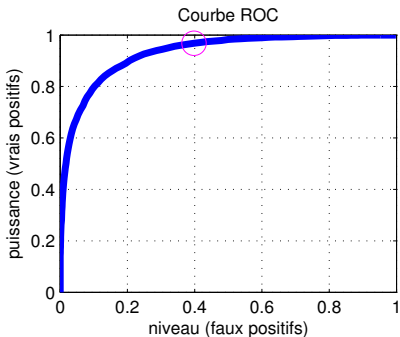
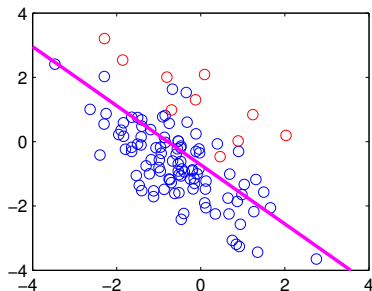
Courbe ROC (Receiver Operating Characteristic)



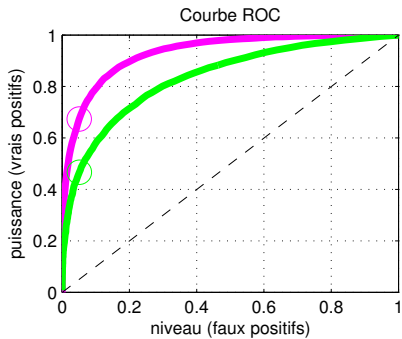
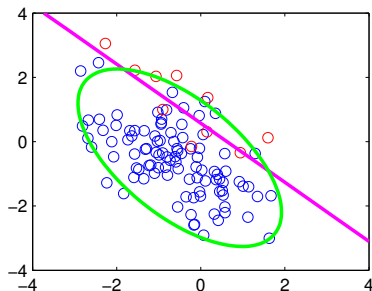
Courbe ROC (Receiver Operating Characteristic)



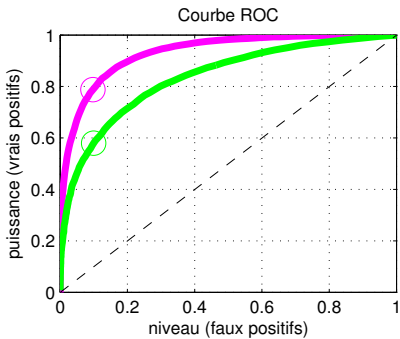
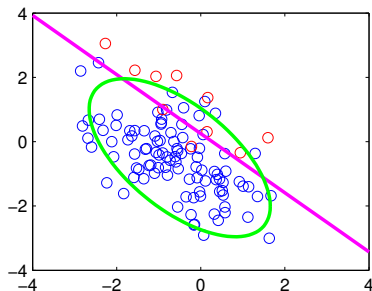
Courbe ROC (Receiver Operating Characteristic)



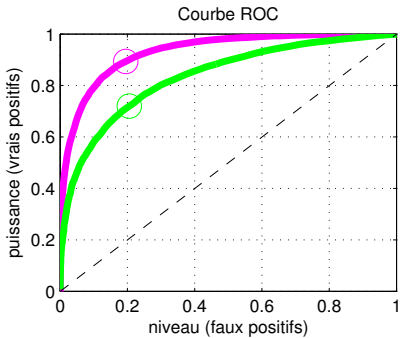
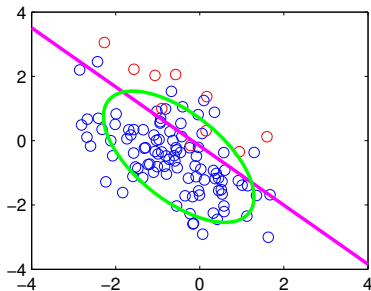
Comparaison de tests de détection



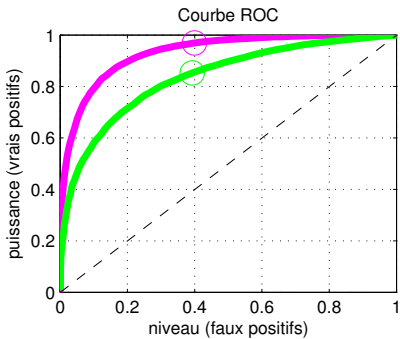
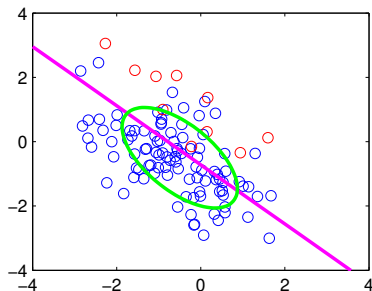
Comparaison de tests de détection



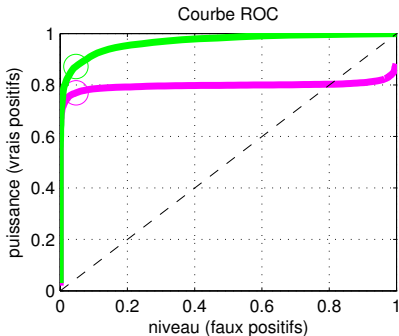
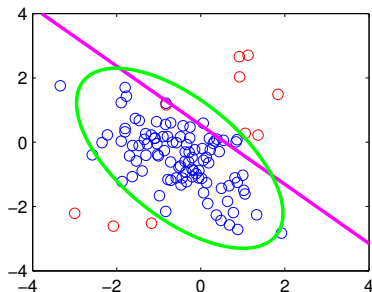
Comparaison de tests de détection



Comparaison de tests de détection

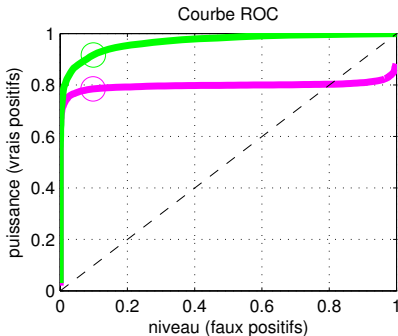
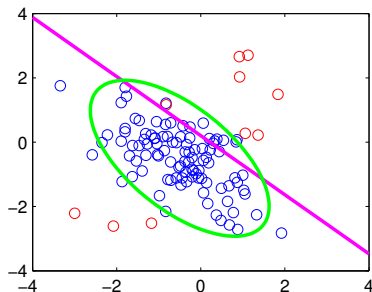


La puissance dépend de la distribution alternative



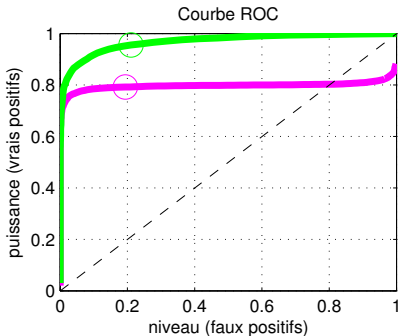
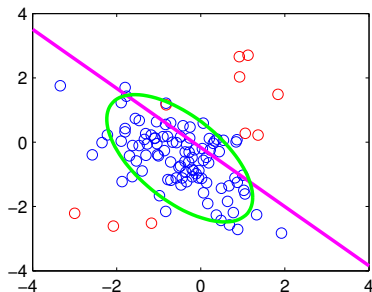
↪ En général, on fixe le seuil t de façon à garantir le niveau du test

La puissance dépend de la distribution alternative



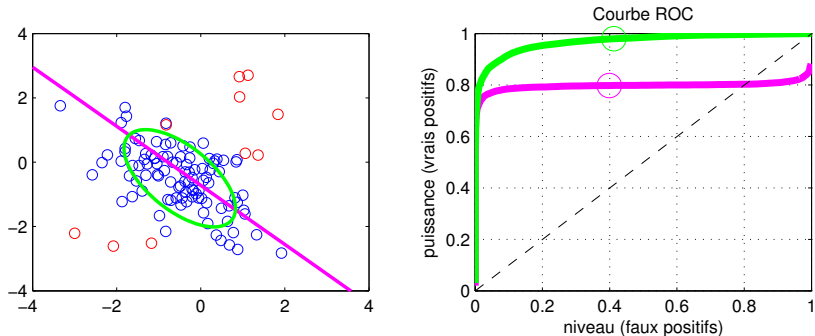
↪ En général, on fixe le seuil t de façon à garantir le niveau du test

La puissance dépend de la distribution alternative



↪ En général, on fixe le seuil t de façon à garantir le niveau du test

La puissance dépend de la distribution alternative



↪ En général, on fixe le seuil t de façon à garantir le niveau du test

Test de rapport de vraisemblance (généralisé)

Lorsqu'on dispose d'un **modèle statistique** des observations

- $\{p_0(x; \theta)\}_{\theta \in \Theta_0}$ sous l'hypothèse de référence H_0
- $\{p_1(x; \theta)\}_{\theta \in \Theta_1}$ sous l'hypothèse de ~~référence~~ H_1 Alternative

Rapport de vraisemblance

La statistique de test du **rapport de vraisemblance** est définie par

$$S_n^{\text{LR}} = \frac{\max_{\theta \in \Theta_1} \prod_{i=1}^n p_1(X_i; \theta)}{\max_{\theta \in \Theta_0} \prod_{i=1}^n p_0(X_i; \theta)}$$

- elle est optimale dans certains cas (par ex., si Θ_0 et Θ_1 sont réduits à une seule valeur de paramètre)
- la loi *asymptotique* de $\log S_n^{\text{LR}}$ est connue (sous des hypothèses assez générales) et peut être utilisée pour fixer le **niveau asymptotique** du test

- 1 A propos de ce cours
- 2 Éléments de théorie des tests statistiques
- 3 Tests d'adéquation**
- 4 Tests à deux échantillons
- 5 Détection de changements

Rappel sur la loi gaussienne multivariée

Densité gaussienne multivariée

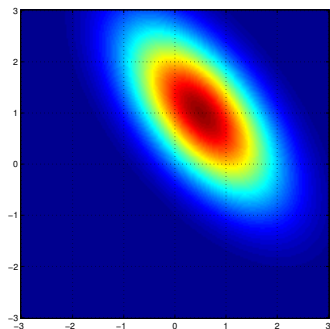
Si $X \in \mathbb{R}^k$ suit une loi gaussienne multivariée non-dégénérée, ce qu'on notera $X \sim \mathcal{N}(\mu, \Sigma)$, sa densité de probabilité est donnée par

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

où

- $\mu \in \mathbb{R}^k$ est l'**espérance** de X
- Σ est la **matrice de covariance** (ou matrice de variances-covariances) de X (matrice $k \times k$ définie positive)

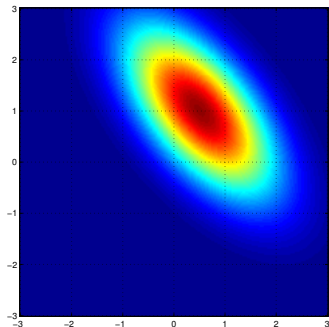
Densité gaussienne en 2D



Densité de

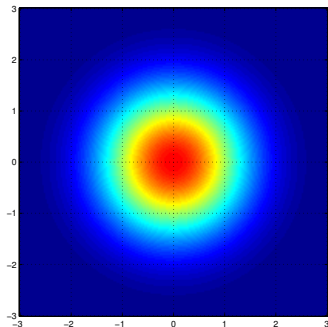
$$\mathcal{N}\left(\begin{bmatrix} 0.5 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1.3 \end{bmatrix}\right)$$

Densité gaussienne en 2D



Densité de

$$\mathcal{N}\left(\begin{bmatrix} 0.5 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1.3 \end{bmatrix}\right)$$



$F(X - \mu)$, où $F^T F = \Sigma^{-1}$ (décomp. de Cholevski de Σ^{-1}) est de loi

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

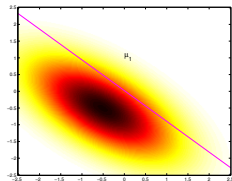
Test d'adéquation à moyennes connues

Modèle statistique

$$H_0 : \mathcal{N}(\mu_0, \Sigma)$$

$$H_1 : \mathcal{N}(\mu_1, \Sigma)$$

où μ_0 et μ_1 sont supposés connus




Test du rapport de vraisemblance à moyennes connues

Le rapport de vraisemblance est équivalent à la statistique

$$\sqrt{n}(\bar{X}_n - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)$$

où $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$

↪ La région de rejet est un hyperplan

 Requiert des connaissances fortes concernant l'hypothèse alternative

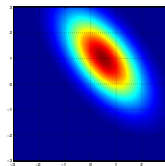
Test d'adéquation à μ_1 inconnue

Modèle statistique

$$H_0 : \mathcal{N}(\mu_0, \Sigma)$$

$$H_1 : \mathcal{N}(\mu_1, \Sigma)$$

où μ_1 est inconnue



"Distance" de Mahalanobis

Le rapport de vraisemblance est équivalent à la statistique

$$n(\bar{X}_n - \mu_0)^T \Sigma^{-1} (\bar{X}_n - \mu_0)$$

↪ La région de rejet est le complémentaire d'une ellipse



On peut fixer le niveau du test en utilisant une loi du khi-deux à k degrés de liberté

↪ $n(\bar{X}_n - \mu_0)^T \Sigma^{-1} (\bar{X}_n - \mu_0)$ suit la loi χ_k^2 sous H_0

Loi du khi-deux

Définition χ_k^2 est la loi de $X^T X = \|X\|^2 = \sum_{j=1}^k X^2(j)$ lorsque $X \sim \mathcal{N}(0, \text{Id}_k)$

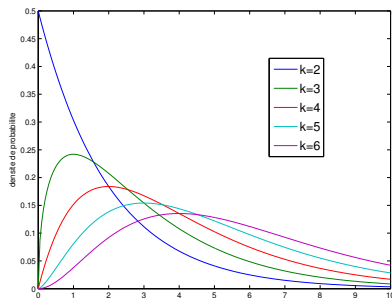


Figure: Densité de probabilité de la loi χ_k^2 pour $k=2, \dots, 6$

Loi du khi-deux

Définition χ_k^2 est la loi de $X^T X = \|X\|^2 = \sum_{j=1}^k X^2(j)$ lorsque $X \sim \mathcal{N}(0, \text{Id}_k)$

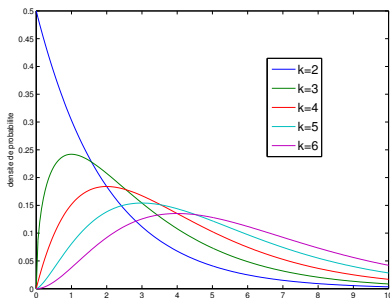


Figure: Densité de probabilité de la loi χ_k^2 pour $k=2, \dots, 6$

Propriété C'est aussi la loi de $(X - \mu)^T \Sigma^{-1} (X - \mu)$ quand $X \sim \mathcal{N}(\mu, \Sigma)$

Preuve : $F(X - \mu) \sim \mathcal{N}(0, \text{Id}_k)$ où $F^T F = \Sigma^{-1}$



Détermination du seuil du test

Si $S_n \sim \chi_k^2$ sous H_0 , on fixe le seuil t de façon à ce que

$$P(Z > t) = \alpha \quad \text{pour} \quad Z \sim \chi_k^2$$

où α est le niveau souhaité du test

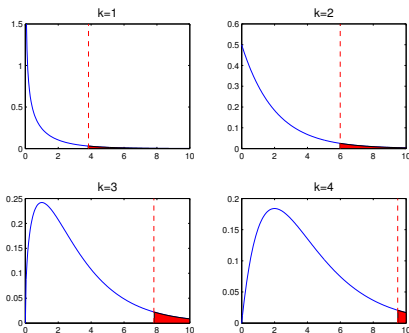


Figure: Seuil de niveau 5% pour différentes valeurs de k

Cas d'observations discrètes

Le cas d'observations catégorielles est très important (traitement du langage naturel, traitement d'images, etc.)

Dans le cas où l'observation X_i correspondent à une catégorie discrète parmi k on définit \bar{X}_n comme le **vecteur des fréquences empiriques des différentes catégories**

$$\bar{X}_n = \frac{1}{n} \begin{pmatrix} \text{nombre d'observations de la catégorie 1} \\ \vdots \\ \text{nombre d'observations de la catégorie k} \end{pmatrix}$$

Ici il n'est pas approprié de supposer que les X_i sont gaussiennes ($\bar{X}_n \geq 0$ et $\sum_{j=1}^k \bar{X}_n(j) = 1$)

Cas d'observations discrètes (suite)

Test de Pearson (dit également du khi-deux)

Sous l'hypothèse H_0 que les probabilités des k catégories sont données par p_1, \dots, p_k , le test du rapport de vraisemblance est asymptotiquement équivalent à

$$n \left(\bar{X}_n - \begin{bmatrix} p_1 \\ \vdots \\ p_k \end{bmatrix} \right)^T \begin{pmatrix} p_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & p_k \end{pmatrix}^{-1} \left(\bar{X}_n - \begin{bmatrix} p_1 \\ \vdots \\ p_k \end{bmatrix} \right) = n \sum_{j=1}^k \frac{(\bar{X}_n(j) - p_j)^2}{p_j}$$



On peut fixer le niveau asymptotique du test en utilisant une loi du khi-deux à $k-1$ degrés de liberté

- 1 A propos de ce cours
- 2 Éléments de théorie des tests statistiques
- 3 Tests d'adéquation
- 4 Tests à deux échantillons**
- 5 Détection de changements

On dispose dorénavant de deux échantillons de tailles comparables

- X_1, \dots, X_{n_x}

- Y_1, \dots, Y_{n_y}

→ on souhaite tester leur compatibilité de façon symétrique

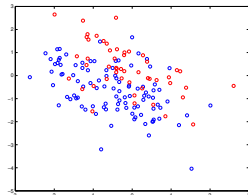
Modèle statistique

$$H_0 : X_i, Y_i \sim \mathcal{N}(\mu, \Sigma)$$

$$H_1 : X_i \sim \mathcal{N}(\mu_x, \Sigma) \quad Y_i \sim \mathcal{N}(\mu_y, \Sigma)$$

$$\text{avec } \mu_x \neq \mu_y$$

où μ, μ_x, μ_y sont supposés inconnus



Test gaussien à deux échantillons

La statistique du rapport de vraisemblance est donnée par

$$\frac{n_x n_y}{n_x + n_y} (\bar{X}_{n_x} - \bar{Y}_{n_y})^T \Sigma^{-1} (\bar{X}_{n_x} - \bar{Y}_{n_y})$$

où $\bar{X}_{n_x} = \frac{1}{n} \sum_{i=1}^{n_x} X_i$, $\bar{Y}_{n_y} = \frac{1}{n} \sum_{i=1}^{n_y} Y_i$



Sous H_0 elle suit une loi du khi-deux à k degrés de liberté

Si Σ est inconnu on peut la remplacer par un estimateur comme

$$\hat{\Sigma}_n = \frac{\sum_{i=1}^{n_x} (X_i - \bar{X}_{n_x})(X_i - \bar{X}_{n_x})^T + \sum_{i=1}^{n_y} (Y_i - \bar{Y}_{n_y})(Y_i - \bar{Y}_{n_y})^T}{n_x + n_y}$$

(Test d'Hotelling)

- 1 A propos de ce cours
- 2 Éléments de théorie des tests statistiques
- 3 Tests d'adéquation
- 4 Tests à deux échantillons
- 5 Détection de changements**

Modèle statistique

$$H_0: X_1, \dots, X_n \sim \mathcal{N}(\mu, \Sigma)$$

$$H_1: X_1, \dots, X_\tau \sim \mathcal{N}(\mu_1, \Sigma)$$

$$\text{et } X_{\tau+1}, \dots, X_n \sim \mathcal{N}(\mu_2, \Sigma) \text{ avec } \mu_1 \neq \mu_2$$

où μ, μ_1, μ_2 et τ sont inconnus.

Test de détection à un changement

$$\max_{\tau \in \{1, \dots, n-1\}} \underbrace{\frac{\tau(n-\tau)}{n} (\bar{X}_1(\tau) - \bar{X}_2(\tau))^T \Sigma^{-1} (\bar{X}_1(\tau) - \bar{X}_2(\tau))}_{S_n(\tau)}$$

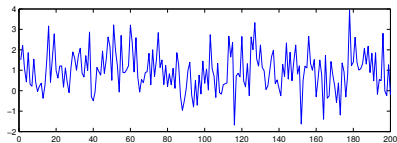
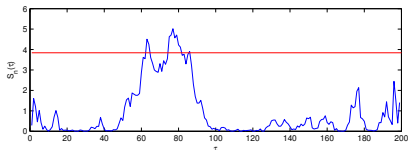
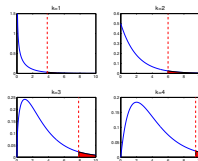
où

$$\begin{cases} \bar{X}_1(\tau) &= \frac{1}{\tau} \sum_{i=1}^{\tau} X_i \\ \bar{X}_2(\tau) &= \frac{1}{n-\tau} \sum_{i=\tau+1}^n X_i \end{cases}$$

↪ Si le test est significatif l'argmax donne une estimation de la position du changement



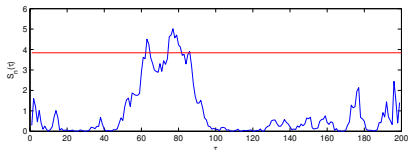
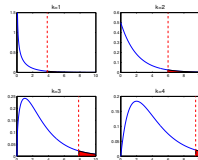
Le fait d'optimiser sur la position de changement τ implique de *relever le seuil de détection*, par rapport au test à τ connu



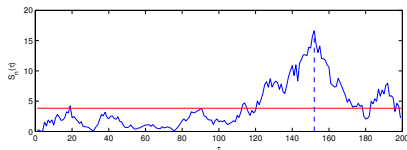
en l'absence de changement



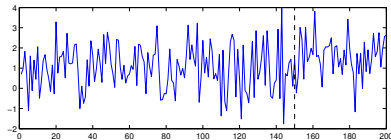
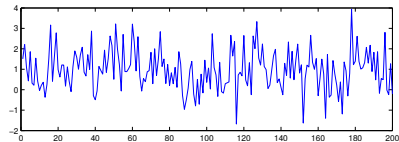
Le fait d'optimiser sur la position de changement τ implique de *relever le seuil de détection*, par rapport au test à τ connu



en l'absence de changement



avec un changement d'amplitude 0.5σ à $t = 150$



Extensions

- Il existe une théorie permettant de fixer le seuil de façon plus précise dans le modèle à un changement
- Il existe un algorithme **de programmation dynamique** efficace (de complexité quadratique en n) permettant de rechercher la position de plus de un changement

- 1 A propos de ce cours
- 2 Éléments de théorie des tests statistiques
- 3 Tests d'adéquation
- 4 Tests à deux échantillons
- 5 Détection de changements