Statistique Mathématique 2 (MATH-F-309, Chapitre #3)

Thomas Verdebout

Université Libre de Bruxelles

2015/2016

Plan du cours

1. Vecteurs aléatoires.
2. Loi normale multivariée.
3. Inférence dans les modèles gaussiens.
4. Méthodes classiques de l'analyse multivariée.
5. Données directionelles.
6. Modèle linéaire.

3. Inférence dans les modèles gaussiens.

3.1. Sur le paramètre de position.

3.1.1. Estimateurs MLE.

3.1.2. Tests de Hotelling.

3.1.3. Zones de confiance.

3.1.4. Problèmes à plusieurs échantillons.

3.2. Sur le paramètre de dispersion.

3.3. Autres types de problèmes.

MLE

Le résultat suivant donne les estimateurs du maximum de vraisemblance de μ et Σ pour un échantillon gaussien p-varié.

Théorème: *soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}_p(\mu, \Sigma)$. Alors les estimateurs du maximum de vraisemblance de μ et de Σ sont respectivement*

$$\hat{\mu} = \bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{et} \quad \hat{\Sigma} := \frac{1}{n} W := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'.$$

Preuve: la vraisemblance de cet échantillon est donnée par

$$L_{\mu,\Sigma}^{(n)} = \prod_{i=1}^{n}\left[\left(\frac{1}{2\pi}\right)^{\frac{p}{2}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X_i - \mu)'\Sigma^{-1}(X_i - \mu)\right)\right],$$

de sorte que la log-vraisemblance est

$$\log L_{\mu,\Sigma}^{(n)} = C - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\left[(X_i - \mu)'\Sigma^{-1}(X_i - \mu)\right].$$

MLE

En décomposant $X_i - \mu$ en $(X_i - \bar{X}) + (\bar{X} - \mu)$, on obtient

$$\sum_{i=1}^{n} \Big[(X_i - \mu)' \Sigma^{-1} (X_i - \mu) \Big]$$

$$= \sum_{i=1}^{n} \Big[(X_i - \bar{X})' \Sigma^{-1} (X_i - \bar{X}) \Big] + n(\bar{X} - \mu)' \Sigma^{-1} (\bar{X} - \mu)$$

ce qui livre

$$\log L_{\mu,\Sigma}^{(n)} = C - \frac{n}{2} \log |\Sigma|$$

$$- \frac{1}{2} \sum_{i=1}^{n} \Big[(X_i - \bar{X})' \Sigma^{-1} (X_i - \bar{X}) \Big] - \frac{n}{2} (\bar{X} - \mu)' \Sigma^{-1} (\bar{X} - \mu).$$

Puisque Σ (et donc Σ^{-1}) est définie-positive, on en déduit que, pour toute valeur fixée de Σ,

$$\arg\max_{\mu} \log L_{\mu,\Sigma}^{(n)} = \arg\min_{\mu} (\bar{X} - \mu)' \Sigma^{-1} (\bar{X} - \mu) = \bar{X}.$$

MLE

Il ne reste donc qu'à maximiser, en Σ, la quantité

$$\log L_{\bar{X},\Sigma}^{(n)} = C - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\Big[(X_i - \bar{X})'\Sigma^{-1}(X_i - \bar{X})\Big].$$

Pour ce faire, remarquons que

$$\begin{aligned}
&\sum_{i=1}^{n}\left[(X_i - \bar{X})'\Sigma^{-1}(X_i - \bar{X})\right] \\
&\quad = \sum_{i=1}^{n}\mathrm{tr}\left[(X_i - \bar{X})'\Sigma^{-1}(X_i - \bar{X})\right] \\
&\quad = \sum_{i=1}^{n}\mathrm{tr}\left[\Sigma^{-1}(X_i - \bar{X})(X_i - \bar{X})'\right] \\
&\quad = \mathrm{tr}\left[\Sigma^{-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'\right] \\
&\quad = \mathrm{tr}\left[\Sigma^{-1}W\right].
\end{aligned}$$

MLE

Donc

$$\begin{aligned}\log L_{\bar{X},\Sigma}^{(n)} &= \tilde{C} + \frac{n}{2}\log|\Sigma^{-1}(W/n)| - \frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}W\right] \\ &= \tilde{C} + \frac{n}{2}\left[\log|\Sigma^{-1}(W/n)| - \operatorname{tr}\left[\Sigma^{-1}(W/n)\right]\right]\end{aligned}$$

pour une certaine quantité $\tilde{C}$ qui ne dépend pas de Σ.

Comme, en écrivant $W = W^{1/2}(W^{1/2})'$, on a

$$\begin{aligned}\hat{\Sigma} &= \arg\max_{\Sigma} \log L_{\bar{X},\Sigma}^{(n)} = \arg\max_{\Sigma}\left[\log|\Sigma^{-1}(W/n)| - \operatorname{tr}\left[\Sigma^{-1}(W/n)\right]\right] \\ &= \arg\max_{\Sigma}\left[\log|(W^{1/2})'\Sigma^{-1}W^{1/2}/n| - \operatorname{tr}\left[(W^{1/2})'\Sigma^{-1}W^{1/2}/n\right]\right],\end{aligned}$$

le résultat suivant permet de conclure (puisqu'il montre que $\hat{\Sigma}$ est tel que $(W^{1/2})'\hat{\Sigma}^{-1}W^{1/2}/n = I_p$, ce qui livre $\hat{\Sigma} = W/n$).

Lemme: *soit $\mathcal{S}$ la collection des matrices $(p \times p)$ symétriques et définies positives. Alors*

$$\arg\max_{T \in \mathcal{S}} \Big[\log|T| - \operatorname{tr} T \Big] = I_p.$$

Preuve du lemme: décomposons T en $T = O\Lambda O'$, où O est orthogonale et Λ est diagonale (notons $\lambda_i := \Lambda_{ii} > 0$). Alors

$$\log|T| - \operatorname{tr} T = \log|O\Lambda O'| - \operatorname{tr}\Big[O\Lambda O'\Big] = \log(|O||\Lambda||O'|) - \operatorname{tr}\Big[\Lambda O' O\Big]$$

$$= \log|\Lambda| - \operatorname{tr}\Lambda = \log\Big(\prod_{i=1}^{p} \lambda_i\Big) - \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \Big[\log\lambda_i - \lambda_i\Big].$$

Comme $\arg\max_{x>0}(\log x - x) = 1$, on en déduit que le maximum en T de $\log|T| - \operatorname{tr} T$ est atteint pour $\lambda_1 = \ldots = \lambda_p = 1$, c'est-à-dire en $T = OI_pO' = I_p$. □

MLE

Le résultat suivant donne les estimateurs du maximum de vraisemblance de μ et Σ pour un échantillon gaussien p-varié.

Théorème: *soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}_p(\mu, \Sigma)$. Alors les estimateurs du maximum de vraisemblance de μ et de Σ sont respectivement*

$$\hat{\mu} = \bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{et} \quad \hat{\Sigma} := \frac{1}{n} W := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'.$$

Remarques:

- $\hat{\mu}$ est sans biais pour μ; par contre, $\hat{\Sigma}$ est seulement asymptotiquement non biaisé ($\mathrm{E}[\hat{\Sigma}] = \mathrm{E}[\frac{n-1}{n} S] = \frac{n-1}{n}\Sigma$).
- Tout ceci est similaire à ce qui se passe dans le cas univarié ($p = 1$). En particulier, $\hat{\mu} = \bar{X}$ est convergent, normal, UMVU, affine-équivariant, etc.

3. Inférence dans les modèles gaussiens.

3.1. Sur le paramètre de position.

3.1.1. Estimateurs MLE.

3.1.2. Tests de Hotelling.

3.1.3. Zones de confiance.

3.1.4. Problèmes à plusieurs échantillons.

3.2. Sur le paramètre de dispersion.

3.3. Autres types de problèmes.

Tests de Hotelling (Σ connu)

Soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}_p(\mu, \Sigma)$.

Soit μ_0 un p-vecteur fixé. Considérons le problème de test

$$\begin{cases} \mathcal{H}_0 : \mu = \mu_0 \\ \mathcal{H}_1 : \mu \neq \mu_0. \end{cases}$$

Comme pour $p = 1$, il est naturel de baser la règle de décision sur $\bar{X}$ (et plus spécifiquement sur la distance entre $\bar{X}$ et μ_0).

Puisque $\bar{X} \sim \mathcal{N}_p(\mu_0, \frac{1}{n}\Sigma)$ sous $\mathcal{H}_0$, on a que, sous $\mathcal{H}_0$,

$$T_c^2(X) = n(\bar{X} - \mu_0)'\Sigma^{-1}(\bar{X} - \mu_0) = d^2_{\frac{1}{n}\Sigma}(\bar{X}, \mu_0) \sim \chi^2_p.$$

On rejete $\mathcal{H}_0$ pour de grandes valeurs de $T_c^2(X)$.

Il en découle qu'au niveau α, un test convenable est le test ϕ qui consiste à rejeter $\mathcal{H}_0$ ssi $T_c^2(X) > \chi^2_{p;1-\alpha}$.

Tests de Hotelling asymptotique

Bien entendu, ceci requiert que Σ soit connu.

Si Σ est inconnu, il est naturel de remplacer Σ par $\hat{\Sigma} = S$...
$\rightsquigarrow T^2(X) = n(\bar{X} - \mu_0)' S^{-1}(\bar{X} - \mu_0)$ (notation usuelle: T^2).

En utilisant le lemme de Slutzky, on obtient que, sous $\mathcal{H}_0$,

$$T^2 = n(\bar{X} - \mu_0)' \Sigma^{-1}(\bar{X} - \mu_0) + o_P(1) \xrightarrow{\mathcal{L}} \chi^2_p.$$

Donc, un test asymptotique (au niveau asymptotique α) consiste à rejeter $\mathcal{H}_0$ ssi $T^2 > \chi^2_{p;1-\alpha}$.

Remarque: il découle du TCL multivarié que ce test ne requiert pas que la loi commune des X_i soit normale, mais seulement que celle-ci ait des moments finis d'ordre 2.

Test de Student

Pour $p = 1$, cette statistique est simplement

$$T^2 = \left| \frac{\sqrt{n}\,(\bar{X} - \mu_0)}{s} \right|^2, \text{ où } s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

ce qui est le carré de la statistique de Student usuelle.

Si $X_1, \ldots, X_n$ sont i.i.d. $\mathcal{N}_1(\mu_0, \sigma^2)$, le lemme de Fisher implique que

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \sim t_{n-1},$$

de sorte que la loi exacte de T^2 sous $\mathcal{H}_0$ (pour $p = 1$) est $F_{1,n-1}$.

Un test exact (au niveau α) consiste donc à rejeter $\mathcal{H}_0$ ssi $T^2 > F_{1,n-1;1-\alpha}$ (c'est le test de Student usuel).

Remarque: ce test exact, contrairement au précédent, requiert clairement la normalité des X_i.

Tests de Hotelling (loi exacte)

Une question naturelle est:

Pour $p > 1$, quelle est la loi exacte (sous $\mathcal{H}_0$) de la statistique de test

$$T^2 = n(\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0),$$

si les X_i sont i.i.d. de loi normale p-variée ?

Le lemme suivant permet de répondre à cette question:

Lemme: *soient* $Y \sim \mathcal{N}_p(0, \Sigma)$ *et* $V \sim W_p(m, \Sigma)$. *Alors, si* $m \geq p$ *et* $Y \perp\!\!\!\perp V$,

$$\frac{m - p + 1}{p} \, Y' V^{-1} Y \sim F_{p, m-p+1}.$$

Tests de Hotelling (loi exacte)

Soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}_p(\mu, \Sigma)$, où $n \geq p+1$.

En utilisant le lemme de Fisher multivarié, il découle de ce lemme que, sous $\mathcal{H}_0 : \mu = \mu_0$,

$$\frac{n-p}{p(n-1)} T^2 \sim F_{p,n-p}.$$

Un test exact (au niveau α) consiste donc à rejeter $\mathcal{H}_0$ ssi $\frac{n-p}{p(n-1)} T^2 > F_{p,n-p;1-\alpha}$.

Remarque: la version asymptotique de ce test est bien le test asymptotique vu précédemment.

Ce test, qui est appelé test de Hotelling, étend donc au cas multivarié le test de Student usuel.

Tests de Hotelling

Preuve du lemme: comme d'habitude nous supposons que $\Sigma > 0$. Alors $Y'V^{-1}V = (Y^*)'(V^*)^{-1}Y^*$ où $Y^* = \Sigma^{-1/2}Y$ et $V^* = \Sigma^{-1/2}V\Sigma^{-1/2}$. Donc on peut supposer que $Y \sim \mathcal{N}_p(0, I_p)$ et $V \sim W_p(m, I_p) = W_p(m)$.

On peut montrer (c'est un peu délicat), que

$$\frac{a'V^{-1}a}{a'a} \sim \frac{1}{\chi^2_{m-p+1}} \quad \forall a \in \mathbb{R}^p,\ a \neq 0.$$

(Par contre, il est facile de montrer que $\frac{a'Va}{a'a} \sim \chi^2_m \quad \forall a \in \mathbb{R}^p,\ a \neq 0.$)

Nous écrivons

$$Y'V^{-1}Y = \frac{Y'V^{-1}Y}{Y'Y} \times Y'Y = A(Y, V) \times B(Y).$$

Tests de Hotelling

On note F la fonction de répartition de Y. Alors par indépendence de V et Y

$$
\begin{aligned}
& P(A(Y,V) \leq x \cap B(Y) \leq y) \\
&= \int_{\mathbb{R}^p} P(A(h,V) \leq x \cap B(h) \leq y) dF(h) \\
&= \int_{\mathbb{R}^p} P(A(h,V) \leq x) I\{B(h) \leq y\} dF(h) \\
&= P\left(\frac{1}{\chi^2_{m-p+1}} \leq x \right) \int_{\mathbb{R}^p} I\{B(h) \leq y\} dF(h) \\
&= P\left(\frac{1}{\chi^2_{m-p+1}} \leq x \right) \underbrace{P(B(Y) \leq y)}_{P(\chi^2_p \leq y)}.
\end{aligned}
$$

Tests de Hotelling

Il découle des calculs précédants, que

$$
\begin{aligned}
Y' V^{-1} Y &\overset{\mathcal{D}}{=} \frac{\chi^2_p}{\chi^2_{m-p+1}} \\
&= \frac{\chi^2_p / p}{\chi^2_{m-p+1}/(m-p+1)} \frac{p}{m-p+1} \\
&= \frac{p}{m-p+1} F_{p,p-m+1}.
\end{aligned}
$$

□

Tests de Hotelling

Soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}_p(\mu, \Sigma)$.

Le test de Hotelling, pour $\begin{cases} \mathcal{H}_0 : \mu = \mu_0 \\ \mathcal{H}_1 : \mu \neq \mu_0, \end{cases}$

consiste (au niveau α) à rejeter $\mathcal{H}_0$ ssi

$$\frac{n-p}{p(n-1)} T^2 = \frac{(n-p)n}{p(n-1)} (\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0) > F_{p,n-p;1-\alpha}.$$

Quelles sont les propriétés de ce test?

$\rightsquigarrow$ **Théorème**: *le test de Hotelling coïncide avec le test du rapport de vraisemblance (gaussien)* .

Test du rapport de vraisemblance

Preuve: pour rappel, pour le problème de test $\mathcal{H}_0 : \theta \in \Theta_0$ contre $\mathcal{H}_1 : \theta \in \Theta \backslash \Theta_0$, la statistique du test du rapport de vraisemblance est

$$\Lambda^{(n)} = \frac{L_{\tilde{\theta}}}{L_{\hat{\theta}}},$$

où $\tilde{\theta} := \arg\max_{\theta \in \Theta_0} L_\theta$ et $\hat{\theta} := \arg\max_{\theta \in \Theta} L_\theta$ sont respectivement les estimateurs de maximum de vraisemblance contraint et non contraint pour θ.

Et le test associé consiste à rejeter $\mathcal{H}_0 : \theta \in \Theta_0$ (au niveau asymptotique α) ssi

$$-2 \ln \Lambda^{(n)} > \chi^2_{k-k_0;1-\alpha},$$

où k et k_0 sont respectivement les nombres de paramètres libres dans Θ et Θ_0.

Test du rapport de vraisemblance

Ici, $\theta = (\mu, \Sigma)$, $\Theta = \mathbb{R}^p \times \mathcal{V}_p$, où $\mathcal{V}_p$ désigne la collection des matrices $p \times p$ symétriques et définies positives. Et $\Theta_0 = \{\mu_0\} \times \mathcal{V}_p$. $\rightsquigarrow k = p + p(p+1)/2$ et $k_0 = p(p+1)/2$.

Comme on l'a vu, $\hat{\theta} = (\hat{\mu}, \hat{\Sigma}) = (\bar{X}, W/n)$.

Que vaut $\tilde{\theta} = (\tilde{\mu}, \tilde{\Sigma})$? Clairement, $\tilde{\mu} = \mu_0$. Et en utilisant les mêmes arguments que lors du calcul de l'estimateur de maximum de vraisemblance de Σ, on montre que

$$\tilde{\Sigma} := W_0/n, \text{ où } W_0 := \sum_{i=1}^{n} (X_i - \mu_0)(X_i - \mu_0)'.$$

Donc

$$\Lambda^{(n)} = \frac{L_{\tilde{\theta}}}{L_{\hat{\theta}}} = \frac{L_{\mu_0, W_0/n}}{L_{\bar{X}, W/n}}.$$

Test du rapport de vraisemblance

Ceci livre

$$\Lambda^{(n)} = \frac{(2\pi)^{-np/2}|W_0/n|^{-n/2}\exp[-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu_0)'(W_0/n)^{-1}(X_i-\mu_0)]}{(2\pi)^{-np/2}|W/n|^{-n/2}\exp[-\frac{1}{2}\sum_{i=1}^{n}(X_i-\bar{X})'(W/n)^{-1}(X_i-\bar{X})]}.$$

Comme

$$\sum_{i=1}^{n}(X_i-\bar{X})'(W/n)^{-1}(X_i-\bar{X}) = \mathrm{tr}[(W/n)^{-1}W] = \mathrm{tr}[n\,I_p] = np$$

et

$$\sum_{i=1}^{n}(X_i-\mu_0)'(W_0/n)^{-1}(X_i-\mu_0) = \mathrm{tr}[(W_0/n)^{-1}W_0] = \mathrm{tr}[n\,I_p] = np,$$

on obtient que

Test du rapport de vraisemblance

$$\begin{aligned}\Lambda^{(n)} &= \frac{|W_0/n|^{-n/2}}{|W/n|^{-n/2}} \\ &= |W_0 W^{-1}|^{-n/2} \\ &= |(W + n(\bar{X} - \mu_0)(\bar{X} - \mu_0)')W^{-1}|^{-n/2},\end{aligned}$$

où on a obtenu $W_0 = W + n(\bar{X} - \mu_0)(\bar{X} - \mu_0)'$ en décomposant $X_i - \mu_0$ en $(X_i - \bar{X}) + (\bar{X} - \mu_0)$.

Le lemme suivant est très utile:

Lemma

Soit $C \in \mathbb{R}^{p \times p}$ avec $|C| > 0$. Alors pour tout $y \in \mathbb{R}^p$

$$|C + yy'| = |C|(1 + y' C^{-1} y).$$

Test du rapport de vraisemblance

En utilisant le lemme, on voit que

$$\begin{aligned}\Lambda^{(n)} &= |1 + n(\bar{X} - \mu_0)' W^{-1} (\bar{X} - \mu_0)|^{-n/2} \\ &= (1 + (n-1)^{-1} n(\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0))^{-n/2} \\ &= (1 + (n-1)^{-1} T^2)^{-n/2}.\end{aligned}$$

Les statistiques $\Lambda^{(n)}$ et T^2 sont donc en bijection.
Par conséquent, les tests associés sont équivalents. □

Test du rapport de vraisemblance

Remarque:

comme nous l'avons rappelé, le test de rapport du vraisemblance associé consiste à rejeter $\mathcal{H}_0 : \theta \in \Theta_0$ (au niveau asymptotique α) ssi

$$-2 \ln \Lambda^{(n)} > \chi^2_{k-k_0;1-\alpha},$$

c'est-à-dire, dans ce cas, ssi (pour n grand)

$$\begin{aligned} & -2 \ln \Lambda^{(n)} \\ & \quad -2 \ln \left((1 + (n-1)^{-1} T^2)^{-n/2} \right) \\ & \quad = n \ln(1 + (n-1)^{-1} T^2) \\ & \quad \approx T^2 > \chi^2_{p;1-\alpha}, \end{aligned}$$

ce qui n'est rien d'autre que la version asymptotique du test de Hotelling.

Tests de Hotelling

Autres propriétés du test de Hotelling:

- pour $\mathcal{H}_0 : \mu = 0$, la statistique de test T^2 (et par suite, le test lui-même) est invariante par transformations linéaires , ce qui signifie que $T^2(AX_1, \ldots, AX_n) = T^2(X_1, \ldots, X_n)$ pour toute matrice A $(p \times p)$ inversible (interprétation!)
- Cette invariance explique le fait que la loi de T^2 sous $\mathcal{H}_0$ ne dépende pas de Σ...
- Par contre, il n'y a pas invariance par rapport au groupe des translations ($T^2(X_1 + b, \ldots, X_n + b) = T^2(X_1, \ldots, X_n)$ pour tout p-vecteur b). Heureusement! (commenter).
- Le test de Hotelling est UMPI ("uniformly most poweful invariant"), c'est-à-dire que, pour tout test ϕ de niveau α et invariant par transformations linéaires, la puissance du test de Hotelling est supérieure à celle de ϕ en tout $\mu(\neq \mu_0)$.

Zones de confiance

Les résultats distributionnels de la section précédente permettent de construire des zones de confiance pour μ.

Definition (Zones de confiance)

On appelle un ensemble $C_{1-\alpha}^{(n)} = C_{1-\alpha}(X_1, \ldots, X_n) \subset \mathbb{R}^p$ un zone de confiance pour un parametre θ au niveau $(1-\alpha) \times 100\%$, si

$$P(C_{1-\alpha}^{(n)} \text{ contient } \theta) = 1 - \alpha.$$

En effet, si $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}_p(\mu, \Sigma)$, on a vue

$$P\left[n(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu) \leq \frac{p(n-1)}{n-p} F_{p,n-p;1-\alpha} \right] = 1 - \alpha.$$

Zones de confiance

Par conséquent une zone de confiance (au niveau de confiance $(1-\alpha)\times 100\%$) est donnée par l'ellipsoïde:

$$C_{1-\alpha}^{(n)} := \left\{ \mu \in \mathbb{R}^p \,\middle|\, T^2(\mu) \leq \frac{p(n-1)}{n-p} F_{p,n-p;1-\alpha} \right\}$$
$$= \left\{ \mu \in \mathbb{R}^p \,\middle|\, d_S^2(\bar{X},\mu) \leq \frac{p(n-1)}{n(n-p)} F_{p,n-p;1-\alpha} \right\}.$$

Zones de confiance

De même, le fait que

$$P\left[T^2(\mu) \leq \chi^2_{p;1-\alpha}\right] \to 1-\alpha, \text{ si } n \to \infty,$$

implique, qu'une zone de confiance (au niveau de confiance asymptotique $(1-\alpha)\times 100\%$) est donnée par l'ellipsoïde

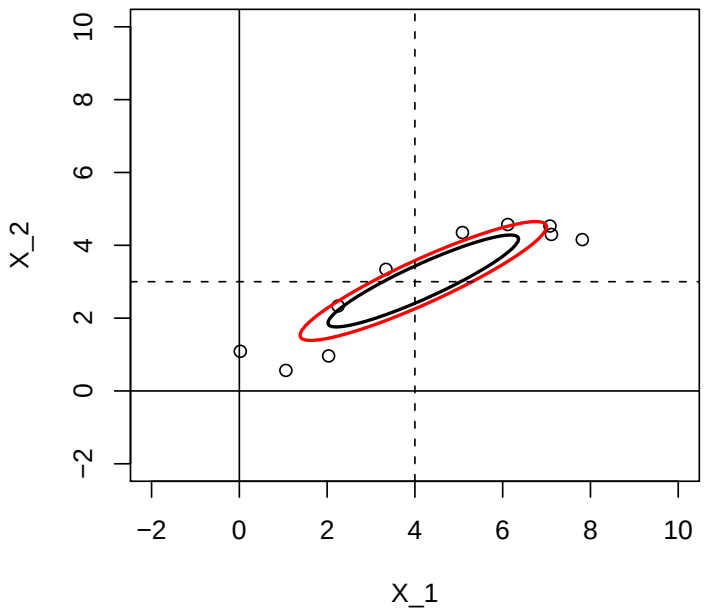
$$C^{(\infty)}_{1-\alpha} := \left\{\mu \in \mathbb{R}^p \;\middle|\; T^2(\mu) \leq \chi^2_{p;1-\alpha}\right\}$$

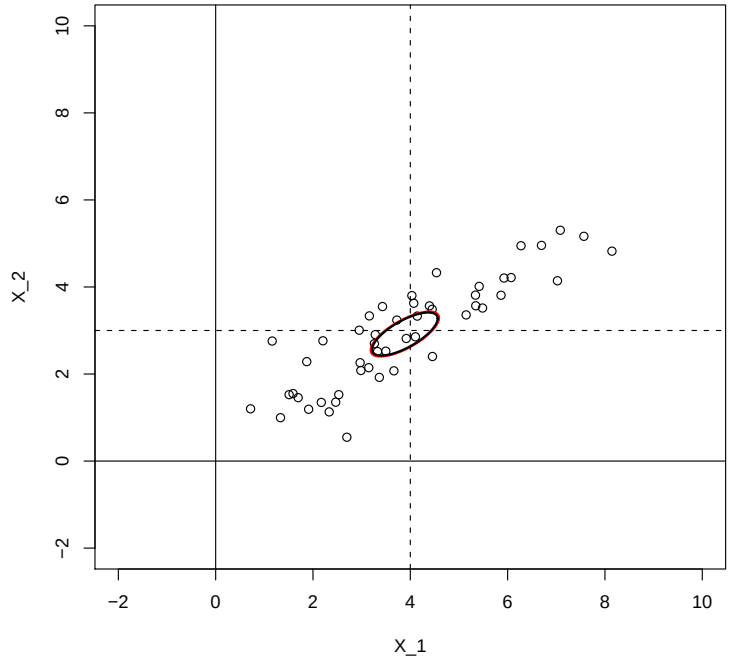
$$= \left\{\mu \in \mathbb{R}^p \;\middle|\; d^2_S(\bar{X}, \mu) \leq \frac{1}{n}\chi^2_{p;1-\alpha}\right\}.$$

Remarque: tout comme le test de Hotelling asymptotique, cette procédure ne requiert pas la normalité des X_i, mais seulement l'existence de moments finis d'ordre 2.

Exemple

Ellipses de confiance exact (rouge) et asymptotique (noir) pour $X_1, \ldots, X_{10}$ $(X_1, \ldots, X_{50})$ $\mathcal{N}_2(\mu, \Sigma)$, où $\begin{pmatrix} 4 \\ 3 \end{pmatrix}$ et $\Sigma = \begin{pmatrix} 5 & 3 \\ 3 & 2.25 \end{pmatrix}$.





Zones de confiance

A ces zones de confiance elliptiques

$$C_{1-\alpha}^{(n)} = \left\{ \mu \in \mathbb{R}^p \;\middle|\; d_S^2(\bar{X}, \mu) \leq \frac{p(n-1)}{(n-p)n} F_{p,n-p;1-\alpha} \right\},$$

il est souvent préféré en pratique des zones “rectangulaires", qui livrent des intervalles de confiance pour chacune des composantes de $\mu = (\mu_1, \ldots, \mu_p)'$.

Bien entendu, il est facile de construire des intervalles de confiance pour toute combili $a'\mu$ des composantes de μ
(ici, a est un p-vecteur non nul fixé).

puisque $a'X_1, \ldots, a'X_n$ sont i.i.d. $\mathcal{N}_1(a'\mu, a'\Sigma a)$.

Zones de confiance

On obtient en effet directement que

$$C^{(n)}_{1-\alpha}(a) := \left\{ t \in \mathbb{R} \,\middle|\, d^2_{a'Sa}(a'\bar{X}, t) \leq \frac{1}{n} F_{1,n-1;1-\alpha} \right\},$$

constitue une zone (un intervalle) de confiance à $(1-\alpha) \times 100\%$ pour $a'\mu$.

Cet intervalle de confiance se réécrit simplement

$$a'\bar{X} \pm \sqrt{\frac{a'Sa}{n} F_{1,n-1;1-\alpha}},$$

ou encore

$$a'\bar{X} \pm \sqrt{\frac{a'Sa}{n}} t_{n-1;1-\alpha/2}.$$

Zones de confiance

Ainsi, un intervalle de confiance à $(1-\alpha)\times 100\%$ pour μ_i $(i=1,\ldots,p)$ est donné par

$$C_{1-\alpha}^{i,(n)} = (\bar{X})_i \pm \sqrt{\frac{(S)_{ii}}{n} F_{1,n-1;1-\alpha}}.$$

Néanmoins, il faut insister sur le fait qu'il s'agit là d'intervalles de confiance individuels, dans le sens où, s'il est vrai que, $\forall i=1,\ldots,p$, $P\left[\mu_i \in C_{1-\alpha}^{i,(n)}\right] \geq 1-\alpha$, il est faux (pour $p \geq 2$) que

$$P\left[\forall i=1,\ldots,p,\ \mu_i \in C_{1-\alpha}^{i,(n)}\right] \geq 1-\alpha.$$

Le zone rectangulaire $C_{1-\alpha}^{1,(n)} \times \ldots \times C_{1-\alpha}^{p,(n)}$ n'est donc pas une zone de confiance à $(1-\alpha)\times 100\%$ pour μ.

Zones de confiance

Question naturelle:

Comment construire des intervalles de confiance simultanés ?

Nous aurons besoin du lemme suivant:

Lemme *Soit M une matrice $p \times p$ symétrique et définie positive. Alors, $\forall a, b \in \mathbb{R}^p$, $(a'b)^2 \leq (a'Ma)(b'M^{-1}b)$.*

Preuve: Notons que $a'Ma = \|M^{1/2}a\|^2$ et que l'inégalité de Cauchy-Schwarz donne

$$\begin{aligned}(a'b)^2 &= (a'M^{1/2}M^{-1/2}b)^2 \\ &= \langle M^{1/2}a, M^{-1/2}b\rangle^2 \\ &\leq \|M^{1/2}a\|^2\|M^{-1/2}b\|^2.\end{aligned}$$

□

Zones de confiance

Conséquence: pour tout $a \in \mathbb{R}^p$, on a

$$(a'(\bar{X} - \mu))^2 \leq (a'Sa)((\bar{X} - \mu)'S^{-1}(\bar{X} - \mu)),$$

ou encore

$$\frac{(a'(\bar{X} - \mu))^2}{a'Sa} \leq \frac{1}{n} T^2(\mu),$$

de sorte que

$$P\left[\sup_a \frac{(a'(\bar{X} - \mu))^2}{a'Sa} \leq \frac{(n-1)p}{n(n-p)} F_{p,n-p;1-\alpha}\right] \geq 1 - \alpha.$$

Zones de confiance

Des intervalles de confiance simultanés (pour tout $a \in \mathbb{R}^p$) pour $a'\mu$ à $(1-\alpha) \times 100\%$ sont donc donnés par

$$a'\bar{X} \pm \sqrt{\frac{(n-1)p}{n(n-p)}(a'Sa)\, F_{p,n-p;1-\alpha}}.$$

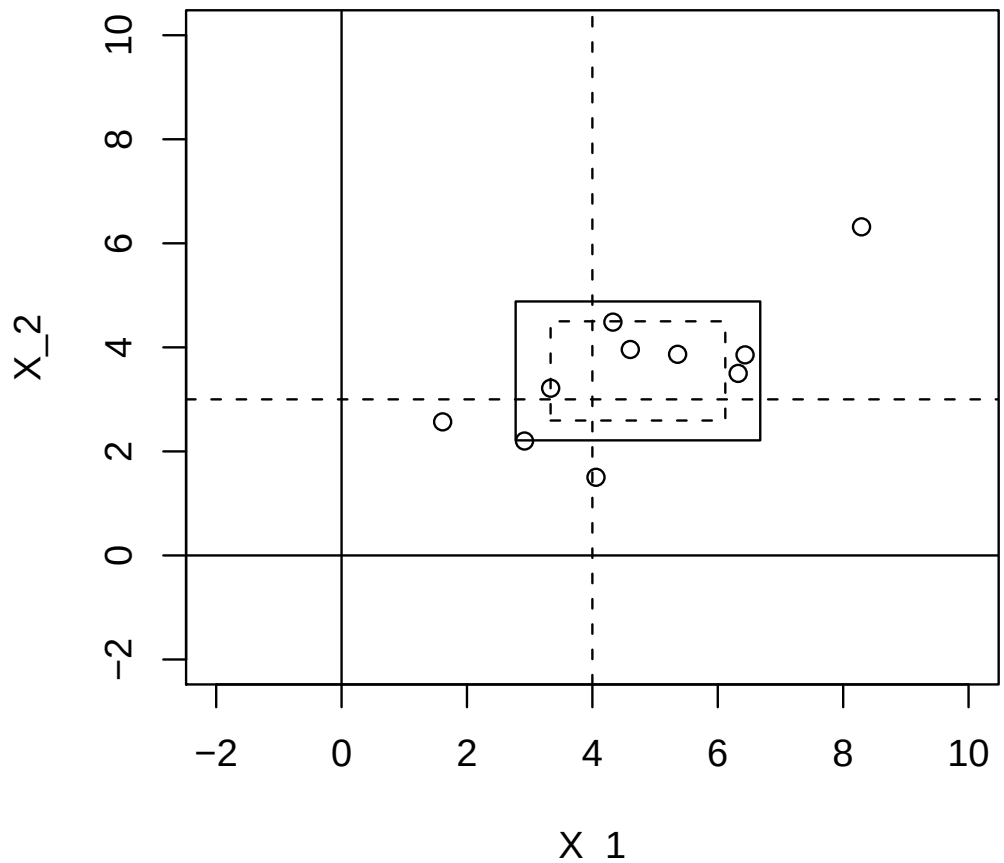
Ceux-ci sont à comparer aux intervalles de confiance individuels

$$a'\bar{X} \pm \sqrt{\frac{a'Sa}{n}\, F_{1,n-1;1-\alpha}},$$

qui ont été obtenus plus haut.

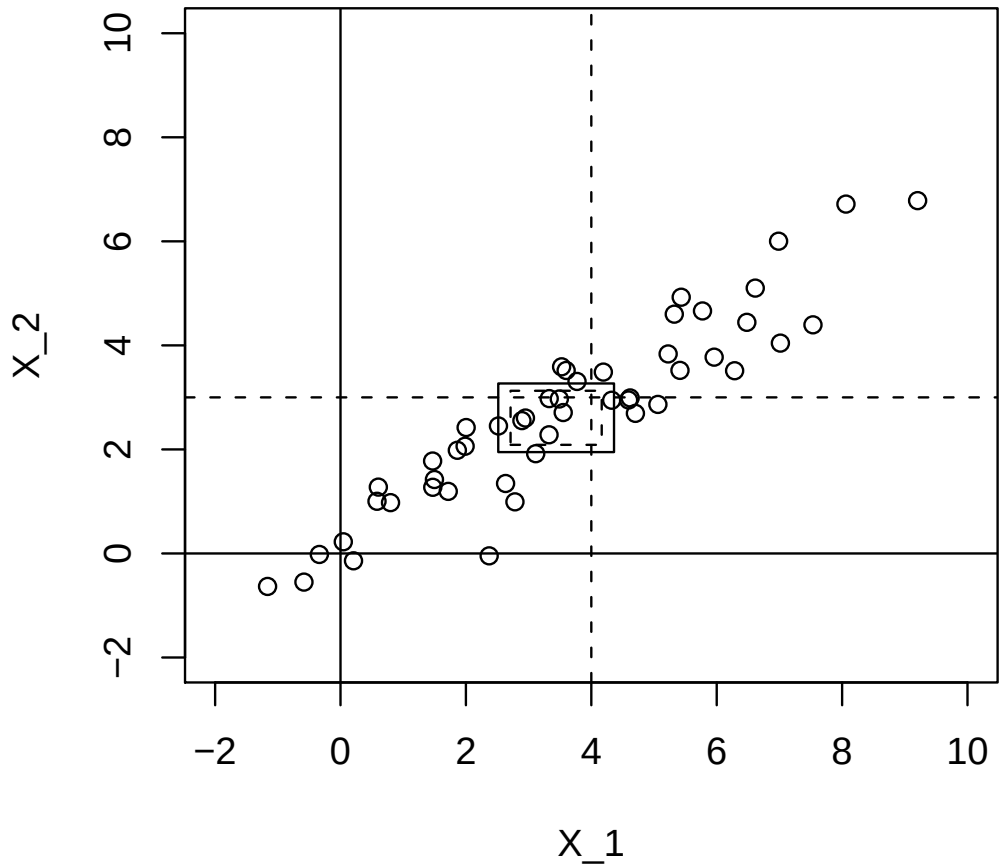
Exemple

Intervalles de confiance simultanés (noir) et individuels (rouge!) pour $X_1, \ldots, X_{10}$ i.i.d. $\mathcal{N}_2(\mu, \Sigma)$, où $\mu = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$ et $\Sigma = \begin{pmatrix} 5 & 3 \\ 3 & 2.25 \end{pmatrix}$.



Exemple

Intervalles de confiance simultanés (noir) et individuels (rouge!) pour $X_1, \ldots, X_{50}$ i.i.d. $\mathcal{N}_2(\mu, \Sigma)$, où $\mu = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$ et $\Sigma = \begin{pmatrix} 5 & 3 \\ 3 & 2.25 \end{pmatrix}$.



3. Inférence dans les modèles gaussiens.

3.1. Sur le paramètre de position.

3.1.1. Estimateurs MLE.

3.1.2. Tests de Hotelling.

3.1.3. Zones de confiance.

3.1.4. Problèmes à plusieurs échantillons.

3.2. Sur le paramètre de dispersion.

3.3. Autres types de problèmes.

Problèmes à plusieurs échantillons

Soient deux échantillons indépendants:
$X_1, \ldots, X_{n_1}$ i.i.d. $\mathcal{N}_p(\mu_1, \Sigma)$ et $Y_1, \ldots, Y_{n_2}$ i.i.d. $\mathcal{N}_p(\mu_2, \Sigma)$.

Nous considérons le problème de test

$$\begin{cases} \mathcal{H}_0 : \mu_1 = \mu_2 \\ \mathcal{H}_1 : \mu_1 \neq \mu_2. \end{cases}$$

Remarque: plus généralement, on pourrait traiter le cas où les matrices de variance-covariance des deux échantillons sont différentes.

Dans ce cas, les tests gaussiens fondent la règle de décision sur $\bar{X} - \bar{Y}$ (et plus spécifiquement sur la distance entre $\bar{X}$ et $\bar{Y}$).

Tests de Hotelling

$\rightsquigarrow$ la statistique du test de Hotelling pour deux échantillons est

$$T^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} (\bar{X} - \bar{Y})' S_{\text{pool}}^{-1} (\bar{X} - \bar{Y}) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} d^2_{S_{\text{pool}}}(\bar{X}, \bar{Y}),$$

$$\text{où } \bar{X} := \frac{1}{n_1} \sum_{i=1}^{n_1} X_i, \;\; W_x := \sum_{i=1}^{n_1} (X_i - \bar{X})(X_i - \bar{X})',$$

$$\bar{Y} := \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i, \;\; W_y := \sum_{i=1}^{n_2} (Y_i - \bar{Y})(Y_i - \bar{Y})',$$

et

$$S_{\text{pool}} := \frac{W_x + W_y}{n_1 + n_2 - 2}.$$

Et il convient de rejeter $\mathcal{H}_0 : \mu_1 = \mu_2$ pour de grandes valeurs de T^2.

Tests de Hotelling (loi exacte)

Le résultat suivant précise la loi exacte (sous $\mathcal{H}_0$) de la statistique de test de Hotelling:

Proposition: *supposons que* $n_1 + n_2 \geq p + 2$. *Alors sous* $\mathcal{H}_0$,

$$\frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} \, T^2 \sim F_{p, n_1+n_2-p-1}.$$

Le test de Hotelling exact consiste donc (au niveau α) à rejeter $\mathcal{H}_0 : \mu_1 = \mu_2$ ssi $\frac{n_1+n_2-p-1}{p(n_1+n_2-2)} \, T^2 > F_{p, n_1+n_2-p-1; 1-\alpha}$.

Dans sa version asymptotique, ce test rejette $\mathcal{H}_0 : \mu_1 = \mu_2$ ssi $T^2 > \chi^2_{p;1-\alpha}$. Dans ce cas, comme pour le problème à un échantillon, la normalité n'est pas requise (seules l'existence de moments finis d'ordre 2 et l'égalité des matrices de variance-covariance population le sont). Exercice: vérifier ceci en utilisant le TCL.

Tests de Hotelling

Preuve de la proposition: comme dans le cas à un échantillon, la loi (sous $\mathcal{H}_0$) de la statistique T^2 découle du lemme suivant:

Lemme: *soient* $Y \sim \mathcal{N}_p(0, \Sigma)$ *et* $V \sim W_p(m, \Sigma)$. *Alors, si* $m \geq p$ *et* $Y \perp\!\!\!\perp V$, $\frac{m-p+1}{p}\, Y' V^{-1} Y \sim F_{p,m-p+1}$.

En effet, sous $\mathcal{H}_0$, $\bar{X}$ et $\bar{Y}$ sont indépendantes et de loi respective $\mathcal{N}_p(\mu, \frac{1}{n_1}\Sigma)$ et $\mathcal{N}_p(\mu, \frac{1}{n_2}\Sigma)$ (où μ est la valeur commune de μ_1 et μ_2). Donc $\bar{X} - \bar{Y} \sim \mathcal{N}_p(0, (\frac{1}{n_1} + \frac{1}{n_2})\Sigma)$.

D'autre part, $W_x \sim W_p(n_1 - 1, \Sigma)$ et $W_y \sim W_p(n_2 - 1, \Sigma)$ sont aussi indépendantes, de sorte que

$$(n_1 + n_2 - 2) S_{\text{pool}} = W_x + W_y \sim W_p(n_1 + n_2 - 2, \Sigma).$$

Le lemme fournit alors le résultat en prenant $Y := (\frac{1}{n_1} + \frac{1}{n_2})^{-1/2}(\bar{X} - \bar{Y})$ et $V := (n_1 + n_2 - 2) S_{\text{pool}}$. □

Propriétés d'invariance

La statistique de test T^2 (et par suite, le test lui-même) est ici invariante par transformations linéaires et par translations :
pour toute matrice A $(p \times p)$ inversible et pour tout p-vecteur b,

$$\begin{aligned} &T^2(AX_1 + b, \ldots, AX_{n_1} + b, AY_1 + b, \ldots, AY_{n_2} + b) \\ &\quad = T^2(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}). \end{aligned}$$

Cette invariance affine explique le fait que la loi de T^2 sous $\mathcal{H}_0$ ne dépende

- ni de la valeur de Σ,
- ni de la valeur commune de $\mu_1 = \mu_2$.

Test du rapport de vraisemblance

Comme dans le cas à un échantillon, le test de Hotelling est essentiellement celui du rapport de vraisemblance gaussien:

Théorème: *soit* $\Lambda^{(n_1,n_2)}$ *la statistique du test du rapport de vraisemblance. Alors*

$$\Lambda^{(n_1,n_2)} = \left(1 + \frac{T^2}{n_1 + n_2 - 2}\right)^{-(n_1+n_2)/2}.$$

Preuve: exercice.

Remarque

Pour ce problème, on a constamment supposé que les deux échantillons $(X_1, \ldots, X_{n_1})$ et $(Y_1, \ldots, Y_{n_2})$ sont indépendants.

Si ce n'est pas le cas, tout ce qui a été fait plus haut s'effondre...

Exemple classique:

Supposons que les deux échantillons soient pairés : $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$, où X_i et Y_i reprennent p mesures effectuées, avant et après traitement respectivement, sur un même individu.

Dans ce cas, si on veut tester $\mathcal{H}_0 : \mu_1 = \mu_2$, il convient d'effectuer un test à un échantillon de $\mathcal{H}_0 : \mu = 0$ sur la série des différences $(Y_1 - X_1, \ldots, Y_n - X_n)$.

3. Inférence dans les modèles gaussiens.

3.1. Sur le paramètre de position.

3.1.1. Estimateurs MLE.

3.1.2. Tests de Hotelling.

3.1.3. Zones de confiance.

3.1.4. Problèmes à plusieurs échantillons.

3.2. Sur le paramètre de dispersion.

3.3. Autres types de problèmes.

Test d'adéquation sur Σ

Tous les tests suivants sont des test de rapport de vraisemblance. Je laisse les preuves pour les TP.

Soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}_p(\mu, \Sigma)$. Considérons le problème de test

$$\begin{cases} \mathcal{H}_0 : \Sigma = \Sigma_0 \\ \mathcal{H}_1 : \Sigma \neq \Sigma_0, \end{cases}$$

où Σ_0 une matrice $p \times p$ symétrique et définie positive fixée.

Dans ce cas, le test de rapport de vraisemblance rejette $\mathcal{H}_0$ (au niveau asymptotique α) si

$$-2 \ln \Lambda^{(n)} > \chi^2_{p(p+1)/2;1-\alpha},$$

où

$$\Lambda^{(n)} = e^{np/2} \, |\Sigma_0^{-1}\hat{\Sigma}|^{n/2} \exp\left[-\frac{n}{2} \operatorname{tr}(\Sigma_0^{-1}\hat{\Sigma}) \right].$$

Problème à deux échantillons

Soient deux échantillons indépendants:
$X_1, \ldots, X_{n_1}$ i.i.d. $\mathcal{N}_p(\mu_1, \Sigma_1)$ et $Y_1, \ldots, Y_{n_2}$ i.i.d. $\mathcal{N}_p(\mu_2, \Sigma_2)$.

Pour le problème de test

$$\begin{cases} \mathcal{H}_0 : \Sigma_1 = \Sigma_2 \\ \mathcal{H}_1 : \Sigma_1 \neq \Sigma_2 \end{cases}$$

le test de rapport de vraisemblance rejette $\mathcal{H}_0$ (au niveau asymptotique α) si

$$-2 \ln \Lambda^{(n_1,n_2)} > \chi^2_{p(p+1)/2;1-\alpha},$$

où

$$\Lambda^{(n_1,n_2)} = \frac{|W_x/n_1|^{n_1/2}|W_y/n_2|^{n_2/2}}{|(W_x + W_y)/(n_1+n_2)|^{(n_1+n_2)/2}}.$$

Test de sphéricité

Soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}_p(\mu, \Sigma)$.

Considérons le problème de test

$$\begin{cases} \mathcal{H}_0 : \exists \lambda > 0 \text{ tel que } \Sigma = \lambda I_p \\ \mathcal{H}_1 : \forall \lambda > 0,\ \Sigma \neq \lambda I_p, \end{cases}$$

qui consiste à tester la sphéricité des contours d'équidensité sous-jacents.

Dans ce cas, le test de rapport de vraisemblance rejette $\mathcal{H}_0$ (au niveau asymptotique α) si

$$-2 \ln \Lambda^{(n)} > \chi^2_{\frac{p(p+1)}{2}-1;1-\alpha}, \quad \text{où} \quad \Lambda^{(n)} = \left(\frac{|S|^{1/p}}{\frac{1}{p}(\operatorname{tr} S)} \right)^{np/2}.$$

Test de sphéricité

Remarque: en écrivant

$$S = O\Lambda O', \quad \text{où} \quad \Lambda := \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}$$

et où O est orthogonale, on obtient que

$$\left(\Lambda^{(n)}\right)^{2/(np)} = \frac{\prod_i \lambda_i^{1/p}}{\frac{1}{p}\sum_i \lambda_i},$$

qui n'est autre que le quotient de la moyenne géométrique des valeurs propres de S par leur moyenne arithmétique (intuition).

3. Inférence dans les modèles gaussiens.

3.1. Sur le paramètre de position.

3.1.1. Estimateurs MLE.

3.1.2. Tests de Hotelling.

3.1.3. Zones de confiance.

3.1.4. Problèmes à plusieurs échantillons.

3.2. Sur le paramètre de dispersion.

3.3. Autres types de problèmes.

Test d'indépendance

Soient $Z_1 = (X_1', Y_1')', \ldots, Z_n = (X_n', Y_n')'$ i.i.d. $\mathcal{N}_{p_1+p_2}(\mu, \Sigma)$, où

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{et} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Considérons le problème de test

$$\begin{cases} \mathcal{H}_0 : \Sigma_{12} = 0 \\ \mathcal{H}_1 : \Sigma_{12} \neq 0 \end{cases}$$

qui (dans cette situation gaussienne) consiste à tester l'indépendance entre X_1 et Y_1.

Test d'indépendance

Le test de rapport de vraisemblance rejette ici $\mathcal{H}_0$ (au niveau asymptotique α) si

$$-2 \ln \Lambda^{(n)} > \chi^2_{p_1 p_2; 1-\alpha},$$

où

$$\Lambda^{(n)} = \left(\frac{|S_z|}{|S_x||S_y|} \right)^{n/2},$$

avec

$$S_z := \frac{1}{n-1} \sum_{i=1}^{n} \begin{pmatrix} X_i - \bar{X} \\ Y_i - \bar{Y} \end{pmatrix} \begin{pmatrix} X_i - \bar{X} \\ Y_i - \bar{Y} \end{pmatrix}' =: \begin{pmatrix} S_x & S_{xy} \\ S_{yx} & S_y \end{pmatrix}.$$