3.15 Théorème de Cochran et test du χ^2

Leçons : 261, 262, 266.

Références : [GK], [RS].

Prérequis : matrice de projection orthogonale, vecteurs gaussiens, TCL multi-dimensionnel, loi du χ^2.

Théorème (Cochran) : Soit $X \sim \mathcal{N}_n(0, I_n)$. Soit $\mathbb{R}^n = E_1 \oplus ... \oplus E_p$ une décomposition en somme directe orthogonale. On note $d_i = \dim(E_i)$. Soit π_i la matrice de projection orthogonale sur E_i, et $Y_i = \pi_i X$. Alors :

1. Les Y_i sont des vecteurs gaussiens indépendants et $Y_i \sim \mathcal{N}_{d_i}(0, \pi_i)$.
2. Les VA $\|Y_i\|^2$ sont indépendants, et $\|Y_i\|^2 \sim \chi^2(d_i)$.

Corollaire : Soit $(X_1, ...X_n)$ des VA i.i.d. à valeurs dans $\{1, ..., m\}$. On note pour tout $i \in \{1, ..., m\}$: $N_i(n) = \sum_{k=1}^n 1_{\{X_k = i\}}$ et $\pi_i = \mathbb{P}(X_1 = i)$. Alors :

$$D_n(\pi) = \sum_{i=1}^m \frac{(N_i(n) - n\pi_i)^2}{n\pi_i} \underset{n\to\infty}{\mathcal{L}} \chi^2(m-1).$$

Preuve du théorème de Cochran. Pour tout $k \in \{1, ..., p\}$ on note $(e_{k,1}, ..., e_{k,d_k})$ une base orthonormée de E_k, et P la matrice de passage de la base canonique à la base $(e_{1,1}, ..., e_{p,d_p})$. On pose $Y = {}^tPX$. Alors d'après les propriétés des vecteurs gaussiens, $Y \sim \mathcal{N}_n(0, {}^tPI_nP) = \mathcal{N}_n(0, I_n)$, donc Y est diagonale et les composantes de Y, qui s'écrivent ${}^te_{i,j}X$, sont des variables indépendantes. Or la matrice de projection sur E_i s'écrit : $\pi_i = (e_{i,1}...e_{i,d_i})\,{}^t(e_{i,1}...e_{i,d_i})$, donc $Y_i = \pi_i X = \sum_{j=1}^{d_i} ({}^te_{i,j}X)e_{i,j}$, qui ne dépend que des ${}^te_{i,j}X$. Ainsi, les (Y_i) sont indépendantes, donc les $\|Y_i\|^2$ aussi. En outre,

$$Y_i = \pi_i X \sim \mathcal{N}_{d_i}(0, \pi_i I_n\, {}^t\pi_i) = \mathcal{N}_{d_i}(0, \pi_i),$$

car π_i est une matrice de projection, et puisque les $(e_{i,j})$ forment une base orthonormale :

$$\|Y_i\|^2 = \sum_{j=1}^{d_i} ({}^te_{i,j}X)^2,$$

et les variables ${}^te_{i,j}X \sim \mathcal{N}(0,1)$ sont indépendantes, donc $\|Y_i\|^2 \sim \chi^2(d_i)$. D'où le théorème de Cochran. □

Preuve du corollaire. On pose $\pi = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_m \end{pmatrix}$, et $N(n) = \begin{pmatrix} N_1(n) \\ \vdots \\ N_m(n) \end{pmatrix}$. On pose aussi, pour tout $k \geqslant 1$, $Y_k = \begin{pmatrix} 1_{\{X_k=1\}} \\ \vdots \\ 1_{\{X_k=m\}} \end{pmatrix}$, de sorte que $N(n) = \sum_{k=1}^n Y_k$. Les vecteurs aléatoires $(Y_k)_{k\geqslant 1}$ sont i.i.d., d'espérance commune π, et de matrice de covariance commune donnée par :

$$\begin{aligned} \forall 1 \leqslant i, j \leqslant m,\ \mathrm{Cov}[Y_1]_{i,j} &= \mathrm{Cov}(1_{\{X_k=i\}}, 1_{\{X_k=j\}}) \\ &= \mathbb{E}[(1_{\{X_k=i\}} - \pi_i)(1_{\{X_k=j\}} - \pi_j)] \\ &= \mathbb{E}[\delta_{i,j} 1_{\{X_k=i\}}] - \pi_i\pi_j, \end{aligned}$$

soit $\mathrm{Cov}[Y_1] = \Delta_\pi - \pi\,{}^t\pi =: \Gamma$, où $\Delta_\pi = \begin{pmatrix} \pi_1 & & \\ & \ddots & \\ & & \pi_m \end{pmatrix}$. Ainsi, d'après le théorème central limite multi-dimensionnel, on a

$$\frac{N(n) - n\pi}{\sqrt{n}} \underset{n\to\infty}{\mathcal{L}} \mathcal{N}_m(0, \Gamma).$$

On pose $f : \mathbb{R}^m \to \mathbb{R}$ définie par $f(x) = \sum_{k=1}^m \frac{x_k^2}{\pi_i}$. La fonction f est continue, donc on en déduit

$$D_n(\pi) = f\left(\frac{N(n) - n\pi}{\sqrt{n}}\right) \underset{n\to\infty}{\mathcal{L}} f(Z),$$

où $Z \sim \mathcal{N}_m(0,\Gamma)$. Pour conclure, il reste donc à trouver la loi de $f(Z)$. On peut réécrire $f(Z) = \|U\|^2$ où $U \sim \mathcal{N}_m(0, I_m - \sqrt{\pi}\,{}^t\sqrt{\pi})$. Or, on a la décomposition en somme directe orthogonale $\mathbb{R}^m = E_1 \oplus E_2$, où $E_1 = \text{Vect}(\sqrt{\pi})$ et $E_2 = E_1^\perp$. Une matrice de projection orthogonale sur E_1 est $\sqrt{\pi}\,{}^t\sqrt{\pi}$ et une matrice de projection orthogonale sur E_2 est $I_m - \sqrt{\pi}\,{}^t\sqrt{\pi}$, donc d'après les notations du théorème de Cochran, U a la même loi que Y_2, et donc $\|U\|^2$ a la même loi que $\|Y_2\|^2 \sim \chi^2(m-1)$. D'où finalement $D_n(\pi) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \chi^2(m-1)$.

□

Remarques :

1. Une autre application importante du théorème de Cochran est la construction d'un intervalle de confiance de l'espérance d'un n-échantillon de variables de loi normale, lorsque la variance est inconnue (voir Probabilités et statistiques pour l'épreuve de modélisation, Chabanol).
2. Une application importante du corollaire est le test du chi-deux, qui est un test asymptotique : on suppose que l'on dispose d'un n-échantillon $X_1, ..., X_n$ de variables à valeurs dans un ensemble discret et fini, que l'on note $\{1, ..., m\}$ pour simplifier. On souhaite mettre en place un test d'adéquation de la loi à une loi donnée : plus précisément, en gardant les notations du corollaire, on souhaite tester
$$(H_0) : \pi = \pi^0 \text{ contre } (H_1) : \pi \neq \pi^0,$$
où π^0 est donné. Sous l'hypothèse (H_0), on a
$$D_n(\pi^0) = \sum_{i=1}^{m} \frac{(N_i(n) - n\pi_i^0)^2}{n\pi_i^0} \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \chi^2(m-1),$$
et sinon, $D_n(\pi^0) \underset{n\to\infty}{\longrightarrow} +\infty$ presque sûrement (d'après la loi des grands nombres). On va donc prendre $D_n(\pi^0)$ comme statistique de test, et la zone de rejet sera de la forme $[s, +\infty[$. Pour trouver s, on utilise en pratique l'approximation de la loi de $D_n(\pi^0)$, lorsque $\pi = \pi^0$, par la loi $\chi^2(m-1)$, et pour un test de niveau α, on choisit s tel que $\mathbb{P}(K \geqslant s) = \alpha$, où $K \sim \chi^2(m-1)$.

Questions :

1. Détailler la formule pour la matrice de projection.
2. Pourquoi si la matrice de covariance d'un vecteur gaussien est diagonale, les variables sont indépendantes ?
3. En général, est-ce-que des variables aléatoires non corréelées sont indépendantes ?
4. On sait que si des variables gaussiennes réelles sont indépendantes, le vecteur associé est gaussien. Donner un contre-exemple si les variables ne sont pas indépendantes.

Réponses :

1. On utilise la propriété : la matrice de projection orthogonale sur un espace engendré par une base orthogonale $(e_1, ..., e_d)$ est $P = A({}^tAA)^{-1}\,{}^tA$ où $A = (e_1...e_d)$.
2. Dans le cas de deux variables aléatoires X et Y, si (X, Y) est un vecteur gaussien et $\text{Cov}(X, Y) = 0$, on montre :
$$\forall x, y \in \mathbb{R},\ \varphi_{(X,Y)}(x, y) = \varphi_X(x)\varphi_Y(y),$$
ce qui suffit à démontrer l'indépendance.
3. Non : si X est de loi uniforme sur $\{-1, 0, 1\}$ et $Y = X^2$, alors on montre que les variables X et Y sont non corrélées (i.e. de covariance nulle), mais elles ne sont pas indépendantes.
4. On considère $X \sim \mathcal{N}(0, 1)$, ε une variable qui suit une loi de Rademacher (i.e. de loi uniforme sur $\{-1, 1\}$), et $Y = \varepsilon X$, alors on montre que X et Y sont gaussiens non indépendants, mais de somme non gaussienne (il y a un atome en 0, ce qui n'est pas possible pour une loi normale), donc le vecteur (X, Y) n'est pas gaussien.