Tests d'homogénéité dans les modèles de mélange

A. Autin, C. Pouet

Université de Provence

Rennes, 29 Août 2008

Plan

1. Cadre du problème
2. Résultats non-adaptatifs et adaptatifs
3. Idées des preuves
4. Simulations

Modèle de mélange

Soit X une variable aléatoire dont la densité est

$$f(x) = \sum_{u=1}^{M} \omega_u p_u(x),$$

où M est le nombre de sous-populations, p_u la densité de la sous-population u et ω_u la proportion de cette sous-population dans la population totale.

Problème d'identifiabilité : Hall et Zhou (2003).

Quelques domaines d'application : médecine, biologie, enquêtes en sciences humaines, ...

Cadre général

Problème de test d'homogénéité avec deux échantillons : Butucea et Tribouley (2006)

$$Y_1, \ldots, Y_n \quad \text{avec} \quad Y_i \sim f_i(\cdot) = \sum_{u=1}^{M} \omega_u(i)\, p_u(\cdot)$$

$$Z_1, \ldots, Z_n \quad \text{avec} \quad Z_i \sim g_i(\cdot) = \sum_{u=1}^{M} \sigma_u(i)\, q_u(\cdot).$$

Problème de test d'homogénéité avec un échantillon : Maiboroda (2000)

Propriété : Maiboroda (2000), Pokhyl'ko (2005)

Les jeux de poids $\omega_u(i)$ et $\sigma_u(i)$ varient en fonction de i.

Problème de test

$$H_0 \quad : \quad \forall l = 1, \dots, M, \ p_l = q_l \in \mathcal{S}(R) = \mathbb{L}_\infty(R) \cap \mathbb{L}_2(R)$$

$$\text{contre } H_1 \quad : \quad p_l - q_l \in \mathcal{B}^s_{2,\infty}(R), \quad \sum_{l=1}^{M} \|p_l - q_l\|_2^2 \geq C^2 \psi(n)^2 .$$

Approche minimax

La suite $\psi_s(n)$ est la vitesse de test si pour tout $\gamma > 0$:
(i) il existe une constante C_0 et un test $\Delta_n^\star$ tels que

$$\forall C > C_0 : \lim_{n \to +\infty} \mathbb{P}_0(\Delta_n^\star = 1) + \sup_{f \in \Lambda_n(s,R,C,\psi_s(n))} \mathbb{P}_f(\Delta_n^\star = 0) \leq \gamma.$$

(ii) il existe une constante c_0 telle que

$$\forall C < c_0 : \lim_{n \to +\infty} \inf_{\Delta_n} \left[\mathbb{P}_0(\Delta_n = 1) + \sup_{f \in \Lambda_n(s,R,C,\psi_s(n))} \mathbb{P}_f(\Delta_n = 0) \right] \geq \gamma.$$

Perte en adaptation (Spokoiny, 1997)

Soit $\mathcal{T}$ un ensemble non-vide pour le paramètre s.
La perte en adaptation est définie comme la suite t_n tendant vers 0 telle que
Si $\lim \frac{t'_n}{t_n} = 0$, alors

$$\lim_{n\to+\infty} \inf_{\Delta_n} \left[\mathbb{P}_0 \left(\Delta_n = 1\right) + \sup_{s\in\mathcal{T}} \sup_{f\in\Lambda(s,R,C,\psi_s(nt'_n))} \mathbb{P}_f \left(\Delta_n = 0\right) \right] = 1.$$

Il existe une constante C_1 et un test $\Delta_n^\star$ tels que

$$\lim_{n\to+\infty} \mathbb{P}_0 \left(\Delta_n^\star = 1\right) + \sup_{s\in\mathcal{T}} \sup_{f\in\Lambda_n(s,R,C_1,\psi_s(nt_n))} \mathbb{P}_f \left(\Delta_n^\star = 0\right) = 0.$$

Hypothèses

Soient $\Omega = (\Omega_{jl})_{1\leq j\leq n, 1\leq l\leq M} = (\omega_l(j))_{1\leq j\leq n, 1\leq l\leq M}$ et $\Sigma = (\Sigma_{jl})_{1\leq j\leq n, 1\leq l\leq M} = (\sigma_l(j))_{1\leq j\leq n, 1\leq l\leq M}$.

Alors

H.1 Le rang des matrices $\Omega^\star$ et $\Sigma^\star$ est M.

H.2 La plus petite des valeurs propres des matrices $\Omega\Omega^\star$ et $\Sigma\Sigma^\star$ est supérieure à $\frac{K}{n}$.

On définit $a_l(i)$ et $b_l(i)$ tels que

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \omega_k(i)\, a_l(i) &= \delta_{kl}, \\ \frac{1}{n}\sum_{i=1}^{n} \sigma_k(i)\, b_l(i) &= \delta_{kl}. \end{aligned}$$

Statistique de test

Soit $\psi_s(n) = n^{-\frac{2s}{4s+1}}$.
Pour un paramètre j, on définit le test

$$\Delta_j = \begin{cases} 1 & \text{si } T_j > t\ \psi_s(n)^2, \\ 0 & \text{si } T_j \leq t\ \psi_s(n)^2. \end{cases}$$

La statistique de test est

$$\begin{aligned} T_j &= \frac{1}{n^2} \sum_{l=1}^{M} \sum_{k} \sum_{i_1 \neq i_2} [a_l(i_1)\phi_{jk}(Y_{i_1}) - b_l(i_1)\phi_{jk}(Z_{i_1})] \\ &\qquad [a_l(i_2)\phi_{jk}(Y_{i_2}) - b_l(i_2)\phi_{jk}(Z_{i_2})]\,. \end{aligned}$$

Test adaptatif

Soient une grille de taille $\mathcal{O}(\ln n)$ pour le paramètre s et la grille équivalente pour le paramètre j.
On définit le test adaptatif

$$\tilde{\Delta}_n = \begin{cases} 1 & \text{si il existe } j \text{ tel que } T_j > t\ \psi_s(n(\ln\ln n)^{-\frac{1}{2}})^2, \\ 0 & \text{si pour tout } j,\ T_j < t\ \psi_s(n(\ln\ln n)^{-\frac{1}{2}})^2. \end{cases}$$

Théorèmes

Théorème non-adaptatif : Soit j_n le plus petit entier tel que $2^{-j_n} \leq n^{-\frac{2}{4s+1}}$. Soient C_0 et t les solutions de

$$\frac{2K_T}{\left(t - \frac{8LMR^2}{K}\right)^2} = \frac{\gamma}{2};$$

$$\frac{3K_T}{\left(C_\gamma^2 - \frac{8LR^2}{KM} - MR - t\right)^2} = \frac{\gamma}{2}.$$

Alors la vitesse minimax asymptotique de test est

$$\psi_s(n) = n^{-\frac{2s}{4s+1}}$$

et le test Δ_{j_n} est le test minimax.

Théorème adaptatif : Si les jeux de poids sont égaux, alors le test $\tilde{\Delta}_n$ est adaptatif et la perte en adaptation est $t_n = (\ln \ln n)^{-\frac{1}{2}}$.

Borne supérieure

Cas non-adaptatif : inégalité de Bienaymé-Chebyshev

$$\begin{aligned}\mathbb{E}_{f,g}(T_j) &= \sum_{l=1}^{M}\sum_{k}\left(\int_{\mathbb{R}}(p_l - q_l)\phi_{jk}\right)^2 \\ &\quad - \frac{1}{n^2}\sum_{l=1}^{M}\sum_{k}\sum_{i=1}^{n}\left(\int_{\mathbb{R}}(a_l(i)f_i - b_l(i)g_i)\phi_{jk}\right)^2, \\ \mathrm{Var}_{f,g}(T_j) &= K_T\left(\frac{2^j}{n^2} + \frac{1}{n}\sum_{l=1}^{M}\|p_l - q_l\|_2^2 + \sqrt{\frac{2^{j\,3}}{n}}\sum_{l=1}^{M}\|p_l - q_l\|_2\right).\end{aligned}$$

Cas adaptatif : inégalité de Berry-Esseen adaptée (voir Petrov, 1995) afin de contrôler l'erreur de 1ère espèce.

Borne inférieure

Cas non-adaptatif : preuve classique avec une loi a priori sur les fonctions de l'alternative.

Cas adaptatif : utilisation d'une famille de taille $K_1 \ln n$ de lois a priori indexée par le paramètre j.

$$\begin{aligned}
\gamma_n &\geq 1 - \frac{1}{2} \left| \mathbb{P}_{f,f} - \frac{1}{K_1 \ln n} \sum_j \mathbb{P}_{\pi_j} \right|_1 \\
&\geq 1 - \frac{1}{2K_2 \ln\ln n} \sum_{\mathcal{J}} \left| \mathbb{P}_{f,f} - \frac{K_2 \ln\ln n}{K_1 \ln n} \sum_{j \in \mathcal{J}} \mathbb{P}_{\pi_j} \right|_1 \\
&\geq 1 - \frac{1}{2K_2 \ln\ln n} \sum_{\mathcal{J}} \left| \mathbb{P}_{f,f} - \frac{K_2 \ln\ln n}{K_1 \ln n} \sum_{j \in \mathcal{J}} \mathbb{P}_{\pi_j} \right|_2 .
\end{aligned}$$

Simulations

Cas du problème de test non-adaptatif pour un mélange de 2 lois gaussiennes.

Taille	Loi I de Y	Loi II de Y	Loi I de Z	Loi II de Z	Erreur	Puissance
500 (1)	$\mathcal{N}(0,1)$	$\mathcal{N}(2,1)$	$\mathcal{N}(2,1)$	$\mathcal{N}(0,1)$	0.07	0.87
500 (1)	$\mathcal{N}(1,1)$	$\mathcal{N}(2,1)$	$\mathcal{N}(2,1)$	$\mathcal{N}(0,1)$	0.09	0.47
1000 (1)	idem	idem	idem	idem	0.09	0.63
5000 (1)	idem	idem	idem	idem	0.09	0.85
5000 (2)	$\mathcal{N}(5,1)$	$\mathcal{N}(2,1)$	$\mathcal{N}(-1,1)$	$\mathcal{N}(0,1)$	0.07	0.89
5000 (2)	$\mathcal{N}(5,1)$	$\mathcal{N}(2,1)$	$\mathcal{N}(5,1)$	$\mathcal{N}(0,1)$	0.09	0.3

Essai	Poids I de Y	Poids II de Y	Poids I de Z	Poids II de Z
1	0.8	0.2	0.75	0.25
1	0.25	0.75	0.3	0.7
2	0.5	0.5	0.75	0.25
2	0.25	0.75	0.3	0.7