Divergence de Kullback-Leibler et sélection de modèles de mélange de régression multivariée

Abdelaziz El Matouat [1], Hassania Hamzaoui [2] & Abdelaziz Aloui [3]

[1] *MASI, ENS-USMBA, Fès, Maroc / LMAH, Université Le Havre Normandie, France, abdelaziz.el-matouat@univ-lehavre.fr*
[2] *LIMS, FSDM, Université de Fès, Maroc, hassania.hamzaoui@usmba.ac.ma*
[3] *FSDM, Université de Fès, Maroc, abdelaziz.aloui@usmba.ac.ma*

Résumé : Nous nous intéressons dans ce travail à l'estimation du nombre de composantes d'un modèle de mélange de régression multivariée. Nous généralisons les résultats obtenus par Hafidi et al.[4] qui ont proposé un critère MRC_{sd} en se basant sur la divergence de Kulback-Leibler symétrique pour identifier un modèle de mélange de régression univarié. Les simulations mettent en évidence la performance du critère proposé.

Mots-clés. Divergence de Kulback-Leibler symétrique, MRC_{sd}, BIC, AIC, régression multivariée.

Abstract : We are interested to estimate the number of components of multivariate regression mixture model. In this work, we generalize the results obtained by Hafidi and al. [4] who proposed a criterion MRC_{sd} based on the divergence of symmetric Kulback-Leibler to identify an univariate regression mixture model. Simulation results show the perfermance of the proposed criterion.

Keywords : Kullback-Leibler symmetric divergence, MRC_{sd}, BIC, AIC, multivariate regression.

1 Introduction

Un modèle de mélange linéaire multivarié consiste à écrire un vecteur y en fonction des covariables x de la façon suivante :

$$y = \begin{cases} \beta_1' x + \epsilon_1 & \text{avec la probabilité } \alpha_1 \\ \beta_2' x + \epsilon_2 & \text{avec la probabilité } \alpha_2 \\ . & \\ . & \\ \beta_K' x + \epsilon_K & \text{avec la probabilité } \alpha_K \end{cases}$$

où $y = (y_1, ..., y_m)'$, $x = (x_1, ..., x_p)'$, β_k une matrice d'ordre $p \times m$, ϵ_k désigne le bruit blanc de dimension m et de distribution gaussienne multivariée de moyenne nulle et de

matrice variance-covariance Σ_k.
La fonction densité relative à la loi de y conditionnellement à x est donnée par :

$$f(y, x, \phi) = \sum_{k=1}^{K} \alpha_k f_k(y, x, \beta_k, \Sigma_k) \tag{1}$$

où $\phi = \{(\alpha_k, \beta_k, \Sigma_k), k = 1, ..., K\}$ est l'ensemble des paramètres tel que $0 < \alpha_k < 1$ et $\Sigma_{k=1}^{K} \alpha_k = 1$, K le nombre de composantes et f_k la fonction densité d'une loi normale $N(\beta_k' x, \Sigma_k)$.
Soit $\{(y_1, x_1), ..., (y_n, x_n)\}$ un échantillon observé issu du modèle (1), La fonction log-vraisemblance s'écrit :

$$\begin{array}{rcl} L(\phi, Z, Y, X) & = & \log \prod_{j=1}^{n} \prod_{k=1}^{K} \{\alpha_k f_k(y_j, x_j, \beta_k, \Sigma_k)\}^{z_{jk}} \\ & = & \sum_{j=1}^{n} \sum_{k=1}^{K} z_{jk} \{log\alpha_k + log f_k(y_j, x_j, \beta_k, \Sigma_k)\} \end{array} \tag{2}$$

où $Y = (y_1, ..., y_n)'$, Z une variable latente d'ordre $n \times K$ telle que z_{jk} vaut 1 lorsque y_j provient de la k$^{\text{ième}}$ composante du mélange et 0 sinon, et $X = (x_1, ..., x_n)'$ la variable explicative.
On note f° la densité du vrai modèle telle que :

$$f^\circ(y, x, \phi^\circ) = \sum_{k=1}^{K^\circ} \alpha_k^\circ f_k^\circ(y, x, \beta_k^\circ, \Sigma_k^\circ)$$

L° est la log-vraisemblance associée avec $1 < K^\circ < K$ et $1 < p^\circ < p$. Sous cette hypothèse, les colonnes de X peuvent être réorganisées afin que $X^\circ \beta_k^\circ = X \beta_k^*$ avec $\beta_k^* = ((\beta_k^\circ)', (\beta_k^1)')'$ et β_k^1 est un vecteur nul d'ordre $(p - p^\circ) \times m$ [3].

2 Identification d'un modèle de mélange de régression multivariée

2.1 Estimation des paramètres

On applique la méthode itérative de l'algorithme EM (Expectation-Maximization)[1] pour estimer le vecteur paramètre $\phi = (\alpha_k, \beta_k, \Sigma_k), k = 1, ..., K$. L'estimateur $\phi^{(q+1)} = (\alpha_k^{q+1}, \beta_k^{q+1}, \Sigma_k^{q+1})$ du vecteur ϕ à l'itération $(q + 1)$ est donné par :

1. $\alpha_k^{(q+1)} = \sum_{j=1}^{n} \frac{\tau_{jk}^{(q)}}{n}$
2. $\beta_k^{(q+1)} = (\tilde{X}_k^{(q)'} \tilde{X}_k^{(q)})^{-1} \tilde{X}_k^{(q)'} \tilde{Y}_k^{(q)}$
3. $\Sigma_k^{(q+1)} = \tilde{Y}_k^{(q)'} (I - \tilde{H}_k^{(q)}) \tilde{Y}_k^{(q)} / tr(W_k^{(q)})$

où $\tau_{jk}^{(q)} = E(z_{jk}/y_j) = P(z_{jk} = 1/y_j$, $W_k^{(q)} = diag(\tau_k^{(q)})$, $\tau_k^{(q)} = (\tau_{1k}^{(q)}, ..., \tau_{nk}^{(q)})'$
$\tilde{Y}_k^{(q)} = (W_k^{(q)})^{1/2} Y$, $\tilde{X}_k^{(q)} = (W_k^{(q)})^{1/2} X$ et $\tilde{H}_k^{(q)} = \tilde{X}_k^{(q)} (\tilde{X}_k^{(q)'} \tilde{X}_k^{(q)})^{-1} \tilde{X}_k^{(q)'}$.
Les itérations successives du vecteur paramètre $\phi^{(q)}$ convergent vers l'estimateur du maximum de vraisemblance $\hat{\phi}$.

2.2 Dérivation du critère MRC_{sd} dans le cas multivarié

Pour identifier un modèle de mélange de régression multivariée, nous proposons une généralisation du critère MRC_{sd} proposé par Hafidi et al. en 2010 [4] dans le cas de mélange de régression univariée. Le modèle de paramètre inconnu $\theta = (\hat{\phi}, \hat{\tau})$, appelé modèle candidat, a été ajusté en utilisant l'échantillon observé Y et les paramètres estimés $\hat{\theta} = (\hat{\phi}, \hat{\tau})$ par l'algorithme EM. Soit $Y^* = (y_1^*, ..., y_n^*)'$ un échantillon de prédiction à partir du vrai modèle et indépendant de Y. La divergence de Kullack-Liebler symétrique entre le vrai modèle et le modèle candidat est définie par :

$$\begin{aligned} J(\theta^\circ, \theta) &= I(\theta^\circ, \theta) + I(\theta, \theta^\circ) \\ &= E_{Y^*/\theta^\circ}\{L^\circ(\phi^\circ, Z^*, Y^*, X) - L(\phi, \tau, Y^*, X)\} \\ &+ E_{Y^*/\theta}\{L(\phi, \tau, Y^*, X) - L^0(\phi^\circ, Z^*, Y^*, X)\} \end{aligned}$$

où Z^* est une matrice d'ordre $n \times K^\circ$ telle que z_{jk}^* vaut 1 lorsque y_j^* provient de la k$^{\text{ième}}$ composante du mélange et 0 sinon, et E_{Y^*/θ° l'espérance conditionnelle sachant $\theta^\circ = (\phi^\circ, \tau^\circ)$.
En ne considérant que les termes dépendant du modèle candidat, la divergence symétrique devient :
$K(\theta^\circ, \theta) = E_{Y^*/\theta^\circ}\{-L(\phi, \tau, Y^*, X)\} + E_{Y^*/\theta}\{L(\phi, \tau, Y^*, X) - L^\circ(\phi^\circ, Z^*, Y^*, X)\}$

Proposition :

Sous des conditions de régularité, un estimateur de $\Omega = E_{Y/\theta^0}\{2K(\theta^0, \hat{\theta})\}$ est :

$$MRC_{sd} = \sum_{k=1}^{K} \hat{n}_k log(|\hat{\Sigma}_k|) - 2\sum_{k=1}^{K} \hat{n}_k log(\hat{\alpha}_k) + \sum_{k=1}^{K} \hat{d}_k m(p_k + \hat{n}_k) + \sum_{k=1}^{K} (mp_k + \frac{m(m+1)}{2})$$

avec $\hat{\theta} = (\hat{\phi}, \hat{\tau})$, $\hat{n}_k = tr(\hat{W}_k)$, $\hat{d}_k = \frac{\hat{n}_k}{\hat{n}_k - (m + p_k + 1)}$ et $p_k = tr(\hat{H}_k)$

Nous considérons le critère MRC_{sd} pour identifier un modèle de mélange de régression multivariée, l'ordre estimé étant obtenu par une minimisation de MRC_{sd}.

Remarque :

– Si $m = 1$, on a la formulation de Hafidi et al.[4] :

$$MRC_{sd} = \sum_{k=1}^{K} \hat{n}_k \log(\hat{\sigma}_k^2) + \sum_{k=1}^{K} \frac{\hat{n}_k(\hat{n}_k + p_k)}{(\hat{n}_k - p_k - 2)} - 2\sum_{k=1}^{K} \hat{n}_k \log(\hat{\alpha}_k) + \sum_{k=1}^{K} (p_k + 1)$$

– Si $K = 1$, on a la formulation de Hafidi et al.[3] :

$$MRC_{sd} = n(\log(|\hat{\Sigma}|) + m) + \frac{d(3n - m - p - 1)}{n - m - p - 1} = KIC_c$$

3 Résultats numériques

Pour effectuer les tests numériques, nous simulons 500 échantillons de taille $n = 30$ et $n = 300$ du modèle de mélange de régression multivariée de trois composantes $K^\circ = 3$, avec :

$$Y_k = X_k^\circ \beta_k^\circ + \epsilon_k^\circ$$

$\epsilon_k^\circ \sim \aleph(0, I_3)$ et $k = 1, 2, 3$

$$\beta_1^\circ = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \beta_2^\circ = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \\ 4 & 4 \end{pmatrix}, \beta_3^\circ = \begin{pmatrix} 5 & 5 \\ 6 & 6 \\ 7 & 7 \\ 8 & 8 \end{pmatrix},$$

les éléments de X_1°, X_2° et X_3° sont générés respectivement des distributions uniformes $U(0, 5)$, $U(5, 10)$ et $U(10, 15)$.
La sélection par les critères MRC_{sd}, BIC et AIC est présentée dans les tableaux 1 et 2 ci-dessous.
Le tableau1 montre l'effectif de sélection de l'ordre pour $p = 2, ..., 7$ et $K = 1, ..., 5$. Le tableau2 résume les effectifs de bonne sélection obtenus par les critères MRC_{sd}, BIC et AIC pour des échantillons de tailles $n = 30$ et $n = 300$.
Nous constatons que la fréquence de bonne sélection par le critère MRC_{sd} est supérieure à celle qu'on obtient avec les critères BIC et AIC pour les échantillons de petite taille ($n = 30$). Pour une taille assez grande ($n = 300$), les critères MRC_{sd} et BIC fournissent le même résultat avec un pourcentage de sélection de l'ordre exact de $99,6\%$, le critère AIC est moins performant ($89,4\%$ de bon sélection de l'ordre).

K	p					
	2	3	p°=4	5	6	7
1	0	0	0	0	0	0
2	0	0	0	0	0	0
$K^\circ = 3$	0	0	498	2	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

Tableau1 : Critère MRC_{sd}, $n = 300$

Critères	n=30	n=300
MRC_{sd}	451	498
AIC	51	447
BIC	319	498

Tableau2 : Comparaison des critères MRC_{sd}, BIC et AIC

4 Conclusion

Nous avons considéré dans ce travail la sélection d'un modèle de mélange multivarié par le critère MRC_{sd}, c'est une généralisation du critère qui a été développé par Hafidi et al. dans le cas univarié. Pour la construction du nouveau critère, nous avons considéré la divergence symétrique entre le vrai modèle et le modèle candidat. Les résultats de simulation montrent que le critère proposé permet d'améliorer la sélection de l'ordre, principalement pour les petits échantillons, par rapport aux critères BIC et AIC.

5 Bibliographie

[1] Dempster, A.P, Laird, N.M and Rubin, D.P. (1977) Maximum likelihood from incomplete data via the EM algorithm.*Journal of the Royal Statistical Society Series B*, 39,1–38.

[2] Edward J. Bedrick and Chih-Ling Tsai. (1994) Model Selection for Multivariate Regression in small Samples ,*Biometrics*, 50, 226–231.

[3] Hafidi, B. Mkhadri,A. (2006) A corrected Akaike criterion based on KullbacK symmetric divergence : Applications in times series, multiple and multivariate regression. Computationnal Statistic and Data Analysis 50 (6), 1524-1550.

[4] Hafidi, B. Mkhadri,A. (2010) The Kullback information criterion for mixture regression models. Statistics and Probability Letters, 80(9-10) : 807-815.

[5] Muirhead,R.J.(1982) Aspects of Multivariate Statistical Theory. Wiley, New York.