Modèles adaptatifs pour les mélanges de régressions

Charles Bouveyron, Julien Jacques

Modèles adaptatifs pour les mélanges de régressions

Charles Bouveyron[†] & Julien Jacques[‡]

[†] *SAMOS-MATISSE, CES, Université Paris 1 (Panthéon-Sorbonne), Paris*
[‡] *Laboratoire Paul Painlevé, UMR CNRS 8524, Université Lille 1, Villeneuve d'Ascq*

Résumé : Partant de l'estimation d'un modèle de mélange de régressions sur des données prélevées dans une situation donnée, ce travail montre comment il est possible de transférer cette information vers une nouvelle situation. Pour ce faire, des modèles parcimonieux de transformation sont mis en évidence entre les deux modèles de mélange de régressions. L'estimation des ces transformations *via* l'algorithme EM permet alors de déduire par *plug-in* les paramètres du nouveau mélange de régression. Cette stratégie s'avère être très efficace quand on ne dispose que de peu de données pour modéliser la nouvelle situation.

Abstract : Based on the estimation of a regression mixture model on data collected in a given situation, this work shows how it is possible to transfer this information to a new situation. For this, parsimonious transformation models are proposed between both regression mixture models. The estimation of the new regression mixture model can then be inferred by plug-in from the estimation of the transformation parameters obtained using the EM algorithm. The interest of such an approach is that the transformation model is more pasimonious than the whole regression mixture model.

Mots-Clés : mélange de régressions, modèles adaptatifs, algorithme EM.

1 Introduction

Lorsqu'une société à recours à des modèles statistiques, le principal coût est celui engendré par la collecte de données, nécessaire à l'estimation du modèle. De plus, lorsqu'un un modèle donné est utilisé dans un objectif particulier, ses paramètres devront être adapté à chaque situation d'utilisation. Considérons par exemple le cas d'une compagnie immobilière utilisant un modèle de régression pour estimer le prix de vente de ses logements : les paramètres du modèle devront être estimés pour chaque ville dans laquelle la compagnie est implantée. En effet il est raisonable de penser que le prix des logements à Paris n'est pas régi par le même modèle de régression qu'en province. Le coût engendré par la collecte des données est alors mutliplié par autant de situations dans lesquelles le modèle souhaite être appliqué.

Ce travail se base sur l'hypothèse suivante : même si les modèles ne sont pas exactement les mêmes dans les deux situations, ils ne sont néanmoins pas totalement

indépendants et il est possible de transférer l'information d'une situation à l'autre. En supposant qu'un modèle de référence soit bien identifié et connu dans une situation donnée, nous proposons d'estimer les modèles de la nouvelle situation comme une transformation de ce modèle de référence. Ainsi, la connaissance sur la situation de référence sera utilisée pour la nouvelle situation, et le nombre données qu'il sera nécessaire de collecter en sera bien souvent très réduit. Le transfert de connaissance d'une situation vers une autre que l'on vient de décrire a déjà été étudié dans différentes situations : en classification de données continues par Biernacki *et al.* (2002), binaires par Jacques et Biernacki (2007), mais également pour des modèles linéaires de régression par Bouveyron et Jacques (2008). Chacun de ces travaux a donné des résultats très prometteurs, c'est pourquoi nous proposons de les étendre au cas des modèles de mélange de régressions afin de pouvoir être utilisés dans des situations complexes.

2 Les mélanges de régressions

Soit y une quantité d'intérêt observée conjointement à un ensemble de covariables $x = (x_0, x_1, \ldots, x_p)$, dans lequel nous supposons $x_0 = 1$. Dans bien des domaines comme la chimiométrie ou l'économétrie, les covariables x sont hétérogènes, et il est donc très difficile d'exhiber une relation entre y et x. Pour résoudre ce problème, les modèles de mélange de régressions ont été introduits, connus sous le nom de *switching regression* en économétrie (Goldfeld et Quandt (1973), Hurn *et al.* (2003)). Un modèle de mélange de régressions consiste à écrire y en fonction des covariables x de la façon suivante :

$$y = x^t \beta_k + \sigma_k \epsilon \tag{1}$$

où ϵ est centré réduit, généralement supposé gaussien, et $(\beta_k, \sigma_k) \in \mathbb{R}^{p+1} \times \mathbb{R}$ pouvant prendre ses valeurs parmi un ensemble de K valeurs avec probabilité π_k $(1 \leq k \leq K)$.
La loi de y conditionnellement à x est donnée par :

$$y|x \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(x^t \beta_k, \sigma_k^2).$$

Soit $S = (y_i, x_{1,i}, \ldots, x_{p,i})_{1 \leq i \leq n}$ un échantillon observé sur une population d'intérêt P. L'estimation des paramètres du modèle (1) à partir de l'échantillon S est généralement réalisée par maximum de vraisemblance. Ce dernier ne pouvant pas être calculé directement puisque l'on ne sait pas de quelle composante du mélange provient chacune des observations, l'algorithme iteratif EM est employé (McLachlan et Krishnan (1997)).

3 Formulation du problème

La situation que nous proposons d'étudier est la suivante : le modèle (1) a été estimé pour la population étudiée P à partir de l'échantillon S, dont la taille est supposée suffisament

grande pour que l'on ait confiance en cette estimation. Supposons que l'on cherche à étudier une nouvelle population P^*, mesurée sur les mêmes variables, mais pour laquelle nous ne disposons que d'un petit échantillon d'observations, de taille insuffisante pour permettre d'estimer avec confiance les paramètres du modèle de mélange de régressions :

$$y^*|x^* \sim \sum_{k=1}^{K} \pi_k^* \mathcal{N}(x^{*t}\beta_k^*, \sigma_k^{*2}). \tag{2}$$

Les populations P et P^* n'étant pas statistiquement les mêmes, le modèle estimé à partir des données de P ne conviendra probablement pas pour P^*. Néanmoins il existe certainement un lien entre ceux deux populations puisque ce sont les mêmes variables qui sont mesurées, mais dans une situation différente. Il devrait donc être intéressant d'utiliser l'information connue sur P pour estimer le modèle de mélange de régressions (2) sur P^*.

4 Modèles de lien entre les deux populations

Pour cela, il faut définir un lien entre ces deux modèles de mélange de régressions. En suivant les travaux de Bouveyron et Jacques (2008) dans le cas des modèles linéaires de régression, nous supposons que le lien entre les deux modèles (1) et (2) oppère au niveau des paramètres de régression de la façon suivante :

$$\beta_k^* = \Lambda_k \beta_k \qquad \forall 1 \leq k \leq K \tag{3}$$

où Λ_k est une matrice $(p+1) \times (p+1)$. Nous supposons de plus que les liens entre paramètres de régression se font uniquement variable par variable, ce qui revient à supposer les matrices Λ_k diagonales. Le nombre de paramètres à estimer est ainsi $K(p+1)$, et estimer le lien entre les deux modèles de régressions est alors équivalent à estimer directement le modèle (2). L'introduction d'hypothèses supplémentaires sur ce lien permet de définir des modèles parcimonieux :

- M_1 : $\Lambda_k = I_d$ les deux populations sont identiques,
- M_2 : $\Lambda_k = \lambda I_d$ le lien entre les populations est indépendant des variables et des composantes du mélange,
- M_3 : $\Lambda_k = \lambda_k I_d$ le lien entre les populations est indépendant des variables,
- M_4 : $\Lambda_k = \Lambda$ le lien entre les populations est indépendant des composantes du mélange,
- M_5 : Λ_k non contraint, ce qui revient à ne pas utiliser les données sur la population de référence P.

modèle	M_1	M_2	M_3	M_4	M_5
nb. paramètres	0	1	K	p	Kp

Table 1: Complexité (en nombre de paramètres) des modèles de lien entre populations

Le nombre de paramètres de ces modèles est présenté dans la Table 1. De plus, nous différencions les cas où les proportions π_k des composantes du mélange de régressions sont identiques ou différentes entre les deux populations, ce qui conduit à définir 10 modèles de lien. Le nombre de paramètres à estimer pour les modèles à proportions différentes est alors obtenu en ajoutant $K-1$ aux nombres de la Table 1.

5 Estimation et choix de modèles

La situation considérée dans ce papier suppose que les paramètres $\theta = (\beta_k, \sigma_k, \pi_k)_{1\leq k\leq K}$ du premier mélange de régression sont connus (ils sont estimés via l'algorithme EM en pratique). L'estimation des paramètres du nouveau mélange de régression θ^* se fait alors en deux étapes :

- estimation des paramètres de liens Λ_k via l'algorithme EM,
- estimation de θ^* par *plug-in* en injectant les estimations de Λ_k dans l'équation (3).

Les critères de choix de modèles BIC et AIC sont utilisés pour sélectionner le modèle de lien conduisant au meilleur modèle de mélange de régressions pour la population cible P^*.

Bibliographie

[1] Bouveyron, C. and Jacques, J. (2008) *Adaptive linear models in regression for the modeling of housing market in different U.S. cities*, Computational Methods for Modelling and Learning in Social and Human Sciences (MASHS2008), Créteil, France.

[2] Biernacki, C. and Beninel, F. and Bretagnolle, V. (2002) *A generalized discriminant rule when training population and test population differ on their descriptive parameters*, Biometrics, 58(2), 387–397.

[3] J.Jacques and C.Biernacki (2007) *Classement de données binaires lorsque les populations d'apprentissage et de test sont différentes*, Revue des Nouvelles Technologies de l'Information, Data Mining et apprentissage statistique : application en assurance, banque et marketing, A1, 109–130.

[4] Goldfeld, M. and Quandt, R.E. (1973) *A markov model for switching regressions*, Journal of Econometrics, 1, 3–16.

[5] Hurn, M. and Justel, A. and Robert, C.P. (2003)*Estimating mixtures of regressions*, J. Comput. Graph. Statist., 12(1), 55–7.

[6] McLachlan, G.J. and Krishnan, T. (1997) *The EM algorithm and extensions*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, New York.