Modèles linéaires généralisés à facteurs: une estimation par algorithme EM local

Xavier Bry, Christian Lavergne, Mohamed Saidane

Modèles linéaires généralisés à facteurs: une estimation par algorithme EM local

Xavier Bry[†]; Christian Lavergne[†‡] & Mohamed Saidane[§]

[†] *I3M Université Montpellier II - CC 051 - Place Eugène Bataillon*
34095 Montpellier Cedex 5, France
[‡] *Université Paul Valéry, Montpellier III - route de Mende*
34199 Montpellier Cedex 5, France
[§] *Université du 7 Novembre à Carthage ISCC de Bizerte*
Zarzouna 7021 - Bizerte - Tunisie
[bry, lavergne]@math.univ-montp2.fr
Mohamed.Saidane@isg.rnu.tn

Résumé: Les modèles à facteurs ont été développés et étudiés dans le cas où les observations sont supposées être de loi normale. Nous considérons ici le contexte plus large où les observations sont supposées suivre une loi de la famille exponentielle. On obtient ainsi une nouvelle classe de modèles à facteurs: les modèles linéaires généralisés à facteurs (GLFM). Les GLFM permettent la modélisation multivariée de données discrètes (binômiale, Poisson...), mais aussi de certaines données continues non normales (gamma, par exemple). Ils permettent notamment l'étude conjointe de variables de types différents, supposées dépendre de facteurs communs. Les GLFM sont, formellement, une synthèse des GLM et des modèles factoriels standards. Nous proposons une méthode d'estimation des paramètres et des facteurs de ces modèles, en combinant l'algorithme des scores de Fisher pour les GLM avec un algorithme itératif de type EM. Nous étudions les performances de cette méthode en l'appliquant sur des données simulées.

Mots clés: Modèles linéaires généralisés, Modèles à facteurs, Algorithme des scores, Algorithme EM, Simulations.

Abstract: Factor models have been developed and studied in the case where observations are assumed to be normally distributed. Here, we consider the less restrictive framework in which the distribution of the observations is assumed to belong to the exponential family. Thus, we introduce a new class of factor models: Generalized Linear Factor Models (GLFM). These allow multivariate modeling of discrete data (Binomial, Poisson...), but also of non-normal continuous data (gamma, for instance). GLFM's are built up combining standard Factor Models with Generalized Linear Models (GLM). We propose to estimate the parameters and factors of GLFM's by combining Fisher's Score algorithm for GLM with an EM type iterative algorithm. We study the performance of our algorithm on simulated data.

Keywords: Factor Models; Generalized Linear Models; EM Algorithm; Scores Algorithm; Simulations.

1 Introduction

Nous nous plaçons dans le contexte des modèles à facteurs (FM): q variables aléatoires observées $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q\}$ sont supposées être engendrées par un plus petit nombre ($k < q$) de variables non observées (latentes) $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_k\}$ appelées facteurs. Jusqu'ici, la plupart des développements sur les FM étaient restreints par une hypothèse de normalité sur $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_k\}$, cette normalité étant utilisée dans l'estimation par algorithme EM. Nous nous proposons ici d'étendre les FM à toute loi appartenant à la famille exponentielle: binomiale, gamma, Poisson, etc. Pour cela, nous devons aussi assurer le pont avec le contexte des modèles linéaires généralisés (GLM), dans lequel les variables explicatives sont observées.

Jusqu'ici, les FM et les GLM ont été formulés et étudiés indépendamment. Nous proposons une classe de modèles, les modèles linéaires généralisés à facteurs (GLFM) dans lesquels chaque variable $\mathbf{y}_i$ suit, conditionnellement aux facteurs $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_k\}$, un GLM. Pour assurer l'identifiabilité, les facteurs sont pris décorrélés et de loi normale centrée réduite. En outre, nous supposons que, conditionnellement aux facteurs $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_k\}$ les variables $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q\}$ sont indépendantes. Le problème posé par l'estimation des GLFM est que l'algorithme EM - qui utilise le calcul analytique de l'espérance de la log-vraisemblance complétée des paramètres conditionnellement observations - ne s'étend pas directement aux $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q\}$ non gaussiennes. Pour contourner cette difficulté, nous considérons l'algorithme d'estimation des GLM qui linéarise itérativement le modèle et procède aux moindres carrés généralisés sur le modèle linéarisé. Nous proposons d'appliquer l'algorithme EM sur le GLM linéarisé à chaque étape, i.e. "localement".

2 Structure générale d'un modèle linéaire généralisé à facteurs:

2.1 Modèle de la variable dépendante Y conditionnellement aux facteurs

Considérons n observations $\{1, ..., t, ..., n\}$. On notera respectivement $\mathbf{y}_t = (y_{it})_{i=1,q}$ et $\mathbf{f}_t = (f_{jt})_{j=1,k}$ le vecteur des variables observées $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q\}$ et celui des facteurs latents $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_k\}$ pour l'observation t. Conditionnellement aux facteurs $\mathbf{f}_t$, $(y_{it})_{i=1,q}$ sont indépendamment distribuées selon un modèle ayant une structure exponentielle (Nelder et Wedderburn, 1972):

$$\ell_i(y_{it}|\delta_{it}, \phi) = \exp\left\{\frac{(y_{it}\delta_{it} - b_i(\delta_{it}))}{a_{it}(\phi)} + c_i(y_{it}, \phi)\right\}$$

On rappelle des résultats classiques concernant cette structure:

$$\mu_{it} = \mathbb{E}(y_{it}) = b_i'(\delta_{it}); \quad Var(y_{it}) = a_{it}(\phi) b_i''(\delta_{it}) = a_{it}(\phi) b_i'' \left[b_i'^{-1}(\mu_{it}) \right]$$

Notons $\nu_i = b_i'' \left[b_i'^{-1}(\mu_{it}) \right]$. L'indépendance de $(y_{it})_{i=1,q}$ conditionnellement à $\mathbf{f}_t$ implique leur variance conditionnelle:

$$Var(\mathbf{y}_i) = diag \left\{ a_{it}(\phi) v_i(\mu_{it}) \right\}_{t=1,...,n}$$

2.2 Prédicteurs linéaires

En empilant les vecteurs $\mathbf{f}_t$, on obtient la matrice (n, k) des facteurs: $\mathcal{F} = [\mathbf{f}_1, ..., \mathbf{f}_t, ..., \mathbf{f}_n]'$. Sous-jacentes aux variables $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q\}$, on considère qu'il se trouve respectivement q combinaisons des facteurs, $\{\eta_1, \eta_2, ..., \eta_q\}$, que nous appelons prédicteurs linéaires.

Soit $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)'$ le vecteur des effets fixes. De façon générale, ce vecteur peut dépendre de covariables, mais pour simplifier nos développements, nous les prenons ici constants. Soit, pour tout i: $\widetilde{\boldsymbol{\theta}}_i = \theta_i \mathbf{1}_n$. Alors, le prédicteur linéaire de $\mathbf{y}_i$ conditionnellement à $\mathcal{F}$ peut s'écrire comme vecteur de $\mathbb{R}^n$: $\eta_i = \widetilde{\boldsymbol{\theta}}_i + \mathcal{F} a_i$, où a_i est un vecteur de k coefficients. Notons $A = (a_1, ..., a_q)'$ la matrice (q, k) des coefficients. On peut écrire matriciellement: $\eta = \boldsymbol{\theta} \mathbf{1}_n' + A\mathcal{F}'$. La colonne t correspond à l'observation t: $\eta_t = \boldsymbol{\theta} + A\mathbf{f}_t$. L'hypothèse de distribution des facteurs est telle que: $\forall t,\ \mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$.

2.3 Fonction de lien

Le prédicteur linéaire et l'espérance conditionnelle de la variable dépendante sont liés par une fonction g_i appelée *fonction de lien*: $\forall i, t:\ \eta_{it} = g_i(\mu_{it})$. Parmi toutes les fonctions de lien envisageables, celle qui permet d'égaler le prédicteur au paramètre canonique est appelée *lien canonique*. Comme: $\mu_{it} = b_i'(\delta_{it}) \ \Rightarrow\ \eta_{it} = g_i(b_i'(\delta_{it}))$. La fonction de lien canonique est: $g_i = b_i'^{-1}$.

3 Estimation d'un GLFM

Attendu que, conditionnellement aux facteurs, le GLFM se réduit à un GLM, on rappelle d'abord la structure d'un algorithme d'estimation des GLM, ce qui permet d'introduire quelques notations. Dans un second temps, on rend aux facteurs leur caractère aléatoire, et on adapte la procédure à la situation en incluant à son itération courante une étape EM.

3.1 Estimation d'un GLM

Soit le GLM d'une variable y, avec $\mu = \mathbb{E}(y)$. Soit $X = (x_1, ..., x_t, ..., x_n)'$ la matrice (n, k) des variables explicatives observées. Soient enfin g la fonction de lien et η le prédicteur linéaire: $\eta = X\beta$, $\beta \in \mathbb{R}^k$. Pour chaque observation t, on a:

$$\eta_t = g(\mu_t) \;\Rightarrow\; x_t'\beta = g\left(b'(\delta_t)\right)$$

Le problème est d'estimer β. La log-vraisemblance du modèle est:

$$\mathcal{L}(\delta; y) = \sum_{t=1}^{n} \mathcal{L}_t(\delta_t; y_t) = \sum_{t=1}^{n} \left[\frac{y_t\delta_t - b(\delta_t)}{a_t(\phi)} + c(y_t, \phi) \right]$$

Soit: $W_\beta = diag\left[g'(\mu_t)^2 V(y_t)\right]_{t=1,n} = diag\left[g'(\mu_t)^2 a_t(\phi) v(\mu_t)\right]_{t=1,n}$ et

$$\frac{\partial \eta}{\partial \mu} = diag\left(\frac{\partial \eta_t}{\partial \mu_t}\right)_{t=1,n} = diag\left(g'(\mu_t)\right)_{t=1,n}$$

Les équations de vraisemblance s'écrivent alors:

$$X' W_\beta^{-1} \frac{\partial \eta}{\partial \mu}(y - \mu) = 0 \tag{1}$$

Ce système d'équations n'étant pas linéaire en β, on le résout itérativement, à l'aide de *l'algorithme des scores de Fisher*. En notant $m^{[e]}$ la valeur courante de l'élément m en sortie de l'itération k, on pose:

$$\beta^{[e+1]} = \left(X' W_{\beta^{[e]}}^{-1} X\right)^{-1} X' W_{\beta^{[e]}}^{-1} z^{[e]} \tag{2}$$

où $z^{[e]} = X\beta^{[e]} + \left(\frac{\partial \eta}{\partial \mu}\right)^{[e]} \left(y - \mu^{[e]}\right)$. Cet algorithme peut aussi être interprété comme suit. On pose:

$$z_\beta = \eta + \frac{\partial \eta}{\partial \mu}(y - \mu) = X\beta + \frac{\partial \eta}{\partial \mu}(y - \mu) \tag{3}$$

Alors, (1) s'écrit:

$$X' W_\beta^{-1}(z_\beta - X\beta) = 0 \tag{4}$$

Les équations (4), en considérant z_β fixé, peuvent être interprétées comme les équations normales issues des moindres carrés généralisés (MCG) sur le modèle linéaire suivant:

$$\mathcal{M} \quad : \qquad z_\beta = X\beta + \zeta\,, \quad \text{où : } \mathbb{E}(\zeta) = 0\ ;\ V(\zeta) = W_\beta$$
$$\text{(en effet: } V(\zeta_t) = V(z_{\beta,t}) = g'(\mu_t)^2 Var(y_t))$$

Ainsi, l'itération courante e de l'algorithme d'estimation consiste à résoudre $X'W^{-1}_{\beta^{[e]}}(z_{\beta^{[e]}} - X\beta) = 0$ en β, puis à mettre à jour β dans W_β et z_β. Le modèle: $\mathcal{M}^{[e]} : z_{\beta^{[e]}} = X\beta + \zeta^{[e]}$; $\mathbb{E}(\zeta^{[e]}) = 0$; $V(\zeta^{[e]}) = W_{\beta^{[e]}}$ sera appelé *modèle linéarisé* courant. L'estimation MCG de ce modèle n'est autre qu'une estimation par quasivraisemblance (QV). La maximisation de la QV imite, à chaque étape, celle d'une vraisemblance des $z_{\beta,t}$, sous contrainte de normalité, d'indépendance et de structure de covariance fixe.

3.2 Estimation du GLFM

L'estimation d'un FM classique est aisément réalisée par l'algorithme EM, qui requiert alors la normalité des variables, et maximise l'espérance, intégrée sur les facteurs et conditionnelle aux observations, de la log-vraisemblance complétée. En vertu du paragraphe précédent, cette hypothèse de normalité peut être formellement utilisée, dans l'étape e d'estimation d'un GLM, sur le modèle linéarisé courant, en tant que les MCG miment la maximisation de sa vraisemblance. Concernant un GLFM, à présent, la stratégie d'estimation que nous proposons est informellement assez directe. Elle consiste à considérer en alternance le modèle comme:

- un GLM conditionnellement à $\mathcal{F} = (\mathbf{f}_1, ..., \mathbf{f}_t, ..., \mathbf{f}_n)'$
- un FM lors de l'étape courante d'estimation de ce GLM, puisque cette étape utilise une version linéarisée du GLM.

Plus précisément, conditionnellement aux valeurs courantes de $\boldsymbol{\theta}$, A, $\mathcal{F}$, et en vertu de (3), nous introduisons la variable de travail z (pseudo-variable dépendante), qui est alors connue:

$$\begin{aligned} z_{i,\mathcal{F}} &= \widetilde{\theta}_i + \mathcal{F}a_i + \frac{\partial \eta_{i,\mathcal{F}}}{\partial \mu_{i,\mathcal{F}}}(\mathbf{y}_i - \mu_{i,\mathcal{F}}) = \widetilde{\theta}_i + \mathcal{F}a_i + g'(\mu_{i,\mathcal{F}})(\mathbf{y}_i - \mu_{i,\mathcal{F}}) \\ &\text{Soit} \quad \zeta_{i,\mathcal{F}} = g'(\mu_{i,\mathcal{F}})\varepsilon_{i,\mathcal{F}} \quad \text{où} \quad \varepsilon_{i,\mathcal{F}} = \mathbf{y}_i - \mu_{i,\mathcal{F}} \end{aligned}$$

Cette variable intermédiaire z est utilisée dans l'algorithme d'estimation suivant. Posons: $\forall\, t\ z_t = (z_{1t}, ..., z_{qt})'$, et $Z = (z_1, ..., z_t, ..., z_n)'$:

(i) Étant données Z et $V(\zeta)$, le modèle - qu'on appellera *modèle marginal linéarisé* - est:

$$\forall\, t = 1, n\ z_t = \boldsymbol{\theta} + A\mathbf{f}_t + \zeta_t$$

Il est considéré comme un FM (non-standard), et estimé via une étape EM, qui fournit $\mathcal{F}$. Comme $\mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$, on a: $V(z_t) = \boldsymbol{\Sigma} = AA' + \boldsymbol{\Psi}$avec $\boldsymbol{\Psi} = \mathbb{E}(\boldsymbol{\Psi}_t) = \mathbb{E}\left(diag(g'(\mu_{i,\mathbf{f}_t})^2 var(\varepsilon_{it}|\mathbf{f}_t))\right)_{i=1,q}$. Si g est la fonction de lien canonique, on a: $\boldsymbol{\Psi} = \mathbb{E}(\boldsymbol{\Psi}_t) = \mathbb{E}\left(diag(a_{it}(\phi)g'(\mu_{i,\mathbf{f}_t}))\right)_{i=1,q}$. La matrice $\boldsymbol{\Sigma}$ est la matrice de variance utilisée dans EM. Elle peut être calculée analytiquement pour toutes les fonctions de lien canoniques des GLM classiques.

(*ii*) Étant donnée $\mathcal{F}$, le modèle - appelé *modèle conditionnel linéarisé* - est considéré comme GLM, et ses paramètres $\boldsymbol{\theta}$ et A sont mis à jour via l'algorithme des scores de Fisher. Celui-ci utilise la matrice de variance de ζ *conditionnellement* à $\mathcal{F}$: $V(z_t|\mathbf{f}_t) = V(\zeta_t) = \boldsymbol{\Psi}_t$.

(*ii*) La matrice de variance $V(\zeta)$ et z sont alors mises à jour.

4 Résultats expérimentaux

Nous présentons des simulations d'un GLFM à deux facteurs, fondé sur une loi de Poisson ($g = \log$). Nous avons, selon ce schéma, simulé un tableau de données de dimensions $(n, q) = (400, 40)$. Le seuil de convergence a été fixé à 10^{-5}. Les valeurs initiales des paramètres pour l'étape EM ont été calculées par perturbation aléatoire des vraies valeurs. EM nécessitant par ailleurs une valeur initiale de z, nous avons utilisé l'approximation suivante:

$$\forall\, i = 1, q\,; t = 1, n \quad z_{it}^{[0]} = \log\left(\alpha y_{it} + (1-\alpha)\overline{\mathbf{y}}_i\right)\,, \quad \text{with } \alpha = 0.5$$

L'utilisation d'une valeur $\alpha < 1$ permet d'éviter les difficultés posées par les valeurs nulles dans les données. Nos tests ont mis en évidence un bon comportement de l'algorithme, tant pour l'estimation des paramètres que pour celle des facteurs. Le seuil de convergence a été atteint après 7 itérations en moyenne, et les corrélations entre les vrais facteurs et leur estimation ont été très proche de 1.

Nous présentons également des simulations impliquant des variables $\mathbf{y}_i$ de lois appartenant à des familles distinctes, mais conditionnées par les mêmes facteurs.

Bibliographie

[1] Engel B., et Keen A. (1994): A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica* **48**, 1–22.

[2] Fahrmeir, L. et Tutz, G. (1994): *Multivariate Statistical Modeling Based on Generalised Linear Models*. Springer-Verlag, New York.
[3] Lavergne C. et Trottier C. (1997): From a conditional to a marginal point of view in GL2M. *Proceedings in Good Statistical Practice*, Seeber, pp. 205–209.
[4] McCullagh, P., et Nelder, J.A. (1989): *Generalized linear models*. 2 Ed., Chapman and Hall, New York, New York, USA.
[5] McLachlan, G.J., et Krishnan, T. (2008): *The EM algorithm and extensions*. Wiley Series in Probability and Statistics. New York, NY: John Wiley & Sons, Inc.
[6] Nelder, J.A., et Wedderburn, R.W.M. (1972): Generalized linear models. *Journal of the Royal Statistical Society: Series A* **135**, 370–384.
[7] Saidane, M. et Lavergne, C. (2009): *Modelling and Forecasting Volatility Dynamics Using Quadratic GARCH-Factor Models: Empirical Evidence from International Foreign Exchange Markets* in Stock Returns: Cyclicity, Prediction and Economic Consequences (Editor: George I. Ellison), Nova Science Publishers, Inc. ISBN: 978-1-60741-458-2.