Une relaxation continue du rasoir d'Ockham pour la régression en grande dimension

Pierre-Alexandre Mattei [1], Pierre Latouche [2], Charles Bouveyron [1] & Julien Chiquet [3]

[1] *Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes,*
pierrealex.mattei@gmail.com, charles.bouveyron@parisdescartes.fr
[2] *Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne*
pierre.latouche@univ-paris1.fr
[3] *Laboratoire LaMME, UMR CNRS 8071/UEVE, USC INRA, Évry, France*
julien.chiquet@genopole.cnrs.fr

Résumé. Nous considérons le problème de la régression parcimonieuse bayésienne. Dans ce cadre, un modèle génératif est proposé dans lequel un *a priori* de type *spike-and-slab* est supposé sur le paramètre de régression en multipliant un vecteur déterministe binaire, traduisant la parcimonie du problème, avec un vecteur aléatoire gaussien. Notre principale contribution consiste en l'utilisation d'une méthode d'inférence approchée basée sur une relaxation continue simple du modèle ainsi qu'un algorithme de type *expectation-maximization*. Nous pouvons ainsi maximiser la vraisemblance marginale des données avant de sélectionner quelles variables sont pertinentes grâce au rasoir d'Ockham. Des comparaisons numériques entre notre méthode (appelée spinyReg) et d'autres procédés de régression parcimonieuse (lasso, lasso adaptatif, *stability selection* et *spike-and-slab*) sont présentées. Que ce soit sur données réelles ou simulées, l'approche choisie se révèle être particulièrement efficace, tant en performances de prédiction que de sélection. Une nouvelle base de données de régression en grande dimension est également présentée : il s'agit de prédire le nombre de visiteurs du musée d'Orsay à une certaine heure en observant l'activité des 1200 stations *Vélib'* de Paris. Dans ce cas, spinyReg permet de sélectionner efficacement quelles stations sont particulièrement liées à la fréquentation du musée. Un paquet R implémentant l'algorithme spinyReg est en cours de développement et est accessible à l'adresse `https://r-forge.r-project.org/projects/spinyreg`.

Mots-clés. Régression linéaire, parcimonie, sélection bayésienne de variables, algorithme EM, données ouvertes

Abstract. We address the problem of Bayesian variable selection for high-dimensional linear regression. We consider a generative model that uses a spike-and-slab-like prior distribution obtained by multiplying a deterministic binary vector, which traduces the sparsity of the problem, with a random Gaussian parameter vector. The originality of the work is to consider inference through relaxing the model and using a type-II log-likelihood maximization based on an EM algorithm. Model selection is performed afterwards relying on Occam's razor and on a path of models found by the EM algorithm. Numerical

comparisons between our method, called spinyReg, and state-of-the-art high-dimensional variable selection algorithms (such as lasso, adaptive lasso, stability selection or spike-and-slab procedures) are reported. Competitive variable selection results and predictive performances are achieved on both simulated and real benchmark data sets. An original regression data set involving the prediction of the number of visitors of the Orsay museum in Paris using bike-sharing system data is also introduced, illustrating the efficiency of the proposed approach. An R package implementing the spinyReg method is currently under development and is available at `https://r-forge.r-project.org/projects/spinyreg`.

Keywords. Linear regression, sparsity, Bayesian variable selection, EM algorithm, open-data

1 Introduction

Ces dernières décennies, diverses approches parcimonieuses ont permis l'établissement de méthodes statistiques adaptées au traitement des données de très grande dimension (Candès, 2014). Dans un cadre de régression linéaire, l'obtention d'un paramètre parcimonieux permet d'obtenir des résultats interprétables et de simplifier grandement les problèmes mal posés dans lesquels le nombre de prédicteurs surpasse largement la quantité d'observations.

Depuis la deuxième moitié des années 1990, l'obtention de tels paramètres a été essentiellement obtenue à l'aide de méthodes fréquentistes maximisant des critères de vraisemblance pénalisée (Tibshirani, 1996; Zou et Hastie, 2005; Meinshausen et Bühlmann, 2010), ou bien grace à des méthodes de sélection de variables bayésienne (Mitchell et Beauchamp, 1988; Liang *et al.*, 2008; Hernández-Lobato *et al.*, 2013).

Les méthodes fréquentistes étant généralement plus rapides, mais moins efficaces que les procédés bayésiens, l'approche proposée ici s'efforcera de concilier ces deux visions. Evitant ainsi l'utilisation d'algorithmes de type MCMC, nous proposons une méthode rapide et précise de régression parcimonieuse. Pour plus de détails concernant ce travail, voir la prépublication (Latouche *et al.*, 2014).

2 Un modèle génératif de régression

2.1 Le modèle

On considère le modèle de régression suivant :

$$\begin{cases} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\beta} &= \mathbf{z} \odot \mathbf{w}, \end{cases} \tag{1}$$

dans lequel $\mathbf{Y} \in \mathbb{R}^n$ correspond aux n observations de la variable réponse et $\mathbf{X} \in \mathcal{M}_{n,p}(\mathbb{R})$ est la matrice des p variables explicatives. Le vecteur résiduel $\boldsymbol{\varepsilon}$ a une loi supposée gaussienne $p(\boldsymbol{\varepsilon}|\gamma) = \mathcal{N}(\boldsymbol{\varepsilon}; 0, I_n/\gamma)$. On suppose de plus que le paramètre de régression a une loi a priori gaussienne isotrope $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}; 0, I_p/\alpha)$. De plus, $\mathbf{z} \in \{0,1\}^p$ est un vecteur déterministe binaire, dont le support correspond aux variables actives du modèle de régression. On note q le cardinal du support de $\mathbf{z}$. On note $\odot$ le produit de Hadamard entre deux vecteurs ou matrices.

2.2 Inférence

Afin d'estimer les paramètres $\mathbf{z}$, α et γ du modèle, on adopte une approche bayésienne empirique : il s'agira de maximiser la *vraisemblance marginale* (parfois appelée *évidence* ou *vraisemblance de type II*) des données :

$$p(\mathbf{Y}|\mathbf{z}, \alpha, \gamma) = \int_{\mathbb{R}^p} p(\mathbf{Y}|\mathbf{w}, \mathbf{z}, \alpha, \gamma) p(\mathbf{w}|\alpha) d\mathbf{w}. \tag{2}$$

Afin d'effectuer cette optimisation en utilisant un algorithme EM (Dempster *et al.*, 1977), nous considérons que $\mathbf{w}$ est une variable latente. Cependant, comme $\mathbf{z}$ est une variable discrète, une procédure EM classique se heurtera au caractère combinatoire de l'optimisation vis-à-vis de $\mathbf{z}$ ainsi qu'au fait que tous les résultats de convergence de l'algorithme EM concernent des espaces paramétriques continus (McLachlan et Krishnan, 2008).

2.3 Relaxation continue

Afin de pallier ces deux problèmes, nous considérons une relaxation simple du modèle en remplaçant le paramètre de modèle par un vecteur $\mathbf{z}^{\text{relâché}} \in [0,1]^p$ continu. Dans ce nouveau cadre, nous pouvons trouver l'algorithme EM suivant (en notant $\mathbf{Z} = \text{diag}(\mathbf{z}^{\text{relâché}})$) :

Etape E.

$$\mathbf{S} = (\gamma \mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z} + \alpha I_p)^{-1} \tag{3}$$

$$\mathbf{m} = \gamma \mathbf{S}\mathbf{Z}\mathbf{X}^T\mathbf{Y} \tag{4}$$

Etape M.

$$\hat{\gamma}^{-1} = \frac{1}{n}\left\{\mathbf{Y}^T\mathbf{Y} + \mathbf{z}^{\text{relâché}^T}(\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Sigma})\mathbf{z}^{\text{relâché}} - 2\mathbf{z}^{\text{relâché}^T}(\mathbf{m} \odot (\mathbf{X}^T\mathbf{Y}))\right\} \tag{5}$$

$$\hat{\alpha} = \frac{p}{\text{Tr}(\boldsymbol{\Sigma})} \tag{6}$$

$$\hat{\mathbf{z}}^{\text{relâché}} = \underset{\mathbf{u}\in[0,1]^p}{\text{argmax}} \left\{-\frac{1}{2}\mathbf{u}^T(\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Sigma})\mathbf{u} + \mathbf{u}^T(\mathbf{m} \odot (\mathbf{X}^T\mathbf{Y}))\right\} \tag{7}$$

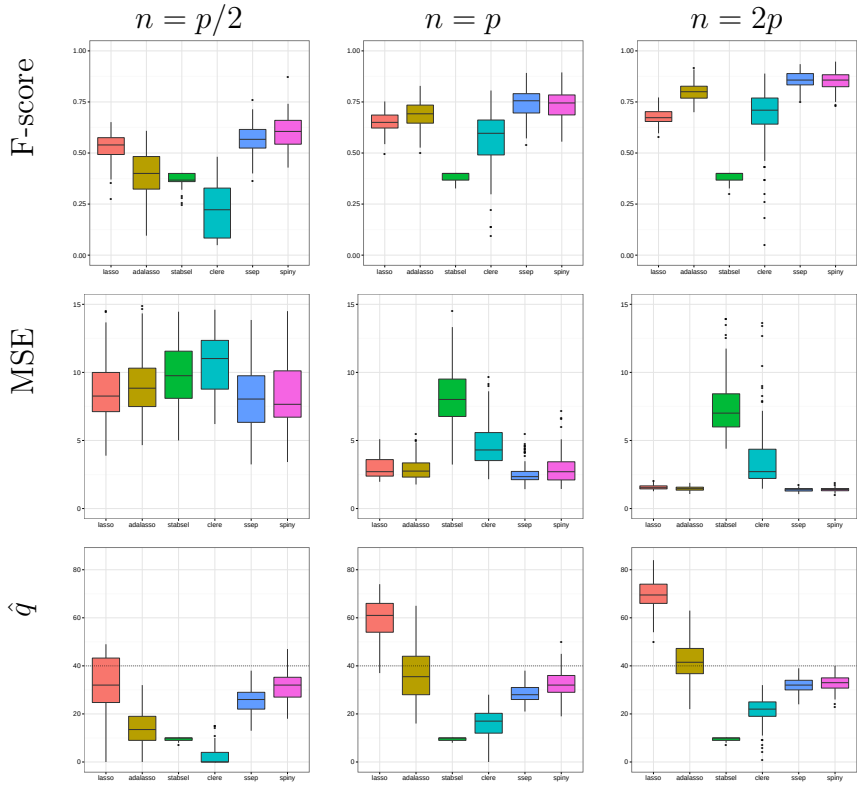

FIGURE 1 – Résultats numériques sur données simulées.

2.4 Sélection de variables

Après convergence, le vecteur $\mathbf{z}^{\text{relâché}}$ doit être binarisé afin d'obtenir un estimateur de $\mathbf{z}$. Afin d'effectuer cette binarisation, on calcule la valeur de la vraisemblance marginale évaluée en les valeurs de $\hat{\gamma}$ et $\hat{\alpha}$ trouvées par l'algorithme EM ainsi que qu'en les p valeurs de $\mathbf{z}$ obtenues en conservant uniquement les k plus grands coefficients de $\mathbf{z}^{\text{relâché}}$ pour $k \leq p$. Le vecteur $\mathbf{z}$ ayant la plus grande vraisemblance est alors choisi. Plus précisément, en notant $(\hat{\mathbf{z}}^{(k)})_{k \leq p}$ l'ensemble des vecteurs binaires obtenus en fixant à 1 exactement les coefficients correspondant aux k plus grandes valeurs de $\mathbf{z}^{\text{relâché}}$, l'estimateur du paramètre de modèle sera défini comme :

$$\hat{q} = \text{argmax}_{1 \leq k \leq p} p(\mathbf{Y} | \hat{\mathbf{z}}^{(k)}, \hat{\alpha}, \hat{\gamma}) \quad \text{and} \quad \hat{\mathbf{z}} = \hat{\mathbf{z}}^{(\hat{q})}. \tag{8}$$

3 Résultats numériques

3.1 Données simulées

On considère le schéma de simulation suivant : les données sont générées selon le modèle (1) avec $p = 100$, $n \in \{p/2, p, 2p\}$ $q = 25$, $\mathbf{w} \sim \mathcal{N}(0, I_p)$ et $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$. Chaque ligne de la matrice $\mathbf{X}$ est simulée suivant une loi gaussienne de moyenne nulle et de matrice de covariance R où $R = \text{diag}(R_1, ..., R_4)$ est une matrice diagonale par blocs telle que R_ℓ vérifie $r_{\ell ii} = 1$ et $r_{\ell ij} = 0.75$ pour $i, j = 1, \ldots, p/4$ et $i \neq j$. D'autres schémas de simulation sont disponibles dans la prépublication (Latouche *et al.*, 2014).

Notre algorithme, appelé spinyReg, est comparé à plusieurs méthodes constituant l'état de l'art en régression parcimonieuse : le lasso de Tibshirani (1996), le lasso adaptatif de Zou (2006), la *stability selection* de Meinshausen et Bühlmann (2010) ainsi que deux approches de type *spike-and-slab*, CLERE (Yengo *et al.*, 2014) et SSEP (Hernández-Lobato *et al.*, 2013). Les critères de comparaison sont le F-score (moyenne harmonique de la précision et du rappel), l'erreur quadratique (MSE) ainsi que le nombre de variables actives estimées. Dans ce schéma de simulation, les performances de spinyReg sont très bonnes, tant en prédiction qu'en sélection.

3.2 Données réelles

Une comparaison entre les algorithmes précédents sur des données réelles est disponible dans la prépublication (Latouche *et al.*, 2014). Elle comprend notamment l'introduction d'une nouvelle base de données de régression en grande dimension ($p = 1158$, $n = 316$) permettant de prédire le nombre de visiteurs du musée d'Orsay à l'aide de l'activité des stations *Vélib'*.

Comme sur données simulées, les performances de spinyReg sont très bonnes, en particulier lorsque p est plus grand que n.

Références

Candès, E. 2014, «Mathematics of sparsity (and a few other things)», dans *Proceedings of the International Congress of Mathematicians, Seoul, South Korea.*

Dempster, A., N. Laird et D. Rubin. 1977, «Maximum likelihood for incomplete data via the em algorithm», *Journal of the Royal Statistical Society*, vol. B39, p. 1–38.

Hernández-Lobato, D., J. M. Hernández-Lobato et P. Dupont. 2013, «Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation», *The Journal of Machine Learning Research*, vol. 14, n° 1, p. 1891–1945.

Latouche, P., P.-A. Mattei, C. Bouveyron et J. Chiquet. 2014, «Combining a Relaxed EM Algorithm with Occam's Razor for Bayesian Variable Selection in High-Dimensional Regression», URL `https://hal.archives-ouvertes.fr/hal-01003395`.

Liang, F., R. Paulo, G. Molina, M. A. Clyde et J. O. Berger. 2008, «Mixtures of g-priors for bayesian variable selection», *Journal of the American Statistical Association*, vol. 103, n° 481.

McLachlan, G. et T. Krishnan. 2008, *The EM Algorithm and Extensions. Second Edition.*, John Wiley & Sons, New York.

Meinshausen, N. et P. Bühlmann. 2010, «Stability selection», *Journal of the Royal Statistical Society : Series B*, vol. 27.

Mitchell, T. J. et J. J. Beauchamp. 1988, «Bayesian variable selection in linear regression (with discussion)», *Journal of the American Statistical Association*, vol. 83, p. 1023–1036.

Tibshirani, R. 1996, «Regression shrinkage and selection via the lasso», *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)*, vol. 58, n° 1, p. 267–288.

Yengo, L., J. Jacques et C. Biernacki. 2014, «Variable clustering in high dimensional linear regression models», *Journal de la Société Française de Statistique*, vol. 155, n° 2, p. 38–56.

Zou, H. 2006, «The adaptive lasso and its oracle properties», *Journal of the American Statistical Association*, vol. 101, n° 476, p. 1418–1429.

Zou, H. et T. Hastie. 2005, «Regularization and variable selection via the elastic net», *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)*, vol. 67, n° 2, p. 301–320.