Sélection bayésienne de variables pour les modèles d'état dans le cadre de reconstructions climatiques

Ophélie Guin, Philippe Naveau

Sélection bayésienne de variables pour les modèles d'état dans le cadre de reconstructions climatiques

Ophélie Guin & Philippe Naveau

Laboratoire de Sciences du Climat et de l'Environnement, IPSL-CNRS, France

Résumé

De nombreuses variantes sur la sélection de variables pour un modèle de régression sous l'approche bayésienne ont été proposées dans la littérature ; voir Mitchell and Beauchamp (1988) and George and McCulloch (1993, 1995, 1997), Smith and Kohn (1996), George (2000), Kohn et al. (2001), Nott and Green (2004), Schneider and Corcoran (2004), Casella and Moreno (2004), Celeux et al. (2006) entre autres références. Dans ces différents articles, on considère une variable aléatoire Y et un jeu $\{x_1, x_2, ..., x_p\}$ de p régresseurs potentiels à explorer. Soit $q = 0, 1, ..., p$ et $\{i_1, i_2, ..., i_q\}$ une combinaison d'indices, il est supposé que chaque modèle de régression avec les régresseurs $\{x_{i_1}, x_{i_2}, ..., x_{i_q}\}$ est *a priori* un modèle plausible pour expliquer la variable Y. Le problème consiste donc à choisir l'un de ces modèles à partir de l'information fournie par l'échantillon $(Y, x_1, ...x_p)$.

De manière générale, on pose $X = [1_n, x_1, ..., x_p]$ une matrice $n \times p$ et β un vecteur représentant les coefficients de régression. Soit $\gamma = (\gamma_1, \gamma_2, ..., \gamma_p)^T$ un vecteur binaire, on écrit $q_\gamma = \sum_i \gamma_i$ le nombre d'éléments de γ différents de zéro. Soit X_γ une matrice $n \times q_\gamma$ obtenue en enlevant les colonnes i de X pour lesquelles $\gamma_i = 0$. De la même manière, β_γ est le sous-vecteur de β obtenu en enlevant les composantes β_i de β pour lesquelles $\gamma_i = 0$. On suppose que

$$y|\gamma, X_\gamma, \beta_\gamma, \sigma^2 \sim N(X_\gamma \beta_\gamma, \sigma^2 I).$$

Afin d'effectuer l'inférence bayésienne sur les paramètres du modèle on utilise un prior hiérarchique. Le prior pour β_γ sachant γ et σ^2 est normal,

$$\beta_\gamma|\gamma, \sigma^2 \sim N(\tilde{\beta}, c\sigma^2(X_\gamma^T X_\gamma)^{-1}).$$

Le prior de σ^2 est $p(\sigma^2) \propto \sigma^{-2}$, et pour le prior sur γ on utilise $p(\gamma) = 2^{-p}$. On s'intéresse à la distribution a posteriori de γ,

$$p(\gamma|y) \propto p(\gamma)p(y|\gamma).$$

Pour p relativement petit, on peut calculer exactement la loi a posteriori de $p(\gamma|y)$. On obtient la constante de normalisation de l'équation précédente en sommant sur toutes les valeurs possible de γ. En revanche si p est grand cela n'est pas faisable du fait du nombre de termes dans la somme. On utilise alors des algorithmes de Monte Carlo par chaîne de Markov afin identifier les probabilités a posteriori des modèles.

Dans cette présentation, nous adaptons cette méthode de sélection de variables à un modèle d'état. Les modèles d'états ont une structures très proche des modèles de régression. En effet, si l'on conserve les même notations que précédemment, la relation entre y_t et $X_t = (1, x_{t1}, ...x_{tp})$ est linéaire et est spécifiée par l'équation suivante,

$$y_t = \alpha_t X_t + v_t,$$

L'erreur v_t est supposée normale de moyenne zéro et de variance V_t. La différence essentielle entre les modèles d'état et les modèles linéaires classiques est que X a une structure temporelle :

$$X_t = \theta_t X_{t-1} + w_t,$$

avec $w_t \sim N(0, W_t)$.

On applique cette méthode à un problème de reconstruction climatique. En effet, afin de comprendre si le réchauffement climatique actuel est plus important que la variabilité climatique naturelle, il est nécessaire de d'avoir de longues séries climatiques. Seulement, des mesures directes de températures et précipitations manquent, en particulier pour les période les plus anciennes, et il est nécessaires d'utiliser des proxis climatiques afin de reconstruire des chronologies passées. Un proxy bien connu est la croissance des cernes d'arbres. Afin de comprendre les relations existantes entre ce proxy est des variables climatiques on essaye d'expliquer la croissance des cernes d'arbres avec la meilleures combinaison possible de séries de températures et de précipitations.

Mots clés : modèle d'état, sélection bayésienne de variables, modèles hiérarchiques, échantillonnage de Gibbs, cernes d'arbres, reconstruction climatique

Abstract

Many variants of the Bayesian approach to variable selection in regression models have been proposed in the litterature ; see Mitchell and Beauchamp (1988), George and McCulloch (1993, 1995, 1997), Smith and Kohn (1996), George (2000), Kohn et al. (2001), Nott and Green (2004), Schneider and Corcoran (2004), Casella and Moreno (2004), Celeux et al. (2006) for further references. In this different papers we consider a dependent random variable $Y = (y_1, y_2, ..., y_n)T$ and a set $\{x_1, x_2, ..., x_p\}$ of p potential explanatory regressors. It is assume that every regression model with regressors $\{x_{i_1}, x_{i_2}, ..., x_{i_q}\}$, where $q = 0, 1, ..., p$ and $\{i_1, i_2, ..., i_q\}$ is a combinaison of the set indices $\{1, 2, ..., p\}$ is *a priori* a plausible model to explain the variable Y. The problem consists of choosing one of the above alternative models based on the information provided by a sample $(Y, x_1, ...x_p)$.

Generally, we put $X = [1_n, x_1, ..., x_p]$ a $n \times p$ matrix and β for a vector of regression coefficients. Let $\gamma = (\gamma_1, \gamma_2, ..., \gamma_p)^T$ be a binary vector, and write $q_\gamma =$

$\sum_i \gamma_i$ for the number of nonzero elements of γ. Let X_γ be the $n \times q_\gamma$ matrix obtained by removing those columns i from X for which $\gamma_i = 0$. Similarly let β_γ be the subvector of β obtained by removing components β_i of β for which $\gamma_i = 0$. We assume that

$$y|\gamma, X_\gamma, \beta_\gamma, \sigma^2 \sim N(X_\gamma \beta_\gamma, \sigma^2 I).$$

For Bayesian inference on the model parameters we use a hierarchical prior. The prior for β_γ given γ and σ^2 is normal,

$$\beta_\gamma|\gamma, \sigma^2 \sim N(\tilde{\beta}, c\sigma^2(X_\gamma^T X_\gamma)^{-1}).$$

In many papers, the prior on σ^2 is $p(\sigma^2) \propto \sigma^{-2}$, and for the prior on γ we use $p(\gamma) = 2^{-p}$. We are interested in the posterior distribution on γ,

$$p(\gamma|y) \propto p(\gamma)p(y|\gamma).$$

For p relatively small, we can compute the posterior $p(\gamma|y)$ exactly, obtaining the normalizing constant in the previous equation by summing over all possible values of γ. For large p, this is not feasible due to the number of terms in the sum, and we use Markov chain Monte Carlo algorithms to identify high posterior probability models.

In this presentation, we adapt this type of variable selection method to state space models. State space models are closely to regression models. If we keep the same notations than before, the relation between y_t and $X_t = (1, x_{t1}, ...x_{tp})$ is linear and is specified by the *observation equation*

$$y_t = \alpha_t X_t + v_t,$$

The *observation error* v_t is assumed to be normally distributed with mean zero and a variance V_t. The essential difference between state space models and conventional linear models representation is that X have a temporal structure :

$$X_t = \theta_t X_{t-1} + w_t,$$

with $w_t \sim N(0, W_t)$.

We apply this variable selection to a climatic reconstruction problem. In order to understand if the recent climate warming is greater than its natural variability, it is necessary to construct long climatic series that reach back to pre-industrial times. But, temperatures or precipitation direct measurements are missing, especially in the old days, and proxies are necessary to reconstruct past chronologies. One of the known proxy is tree-ring growths. To understand the relations between this proxy and climatic variables we try to explain tree-ring growths with the best combination of temperatures and precipitation series.

Key words : state sapce model, bayesian variable selection, hierarchical model, Gibbs sampler, tree-ring growths, climatic reconstruction

Bibliographie

[1] Casella G. and Moreno E. (2004) Objective Variable Selection *Journal of the American Statistical Association 101(473)*, 157-167.
[2]Celeux Gilles, Marin Jean-Michel, Robert Christian P. (2006) Sélection bayésienne de variables en régression linéaire *Journal de la société française de statistique 147(1)*, 59-79.
[3] George Edward I. and McCulloch Robert E. (1993) Variable selection via Gibbs sampling *Journal of American Statistical Associaltion, Vol. 88, No. 423*, 881-889.
[4] George Edward I. and McCulloch Robert E. (1997) Approaches for bayesian variable selection *Statistica Sinica 7*, 339-373.
[5] George Edward I. (2000) Calibration and empirical Bayes variable selection *Biometrika, 87, 4*, pp 731-747.
[6] Geweke J. (1994) Variable Selection and Model Comparison in Regression *technical report.*
[7] Kohn R., Smith M. and Chan D. (2001) Nonparametric regression using linear combinations of basis functions *Statistics and Computing, 11*, 313-322.
[8] Meinhold Richard J. and Singpurwalla Nozer D. (1983) Undersanding the Kalman Filter *The American Satistician, Vol. 37, No. 2*, pp 123-127.
[9] Mitchell T.J. and Beauchamp J.J. (1988) Bayesian variable selection in linear regression *Journal of the American Satistical Association 90*, 1257-1270.
[10] Nott D.J. and Green P.J. (2004) Bayesian variable selection and the Swendsen-Wang algorithm *J. Comput. Graph. Statist., 13*, 1-17.
[11] Raftery A., Madigan D. and Hoeting J. (1993) Model selection and accounting for model uncertainty in linear regression models *Working paper.*
[12] Smith Michael and Kohn Robert (1996) Nonparametric regression using Bayesian variable selection *Journal of Econometrics 75*, 317-343.