European Journal of Operational Research 220 (2012) 414–422

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Bayesian variable selection in generalized linear models using a combination of stochastic optimization methods

D. Fouskakis *

Department of Mathematics, National Technical University of Athens, Zografou Campus, Athens 15780, Greece

ARTICLE INFO

Article history:
Received 20 July 2011
Accepted 20 January 2012
Available online 28 January 2012

Keywords:
Bayesian variable selection
Genetic algorithm
Laplace approximation
Simulated annealing
Stochastic optimization
Tabu search

ABSTRACT

In this paper the usage of a stochastic optimization algorithm as a model search tool is proposed for the Bayesian variable selection problem in generalized linear models. Combining aspects of three well known stochastic optimization algorithms, namely, simulated annealing, genetic algorithm and tabu search, a powerful model search algorithm is produced. After choosing suitable priors, the posterior model probability is used as a criterion function for the algorithm; in cases when it is not analytically tractable Laplace approximation is used. The proposed algorithm is illustrated on normal linear and logistic regression models, for simulated and real-life examples, and it is shown that, with a very low computational cost, it achieves improved performance when compared with popular MCMC algorithms, such as the MCMC model composition, as well as with "vanilla" versions of simulated annealing, genetic algorithm and tabu search.

1. Introduction

Generalized linear models (GLMs) are widely used to model the dependence of a response variable Y on a set of possible explanatory variables (or predictors) $X_1, \ldots, X_p$. They assume that the conditional distribution of $Y|(X_1,\ldots,X_p)$ belongs to the exponential family with mean μ related to the explanatory variables through the linear predictor $\eta = g(\mu)$ where $g(\cdot)$ is the link function. When there is no uncertainty about the structural properties (such as the response distribution, or the link function; for a link function selection see for example Ntzoufras et al., 2003; Hahn, 2006) of the GLM, the later can be fully described by a vector $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_p)^T \in \{0,1\}^p$, denoting which explanatory variables are present in the linear predictor. The γ_j, $j = 1, \ldots, p$, takes the value 1 if explanatory variable j is included in the model and 0 otherwise. The variable selection problem deals with the issue of choosing the appropriate $\boldsymbol{\gamma}$, i.e. a parsimonious and reasonable subset of the explanatory variables. When p is large, a computationally demanding optimization problem arises, since the size of the model space $\mathcal{M} = \{0,1\}^p$ can be enormous.

The plan of the paper is as follows. In Section 2 a brief description of how the Bayesian community deals with variable selection problems in GLMs is presented, in Section 3 three well know optimization algorithms: the simulated annealing, the genetic algorithm and the tabu search are re-visited and a new algorithm is formed by combining ideas from those techniques. In Section 4 the proposed algorithm is illustrated with various examples and Section 5 contains some final remarks. Finally, there is an Appendix with the pseudocode of the proposed algorithm, together with details concerning a data-set used for illustration.

2. Bayesian variable selection in generalized linear models

Letting $\boldsymbol{y} = (y_1,\ldots,y_n)^T$ denote the available observations for the response variable and X_{ij} the value of the explanatory variable j ($j = 1,\ldots,p$) for observation i ($i = 1,\ldots,n$). Let also X represent the $n \times p$ data matrix with elements X_{ij}. Within the Bayesian framework the identification of the "best model" (identification of the "best set of explanatory variables" in our case) between the 2^p competitors, $\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_{|\mathcal{M}|}\}$, is equivalent (assuming a zero-one loss function) to find the model with the highest posterior model probability, defined as

$$f(\boldsymbol{\gamma}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\gamma})f(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}_\ell \in \mathcal{M}} f(\boldsymbol{y}|\boldsymbol{\gamma}_\ell)f(\boldsymbol{\gamma}_\ell)}, \tag{1}$$

where $\boldsymbol{\gamma} \in \mathcal{M}$, $f(\boldsymbol{y}|\boldsymbol{\gamma})$ is the marginal likelihood under model $\boldsymbol{\gamma}$ and $f(\boldsymbol{\gamma})$ is the prior model probability of model $\boldsymbol{\gamma}$. The marginal likelihood function in the above calculation can be further expanded to include the effect of the model parameters:

$$f(\boldsymbol{y}|\boldsymbol{\gamma}) = \int f(\boldsymbol{y}|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}) f(\boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}) d\boldsymbol{\theta}_\gamma, \tag{2}$$

where $f(\boldsymbol{y}|\boldsymbol{\theta}_\gamma,\boldsymbol{\gamma})$ is the likelihood under model $\boldsymbol{\gamma}$ with parameters $\boldsymbol{\theta}_\gamma$ and $f(\boldsymbol{\theta}_\gamma|\boldsymbol{\gamma})$ is the prior distribution of model parameters given model $\boldsymbol{\gamma}$.

* Tel.: +30 210 7721702.
E-mail address: fouskakis@math.ntua.gr

0377-2217/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2012.01.040

Pairwise comparison of any two models, $\gamma_{\ell'}$ and γ_ℓ is given by the posterior odds

$$PO_{\gamma_{\ell'},\gamma_\ell} \equiv \frac{f(\gamma_{\ell'}|\boldsymbol{y})}{f(\gamma_\ell|\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\gamma_{\ell'})}{f(\boldsymbol{y}|\gamma_\ell)} \times \frac{f(\gamma_{\ell'})}{f(\gamma_\ell)}. \tag{3}$$

Closed form expression of the marginal likelihood (2), and therefore of the posterior model probability (1) is available only in special cases, such as the normal linear regression with conjugate or semi-conjugate prior for the model parameters (e.g., Marin and Robert, 2007). Therefore, a combination of Laplace approximations (e.g., Bernardo and Smith, 1994; Raftery, 1996) and Markov Chain Monte Carlo (MCMC) methodology (e.g., Green, 1995; Han and Carlin, 2001; Chipman et al., 2001; Dellaportas et al., 2002) is usually used.

One important issue in Bayesian model evaluation using posterior model probabilities is their sensitivity to the prior variance of the model parameters: large prior variance of the $\boldsymbol{\theta}_\gamma$ (used to represent prior ignorance) will increase the posterior probabilities of the simpler models considered in the model space $\mathcal{M}$ (e.g., Bartlett, 1957; Lindley, 1957). Therefore, specifying the prior distribution is pivotal for the a-posteriori support of the models examined. Here a family of prior distributions based on the Zellner's g-prior (Zellner, 1986) is used after applying ideas proposed by Ntzoufras et al. (2003). For the normal regression case, for any model γ of dimension d_γ and parameters $\boldsymbol{\theta}_\gamma = (\beta_0, \boldsymbol{\beta}_\gamma, \sigma^2)$, a prior of the form

$$\begin{aligned} &f(\boldsymbol{\beta}_\gamma|\gamma,\sigma^2) = N\left[\boldsymbol{0}, n\sigma^2\left(\mathrm{X}_\gamma^T\mathrm{X}_\gamma\right)^{-1}\right], \\ &f(\beta_0,\sigma^2|\gamma) \propto \sigma^{-2} \end{aligned} \tag{4}$$

is used (see for example Liang et al., 2008), where β_0 is an intercept that is common to all models, $\boldsymbol{\beta}_\gamma$ is the d_γ-dimensional vector of nonzero regression coefficients included in the model specified by γ, σ^2 is the error variance of any model and X_γ is the data matrix corresponding to model γ (i.e. the submatrix of X with columns corresponding to explanatory variables included in the model specified by γ). For the logistic regression case, for any model γ with parameters $\boldsymbol{\theta}_\gamma = (\beta_0, \boldsymbol{\beta}_\gamma)$, the prior of the form

$$f(\boldsymbol{\theta}_\gamma|\gamma) = N\left[\boldsymbol{0}, 4n\left(\widetilde{\mathrm{X}}_\gamma^T\widetilde{\mathrm{X}}_\gamma\right)^{-1}\right] \tag{5}$$

is used, where $\widetilde{\mathrm{X}}_\gamma = (X_{ih},\ i = 1,\ldots,n;\ h = 0,\ldots,p)$ is the design matrix corresponding to model γ, with $X_{i0} = 1$ for all $i = 1, \ldots, n$.

Regarding the prior on model space, following the work by Ley and Steel (2009) a beta-binomial hierarchical prior is used. For any model $\gamma \in \mathcal{M}$, this prior has the following form

$$\begin{aligned} &d\gamma \equiv \sum_{j=1}^{p} \gamma_j \sim Bin(p,\pi), \\ &\pi \sim Beta(\alpha,\beta). \end{aligned} \tag{6}$$

The above prior depends on two hyperparameters, (α, β). For fixed $\alpha = 1$ the above prior could be elicited in terms of the prior mean model size, m. The choice of m will then determine β through the formula $\beta = (p - m)/m$. By setting $m = p/2$ will get $\beta = 1$ and therefore a discrete uniform prior for model size is obtained. It is easy to show that the prior probability for any model γ_ℓ of size k, using the prior (6), with $\alpha = 1$, is given by:

$$f(\gamma_\ell) = \frac{\Gamma\left(1+\frac{p-m}{m}\right)}{\Gamma\left(\frac{p-m}{m}\right)} \frac{\Gamma(1+k)\Gamma\left(\frac{p-m}{m}+p-k\right)}{\Gamma\left(1+\frac{p-m}{m}+p\right)}. \tag{7}$$

2.1. Model search algorithms

The number of models under consideration is equal to 2^p and therefore when p is even moderately large, this number grows tremendously. As a result, visiting every possible competing model becomes infeasible. This motivates the need for global optimization methods, for example stochastic optimization techniques such as simulated annealing (Kirkpatrick et al., 1983), genetic algorithms (e.g., Holland, 1975), and tabu search (e.g., Glover, 1989). Alternatively, popular MCMC methods can be used, such as Markov chain Monte Carlo model composition (MC^3; Madigan and York, 1995), or RJMCMC (Green, 1995), as model search algorithms, to trace the most important models. RJMCMC is a quite general approach that can be used when the model parameters can not be integrated out and is also effective when Laplace approximation does not work well. When using MCMC methods, posterior model probabilities and posterior model odds can be directly estimated from their output, but with a realistically large number of explanatory variables, those estimates can be poor in a reasonable amount of CPU time.

The above mentioned MCMC techniques may suffer from poor mixing in high dimensional spaces, and as a result they may be unable to explore the full support of the posterior distribution. To avoid this problem, stochastic optimization algorithms have been used for variable selection problems in order to form more powerful MCMC algorithms. Liang and Wong (2000) have proposed a new MCMC algorithm, called an evolutionary Monte Carlo algorithm, that has incorporated several attractive features of genetic algorithms and simulated annealing into the framework of MCMC. Liang et al. (2001) have used evolutionary Monte Carlo to sample from the posterior distributions for a multiple linear regression setup. They have shown that sampling from the posterior distribution is approximately equivalent to sampling from a Boltzmann distribution defined on C_p values. Furthermore, Bottolo and Richardson (2010) have proposed a new sampling algorithm based upon evolutionary Monte Carlo that is designed to work under the "large p, small n" paradigm. Finally, Clyde et al. (2011) have considered sampling the model space without replacement and this in some sense implements a similar idea to tabu search by setting all previously visited models "tabu".

Furthermore, stochastic optimization algorithms have been used as model search tools in order to find "good" models. Unler and Murat (2010) have used a discrete particle swarm optimization method for feature selection in binary classification problems and Piramuthu (2004) has evaluated several inter-class as well as probabilistic distance-based feature selection methods as to their effectiveness in preprocessing input data for inducing decision trees. Chatterjee et al. (1996) have used the genetic algorithm for solving various discrete optimization problems in statistical modeling, while Tolvi (2004) has used genetic algorithm for outlier detection and variable selection in linear regression models. Mills et al. (2005) and Pacheco et al. (2009) have performed variable selection based on tabu search, while Meiri and Zahavi (2006) have used simulated annealing to optimize the variable selection problem in marketing applications. Pacheco et al. (2006) have compared a series of techniques based on metaheuristic strategies, among them a genetic local search method and a tabu search, on a variable selection problem with stepwise methods. Brusco and Steinley (2011) have used tabu search for variable selection in linear discriminant analysis. Finally, Cadima et al. (2004) have solved the combinatorial optimization problem of variable selection for three different objective functions with simulated annealing, genetic algorithm and a restricted local search.

Within the Bayesian variable selection literature, Soyer and Tanyeri (2006) have presented a simulation-based method for a Bayesian portfolio selection problem, Brooks et al. (2003) have performed model selection via simulated annealing, while Fearn et al. (2002), Draper and Fouskakis (2000) and Fouskakis and Draper (2008) have used stochastic optimization methods for a Bayesian decision theory approach to variable selection. Finally, Kapetanios