ARTICLE IN PRESS

Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Bayesian variable selection with strong heredity constraints

Joungyoun Kim [a], Johan Lim [b], Yongdai Kim [b], Woncheol Jang [b,*]

[a] Department of Information & Statistics, Chungbuk National University, Cheongju, Republic of Korea
[b] Department of Statistics, Seoul National University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 2 April 2017
Accepted 25 March 2018
Available online xxxx

AMS 2000 subject classifications:
62J05
62C10

Keywords:
Heredity principle
Interaction
Shotgun stochastic search
Strong heredity
Variable selection

ABSTRACT

In this paper, we propose a Bayesian variable selection method for linear regression models with high-order interactions. Our method automatically enforces the heredity constraint, that is, a higher order interaction term can exist in the model only if both of its parent terms are in the model. Based on the stochastic search variable selection George and McCulloch (1993), we propose a novel hierarchical prior that fully considers the heredity constraint and controls the degree of sparsity simultaneously. We develop a Markov chain Monte Carlo (MCMC) algorithm to explore the model space efficiently while accounting for the heredity constraint by modifying the shotgun stochastic search algorithm Hans et al. (2007). The performance of the new model is demonstrated through comparisons with other methods. Numerical studies on both real data analysis and simulations show that our new method tends to find relevant variable more effectively when higher order interaction terms are considered.

1. Introduction

Suppose that we observe $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ from n subjects, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ are the predictors and y_i is the response. To capture the relationship between the response and the predictors, one may consider the following linear model:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\epsilon_i \sim N(0, \sigma^2)$. One of the main interests of regression analysis is to select a subset of predictors that are relevant to the response.

When the main terms $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are not sufficient to capture the relationship between the response and the predictors, it can be helpful to add higher order interactions to the model. An interaction, as a product of a pair of predictors, can be considered when one predictor has different effects on the response depending on values of the other predictor. For example, interactions between genetic markers and environmental factors, denoted as G×E, are emphasized as one potential source of missing genetic variations for human disease risk (Khoury & Wacholder, 2009). There are numerous examples of the effects of G×Es on disease risk such as for bladder cancer and skin cancer (Green & Trichopoulos, 2002; Hung et al., 2004).

In this paper, we consider a regression model with main terms and all possible second-order interaction terms:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j<k} \alpha_{jk} x_{ij} x_{ik} + \epsilon_i. \tag{2}$$

* Corresponding author.
E-mail addresses: joungyoun@chungbuk.ac.kr (J. Kim), johanlim@snu.ac.kr (J. Lim), ydkim903@snu.ac.kr (Y. Kim), wcjang@snu.ac.kr (W. Jang).

https://doi.org/10.1016/j.jkss.2018.03.003
1226-3192/

The main goal is to identify which terms, especially which interaction terms, have an important effect on the response. By adding interaction terms to the model, we may face two challenges: (1) the increased number of terms and (2) the complexity in the relations among predictors. First, having 2nd-order interaction terms increases the number of terms significantly from p to $p + p(p-1)/2$, which can be easily larger than the sample size n. Hence, we would like to select a small subset of predictors if possible. Second, when interaction terms exist, there is a hierarchical relation among the predictors, that is, main terms are more important than interaction terms. In other words, we may include an interaction term only if one (called weak heredity) or both (called strong heredity) of the parental main terms are also included in the model (Bien, Taylor, & Tibshirani, 2013; Wu & Hamada, 2000). In this paper, we focus on variable selection in applications where such a hierarchical relation is desired.

Traditional variable selection methods include stepwise selection and criterion-based approaches such as Mallows' C_p (Mallows, 1973). However, methods based on criteria are unstable when the number of predictors is large (Miller, 2002). Alternatively, penalized approaches, such as LASSO, have been studied extensively (Breiman, 1995; Fan & Li, 2001; Tibshirani, 1996). However, penalized methods do not necessarily obey the heredity constraint. There have been efforts to extend LASSO for hierarchical variable selection (Bien et al., 2013; Choi, Li, & Zhu, 2010); however, these methods are not easy to implement, and they may not be consistent of the model choice in certain situations (Meinshausen & Buhlmann, 2006; Zou, 2006).

In this paper, we propose a Bayesian variable selection method enforcing heredity constraints, herein focusing on strong heredity, based on the stochastic search variable selection (SSVS), proposed by George and McCulloch (1993). We propose a novel hierarchical prior for SSVS, which fully considers the strong heredity and controls the degree of sparsity simultaneously. In addition, we develop a computational algorithm to explore the model space efficiently under the strong heredity by modifying the shotgun stochastic search (SSS) algorithm (Hans, Dobra, & West, 2007). The SSS algorithm has an advantage over other SSVS algorithms (Madigan & York, 1995; Raftery, Madigan, & Hoeting, 1997) in the sense that the SSS algorithm can be implemented through parallel processing. The proposed algorithm also has this advantage.

The remainder of this paper is as follows: In Section 2, we review SSVS and present our method and prior distributions. Section 3 extends the SSS algorithm to apply it to variable selection under the strong heredity principle. In Sections 4 and 5, we present numerical studies, including a simulation and a real data example, to show that our method outperforms existing methods. Concluding remarks are given in Section 6.

2. Bayesian models

2.1. Stochastic search variable selection for main terms

In this section, we present a short review on stochastic search variable selection (SSVS). Here, we consider the regression model with only main terms: As in (1), we have p predictors, $X_1, X_2, \ldots, X_p$, and a response $\mathbf{y}$ from n subjects, where $X_j = (x_{1,j}, \ldots, x_{n,j})$ and $\mathbf{y} = (y_1, \ldots, y_n)$. In SSVS, George and McCulloch (1993) introduced a binary latent variable γ_j to identify whether the corresponding jth predictor X_j should be included in the model and used the prior

$$\pi(\gamma_j \mid w_j) = w_j^{\gamma_j}(1-w_j)^{1-\gamma_j},$$

where w_j is the inclusion probability. Often $w_j = 1/2$ is assumed. In Bayesian inference, the posterior mean of γ_j, denoted by $\bar{\gamma}_j$, represents the posterior inclusion probability of predictor X_j in the regression models. A higher $\bar{\gamma}_j$ indicates that the covariate X_j is more important in predicting the response.

Conditioning on γ_j, the prior distribution of β_j is given as a normal mixture:

$$\beta_j|\gamma_j \sim \gamma_j N(0, \sigma^2\tau^2) + (1-\gamma_j)N(0, \sigma^2\nu^2), \text{ for } j = 1, \ldots, p,$$

where $0 < \nu^2 \ll \tau^2$. Note that σ^2 is the noise variance in (1). If $\gamma_j = 1$, β_j has a normal prior with a large variance that is sufficient to stay away from zero. In contrast, if $\gamma_j = 0$, β_j has a normal prior with a narrow peak at zero, which results in β_j being close to zero.

For the remaining parameters, we use the following improper priors

$$\pi(\sigma^2) = \frac{1}{\sigma^2}, \quad \pi(\nu^2, \tau^2 \mid \sigma^2) = \frac{1}{\sigma^2}\left(1+\frac{\nu^2}{\sigma^2}\right)^{-2} \times \frac{1}{\sigma^2}\left(1+\frac{\tau^2}{\sigma^2}\right)^{-2} \times \mathrm{I}_{\{\nu^2<\tau^2\}}$$

that are used by Scott and Berger (2006).

2.2. Stochastic search variable selection for interaction terms

Now, we consider the extended model (2), which includes main terms and interaction terms. By including interaction terms, we require additional binary variables $\gamma_{j,k}$ ($j = 1, \ldots, p-1$ and $k = 1, \ldots, p$), which indicates the inclusion state of an interaction term X_jX_k in a regression model. The role of $\gamma_{j,k}$ is the same as γ_j, indicating the importance of the covariate X_jX_k. Each model can be identified by the binary latent vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p, \gamma_{1,2}, \ldots, \gamma_{p-1,p})$. For a regression coefficient $\alpha_{j,k}$, we assume the following normal mixture prior distribution:

$$\alpha_{j,k}|\gamma_{j,k} \sim \gamma_{j,k}N(\alpha_{j,k}; 0, \sigma^2\tau^2) + (1-\gamma_{j,k})N(\alpha_{j,k}; 0, \sigma^2\nu^2)$$

with the same hyper-parameters σ^2, ν^2 and τ^2 in the prior distribution of β_j.

Please cite this article in press as: Kim, J., et al., Bayesian variable selection with strong heredity constraints. Journal of the Korean Statistical Society (2018), https://doi.org/10.1016/j.jkss.2018.03.003.