Computational Statistics and Data Analysis 103 (2016) 401–412

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

The use of random-effect models for high-dimensional variable selection problems

Sunghoon Kwon [a], Seungyoung Oh [b], Youngjo Lee [b,*]

[a] *Department of Applied Statistics, Konkuk University, Republic of Korea*

[b] *Department of Statistics, Seoul National University, Republic of Korea*

ARTICLE INFO

Article history:
Received 19 November 2015
Received in revised form 16 May 2016
Accepted 24 May 2016
Available online 8 June 2016

Keywords:
Generalized linear model
Hierarchical likelihood
High-dimension
Random effect
Unbounded penalty
Variable selection

ABSTRACT

We study the use of random-effect models for variable selection in high-dimensional generalized linear models where the number of covariates exceeds the sample size. Certain distributional assumptions on the random effects produce a penalty that is non-convex and unbounded at the origin. We introduce a unified algorithm that can be applied to various statistical models including generalized linear models. Simulation studies and data analysis are provided.

1. Introduction

There has been increasing interest in the analysis of high-dimensional data, where the number of covariate p is greater than the sample size n. Identifying relevant covariates enhances not only the interpretation but also the prediction accuracy of the fitted model. Tibshirani (1996) showed that the least absolute shrinkage and selection operator (LASSO) can simultaneously execute parameter estimation and variable selection. However, the LASSO has been criticized because it typically ends up selecting a larger size of sub-model than necessary (Radchenko and James, 2008), and cannot achieve a prediction accuracy and selection consistency simultaneously (Leng et al., 2006). Zou and Hastie (2005) proposed the elastic net that combines the LASSO and ridge penalties. However, the elastic net shares common deficiency with the LASSO in variable selection.

To overcome the deficiency problem, various penalties have been proposed. Fan and Li (2001) introduced the smoothly clipped absolute deviation (SCAD) penalty, which achieves the oracle property. They showed that the resulting estimator is asymptotically equivalent to the ordinary least square estimator that is obtained by using the relevant covariates only. The results hold for high-dimensional cases (Kim et al., 2008) and can be extended to the maximum likelihood estimation (Fan and Peng, 2004; Kwon and Kim, 2012). Various other penalties have been studied to achieve the oracle property. Zhang (2010) proposed the minimax concave penalty, which guarantees higher chance of having unique solution. Huang et al. (2008) proposed the bridge penalty but it naturally produces biased estimator. Adaptive modifications are proposed for the LASSO (Zou, 2006) and elastic net (Zou and Zhang, 2009) to achieve the oracle property. However, these methods require a weight vector that is not known yet how to compute in high-dimensional cases. For a recent advances of penalized estimation in the linear regression model see Zhang and Zhang (2012).

* Corresponding author.
E-mail address: youngjomail@gmail.com (Y. Lee).

http://dx.doi.org/10.1016/j.csda.2016.05.016
0167-9473/

Recently, Lee and Oh (2009) proposed to use a random-effect model approach for variable selection in linear regression models, which is very flexible and has been applied to various high-dimensional numerical studies: Lee et al. (2010) for principal component analysis, Lee et al. (2011a) for canonical correlation analysis, and Lee et al. (2011b) for partial least squares. However, these applications have been restricted to the problems that can be reformulated into low-dimensional linear regression problems.

In this paper, we study the use of random-effect models for high-dimensional generalized linear models. First, we show that the problem of variable selection in high-dimensional generalized linear models can be cast into the framework of the hierarchical likelihood (h-likelihood) estimation (Lee and Nelder, 2006). The use of random-effect models naturally includes a process of selecting variables by assuming some distributional assumptions only. Although the rest of the paper is written in the penalized estimation framework, this is distinct property of the proposed method compared with other existing non-convex penalties such as SCAD and minimax concave penalties.

Second, we show that the random-effect model approach gives a nice way of combining shrinkage estimation with variable selection. The random-effect model approach via the h-likelihood estimation (Lee and Nelder, 2006) can be interpreted as penalized estimations, and the resulting estimator has its own distinct properties compared with other existing penalized estimators. In fact, by choosing values of two variance components, the random-effect model approach can control the sparsity of the estimator and the amount of shrinkage of nonzero estimated coefficients independently, which we cannot expect from other penalties such as the LASSO, SCAD and capped-ℓ_1 (Zhang and Zhang, 2012). It is well known that for prediction, shrinkage estimations such as the ridge would be preferred (Efron and Morris, 1975; Casella, 1985) while for interpretation, sparse estimations such as the SCAD would be preferred. Hence the random-effect model approach can be used for prediction and selection simultaneously in various high-dimensional data analysis.

Third, we propose a unified computational algorithm. Lee and Oh (2009) used a modified iterative weighted least square (MIWLS) algorithm of Hunter and Li (2005). However, the MIWLS algorithm is intractable to implement in high-dimensional problems because it requires non-trivial $p \times p$ matrix inversions. Further, the solution from the NIWLS can be significantly different from the original solution since it uses another thresholding level to distinguish the solution from zero which is similar to the algorithm used in Fan and Li (2001). By combining the convex concave procedure (CCCP) of Yuille and Rangarajan (2003) and current existing algorithms of the LASSO, the proposed algorithm finds an exact local solution efficiently for various high-dimensional statistical models.

Section 2 shows how to use random-effect models for high-dimensional generalized linear models, and Section 3 studies properties of the proposed penalty, introducing an efficient algorithm for non-convex optimization. Section 4 presents simulation studies to demonstrate the usefulness of the new estimators in high-dimensional problems, and Section 5 reports several analysis results of real data sets to show the differences among various methods, and discussions are provided in Section 6.

2. Use of random-effect models for variable selection

2.1. Random-effect estimation via hierarchical likelihood

Consider a generalized linear model with random effects:

$$z_i|\mathbf{v} \sim f_\phi(z_i|\mathbf{v}), \quad i = 1, \ldots, n,$$

where $z_i = (y_i, \mathbf{x}_i)$ is a pair of independent response y_i and p-dimensional covariate vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$, $\mathbf{v} = (v_1, \ldots, v_p)^T \in \mathbb{R}^p$ is a vector of random effects and f_ϕ is a density from an exponential family with a dispersion parameter ϕ. For example, in normal regression models, $\phi > 0$, while in Poisson and binomial regression models, $\phi = 1$.

For random effects $\mathbf{v}$, Lee and Oh (2009) considered the scale gamma mixture of normal model as follows.

$$v_j|w_j \sim f_\sigma(v_j|w_j), \qquad w_j \sim f_\tau(w_j), \quad j = 1, \ldots, p, \tag{1}$$

independently, where

$$f_\sigma(v_j|w_j) = (2\pi\sigma \exp w_j)^{-1/2} \exp(-v_j^2/(2\sigma \exp w_j))$$

is the density function of a normal distribution with mean 0 and variance $\sigma \exp w_j$, and

$$f_\tau(w_j) = (1/\tau)^{1/\tau} \Gamma(1/\tau)^{-1} \exp(w_j/\tau) \exp(-(\exp w_j)/\tau)$$

is the density function of $w_j = \log u_j$, where u_j is distributed as a gamma distribution with mean 1 and variance τ.

Let $g(\mathbf{v}, \mathbf{w})$ be the logarithm of the joint density function of $\mathbf{v} = (v_1, \ldots, v_p)^T$ and $\mathbf{w} = (w_1, \ldots, w_p)^T$, then

$$g(\mathbf{v}, \mathbf{w}) = \sum_{j=1}^{p} \left\{ \log f_\sigma(v_j|w_j) + \log f_\tau(w_j) \right\}.$$

Let $\mathbf{w}(\mathbf{v})$ be the solution of

$$\partial g(\mathbf{v}, \mathbf{w})/\partial \mathbf{w} = \mathbf{0}_p,$$