Computational Statistics and Data Analysis 63 (2013) 42–49

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Robust variable selection through MAVE

Weixin Yao [a], Qin Wang [b,*]

[a] Department of Statistics, Kansas State University, Manhattan, KS 66506, USA

[b] Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, VA 23284, USA

ARTICLE INFO

Article history:
Received 16 October 2011
Received in revised form 21 January 2013
Accepted 22 January 2013
Available online 4 February 2013

Keywords:
Sufficient dimension reduction
MAVE
Shrinkage estimation
Robust estimation

ABSTRACT

Dimension reduction and variable selection play important roles in high dimensional data analysis. The *sparse MAVE*, a model-free variable selection method, is a nice combination of shrinkage estimation, *Lasso*, and an effective dimension reduction method, *MAVE (minimum average variance estimation)*. However, it is not robust to outliers in the dependent variable because of the use of least-squares criterion. A robust variable selection method based on *sparse MAVE* is developed, together with an efficient estimation algorithm to enhance its practical applicability. In addition, a robust cross-validation is also proposed to select the structural dimension. The effectiveness of the new approach is verified through simulation studies and a real data analysis.

1. Introduction

The explosion of massive data in the last decades has generated considerable challenges and interests in the development of statistical modeling. Practically, only part of these observed variables are believed to be truly relevant to the response. Thus, variable selection plays an important role in analyzing these high dimensional data, not only for better model interpretation but also for higher prediction accuracy (Fan and Li, 2006). A lot of research efforts have been devoted to this area. Many traditional model-based variable selection criteria have been advocated and strengthened in the literature, such as C_p, AIC, BIC, etc. Recently a family of regularization approaches, including Nonnegative Garrote (Breiman, 1995), Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), Lars (Efron et al., 2004) and Elastic Net (Zou and Hastie, 2005), was proposed to automatically select informative variables through continuous shrinkage. However, because of the so-called 'curse of dimensionality' (Bellman, 1961), it is very difficult or even infeasible to formulate and validate a parametric model with a large number of covariates. So it is desirable to have a set of model-free variable selection approaches.

Sufficient dimension reduction (Li, 1991; Cook, 1998) provides such a model-free alternative to variable selection. The basic idea of sufficient dimension reduction is to replace the original high dimensional predictor vector with its appropriate low dimensional projection, while preserving full regression information. Each direction in the low dimensional subspace is a linear combination of original predictors. Cook (2004) and Li et al. (2005) proposed several testing procedures to evaluate the contribution of each covariate. Similar to the model-based subset selection procedures, these methods are not stable because of their inherent discreteness (Breiman, 1996). Ni et al. (2005), Li and Nachtsheim (2006), Li (2007), Zhou and He (2008) and (Bondell and Li, 2009) used regularization paradigm to incorporate shrinkage estimation into inverse regression dimension reduction methods. Along the same line, Wang and Yin (2008) combined shrinkage estimation and a forward regression dimension reduction method, MAVE (minimum average variance estimation, Xia et al., 2002), and proposed sparse MAVE to select informative covariates. Compared to the previous work, sparse MAVE is model-free and requires no strong probabilistic assumptions on the predictors. However, MAVE and sparse MAVE are not robust to outliers in the

* Corresponding author.
E-mail addresses: wxyao@ksu.edu (W. Yao), qwang3@vcu.edu (Q. Wang).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2013.01.021

dependent variable because of the use of least-squares criterion. Čížek and Härdle (2006) gave a comprehensive study of the sensitivity of MAVE to outliers and proposed a robust enhancement to MAVE by replacing the local least squares with local L- or M- estimation.

In this article, we extend the robust estimation to variable selection and propose a robust sparse MAVE. It can exhaustively estimate directions in the regression mean function and select informative covariates simultaneously, while being robust to the existence of possible outliers in the dependent variable. In addition, a robust cross-validation is also proposed to select the structural dimension. The effectiveness of the new approach is verified through simulation studies and a real data analysis.

The rest of the article is organized as follows. In Section 2, we briefly review the methods MAVE and sparse MAVE. The robust extension of sparse MAVE is detailed in Section 3. Simulation studies and comparison with some existing methods are presented in Section 4. In Section 5, we apply the proposed robust sparse MAVE to a logo design data collected by Henderson and Cote (1998). Finally, in Section 6, we conclude the article with a short discussion.

2. A brief review of MAVE and sparse MAVE

The regression-type model of a response $y \in \mathcal{R}^1$ on a vector $\mathbf{x} \in \mathcal{R}^p$ can be written as

$$y = g(\mathbf{B}^T\mathbf{x}) + \varepsilon, \tag{1}$$

where $g(\cdot)$ is an unknown smooth link function, $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ is a $p \times d$ orthogonal matrix ($\mathbf{B}^T\mathbf{B} = I_d$) with $d < p$ and $\mathrm{E}(\varepsilon \mid \mathbf{x}) = 0$ almost surely. Xia et al. (2002) defined the d-dimensional subspace $\mathbf{B}^T\mathbf{x}$ the effective dimension reduction (EDR) space, which captures all the information of $E(y|\mathbf{x})$. The d is usually called the structural dimension of the EDR space. Given a random sample $\{(\mathbf{x}_i, y_i),\ i = 1, \ldots, n\}$, the MAVE estimates the EDR directions by solving the following minimization problem

$$\min_{\mathbf{B}, a_j, \mathbf{b}_j, j=1,\ldots,n} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \left[y_i - \left\{ a_j + \mathbf{b}_j^T \mathbf{B}^T (\mathbf{x}_i - \mathbf{x}_j) \right\} \right]^2 w_{ij} \right), \tag{2}$$

where $\mathbf{B}^T\mathbf{B} = I_d$ and the weight w_{ij} is a function of the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. The minimization of (2) can be solved iteratively with respect to $\{(a_j, \mathbf{b}_j),\ j = 1, \ldots, n\}$ and $\mathbf{B}$ separately. The estimation of MAVE is very efficient since only two quadratic programming problems are involved and both have explicit solutions. To improve the estimation accuracy, a lower dimensional kernel weight $\tilde{w}_{ij}$ as a function of $\tilde{\mathbf{B}}^T(\mathbf{x}_i - \mathbf{x}_j)$ can be used after an initial estimate $\tilde{\mathbf{B}}$ was obtained (the refined MAVE).

Note that, each reduced variable in $\mathbf{B}^T\mathbf{x}$ is a linear combination of all original predictors. But it is not uncommon in practice that some covariates are irrelevant among a large number of candidates. To effectively select those informative variables can improve both the model interpretability and the prediction accuracy, Wang and Yin (2008) proposed sparse MAVE to incorporate an L_1 penalty into the above estimation. The constrained optimization is as follows,

$$\min_{\mathbf{B}, a_j, \mathbf{b}_j, j=1,\ldots,n} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \left[y_i - \left\{ a_j + \mathbf{b}_j^T \mathbf{B}^T (\mathbf{x}_i - \mathbf{x}_j) \right\} \right]^2 w_{ij} + \sum_{k=1}^{d} \lambda_k |\boldsymbol{\beta}_k|_1 \right), \tag{3}$$

where $|\cdot|_1$ represents the L_1 norm and $\{\lambda_k,\ k = 1, \ldots, d\}$ are nonnegative regularization parameters which control the amount of shrinkage. Through penalizing on the L_1 norm of the parameter estimates, we can achieve the goal of variable selection when the true direction has a sparse representation. The minimization of (3) can be solved by a standard Lasso algorithm. More details can be found in Wang and Yin (2008).

3. Robust sparse MAVE

3.1. Robust estimation

Note that, in (2) and (3), the least-squares criterion is used between the response and the regression function to evaluate how well the model fits. It corresponds to the *maximum likelihood estimation* (MLE) when the error is normally distributed. However, it is not robust to outliers in the dependent variable y and to the violation of distribution assumptions on ε, such as heavy-tailed errors. To achieve the robustness in estimation, Čížek and Härdle (2006) proposed to replace the local least squares with local L- or M- estimation. The robust MAVE estimates the EDR directions by minimizing

$$\min_{\mathbf{B}, a_j, \mathbf{b}_j, j=1,\ldots,n} \sum_{j=1}^{n} \sum_{i=1}^{n} \rho \left(y_i - \left\{ a_j + \mathbf{b}_j^T \mathbf{B}^T (\mathbf{x}_i - \mathbf{x}_j) \right\} \right) w_{ij}, \tag{4}$$

where $\rho(\cdot)$ is a robust loss function. Note that, the traditional least squares criterion corresponds to $\rho(t) = t^2$, and the median regression uses L_1 loss where $\rho(t) = |t|_1$. Its derivative $\psi(\cdot) = \rho'(\cdot)$ is proportional to the influence function.