Dimension reduction based linear surrogate variable approach for model free variable selection

Pengjie Dai*
School of Business, Renmin University of China, Beijing, China,
daipengjie@rbs.org.cn

Xiaobo Ding
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

Qihua Wang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China,
Yunnan University, Kunming, China

variable selection methods are popular to be applied in the problem of high dimensional data sets in the past years. Most of those methods depend on the model assumptions, while sufficient dimension reduction is a nonparametric method to deal with high dimensional data. In this topic, We aim at integrating sufficient dimension reduction into variable selection. A two stage procedure is proposed. First, we obtain dimension reduction directions and integrate them to construct a variable which is linearly dependent on predictors. Then by treating this constructed variable as a new response, we use the traditional variable selection methods such as adaptive LASSO to conduct variable selection. We call such a procedure as dimension reduction based linear surrogate variable (LSV) method. To illustrate that it has wide application, we also apply it to variable selection for the problem of missing responses. Extensive simulation studies show that it is more robust than the variable selection methods depending on model assumptions, and more efficient than the other model-free variable selection methods. Another advantage of the LSV is that it can be easily implemented. We also consider about the situation of missing data, within which LSV still has a good performance. A real example is given to illustrate the proposed method.

Key Words: adaptive LASSO, central subspace, linear surrogate variable, sufficient dimension reduction, variable selection.

1 Introduction

High dimensional data sets are encountered in many practical situations. However it is often the case that only some of the predictors are relevant for the response, and we want to pick out these relevant predictors. This raises the variable selection issue. Most of the existing variable selection methods rely on model assumptions, such as LASSO (Tibshirani 1996, 1997), adaptive LASSO (ALASSO, Zou 2006), SCAD (Fan and Li 2001) and so forth. When model assumptions are correctly specified, the corresponding variable selection methods can have some good properties. However, it is difficult to establish reasonable model assumptions especially for the high dimensional data.

The existing nonparametric variable selection methods are based on the sufficient dimension reduction (SDR), which is a nonparametric technique for dealing with high dimensional data. Attracted by the model-free property of SDR, many statisticians start to work on the variable selection under the framework of SDR. There are two categories of model-free variable selection approaches. The first category of approaches are test based, including, for example, the approximate SIR based t-test (Chen and Li, 1998), the marginal co-ordinate test (Cook, 2004) and the gridded χ^2-test (Li, Cook and Nachtsheim, 2005). The tests are typically incorporated into a variable subset search procedure, e.g. a stepwise backward or forward search. However, such subset selection methods are not only computationally intensive but may also be unsatisfactory in terms of prediction accuracy and stability (Breiman, 1995). An alternative class of model-free selection methods integrate SDR with the regularization paradigm, and examples include shrinkage SIR (Ni et al, 2005), sparse SDR method (Li, 2007), shrinkage ridge SIR (Li and Yin, 2008), and shrinkage inverse regression (Bondell and Li, 2009). It should be pointed out that all such methods are implemented by iterative processes, which are definitely time-consuming. In addition, due to the complicated algorithms, some efficiency of the methods might be lost.

The target of this paper is to find a new model-free variable selection approach which can select the relevant predictors efficiently. We propose the so called dimension reduction based linear surrogate variable (LSV) method. Specifically, we first, based on SDR, construct a LSV, say U, which is linearly dependent on $\mathbf{X}$, and then apply the variable selection methods for linear models to the linear regression of U on $\mathbf{X}$. Unlike the methods mentioned above, the LSV method employs the SDR and variable selection methods respectively in two individual steps, such that it has some advantages. First, the computation is quite simple and fast. Second, since the variable selection methods are best developed for linear models, which are incorporated into the LSV method, the LSV method might gain more efficiency than the other model-free methods mentioned above. Third, the LSV method can easily incorporate the SDR and variable selection methods. So it can be easily extended to more complicated cases as long as there exist the SDR and variable selection methods adaptive for these cases. For example, the LSV method can handle this case of missing responses by using the two-stage procedure proposed by Ding and Wang (2011) to estimate the central subspace.

2 The LSV method

2.1 Methodology

As stated in Section 1, SIR is a popular SDR method. In this paper it is employed to estimate $\mathcal{S}_{Y|\mathbf{X}}$. It is well known that SIR requires the so-called linear condition that $E\left(\mathbf{X}|B^\top\mathbf{X}\right)$ is a linear function of $B^\top\mathbf{X}$, where B is a basis of central subspace $\mathcal{S}_{Y|\mathbf{X}}$. Denote $\mathcal{M} = \mathrm{cov}\left(E\left(\mathbf{X}|Y\right)\right)$, which is called the kernel matrix of SIR. When the linear condition holds, then Span$(\mathcal{M}) \subseteq \mathcal{S}_{Y|\mathbf{X}}$, where Span$(\mathcal{M})$ refers to the column space of $\mathcal{M}$. Note that Span$(\mathcal{M})$ is a proper subspace of $\mathcal{S}_{Y|\mathbf{X}}$ only in exceptional cases. Throughout this paper we assume Span$(\mathcal{M}) = \mathcal{S}_{Y|\mathbf{X}}$. Denote the dimension of $\mathcal{S}_{Y|\mathbf{X}}$ by K. The eigenvalues of $\mathcal{M}$ are ranked in decreasing order and denoted by $\lambda_1 \geq \cdots \geq \lambda_K > \lambda_{K+1} = \cdots = \lambda_p = 0$. The corresponding eigenvectors are denoted by $\nu_i = \left(\nu_{i1}, \cdots, \nu_{ip}\right)^\top$, $i = 1, \ldots, p$. Then $\nu_i \in \mathcal{S}_{Y|\mathbf{X}}$ for $i \leq K$ and ν_i belongs to the orthogonal complement subspace of $\mathcal{S}_{Y|\mathbf{X}}$ for $i > K$.

By the definition of relevant predictors, we have that if X_j is a relevant predictor, i.e., $j \in \mathscr{A}$, then $\nu_{ij} \neq 0$ for some $1 \leq i \leq K$, and if $j \in \mathscr{A}^C$, then $\nu_{ij} = 0$ for $1 \leq i \leq K$. To select the relevant predictors from the K dimension reduction directions, we let

$$\beta^* = \sum_{i=1}^{K} \lambda_i |\nu_i|, \tag{1}$$

where $|\nu_i| = (|\nu_{i1}|, |\nu_{i2}|, \cdots, |\nu_{ip}|)^\top$. Let β_j^* be the jth element of β^*. It is obvious that $\{j : \beta_j^* \neq 0\} = \mathscr{A}$. Then we construct a LSV

$$U = \mathbf{X}^\top \beta^*. \tag{2}$$

Immediately the variable selection methods for linear models can be applied for the regression of U on $\mathbf{X}$. Since the ALASSO proposed by Zou (2006) is known as an efficient method for linear models, we employ it in this paper.

In practice K is unknown. There are many methods to estimate K such as the sequential test (Li 1991), Bayesian information criterion (BIC, Zhu et al. 2006), penalized spectral decomposition (Zhu et al. 2010) and so forth. In this paper we introduce the modified BIC proposed by Zhu et al. (2010) since it can estimate K consistently.

Note that a larger absolute value of ν_{ij} indicates that X_j is more important for the direction ν_i. To remain the importance we take absolute values of the dimension reduction directions for the construction of β^*. Because if we do not take absolute values, then there might be exist $j \in \mathscr{A}$ such that β_j^* is equal to 0 or very small even when the absolute values of ν_{ij}, $i = 1, \ldots, K$ are large, because they can cancel each other. Of course, one might take square of elements of ν_i as an alternative method.

2.2 *Sample estimator and asymptotic properties*

Suppose that $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ are independently drawn from $(\mathbf{X}, Y)$. Then the proposed LSV method works as follows.

Step 1. Slice the range of Y into H pieces, say $I_1, \ldots, I_H$.

Step 2. Estimate the kernel matrix by $\hat{\mathcal{M}} = H^{-1} \sum_{h=1}^{H} \hat{p}_h \hat{m}_h \hat{m}_h^\top$, where $\hat{p}_h = n^{-1} \sum_{j=1}^{n} I(y_j \in I_h)$ and $\hat{m}_h = (n\hat{p}_h)^{-1} \sum_{j=1}^{n} \mathbf{x}_j I(y_j \in I_h)$.

Step 3. Conduct spectral decomposition of $\hat{\mathcal{M}}$ and obtain the eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ and the corresponding eigenvectors $\hat{\nu}_1, \hat{\nu}_2, \ldots, \hat{\nu}_p$.

Step 4. Use the BIC proposed by Zhu et al. (2010) to obtain an estimator of K, i.e., $\hat{K} = \arg\max_{k=1,\cdots,p} \{G(k)\}$, where $G(k)$ is defined as

$$G(k) = \frac{n}{2} \times \frac{\sum_{l=1}^{k} \left\{\log\left(\hat{\lambda}_l + 1\right) - \hat{\lambda}_l\right\}}{\sum_{l=1}^{p} \left\{\log\left(\hat{\lambda}_l + 1\right) - \hat{\lambda}_l\right\}} - 2 \times C_n \times \frac{k(k+1)}{2p}. \tag{3}$$

Here C_n is a value dependent on n. Usually we can set $C_n = n^{1/2}$.

Step 5. Obtain that

$$\hat{\beta} = \sum_{i=1}^{\hat{K}} \hat{\lambda}_i |\hat{\nu}_i| \tag{4}$$

and $\hat{u}_i = \mathbf{x}_i^\top \hat{\beta}$ for $i = 1, \ldots, n$.

Step 6. Apply the ALASSO to the data $(\hat{u}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ and select the relevant predictors. That is,

$$\hat{\beta}_{\text{alasso}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n} \left(\hat{u}_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \xi_n \sum_{j=1}^{p} \hat{w}_j |\beta_j| \right\}, \tag{5}$$

where ξ_n is a tuning parameter, x_{ij} is the jth element of $\mathbf{x}_i$, and the weight $\hat{w}_j = |\hat{\beta}_{\text{ols},j}|^{-\gamma}$ with $\gamma > 0$ and $\hat{\beta}_{\text{ols}}$ is the ordinary least square solution of U on $\mathbf{X}$. Then we obtain that $\widehat{\mathscr{A}} = \{j : \hat{\beta}_{\text{alasso},j} \neq 0\}$.

And we also prove the oracle properties of LSV method.

2.3 Application to data with missing response at random

The existing model-free variable selection methods only handle the cases where all the subjects are completely observed. However, data with missing responses are a common problem in practice. Some procedures have been proposed for handling this problem, see, e.g., Garcia, Ibrahim and Zhu (2010), Heymans et al. (2007) and so forth. But these methods are somewhat complicated and based on some model assumptions. So here we extend the LSV method to this case. We applied a simple general two-stage procedure(Ding and Wang (2011)), called Fusion–Refinement (FR), and the asymptotic results are also nice.

All proposed methods were tested by a plenty of simulation studies, and showed good performances.

3 Concluding Remarks

In this paper we use SIR to obtain the dimension reduction directions to construct the LSV, and then employ ALASSO to conduct variable selection. It should be emphasized that other SDR and variable selection methods can also be employed to implement the LSV procedure. For the SDR methods with spectral decomposition of a matrix, such as SAVE, the eigenvalues can be also used as weights to construct LSV. For other type of SDR methods, such as sliced regression (Wang and Xia 2008), how to measure the importance of the obtained dimension reduction directions should be considered so that suitable weights to every direction can be assigned.

References

Bondell, H. D. and Li, L. (2009). Shrinkage inverse regression estimation for model-free variable selection. *J. R. Statist. Soc.* B **71**, 287–299.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32**, 1062–1092.

Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika* **94**, 569–584.

Ding, X. and Wang, Q. (2011). Fusion–refinement procedure for dimension reduction with missing response at random. *J. Amer. Statist. Assoc.* **106**, 1193–1207.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc.* B **70**, 849–911.

Garcia, R. I., Ibrahim, J. G. and Zhu, H. (2010). Variable selection for regression models with missing data. *Statist. Sinica* **20**, 149–165.

Heymans, M. W., Van Buuren, S., Knol, D. L, Van Mechelenen, W., and de Vet, H. C. W. (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Medical Research Methodology **7**, 33-42.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 997–1008.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316–342.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94**, 603–613.

Li, L., Cook, R.D. and Nachtsheim C.J. (2005). Model-free variable selection. *J. R. Statist. Soc.* B **67**, 285–299.

Li, L., and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64**, 124–131.

Lu, W. and Li, L. (2011). Sufficient dimension reduction for censored regressions. *Biometrics* **67**, 513–523.

Ni, L., Cook, R.D. and Tsai, C.L. (2005). A note on shrinkage sliced inverse regression. *Biometrika* **92**, 242–247.

Tibshirani, R. J. (1997). The LASSO method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc.* **103**, 811–821.

Zhu, L. P., Yu, Z. and Zhu, L. X. (2010). A sparse eigen-decomposition estimation in semi-parametric regression. *Comput. Statist. Data Anal.* **54**, 976–986.

Zhu, L. P., Zhu, L. X. and Feng, Z. H. (2010). Dimension reduction in regressions through cumulative slicing estimation *J. Amer. Statist. Assoc.* **105**, 1455–1466.

Zhu, L. X., Miao, B. Q. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Amer. Statist. Assoc.* **101**, 630–643.

Zhu, L. X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5**, 727–736.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.