Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Fang, J Biom Biostat 2016, 7:4
DOI: 10.4172/2155-6180.1000318

Mini Review Open Access

A Pass to Variable Selection

Yixin Fang*

Department of Mathematical Sciences, New Jersey Institute of Technology, USA

Introduction

Many regularized procedures produce sparse solution and therefore are sometimes used for variable selection in linear regression. It has been showed that regularized procedures are more stable than subset selection. Such procedures include LASSO, SCAD, and adaptive LASSO, to name just a few. However, their performance depends crucially on the tuning parameter selection. For the purpose of prediction, popular methods for the tuning parameter selection include C_p, cross-validation, and generalized cross-validation. For the purpose of variable selection, the most popular method for the tuning parameter selection is BIC. The selection consistency of BIC for some regularized procedures have been shown. However, knowing degrees of freedom is required in the use of BIC. For many regularized procedures, such as those for graphical models and clustering algorithms, the formulae for degrees of freedom do not exist.

Recently, stability selection has become another popular method for variable selection [1,2]. However, most methods based on stability depend on some hyper-tuning parameter explicitly. For example, the method in [1] depends on a threshold (pre-set as 0.8 in [1]) and the method in [2] depends also on a threshold (pre-set as 0.9 in [2]). Therefore, it is desirable to propose some method to avoid such hyper-tuning parameter in stability selection methods. One suggestion is to combine the strength of both stability selection and cross-validation. Since cross-validation is one variable selection method based on prediction, the new method is referred as the prediction and stability selection (PASS).

Prediction and Stability Selection (PASS)

Consider variable selection in linear regression, $y_i = x_i\beta + \varepsilon_i$, $i = 1,\ldots,n$. Assume $\beta = (\beta_1,\ldots,\beta_p)'$ is sparse in the sense that $|\mathcal{A}| = q < p$, where $\mathcal{A} = \{j\colon \beta_j \neq 0\}$. Without loss of generality, assume $\mathcal{A} = \{1,\ldots,q\}$. A general framework for the regularized regression is $\hat{\beta}_\lambda = \arg\min_{\gamma\in\mathbb{R}p} \sum_{i=1}^{n}\left(y_i - x_i'\gamma\right)2/n + \sum_{j=1}^{p} p\lambda\left(|\gamma_j|\right)$. This framework includes LASSO, SCAD, and adaptive LASSO. If $\widehat{\mathcal{A}}_\lambda = \{j : \widehat{\beta}_{\lambda j} \neq 0\}$ is used to estimate $\mathcal{A}$, most regularized procedures have been shown to be selection consistent with appropriate $\lambda = \lambda_n$, emphasizing its dependence on data. In general, as shown in [3], there are five cases:

Case 1: If $\lambda_n \lambda_n \succ S_n$, then $\widehat{\beta}_{\lambda_n} = 0$ with probability tending to one.

Case 2: If $\lambda_n \asymp s_n$, then $\widehat{\beta}_{\lambda_n} \to \gamma_0 \neq \beta$, where γ_0 is fixed and its sign pattern may or may not be the same as that of β.

Case 3: If $r_n \prec \lambda_n \prec S_n$, then $\widehat{\beta}_{\lambda_n} \to \beta$ and the sign pattern of $\widehat{\beta}_{\lambda_n}$ is consistent with that of β with probability tending to one.

Case 4: If $\lambda_n \asymp r_n$, then the sign pattern of $\widehat{\beta}_{\lambda_n}$ is consistent with that of β on $\mathcal{A}$ with probability tending to one, while for all sign patterns consistent with that of β on $\mathcal{A}$, the probability of obtaining this pattern is tending to a limit in (0,1).

Case 5: If $\lambda_n \prec r_n$, then $\widehat{\beta}_{\lambda_n} \to \beta$ and $\widehat{\mathcal{A}}_{\lambda_n} = \{1,\cdots,p\}$ with probability tending to one.

A good criterion should intend to select λ_n from case 3; selecting λ_n from cases 1 or 2 might lead to under-fitting while from cases 4 or 5 might lead to over-fitting. If the two degenerate cases (1 and 5) are pre-excluded, the criterion, referred to PASS, incorporates cross-validation, which avoids under-fitting, and Kappa selection proposed in [2], which avoids over-fitting. To describe this criterion, consider any regularized procedure with λ and randomly partition the dataset $\{(y_1,x_1\{(y_1,\ x_1),\ldots,(y_n,x_n)\}$ into two halves, $Z_1 = \{(y_1^*,x_1^*),\cdots,(y_m^*,x_m^*)\}$ and $Z_2 = \{(y_{m+1}^*,x_{m+1}^*),\cdots,(y_n^*,x_n^*)\}$, where $m = \lfloor n/2 \rfloor$. Based on Z_1 and Z_2 respectively, $\widehat{\beta}_{k\lambda}$ is obtained and then submodel $\widehat{\mathcal{A}}_{k\lambda}$ is selected, $k = 1,2$.

If λ is from Case 4, both submodels, $\widehat{\mathcal{A}}_{k\lambda}, k=1,2,$ would include non-informative variables randomly. The agreement of these two submodels can be measured by Cohen's Kappa Coefficient, $\kappa(\widehat{\mathcal{A}}_{1\lambda},\widehat{\mathcal{A}}_{2\lambda})$. On the other hand, if λ is from Case 2, either submodels, $\widehat{\mathcal{A}}_{k\lambda}, k=1,2,$ might exclude some informative variable. To avoid such under-fitting, consider cross-validation, CV(Z_1,Z_2; λ). Now we are ready to describe the PASS algorithm, which runs the following five steps.

Step 1: Randomly partition the original dataset into Z_1^{*b} and Z_2^{*b}.

Step 2: Based on Z_2^{*b} and Z_2^{*b} respectively, two sub-models, $\widehat{\mathcal{A}}_{1\lambda}^{*b}$ and $\widehat{\mathcal{A}}_{1\lambda}^{*b}$, are selected.

Step 3: Calculate $\kappa(\widehat{\mathcal{A}}_{1\lambda}^{*b},\widehat{\mathcal{A}}_{2\lambda}^{*b})$ and $CV(Z_1^{*b},Z_2^{*b};\lambda)$.

Step 4: Repeat Steps 1-3 for B times and obtain the following ratio,

$$PASS(\lambda) = \sum_{b=1}^{B}\kappa(\widehat{\mathcal{A}}_{1\lambda}^{*b},\widehat{\mathcal{A}}_{2\lambda}^{*b}) / \sum_{b=1}^{B} CV(Z_1^{*b},Z_2^{*b};\lambda). \tag{1}$$

Step 5: Compute *PASS*(λ) on a grid of λ and select $\hat{\lambda} = \arg\max_\lambda PASS(\lambda)$.

Discussion

The new criterion has several advantages. First, it does not depend on any hyper-tuning parameter. Second, the implementation is straightforward. Third, it can be applied to variable selection in any models such as linear model, generalized linear model, and Cox's proportional hazard model. Fourth, it can also be applied to variable selection in both supervised learning and unsupervised learning.

References

1. Meinshausen N, Buhlmann P (2010) Stability selection (with discussion). Journal of the Royal Statistical Society 72: 417-473.
2. Sun W, Wang J, Fang Y (2013) Consistent selection of tuning parameters via variable selection stability. Journal of Machine Learning Research 14: 3419-3440.
3. Bach F (2008) Bolasso: model consistent lasso estimation through the bootstrap. Proceedings of 25th International Conference of Machine Learning: 33-40.

***Corresponding author:** Fang Y, Department of Mathematical Sciences, New Jersey Institute of Technology, USA, Tel: +1 973-596-3281; E-mail: yixin.fang@njit.edu

Received September 06, 2016; **Accepted** September 30, 2016; **Published** October 07, 2016

Citation: Fang Y (2016) A Pass to Variable Selection. J Biom Biostat 7: 318. doi:10.4172/2155-6180.1000318