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 shown 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 hypothesis-testing 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, $y_i = x_i \beta + \epsilon_i, i = 1, \ldots, n$. Assume $\beta = (\beta_1, \ldots, \beta_p)'$ is sparse in the sense that $|A| = q \ll p$, where $A = \{i : \beta_i \neq 0\}$. Without loss of generality, assume $A = \{1, \ldots, q\}$. A general framework for the regularized regression is $\hat{\beta}_A = \text{arg}(\lambda \sum_{i=1}^{n} (y_i-x_i \beta)^2 + \sum_{j=1}^{p} \lambda_j \beta_j, \|\beta\|_0 \leq q \}$ [3]. This framework includes LASSO, SCAD, and adaptive LASSO. If $\hat{\beta}_A = \{i : \hat{\beta}_i \neq 0\}$ is used to estimate $A$, most regularized procedures have been shown to be selection consistent with appropriate $\lambda = \lambda_q$ emphasizing its dependence on data. In general, as shown in [3], there are five cases:

Case 1: If $\lambda \lambda_1 \lambda S_1$, then $\hat{\beta}_A = 0$ with probability tending to one.

Case 2: If $\lambda \lambda_2 \lambda S_2$, then $\hat{\beta}_A \rightarrow \gamma \beta$, where $\gamma$ is fixed and its sign pattern may or may not be the same as that of $\beta$.

Case 3: If $\lambda \lambda_3 \lambda S_3$, then $\hat{\beta}_A \rightarrow \beta$ and the sign pattern of $\hat{\beta}_A$ is consistent with that of $\beta$ with probability tending to one.

Case 4: If $\lambda \lambda_4 \lambda S_4$, then the sign pattern of $\hat{\beta}_A$ is consistent with that of $\beta$ on $A$ with probability tending to one, while for all sign patterns consistent with that of $\beta$ on $A$, the probability of obtaining this pattern is tending to a limit in $(0, 1)$.

Case 5: If $\lambda \lambda_5 \lambda S_5$, then $\hat{\beta}_A \rightarrow \beta$ and $\hat{A}_A = \{1, \ldots, p\}$ with probability tending to one.

A good criterion should intend to select $\lambda_q$ from case 3; selecting $\lambda_q$ 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-cluded, 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 $\lambda$ and randomly partition the dataset $\{(y_1, x_1, \ldots, y_n, x_n)\}$ into two halves, $Z_1 = (\{y_1, \ldots, y_{n_1}\}, \{x_1, \ldots, x_{n_1}\})$ and $Z_2 = (\{y_{n_1+1}, \ldots, y_{n}\}, \{x_{n_1+1}, \ldots, x_{n}\})$, where $m = \lfloor n/2 \rfloor$. Based on $Z_1$ and $Z_2$ respectively, $\hat{\beta}_{A_{Z_1}}$ is obtained and then submodel $\hat{A}_{Z_1}$ is selected, $k = 1, 2$.

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

Step 1: Randomly partition the original dataset into $Z_1^n$ and $Z_2^n$.

Step 2: Based on $Z_1^n$ and $Z_2^n$ respectively, two sub-models, $\hat{A}_{Z_1}$ and $\hat{A}_{Z_2}$, are selected.

Step 3: Calculate $\kappa(\hat{A}_{Z_1}, \hat{A}_{Z_2})$ and $CV(Z_1^n, Z_2^n; \lambda)$.

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

$PASS(\lambda) = \frac{\sum_{i=1}^{B} \kappa(\hat{A}_{Z_1}, \hat{A}_{Z_2})}{\sum_{i=1}^{B} CV(Z_1^n, Z_2^n; \lambda)}$.

Step 5: Compute $PASS(\lambda)$ on a grid of $\lambda$ and select $\lambda = \text{arg max,}PASS(\lambda)$.

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