# LASSO Tuning Parameter Selection

# LASSO TUNING PARAMETER SELECTION

***Lisa-Ann Kirkland, Frans Kanfer and Sollie Millard***
University of Pretoria

***Key words:*** LASSO, model selection, penalized least squares, tuning parameter, variable selection.

***Summary:*** The LASSO is a penalized regression method which simultaneously performs shrinkage and variable selection. The output produced by the LASSO consists of a piecewise linear solution path, starting with the null model and ending with the full least squares fit, as the value of a tuning parameter is decreased. The performance of the selected model therefore depends greatly on the choice of this parameter. This paper attempts to provide an overview of methods which are available to select the value of the tuning parameter for either prediction or variable selection purposes. A simulation study provides a comparison of these methods and assesses their performance.

## 1. Introduction

The least angle shrinkage and selection operator (LASSO) was proposed by Tibshirani (1996) as a linear regression method which yields interpretable models with high accuracy. When the true model is sparse, the LASSO often yields a lower prediction error than the full least squares model. The residual sum of squares (RSS) is penalized, yielding coefficients which are shrunk towards zero, trading a slight increase in bias for a substantial reduction in variance (Seber and Lee, 2003; Hastie, Tibshirani and Friedman, 2009). By utilizing the $\ell_1$ penalty, which is non-differentiable at zero, some of the coefficients are set exactly to zero so that the LASSO performs estimation and variable selection simultaneously (Fu, 1998). Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be the response vector, $\mathbf{x}_j = \left(x_{1j}, \ldots, x_{nj}\right)^T$ be the predictor vectors and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ be the predictor matrix. The LASSO estimates are given by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \; \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} \left|\beta_j\right|$ is the $\ell_1$ norm and $\lambda \geqslant 0$ is a tuning parameter. The predictor variables are assumed to be centred and scaled to have unit $\ell_2$ norm, that is, $\sum_{i=1}^{n} x_{ij} = 0$ and $\sum_{i=1}^{n} x_{ij}^2 = 1$. Without loss of generality, the response variable is centred so that the intercept term is $\bar{y} = 0$.

The LASSO's performance relies heavily on the choice of tuning parameter $\lambda$ to select the optimal model. For prediction purposes, the prediction error (PE) is estimated using either cross-validation (CV) methods or information criteria (Hastie et al., 2009). A drawback of using information criteria is that the degrees of freedom (DF) and error variance $\sigma^2$ must be known. Recent studies (Efron, Hastie, Johnstone and Tibshirani, 2004; Zou, Hastie and Tibshirani, 2007; Tibshirani and Taylor, 2012; Dossal, Kachour, Fadili, Peyre and Chesneau, 2013) have shown that the LASSO uses DF equal to the number of non-zero coefficients in the model. Choosing the tuning parameter for variable selection is more difficult since the prediction optimal value is inconsistent for selection (Wang, Li and Leng, 2009). Recent advances have been made to stabilize the selection and entail some form of resampling or multiple sample splitting.

This paper serves as a selective overview and comparison of methods which are available to select the tuning parameter for the LASSO. Formulae for DF of the LASSO are given in Section 2. Methods used to select the best predictive model are discussed in Section 3, while those more appropriate for variable selection are discussed in Section 4. A simulation study in Section 5 provides a comparison of all the methods and some concluding remarks can be found in Section 6.

# 2. Degrees of Freedom

Let $\hat{\boldsymbol{\beta}}(\lambda)$ be the LASSO coefficient estimates at $\lambda$. Tibshirani (1996) proposed writing the LASSO penalty as $\sum|\beta_j| = \sum \beta_j^2 \big/ |\beta_j|$, yielding a ridge approximation for the LASSO which could be used to approximate its DF. However, Efron et al. (2004) discovered that the solution path of the LASSO is piecewise linear and modified the least angle regression (LAR) algorithm to compute its solution. They claimed that the DF is well approximated by the number of non-zero coefficients,

$$\widehat{df}(\lambda) = \left|\hat{\mathscr{A}}(\lambda)\right|,$$

where $\hat{\mathscr{A}}(\lambda) = \left\{j : \hat{\beta}_j(\lambda) \neq 0\right\}$ is the subset of non-zero coefficients (active set) and $\left|\hat{\mathscr{A}}(\lambda)\right|$ is its cardinality. Zou et al. (2007) proved that if $\mathbf{y} \sim N\left(\boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$ and $rank(\mathbf{X}) = p$ then this estimate is unbiased and consistent, and the optimal value of $\lambda$ is one of the transition points in the LASSO path. Tibshirani and Taylor (2012) and Dossal et al. (2013) generalized the result (independently) so that the full rank assumption is not needed. Their estimates can therefore be used when $p > n$, and are given by, respectively,

$$\widehat{df}(\lambda) = rank\left(\mathbf{X}_{\hat{\mathscr{A}}(\lambda)}\right) \quad \text{and} \quad \widehat{df}(\lambda) = \left|\hat{\mathscr{A}}^*(\lambda)\right|,$$

where $\hat{\mathscr{A}}^*(\lambda) = \left\{j : \hat{\beta}_j^*(\lambda) \neq 0\right\}$ and $\hat{\boldsymbol{\beta}}^*(\lambda)$ is a solution such that $\mathbf{X}_{\hat{\mathscr{A}}^*(\lambda)}$ has full rank. They show that $\hat{\mathscr{A}}^*(\lambda)$ is the minimum size of all active sets of LASSO solutions.

# 3. Prediction

Let the LASSO fit be denoted by $\hat{\boldsymbol{\mu}}(\mathbf{X},\lambda) = \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda)$. Prediction accuracy of a model can be assessed by calculating its PE, that is, the error when the model is used to predict a new sample of, say $m$, observations $(\mathbf{y}_0, \mathbf{X}_0)$,

$$PE(\lambda) = PE(\hat{\boldsymbol{\mu}}(\mathbf{X}_0,\lambda)) = E\left\|\mathbf{y}_0 - \hat{\boldsymbol{\mu}}(\mathbf{X}_0,\lambda)\right\|^2 = MSE(\lambda) + m\sigma^2,$$

where the mean squared error (MSE) of the model is given by

$$MSE(\lambda) = MSE(\hat{\boldsymbol{\mu}}(\mathbf{X}_0,\lambda)) = \left(\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}\right)^T \mathbf{C}\left(\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}\right),$$

with $\mathbf{C}$ being the covariance matrix of $\mathbf{X}_0$. Greenshtein and Ritov (2004) showed that the LASSO is persistent (consistent for prediction error) in high-dimensional settings.

## 3.1. Cross-validation

Model selection can be performed by splitting the available data into three sets, one for training, one for validation and one for testing. If one does not have enough data for a validation set, CV can be used as an alternative. The data is resampled to form an estimate of the PE called the CV error. Three different types of CV are commonly encountered: $K$-fold CV, leave-one-out CV (LOOCV) and generalized CV (GCV). They are discussed in many texts, for example Hastie et al. (2009), and can be calculated using the formulae in Table 1, where $\hat{\mu}^{-k}(\mathbf{x}_i,\lambda)$ is the prediction of the $i$-th observation by the model estimated using all observations except those in the $k$-th part. The LASSO is shown to be persistent when using $K$-fold CV (Homrighausen and McDonald, 2013) and LOOCV (Homrighausen and McDonald, 2014).

| | |
|---|---|
| $K$-fold CV | $CV_K(\lambda) = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{n_k}\sum_{i=1}^{n_k}\left(y_i - \hat{\mu}^{-k}(\mathbf{x}_i,\lambda)\right)\right]$ |
| LOOCV | $CV_n(\lambda) = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \hat{\mu}^{-k}(\mathbf{x}_k,\lambda)\right)^2$ |
| GCV | $GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}(\mathbf{x}_i,\lambda))^2 \Big/ \left(1 - \widehat{df}(\lambda)\Big/ n\right)^2$ |

***Table 1**: Formulae for calculating different types of CV error*

## 3.2. Information Criteria

An advantage of information criteria is that they have considerably less computational expense than CV. Once the models are estimated, it is simply a matter of evaluating an expression for each model. However, there are some drawbacks: the DF must be known and a model that is roughly correct is necessary to estimate $\sigma^2$. Mallow's $C_p$ (Mallows, 1973) and Akaike's information criteria (AIC) (Akaike, 1973) are well-known information criteria. They are suitable for prediction purposes and select identical models. For use with the LASSO, Zou et al. (2007) propose using the formulae,

$$C_p(\lambda) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}(\mathbf{X},\lambda)\|^2}{n} + \frac{2\widehat{df}(\lambda)}{n}\hat{\sigma}^2 \quad \text{and} \quad AIC(\lambda) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}(\mathbf{X},\lambda)\|^2}{n\hat{\sigma}^2} + \frac{2}{n}\widehat{df}(\lambda),$$

where an appropriate estimate from Section 2 is used as $\widehat{df}(\lambda)$ and $\hat{\sigma}^2 = \|\mathbf{y} - \hat{\boldsymbol{\mu}}_{LS}(\mathbf{X})\|^2 \Big/ p$ is the estimate obtained from the full least squares fit $\hat{\boldsymbol{\mu}}_{LS}(\mathbf{X})$.

# 4. Variable Selection

Variable selection performance can be assessed by the probability of choosing the correct subset (PCS). An alternative measure, which is good for variable screening, is the probability that the model includes the correct subset (PIS). If the true model includes the subset of predictors $\mathscr{A}$ and the model selects the subset $\hat{\mathscr{A}}(\lambda)$, then these measures are given by

$$PCS(\lambda) = P\left(\hat{\mathscr{A}}(\lambda) = \mathscr{A}\right) \quad \text{and} \quad PIS(\lambda) = P\left(\hat{\mathscr{A}}(\lambda) \supseteq \mathscr{A}\right).$$

Bühlmann and van de Geer (2011) show that the LASSO is suitable for variable screening under an assumption called the restricted eigenvalue condition (Bickel, Ritov and Tsybakov, 2009). Under a stricter assumption, the irrepresentable condition, the LASSO is consistent for variable selection (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Zou, 2006; Yuan and Lin, 2007).

## 4.1. Information Criteria

The Bayesian information criterion (BIC) (Schwarz, 1978) is commonly used to uncover the correct model. For use with the LASSO, Zou et al. (2007) propose

$$BIC(\lambda) = \frac{\|\mathbf{y} - \hat{\mu}(\lambda)\|^2}{n\hat{\sigma}^2} + \frac{\log(n)}{n}\widehat{df}(\lambda),$$

where $\widehat{df}(\lambda)$ and $\hat{\sigma}^2$ are as described in Section 3.2. Wang et al. (2009) constructed a modified *BIC* criterion for the LASSO,

$$\text{Modified } BIC(\lambda) = \log\left(\frac{\|\mathbf{y} - \hat{\mu}(\lambda)\|^2}{n}\right) + |\mathscr{A}(\lambda)|\,\frac{\log(n)}{n}C_n,$$

where $C_n > 0$ is a positive constant. They prove that it is consistent for variable selection when $p < n$ and use the value of $C_n = \ln(\ln p)$ in their studies. Chand (2012) show by simulations that $C_n = \sqrt{n}\big/ p$ leads to consistent selection. A more general information criterion is proposed by Fan and Tang (2013), the generalized information criteria (GIC), which can be used when $p \gg n$.

## 4.2. Sample Splitting and Resampling

CV does not lead to consistent selection since the chosen $\hat{\lambda}$ is not large enough to set coefficients to zero. Breiman, Friedman, Ohlsen and Stone (1984) proposed using CV with the one-standard error (1SE) rule for pruning classification and regression trees. The idea is to stabilize the selection and promote parsimony without losing accuracy by selecting the smallest model for which $CV_K$ lies within one standard error of the minimum $CV_K$. Hastie et al. (2009) recommend using the 1SE rule for subset selection but there is no supporting literature for its use with the LASSO.

Roberts and Nowak (2014) propose the percentile CV, a method that repeatedly performs $K$-fold CV to stabilize the variability due to different fold allocations. $K$-fold CV is applied for $M$ repetitions and the $\theta$-th percentile of the vector $\left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_M\right)$ is used as the tuning parameter. They suggest using $M \geqslant 10$ and $\theta = 0.95$. They show that the number of false inclusions and the variability of the model size is greatly reduced.

Sun, Wang and Fang (2013) propose using the kappa coefficient, which measures the agreement between two independent sets. Suppose that $P(\text{obs})$ is the relative observed agreement and $P(\text{chance})$ is the probability of chance agreement, then the kappa coefficient is given by

$$\kappa(\mathscr{A}_1, \mathscr{A}_2) = \frac{P(\text{obs}) - P(\text{chance})}{P(\text{chance})}.$$

The procedure randomly splits the data into two equally sized samples and calculates the kappa coefficient between the LASSO subsets selected at each $\lambda$. The process is repeated a number of times and the average kappa coefficient is calculated for each $\lambda$ as $\bar{\kappa}(\lambda)$. They propose using

$$\hat{\lambda} = \min\left\{\lambda \,\middle|\, \frac{\bar{\kappa}(\lambda)}{\kappa_{\max}} \geqslant 1 - \theta\right\},$$

where $\kappa_{\max} = \max_{\lambda}\{\bar{\kappa}(\lambda)\}$. They recommend using a small value of $\theta$ such as 0.1 and prove that the method leads to consistent variable selection.

Fang, Wang and Sun (2013) propose combining the performance of variable selection and prediction by a criterion which they call prediction and stability selection (PASS). It is the ratio of the average kappa coefficient to the CV error, $PASS(\lambda) = \bar{\kappa}(\lambda)/CV(\lambda)$.

# 5. Simulation Study

The simulation study analyses the selection and prediction performance of the LASSO when using different methods for choosing the tuning parameter $\lambda$. The data generating process is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ and $\mathbf{X} \sim N(\mathbf{0}, \mathbf{C})$ with $\mathbf{C}_{ij} = \rho^{|j-k|}$ for $j, k = 1, \ldots, 8$. The true parameter vector is $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ so that the correct subset of predictors is given by $\mathscr{A} = \{1, 2, 5\}$. This example was studied by Tibshirani (1996) and appears in a number of papers under various scenarios. In this study the sample size is varied $n \in \{25, 50, \ldots, 500\}$ to test consistency, the correlation is $\rho = 0.5$ and the noise is $\sigma = 3$. Estimation is carried out by generating $N = 100$ samples from this process. The validation set used for model selection has a sample size of 500.

Zhao and Yu (2006) showed that correlation matrices such as $\mathbf{C}$ satisfy the irrepresentable condition so the LASSO should be consistent for variable selection. The left panel of figure 1 shows the maximum irrepresentable condition over the $N$ samples for each sample size. The condition is not satisfied for a small percentage of samples when the sample size is small. When $n \geqslant 150$, it is certain that the condition holds. The LASSO path was fitted to each sample using the cyclic coordinate descent algorithm by Friedman, Hastie and Tibshirani (2010) available in the R package `glmnet`. The right panel of Figure 1 shows the percentage of times over the $N$ samples when the LASSO path contains the true model. When $n \geqslant 175$, the LASSO should be able to recover the correct model since it is contained in the set of candidate models with certainty.

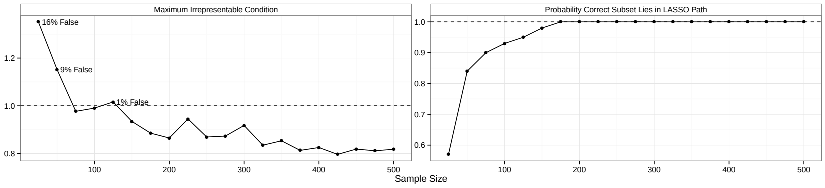

***Figure 1**: The maximum irrepresentable condition (left) and the probability that the LASSO path contains the correct subset (right).*

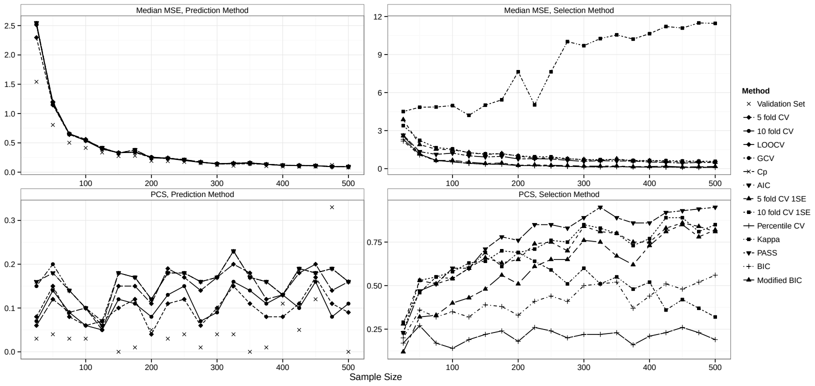

***Figure 2**: As the sample size increases, the median MSE (top) and PCS (bottom) for the prediction optimal methods (left) and methods best for selection purposes (right).*

Figure 2 depicts the selection and prediction consistency of the LASSO when using each method. The left panels show the prediction optimal methods and the right panels show methods used to recover the correct subset. The median MSE is shown in the top panels. The MSE is similar for all the prediction methods and the persistence of the LASSO is clear. The LASSO also appears to achieve persistence when using the percentile CV, BIC and the modified BIC. The MSE is slightly higher when using CV with the 1SE rule or PASS but the kappa coefficient yields highly inaccurate models. The bottom panels show the PCS. It is clear that the prediction methods do not achieve selection consistency. Among the selection methods, PASS is selection consistent, while CV with the 1SE rule and the modified BIC also look promising. The ordinary BIC, percentile CV and kappa do not appear consistent for selection with the LASSO. Table 2 and Figure 3 show the simulation results for a small sample $n = 25$ and a large sample $n = 200$. The accuracy of the median MSE is assessed by bootstrap standard errors, calculated using 200 bootstrap replications.

The prediction methods all produce models with low MSE and their performance is excellent when $n = 200$. Using a large validation set slightly outperforms any of the other methods. Although they perform poorly in choosing the correct model, the chosen model contains the correct model with high probability. In each case, roughly two noise variables are included in the model.

| | $n = 25$ | | | | | $n = 200$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | $\lambda$ | DF | MSE (SE) | PCS | PIS | $\lambda$ | DF | MSE (SE) | PCS | PIS |
| | Prediction Methods | | | | | | | | | |
| Validation Set | 2.20 | 5.23 | 1.54 (0.16) | 0.03 | 0.90 | 1.83 | 5.68 | 0.19 (0.01) | 0.05 | 1.00 |
| 5 fold CV | 1.77 | 5.27 | 2.30 (0.31) | 0.08 | 0.86 | 1.76 | 5.48 | 0.24 (0.03) | 0.04 | 1.00 |
| 10 fold CV | 1.61 | 5.27 | 2.52 (0.24) | 0.07 | 0.85 | 1.77 | 5.34 | 0.26 (0.02) | 0.08 | 1.00 |
| LOOCV | 1.50 | 5.36 | 2.51 (0.16) | 0.06 | 0.87 | 1.56 | 5.25 | 0.25 (0.02) | 0.11 | 1.00 |
| GCV | 1.54 | 4.72 | 2.55 (0.26) | 0.15 | 0.86 | 1.38 | 5.15 | 0.24 (0.02) | 0.12 | 1.00 |
| Cp | 1.47 | 4.83 | 2.55 (0.27) | 0.16 | 0.88 | 1.38 | 5.15 | 0.24 (0.02) | 0.12 | 1.00 |
| AIC | 1.47 | 4.83 | 2.55 (0.24) | 0.16 | 0.88 | 1.38 | 5.15 | 0.24 (0.02) | 0.12 | 1.00 |
| | Selection Methods | | | | | | | | | |
| 5 fold CV 1SE | 5.66 | 3.53 | 3.87 (0.46) | 0.28 | 0.67 | 8.71 | 3.45 | 0.99 (0.09) | 0.65 | 1.00 |
| 10 fold CV 1SE | 5.22 | 3.69 | 3.39 (0.33) | 0.23 | 0.69 | 8.87 | 3.37 | 1.01 (0.08) | 0.69 | 1.00 |
| Percentile CV | 2.94 | 4.24 | 2.32 (0.24) | 0.17 | 0.81 | 3.01 | 4.56 | 0.24 (0.02) | 0.18 | 1.00 |
| Kappa | 6.80 | 3.11 | 4.51 (0.62) | 0.29 | 0.57 | 25.1 | 2.65 | 7.64 (0.92) | 0.69 | 0.70 |
| PASS | 4.18 | 3.73 | 2.65 (0.42) | 0.23 | 0.72 | 8.19 | 3.29 | 0.77 (0.08) | 0.76 | 1.00 |
| BIC | 2.18 | 4.12 | 2.18 (0.28) | 0.20 | 0.80 | 2.93 | 4.01 | 0.26 (0.02) | 0.33 | 1.00 |
| Modified BIC | 1.18 | 5.14 | 2.59 (0.21) | 0.12 | 0.88 | 4.14 | 3.59 | 0.26 (0.03) | 0.51 | 1.00 |

***Table 2**: Small and large sample results showing the average $\lambda$ and DF, median MSE and its bootstrap standard error, PCS and PIS.*

The selection methods choose a larger value for $\lambda$ in order to set more coefficients to zero. The percentile CV does not perform significantly better than the prediction methods in recovering the correct model. The BIC and modified BIC methods have similar MSE to the prediction methods but perform variable selection considerably better when $n = 200$, with the modified BIC outperforming the original BIC. However, they do tend to include at least one noise variable 50% of the time. Using CV with the 1SE rule does improve variable selection but has larger and more variable MSE. The kappa coefficient performs similarly to the 1SE rule in terms of variable selection but it tends to underestimate the model 30% of the time and is extremely variable in terms of prediction accuracy. PASS yields the best selection performance with lower MSE than the 1SE rule.

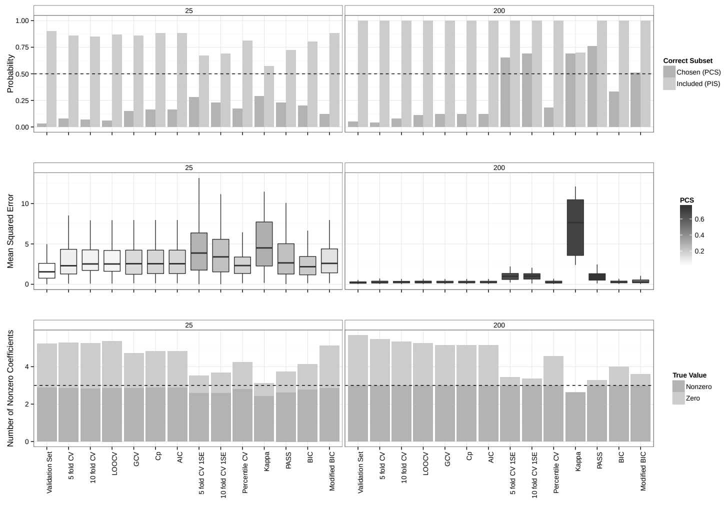

***Figure 3**: Top: PCS (dark) and PIS (light). Middle: Box plots of MSE with fill corresponding to PCS. Bottom: Number of coefficients estimated to be non-zero correctly (dark) and incorrectly (light).*

## 6. Concluding Remarks

The LASSO is persistent with the use of methods like CV, GCV or AIC. The PASS method and using the 1SE rule with CV improves its variable selection performance with only slightly larger MSE. However, the estimates obtained using the other selection methods suffer from large bias. To avoid the bias, the LASSO could be used for variable selection followed by least squares for estimation (Hastie et al., 2009). Alternatively, two-stage LASSO methods have been proposed, including the adaptive LASSO (Zou, 2006) and relaxed LASSO (Meinshausen, 2007). Other work focusing on variable selection with the LASSO is the $p$-value method (Meinshausen, 2009) and stability selection (Meinshausen and Bühlmann, 2010). Also worth mentioning is the significance test of Lockhart, Taylor, Tibshirani and Tibshirani (2014) which can be used as a stopping rule for the LAR algorithm. Further developments may still be necessary before these methods have mainstream appeal. Better standard errors of estimates and confidence intervals remain a topic for further research. Testing the overall significance of a predictor and the goodness of fit of the model still need to be uncovered.

## References

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. *In* PETROV, B. N. AND CSAKI, F. (Editors) *Second International Symposium on Information Theory*. Akademiai Kiado, Budapest, pp. 267–281.

BICKEL, P., RITOV, Y., AND TSYBAKOV, A. (2009). Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, **37**, 1705–1732.

BREIMAN, L., FRIEDMAN, J. H., OHLSEN, R. A., AND STONE, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton, FL, USA.

BÜHLMANN, P. AND VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, Heidelberg, Germany.

CHAND, S. (2012). On tuning parameter selection of lasso-type methods - a Monte Carlo study. *In IBCAST 2012: Proceedings of the 9th International Bhurban Conference on Applied Sciences & Technology*. pp. 120–129.

DOSSAL, C., KACHOUR, M., FADILI, J., PEYRE, G., AND CHESNEAU, C. (2013). The degrees of freedom of the LASSO for general design matrix. *Statistica Sinica*, **23**, 809–828.

EFRON, B., HASTIE, T., JOHNSTONE, I., AND TIBSHIRANI, R. (2004). Least angle regression. *The Annals of Statistics*, **32** (2), 407–451.

FAN, Y. AND TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75** (3), 531–552.

FANG, Y., WANG, J., AND SUN, W. (2013). A note on selection stability: combining stability and prediction. Technical report, New York University, University of Illinois and Purdue University.

FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33** (1), 1–22.

FU, W. J. (1998). Penalized regressions: The bridge versus the LASSO. *Journal of Computational and Graphical Statistics*, **7** (3), 397–416.

GREENSHTEIN, E. AND RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, **10** (6), 971–988.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics, 2nd edition. Springer, New York, NY, USA.

HOMRIGHAUSEN, D. AND MCDONALD, D. (2013). The lasso, persistence, and cross-validation. *In ICML 2013: Proceedings of the 30th International Conference on Machine Learning*. pp. 103–1039.

HOMRIGHAUSEN, D. AND MCDONALD, D. J. (2014). Leave-one-out cross-validation is risk consistent for lasso. *Machine Learning*, **97**, 65–78.

LOCKHART, B. R., TAYLOR, J., TIBSHIRANI, R. J., AND TIBSHIRANI, R. (2014). A significance test for the LASSO. *The Annals of Statistics*, **42** (2), 413–468.

MALLOWS, C. L. (1973). Some comments on Cp. *Technometrics*, **15** (4), 661–675.

MEINSHAUSEN, N. (2007). Relaxed LASSO. *Computational Statistics & Data Analysis*, **52** (1), 374–393.

MEINSHAUSEN, N. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, **104**, 1671–1681.

MEINSHAUSEN, N. AND BÜHLMANN, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 417–473.

MEINSHAUSEN, N. AND BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the LASSO. *The Annals of Statistics*, **34** (3), 1436–1462.

ROBERTS, S. AND NOWAK, G. (2014). Stabilizing the lasso against cross-validation variability. *Computational Statistics & Data Analysis*, **70**, 198–211.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6** (2), 461–464.

SEBER, G. A. F. AND LEE, A. J. (2003). *Linear Regression Analysis.* Wiley Series in Probability and Statistics, 2nd edition. John Wiley & Sons, Hoboken, NJ, USA.

SUN, W., WANG, J., AND FANG, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research*, **14**, 3419–3440.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58** (1), 267–288.

TIBSHIRANI, R. J. AND TAYLOR, J. (2012). Degrees of freedom in LASSO problems. *The Annals of Statistics*, **40** (2), 1198–1232.

WANG, H., LI, B., AND LENG, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71** (3), 671–683.

YUAN, M. AND LIN, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69** (2), 143–161.

ZHAO, P. AND YU, B. (2006). On model selection consistency of LASSO. *Journal of Machine Learning Research*, **7**, 2541–2563.

ZOU, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, **101** (476), 1418–1429.

ZOU, H., HASTIE, T., AND TIBSHIRANI, R. (2007). On the "degrees of freedom" of the LASSO. *The Annals of Statistics*, **35** (5), 2173–2192.