# Variable Selection and Regularization

**Xiaogang Su, Ph.D.**
Department of Mathematical Sciences
University of Texas at El Paso (UTEP)
xsu@utep.edu

April 2, 2018

## Contents

**1 Why Variable Selection? 2**

**2 All Possible Regressions 3**
2.1 PRESS and GCV . . . 4
2.2 Mellow's $C_p$ . . . 4
2.3 AIC and BIC . . . 5

**3 Stepwise Procedures 6**

**4 Regularization 7**
4.1 $\ell_0$-Regularization . . . 9
4.2 LASSO and Other Extensions . . . 9
4.2.1 LASSO . . . 9
4.2.2 Adaptive LASSO and Others . . . 11
4.3 SCAD and MCP . . . 12
4.4 MIC . . . 13

**A Quick Derivation of BIC 15**

# 1 Why Variable Selection?

In the previous linear model specification, we have assumed that the true regression function $\boldsymbol{\mu} = [E(y_i|\mathbf{x}_i)] = [\mu(\mathbf{x}_i)]$ is in the linear form specified by the the linearity assumption $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. This is unlikely to be true in reality, where model misspecification can occur in various ways. The underlying regression function $\mu(\cdot)$ can be curvilinear. Even if it is linear, model specification is still under the risk of overfitting or underfitting or both, meaning that important predictors have been missed out or irrelevant variables are included in the model.

To study the adverse effects of underfitting and overfitting on model estimation and prediction, a simplified setting is employed by partition the columns of $\mathbf{X}$ into $(\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ is $n \times (k+1)$ and $\mathbf{X}_2$ is $n \times (p-k)$. Rewrite model $\mathbf{y} = +\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}. \tag{1}$$

At the same time, we consider a reduced model that uses $\mathbf{X}_1$ only

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}. \tag{2}$$

Note that we have sightly abused notations by not distinguishing $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$, as well as the error variance $\sigma^2$, between the above two models. With this setting, *underfitting* is committed when model (2) is fit whereas (1) is the true model. Let $\hat{\boldsymbol{\beta}}_1$ denote the LSE of $\boldsymbol{\beta}_1$ obtained from fitting (2). It can be shown that $\hat{\boldsymbol{\beta}}_1$ is biased for $\boldsymbol{\beta}_1$ in model (1), i.e., $E(\hat{\boldsymbol{\beta}}_1) \neq \boldsymbol{\beta}_1$, although it has a smaller variance than the LSE of $\boldsymbol{\beta}_1$ obtained from fitting the true model (1). On the other hand, *overfitting* is committed if we fit model (1) whereas the true model is (2). In this case, the LSE $\boldsymbol{\beta}_1$, obtained as a subcomponent of the LSE from fitting (1), remains unbiased for $\boldsymbol{\beta}_1$ in model (2); however, its variance are inflated when compared to the LSE of $\boldsymbol{\beta}_1$ obtained from fitting the true model(2).

In sum, underfitting leads to bias while overfitting inflates variance. The same conclusion can be arrived from model prediction. Regression usually has two goals, either for predicting future observations or for studying the relationship between the response and predictors. The latter goal is more related to model interpretation. While one is sometimes more emphasized than the other in specific applications, these two goals are closely related with each other. Reliable interpretation should be based on a model that generalizes well to further observations. Consider prediction at one new observation $(y_0, \mathbf{x}_0)$ from true model $y = m(\mathbf{x}) + \varepsilon = \mathrm{E}(y|\mathbf{x}) + \varepsilon$ with $\varepsilon \sim (0, \sigma^2)$. Let $\hat{y}_0 = \mathbf{x}_0^T\hat{\boldsymbol{\beta}}$ denote the predicted value based on a fitted linear model. Then the prediction mean squared error (MSE) of $\hat{y}_0$ can be decomposed as

$$\begin{aligned} \mathrm{E}(y_0 - \hat{y}_0)^2 &= \mathrm{E}_{\mathbf{x}_0, y_0}\left\{y_0 - m(\mathbf{x}_0)\right\}^2 + \mathrm{E}_{\mathbf{x}_0}(m(\mathbf{x}_0) - E(\hat{y}_0))^2 + \mathrm{var}(\hat{y}_0) \\ &= \sigma^2 + \mathrm{bias}^2 + \mathrm{var}, \end{aligned}$$

which essentially involves the squared bias of $\hat{y}_0$ and its variance. An overfitted model tends to provide prediction with a larger variance in spite of a smaller bias while an underfitted model tends to provide prediction with a smaller variance yet with a larger bias, a phenomenon often referred to as the "bias-variance tradeoff". A reasonably good prediction with a small MSE balances off between bias and variance.

An empirical illustration of the bias-variance tradeoff is made as follows. Let $\mathcal{D}_0$ denote a new data set consisting of future observations, which are independent of current data $\mathcal{D}$. We fit a number

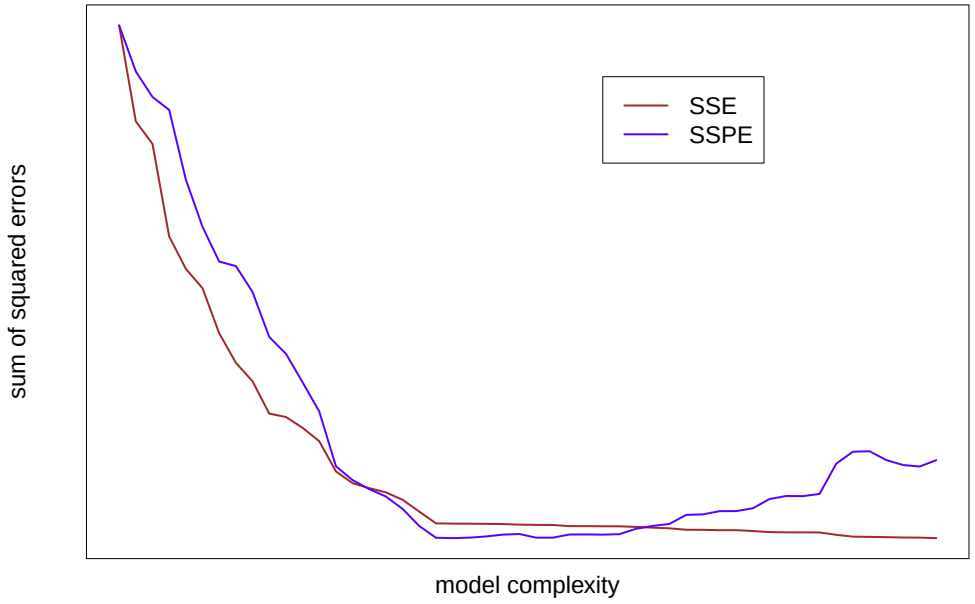


Figure 1: Illustration of the Bias-Variance Tradeoff.

of nested models with increasingly more predictors included, gradually ranging from underfitting to overfitting scenarios, and compute the resultant sum of squared errors for prediction (SSPE) with $\mathcal{D}_0$. It is important to distinguish between

$$\text{SSPE} = \sum_{i \in \mathcal{D}_0} (y_i - \hat{y}_i)^2 \quad \text{and} \quad \text{SSE} = \sum_{i \in \mathcal{D}} (y_i - \hat{y}_i)^2.$$

Figure (1) plots SSE and SSPE versus the model complexity (measured by number of parameters used in the model) based on a simulated data. It can be seen that SSE always decreases with more predictors added in, even they have no predictive power. However, SSPE decreases as important variables are gradually added in, hits its minimum around the best model, and then starts to increase when irrelevant variables are included. The graph also suggests that underfitting causes more concerns than overfitting if prediction is the primary goal. This is because the inflation amount in SSPE caused by slightly overfitting is relatively smaller than that caused by underfitting. Nevertheless, a simpler model is much easier to interpret.

The goal of model selection is to find a parsimonious model that does reasonably well in prediction. There are three groups of methods for this task, which are discussed in order.

## 2 All Possible Regressions

The method of *all possible regressions* tries out all possible subsets of predictors and select the best according to some model selection criteria. Note that there are $2^p$ model choices to consider.

Clearly this method only applies to scenarios when $p$ is small, although there are some methods designed to reduce the computational burden. For example, the *best subsets algorithms* attempt to sort out good model choices while avoiding the evaluation of all models.

Given a model with $k$ predictors and predictor space $\mathbb{V}_0$, a few popular model selection criteria for evaluating its model performance are listed below.

## 2.1 PRESS and GCV

The first criterion is the SSPE by using an additional independent sample. However, this method is only available when data are rich. When data are limited, an approximate version of SSPE can be computed via cross-validation (CV) such as the v-fold CV or jackknife technique. One commonly used criterion, PRESS for *prediction sum of squares* (Allen, 1974), is computed via the leave-on-out or jackknife technique, in which each observation is left out in turn and its prediction is computed using the remaining $(n-1)$ observations.

$$\text{PRESS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(-i)}\right)^2, \tag{3}$$

where $\hat{y}_{(-i)} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(-i)}$ denotes the predicted value for $y_i$ by least squares fit on data that leave the $i$th observation out and $\hat{\boldsymbol{\beta}}_{(-i)}$ denotes the resultant LSE of $\boldsymbol{\beta}$. PRESS can be conveniently computed from the LS fit with the whole data

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2, \tag{4}$$

where $h_i$ is the $i$th diagonal element of the projection matrix $\mathbf{H}$. This is because

$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1} \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix} \frac{e_i}{1 - h_i}.$$

Replacing $h_i$'s by their average $\text{trace}(\mathbf{H})/n$ in $\text{PRESS}/n$, Craven and Wahba (1979) obtained the generalized cross-validation (GCV) criterion

$$\text{GCV} = \frac{n \cdot \text{SSE}}{\{n - \text{trace}(\mathbf{H})\}^2} = \frac{n \cdot \text{SSE}}{\{n - (k+1)\}^2}. \tag{5}$$

GCV is extensively used in modern regression methods.

## 2.2 Mellow's $C_p$

Mallow (1973) derived a $C_p$ criterion by examining the expected model error. Given a linear model choice with predictor space $\mathbb{V}_0$, recall that $\text{SSE} = \| \mathbf{y} - \hat{\mathbf{y}} \|^2 = \mathbf{y}^T \text{P}_{\mathbb{V}_0} \mathbf{y}$ provide a measure of the empirical distance between the observed and predicted responses with

$$E(\text{SSE}) = \boldsymbol{\mu}^T \text{P}_{\mathbb{V}_0} \boldsymbol{\mu} + (n - k - 1)\sigma^2. \tag{6}$$

The expected model error is concerned about

$$\begin{aligned} E \| \boldsymbol{\mu} - \hat{\mathbf{y}} \|^2 &= E \| \boldsymbol{\mu} - \text{P}_{\mathbb{V}_0} \mathbf{y} \|^2 = E \| \boldsymbol{\mu} - \text{P}_{\mathbb{V}_0} (\boldsymbol{\mu} + \boldsymbol{\varepsilon}) \|^2 \\ &= \boldsymbol{\mu}^T \text{P}_{\mathbb{V}_0^\perp} \boldsymbol{\mu} + (k+1)\sigma^2 = E(\text{SSE}) + \{2(k+1) - n\}\, \sigma^2 \end{aligned}$$

The $C_p$ criterion is defined as an estimate of $E \parallel \boldsymbol{\mu} - \hat{\mathbf{y}} \parallel^2 / \sigma^2$,

$$C_p = \frac{\text{SSE}}{\hat{\sigma}^2} + 2(k+1) - n, \tag{7}$$

where $\hat{\sigma}^2$ is hoped to be a reliable estimate of the true error variance $\sigma^2$. In common practice, $\hat{\sigma}^2$ is obtained from the full model with all predictors included. If a model fits well so that $\boldsymbol{\mu} \in \mathbb{V}_0$ approximately, then $E(C_p) \approx k+1$. Mellow suggested plotting $C_p$ versus $k$ for all possible models and considering models with $C_p \approx k+1$ as favorable choices.

### 2.3 AIC and BIC

Akaike (1974) derived a criterion from information theories, known as Akaike information criterion (AIC). As an approximation to the Kullback-Leibler discrepancy function between a candidate model distribution and the true model distribution, AIC, given by

$$\text{AIC} \simeq n \cdot \log(\text{SSE}) + 2 \cdot k$$

up to a constant, penalizes the goodness-of-fit with model complexity. The GCV criterion can be easily shown to be asymptotically equivalent to AIC if $\lim_{n\to\infty} k/n = 0$, by considering

$$\begin{aligned} \log(GCV) &\simeq \log(SSE) - 2\log\left(1 - \frac{k+1}{n}\right) \\ &\simeq \log(SSE) + 2\,(k+1)/n, \text{ using } \log(1+x) = x \text{ as } x \approx 0 \\ &\simeq n\log(SSE) + 2\,(k+1) \end{aligned}$$

where the symbol '$\simeq$' is used for 'up to some constant' so that irrelevant terms could be added or eliminated freely. In the above derivation, we assume $n \gg k$ so that $(k+1)/n \approx 0$.

In fact, AIC was the first model selection criterion derived from information theory. Let $f(y)$ denote the true density of $y$ and $g(y|\boldsymbol{\theta})$ denote the density of $y$ under the approximating model, which is the linear model $y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$ in linear regression with $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$. The Kullback-Leibler (KL) divergence between $f(y)$ and $g(y|\boldsymbol{\theta})$ is

$$\begin{aligned} D(f\|g(\cdot|\boldsymbol{\theta}) &= \int_{\mathcal{Y}} f(y)\log\frac{f(y)}{g(y|\boldsymbol{\theta}}dy \\ &= \int_{\mathcal{Y}} f(y)\log f(y)dy - \int_{\mathcal{Y}} f(y)\log g(y|\boldsymbol{\theta})dy, \end{aligned}$$

where the last term is the relative information. In practice, the parameter $\boldsymbol{\theta}$ must be estimated from data $\mathcal{D}$; replacing $\boldsymbol{\theta}$ with its estimator $\widehat{\boldsymbol{\theta}}$ leads to $D(f\|g(\cdot|\widehat{\boldsymbol{\theta}})$, which is random involving data $\mathcal{D}$. The best approximating model would solve

$$\min_{g\in\mathcal{F}} \mathrm{E}_{\mathcal{D}}\, D(f\|g(\cdot|\widehat{\boldsymbol{\theta}}) \quad \text{or, equivalently,} \quad \max_{g\in\mathcal{F}} \mathrm{E}_{\mathcal{D}}\mathrm{E}_y \log(g(y|\widehat{\boldsymbol{\theta}}).$$

Akaike (1973) obtained $\log L(\widehat{\boldsymbol{\theta}}) - k$, which essentially leads to $AIC = -2\log L(\widehat{\boldsymbol{\theta}}) + 2k$, as an approximately unbiased estimate of the relative information $\mathrm{E}_{\mathcal{D}}\mathrm{E}_y \log(g(y|\widehat{\boldsymbol{\theta}})$ if $n$ is large and $g(\cdot|\boldsymbol{\theta})$ is a 'good' model close to the true model $f(y)$.

Within the Bayesian framework, Schwarz (1978) developed a Bayesian information criterion (BIC), given by

$$\text{BIC} \simeq n \cdot \log(\text{SSE}) + \log(n) \cdot k$$

up to a constant. Since $\log(n) \geq 2$ for $n \geq 8$, BIC applies a larger penalty for model complexity.

In large samples, a model selection criterion is said to be asymptotically *efficient* if it selects the model with minimum mean squared error, and *consistent* if it selects the true model with probability one. To be more precise, there are different assumptions involved in either definition. For asymptotic efficiency, it is assumed that the true underlying model is nonparametric, unspecified, or infinitely dimensional so that it lies outside of the set of candidate models. For asymptotic consistency, it is assumed that the true mode is of finite dimension and is included in the set of candidate models. It has been shown (Yang, 2005; *Biometrika*) that no criterion could be both consistent and efficient. Based on this categorization, PRESS, GCV , $C_p$, and AIC are efficient while BIC is consistent. Empirically PRESS, GCV, and AIC work well with moderately-sized samples while BIC works best with large samples with strong signals.

# 3 Stepwise Procedures

Stepwise procedure can be viewed as a surrogate algorithmic approach to the best subset selection. When $p$ is large, stepwise procedures offer a more feasible method for variable selection by adding or removing variables one at a time. In each step of the procedure, comparison is made only among models that has the same complexity. Such a comparison can be simply based on SSE. It is worth noting that dummy variables created for explaining one categorical predictor can be treated as individual variables in the selection process. This essentially involves level merging. Alternatively, this set of dummy variables can be bound together so that we either drop or include them all. In this case, either $F$ test or model selection criterion such as $AIC$ or $BIC$ can be used.

Stepwise procedures can be executed in three ways: backward elimination, forward addition, or stepwise selection. Suppose that $F$ test is used for model comparisons, as implemented in SAS (2011). In backward elimination, one starts with the whole model with all predictors included and removes the least significant variable at each step till predictors remaining in the model are all significant. In forward addition, one starts with the null model and adds the most significant variable at each step till no additional variable can be significant in the current model. Note that any predictor that has been removed in backward elimination has no chance to reenter the model even if its addition to the current model became significant. Similarly, any predictor that has been added in forward addition will not be removed even if its effect becomes insignificant in the current model. Stepwise selection is meant to to correct these problem. It resembles the forward addition, but takes one extra check at each step to remove insignificant variables from the current model. In terms of computational speed, backward elimination is the fastest, followed by forward addition. Yet stepwise selection offers the best performance comparatively.

Despite its popular use in applications, stepwise procedures have been widely recognized as suboptimal in the statistical society. Due to the multiplicity issue and lack of validation, the selected model is under considerable risk of misidentification and often does not generalize well to new data.

# 4 Regularization

The third group of methods is regularization or shrinkage. In the methods of all possible regressions and stepwise procedures, variable selection is a discrete process, in which a variable is either used or unused. Shrinkage methods does the variable selection in a continuous fashion.

In fact, the best subset selection can be formulated into a regularization problem with the so-called $\ell_0$ norm penalty. Common shrinkage methods optimize the least squares criterion while shrinking the size or length of the regression coefficients. One motivation for this approach is that $E \parallel \hat{\boldsymbol{\beta}} \parallel^2 \geq \parallel \boldsymbol{\beta} \parallel^2$ despite LSE $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$. In general, a regularized or penalized estimator $\tilde{\boldsymbol{\beta}}$ in linear regression can be stated as follows

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} g(|\beta_j|) \leq t, \tag{8}$$

for some convex function $g(\cdot)$ and constant $t$. Note that the intercept term $\beta_0$ can be suppressed by working with centered data. This is the BRIDGE regression introduced by Frank and Friedman (1993). In its Lagrangian form,

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} g(|\beta_j|) \right\}, \tag{9}$$

where $\lambda > 0$ is a penalty or regularization parameter that controls the amount of the shrinkage.

Power functions $g(x) = x^q$ with $q \geq 0$ is most often used. $\tilde{\boldsymbol{\beta}}$ corresponds to LSE when $q = 0$; *lasso* (Tibshirani, 1996) estimator when $q = 1$; and the ridge estimator (Hoerl and Kennard, 1970) when $q = 2$. The ridge solution has a simple form

$$\tilde{\boldsymbol{\beta}}_{L_2} = \left( \mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{y}. \tag{10}$$

It can be seen that $\tilde{\boldsymbol{\beta}}_{L_2}$ is biased for $\boldsymbol{\beta}$. However, its MSE can be smaller than that of $\hat{\boldsymbol{\beta}}$ with appropriate choice of $\lambda$. Furthermore, ridge estimator can be minimax under some conditions (Casella, 1980). While the lasso estimator $\tilde{\boldsymbol{\beta}}_{L_1}$ does not have an explicit form, its entire solution path for any $\lambda$ can be efficiently obtained via the LARS (Efron et al., 2004) algorithm. The GCV criterion in (5) is often used to determine the optimal $\lambda$ in both ridge regression and lasso.

From the perspective of predictive modeling, variable selection is not necessary with regularization, except for the determination of the penalty parameter $\lambda$. In applications where the effects of predictors are under study, inclusion of irrelevant variables complicate model interpretability, although their effects are shrunk. Both ridge regression and lasso usually provide competitive predictive performance. However, there is a critical difference between the ridge and the lasso estimators. As $\lambda$ increases, Lasso, as well as its various variants, effectively does variable selection by setting some coefficients to be exactly zero.

To gain insight, first observe that

$$\begin{aligned} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 &= \parallel \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta} \parallel^2 \\ &= \parallel \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \parallel^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \end{aligned}$$

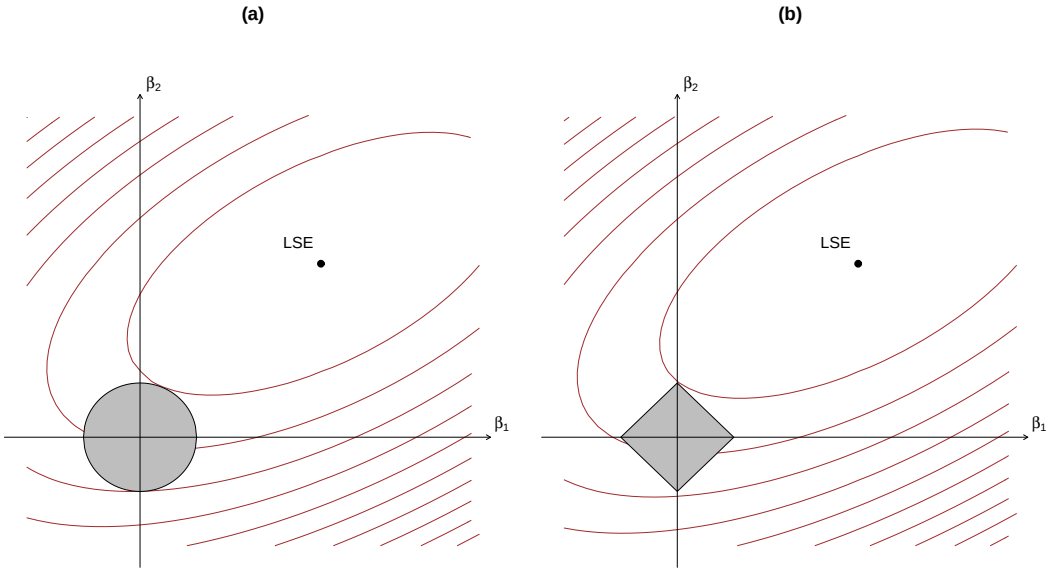


Figure 2: Illustration of shrinkage Estimators in the two-dimensional case: (a) ridge (q=2) and (b) lasso (q=1).

where the first term does not involve $\boldsymbol{\beta}$. Thus we can rewrite the optimization problem in (8) as

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j|^q \leq t. \tag{11}$$

The objective function is a hyper ellipsoid centered at the LSE $\hat{\boldsymbol{\beta}}$, while the constrain is a disk when $q = 2$ and a diamond when $q = 1$. A graphical illustration for the two dimensional case is given in Figure 2. In either case of ridge and lasso, the solution occurs when the elliptical contours hit the constraint region. Compared to the disk case, the diamond constraint region has corners and is more likely to have solution at a corner. When this happens, one parameter estimate becomes zero. When $q > 2$ in (8), the constraint region becomes a polyhedron with many corners, flat edges, and faces, and hence it is more likely to have zero-valued coefficients.

The lasso method has been shown quite successful in both predictive modeling and variable selection. Since it inception, intensive research efforts have been devoted to this direction. Various lasso variants have been developed and shown to be consistent in both variable selection and estimation. However, there has not been much optimality result established. Thus, as recommended by Efron et al. (2004), it is advisable to use lasso for variable selection only and switch back to the LSE for model interpretation once the best model is identified.

### 4.1 $\ell_0$-Regularization

The best subset selection can be viewed a regularization problem that solves

$$\min_{\boldsymbol{\beta}} \quad \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2 \quad \text{s.t.} \quad \| \boldsymbol{\beta} \|_0 \leq k, \tag{12}$$

where $\| \cdot \|_0$ is the $\ell_0$ norm defined as the number of nonzero components in $\boldsymbol{\beta}$ or its cardinality, namely,

$$\| \boldsymbol{\beta} \|_0 = \text{card}(\boldsymbol{\beta}) = \sum_{j=1}^{p} I\{\beta_j \neq 0\}.$$

This is a nonconvex, discrete, NP-hard problem. As the tuning parameter $k$ goes from 0 to $p$, essentially all possible models are examined. Alternatively, its Lagrangian form can be considered

$$\min_{\boldsymbol{\beta}} \quad \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2 + \lambda \| \boldsymbol{\beta} \|_0, \tag{13}$$

To choose the best tuning parameter $k^\star$ or $\lambda^\star$, one needs to resort to minimum GCV, AIC, or BIC.

There are faster algorithms available for solving (12). The first one is the branch-and-bound method (Furnival and Wilson, 1974), which is an combinatorial optimization method. The second one is iterative hard thresholding (IHT) algorithm (Blumensath and Davies, 2009) that iterates with updating formula

$$\boldsymbol{\beta} \ := \ H_k\{\boldsymbol{\beta} + \alpha\, g(\boldsymbol{\beta})\},$$

where $g(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is the gradient evaluated at the current estimate $\boldsymbol{\beta}$; $\alpha > 0$ is the step size; and $H_k(\cdot)$ is an hard threshold operator that takes only the first $k$ $\beta_j$'s of largest magnitude (i.e., largest absolute value) and nullifies all $(p-k)$ remaining ones. There are strategies for determining the step size; one common way is to optimize the objective function with respect to $\alpha$ treating other parameters $\boldsymbol{\beta}$ fixed at their current values (which is a one-dimensional optimization problem). It can be shown that IHT converges to a local optimum for (12).

Despite all the above efforts, the best subset selection method is not feasible for even moderately large $p$. It is worth nothing that what is essentially involved in (12) and (13) is the indicator penalty function $I\{\beta_j \neq 0\}$ in the $\ell_0$ norm. Figure 3 plots this $\ell_0$ penalty function, as well as a few continuous or smooth relaxations or approximations, which will be discussed in the following. The $\ell_1$ or $\ell_2$ regularization can be viewed as convex relaxation of the problem.

### 4.2 LASSO and Other Extensions

The LASSO or lasso (least absolute shrinkage and selection operator) is another shrinkage method like ridge regression, yet with an important and attractive feature in variable selection.

#### 4.2.1 LASSO

To motivate, we continue with the comparison between ridge regression and PCR from the variable selection perspective. PCR, same as subset selection procedures, is a discrete selecting process - regressors or the principal components of them are either fully retained or completely dropped from the model. Comparatively, the ridge regression makes the selection process continuous by varying shrinkage parameter $k$ and hence is more stable. On the other hand, since ridge regression does

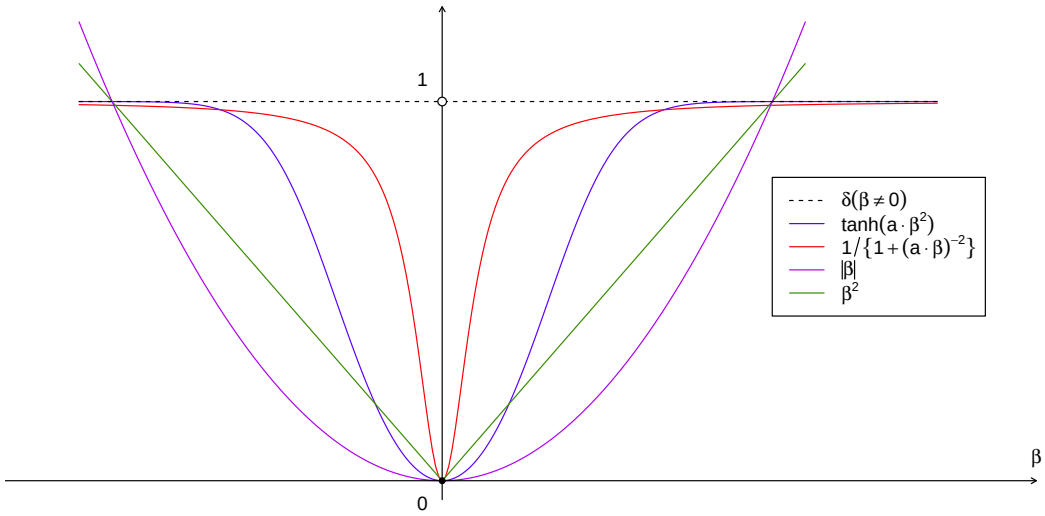


Figure 3: The penalty function $I\{\beta_j \neq 0\}$ in $\ell_0$-regularization, together with smooth approximation and convex relaxation.

not set any coefficients to 0, it does not give an easily interpretable model as in subset selection. The lasso technique is intended to balance off in between and retains the favorable features of both subset selection and ridge regression by shrinking some coefficients and setting others to 0.

The lasso estimator of $\boldsymbol{\beta}$ is obtained by

$$\text{minimizing } \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2, \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s. \tag{14}$$

Namely, the $L_2$ penalty $\sum_j \beta_j^2$ in ridge regression is replaced by the $L_1$ penalty $\sum_j |\beta_j|$ in lasso. If $s$ is chosen greater than or equal to $\sum_j |\beta_j^{\text{LS}}|$, then the lasso estimates are the same as the LSE; if $s$ is chosen to be smaller, then it will cause shrinkage of the solutions towards 0.

The lasso solution is generally competitive with ridge solution yet with many zero coefficient estimates. Insight about the nature of the lasso can be further gleaned from orthonormal designs where $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. In this case, the lasso estimator can be shown to be

$$\hat{\beta}_j^{\text{lasso}} = \text{sign}(\hat{\beta}_j^{\text{LS}}) \left\{ |\hat{\beta}_j^{\text{LS}}| - \gamma \right\}_+, \tag{15}$$

where $\gamma$ is determined by the condition $\sum_j |\hat{\beta}_j^{\text{lasso}}| = s$. Thus, coefficients less than the threshold $\gamma$ would be automatically suppressed to 0 while coefficients larger than $\gamma$ would be shrunk by a unit of $\gamma$. Hence, the lasso technique performs as a variable selection operator. By increasing $s$ in discrete steps, one obtains a sequence of regression coefficients where those nonzero coefficients at each step correspond to selected predictors.

### 4.2.2 Adaptive LASSO and Others

Extensions and further explorations of LASSO are currently under intensive research. Zou (2006) proposed an 'adaptive LASSO' method with slight changes yet impressively improvements over LASSO. These improvements include both better empirical and theoretical variable selection. The path from adaptive lasso can be conveniently obtained from the LARS algorithm without added difficulty. It empirically works better than lasso by not penalizing the regression coefficients for important predictors. See Figure 4 for a comparison of the threshold functions used in lasso, SCAD, and the adaptive lasso. Moreover, using the similar techniques in Fan and Li (SCAD; 2001), Zou proved the 'oracle' property of adaptive lasso, i.e., asymptotic consistency in correct variable selection with large samples when selecting the penalty parameter $\lambda \equiv \lambda(n)$ appropriately.

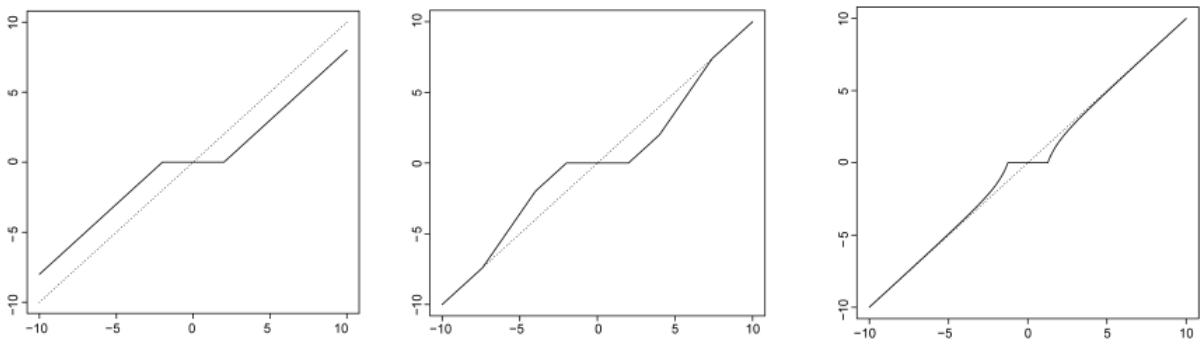


Figure 4: Comparison of the Thresholding Functions used in LASSO, SCAD, and Adaptive LASSO with $\gamma = 2$ (from left to right).The figure is taken from Zou (*JASA*, 2006).

The adaptive lasso solution $\tilde{\boldsymbol{\beta}}$ is generally formulated as follows

$$\arg\min_{\boldsymbol{\beta}} \quad \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \cdot \sum_{j=1}^{p} w_j |\beta_j|, \tag{16}$$

where $\mathbf{w} = (w_j)$ is a *known* weights vectors. Typically $\hat{\mathbf{w}} = \left|\hat{\boldsymbol{\beta}}\right|^{-\gamma}$ for some $\gamma > 0$ and $\gamma = 1$ is most often used. Again, recall that $\hat{\boldsymbol{\beta}}$ denotes the LSE of $\boldsymbol{\beta}$. It is an immediate question how this would work with ultra-high dimensional data where $\hat{\boldsymbol{\beta}}$ is not easily available owing to multi-collinearity and high dimension.

To solve (19) computationally, one can modify the covariates $\tilde{x}_{ij} := x_{ij}/w_j$ for $i = 1, \ldots, n$ and

$j = 1, \ldots, p$. Then the adaptive lasso problem can be solved in two steps: first finding $\tilde{\tilde{\boldsymbol{\beta}}}$ as

$$\arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta} \right\|^2 + \lambda \cdot \sum_{j=1}^{p} |\beta_j|, \tag{17}$$

and then $\tilde{\beta}_j = \tilde{\tilde{\beta}}_j / w_j$ for $j = 1, \ldots, p$.

Other interesting proposed versions of modified $L_1$ regularization are listed below. Tibshirani et al. (2005) proposed the fused lasso, which is formulated as follows

$$\text{minimizing } \left\| \mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta} \right\|^2 \quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \le s_1 \text{ and } \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \le s_2. \tag{18}$$

One limitation of LASSO is that the number of variables selected by the lasso is limited by the number of observations. In other words, lasso can only select at most $n$ genes in $n \times p$ gene expression data ($n \ll p$). To address this problem, Zou and Hastie (2005) proposed the *elastic net* estimator $\hat{\boldsymbol{\beta}}_{en}$ as a combination of a lasso and ridge estimate:

$$\underset{\boldsymbol{\beta}}{\text{argmin}} \quad \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2, \tag{19}$$

preferably, with additional correction $\hat{\boldsymbol{\beta}}_{en} := (1 + \lambda_2)\hat{\boldsymbol{\beta}}_{en}$. Instead of combining the lasso and ridge estimates, their penalties are combined in the elastic net estimator.

### 4.3 SCAD and MCP

The statistical properties of LASSO have been explored in depth. It has been found that $\ell_1$ regularization may not provide satisfactory variable selection especially when predictors are highly correlated. Moreover, stringent conditions are needed in order for LASSO to enjoy oracle properties.

From the perspective of sparse estimation, a desirable penalty function is expected to help achieve three key goals as spelled out in Fan and Li (2001): unbiasedness in estimating nonzero parameters, sparsity in terms of enforcing zero estimates, and continuity in terms of the model spectrum under consideration. These motivate their proposal of the SCAD (short for smoothly clipped absolute deviation) nonconvex penalty. Later on, Zhang (2010) proposed a similarly shaped penalty – MCP (short for minimax concave penalty). Both SCAD and MCP yield very similar empirical performance of sparse estimation, leading to substantial improvement over LASSO. Both being even functions, the SCAD penalty function is given by, for $\beta > 0$,

$$w_{a,b}(\beta) = \begin{cases} a\beta & \text{if } \beta \le a, \\ \dfrac{2ab\beta - a^2 - \beta^2}{2(b-1)} & \text{if } a < \beta \le ab, \\ a^2(b+1)/2 & \text{if } \beta > ab, \end{cases} \tag{20}$$

with first derivative $\dot{w}_{a,b}(\beta) = a\left\{ I(\beta \le a) + \dfrac{(ab-\beta)_+}{a(b-1)} I(\beta > a) \right\}$, for $a \ge 0$ and $b > 2$. It corresponds to a quadratic spline function with knots at $a$ and $ab$. For $\beta > 0$, the MCP penalty is given by

$$w_{a,b}(\beta) = \begin{cases} a\beta - \dfrac{\beta^2}{2b} & \text{if } \beta \le ab, \\ b\,a^2/2 & \text{if } \beta > ab, \end{cases} \tag{21}$$

with first derivative $\dot{w}_{a,b}(\beta) = (a - \beta/b)\, I\,(\beta \le a\, b)$ for $a \ge 0$ and $b > 1$. Both SCAD and MCP penalties are smooth in $\beta \ge 0$ with singularity at $\beta = 0$.

The SCAD or MCP estimator $\widetilde{\boldsymbol{\beta}}$ solves

$$\min_{\boldsymbol{\beta}} \quad \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} w_{a,b}(\beta_j).$$

The oracle properties of $\widetilde{\boldsymbol{\beta}}$, which imply both consistency in variable selection and efficiency in estimating the nonzero coefficients, have been established.

### 4.4 MIC

One latest development in sparse estimation is the Minimum approximated Information Criterion (MIC; Su, 2015). To motivate the method, recall that the best subset selection (BSS) with, e.g., BIC, which solves $\min_{\boldsymbol{\beta} \in \mathcal{B}} \; n \log \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2 + \ln(n) \cdot \| \boldsymbol{\beta} \|_0$ . Owing to the discrete and nonconvex nature of cardinality, BSS is known as an NP-hard. While $\ell_1$ regularization changes the discrete selection into a continuous process via convex relaxation, a loss of track of the complexity penalty $\ln(n)$ is incurred. Hence, it is replaced with a so-called tuning parameter $\lambda$, whose 'best' choice is yet to be determined. As a result, the common practice of regularization involves two steps: first compute the entire regularization path, i.e., solutions of $\boldsymbol{\beta}$ for every $\lambda \ge 0$; and then select the best $\lambda$ according to a model selection criterion, say, BIC. It is noteworthy that the regularization path $\{\widetilde{\boldsymbol{\beta}}(\lambda) : \lambda \ge 0\}$ is merely a one-dimensional curve in the $p$-dimensional parameter space $\mathbb{R}^p$. Thus these regularization methods essentially seek to optimize AIC or BIC over the much reduced search space.

While regularization is mainly motivated from optimization, MIC is more aligned with approximation. MIC bridges BBS and regularization by first approximating the information criterion (which involves approximating the indicator function $I\{\beta_j \neq 0\}$) in $\ell_0$ penalty with, e.g., a smooth hyperbolic tangent function $\tanh(a\beta_j^2)$) and then minimizing the approximated criterion directly. Besides, a reparameterization trick is then employed to enforce sparsity of parameter estimates while maintaining smoothness of the objective function.

In formulation, MIC solves

$$\min_{\boldsymbol{\beta}} \;\; n \, \log \left\{ \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} x_{ij} \gamma_j w(\gamma_j) \right)^2 \right\} \;\; + \;\; \lambda_0 \cdot \sum_{j=1}^{p} w(\gamma_j), \tag{22}$$

where

$$w(\gamma_j) = \tanh\left(a \cdot \gamma_j^2\right) \qquad \text{and} \qquad \beta_j = \gamma_j \cdot w(\gamma_j).$$

for $a > 0$ and $j = 1, \ldots, p$. In matrix form, (22) can be rewritten as

$$\min_{\boldsymbol{\gamma}} \;\; n \, \log \| \mathbf{y} - \mathbf{X}\mathbf{W}\boldsymbol{\gamma} \|^2 \;\; + \;\; \lambda_0 \cdot \mathrm{tr}(\mathbf{W}), \tag{23}$$

where $\mathbf{W} = \mathrm{diag}(w_j)$ with $w_j = w(\gamma_j)$ and the original parameter $\boldsymbol{\beta}$ has been reparameterized as $\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\gamma}$. The shape parameter $a > 0$ controls the sharpness of the approximation to the indicator function $I\{\beta_j \neq 0\}$). The performance of MIC is rather robust with respect to the choice of $a$. In practice, $a$ is recommended to be fixed as a constant in $[10, 50]$.

MIC offers several advantages. First, it is free of tuning parameters and hence is computationally more efficient. Secondly, it is aimed to minimize BIC, albeit approximated, without reducing the search space as in regularization. Hence, it often yields superior performance in terms of minimizing BIC. Essentially, MIC extends BSS to scenarios when $p$ is large. Thirdly, it facilitates a convenient way of circumvening post-selection inference since estimation and variable selection are completed simultaneously in one step. See Su et al. (2018) for details. On the other hand, MIC has a nonconvex formulation, same as SCAD and MCP.

## References


Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Procedings of Second International Symposium on Information Theory*, (eds B. N. Petrov and F. Csaki). Budapest: Akademiai Kiado, 267–281.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**: 125–127.

Blumensath, T. and Davies, M. E. (2009). Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, **27**: 265–274.

Casella, G. (1980). Minimax Ridge Regression Estimation. *Annals of Statistics*, **8**(5): 1036–1056.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression (with discussion). *Annals of Statistics*, **32**: 407–499.

Fan, J. and Li, R. (2001), Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**: 1348–1360.

Frank, I. E. and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**: 109–148.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *Elements of Statistical Learning*, 2nd Edition. Chapman and Hall.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased rstimation for nonorthogonal problems. *Technometrics*, **12**: 55–67.

Huang, J., Zhang, T., and Metaxas, D. (2011). Learning with structured Sparsity. *Journal of the Machine Learning Research*, **12**: 3371–3412.

Mallows, C. L. (1973). Some comments on $C_p$ *Technometrics*, **15**: 661–675.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, **92**: 179–191.

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

Su, X. G. (2015). Variable selection via subtle uprooting. *Journal of Computational and Graphical Statistics*, **24**: 1092-1113.

Su, X. G., Fan, J. J., Levine, R. A., Nunn, M. E., and Tsai, C.-L. (2018). Sparse estimation of generalized linear models (GLM) via approximated information criteria. To appear, *Statistica Sinica*.

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

Tibshirani, R., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, **67**: 91–108.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, **92**(4): 937–950.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**: 1418–1429.

Zou, H. and Hastie, T. (2003). Regression Shrinkage and Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, **67**: 301–320.

# APPENDIX

## A Quick Derivation of BIC

In the following short notes, I outlined a simplified version of Bhat and Kumar (2010). BIC intends to maximize the posterior probability of a model $M \in \mathcal{M}$ given the data $\mathbf{y} = \{y_i\}_{i=1}^n$. If all candidate models are equally likely to be selected, then

$$p(M|\mathbf{y}) = \frac{p(y|M)p(M)}{p(\mathbf{y})} \ \propto \ p(\mathbf{y}|M).$$

Maximizing $p(M|\mathbf{y})$ is equivalent to maximizing $p(\mathbf{y}|M)$. Assume that model $M$ is parameterized with $\boldsymbol{\theta}_M$ or, for simplicity, $\boldsymbol{\theta}$. Consider

$$\begin{aligned}
p(\mathbf{y}|M) &= \int_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}) g_M(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int \exp\{\log p(\mathbf{y}|\boldsymbol{\theta})\} d\boldsymbol{\theta}, \text{ assuming flat prior } g_M(\boldsymbol{\theta}) = 1 \\
&\approx \int \exp\left\{ L(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \hat{\mathbf{I}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right\} d\boldsymbol{\theta}, \text{ applying Taylor's expansion at MLE } \hat{\boldsymbol{\theta}}. \\
&\quad \text{where } \log p(\mathbf{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}) \text{ is the log-likelihood;} \\
&\quad\quad \hat{\mathbf{I}} = \hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}) = - \left. \frac{d^2 L}{d\boldsymbol{\theta} d\boldsymbol{\theta}^T} \right|_{\widehat{\boldsymbol{\theta}}} \text{ is observed Fisher information matrix;} \\
&\quad\quad \text{The gradient } dL/d\boldsymbol{\theta} \text{ is } \mathbf{0} \text{ at } \hat{\boldsymbol{\theta}} \text{ and hence the linear term has been dropped.} \\
&= \exp\{L(\hat{\boldsymbol{\theta}})\} \cdot \int \exp\left\{ \frac{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbf{I} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})}{2} \right\} d\boldsymbol{\theta} \\
&= \exp\{L(\hat{\boldsymbol{\theta}})\} \cdot \frac{(2\pi)^{p/2}}{|\mathbf{I}|^{1/2}} \quad \text{using multivariate normal density } \mathcal{N}(\widehat{\boldsymbol{\theta}}, \mathbf{I}^{-1}) \\
&= \exp\{L(\hat{\boldsymbol{\theta}})\} \cdot \frac{(2\pi)^{p/2}}{n^{p/2} |\mathbf{I}_0|^{1/2}},
\end{aligned}$$

where $p = |\boldsymbol{\theta}|$ denotes the cardinality of $\boldsymbol{\theta}$ and $\mathbf{I}_0$ denotes the Fisher information matrix based on one single $y_i$ observation so that $\mathbf{I} = n \cdot \mathbf{I}_0$. Note that $\det(n \cdot \mathbf{I}_0) = n^p \cdot \det(\mathbf{I}_0)$. We may use the expected Fisher information instead by invoking weak law of large numbers (WLLN).

Now taking logarithm and multiplying by $(-2)$ on both sides lead to, when $n \to \infty$,

$$-2 \log p(\mathbf{y}|M) \quad \propto \quad -2L(\widehat{\boldsymbol{\theta}}) + p \log(n) \quad \triangleq \quad BIC,$$

after removing all terms not involving $n$. The model with a smaller BIC is preferable. The Taylor expansion part within the integral is often termed Laplace approximation.