

Simulation Study Set-up from George and Foster (2000)

Draft: January 8, 2003

1 Simulation Studies

1.1 Methods of Selection

The goal of the simulation studies found in George and Foster (2000) is to investigate the performance of different methods of variable selection. Some of the methods considered were:

AIC (Akaike Information Criterion)

$$C_p$$

BIC (Bayesian Information Criterion)

RIC (Risk Inflation Criterion) Asymptotically minimizes the maximum predictive risk inflation due to selection when \mathbf{X} is orthogonal.

$$\text{MML } \text{MML} = SS_\gamma/\sigma^2 - F(\hat{c}, \hat{w})q_\gamma$$

$$\text{CML } \text{CML} = SS_\gamma/\sigma^2 - B(SS_\gamma/\sigma^2) - R(q_\gamma)$$

where $SS_\gamma = \hat{\beta}_\gamma$, $\hat{c}_\gamma = (SS_\gamma/\sigma^2 q_\gamma - 1)_+$, $\hat{w}_\gamma = q_\gamma/p$.

1.2 Basic Simulation Set-up

To study the these different methods of selection, data was simulated from the normal linear model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2\mathbf{I})$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$. Now suppose that some of the β_j 's have nonzero coefficients, with rest being zero. The goal was then to identify this unknown subset of non-zero coefficients through several methods of variable selection.

To simplify the simulation methods and reduce noise by independent estimation, σ^2 was assumed to be known and equal to one. To find the disparity between the selected model, $\hat{\gamma}$ and the correct model, γ underlying $\boldsymbol{\beta}$ the predictive loss should be calculated for each selection criterion at each iteration.

$$L(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}(\hat{\gamma})) = (\mathbf{X}\hat{\boldsymbol{\beta}}(\hat{\gamma}) - \mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\hat{\boldsymbol{\beta}}(\hat{\gamma}) - \mathbf{X}\boldsymbol{\beta}) \tag{2}$$

such that

$$\hat{\beta}_i(\hat{\gamma}) = \begin{cases} \beta_{i(OLS)} & \text{when } x_i \in \hat{\gamma} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

1.3 The Simple Orthogonal Case

Let $\mathbf{X} = \mathbf{I}$ so that the regression equation, (1), can be written as $\mathbf{Y} = \boldsymbol{\beta} + \epsilon$ with $p = n$. In this instance, the variable selection problem has been reduced to problem of locating or identifying the non-zero elements of a multivariate normal mean.

To simulate values of \mathbf{Y}

1. Fix the values of c and $q \leq p$.
2. Generate values for $\boldsymbol{\beta}$
 - Generate the first q components of $\boldsymbol{\beta}$ as $\beta_1, \dots, \beta_q \sim N(0, c)$ independently.
 - Set $\beta_{q+1}, \dots, \beta_p = 0$
 - Add $\epsilon \sim N_p(0, \mathbf{I})$ to each β .
3. Apply each selection criterion to \mathbf{Y} and evaluate the loss (2) using $\hat{\beta}_i = y_i \mathbf{I}(x_i \in \hat{\gamma})$, which is the least squares estimate of β_i under the selected model, $\hat{\gamma}$.
4. Evaluate the loss for the shrinkage estimators.

Begin simulation by setting $p = n = 1000$ and then repeat 2500 times for $c = 5, 25$ and $q = 0, 10, 25, 50, 100, 200, 300, 400, 500, 750, 1000$. For each pair of c and q , the loss is averaged over the 2500 repetitions.

1.4 Correlated Predictors and Fixed β values

When the variable selection cannot be simplified because the predictors are correlated and the values of $\boldsymbol{\beta}$ are fixed, both must first be simulated before evaluating (1) can even be considered. First, the n rows of \mathbf{X} need to be independently simulated from a $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where ij^{th} element of $\boldsymbol{\Sigma}$ is $\rho^{|i-j|}$. The simulation was conducted using $n = 200$, $p = 50$ and $\rho = -0.5, 0, 0.5$.

Next, to simulate values for $\boldsymbol{\beta}$ consider using only 11 choices for each β , such that each $\boldsymbol{\beta}$ consists of five replications of $(\beta_1, \dots, \beta_{10})'$ such that $\beta_i = \beta_{i-10}$ for $i = 11, \dots, 50$