Two-Step Bootstrap Model Selection

Guangbin Hong[1] and En Hua Hu[1]

[1] *University of Toronto*

Abstract

In Jun Shao (1995), it is shown that a consistent bootstrap model selection (BMS) procedure can be obtained by minimizing the bootstrap estimates of the prediction error. However, for consistency, one must bootstrap a sample of size m which is less than the original sample size of N. The method is consistent when $m \to \infty$ and $m/N \to 0$. As the optimal choice of m depends on the true parameters, it cannot be determined ex-ante, yet it greatly impacts the viability of the selection procedure. Here, we present a two-step bootstrap model selection (2SBMS) process which avoids the problem of having to select an optimal m.

1 Introduction

When the relationship between the dependent variable y and independent variables $\mathbf{x}$ is linear, there are many model/variable selection procedures. For instance, the Aike and Bayesian Information Criteria and their generalized counterparts(Aike 1970; Schwarz 1978; Rao and Wu 1989). Different Cross-Validation methods and Lasso techniques can also be considered(Allen 1974; Stone 1974; Tibshirani 1996). This paper studies Jun Shao (1995)'s bootstrap model selection (BMS) which has the problem of picking an optimal bootstrap sample size. We propose the two-step bootstrap model selection (2SBMS), show that it retains consistency and show by simulation that it can improve success rate significantly.

2 Bootstrap Model Selection

2.1 Linear Framework

Let y denote a vector of the variable of interest, and let $\mathbf{X}$ be a matrix of explanatory variables. Suppose $\mathbf{x}$ has p many variables, which is independent of sample size. Then we assume that $\mathbf{X} = (x_1, .., x_N)'$ is full rank and:

$$\mu_i = E(y_i|\mathbf{x_i}) = \mathbf{x_i}'\beta, \quad \text{var}(y_i|\mathbf{x_i}) = \sigma^2 \tag{1}$$

where β is a p vector of unknown parameters.

We denote α to be a subset of $\{1, .., p\}$ with size p_α. Thus α is a potential model. We say that a model α is *correct* if it contains all nonzero elements of β. We call the smallest correct model the *optimal* model, denoted by α_0. We let each model α be fitted by minimizing the least squared error, and denote the estimated model parameter by $\hat{\beta_\alpha}$.

The efficiency of a model can be measured by the average loss:

$$L_N(\alpha) = \frac{1}{N}\sum_{i=1}^{N}(\mu_i - \mathbf{x}'_{\mathbf{i}\alpha}\hat{\beta}_\alpha)^2 = \frac{1}{N}||\mu - \hat{\mu}_\alpha||^2 \tag{2}$$

Suppose $y = \mathbf{x}'\beta + \epsilon$, then an useful way to rewrite $L_N(\alpha)$ is:

$$L_N(\alpha) = \Delta_N(\alpha) + \frac{1}{N}||\mathbf{H}_\alpha\epsilon||^2 - \frac{2}{N}(\mu - \mathbf{H}_\alpha\mu)'\epsilon \tag{3}$$

Where $\Delta_N(\alpha) = \frac{1}{N}||\mu - \mathbf{H}_\alpha\mu||^2$ and $\mathbf{H}_\alpha = \mathbf{X}_\alpha(\mathbf{X}'_\alpha\mathbf{X}_\alpha)'\mathbf{X}'_\alpha$

For any incorrect model $\Delta_N(\alpha) > 0$ as $N \to \infty$. It is straightforward then to see that the optimal model minimizes $L_N(\alpha)$ for N sufficiently large. Thus the success of a model selection technique can be judged by success of selecting the optimal α_0.

Denote the model selected by a selection technique $\hat{\alpha}$, then the selection technique is called *consistent* if $\lim_{N\to\infty} P(\hat{\alpha} = \alpha_0) = 1$.

2.2 Jun Shao Bootstrap Model Selection

Jun Shao (1995) considers the prediction error of a model:

$$\Gamma_N(\alpha) = E\left[\frac{1}{N}\sum_{i=1}^{N}(z_i - \mathbf{x}'_{\mathbf{i}\alpha}\hat{\beta}_\alpha)^2\right] = \sigma^2 + L_N(\alpha) \tag{4}$$

Where $\hat{\beta}_\alpha$ is derived using a sample of size N and used to make a prediction on a new sample of z_is and $\mathbf{x_i}$s. Then minimizing the prediction error is the same as minimizing the average loss and the optimal model has the lowest prediction error.

It is shown in Bunke and Droge (1984) that an almost unbiased estimator for $\Gamma_N(\alpha)$ can be found. This can be done using Efron (1982,1983)'s expected excess error. However, Shao shows that minimizing this estimate does not lead to a consistent model selection technique. Rather, he proposes to minimize an estimate of $E[\Gamma_m(\alpha)]$, for some $m < N$. This leads to a consistent model selection technique as $m \to \infty$ and $\frac{m}{N} \to 0$.

Let $y^*, X^*_{\alpha,m}$ be a bootstrap draw by pairs of size m for model α. Then Jun Shao proposes to estimate $E[\Gamma_m(\alpha)]$ by:

$$\hat{\Gamma}_{N,m}(\alpha) = E_*\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_{i\alpha}\tilde{\beta}^*_{\alpha,m})^2\right] \tag{5}$$

where $\tilde{\beta}^*_{\alpha,m} = (X^*_{\alpha,m}{}'X^*_{\alpha,m})^{-1}X^*_{\alpha,m}{}'y^*$, the estimated β_α from a bootstrap draw of size m and E_* is the expectation with respect to bootstrap sampling.

This procedure is highly accurate, even in small samples, for finding the true model *if* one can pick the optimal m. However, as Shao and our simulation shows, large discrepancies occur for different choices of m. In general, a large m is desirable for ruling out models which underfit while a small m rules out better models which overfit. Therefore, without knowing how the optimal model is positioned among the set of possible models, there is no a priori way of choosing m optimally.

Lastly, Jun Shao considers both bootstrapping pairs and residuals which lead to very similar results. In the following, we consider only bootstrapping pairs because, as Shao points out, unless there is a special structure in the x_i (e.g. Hall 1990 when $x_i = \frac{i}{N}$), it is not clear how to bootstrap residual with bootstrap sample $m < N$.

3 Two Step Bootstrap Model Selection

We propose a two step method to overcome Shao's difficulty. The difficulty was mainly that an optimal m could not be picked ex ante without knowledge of distribution of models. If there were mostly models which overfit, we would like a small m and vice versa. Our proposed technique is a two step method. In this first step, we can eliminate incorrect models which underfit. In the second step, we can use Shao's technique with small m to rule out models which overfit, or in the linear case, simply pick the simplest model.

In the following, we present the technique. Let us define the bootstrap sample mean squared error, $K_m(\alpha)$, as:

$$K_m(\alpha) = E_*\big[\frac{1}{m}\sum_{i=1}^{m}(y_i^* - x^*_{i\alpha}\tilde{\beta}^*_{\alpha,m})^2\big] \tag{6}$$

Let α_p be the most overfit model, with subset of size p, we consider the following ratio:

$$R_\alpha(m) = \frac{K_N(\alpha) - K_N(\alpha_p)}{K_m(\alpha) - K_m(\alpha_p)} \tag{7}$$

One can show that asymptotically, if α is an incorrect model then $R_\alpha(m)$ converges to 1 as m and N grows large. However, if α is a correct model, $R_\alpha(m)$ converges to $\frac{2m}{N+m}$. Proofs in appendix.

We could compute $R_\alpha(m)$ and reject models for which $\frac{2m}{N+m}$ is much closer to 1 than $R_\alpha(m)$. However, for $\frac{2m}{N+m}$ to be sufficiently different from 1 we would need to pick a small m. Recall that in Shao's original method, a smaller m does not allow one to rule out efficiently models which underfit. Similarly, using $R_\alpha(m)$ with small m will be inefficient for rejecting models which underfit. This can be intuitively seen in Figure

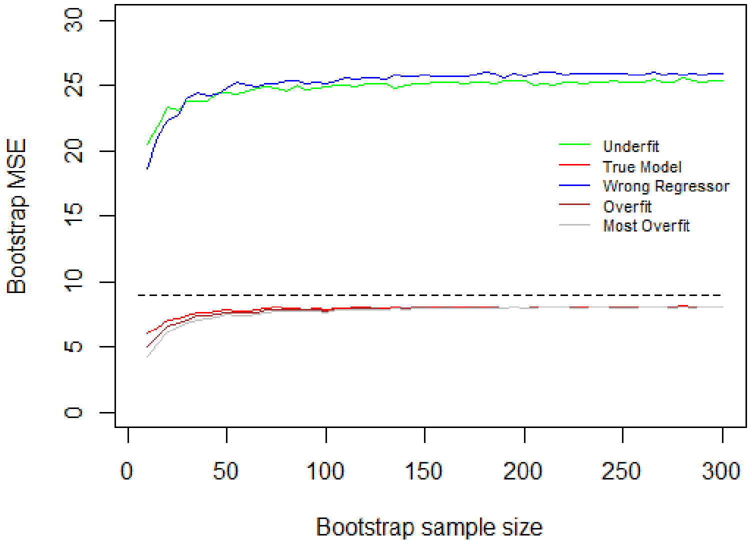

Figure 1: Convergence of Bootstrap MSE

1. When m is small, even underfitting models will have rates of change proportional to α_p's rate.

Thus we wish to pick a large m to maximize chances of ruling out underfitting models. To do this, we note that $R_\alpha(m)$ can be estimated to be:

$$R_\alpha(m) = \frac{\Delta_N(\alpha) - \epsilon'(H_\alpha - H_{\alpha_p})\epsilon/N - (p_\alpha - p)\sigma^2/N + o_p(1/N)}{\Delta_N(\alpha) - \epsilon'(H_\alpha - H_{\alpha_p})\epsilon/N - (p_\alpha - p)\sigma^2/m + o_p(1/m)} \tag{8}$$

$\Delta_N(\alpha)$ is 0 for a correct model, and converges to a constant as $N \to \infty$ for an incorrect model. This, coupled with the observation that the asymptotic expectation of the second term is $\sigma^2(p_\alpha - p)$ leads to our claims earlier.

However, because the second term can be highly volatile and $\frac{2m}{N+m}$ is too close to one for larger m, we consider instead $R^*_\alpha(m)$:

$$R^*_\alpha(m) = \frac{K_N(\alpha) - K_N(\alpha_p) - (p_\alpha - p)\hat{\sigma}^2_\alpha/N}{K_m(\alpha) - K_m(\alpha_p) - (p_\alpha - p)\hat{\sigma}^2_\alpha/m} \tag{9}$$

Where $\hat{\sigma}^2_\alpha$ is the estimated variance of error from the model α using the full original sample by running an OLS regression and computing the residuals.

Asymptotically, $R^*_m(\alpha)$ still converges to 1 for incorrect models while it will converge to $\frac{3m}{m+2N}$ for correct models. We propose to pick m large, such as $0.7N$ and reject models for which $R^*_\alpha(m) > \frac{1}{2}(\frac{3m}{m+2N} + 1)$. Alternatively, one could use a double bootstrap method to generate a confidence interval for $R^*_\alpha(m)$ and reject models whose interval lies above $\frac{3m}{m+2N}$. However it becomes very computationally costly and from our simulations, there is not much accuracy to be gained.

Finally, given the convergence results, the following holds:

Theorem: If $\liminf_{N\to\infty} \Delta_N(\alpha) > 0$ then the two step selection technique is consistent.

Proof in Appendix.

4 Simulation Results

In the following we present some simulation results. We simulate 4 regressors x_1, x_2, x_3, x_4 which are drawn from $\mathcal{N}(1,4)$. We then generate all possible linear combinations of the 4 regressors. Then y is generated by the true model as specified with $\epsilon \sim \mathcal{N}(0,2)$, $y = x'\beta + \epsilon$. Finally, we consider also a probit model where $y = \nVdash\{x'\beta + \epsilon - 2.5 > 0\}$.

We simulate the BMS using 100 bootstrap draws. For the 2SBMS we draw 500 samples of size $m = 0.7N$ and compute their $R^*_\alpha(m)$ in the first step. We keep all those for which $R^*_\alpha(m) < \frac{1}{2}(\frac{3m}{m+2N} + 1) \approx 0.889$. Then we select the model which minimizes the bootstrap prediction error for 100 draws of size $m = \frac{1}{5}$ for the second step. Thus our first step uses $R^*_\alpha(m)$ to rule out some models, and then we use Jun Shao's BMS method in a second step for $m = \frac{1}{5}$ to select among the remaining one.

We run three specifications. The three specification represent three scenarios. The true model can be small, medium, or large compared to other considered models. When the true model is small, as per specification 1, using a small m yields the best result. However, when the true model is large, we see that using a small m is no longer beneficial. Although picking an optimal m is non-trivial, Jun Shao's method does perform comparably to BIC, a common choice among researchers.

Table 1: Simulation Results, Linear Model, 1000 runs

True Model $(\beta_1, \beta_2, \beta_3, \beta_4)$	**BIC**	**BMS** (m=$\frac{1}{2}N$)	**BMS** (m=$\frac{1}{5}$N)	**2SBMS**
$(2,0,0,0), N = 50$	0.848	0.628	0.973	0.966
$(2,2,0,0), N = 100$	0.921	0.699	0.969	0.956
$(2,\frac{2}{5},\frac{2}{5},\frac{1}{3}), N = 200$	0.971	0.834	0.684	0.962

As we have argued, the 2SBMS has the advantage of not having to worry about an optimal m. Given that the second step of the technique uses Jun Shao's BMS for $m = \frac{1}{5}$ it is natural to compare the results with those of this BMS. We see that in

Table 2: Simulation Results, Probit Model, 1000 runs

True Model $(\beta_1, \beta_2, \beta_3, \beta_4)$	**BIC**	**BMS** (m=$\frac{1}{2}N$)	**BMS** (m=$\frac{1}{5}$N)	**2SBMS**
$(1, 0, 0, 0), N = 50$	0.833	0.356	0.856	0.895
$(\frac{1}{2}, \frac{1}{2}, 0, 0), N = 100$	0.937	0.091	0.707	0.912
$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}), N = 400$	0.694	0.989	0.765	0.952

specification 1 and 2, where $m = \frac{1}{5}$ performs very well, the 2SBMS performs slightly worse but remains very comparable and above the BIC. However, in specification 3, the 2SBMS outperforms the BMS for both small and large m. The explanation for the two observation lies within the first step of the 2SBMS. First, there is a small chance that the true model is rejected in the first step, this explains why the 2SBMS performs just slightly worse in specification 1 and 2 compared to the BMS with an optimal m. Second, the first step of the 2SBMS can eliminate underfitting models who have a lower bootstrap prediction error. Thus preventing the BMS from making a mistake in the second step. This explains why it can perform significantly better in the third specification.

To give the reader an idea of the gains and losses of this first step process, we recorded the number of times the true model was rejected in the first step for the first and second specification. In the first specification 30 times out of 1000, the true model was rejected in the first step, however, the 2SBMS only has 7 less correct guesses than the BMS using the same $m = \frac{1}{5}N$. In the second specification, 15 times out 1000 the true model was rejected, but again the 2SBMS only has 7 less correct guesses. Thus we see that the loss is relatively small, but as seen in the third specification, the gains can be significant.

Lastly, we comment on the probit simulations. While we currently do not have consistency results for 2SBMS under probit, we see that it performs well in simulation. Furthermore, BMS's problem with picking an optimal m is further aggravated in the probit case, therefore making 2SBMS's results even more comforting.

5 Conclusion

To summarize, we exploit the rate of change of the bootstrap MSE to improve upon Shao's bootstrap selection technique. Our first uses $R^*_\alpha(m)$ to eliminate underfitting models, and then runs Shao's selection method using a small m to efficiently rule out overfitting models.

The main observation we make is that the difference between the bootstrap MSE of a correct model and the most overfitting model decreases at a rate that is close to m/N. This is because both converge to σ^2 at similar rates, thus the difference converges to 0 at a stable rate. However, an incorrect model will have a term $\Delta_N(\alpha)$ which is remains constant, thus the differences converges to the same constant and $R^*_\alpha(m)$ converges to 1.

This observation allows us to construct the first step of the 2SBMS. We believe that we have not fully exploited this observation as the selection method is still quite primitive during the first stage. However, we believe that this method is very promising as the early simulation results show. We also believe it to be generalizable to non-linear, generalized linear and autoregression models in the same way as Jun Shao's original technique was.

Appendix

All results of the paper can be derived once the asymptotic convergence $K_m(\alpha)$ is known. Let E_* and var_* be the asymptotic expectation and variance.

First, we cite some results from Shao (1995):

1) $\mathrm{var}_*\tilde{\beta}^*_{\alpha,N} = (\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}\sum_{i=1}^N \mathbf{x}_{\mathbf{i}\alpha}\mathbf{x}_{\mathbf{i}\alpha}(y_i - \mathbf{x}'_{\mathbf{i}\alpha}\hat{\beta}_\alpha)^2(\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}[1+o_p(1)]$
$\quad = \sigma^2(\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}[1+o_p(1)]$

2) $\mathrm{var}_*\tilde{\beta}^*_{\alpha,m} \approx \frac{N}{m}\mathrm{var}_*\tilde{\beta}^*_{\alpha,N}$

3) $\tilde{\beta}^*_{\alpha,m} - \hat{\beta}_\alpha = (\mathbf{X}^*_{\alpha,\mathbf{m}}{}'\mathbf{X}^*_{\alpha,\mathbf{m}})^{-1}\sum_{i=1}^m \mathbf{x}^*_{\mathbf{i}\alpha}(y^*_i - \mathbf{x}^*_{\mathbf{i}\alpha}{}'\hat{\beta}_\alpha)$

Then we derive the asymptotic convergence of $K_m(\alpha)$ as:

$$\begin{aligned} K_m(\alpha) &= E_*\big[\frac{1}{m}\sum_{i=1}^m(y^*_i - \mathbf{x}^*_{\mathbf{i}\alpha}\tilde{\beta}^*_{\alpha,m})^2\big] \\ &= E_*\big[\frac{1}{m}\sum_{i=1}^m(y^*_i - x^*_{i\alpha}\hat{\beta}_\alpha)^2\big] + E_*\big[\frac{1}{m}\sum_{i=1}^m(\mathbf{x}^*_{\mathbf{i}\alpha}{}'(\tilde{\beta}^*_{\alpha,m} - \hat{\beta}_\alpha)^2)\big] - \frac{2}{m}E_*\big[\sum_{i=1}^m(y^*_i - \mathbf{x}^*_{\mathbf{i}\alpha}\hat{\beta}_\alpha)\mathbf{x}^*_{\mathbf{i}\alpha}{}'(\tilde{\beta}^*_{\alpha,m} - \hat{\beta}_\alpha)\big] \end{aligned}$$

We proceed by solving for each of the three terms:

$$\begin{aligned} 1)E_*\big[\frac{1}{m}\sum_{i=1}^m(y^*_i - x^*_{i\alpha}\hat{\beta}_\alpha)^2\big] &= \frac{1}{N}\sum_{i=1}^N(y_i - \mathbf{x}'_{\mathbf{i}}\beta)^2 \\ &= \frac{1}{N}[||\epsilon||^2 - \epsilon'\mathbf{H}_\alpha\epsilon] + \Delta_N(\alpha) \end{aligned}$$

$$
\begin{aligned}
2)E_*\big[\frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}^*_{\mathbf{i}\alpha}{}'(\tilde{\beta}^*_{\alpha,m}-\hat{\beta}_\alpha)^2)\big] &= \frac{1}{m}E_*[(\tilde{\beta}^*_{\alpha,m}-\hat{\beta}_\alpha)'\sum_{i=1}^{m}\mathbf{x}^*_{\mathbf{i}\alpha}\mathbf{x}^*_{\mathbf{i}\alpha}{}'(\tilde{\beta}^*_{\alpha,m}-\hat{\beta}_\alpha)] \\
&= \frac{1}{m}E_*[(\tilde{\beta}^*_{\alpha,m}-\hat{\beta}_\alpha)'\mathbf{X}^*_{\alpha,\mathbf{m}}{}'\mathbf{X}^*_{\alpha,\mathbf{m}}(\tilde{\beta}^*_{\alpha,m}-\hat{\beta}_\alpha)] \\
&= \frac{1}{m}E_*[(\tilde{\beta}^*_{\alpha,m}-\hat{\beta}_\alpha)'\frac{m}{N}\mathbf{X}'_\alpha\mathbf{X}_\alpha(\tilde{\beta}^*_{\alpha,m}-\hat{\beta}_\alpha)](1+o_p(m)) \\
&= \frac{1}{N}\mathrm{Tr}((\mathbf{X}'_\alpha\mathbf{X}_\alpha)\mathrm{var}_*\tilde{\beta}^*_{\alpha,m}) \\
&= \frac{p_\alpha\sigma^2}{m}+o_p(m)
\end{aligned}
$$

$$
\begin{aligned}
3)\frac{2}{m}E_*\big[\sum_{i=1}^{m}(y^*_i-\mathbf{x}^*_{\mathbf{i}\alpha}\hat{\beta}_\alpha)\mathbf{x}^*_{\mathbf{i}\alpha}{}'(\tilde{\beta}^*_{\alpha,m}-\hat{\beta}_\alpha)] &= \frac{2}{m}E_*[(\tilde{\beta}^*_{\alpha,m}-\hat{\beta}_\alpha)'\mathbf{X}^*_{\alpha,\mathbf{m}}{}'\mathbf{X}^*_{\alpha,\mathbf{m}}(\tilde{\beta}^*_{\alpha,m}-\hat{\beta}_\alpha)] \\
&= \frac{2p_\alpha\sigma^2}{m}+o_p(m) \quad \text{(from same procedure as above)}
\end{aligned}
$$

Putting those three together gives us $R_\alpha(m)$'s asymptotic behavior, our results follow from it along with the fact that $\Delta_N(\alpha)=0$ for a correct model as $N\to\infty$.

The asymptotic behavior of $R_\alpha(m)$ also gives us the theorem. We know already that BMS is consistent, therefore, suffice to notice that as $N\to\infty$, the probability of rejecting the optimal model in the first step goes to 0.

References

[1] Akaike, H., 1974. A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), pp.716-723.

[2] Allen, D.M., 1974. The relationship between variable selection and data augumentation and a method for prediction. Technometrics, 16(1), pp.125-127.

[3] Bunke, O. and Droge, B., 1984. Bootstrap and cross-validation estimates of the prediction error for linear regression models. The Annals of Statistics, 12(4), pp.1400-1424.

[4] Efron, B., 1982. The jackknife, the bootstrap, and other resampling plans (Vol. 38). Siam.

[5] Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American statistical association, 78(382), pp.316-331.

[6] Hall, P. and Pittelkow, Y.E., 1990. Simultaneous bootstrap confidence bands in regression. Journal of Statistical Computation and Simulation, 37(1-2), pp.99-113.

[7] Schwarz, G., 1978. Estimating the dimension of a model. The annals of statistics, 6(2), pp.461-464.

[8] Shao, J., 1996. Bootstrap model selection. Journal of the American Statistical Association, 91(434), pp.655-665.

[9] Stone, M., 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), pp.44-47.

[10] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp.267-288.

[11] Rao, R. and Wu, Y., 1989. A strongly consistent procedure for model selection in a regression problem. Biometrika, 76(2), pp.369-374.