ABSTRACT

DUAN, JIANGTAO. Bootstrap-Based Variance Estimators for a Bagging Predictor. (Under the direction of Dennis Boos and Leonard Stefanski.)

Bagging (**b**ootstrap **agg**regat**ing**) was introduced by Leo Breiman as a method to improve predictions based on regression variable selection procedures that are known to be sensitive to small changes in the observed data. A bagging estimator is an average of estimators computed from bootstrap samples. For variable selection in linear models, a bagging prediction is simply an average of predictions over bootstrap samples. The jackknife-after-bootstrap and the bootstrap-after-bootstrap appear to be the only available methods for estimating the variance of a bootstrap average. However, their performance in the context of regression variable selection is unknown. The parallel bootstrap, introduced in this thesis, is a new method for estimating the variance of a bagging estimator. It is based on the usual iterated variance decomposition and takes advantage of the simple result that the variance of a sample average of independently and identically distributed random variables is the variance of a single random variable divided by n. Two versions of the parallel bootstrap are derived: the "difference-of-variances" (DoV) version comes directly from this variance decomposition; the "variance-of-difference" (VoD) version is a modification that has the advantage of always being nonnegative. Simulations in the linear regression context are carried out for two variable selection procedures: forward selection with model size chosen by the fast False Selection Rate (FSR) method, and the adaptive LASSO. These simulations demonstrate that the VoD version of the parallel bootstrap performs very well and is usually superior to the the bootstrap-after-bootstrap method. The jackknife-after-bootstrap is biased too high and is not competitive with the VoD version and the bootstrap-after-bootstrap. The DoV version of the parallel bootstrap is biased too low and is also not competitive.

Bootstrap-Based Variance Estimators for a Bagging Predictor

by
Jiangtao Duan

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2011

APPROVED BY:

Hao (Helen) Zhang	Yichao Wu
Dennis Boos Co-Chair of Advisory Committee	Leonard Stefanski Co-Chair of Advisory Committee

DEDICATION

To my parents.

BIOGRAPHY

Jiangtao Duan was born in a tranquil village (Tongshu) of Hubei province, China. He loved the fresh air, simple personal relationship and hard-working people in his hometown. He obtained a bachelor degree and a master degree in Applied Mathematics. He moved to North Carolina State University in 2006 to pursue a Ph.D. degree in Statistics.

ACKNOWLEDGEMENTS

I am very grateful to my advisors, Dr. Dennis D. Boos and Dr. Leonard A. Stefanski. They have provided me with valuable ideas, insightful comments and continuous support. They are always available when I need their help. This dissertation could not be possible without their guidance. I am privileged to have this opportunity to work with them and learn from them. They have taught me a lot and their impact on me will last far beyond this dissertation.

I also would like to thank Dr. Helen H. Zhang and Dr. Yichao Wu for being on my thesis committee and providing useful suggestions to improve the quality of this work. I also want to thank Dr. Lingjuan W. Li for graciously agreeing to be the graduate representative. My thanks to Dr. Bibhuti Bhattacharyya for his excellent job of teaching me Probability Theory and his recommendation of Dr. Boos and Dr. Stefanski as my future advisors. I also appreciate the Department of Statistics for creating a very supportive environment for our graduate students.

Special thanks to my wife, Jing, for her patience and continuous support. She always stands by me during my ups and downs. She does a wonderful job taking care of my family. I also want to thank my two wonderful little ones, Warren and William, for the joy and happiness they always bring to me.

I must thank my mom and my dad for their love and encouragement. They always have confidence in me, respect my decisions, and have done everything they could to help me pursue my dreams. Without them I could not have been what I am today. I also want to thank my sisters for taking care of my parents while I study abroad. Their support carried me through hard times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Overview

In this thesis we introduce a new method for estimating the variance of a bagging estimator defined by a bootstrap step,

$$\widehat{\gamma}_{BAG} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}_b, \tag{1.1}$$

where $\widehat{\gamma}_b$ is the estimator $\widehat{\gamma}$ computed from the bth bootstrap sample, and B is the total number of bootstrap replications. In the motivating application, $\widehat{\gamma}$ is a regression predictor $\widehat{\beta}_0 + \boldsymbol{x}^T\widehat{\boldsymbol{\beta}}$, where $\widehat{\beta}_0$ is an estimated intercept and $\widehat{\boldsymbol{\beta}}$ is a vector of estimated linear regression coefficients from a model selection procedure like stepwise regression or the adaptive least absolute shrinkage and selection operator (ALASSO), and $\boldsymbol{x}$ is a particular vector of predictor variables. However, $\widehat{\gamma}$ could be any type of estimator where $\widehat{\gamma}_{BAG}$ is of interest. In order to make statistical inference for $\widehat{\gamma}_{BAG}$, it is necessary to estimate the variance of $\widehat{\gamma}_{BAG}$, $\mathrm{Var}(\widehat{\gamma}_{BAG})$.

Currently, the only methods available to estimate $\mathrm{Var}(\widehat{\gamma}_{BAG})$ are the jackknife-after-bootstrap and the bootstrap-after-bootstrap. However, the bootstrap-after-bootstrap requires a lot of computation, and the jackknife-after-bootstrap has not been studied very

much. We introduce two new bootstrap-based methods, the parallel bootstrap "variance-of-difference" (VoD) and the parallel bootstrap "difference-of-variances" (DoV) variance estimator. In addition we modify the bootstrap-after-bootstrap to use the residual-based bootstrap in the outer loop and use the random-pair bootstrap in the inner loop. Simulations show that the performance of the parallel bootstrap VoD and the modified bootstrap-after-bootstrap method are both approximately unbiased but that the VoD has lower variance for the same computational cost. The jackknife-after-bootstrap is not approximately unbiased and can not be recommended.

The bagging estimator in (1.1) was proposed by Breiman (1996b) as a method to improve the stability of predictions based on variable selection procedures like Forward Addition or Backward Elimination. Breiman (1996a) pointed out that standard variable selection procedures are unstable in the sense that a small change in the data set can lead to different models and predictions. The bagging estimator in (1.1), however, is able to stabilize the prediction and reduce model error.

Now we briefly introduce this new parallel bootstrap variance estimator in the context of linear regression. We assume a linear regression model with k_T covariates $X_1, \cdots, X_{k_T}$ available,

$$\boldsymbol{Y} = \beta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ is an $n \times 1$ vector of response values, β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{k_T})^T$ is the regression coefficient vector, $\boldsymbol{X}$ is an $n \times k_T$ matrix with the ith row $\boldsymbol{x}_i = (x_{i,1}, \cdots, x_{i,k_T})$, and $\boldsymbol{\epsilon}$ is an $n \times 1$ error vector with each element independently drawn from a distribution with mean 0 and variance σ^2. The covariate X_j is informative if $\beta_j \neq 0$ and is uninformative if $\beta_j = 0$. The bagging estimator with predictor vector $\boldsymbol{x}$ is

$$\widehat{\gamma}_{BAG} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\gamma}_b = \frac{1}{B}\left\{ \sum_{b=1}^{B} \left(\widehat{\beta}_{0,b} + \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}_b \right) \right\},$$

where $\widehat{\beta}_{0,b}$ is the estimated intercept and $\widehat{\boldsymbol{\beta}}_b$ is the estimated regression coefficient vector from a linear model chosen by variable selection procedures with the bth bootstrap sample.

The parallel bootstrap variance estimator works as follows:

1. Draw a level-1 bootstrap sample $\mathcal{L}^{(b)} = (\boldsymbol{Y}^{(b)}, \boldsymbol{X}^{(b)})$ from the original data set $\mathcal{L} = (\boldsymbol{Y}, \boldsymbol{X})$ using independent random-pair bootstrap method and compute the basic statistic $\widehat{\gamma}_b$.

2. Draw a level-2 bootstrap sample $\mathcal{L}^{(b,1)} = (\boldsymbol{Y}^{(b,1)}, \boldsymbol{X}^{(b,1)})$ from each level-1 bootstrap sample $\mathcal{L}^{(b)}$ and compute the basic statistic $\widehat{\gamma}_{b,1}$. In our work the level-2 sample is drawn by the residual-based bootstrap (more on this in Section 1.2.1).

3. Repeat the first two steps B times to obtain $\widehat{\gamma}_1, \cdots, \widehat{\gamma}_B$, and $\widehat{\gamma}_{1,1}, \cdots, \widehat{\gamma}_{B,1}$.

4. Compute the parallel bootstrap VoD and the parallel bootstrap DoV variance estimator as

$$\widehat{\mathrm{Var}}_{VoD}(\widehat{\gamma}_{BAG}) = \frac{1}{B-1}\sum_{b=1}^{B}\{(\widehat{\gamma}_{b,1} - \widehat{\gamma}_b) - (\overline{\widehat{\gamma}}_{\cdot,1} - \overline{\widehat{\gamma}}_{\cdot})\}^2, \tag{1.2}$$

$$\widehat{\mathrm{Var}}_{DoV}(\widehat{\gamma}_{BAG}) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_{b,1} - \overline{\widehat{\gamma}}_{\cdot,1})^2 - \left(1 - \frac{1}{B}\right)\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_b - \overline{\widehat{\gamma}}_{\cdot})^2, \tag{1.3}$$

where $\overline{\widehat{\gamma}}_{\cdot} = B^{-1}\sum \widehat{\gamma}_b$ and $\overline{\widehat{\gamma}}_{\cdot,1} = B^{-1}\sum \widehat{\gamma}_{b,1}$.

As explained in Chapter 3, the idea behind (1.2) and (1.3) is the usual iterated variance calculation

$$\begin{aligned}
\mathrm{Var}(\widehat{\gamma}_{BAG}) &= \mathrm{Var}\left\{\mathrm{E}(\widehat{\gamma}_{BAG} \mid \mathcal{L})\right\} + \mathrm{E}\left\{\mathrm{Var}(\widehat{\gamma}_{BAG} \mid \mathcal{L})\right\} \\
&= \mathrm{Var}\left\{\mathrm{E}(\widehat{\gamma}_1 \mid \mathcal{L})\right\} + \mathrm{E}\left\{\frac{1}{B}\mathrm{Var}(\widehat{\gamma}_1 \mid \mathcal{L})\right\} \\
&= \mathrm{Var}(\widehat{\gamma}_\infty) + \frac{1}{B}\mathrm{E}\left\{\mathrm{Var}(\widehat{\gamma}_1 \mid \mathcal{L})\right\},
\end{aligned}$$

where $\widehat{\gamma}_\infty = \lim_{B\to\infty}\widehat{\gamma}_{BAG}$.

This thesis is organized as follows. In the remainder of Chapter 1, we review the random-pair bootstrap and the residual-based bootstrap methods and the bagging technique. We also review some typical variable selection procedures, such as the fast false selection rate method (FFSR) and the adaptive least absolute shrinkage and selection operator method (ALASSO). In Chapter 2, we describe how the jackknife-after-bootstrap and the bootstrap-after-bootstrap can be used to estimate the variance of bagging estimators. We also gain insight into the effect of bootstrap sample sizes on the bootstrap-after-bootstrap variance estimator for bagging predictors. In Chapter 3, we discuss the parallel bootstrap variance estimators in detail. We also show some properties of our methods in two special situations. In Chapter 4, we compare the parallel bootstrap variance estimators with the bootstrap-after-bootstrap method by Monte Carlo simulation in the linear regression model. Chapter 5 includes simulations to demonstrate the performance of the jackknife-after-bootstrap and some additional simulation results. In Chapter 6 we illustrate the application of the proposed parallel bootstrap and the modified bootstrap-after-bootstrap with data from a study of college graduation rate.

1.2 Bootstrap Methods and Bagging Technique

1.2.1 The Bootstrap

Efron (1979) introduced the bootstrap to estimate the sampling distribution of statistics. One primary use of the bootstrap is to assess the accuracy of estimators. The computing-based version of the bootstrap generates many, say B, bootstrap samples from an estimate of the population or distribution from which the data were obtained. For each bootstrap sample, the same estimator is calculated. The empirical distribution of these B estimates approximates the true sampling distribution of the estimator. The standard bootstrap variance estimate of the estimator is just the sample variance of these B estimates.

When the data are i.i.d. vectors $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, the nonparametric bootstrap is equiv-

alent to drawing individual members of the bootstrap sample independently and with replacement from the set of values $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$. When a parametric assumption is used to estimate the underlying distribution, then the parametric bootstrap draws samples independently from that estimated distribution.

In the linear regression context, there are several ways to generate bootstrap samples. For regression, the data consist of n random pairs $(Y_1, \boldsymbol{x}_1), \ldots, (Y_n, \boldsymbol{x}_n)$. The nonparametric bootstrap that draws samples with replacement from the data is called the random-pair bootstrap or paired bootstrap (p. 291 of Shao and Tu, 1995).

The residual-based bootstrap generates a bootstrap sample as follows:

1. Assume the regression model is $\mathrm{E}(Y|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$.

2. Use the original data to fit the regression model; estimate the regression coefficient vector $\widehat{\boldsymbol{\beta}}$; calculate the fitted values $\widehat{Y}_i = \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}$ and the residuals $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i$, $i = 1, \cdots, n$.

3. Randomly draw with replacement from the n residuals to obtain $\widehat{\epsilon}_i^*, i = 1, \cdots, n$.

4. Form a bootstrap sample $(Y_1^*, \boldsymbol{x}_1), \cdots, (Y_n^*, \boldsymbol{x}_n)$, where $Y_i^* = \widehat{Y}_i + \widehat{\epsilon}_i^*$.

If an intercept is not included in the X matrix, then the residuals in Step 3. need to be replaced by centered residuals $\widehat{e}_i - \overline{\widehat{e}}$. It is also recommended, that the residuals or centered residuals be scaled by dividing by $\sqrt{1 - p/n}$ (p. 290 of Shao and Tu, 1995).

Note that the residual-based bootstrap could also be called model-based because it depends on the model assumption of $\mathrm{E}(Y|\boldsymbol{x})$. Nevertheless, it is still nonparametric in the sense that no parametric assumptions about the errors ϵ are made.

1.2.2 Motivation for Using Residual-based Bootstrap with Variable Selection Procedures

As we will see later, the proposed parallel bootstrap VoD variance estimator uses the random-pair bootstrap and residual-based bootstrap. It uses the random-pair bootstrap to draw level-1 samples and the residual-based bootstrap to draw level-2 samples. In this subsection we investigate a special situation to gain insight into why we use residual-based bootstrap to estimate variances for estimators with variable selection procedures.

The standard random-pair bootstrap will typically not be consistent for estimating the distribution of estimators based on variable selection procedures. To see this, we consider a truncated estimator similar to the Hodges super-efficient estimator at a single point. Leeb and Potscher (2005) and others have used these type of examples as well.

Let $Y_1, \cdots, Y_n$ be an i.i.d. sample from a normal distribution with mean μ and known variance 1. Consider the estimator

$$\widehat{\mu} = \overline{Y} I(|\sqrt{n}\,\overline{Y}| > 1.96),$$

which is the sample mean $\overline{Y}$ when $|\overline{Y}|$ is suitably large and 0 otherwise. Thus, $\sqrt{n}\widehat{\mu}$ has the same distribution as

$$(\sqrt{n}\mu + Z_1) I(|\sqrt{n}\mu + Z_1| > 1.96),$$

where Z_1 is a standard normal random variable. When $\mu = 0$, this latter expression reduces to

$$Z_1 I(|Z_1| > 1.96). \tag{1.4}$$

The random-pair (nonparametric) bootstrap randomly selects each of $Y_1^*, \cdots, Y_n^*$ independently and with replacement from the sample $\{Y_1, \cdots, Y_n\}$. Conditional on $\boldsymbol{Y} =$

$(Y_1, \cdots, Y_n)^T$, the bootstrap quantity $\sqrt{n}\widehat{\mu}^*$ approximately has the distribution

$$(\sqrt{n}\,\overline{Y} + s_n Z_2) I(|\sqrt{n}\,\overline{Y} + s_n Z_2| > 1.96), \tag{1.5}$$

by the central limit theorem and the fact that $\mathrm{E}(\overline{Y}^*|\boldsymbol{Y}) = \overline{Y}$ and $\mathrm{Var}(\overline{Y}^*|\boldsymbol{Y}) = s_n^2$, where $s_n^2 = n^{-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$ and Z_2 is a standard normal random variable independent of Z_1.

When $\mu = 0$, the distribution of (1.4) is not well approximated by the conditional distributional distribution of (1.5), because $\sqrt{n}\,\overline{Y}$ need not be near 0. In large samples, (1.5) approximately becomes

$$(Z_1 + Z_2) I(|Z_1 + Z_2| > 1.96) = \sqrt{2} Z_3 I(|Z_3| > \frac{1.96}{\sqrt{2}}) \approx \sqrt{2} Z_3 I(|Z_3| > 1.39), \tag{1.6}$$

where Z_1 is independent of Z_2, and Z_3 is a standard normal random variable. Thus, the nonparametric bootstrap distribution does not converge to the correct distribution (1.4).

The residual-based bootstrap in this setting generates

$$Y_i^* = \widehat{\mu} + e_i^*,\ i = 1, \cdots, n, \tag{1.7}$$

where e_i^* are drawn independently with replacement from the set of centered residuals $\{e_1 - \overline{e}, \cdots, e_n - \overline{e}\}$, which reduces to the set $\{Y_1 - \overline{Y}, \cdots, Y_n - \overline{Y}\}$, because the raw residuals are $e_i = Y_i - \widehat{\mu},\ i = 1, \cdots, n$ and $\overline{e} = \overline{Y} - \widehat{\mu}$. In the regression setting these residuals would be adjusted by multiplying by $[n/(n-1)]^{1/2}$, but here it makes no difference for our asymptotic argument. Note that centering is crucial for consistency, because without subtracting $\overline{e} = \overline{Y} - \widehat{\mu}$ this residual-based bootstrap would be the same as the standard nonparametric bootstrap above. (This centering principle is well-known in the bootstrap regression literature, e.g., Shao and Tu, 1995, p.289.)

In large samples, $\sqrt{n}\widehat{\mu}^*$ based on (1.7) has approximately the distribution (conditional

on $\boldsymbol{Y}$)

$$(\sqrt{n}\widehat{\mu} + s_n Z_2) I(|\sqrt{n}\widehat{\mu} + s_n Z_2| > 1.96). \tag{1.8}$$

When $\mu = 0$, we may approximate by substituting (1.4) for $\sqrt{n}\widehat{\mu}$ and 1 for s_n yielding

$$\{Z_1 I(|Z_1| > 1.96) + Z_2\} I\{|Z_1 I(|Z_1| > 1.96) + Z_2| > 1.96\}. \tag{1.9}$$

Because $Z_1 I(|Z_1| > 1.96)$ is not identically 0, this limit for the residual-based bootstrap is not the same as (1.4) and thus not consistent. However, the probability that $Z_1 I(|Z_1| > 1.96)$ equals to 0 is 0.95, and (1.9) is much closer to (1.4) then (1.6) is.

When variable selection procedures are used to build sparse linear regression models, the noninformative covariates behave similar to the truncated estimator above. Thus the standard random-pair bootstrap samples do not estimate the variance of predictors with variable selection, and the residual-based bootstrap should work better for this purpose.

1.2.3 Bagging

As mentioned in Section 1.1, Breiman (1996a) showed that variable selection procedures can be unstable in the sense that small changes in the data can affect models chosen and the resulting predictions. Thus Breiman (1996b) proposed bagging (bootstrap aggregating) to reduce the instability. The simplest version of bagging is to average over estimates from the random-pair bootstrap samples. Considering the linear regression example when the goal is to make predictions from a reduced model built by a variable selection procedure, bagging proceeds as follows.

1. Generate a random-pair bootstrap sample from the original data set $\{(Y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$.

2. Run the variable selection procedure on the bootstrap sample and obtain $\widehat{\beta}_{0,b}$ and $\widehat{\boldsymbol{\beta}}_b$.

3. Repeat step 1 and step 2 B times.

4. Average the bootstrap $\widehat{\beta}_{0,b}$, $\widehat{\boldsymbol{\beta}}_b$ to have bagging regression coefficients, say $\widehat{\beta}_{0,BAG}$ and $\widehat{\boldsymbol{\beta}}_{BAG}$, and predict from them.

Note that with bagging, $\widehat{\boldsymbol{\beta}}_{BAG}$ typically has no zeroes even though each $\widehat{\boldsymbol{\beta}}_b$ may have many zeroes due to the variable selection. Thus, there is no variable selection in the averaged model. Buhlmann and Yu (2002) show that typically bagging adds bias to estimators but can reduce variance and mean square error of estimation in subset methods like forward selection and tree-based approaches. Boos, Stefanski, and Wu (2009) demonstrated that bagging FFSR lowered prediction error in situations where the predictors are correlated and the true beta is not too sparse.

1.3 Variable Selection Procedures

When a large number of predictor variables are available to build statistical models, it is desirable to include only the informative covariates in the model, because a parsimonious model is easier to interpret and also decreases the variation of prediction. The model with only the informative predictor variables will have better prediction accuracy than the full model including all available variables. Moreover, models that do not include weak predictors can have lower mean squared error of prediction than the full model.

In the literature there are many variable selection procedures available that can be used to build models containing only the most important informative variables. These variable selection procedures fall into three categories: subset selection procedures, regularized least square procedures and Bayesian variable selection procedures. We focus on subset selection procedures and regularized procedures in this research, and we briefly view these procedures in this section.

The review is restricted to the following linear regression setting. We assume that there are k_T covariates available, and the true model is $\boldsymbol{Y} = \beta_0\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{Y} =$

$(Y_1, \cdots, Y_n)^T$ is $n \times 1$ vector of response values, β_0 is the intercept, $\boldsymbol{X}$ is a $n \times k_T$ matrix with the ith row corresponding to the predictor variable vector for the ith observation, $\boldsymbol{\beta}$ is the regression coefficient vector, and $\boldsymbol{\epsilon}$ is an $n \times 1$ error vector with each element being an independent sample from $N(0, \sigma^2)$. Also we assume that the design matrix $[\mathbf{1}, \boldsymbol{X}]$ is of full rank. The vector of the j^{th} predictor variable values is the j^{th} column of $\boldsymbol{X}$.

1.3.1 Subset Selection Methods

Subset selection methods are the oldest and possibly the most frequently used variable selection methods in practice. The classic subset selection procedures include forward addition, backward elimination, stepwise selection and all subsets method. Subset selection methods build a sequence of candidate models with difference predictor variables included and then chooses the appropriate model according to some criteria. We give details of how forward addition builds the sequence and refer to Miller (2002) for other subset selection methods.

Forward addition starts from an empty model including only the intercept and builds a sequence of models by adding one predictor variable at a time. Forward addition chooses one variable to add from the variables not in the current model such that the resulting new model reduces the residual sum of squares most. Once a variable is added into the model, it stays in the model. Forward addition does not stop until all the variables have been included or some stopping criterion is satisfied. A related approach is forward stagewise procedure (Efron, et al., 2004) that moves at a much slower pace towards a final model. However, forward stagewise is more stable than the standard forward addition.

Usually forward addition does not require large amounts of computation due to the fact that forward addition starts from the smallest model and usually stops before reaching the full model. Moreover, forward addition is possible even in the case where there are more predictors than the observations, that is, $k_T > n$.

After subset selection methods generate a sequence of candidate models with different

predictor variables, one chooses the best model by some criteria. The most frequently used criteria are model-error-based criterion, information-based criterion, and the false selection rate criterion.

The model-error-based criteria choose the model to minimize the resulting model error. In a fixed design the model error (ME) is defined as ME $= (\widehat{\boldsymbol{Y}} - \boldsymbol{\mu})^T(\widehat{\boldsymbol{Y}} - \boldsymbol{\mu})$, where $\widehat{\boldsymbol{Y}} = \widehat{\beta}_0\mathbf{1} + \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is the predicted mean vector and $\boldsymbol{\mu} = \beta_0\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta}$ is the true mean vector. Wu, et al.(2007) modified the above definition of ME by dividing by the sample size n so that it is more comparable in situations with different sample sizes. Notice that ME is a function of the data set and is thus a random quantity. Model-error-based criteria select subsets of predictor variables by minimizing an estimate of the expected model errors.

Mallow's C_p (Mallows, 1973) estimates the expectation of model errors as $RSS_s - (n-2s)\widehat{\sigma}^2$ and the well-known Mallows' C_p criterion minimizes $RSS_s - (n-2s)\widehat{\sigma}^2$. Equivalently, Mallow's C_p minimizes the statistics, defined as

$$C_p = \frac{RSS_s}{\widehat{\sigma}^2} - (n - 2s), \tag{1.10}$$

where RSS_s is the residual sum of squares in the fitted model with s predictor variables, $\widehat{\sigma}^2$ is an unbiased estimator for σ^2 from the full model (thus requiring $n > k_T$), n is the sample size. When variable selection and estimation of parameters use the same data set, selection bias occurs. The C_p statistic fails to account for the selection bias. In the situation where selection bias is a severe problem, the model selected by Mallow's C_p criterion may be worse than the full model. Modifications of the Mallows' C_p can be found Gilmour (1996) and others.

Prediction accuracy can be measured by another closely related quantity, prediction error (PE), which is defined as PE=E$(\boldsymbol{Y}^{new} - \widehat{\boldsymbol{Y}})^T(\boldsymbol{Y}^{new} - \widehat{\boldsymbol{Y}})$. The expectation is with respect to $\boldsymbol{Y}^{new}$ and $\boldsymbol{Y}^{new}$ is an independent vector from the same distribution as $\boldsymbol{Y}$.

We now briefly explain cross-validation.

Cross-validation randomly divides the original data set into two parts, training data set, and test data set. It uses the training data set to build the model and uses the test data set to validate the model. The V-fold cross-validation method divides the original data evenly or approximate evenly into V parts, say $\text{data}_1, \cdots, \text{data}_V$. Let $\text{data}_{[j]}$ be all the data except for data_j. Then $\text{data}_{[j]}$ is used as the training data while data_j (the part left out) is used as the test data, $j = 1, \cdots, V$. If $\boldsymbol{Y}$ is the original response vector and $\widehat{\boldsymbol{Y}}$ is the prediction for data_j based on fitting the model to $\text{data}_{[j]}$, $j = 1, \cdots, V$, then the estimated prediction error is $\widehat{\text{PE}} = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})/n$. The model with smallest $\widehat{\text{PE}}$ statistic is chosen as the optimal model.

Information-based criteria choose the model to maximize an adjusted log likelihood that is penalized by the model complexity. Different penalty terms lead to different information-based criteria. The most frequently used information criteria are the Akaike information criterion (AIC) (Akaike, 1973) and the Bayesian information criterion (BIC) (Schwartz, 1978). BIC is also often referred as the Schwartz criterion. The AIC and BIC statistics are defined as

$$AIC = -2 \log L + 2s, \tag{1.11}$$

$$BIC = -2 \log L + s \log n, \tag{1.12}$$

where L is the maximized likelihood of the sample data set, s is the number of predictor variables in the model, and n is the sample size. Methods based on AIC and BIC criteria choose the model minimizing the above AIC and BIC statistics. Stone (1977, 1979) showed that AIC and cross-validation are asymptotically equivalent. BIC is selection consistent when there is an exact model and AIC is optimal when there is no exact model. Hurvich and Tsai (1989) modified the AIC statistics in (1.11) to have better

small sample property. Hurvich and Tsai's version of AIC is

$$AIC_c = -2\log L + s + \frac{2s(s+1)}{n-s-1}.$$

When the sequence of candidate models are generated by the forward addition method, the final model can be chosen by specifying the α-to-enter value. Wu, Boos and Stefanksi (2007) chose the α-to-enter by adding uninformative covariates and monitoring the proportion of the selected uninformative variables. For each α-to-enter in a reasonable range, they estimated the proportion of the selected uninformative variables by simulation and used this proportion to approximate the false selection rate (FSR) when the original sample data is used. Thus, an approximate relationship between α-to-enter and the false selection rate is established, and accordingly the FSR method chooses the α-to-enter value.

To be more specific, Wu, Boos and Stefanski (2007) estimates the false selection rate for each α-to-enter as

$$\widehat{\gamma}(\alpha) = \frac{\{k_T - S(\alpha)\}\widehat{\theta}(\alpha)}{1 + S(\alpha)},$$

where k_T is the number of original covariates, $S(\alpha)$ is the total number of covariates (excluding the intercept) in the selected model, and $\widehat{\theta}(\alpha)$ is the estimated rate of the uninformative covariates entering the selected model. Then the FSR α-to-enter is selected by $\widehat{\alpha} = \sup_{\alpha \leq \alpha_{max}} \{\alpha : \widehat{\gamma}(\alpha) \leq \gamma_0\}$, where α_{max} is the largest possible α-to-enter value and γ_0 is the targeted false selection rate. Usually γ_0 is 0.05. The FSR method estimates $\widehat{\theta}(\alpha)$ by simulation as follows:

1. Generate k_P uninformative covariate and adds them to the original k_T covariates.

2. Perform the forward selection and use α-to-enter to select the model.

3. Compute the proportion of uninformative covariates included in the model from Step 2 as $U_{P,b}(\alpha)/k_P$, where $U_{P,b}$ is the number of selected uninformative covariates.

4. Repeat above three steps B times and average the B resulting proportions $U_{P,b}(\alpha)/k_P$ to estimate $\widehat{\theta}(\alpha) = \overline{U}_P/k_P$.

To avoid the above simulation used to estimate the false selection rate, Boos, Stefanski and Wu (2009) proposed the fast FSR method. The fast FSR estimates the false selection rate as $\widehat{\theta}(\alpha) = \alpha$ and chooses the final model size as

$$k(\gamma_0) = \max\left\{ i : \widetilde{p}_i \leq \frac{\gamma_0[1 + S(\widetilde{p}_i)]}{k_T - S(\widetilde{p}_i)} \text{ and } \widetilde{p}_i \leq \alpha_{max} \right\},$$

where $\widetilde{p}_i$ are the monotonized p-values of the forward addition sequence. The fast FSR monotonizes the k_T p-values obtained from each step of forward addition so that $\widetilde{p}_1 \leq \widetilde{p}_2 \leq \cdots \leq \widetilde{p}_{k_T}$ by carrying the largest p-values forward. For instance, if $p_{i+1} < \widetilde{p}_i$ in step $i+1$ of forward addition, $\widetilde{p}_i$ is assigned as the new p-value of the newly added predictor.

1.3.2 Regularized Least Square Methods

Regularized least square procedures minimize a penalized residual sum of squares of the form $\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - x_i^T\widehat{\boldsymbol{\beta}})^2 +$ penalty term. The penalty term controls the model complexity. The regularized least square methods do the variable selection and estimate the regression coefficients at the same time. The most well-known examples are the LASSO (Tibishirani, 1996), the adaptive LASSO (Zou, 2006) and SCAD (Fan and Li, 2001).

Usually the penalty term is a function of the regression coefficient vector $\boldsymbol{\beta}$. When the penalty term is the L_2 norm of $\boldsymbol{\beta}$, that is, penalty term $= \lambda \sum_{i=1}^{k_T} \beta_i^2$, then the regularized method is ridge regression. The regression coefficients are estimated by shrinking the ordinary least square (OLS) estimates by the same factor. However, ridge regression does not shrink OLS estimates to zero and consequently it does not function as variable selection.

In order to take advantage of ridge regression's improvement in prediction accuracy

and also to yield a parsimonious model to increase interpretability, Tibishirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) that uses the L_1 norm of $\boldsymbol{\beta}$ as the penalty term. The LASSO estimates the regression coefficients by minimizing $\sum_{i=1}^{n}\left(Y_i-\beta_0-x_i^T\beta\right)^2+\lambda\sum_{j=1}^{k_T}|\beta_j|$, where λ is a nonnegative tuning parameter. Before applying the LASSO procedure, the vectors of predictor variable need to be centered and scaled. The LASSO regression coefficients are estimated as

$$\widehat{\beta}_0=\overline{\boldsymbol{Y}} \text{ and } \widehat{\boldsymbol{\beta}}_{Lasso}=\arg\min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{n}\left(Y_i-\overline{\boldsymbol{Y}}-\boldsymbol{x}_i^T\beta\right)^2+\lambda\sum_{j=1}^{k_T}|\beta_j|\right\}.$$

The LASSO shrinks the regression coefficients toward 0. If the resulting estimated coefficient is 0, the corresponding variable is not included in the model and thus the LASSO can produce a sparse model. In the above penalized residual sum of squares, every nonzero regression coefficient is assigned the same weight λ, and consequently the LASSO estimator does not have the oracle property (Fan and Li, 2001). The oracle property means that the probability of selecting the true model goes to 1 as the sample size increases (selection consistency) and that the asymptotic normal distribution of the $\widehat{\boldsymbol{\beta}}$ of the selected variables is the same as the ordinary least square estimates using only the informative variables.

Zou (2006) modified the L_1 penalized sum square of residuals by assigning different weights to different nonzero regression coefficients and generated adaptive LASSO (ALASSO) estimators. To be more specific, the ALASSO estimators are

$$\widehat{\beta}_0=\overline{\boldsymbol{Y}} \text{ and } \widehat{\boldsymbol{\beta}}_{Alasso}=\arg\min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{n}\left(Y_i-\overline{\boldsymbol{Y}}-\boldsymbol{x}_i^T\beta\right)^2+\lambda\sum_{j=1}^{k_T}w_j|\beta_j|\right\},$$

where the weight w_j usually takes the value $1/|\widehat{\beta}_{j,ols}|$, and $\beta_{j,ols}$ is the OLS regression coefficient estimate. Zou (2006) showed that the ALASSO estimator has the oracle property.

Zou and Hastie (2005) proposed the elastic net procedure that penalizes the least square term by adding a linear combination of the L_1 and L_2 norm of $\boldsymbol{\beta}$. In doing so, the elastic net can not only produces sparse models, but it also enhances the capability of selecting grouped variables. That is, the elastic net more likely includes the highly correlated predictors in the model. The LASSO and ALASSO both might excessively penalize the large regression coefficients and lead to severe bias problem in prediction.

Another regularized estimator with the oracle property was introduced by Fan and Li (2001), called the smoothly clipped absolute deviation (SCAD) estimator. The SCAD estimator is computationally more challenging than the LASSO and the adaptive LASSO because the penalty term is not convex.

All regularized least square methods simultaneously select the variable and estimate the regression coefficient of the resulting model at the same time. Thus, for a given tuning parameter λ, it is possible to discuss the sampling property of estimators. The tuning parameter is often chosen by cross-validation, generalized cross-validation or BIC.

Chapter 2

Existing Methods for Estimating the Variance of Bagging Estimators

In order to make statistical inference, it is important to obtain a valid estimate for the variance of estimates. The use of variable selection makes it a challenging problem because it is difficult to describe the variable selection process analytically. Ignoring the variable selection process and applying the usual inferential methods to the selected model usually underestimates the variance, and the corresponding nominal confidence interval is too narrow resulting in undercoverage (Zhang, 1992). Also the usual inference methods produce zero variance estimates for the regression coefficient estimates associated with non-selected predictor variables (Tibshirani, 1996; Zou, 2006; Fan and Li, 2001). Shen, et al. (2004) proposed an optimal linear approximation method to estimate the variance. The bootstrap method should automatically account for the variable selection process, but Knight and Fu (2000) showed that the usual residual-based bootstrap variance estimator for the LASSO predictor is not consistent. However, Chatterjee and Lahiri (2010) showed that the residual-based bootstrap estimates the variance of adaptive LASSO estimators consistently.

When we combine the bagging technique with variable selection procedures in or-

der to improve the prediction accuracy, it is more challenging to estimate the variance of bagging predictors. Very little research has been done on this, and it seems to us that the only methods available to estimate the variance of bagging estimators are the jackknife-after-bootstrap and the bootstrap-after-bootstrap. In this chapter we review these two methods and gain some insight into the impact of bootstrap sample sizes on the performance of bootstrap-after-bootstrap in one particular case.

2.1 Jackknife-after-Bootstrap

The jackknife variance estimator for an estimator $\widehat{\gamma}$ from an i.i.d. sample of size n is

$$\widehat{V}_{\text{J}} = \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{\gamma}_{[i]} - \overline{\widehat{\gamma}}_1 \right)^2, \tag{2.1}$$

where $\widehat{\gamma}_{[i]}$ is the "leave-1-out" estimator obtaining by computing $\widehat{\gamma}$ with the ith pair deleted from the sample, and $\overline{\widehat{\gamma}}_1 = n^{-1} \sum_{i=1}^{n} \widehat{\gamma}_{[i]}$. Efron (1992, p. 85) defines the *jackknife influence function* by

$$u_i = (n-1)(\overline{\widehat{\gamma}}_1 - \widehat{\gamma}_{[i]})$$

and rewrites (2.1) as $\widehat{V}_{\text{J}} = \{n(n-1)\}^{-1} \sum_{i=1}^{n} u_i^2$.

In order to get a jackknife variance estimate $\widehat{V}_{\text{J}}$ for $\widehat{\gamma}_{BAG}$, it would appear that we would have to generate B resamples from the original sample with the ith pair deleted, compute $\widehat{\gamma}_{BAG,[i]}$, repeat this for $i = 1, \ldots, n$, and then use (2.1) on these n "leave-1-out" values. This would appear to require $nB = n \times B$ calculations of the original estimator. However, Efron (1992) noticed that bootstrap estimators admit a simple jackknife estimator of variance by just keeping track of which bootstrap samples omit the ith pair, $i = 1, \ldots, n$.

The jackknife-after-bootstrap (Efron, 1992) takes advantage of a useful observation that in the set of B samples there is a subset C_i of samples where the ith data value

does not appear. Denote by B_i the number of samples in C_i. On average, there are $\mathrm{E}(B_i) = (1-1/n)^n B \approx \exp(-1)B = .368B$ such samples. Let $\widehat{\gamma}_1^{(i)}, \ldots, \widehat{\gamma}_{B_i}^{(i)}$ be the underlying estimators $\widehat{\gamma}$ computed from those samples and let $\widehat{\gamma}_{B,[i]}$ be the bootstrap estimator computed from these B_i estimates. $\widehat{\gamma}_{B,[i]}$ is a good approximation to $\widehat{\gamma}_{BAG,[i]}$ as $B \to \infty$. An approximate jackknife influence function is then

$$\widetilde{u}_i = (n-1)\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\gamma}_{B,[i]} - \widehat{\gamma}_{B,[i]}\right),$$

and the jackknife-after-bootstrap variance estimator is

$$\widehat{V}_{\mathrm{JaB}} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\widetilde{u}_i^2.$$

Efron and Tibshirani (page 277-278, 1993) demonstrated by simulation that the bootstrap sample size B needs to be much larger than n in order for $\widehat{V}_{\mathrm{JaB}}$ to be approximately unbiased in a simple situation, where the underlying estimator $\widehat{\gamma}$ is the sample mean. In general, though, there has been very little research on $\widehat{V}_{\mathrm{JaB}}$, certainly none when $\widehat{\gamma}$ involves variable selection. In Chapter 5, we show by simulation that the jackknife-after-bootstrap overestimates the variance of bagging predictors. We believe the reason is related to the simple example give in Section 1.2.2 where the residual-based bootstrap is required.

2.2 Bootstrap-after-Bootstrap

Because the bagging estimator $\widehat{\gamma}_{BAG}$ is a bootstrap estimator, it seems that level-2 bootstrap samples would be necessary to use bootstrap technology to estimate the variance of $\widehat{\gamma}_{BAG}$. Suppose that the bagging estimator is based on B_2 samples. The bootstrap-after-bootstrap variance estimator first draws B_1 level-1 bootstrap samples. Then from each of these level-1 samples, B_2 level-2 samples are drawn and $\widehat{\gamma}_{b,BAG} = \sum_{i=1}^{B_2}\widehat{\gamma}_{b,i}/B_2$ is com-

puted, where $\widehat{\gamma}_{b,i}$ is the estimate of $\widehat{\gamma}$ from the ith level-2 bootstrap sample drawn from the bth level-1 sample. Thus, we obtain B_1 bagging estimators $\widehat{\gamma}_{1,BAG}, \ldots, \widehat{\gamma}_{B_1,BAG}$. The bootstrap-after-bootstrap variance estimator is then just the sample variance of these latter B_1 bagging estimates, i.e.,

$$\widehat{\mathrm{Var}}_{BaB}(\widehat{\gamma}_{BAG}) = \frac{1}{B_1 - 1} \sum_{b=1}^{B_2} (\widehat{\gamma}_{b,BAG} - \overline{\widehat{\gamma}}_{\cdot,BAG})^2.$$

This bootstrap-after-bootstrap approach requires $B_1 B_2 = B_1 \times B_2$ model fits.

Sexton and Laake (2009) studied the bootstrap-after-bootstrap approach (called the "Brute Force Method" there) for use with regression trees. But Sexton and Laake (2009) used the random-pair bootstrap for both levels of bootstrapping. We have found for linear regression that it is important to use the residual-based bootstrap for the level-1 bootstrap samples.

2.3 Bootstrap Selection Matrix

2.3.1 Definition of Bootstrap Selection Matrix

Next we investigate the bootstrap-after-bootstrap variance estimator in a special situation where the basic estimator $\widehat{\gamma}$ is the sample mean. To facilitate the study, we introduce the $n \times n$ **B**ootsrap **S**election **M**atrix (**BSM**), $\boldsymbol{R}$, defined as

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{r}_1^T \\ \vdots \\ \boldsymbol{r}_n^T \end{pmatrix},$$

where $\boldsymbol{r}_1, \cdots, \boldsymbol{r}_1$ are i.i.d. multinomial $(1; 1/n, \cdots, 1/n)$ random vectors. For data vector $\boldsymbol{Y}$, $\boldsymbol{RY}$ is a random-pair bootstrap sample.

The observed data is an i.i.d. sample $Y_1, \cdots, Y_n$, where $\mathrm{E}(Y_i) = \mu$. In analogy with

linear models, $Y_i = \mu + Y_i - \mu$, so that $Y_i - \mu$ are errors. We use residual-based bootstrap sampling to generate level-1 bootstrap samples. From the observed data, the estimated response mean is $\mathbf{1}^T\boldsymbol{Y}/n = \overline{Y}$ and the residual vector is $\boldsymbol{e} = \boldsymbol{Y} - \mathbf{1}\overline{Y}$. Then we treat the residual vector as a finite population and draw a random-pair bootstrap sample from it. Finally, we generate a level-1 $\boldsymbol{Y}^{(b)} = (Y_1^{(b)}, \cdots, Y_n^{(b)})^T$ bootstrap sample as the sum of the predicted vector and the random-pair bootstrap sample of the residual vector. Using the notation of **BSM** matrix, the level-1 bootstrap sample can be written as

$$\begin{aligned}
\begin{pmatrix} Y_1^{(b)} \\ \vdots \\ Y_n^{(b)} \end{pmatrix} &= \mathbf{1}\overline{Y} + \boldsymbol{R}_b\left(\boldsymbol{Y} - \mathbf{1}\overline{Y}\right) \\
&= \mathbf{1}\overline{Y} + \boldsymbol{R}_b\boldsymbol{Y} - \boldsymbol{R}_b\mathbf{1}\overline{Y} \\
&= \mathbf{1}\overline{Y} + \boldsymbol{R}_b\boldsymbol{Y} - \mathbf{1}\overline{Y} \\
&= \boldsymbol{R}_b\boldsymbol{Y},
\end{aligned}$$

where $\boldsymbol{R}_b$ is the **BSM** matrix corresponding to the bth bootstrap sample. The next-to-last step uses the fact that $\boldsymbol{R}_b\mathbf{1} = \mathbf{1}$. Thus, in the sample mean case, the level-1 residual-based bootstrap reduces to a level-1 random-pair bootstrap.

To implement the bootstrap-after-bootstrap approach, we draw B_2 level-2 bootstrap samples by random-pair bootstrapping from each level-1 bootstrap sample $\boldsymbol{R}_b\boldsymbol{Y}$. Using **BSM** notation, a level-2 bootstrap sample from the bth level-1 bootstrap sample can be written as

$$\begin{pmatrix} Y_1^{(b,j)} \\ \vdots \\ Y_n^{(b,j)} \end{pmatrix} = \boldsymbol{R}_{b,j}\begin{pmatrix} Y_1^{(b)} \\ \vdots \\ Y_n^{(b)} \end{pmatrix} = \boldsymbol{R}_{b,j}\boldsymbol{R}_b\boldsymbol{Y}, \text{ for } j = 1, \cdots, B_2,$$

where $\boldsymbol{R}_{b,j}$ and $\boldsymbol{R}_b$ are independent **BSM** matrices. Thus, bootstrap selection matrices

help us define both level-1 and level-2 bootstrap samples.

2.3.2 Properties of Bootstrap Selection Matrices

The following properties of bootstrap selection matrices are useful in our investigation of the bootstrap-after-bootstrap variance estimator in a special situation in the next section.

Let $Y_1, \cdots, Y_n$ be an random i.i.d. sample from a distribution with mean equal to μ and variance equal to σ^2. Let $\boldsymbol{R}$ be a bootstrap selection matrix independent of the i.i.d. sample. Then we have the following facts.

(F1) $\mathrm{E}(\boldsymbol{R}) = \frac{1}{n}\mathbf{1}\mathbf{1}^T$.

Proof. Let $\boldsymbol{r}_1^T$ denote the first row of bootstrap selection matrix $\boldsymbol{R}$. It suffices to show that $\mathrm{E}(\boldsymbol{r}_1^T) = \mathbf{1}^T/n$. Notice that $\boldsymbol{r}_1^T$ takes the values $(1, 0, \cdots, 0), (0, 1, \cdots, 0) \cdots,$ $(0, 0, \cdots, 1)$ with equal probability $1/n$. Hence,

$$\mathrm{E}(\boldsymbol{r}_1^T) = \frac{1}{n}\left\{(1, 0, \cdots, 0) + (0, 1, \cdots, 0) \cdots + (0, 0, \cdots, 1)\right\} = \frac{1}{n}\mathbf{1}^T.$$

■

(F2) $\mathrm{E}(\boldsymbol{R}\boldsymbol{R}^T) = \frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{n-1}{n}\boldsymbol{I}_{n\times n}$.

Proof. Let $\boldsymbol{r}_i^T$ denote the ith row of bootstrap selection matrix $\boldsymbol{R}$, then the ith row jth column element of $\boldsymbol{R}\boldsymbol{R}^T$ is $\boldsymbol{r}_i^T\boldsymbol{r}_j$.

For $i \neq j$, $\mathrm{E}(\boldsymbol{r}_i^T\boldsymbol{r}_j) = \mathrm{E}(\boldsymbol{r}_i^T)\mathrm{E}(\boldsymbol{r}_j) = \frac{1}{n}\mathbf{1}^T \times \frac{1}{n}\mathbf{1} = \frac{1}{n}$.

For $i = j$, $\mathrm{E}(\boldsymbol{r}_i^T\boldsymbol{r}_i) = \mathrm{E}(\boldsymbol{r}_i^T)\mathrm{E}(\boldsymbol{r}_i) + \mathrm{trace}\left\{\mathrm{Cov}(\boldsymbol{r}_i)\right\}$.

From the multinomial distribution, we have

$$\mathrm{Cov}(\boldsymbol{r}_i) = \begin{pmatrix} \frac{n-1}{n^2} & -\frac{1}{n^2} & \cdots & -\frac{1}{n^2} \\ -\frac{1}{n^2} & \frac{n-1}{n^2} & \cdots & -\frac{1}{n^2} \\ \cdot & \cdot & \cdots & -\frac{1}{n^2} \\ -\frac{1}{n^2} & -\frac{1}{n^2} & \cdots & \frac{n-1}{n^2} \end{pmatrix}$$

and $\mathrm{trace}\{\mathrm{Cov}(\boldsymbol{r}_i)\} = \dfrac{n-1}{n}$.

Also, $\mathrm{E}(\boldsymbol{r}_i^T)\mathrm{E}(\boldsymbol{r}_i) = \dfrac{1}{n}\mathbf{1}^T \times \dfrac{1}{n}\mathbf{1} = \dfrac{1}{n}$.

Thus $\mathrm{E}(\boldsymbol{r}_i^T\boldsymbol{r}_i) = 1$.

Therefore, $\mathrm{E}(\boldsymbol{R}\boldsymbol{R}^T) = \dfrac{1}{n}\mathbf{1}\mathbf{1}^T + \dfrac{n-1}{n}\boldsymbol{I}_{n\times n}$. ■

(F3) $\mathrm{E}\left\{\dfrac{1}{n}\mathbf{1}^T\boldsymbol{R}\boldsymbol{Y}|\boldsymbol{Y}\right\} = \overline{Y}$.

Proof.

$$\mathrm{E}\left\{\frac{1}{n}\mathbf{1}^T\boldsymbol{R}\boldsymbol{Y}|\boldsymbol{Y}\right\} = \frac{1}{n}\mathbf{1}^T\left(\frac{\mathbf{1}\mathbf{1}^T}{n}\right)\boldsymbol{Y} = \frac{1}{n}\mathbf{1}^T\boldsymbol{Y} = \overline{Y}\,.$$

■

(F4) $\mathrm{E}\left\{\dfrac{1}{n}\mathbf{1}^T\boldsymbol{R}\boldsymbol{Y}|\boldsymbol{R}\right\} = \mu$.

Proof.

$$\mathrm{E}\left\{\frac{1}{n}\mathbf{1}^T\boldsymbol{R}\boldsymbol{Y}|\boldsymbol{R}\right\} = \frac{1}{n}\mathbf{1}^T\boldsymbol{R}\mathbf{1}\mu = \frac{1}{n}\mathbf{1}\mathbf{1}^T\mu = \mu\,.$$

■

(F5) $\mathrm{E}\left\{\dfrac{1}{n}\mathbf{1}^T\boldsymbol{R}\boldsymbol{Y}\right\} = \mu$.

Proof.

$$\mathrm{E}\left\{\frac{1}{n}\mathbf{1}^T\boldsymbol{R}\boldsymbol{Y}\right\} = \mathrm{E}\left[\mathrm{E}\left\{\frac{1}{n}\mathbf{1}^T\boldsymbol{R}\boldsymbol{Y}|\boldsymbol{R}\right\}\right] = \mathrm{E}(\mu) = \mu\,.$$

■

(F6) $\mathrm{Var}\left\{\dfrac{1}{n}\mathbf{1}^T\boldsymbol{R}\boldsymbol{Y}|\boldsymbol{R}\right\} = \dfrac{\sigma^2}{n^2}\left(\mathbf{1}^T\boldsymbol{R}\boldsymbol{R}^T\mathbf{1}\right)$.

Proof.

$$\text{Var}\left\{\frac{1}{n}\mathbf{1}^T\boldsymbol{RY}|\boldsymbol{R}\right\} = \frac{1}{n^2}\mathbf{1}^T\boldsymbol{R}(\sigma^2\boldsymbol{I})\boldsymbol{R}^T\mathbf{1} = \frac{\sigma^2}{n^2}\left(\mathbf{1}^T\boldsymbol{R}\boldsymbol{R}^T\mathbf{1}\right) .$$

■

(F7) $\text{Var}\left\{\frac{1}{n}\mathbf{1}^T\boldsymbol{RY}|\boldsymbol{Y}\right\} = \frac{n-1}{n^2}S_Y^2 .$

Proof.

$$\begin{aligned}
\text{Var}\left\{\frac{1}{n}\mathbf{1}^T\boldsymbol{RY}|\boldsymbol{Y}\right\} &= \frac{1}{n}\text{Var}(Y^* \mid \boldsymbol{Y}), \text{ where } Y^* \text{ is a random draw from } \boldsymbol{Y}, \\
&= \frac{1}{n}\left\{\frac{1}{n}\sum_{j=1}^{n}(Y_j - \overline{Y})^2\right\} \\
&= \frac{n-1}{n^2}\cdot\left\{\frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2\right\} \\
&= \frac{n-1}{n^2}S_Y^2 .
\end{aligned}$$

■

(F8) $\text{E}\left\{\text{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{RY}|\boldsymbol{R}\right)\right\} = \frac{\sigma^2}{n}\left(\frac{2n-1}{n}\right).$

Proof.

$$\begin{aligned}
\text{E}\left\{\text{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{RY}|\boldsymbol{R}\right)\right\} &= \text{E}\left\{\frac{\sigma^2}{n^2}\left(\mathbf{1}^T\boldsymbol{R}\boldsymbol{R}^T\mathbf{1}\right)\right\} \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T\text{E}(\boldsymbol{R}\boldsymbol{R}^T)\mathbf{1} \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T\left(\frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{n-1}{n}\boldsymbol{I}\right)\mathbf{1} \\
&= \frac{\sigma^2}{n}\left(\frac{2n-1}{n}\right),
\end{aligned}$$

where the next to last step results from **(F2)**. ■

(F9) $\text{E}\left\{\text{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{RY}|\boldsymbol{Y}\right)\right\} = \frac{\sigma^2}{n}\frac{n-1}{n}.$

Proof. From **(F7)**, we have $\mathrm{E}\left\{\mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{RY}|\boldsymbol{Y}\right)\right\} = \frac{n-1}{n^2}\mathrm{E}(S_Y^2) = \frac{\sigma^2}{n}\frac{n-1}{n}$. ■

(F10) $\mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{RY}\right) = \frac{\sigma^2}{n}\left(\frac{2n-1}{n}\right)$.

Proof.

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{RY}\right) &= \mathrm{E}\left\{\mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{RY}|\boldsymbol{R}\right)\right\} + \mathrm{Var}\left\{\mathrm{E}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{RY}|\boldsymbol{R}\right)\right\} \\
&= \frac{\sigma^2}{n^2}\mathrm{E}\left(\mathbf{1}^T\boldsymbol{R}\boldsymbol{R}^T\mathbf{1}\right) + \mathrm{Var}(\mu), \text{ by } \mathbf{(F4)} \text{ and } \mathbf{(F6)} \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T\left(\frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{n-1}{n}\boldsymbol{I}\right)\mathbf{1}, \text{ by } \mathbf{(F2)} \\
&= \frac{\sigma^2}{n}\left(\frac{2n-1}{n}\right).
\end{aligned}$$

■

(F11) $\mathrm{Var}(\boldsymbol{RY}) = \sigma^2\left(\frac{\mathbf{1}\mathbf{1}^T}{n} + \frac{n-1}{n}\boldsymbol{I}\right)$.

Proof.

$$\mathrm{Var}(\boldsymbol{RY}) = \mathrm{E}\{\mathrm{Var}(\boldsymbol{RY}|\boldsymbol{R})\} + \mathrm{Var}\{\mathrm{E}(\boldsymbol{RY}|\boldsymbol{R})\} = \mathrm{E}(\boldsymbol{R}\boldsymbol{R}^T\sigma^2) = \sigma^2\left(\frac{\mathbf{1}\mathbf{1}^T}{n} + \frac{n-1}{n}\boldsymbol{I}\right).$$

■

For the next four results, let $\boldsymbol{W} = \boldsymbol{R}_1\boldsymbol{R}_2$, where $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are independent **BSM** matrices.

(F12) $\mathrm{E}(\boldsymbol{W}) = \frac{\mathbf{1}\mathbf{1}^T}{n}$.

Proof. By the independence of $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$,

$$\begin{aligned} \mathrm{E}(\boldsymbol{W}) &= \mathrm{E}(\boldsymbol{R}_1)\mathrm{E}(\boldsymbol{R}_2) \\ &= \frac{\mathbf{1}\mathbf{1}^T}{n} \times \frac{\mathbf{1}\mathbf{1}^T}{n}, \text{ by } \mathbf{(F1)}, \\ &= \frac{\mathbf{1}\mathbf{1}^T}{n}. \end{aligned}$$

■

(F13) $\mathrm{E}(\boldsymbol{W}\boldsymbol{W}^T) = \dfrac{2n-1}{n^2}\mathbf{1}\mathbf{1}^T + \left(\dfrac{n-1}{n}\right)^2 \boldsymbol{I}\,.$

Proof.

$$\begin{aligned} \mathrm{E}(\boldsymbol{W}\boldsymbol{W}^T) &= \mathrm{E}(\boldsymbol{R}_1\boldsymbol{R}_2\boldsymbol{R}_2^T\boldsymbol{R}_1^T) \\ &= \mathrm{E}\left\{\boldsymbol{R}_1\left(\frac{\mathbf{1}\mathbf{1}^T}{n} + \frac{n-1}{n}\boldsymbol{I}\right)\boldsymbol{R}_1^T\right\}, \text{ by } \mathbf{(F2)} \\ &= \frac{\mathbf{1}\mathbf{1}^T}{n} + \frac{n-1}{n}\mathrm{E}(\boldsymbol{R}_1\boldsymbol{R}_1^T) \\ &= \frac{\mathbf{1}\mathbf{1}^T}{n} + \frac{n-1}{n}\left(\frac{\mathbf{1}\mathbf{1}^T}{n} + \frac{n-1}{n}\boldsymbol{I}\right), \text{ by } \mathbf{(F2)}, \\ &= \frac{2n-1}{n^2}\mathbf{1}\mathbf{1}^T + \left(\frac{n-1}{n}\right)^2\boldsymbol{I}\,. \end{aligned}$$

■

(F14) $\mathrm{Cov}(\boldsymbol{R}_1\boldsymbol{Y}, \boldsymbol{R}_2\boldsymbol{Y}) = \sigma^2\dfrac{\mathbf{1}\mathbf{1}^T}{n}\,.$

Proof.

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{R}_1\boldsymbol{Y}, \boldsymbol{R}_2\boldsymbol{Y}) &= \mathrm{E}\{\boldsymbol{R}_1\boldsymbol{Y}\cdot(\boldsymbol{R}_2\boldsymbol{Y})^T\} - \mathrm{E}(\boldsymbol{R}_1\boldsymbol{Y})\cdot\{\mathrm{E}(\boldsymbol{R}_2\boldsymbol{Y})\}^T \\ \mathrm{E}(\boldsymbol{R}_1\boldsymbol{Y}\cdot\boldsymbol{Y}^T\boldsymbol{R}_2^T) &= \mathrm{E}\boldsymbol{R}_1\cdot\mathrm{E}(\boldsymbol{Y}\boldsymbol{Y}^T)\cdot(\mathrm{E}\boldsymbol{R}_2)^T \\ &= \frac{\mathbf{1}\mathbf{1}^T}{n}\left(\mu^2\mathbf{1}\mathbf{1}^T + \sigma^2\boldsymbol{I}_n\right)\frac{\mathbf{1}\mathbf{1}^T}{n} \\ &= \mu^2\mathbf{1}\mathbf{1}^T + \sigma^2\frac{\mathbf{1}\mathbf{1}^T}{n}\,. \end{aligned}$$

Hence,

$$\text{Cov}(\boldsymbol{R}_1\boldsymbol{Y}, \boldsymbol{R}_2\boldsymbol{Y}) = \mu^2\mathbf{1}\mathbf{1}^T + \sigma^2\frac{\mathbf{1}\mathbf{1}^T}{n} - \mathbf{1}\mathbf{1}^T\mu^2 = \sigma^2\frac{\mathbf{1}\mathbf{1}^T}{n}.$$

■

(F15)

$$\text{Cov}\begin{pmatrix} \boldsymbol{R}_1\boldsymbol{Y} \\ \vdots \\ \boldsymbol{R}_M\boldsymbol{Y} \end{pmatrix} = \frac{\sigma^2}{n}\begin{bmatrix} \mathbf{1}\mathbf{1}^T & \cdots & \mathbf{1}\mathbf{1}^T \\ \vdots & & \vdots \\ \mathbf{1}\mathbf{1}^T & \cdots & \mathbf{1}\mathbf{1}^T \end{bmatrix} + \frac{(n-1)\sigma^2}{n}\begin{bmatrix} \boldsymbol{I}_n & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \boldsymbol{I}_n \end{bmatrix},$$

where the above two matrices are both $nM \times nM$.

Proof. From **(F14)**, we have $\text{Cov}(\boldsymbol{R}_i\boldsymbol{Y}, \boldsymbol{R}_j\boldsymbol{Y}) = \sigma^2\frac{\mathbf{1}\mathbf{1}^T}{n}$ for $i \neq j$.

From **(F11)**, we have $\text{Cov}(\boldsymbol{R}_i\boldsymbol{Y}, \boldsymbol{R}_i\boldsymbol{Y}) = \sigma^2\left(\frac{\mathbf{1}\mathbf{1}^T}{n} + \frac{n-1}{n}\boldsymbol{I}\right)$.

It follows that **(F15)** holds true. ■

2.4 Bootstrap-after-Bootstrap in the Sample Mean Case

Next we study the performance of the bootstrap-after-bootstrap variance estimator in estimating the variance of $\widehat{\gamma}_{BAG}$, when the basic estimator $\widehat{\gamma}$ is the sample mean. Let $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ be an i.i.d. sample from a distribution with $\text{E}(Y) = \mu$ and $\text{Var}(Y) = \sigma^2$. We draw B_1 level-1 bootstrap samples and from each level-1 sample draw B_2 bootstrap samples. In the sample mean case, the residual-based bootstrap and random-pair bootstrap produce the same bootstrap samples, and we denote the level-1 samples as $\boldsymbol{R}_b\boldsymbol{Y}$ and level-2 samples as $\boldsymbol{R}_{b,j}\boldsymbol{R}_b\boldsymbol{Y}$, for $b = 1, \cdots, B_1$ and $j = 1, \cdots, B_2$. B_1 and B_2 are the level-1 and level-2 bootstrap sample sizes, respectively.

The bagging estimator based on the bth level-1 bootstrap sample $\boldsymbol{R}_b\boldsymbol{Y}$ is

$$\widehat{\gamma}_{b,BAG} = \frac{1}{B_2}\sum_{j=1}^{B_2}\widehat{\gamma}_{b,j} = \frac{1}{B_2}\sum_{j=1}^{B_2}\frac{\mathbf{1}^T\boldsymbol{R}_{b,j}\boldsymbol{R}_b\boldsymbol{Y}}{n}, \text{ for } b = 1, \cdots, B_1.$$

The bootstrap-after-bootstrap variance of $\widehat{\gamma}_{BAG}$ is estimated by

$$\widehat{\mathrm{Var}}_{BaB}(\widehat{\gamma}_{BAG}) = \frac{1}{B_1 - 1}\sum_{b=1}^{B_1}(\widehat{\gamma}_{b,BAG} - \overline{\widehat{\gamma}}_{\cdot,BAG})^2, \tag{2.2}$$

where $\overline{\widehat{\gamma}}_{\cdot,BAG}$ is the sample mean of $\widehat{\gamma}_{1,BAG}, \cdots, \widehat{\gamma}_{B_1,BAG}$.

The level-2 bootstrap sample size B_2 is not necessarily the same as the bootstrap sample size B in obtaining the bagging estimator. The reason is that the bagging estimator estimates the parameter of interest, and we use B bootstrap samples to calculate the estimator of interest. The above bootstrap-after-bootstrap procedure estimates the variance of the bagging estimator. The bootstrap-after-bootstrap procedure requires B_1B_2 functional calls. For computational efficiency we may want $B_1 < B$ and $B_2 < B$. The following computation in the sample mean case demonstrates that smaller B_1 and B_2 do not severely affect its performance.

When the basic statistic is sample mean, $\widehat{\gamma}_b = \mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y}/n$. First we have

$$\widehat{\gamma}_{BAG} = \frac{\mathbf{1}^T(\sum_{b=1}^{B}\boldsymbol{R}_b)\boldsymbol{Y}}{nB},$$

and by direct computation

$$\mathrm{Var}(\widehat{\gamma}_{BAG}) = \frac{\sigma^2}{n}\left(1 + \frac{n-1}{nB}\right), \tag{2.3}$$

which will be proved in Chapter 3. In order to compute the expectation of the bootstrap-

after-bootstrap variance estimator we start by noticing that

$$\mathrm{E}\left(\widehat{\gamma}_{b,BAG}\right) = \mathrm{E}\left(\frac{1}{B_2}\sum_{j=1}^{B_2}\frac{\mathbf{1}^T\boldsymbol{R}_{b,j}\boldsymbol{R}_b\boldsymbol{Y}}{n}\right) = \frac{\mathbf{1}^T}{nB_2}\sum_{j=1}^{B_2}\mathrm{E}\left(\boldsymbol{R}_{b,j}\boldsymbol{R}_b\boldsymbol{Y}\right).$$

Since $\boldsymbol{R}_{b,j}$, $\boldsymbol{R}_b$ and $\boldsymbol{Y}$ are mutually independent,

$$\begin{aligned}
\mathrm{E}\left(\widehat{\gamma}_{b,BAG}\right) &= \frac{\mathbf{1}^T}{nB_2}\sum_{j=1}^{B_2}\mathrm{E}\left(\boldsymbol{R}_{b,j}\right)\mathrm{E}\left(\boldsymbol{R}_b\right)\mathrm{E}(\boldsymbol{Y}) \\
&= \frac{\mathbf{1}^T}{nB_2}\sum_{j=1}^{B_2}\left(\frac{\mathbf{1}\mathbf{1}^T}{n}\right)\left(\frac{\mathbf{1}\mathbf{1}^T}{n}\right)\mathbf{1}\mu \\
&= \mu.
\end{aligned}$$

Thus, $\widehat{\gamma}_{b,BAG}$ is unbiased. Our goal is to compare $\mathrm{E}\left\{\widehat{\mathrm{Var}}_{BaB}(\widehat{\gamma}_{BAG})\right\}$ with $\mathrm{Var}(\widehat{\gamma}_{BAG})$ in (2.3).

$$\begin{aligned}
\mathrm{E}\left\{\widehat{\mathrm{Var}}_{BaB}(\widehat{\gamma}_{BAG})\right\} &= \mathrm{E}\left\{\frac{1}{B_1-1}\sum_{b=1}^{B_1}\left(\widehat{\gamma}_{b,BAG} - \overline{\widehat{\gamma}}_{\cdot,BAG}\right)^2\right\} \\
&= \frac{1}{B_1-1}\mathrm{E}\left[\sum_{b=1}^{B_1}\left\{\widehat{\gamma}_{b,BAG} - \mathrm{E}(\widehat{\gamma}_{b,BAG}) + \mathrm{E}(\widehat{\gamma}_{b,BAG}) - \overline{\widehat{\gamma}}_{\cdot,BAG}\right\}^2\right] \\
&= \frac{1}{B_1-1}\mathrm{E}\left[\sum_{b=1}^{B_1}\left\{\widehat{\gamma}_{b,BAG} - \mathrm{E}(\widehat{\gamma}_{b,BAG})\right\}^2 - B_1\left\{\overline{\widehat{\gamma}}_{\cdot,BAG} - \mathrm{E}(\widehat{\gamma}_{b,BAG})\right\}^2\right] \\
&= \frac{B_1}{B_1-1}\left\{\mathrm{Var}(\widehat{\gamma}_{1,BAG}) - \mathrm{Var}(\overline{\widehat{\gamma}}_{\cdot,BAG})\right\}, \qquad (2.4)
\end{aligned}$$

where in this last step we have used the fact that $\mathrm{E}(\widehat{\gamma}_{b,BAG}) = \mathrm{E}(\widehat{\gamma}_{\cdot,BAG})$.

It suffices to compute $\mathrm{Var}(\widehat{\gamma}_{1,BAG})$ and $\mathrm{Var}(\overline{\widehat{\gamma}}_{\cdot,BAG})$ in order to compute the expectation of the bootstrap-after-bootstrap variance estimator.

$$
\begin{aligned}
\mathrm{Var}(\widehat{\gamma}_{1,BAG}) &= \mathrm{Var}\left(\frac{1}{B_2 n}\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\boldsymbol{R}_1\boldsymbol{Y}\right) \\
&= \mathrm{Var}\left\{\mathrm{E}\left(\frac{1}{B_2 n}\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\boldsymbol{R}_1\boldsymbol{Y} \mid \boldsymbol{R}_{1,i}\right)\right\} \\
&\quad +\mathrm{E}\left\{\mathrm{Var}\left(\frac{1}{B_2 n}\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\boldsymbol{R}_1\boldsymbol{Y} \mid \boldsymbol{R}_{1,i}\right)\right\} \\
&= \mathrm{Var}\left(\frac{1}{B_2 n}\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\frac{\mathbf{1}\mathbf{1}^T}{n}\mathbf{1}\mu\right) \\
&\quad +\frac{1}{{B_2}^2 n^2}\mathrm{E}\left\{\mathrm{Var}\left(\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\boldsymbol{R}_1\boldsymbol{Y} \mid \boldsymbol{R}_{1,i}\right)\right\} \\
&= 0+\frac{1}{{B_2}^2 n^2}\mathrm{E}\left\{\left(\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\right)\mathrm{Cov}(\boldsymbol{R}_1\boldsymbol{Y})\left(\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\right)^T\right\} \\
&= \frac{\sigma^2(n-1)}{{B_2}^2 n^3}\mathrm{E}\left\{\left(\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\right)\left(\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\right)^T\right\} \\
&\quad +\frac{\sigma^2}{n^2{B_2}^2}\mathrm{E}\left\{\left(\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\right)\frac{\mathbf{1}\mathbf{1}^T}{n}\left(\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\right)^T\right\} \\
&\stackrel{def}{=} C+D,
\end{aligned}
$$

where C and D are simplified as follows.

$$
\begin{aligned}
D &= \frac{\sigma^2}{{B_2}^2 n^2}\mathrm{E}\left\{\left(\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\right)\frac{\mathbf{1}\mathbf{1}^T}{n}\left(\sum_{j=1}^{B_2}\boldsymbol{R}_{1,j}^T\mathbf{1}\right)\right\} \\
&= \frac{\sigma^2}{{B_2}^2 n^3}\mathrm{E}\left\{\left(\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\mathbf{1}\right)\left(\sum_{j=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,j}^T\mathbf{1}\right)\right\} \\
&= \frac{\sigma^2}{{B_2}^2 n^3}\mathrm{E}\left\{\left(\sum_{i=1}^{B_2}\mathbf{1}^T\mathbf{1}\right)\left(\sum_{j=1}^{B_2}\mathbf{1}^T\mathbf{1}\right)\right\} \\
&= \frac{\sigma^2}{n}.
\end{aligned}
$$

$$
\begin{aligned}
C &= \frac{(n-1)\sigma^2}{{B_2}^2 n^3}\mathrm{E}\left\{\left(\sum_{i=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{1,i}\right)\left(\sum_{i=1}^{B_2}\boldsymbol{R}_{1,i}^T\mathbf{1}\right)\right\} \\
&= \frac{(n-1)\sigma^2}{{B_2}^2 n^3}\mathbf{1}^T\mathrm{E}\left(\sum_{i=1}^{B_2}\boldsymbol{R}_{1,i}\boldsymbol{R}_{1,i}^T + \sum_{1\le i,j\le B_2, i\ne j}\boldsymbol{R}_{1,i}\boldsymbol{R}_{1,j}^T\right)\mathbf{1} \\
&= \frac{(n-1)\sigma^2}{{B_2}^2 n^3}\mathbf{1}^T\left\{\sum_{i=1}^{B_2}\left(\frac{\mathbf{1}\mathbf{1}^T}{n}+\frac{n-1}{n}\boldsymbol{I}_n\right) + B_2(B_2-1)\mathrm{E}(\boldsymbol{R}_{1,2})\mathrm{E}(\boldsymbol{R}_{1,1}^T)\right\}\mathbf{1} \\
&= \frac{(n-1)\sigma^2}{{B_2}^2 n^3}\mathbf{1}^T\left\{\sum_{i=1}^{B_2}\left(\frac{\mathbf{1}\mathbf{1}^T}{n}+\frac{n-1}{n}\boldsymbol{I}_n\right) + B_2(B_2-1)\left(\frac{\mathbf{1}\mathbf{1}^T}{n}\right)\left(\frac{\mathbf{1}\mathbf{1}^T}{n}\right)\right\}\mathbf{1} \\
&= \frac{(n-1)\sigma^2}{{B_2}^2 n^3}\mathbf{1}^T\left\{\frac{B_2\mathbf{1}\mathbf{1}^T}{n}+\frac{(n-1)B_2\boldsymbol{I}_n}{n}+\frac{B_2(B_2-1)\mathbf{1}\mathbf{1}^T}{n}\right\}\mathbf{1} \\
&= \frac{(n-1)\sigma^2}{{B_2}^2 n^3}\mathbf{1}^T\left\{\frac{{B_2}^2\mathbf{1}\mathbf{1}^T}{n}+\frac{(n-1)B_2\boldsymbol{I}_n}{n}\right\}\mathbf{1} \\
&= \frac{(n-1)\sigma^2}{{B_2}^2 n^3}\left\{n{B_2}^2+(n-1)B_2\right\} \\
&= \frac{(n-1)\sigma^2}{n^2}+\frac{(n-1)^2\sigma^2}{B_2 n^3};
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathrm{Var}(\widehat{\gamma}_{1,BAG}) = C + D &= \frac{(n-1)\sigma^2}{n^2}+\frac{(n-1)^2\sigma^2}{B_2 n^3}+\frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n}\left\{1+\frac{n-1}{n}+\frac{(n-1)^2}{B_2 n^2}\right\}
\end{aligned}
$$

Next we compute $\mathrm{Var}(\overline{\widehat{\gamma}}_{\cdot,BAG})$.

$$
\begin{aligned}
\mathrm{Var}(\overline{\widehat{\gamma}}_{\cdot,BAG}) &= \mathrm{Var}\left(\frac{\sum_{b=1}^{B_1}\widehat{\gamma}_{b,BAG}}{B_1}\right) \\
&= \mathrm{Var}\left\{\frac{1}{B_1}\sum_{b=1}^{B_1}\left(\frac{1}{B_2}\sum_{j=1}^{B_2}\frac{\mathbf{1}^T\boldsymbol{R}_{b,j}\boldsymbol{R}_b\boldsymbol{Y}}{n}\right)\right\} \\
&= \frac{1}{{B_1}^2{B_2}^2}\left[\mathrm{E}\left\{\mathrm{Var}\left(\sum_{b=1}^{B_1}\sum_{j=1}^{B_2}\frac{\mathbf{1}^T\boldsymbol{R}_{b,j}\boldsymbol{R}_b\boldsymbol{Y}}{n}\mid \boldsymbol{R}_{b,j}\right)\right\}\right. \\
&\quad \left.+\mathrm{Var}\left\{\mathrm{E}\left(\sum_{b=1}^{B_1}\sum_{j=1}^{B_2}\frac{\mathbf{1}^T\boldsymbol{R}_{b,j}\boldsymbol{R}_b\boldsymbol{Y}}{n}\mid \boldsymbol{R}_{b,j}\right)\right\}\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\overline{\widehat{\gamma}}_{\cdot,BAG}) &= \frac{1}{{B_1}^2{B_2}^2}\left[\mathrm{E}\left\{\mathrm{Var}\left(\sum_{b=1}^{B_1}\sum_{j=1}^{B_2}\frac{\mathbf{1}^T\boldsymbol{R}_{b,j}\boldsymbol{R}_b\boldsymbol{Y}}{n} \mid \boldsymbol{R}_{b,j}\right)\right\}\right. \\
&\quad \left. + \mathrm{Var}\left\{\mathrm{E}\left(\sum_{b=1}^{B_1}\sum_{j=1}^{B_2}\frac{\mathbf{1}^T\boldsymbol{R}_{b,j}}{n}\frac{\mathbf{1}\mathbf{1}^T}{n}\mathbf{1}\mu \mid \boldsymbol{R}_{b,j}\right)\right\}\right] \\
&= \frac{1}{{B_1}^2{B_2}^2}\left[\mathrm{E}\left\{\mathrm{Var}\left(\sum_{b=1}^{B_1}\sum_{j=1}^{B_2}\frac{\mathbf{1}^T\boldsymbol{R}_{b,j}\boldsymbol{R}_b\boldsymbol{Y}}{n} \mid \boldsymbol{R}_{b,j}\right)\right\}\right].
\end{aligned}
$$

Let $A_b = \sum_{j=1}^{B_2}\mathbf{1}^T\boldsymbol{R}_{b,j}/n$, we have

$$
\begin{aligned}
\mathrm{Var}(\overline{\widehat{\gamma}}_{\cdot,BAG}) &= \frac{1}{{B_2}^2{B_1}^2}\left[\mathrm{E}\left\{\mathrm{Var}\left(\sum_{b=1}^{B_1}A_b\boldsymbol{R}_b\boldsymbol{Y} \mid \boldsymbol{R}_{b,j}\right)\right\}\right] \\
&= \frac{1}{{B_2}^2{B_1}^2}\left(\mathrm{E}\left[\mathrm{Var}\left\{(A_1,\cdots,A_{B_1})\left(\boldsymbol{Y}^T\boldsymbol{R}_1^T,\cdots,\boldsymbol{Y}^T\boldsymbol{R}_{B_1}^T\right)^T\right\}\right]\right) \\
&= \frac{1}{{B_2}^2{B_1}^2}\left[\mathrm{E}\left\{(A_1,\cdots,A_{B_1})\,\mathrm{Cov}\left(\boldsymbol{Y}^T\boldsymbol{R}_1^T,\cdots,\boldsymbol{Y}^T\boldsymbol{R}_{B_1}^T\right)^T\right.\right. \\
&\quad \left.\left.(A_1,\cdots,A_{B_1})^T\right\}\right]. \qquad (2.5)
\end{aligned}
$$

Notice

$$
\begin{aligned}
&(A_1,\cdots,A_{B_1})\mathrm{Cov}(\boldsymbol{Y}^T\boldsymbol{R}_1^T,\cdots,\boldsymbol{Y}^T\boldsymbol{R}_{B_1}^T)^T(A_1,\cdots,A_{B_1})^T \\
= &(A_1,\cdots,A_{B_1})\left(\frac{\sigma^2}{n}\begin{bmatrix}\mathbf{1}\mathbf{1}^T & \cdots & \mathbf{1}\mathbf{1}^T \\ \vdots & & \vdots \\ \mathbf{1}\mathbf{1}^T & \cdots & \mathbf{1}\mathbf{1}^T\end{bmatrix}\right. \\
&\left.+\frac{(n-1)\sigma^2}{n}\begin{bmatrix}\boldsymbol{I}_n & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \boldsymbol{I}_n\end{bmatrix}\right)(A_1,\cdots,A_{B_1})^T \\
= &\frac{(n-1)\sigma^2}{n}\sum_{b=1}^{B_1}A_bA_b^T + \frac{\sigma^2}{n}\left(\sum_{b=1}^{B_1}A_b\mathbf{1}\right)\left(\sum_{b=1}^{B_1}\mathbf{1}^TA_b^T\right) \\
= &\frac{(n-1)\sigma^2}{n}\sum_{b=1}^{B_1}A_bA_b^T + \frac{\sigma^2}{n}\left(\sum_{b=1}^{B_1}A_b\mathbf{1}\right)\left(\sum_{b=1}^{B_1}A_b\mathbf{1}\right)^T, \qquad (2.6)
\end{aligned}
$$

$$\sum_{b=1}^{B_1} A_b \mathbf{1} = \sum_{b=1}^{B_1} \left(\sum_{j=1}^{B_2} \frac{\mathbf{1}^T \boldsymbol{R}_{b,j}}{n} \right) \mathbf{1} = \sum_{b=1}^{B_1} \left(\sum_{j=1}^{B_2} \frac{\mathbf{1}^T \mathbf{1}}{n} \right) = B_1 B_2, \tag{2.7}$$

and

$$\begin{aligned}
\mathrm{E}(A_1 A_1^T) &= \mathrm{E}\left\{ \left(\sum_{j=1}^{B_2} \frac{\mathbf{1}^T \boldsymbol{R}_{1,j}}{n} \right) \left(\sum_{k=1}^{B_2} \frac{\boldsymbol{R}_{1,k}^T \mathbf{1}}{n} \right) \right\} \\
&= \mathrm{E}\left(\sum_{j=1}^{B_2} \frac{\mathbf{1}^T \boldsymbol{R}_{1,j} \boldsymbol{R}_{1,j}^T \mathbf{1}}{n^2} + \sum_{1 \le j,k \le B_2, j \ne k} \frac{\mathbf{1}^T \boldsymbol{R}_{1,j} \boldsymbol{R}_{1,k}^T \mathbf{1}}{n^2} \right) \\
&= \frac{B_2}{n^2} \mathbf{1}^T \left(\frac{\mathbf{1}\mathbf{1}^T}{n} + \frac{n-1}{n} \boldsymbol{I}_n \right) \mathbf{1} + \frac{B_2(B_2-1)}{n} \\
&= \frac{B_2}{n} + \frac{B_2(n-1)}{n^2} + \frac{B_2(B_2-1)}{n} \\
&= \frac{(n-1)B_2 + n{B_2}^2}{n^2}.
\end{aligned} \tag{2.8}$$

Substitute (2.6), (2.7) and (2.8) into (2.5) to obtain

$$\begin{aligned}
\mathrm{Var}(\overline{\widehat{\gamma}}_{\cdot,BAG}) &= \frac{1}{{B_2}^2 {B_1}^2} \left\{ \frac{(n-1)\sigma^2}{n} \sum_{b=1}^{B_1} \mathrm{E}(A_b A_b^T) + \frac{\sigma^2}{n} \left(\sum_{b=1}^{B_1} A_b \mathbf{1} \right) \left(\sum_{b=1}^{B_1} A_b \mathbf{1} \right)^T \right\} \\
&= \frac{1}{{B_1}^2 {B_2}^2} \left[\frac{(n-1)\sigma^2}{n} B_1 \left\{ \frac{(n-1)B_2 + n{B_2}^2}{n^2} \right\} + \frac{\sigma^2 {B_1}^2 {B_2}^2}{n} \right] \\
&= \frac{(n-1)\sigma^2 B_1 \left(n - 1 + nB_2\right) B_2}{{B_1}^2 {B_2}^2 n^3} + \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n} + \frac{(n-1)(nB_2 + n - 1)\sigma^2}{B_2 B_1 n^3} \\
&= \frac{\sigma^2}{n} \left\{ 1 + \frac{(n-1)(nB_2 + n - 1)}{B_1 B_2 n^2} \right\}.
\end{aligned}$$

Putting all the results together into (2.4), we have

$$\begin{aligned}
\mathrm{E}\left\{ \widehat{\mathrm{Var}}_{BaB}(\widehat{\gamma}_{BAG}) \right\} &= \frac{B_1}{B_1 - 1} \left\{ \mathrm{Var}(\widehat{\gamma}_{1,BAG}) - \mathrm{Var}(\overline{\widehat{\gamma}}_{\cdot,BAG}) \right\} \\
&= \frac{B_1}{B_1 - 1} \left\{ \frac{(n-1)\sigma^2}{n^2} + \frac{(n-1)^2\sigma^2}{B_2 n^3} + \frac{\sigma^2}{n} - \frac{\sigma^2}{n} \right. \\
&\qquad \left. - \frac{(n-1)(nB_2 + n - 1)\sigma^2}{B_2 B_1 n^3} \right\}
\end{aligned}$$

$$
\begin{aligned}
\mathrm{E}\left\{\widehat{\mathrm{Var}}_{BaB}(\widehat{\gamma}_{BAG})\right\} &= \frac{B_1}{B_1-1}\left(\frac{\sigma^2}{n^2}\right)\left\{n-1+\frac{(n-1)^2}{nB_2}\left(1-\frac{1}{B_1}\right)-\frac{n-1}{B_1}\right\} \\
&\approx \frac{\sigma^2}{n}+\frac{\sigma^2}{n\,B_2}-\frac{\sigma^2}{n\,B_1} \\
&= \frac{\sigma^2}{n}\left(1+\frac{1}{B_2}-\frac{1}{B_1}\right) \qquad (2.9)
\end{aligned}
$$

This expectation value should be compared to the true variance from (2.3),

$$
\mathrm{Var}(\widehat{\gamma}_{BAG}) = \frac{\sigma^2}{n}\left(1+\frac{n-1}{nB}\right).
$$

In practice B_1 and B_2 may be much smaller than B used in computing $\widehat{\gamma}_{BAG}$. In the simulation of Chapter 4, $B = 300$, $B_1 = 30$, $B_2 = 10$, and $n = 150$, so that (2.9) is $1.067\sigma^2/n$ and (2.3) is $1.004\sigma^2/n$. Thus, using smaller B_1 and B_2 result in the bootstrap-after-bootstrap being slightly biased upwards. Also, we can adjust the bootstrap-after-bootstrap variance estimator for using smaller B_1 and B_2 by a factor $(1 + 1/B)/(1 + 1/B_2 - 1/B_1)$ so that the modified bootstrap-after-bootstrap variance estimator is approximately unbiased in the sample mean case.

Chapter 3

The Parallel Bootstrap Variance Estimators

When a variable selection procedure is used to build a statistical model, estimating the variance of bagging predictors is a challenging problem as discussed in Chapter 2. In practice, data analysts sometime ignore the selection process, and estimate the variance of estimated coefficients and predictions based solely on the chosen model. This simplistic approach leads to underestimation of variances.

Our goal is to derive a variance estimation method that is suitable for estimators that combines bootstrap averaging (bagging) with variable selection procedures. In this chapter we introduce a new variance estimation method, the *parallel bootstrap*, for general bagging estimators and anticipate that the method will work on bagging predictions from variable selection procedures. However, in this chapter we present only the basic approach and show how it works in two simple situations: the sample mean of an i.i.d. sample, and predictions from a linear model estimated without variable selection. In Chapter 4 we show by simulation that the parallel bootstrap works reasonably well for bagging variable selection procedures.

3.1 Motivation of Parallel Bootstrap Variance Estimators

Let $\widehat{\gamma}_b$ be the basic estimator calculated from the bth bootstrap sample. The bagging estimator based on B bootstrap samples is

$$\widehat{\gamma}_{BAG} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\gamma}_b. \tag{3.1}$$

The motivation for $\widehat{\gamma}_{BAG}$ is the bagging predicted response $\widehat{\beta}_{0,BAG} + \boldsymbol{x}^T\widehat{\boldsymbol{\beta}}_{BAG}$ in the linear regression setting. However, we derive the parallel bootstrap variance estimator here for the general case of a scalar estimator, $\widehat{\gamma}_{BAG}$, based on data $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$. In regression settings, Y_i would actually consist of the ith response and associated predictor variables. The bootstrap versions of $\widehat{\gamma}$, given by $\widehat{\gamma}_b$, are based on bootstrap samples $\boldsymbol{Y}^{(b)} = \boldsymbol{R}_b\boldsymbol{Y}$, where $\boldsymbol{R}_b$ are i.i.d. bootstrap selection matrices, for $b = 1, \cdots, B$. In regression settings, this type of nonparametric bootstrap is called the random-pair bootstrap as explained in Chapter 1.

To estimate the variance of $\widehat{\gamma}_{BAG}$, we shall also need to draw level-2 bootstrap samples from each of the B level-1 bootstrap samples $\boldsymbol{R}_b\boldsymbol{Y}$. The bootstrap-after-bootstrap requires multiple level-2 bootstrap samples. A computational advantage of the proposed parallel bootstrap is that only one level-2 bootstrap sample is required to be drawn from each level-1 bootstrap sample.

The exact bagging estimator is actually the limit as the bootstrap sample size $B \to \infty$,

$$\widehat{\gamma}_\infty = \lim_{B\to\infty}\widehat{\gamma}_{BAG} = \mathrm{E}\left(\widehat{\gamma}_1 \mid \boldsymbol{Y}\right),$$

where by $\mathrm{E}(\cdot \mid \boldsymbol{Y})$ we mean the expected value in the bootstrap world conditional on the

original data $\boldsymbol{Y}$. The usual iterated variance calculation yields

$$\begin{aligned}
\mathrm{Var}(\widehat{\gamma}_{BAG}) &= \mathrm{Var}\left\{\mathrm{E}(\widehat{\gamma}_{BAG} \mid \boldsymbol{Y})\right\} + \mathrm{E}\left\{\mathrm{Var}(\widehat{\gamma}_{BAG} \mid \boldsymbol{Y})\right\} && (3.2)\\
&= \mathrm{Var}\left\{\mathrm{E}(\widehat{\gamma}_1 \mid \boldsymbol{Y})\right\} + \mathrm{E}\left\{\frac{1}{B}\mathrm{Var}(\widehat{\gamma}_1 \mid \boldsymbol{Y})\right\} && (3.3)\\
&= \mathrm{Var}(\widehat{\gamma}_\infty) + \frac{1}{B}\mathrm{E}\left\{\mathrm{Var}(\widehat{\gamma}_1 \mid \boldsymbol{Y})\right\}. && (3.4)
\end{aligned}$$

The step from (3.2) to (3.3) uses the fact that $\widehat{\gamma}_{BAG}$ is a mean. Letting $B = 1$, (3.4) becomes

$$\mathrm{Var}(\widehat{\gamma}_1) = \mathrm{Var}(\widehat{\gamma}_\infty) + \mathrm{E}\left\{\mathrm{Var}(\widehat{\gamma}_1 \mid \boldsymbol{Y})\right\}. \tag{3.5}$$

Subtracting (3.4) from (3.5) gives

$$\mathrm{Var}(\widehat{\gamma}_1) - \mathrm{Var}(\widehat{\gamma}_{BAG}) = \left(1 - \frac{1}{B}\right)\mathrm{E}\left\{\mathrm{Var}(\widehat{\gamma}_1 \mid \boldsymbol{Y})\right\},$$

and rearranging gives

$$\mathrm{Var}(\widehat{\gamma}_{BAG}) = \mathrm{Var}(\widehat{\gamma}_1) - \left(1 - \frac{1}{B}\right)\mathrm{E}\left\{\mathrm{Var}(\widehat{\gamma}_1 \mid \boldsymbol{Y})\right\}. \tag{3.6}$$

Equation (3.6) suggests the *difference-of-variances* (DoV) estimator for $\widehat{\gamma}_{BAG}$,

$$\widehat{V}_{DoV} = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_{b,1} - \overline{\widehat{\gamma}}_{\cdot,1})^2 - \left(1 - \frac{1}{B}\right)\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_b - \overline{\widehat{\gamma}}_{\cdot})^2, \tag{3.7}$$

where $\widehat{\gamma}_{b,1}$ is the statistic calculated from one level-2 bootstrap data set selected from the bth level-1 bootstrap data set, and $\widehat{\gamma}_b$ is calculated from the bth level-1 bootstrap sample. The subscript DoV indicates this variance estimator defined in (3.7) is a difference of two sample variances. Now we explain why (3.7) is reasonable.

It is clear that

$$\begin{aligned} \mathrm{E}\left\{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_b-\overline{\widehat{\gamma}}_{\cdot})^2\right\} &= \mathrm{E}\left[\mathrm{E}\left\{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_b-\overline{\widehat{\gamma}}_{\cdot})^2 \mid \boldsymbol{Y}\right\}\right] \\ &= \mathrm{E}\left\{\mathrm{Var}(\widehat{\gamma}_1 \mid \boldsymbol{Y})\right\}. \end{aligned}$$

Thus $(B-1)^{-1}\sum_{b=1}^{B}(\widehat{\gamma}_b-\overline{\widehat{\gamma}}_{\cdot})^2$ is an unbiased estimator of $\mathrm{E}\left\{\mathrm{Var}(\widehat{\gamma}_1 \mid \boldsymbol{Y})\right\}$, which appears in the second term of (3.6).

To estimate $\mathrm{Var}(\widehat{\gamma}_1)$ of the first term in (3.6), we imitate the idea of the usual bootstrap variance estimator. That is, generate B level-1 bootstrap samples. From each of these samples, draw one level-2 sample and compute the statistics resulting in $\widehat{\gamma}_{1,1},\ldots,\widehat{\gamma}_{B,1}$. Estimate $\mathrm{Var}(\widehat{\gamma}_1)$ using

$$\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_{b,1}-\overline{\widehat{\gamma}}_{\cdot,1})^2.$$

Putting both estimates into (3.6) results in the variance estimator (3.7).

To understand why $(B-1)^{-1}\sum_{b=1}^{B}(\widehat{\gamma}_{b,1}-\overline{\widehat{\gamma}}_{\cdot,1})^2$ estimates $\mathrm{Var}(\widehat{\gamma}_1)$, we make a comparison with Monte Carlo simulation to get a variance estimate for $\mathrm{Var}(\widehat{\gamma}_1)$. The first level of this two-stage process is Monte Carlo simulation from a known distribution of population (RS=random sampling), and the second stage is a bootstrap step (BSS=bootstrap sampling).

$$\text{Population} \xrightarrow{RS} \begin{cases} \boldsymbol{Y}_1 \xrightarrow{BSS} \boldsymbol{Y}^{(1)} \longrightarrow \widehat{\gamma}_1 = T(\boldsymbol{Y}^{(1)}) \\ \boldsymbol{Y}_2 \xrightarrow{BSS} \boldsymbol{Y}^{(2)} \longrightarrow \widehat{\gamma}_2 = T(\boldsymbol{Y}^{(2)}) \\ \qquad\qquad \cdots \\ \boldsymbol{Y}_k \xrightarrow{BSS} \boldsymbol{Y}^{(k)} \longrightarrow \widehat{\gamma}_k = T(\boldsymbol{Y}^{(k)}) \end{cases}$$

The sample variance of $\widehat{\gamma}_1,\cdots,\widehat{\gamma}_k$ estimates $\mathrm{Var}(\widehat{\gamma}_1)$ unbiasedly and converges in probability to $\mathrm{Var}(\widehat{\gamma}_1)$ as $k\to\infty$.

Of course, we do not really know the population. So the bootstrap uses the sample data $\boldsymbol{Y}$ as a population and then two stages of bootstrap sampling to mimic the above sampling where the population is known. Thus, at the first stage, k level-1 bootstrap

samples are generated. Then from the jth level-1 bootstrap sample, a single level-2 bootstrap sample is generated and $\widehat{\gamma}_{j,1} = T(\boldsymbol{Y}^{(j,1)})$ is computed, $j = 1, \cdots, k$.

$$\boldsymbol{Y} \xrightarrow{BSS} \begin{cases} \boldsymbol{Y}^{(1)} \xrightarrow{BSS} \boldsymbol{Y}^{(1,1)} \longrightarrow \widehat{\gamma}_{1,1} = T(\boldsymbol{Y}^{(1,1)}) \\ \boldsymbol{Y}^{(2)} \xrightarrow{BSS} \boldsymbol{Y}^{(2,1)} \longrightarrow \widehat{\gamma}_{2,1} = T(\boldsymbol{Y}^{(2,1)}) \\ \qquad\qquad \cdots \\ \boldsymbol{Y}^{(k)} \xrightarrow{BSS} \boldsymbol{Y}^{(k,1)} \longrightarrow \widehat{\gamma}_{k,1} = T(\boldsymbol{Y}^{(k,1)}) \end{cases}$$

The sample variance of $\widehat{\gamma}_{1,1}, \cdots, \widehat{\gamma}_{k,1}$ estimates $\mathrm{Var}(\widehat{\gamma}_{1,1})$, which in turn estimates $\mathrm{Var}(\widehat{\gamma}_1)$.

The sample variance of $\widehat{\gamma}_{1,1}, \cdots, \widehat{\gamma}_{k,1}$ estimates the unconditional variance of $\widehat{\gamma}_{1,1}$ in the bootstrap world. Note that $\widehat{\gamma}_{1,1}$ plays the role of $\widehat{\gamma}_1$ in the previous display. This sample variance is unbiased and consistent as $k \to \infty$ in the bootstrap world. However, even as $k \to \infty$, this sample variance is actually a function of the data $\boldsymbol{Y}$ and only converges to $\mathrm{Var}(\widehat{\gamma}_1)$ as $n \to \infty$.

When the first sample variance in (3.7) is less than the second one, a negative variance estimate results and thus the DoV estimate is not range preserving. To overcome this problem, we propose a second version of the parallel bootstrap variance estimator for $\widehat{\gamma}_{BAG}$, the *variance-of-difference* (VoD),

$$\widehat{V}_{VoD} = \frac{\sum_{b=1}^{B} \{(\widehat{\gamma}_{b,1} - \widehat{\gamma}_b) - (\overline{\widehat{\gamma}}_{\cdot,1} - \overline{\widehat{\gamma}}_{\cdot})\}^2}{B-1}, \tag{3.8}$$

where the subscript VoD means this variance estimator is a sample variance of difference.

These two bootstrap variance estimators have small bias when the basic estimator $\widehat{\gamma}$ is the sample mean (which will be proved in the next section). Their biased-adjusted versions are

$$\widetilde{V}_{DoV} = \frac{n}{n-1}\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_{b,1} - \overline{\widehat{\gamma}}_{\cdot,1})^2 - \left(1 - \frac{1}{B}\right)\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_b - \overline{\widehat{\gamma}}_{\cdot})^2$$

and

$$\widetilde{V}_{VoD} = \left(1 + \frac{n-1}{nB}\right)\left(\frac{n}{n-1}\right)^2 \frac{1}{B-1}\sum_{b=1}^{B}\left\{(\widehat{\gamma}_{b,1} - \widehat{\gamma}_b) - \left(\overline{\widehat{\gamma}}_{\cdot,1} - \overline{\widehat{\gamma}}_{\cdot}\right)\right\}^2.$$

It is very difficult to describe the variable selection procedures analytically and thus it is hard to check the properties of our parallel bootstrap variance estimator. Thus we investigate these variance estimators in two simplified situations to gain some insights.

3.2 The Sample Mean Case

In this section, we study the performance of the parallel bootstrap variance estimators DoV and VoD when the basic estimator $\widehat{\gamma}$ is the mean of an i.i.d. sample $Y_1, \cdots, Y_n$, $\widehat{\gamma} = n^{-1}\sum_{i=1}^{n} Y_i$. This corresponds to the situation where there is no predictor variable in the model, and the predicted response is estimated by the sample mean.

Recall that a **BSM** matrix $\boldsymbol{R}$ is defined as

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{r}_1^T \\ \vdots \\ \boldsymbol{r}_n^T \end{pmatrix},$$

where $\boldsymbol{r}_1, \cdots, \boldsymbol{r}_1$ are i.i.d. multinomial $(1; 1/n, \cdots, 1/n)$ random vectors. Note that $\boldsymbol{R}$ is an $n \times n$ matrix. Using the **BSM** matrix notation, we can describe the level-1 and level-2 bootstrap samples drawn by the parallel bootstrap variance estimator as follows.

We draw level-1 bootstrap samples by random-pair bootstrapping. For the bth level-1 bootstrap sample, an element $Y_i^{(b)}$ is randomly drawn from the original sample $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ so that each member has an equal chance of being selected. Repeating this random sampling with replacement n times generates a bootstrap sample $\boldsymbol{Y}^{(b)} = (Y_1^{(b)}, \cdots, Y_n^{(b)})^T$. The bootstrap sample vector $(Y_1^{(b)}, \cdots, Y_n^{(b)})^T$ can be written as a

multiplication of an $n \times n$ matrix $\boldsymbol{R}_b$ with the $n \times 1$ vector $(Y_1, \cdots, Y_n)^T$. That is,

$$\begin{pmatrix} Y_1^{(b)} \\ \vdots \\ Y_n^{(b)} \end{pmatrix} = \boldsymbol{R}_b \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

We use the residual-based bootstrap to generate level-2 bootstrap samples. Actually, the level-2 samples generated by the residual-based bootstrap have the same distribution as those generated by the rand-pair bootstrap in the sample mean case. Using the the bth level-1 bootstrap sample $(Y_1^{(b)}, \cdots, Y_n^{(b)})^T$, we predict the response vector as $\mathbf{1}\mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y}/n$ and compute the residual vector $\boldsymbol{R}_b\boldsymbol{Y} - \mathbf{1}\mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y}/n$. Then treat the residual vector as an original i.i.d. sample and draw a level-1 bootstrap sample from it. Finally, generate a level-2 bootstrap sample $(Y_1^{(b,1)}, \cdots, Y_n^{(b,1)})^T$ as the sum of the predicted vector and the level-1 bootstrap sample of the residual vector.

Using the notation of **BSM** matrix, the level-2 bootstrap sample can be written as

$$\begin{aligned} \begin{pmatrix} Y_1^{(b,1)} \\ \vdots \\ Y_n^{(b,1)} \end{pmatrix} &= \frac{\mathbf{1}\mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y}}{n} + \boldsymbol{R}_{b,1}\left(\boldsymbol{R}_b\boldsymbol{Y} - \frac{\mathbf{1}\mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y}}{n}\right) \\ &= \frac{\mathbf{1}\mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y}}{n} + \boldsymbol{R}_{b,1}\boldsymbol{R}_b\boldsymbol{Y} - \frac{\boldsymbol{R}_{b,1}\mathbf{1}\mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y}}{n} \\ &= \frac{\mathbf{1}\mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y}}{n} + \boldsymbol{R}_{b,1}\boldsymbol{R}_b\boldsymbol{Y} - \frac{\mathbf{1}\mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y}}{n} \\ &= \boldsymbol{R}_{b,1}\boldsymbol{R}_b\boldsymbol{Y}, \end{aligned}$$

where $\boldsymbol{R}_{b,1}$, $\boldsymbol{R}_b$ are two independent **BSM** matrices. The next to last step uses the fact that $\boldsymbol{R}_{b,1}\mathbf{1} = \mathbf{1}$. Thus, in the sample mean case, the level-2 residual-based bootstrapping reduces to the usual level-2 nonparametric bootstrap.

3.2.1 A Theorem in the Sample Mean Case

In this section, we give a theorem on the bias of the parallel bootstrap variance estimators when the basic estimator $\widehat{\gamma}$ is the sample mean.

Theorem 1 (Bagging results for the sample mean). *Suppose* $\boldsymbol{Y}_{n\times 1} = (Y_1, \cdots, Y_n)^T$, *where* $Y_1, \cdots, Y_n$ *are i.i.d. with* $E(Y) = \mu$ *and* $Var(Y) = \sigma^2$. *Assume that* $\boldsymbol{R}_{b,1}, \boldsymbol{R}_b$, $b = 1, \cdots, B$, *are i.i.d.* $n \times n$ *bootstrap selection matrices. Let* $\widehat{\gamma} = \widehat{\gamma}(\boldsymbol{Y}) = \mathbf{1}^T\boldsymbol{Y}/n = \overline{\boldsymbol{Y}}$. *We have*

$$\widehat{\gamma}_b = \frac{1}{n}\mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y} \quad \textit{and} \quad \widehat{\gamma}_{b,1} = \frac{1}{n}\mathbf{1}^T\boldsymbol{R}_{b,1}\boldsymbol{R}_b\boldsymbol{Y}, \quad b = 1, \cdots, B.$$

For $\widehat{\gamma}_{BAG} = B^{-1}\sum_{i=1}^{B}\widehat{\gamma}_b$, *we have*

(I) $E(\widehat{\gamma}_{BAG}) = \mu$,

(II) $Var(\widehat{\gamma}_{BAG}) = \dfrac{\sigma^2}{n}\left(1 + \dfrac{n-1}{nB}\right)$,

(III) $E(\widetilde{V}_{DoV}) = Var(\widehat{\gamma}_{BAG})$,

(IV) $E(\widetilde{V}_{VoD}) = Var(\widehat{\gamma}_{BAG})$.

Proof. To prove **(I)**, note that by **(F1)** and independence of $\boldsymbol{R}_b$ and $\boldsymbol{Y}$

$$\begin{aligned}
\mathrm{E}(\widehat{\gamma}_b) &= \mathrm{E}(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y}) \\
&= \frac{1}{n}\mathbf{1}^T\mathrm{E}(\boldsymbol{R}_{b,1})\mathrm{E}(\boldsymbol{Y}) \\
&= \frac{1}{n}\mathbf{1}^T\left(\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{1}\mu \\
&= \mu.
\end{aligned}$$

Thus, $\mathrm{E}(\widehat{\gamma}_{BAG}) = \mathrm{E}(\widehat{\gamma}_b) = \mu$.

To prove **(II)**, start by noticing that $\mathrm{Var}(\widehat{\gamma}_{BAG}) = \mathrm{Var}(\widehat{\gamma}_1) - \left(1 - \dfrac{1}{B}\right)\mathrm{E}\{\mathrm{Var}(\widehat{\gamma}_1|\boldsymbol{Y})\}$. By the fact **(F9)** and **(F10)**, we have

$$\mathrm{E}\left\{\mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}\boldsymbol{Y}|\boldsymbol{Y}\right)\right\} = \frac{\sigma^2}{n}\left(\frac{n-1}{n}\right)$$

and

$$\mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}\boldsymbol{Y}\right) = \frac{\sigma^2}{n}\left(\frac{2n-1}{n}\right).$$

Thus,

$$\mathrm{Var}(\widehat{\gamma}_{BAG}) = \frac{\sigma^2}{n}\left(\frac{2n-1}{n}\right) - \left(1-\frac{1}{B}\right)\frac{\sigma^2}{n}\left(\frac{n-1}{n}\right) = \frac{\sigma^2}{n}\left(1+\frac{n-1}{nB}\right).$$

To prove **(III)**, start by noticing that $\mathrm{E}\{\mathrm{Var}(\widehat{\gamma}_1|\boldsymbol{Y})\} = \mathrm{E}\left\{(B-1)^{-1}\sum_{b=1}^{B}(\widehat{\gamma}_b - \overline{\widehat{\gamma}}_{\cdot})^2\right\}$. Since $\mathrm{Var}(\widehat{\gamma}_{BAG}) = \mathrm{Var}(\widehat{\gamma}_1) - \left(1-\frac{1}{B}\right)\mathrm{E}\{\mathrm{Var}(\widehat{\gamma}_1|\boldsymbol{Y})\}$, it suffices to show that

$$\frac{n}{n-1}\mathrm{E}\left\{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_{b,1} - \overline{\widehat{\gamma}}_{\cdot,1})^2\right\} = \mathrm{Var}(\widehat{\gamma}_1).$$

Define $S^2_{\widehat{\gamma}_{1,1}} = \dfrac{1}{B-1}\displaystyle\sum_{b=1}^{B}(\widehat{\gamma}_{b,1} - \overline{\widehat{\gamma}}_{\cdot,1})^2$, and thus we want to prove

$$\frac{n}{n-1}\mathrm{E}\left(S^2_{\widehat{\gamma}_{1,1}}\right) = \mathrm{Var}(\widehat{\gamma}_1). \tag{3.9}$$

Note that

$$\mathrm{E}\left(S^2_{\widehat{\gamma}_{1,1}}|\boldsymbol{Y}\right) = \mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}_{b,1}\boldsymbol{R}_b\boldsymbol{Y}|\boldsymbol{Y}\right),$$

and thus,

$$\mathrm{E}\left(S^2_{\widehat{\gamma}_{1,1}}\right) = \mathrm{E}\left\{\mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}_{b,1}\boldsymbol{R}_b\boldsymbol{Y}|\boldsymbol{Y}\right)\right\}.$$

Since $\boldsymbol{R}_{b,1}$ and $\boldsymbol{R}_b$ are independent bootstrap selection matrices, we replace them with $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ for notational convenience and calculate $\tau^2 \overset{\text{def}}{=} \mathrm{E}\left\{\mathrm{Var}\left(\mathbf{1}^T\boldsymbol{R}_1\boldsymbol{R}_2\boldsymbol{Y}/n|\boldsymbol{Y}\right)\right\}$.

Note that

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}_1\boldsymbol{R}_2\boldsymbol{Y}\right) &= \tau^2 + \mathrm{Var}\left\{\mathrm{E}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}_1\boldsymbol{R}_2\boldsymbol{Y}|\boldsymbol{Y}\right)\right\} \\
&= \tau^2 + \mathrm{Var}\left\{\frac{1}{n}\mathbf{1}^T\mathrm{E}(\boldsymbol{R}_1)\mathrm{E}(\boldsymbol{R}_2)\boldsymbol{Y}\right\}, \boldsymbol{R}_1 \text{ and } \boldsymbol{R}_2 \text{ are independent} \\
&= \tau^2 + \mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\frac{\mathbf{1}\mathbf{1}^T}{n}\frac{\mathbf{1}\mathbf{1}^T}{n}\boldsymbol{Y}\right), \text{ by } \mathbf{(F1)} \\
&= \tau^2 + \mathrm{Var}(\overline{\boldsymbol{Y}}) \\
&= \tau^2 + \frac{\sigma^2}{n}.
\end{aligned}$$

Thus,

$$\tau^2 = \mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}_1\boldsymbol{R}_2\boldsymbol{Y}\right) - \frac{\sigma^2}{n} = \theta^2 - \frac{\sigma^2}{n}, \text{ where } \theta^2 = \mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}_1\boldsymbol{R}_2\boldsymbol{Y}\right). \quad (3.10)$$

Using the variance decomposition, we have

$$\begin{aligned}
\theta^2 &= \mathrm{E}\left\{\mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}_1\boldsymbol{R}_2\boldsymbol{Y}|\boldsymbol{R}_1,\boldsymbol{R}_2\right)\right\} + \mathrm{Var}\left\{\mathrm{E}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}_1\boldsymbol{R}_2\boldsymbol{Y}|\boldsymbol{R}_1,\boldsymbol{R}_2\right)\right\} \\
&= \mathrm{E}\left\{\mathrm{Var}\left(\frac{1}{n}\mathbf{1}^T\boldsymbol{R}_1\boldsymbol{R}_2\boldsymbol{Y}|\boldsymbol{R}_1,\boldsymbol{R}_2\right)\right\} + \mathrm{Var}(\mu) \\
&= \mathrm{E}\left\{\frac{1}{n^2}\mathbf{1}^T\boldsymbol{R}_1\boldsymbol{R}_2(\sigma^2\boldsymbol{I})\boldsymbol{R}_2^T\boldsymbol{R}_1^T\mathbf{1}\right\} \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T\left[\mathrm{E}\left\{\boldsymbol{R}_1\mathrm{E}(\boldsymbol{R}_2\boldsymbol{R}_2^T)\boldsymbol{R}_1^T\right\}\right]\mathbf{1}, \text{ by the independence of } \boldsymbol{R}_1 \text{ and } \boldsymbol{R}_2 \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T\left[\mathrm{E}\left\{\boldsymbol{R}_1\left(\frac{n-1}{n}\boldsymbol{I} + \boldsymbol{P}_1\right)\boldsymbol{R}_1^T\right\}\right]\mathbf{1}, \text{ by } \mathbf{(F2)} \text{ and } \boldsymbol{P}_1 = \frac{1}{n}\mathbf{1}\mathbf{1}^T \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T\left\{\frac{n-1}{n}\mathrm{E}(\boldsymbol{R}_1\boldsymbol{R}_1^T) + \mathrm{E}(\boldsymbol{R}_1\boldsymbol{P}_1\boldsymbol{R}_1^T)\right\}\mathbf{1} \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T\left\{\frac{n-1}{n}\left(\frac{n-1}{n}\boldsymbol{I} + \boldsymbol{P}_1\right) + \frac{1}{n}\mathbf{1}\mathbf{1}^T\right\}\mathbf{1}, \text{ by } \mathbf{(F2)} \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T\left\{\left(\frac{n-1}{n}\right)^2\boldsymbol{I} + \frac{n-1}{n}\boldsymbol{P}_1 + \boldsymbol{P}_1\right\}\mathbf{1} \\
&= \frac{\sigma^2}{n}\left\{\left(\frac{n-1}{n}\right)^2 + \left(\frac{2n-1}{n}\right)\right\}. \quad (3.11)
\end{aligned}$$

Substituting equation (3.11) into (3.10), we obtain

$$\begin{aligned}\tau^2 &= \frac{\sigma^2}{n}\left\{\left(\frac{n-1}{n}\right)^2 + \frac{2n-1}{n} - 1\right\} \\ &= \frac{\sigma^2}{n}\left\{\left(\frac{2n-1}{n}\right)\left(\frac{n-1}{n}\right)\right\}.\end{aligned}$$

Therefore, $\mathrm{E}\left(S^2_{\widehat{\gamma}_{1,1}}\right) = \frac{\sigma^2}{n}\left\{\left(\frac{2n-1}{n}\right)\left(\frac{n-1}{n}\right)\right\}$. It follows that

$$\frac{n}{n-1}\mathrm{E}\left(S^2_{\widehat{\gamma}_{1,1}}\right) = \frac{\sigma^2}{n}\left(\frac{2n-1}{n}\right).$$

From **(F10)**, equation (3.9) holds true and we have shown that $\mathrm{E}(\widetilde{V}_{DoV}) = \mathrm{Var}(\widehat{\gamma}_{BAG})$.

To prove **(IV)**, start by noticing

$$\widehat{\gamma}_{b,1} - \widehat{\gamma}_b = \frac{1}{n}\mathbf{1}^T\boldsymbol{R}_{b,1}\boldsymbol{R}_b\boldsymbol{Y} - \frac{1}{n}\mathbf{1}^T\boldsymbol{R}_b\boldsymbol{Y} = \frac{1}{n}\mathbf{1}^T(\boldsymbol{R}_{b,1} - \boldsymbol{I})\boldsymbol{R}_b\boldsymbol{Y} \stackrel{\text{def}}{=} D_b.$$

Also notice that $\mathrm{E}(\widehat{V}_{VoD}|\boldsymbol{Y}) = \mathrm{Var}(D_1|\boldsymbol{Y})$. Thus, $\mathrm{E}(\widehat{V}_{VoD}) = \mathrm{E}\left\{\mathrm{Var}(D_1|\boldsymbol{Y})\right\}$.

Using the variance decomposition $\mathrm{Var}(D_1) = \mathrm{E}\left\{\mathrm{Var}(D_1|\boldsymbol{Y})\right\} + \mathrm{Var}\left\{\mathrm{E}(D_1|\boldsymbol{Y})\right\}$, we see that

$$\mathrm{E}(\widehat{V}_{VoD}) = \mathrm{Var}(D_1) - \mathrm{Var}\{\mathrm{E}(D_1|\boldsymbol{Y})\}.$$

By the independence of $\boldsymbol{R}_{1,1}$ and $\boldsymbol{R}_1$,

$$\begin{aligned}E(D_1|\boldsymbol{Y}) &= \frac{1}{n}\mathbf{1}^T\{\mathrm{E}(\boldsymbol{R}_{1,1}) - \boldsymbol{I}\}\mathrm{E}(\boldsymbol{R}_1)\boldsymbol{Y} \\ &= \frac{1}{n}\mathbf{1}^T(\boldsymbol{P}_1 - \boldsymbol{I})\boldsymbol{P}_1\boldsymbol{Y},\end{aligned}$$

where $\boldsymbol{P}_1 = \mathbf{1}\mathbf{1}^T/n$. But because $\boldsymbol{P}_1$ is a projection matrix and $\boldsymbol{P}_1^2 = \boldsymbol{P}_1$ (idempotent), it follows that

$$\mathrm{E}(D_1|\boldsymbol{Y}) = 0 \text{ and thus } \mathrm{E}(\widehat{V}_{VoD}) = \mathrm{Var}(D_1).$$

Using the variance decomposition, we have

$$\mathrm{Var}(D_1) = \mathrm{E}\left\{\mathrm{Var}(D_1|\boldsymbol{R}_{1,1}, \boldsymbol{R}_1)\right\} + \mathrm{Var}\left\{\mathrm{E}(D_1|\boldsymbol{R}_{1,1}, \boldsymbol{R}_1)\right\}.$$

But

$$\begin{aligned}
\mathrm{E}(D_1|\boldsymbol{R}_{1,1}, \boldsymbol{R}_1) &= \frac{1}{n}\mathbf{1}^T(\boldsymbol{R}_{1,1} - \boldsymbol{I})\boldsymbol{R}_1\mathbf{1}\mu \\
&= \frac{1}{n}\mathbf{1}^T(\boldsymbol{R}_{1,1} - \boldsymbol{I})\mathbf{1}\mu \\
&= \frac{1}{n}\mathbf{1}^T(\mathbf{1} - \mathbf{1}) \\
&= 0,
\end{aligned}$$

and thus

$$\mathrm{Var}(D_1) = \mathrm{E}\left\{\mathrm{Var}(D_1|\boldsymbol{R}_{1,1}\boldsymbol{R}_1)\right\}. \tag{3.12}$$

But

$$\begin{aligned}
\mathrm{Var}(D_1|\boldsymbol{R}_{1,1}, \boldsymbol{R}_1) &= \mathrm{Var}\left\{\frac{1}{n}\mathbf{1}^T(\boldsymbol{R}_{1,1} - \boldsymbol{I})\boldsymbol{R}_1\boldsymbol{Y}|\boldsymbol{R}_{1,1}, \boldsymbol{R}_1\right\} \\
&= \frac{1}{n^2}\mathbf{1}^T(\boldsymbol{R}_{1,1} - \boldsymbol{I})\boldsymbol{R}_1(\sigma^2\boldsymbol{I})\boldsymbol{R}_1^T(\boldsymbol{R}_{1,1} - \boldsymbol{I})^T\mathbf{1} \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T(\boldsymbol{R}_{1,1} - \boldsymbol{I})\boldsymbol{R}_1\boldsymbol{R}_1^T(\boldsymbol{R}_{1,1} - \boldsymbol{I})^T\mathbf{1}. \qquad (3.13)
\end{aligned}$$

Plugging (3.13) into (3.12) to obtain

$$\begin{aligned}
\mathrm{Var}(D_1) &= \frac{\sigma^2}{n^2}\mathbf{1}^T\mathrm{E}\left\{(\boldsymbol{R}_{1,1} - \boldsymbol{I})\mathrm{E}(\boldsymbol{R}_1\boldsymbol{R}_1^T)(\boldsymbol{R}_{1,1} - \boldsymbol{I})^T\right\}\mathbf{1}, \text{ by the iterated expectation} \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T\mathrm{E}\left\{(\boldsymbol{R}_{1,1} - \boldsymbol{I})\left(\frac{n-1}{n}\boldsymbol{I} + \boldsymbol{P}_1\right)(\boldsymbol{R}_{1,1} - \boldsymbol{I})^T\right\}\mathbf{1}, \text{ by } \mathbf{(F2)} \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T\left[\left(\frac{n-1}{n}\right)\mathrm{E}\left\{(\boldsymbol{R}_{1,1} - \boldsymbol{I})(\boldsymbol{R}_{1,1} - \boldsymbol{I})^T\right\}\right. \\
&\quad \left.+\mathrm{E}\left\{(\boldsymbol{R}_{1,1} - \boldsymbol{I})\boldsymbol{P}_1(\boldsymbol{R}_{1,1} - \boldsymbol{I})^T\right\}\right]\mathbf{1} \\
&= \frac{\sigma^2}{n^2}\mathbf{1}^T\left[\left(\frac{n-1}{n}\right)\mathrm{E}\left\{(\boldsymbol{R}_{1,1} - \boldsymbol{I})(\boldsymbol{R}_{1,1} - \boldsymbol{I})^T\right\}\right]\mathbf{1}. \qquad (3.14)
\end{aligned}$$

Equation (3.14) is true, because $\boldsymbol{R}_{1,1}\boldsymbol{P}_1 = \boldsymbol{P}_1$ and thus $(\boldsymbol{R}_{1,1} - \boldsymbol{I})\boldsymbol{P}_1 = 0$. Furthermore,

$$\begin{aligned}
\mathrm{E}\left\{(\boldsymbol{R}_{1,1} - \boldsymbol{I})(\boldsymbol{R}_{1,1} - \boldsymbol{I})^T\right\} &= \mathrm{E}\left\{\boldsymbol{R}_{1,1}\boldsymbol{R}_{1,1}^T - \boldsymbol{R}_{1,1} - \boldsymbol{R}_{1,1}^T + \boldsymbol{I}\right\} \\
&= \left(\frac{n-1}{n}\boldsymbol{I} + \boldsymbol{P}_1\right) - 2\boldsymbol{P}_1 + \boldsymbol{I} \\
&= \frac{n-1}{n}\boldsymbol{I} - \boldsymbol{P}_1 + \boldsymbol{I} \\
&= \frac{2n-1}{n}\boldsymbol{I} - \boldsymbol{P}_1 \,. \qquad (3.15)
\end{aligned}$$

Substituting equation (3.15) into equation (3.14), we see that

$$\begin{aligned}
\mathrm{Var}(D_1) &= \frac{\sigma^2}{n^2}\mathbf{1}^T\left\{\frac{n-1}{n}\left(\frac{2n-1}{n}\boldsymbol{I} - \boldsymbol{P}_1\right)\right\}\mathbf{1} \\
&= \frac{\sigma^2}{n^2}\left\{\frac{n-1}{n}\frac{2n-1}{n}n - \frac{n-1}{n}n\right\} \\
&= \frac{\sigma^2}{n}\frac{n-1}{n}\left\{\frac{2n-1}{n} - \frac{n}{n}\right\} \\
&= \frac{\sigma^2}{n}\left(\frac{n-1}{n}\right)^2 .
\end{aligned}$$

We have shown that $\mathrm{E}(\widehat{V}_{VoD}) = \dfrac{\sigma^2}{n}\left(\dfrac{n-1}{n}\right)^2$. It follows that $\mathrm{E}\left\{\left(\dfrac{n-1}{n}\right)^2\widehat{V}_{VoD}\right\} = \dfrac{\sigma^2}{n}$ and thus that

$$\mathrm{E}\left\{\left(1 + \frac{n-1}{nB}\right)\left(\frac{n-1}{n}\right)^2\widehat{V}_{VoD}\right\} = \frac{\sigma^2}{n}\left(1 + \frac{n-1}{nB}\right) = \mathrm{Var}(\widehat{\gamma}_{BAG}).$$

Hence, $\mathrm{E}(\widetilde{V}_{VoD}) = \mathrm{Var}(\widehat{\gamma}_{BAG})$. ■

3.3 Two Variance Estimators with Structures Similar to $\widehat{V}_{VoD}$ and $\widehat{V}_{DoV}$

In this section, we define two variance estimators with structures similar to $\widehat{V}_{VoD}$ and $\widehat{V}_{DoV}$. Under some conditions the one with the structure of $\widehat{V}_{VoD}$ is more efficient than

the one with the structure of $\widehat{V}_{DoV}$. This result is useful in comparing the efficiency of $\widehat{V}_{VoD}$ and $\widehat{V}_{DoV}$ in the linear model without variable selection in the next section.

Given pairs (V_i, U_i), $i = 1, \cdots, n$, define $W_i = V_i + U_i$. Consider the statistics

$$
\begin{aligned}
s_V^2 &= \frac{\sum_{i=1}^n (V_i - \overline{V})^2}{n-1}, && (3.16)\\
s_{VW} &= \frac{\sum_{i=1}^n V_i (W_i - \overline{W})}{n-1}, && (3.17)\\
s_U^2 &= \frac{\sum_{i=1}^n \{(W_i - V_i) - (\overline{W} - \overline{V})\}^2}{n-1}, && (3.18)\\
s_{VU} &= \frac{\sum_{i=1}^n V_i (U_i - \overline{U})}{n-1}. && (3.19)
\end{aligned}
$$

For the variance of U, we have two different estimators defined as

$$
\begin{aligned}
\widehat{\sigma}_1^2 &= \frac{\sum_{i=1}^n \{(W_i - V_i) - (\overline{W} - \overline{V})\}^2}{n-1}, && (3.20)\\
\widehat{\sigma}_2^2 &= \frac{\sum_{i=1}^n (W_i - \overline{W})^2 - \sum_{i=1}^n (V_i - \overline{V})^2}{n-1}. && (3.21)
\end{aligned}
$$

Note that $\widehat{\sigma}_1^2$ in (3.20) is the same as s_U^2 defined in (3.18). Recall the parallel bootstrap variance estimators are

$$
\widehat{V}_{DoV} = \frac{\sum_{b=1}^B (\widehat{\gamma}_{b,1} - \overline{\widehat{\gamma}}_{\cdot,1})^2}{B-1} - \left(1 - \frac{1}{B}\right) \frac{\sum_{b=1}^B (\widehat{\gamma}_b - \overline{\widehat{\gamma}}_{\cdot})^2}{B-1}
$$

and

$$
\widehat{V}_{VoD} = \frac{\sum_{b=1}^B \{(\widehat{\gamma}_{b,1} - \widehat{\gamma}_b) - (\overline{\widehat{\gamma}}_{\cdot,1} - \overline{\widehat{\gamma}}_{\cdot})\}^2}{B-1}.
$$

Thus, $\widehat{\sigma}_1^2$ and $\widehat{V}_{VoD}$ have the same structure, and $\widehat{\sigma}_2^2$ and $\widehat{V}_{DoV}$ have the same structure.

Using definitions (3.16), (3.17) and (3.19), we have

$$\begin{aligned} s_{VW} &= \frac{\sum_{i=1}^n V_i(W_i - \overline{W})}{n-1} \\ &= \frac{\sum_{i=1}^n V_i(V_i + U_i - \overline{V} - \overline{U})}{n-1} \\ &= \frac{\sum_{i=1}^n V_i(V_i - \overline{V})}{n-1} + \frac{\sum_{i=1}^n V_i(U_i - \overline{U})}{n-1} \\ &= s_V^2 + s_{VU}. \end{aligned} \tag{3.22}$$

Subtracting (3.20) from (3.21) produces

$$\begin{aligned} \widehat{\sigma}_2^2 - \widehat{\sigma}_1^2 &= \frac{\sum_{i=1}^n (W_i - \overline{W})^2 - \sum_{i=1}^n (V_i - \overline{V})^2}{n-1} - \frac{\sum_{i=1}^n (W_i - V_i - \overline{W} + \overline{V})^2}{n-1} \\ &= \frac{\sum_{i=1}^n (W_i - \overline{W})^2 - \sum_{i=1}^n (V_i - \overline{V})^2}{n-1} - \frac{\sum_{i=1}^n \{(W_i - \overline{W}) - (V_i - \overline{V})\}^2}{n-1} \\ &= \frac{\sum_{i=1}^n (W_i - \overline{W})^2 - \sum_{i=1}^n (V_i - \overline{V})^2}{n-1} - \frac{\sum_{i=1}^n (W_i - \overline{W})^2}{n-1} - \frac{\sum_{i=1}^n (V_i - \overline{V})^2}{n-1} \\ &\quad +2\,\frac{\sum_{i=1}^n (W_i - \overline{W})(V_i - \overline{V})}{n-1} \\ &= 2\,\frac{\sum_{i=1}^n (W_i - \overline{W})(V_i - \overline{V})}{n-1} - 2\,\frac{\sum_{i=1}^n (V_i - \overline{V})^2}{n-1} \\ &= 2(s_{VW} - s_V^2) \end{aligned} \tag{3.23}$$

The next theorem states certain conditions under which the conditional means of $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ are equal and the conditional variance of $\widehat{\sigma}_1^2$ is smaller than that of $\widehat{\sigma}_2^2$. Consequently, $\widehat{\sigma}_1^2$ is more efficient than $\widehat{\sigma}_2^2$ in estimating the variance of U.

Theorem 2. *Assume that there exists a random vector $\boldsymbol{T}$ such that the random pairs $(V_i, U_i)_{i=1}^n$ are conditionally i.i.d. given $\boldsymbol{T}$. Define $W_i = V_i + U_i$, and let s_U^2, s_{VU}, $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ be defined as in* (3.18), (3.19), (3.20) *and* (3.21). *Also assume:*

(A1) $E(U_1|V_1, \boldsymbol{T}) = 0$,

(A2) $E(s_U^2 s_{VU}|\boldsymbol{T}) \geq 0$,

(A3) $E(V_1^4) + E(U_1^4) < \infty$.

Then, $E(\widehat{\sigma}_1^2|\boldsymbol{T}) = E(\widehat{\sigma}_2^2|\boldsymbol{T}) = Var(U|\boldsymbol{T})$ *and* $Var(\widehat{\sigma}_1^2|\boldsymbol{T}) \leq Var(\widehat{\sigma}_2^2|\boldsymbol{T})$.

Proof. From assumption **(A1)**, we obtain that

$$\mathrm{E}(U_1|\boldsymbol{T}) = \mathrm{E}\{\mathrm{E}(U_1|V_1, \boldsymbol{T})|\boldsymbol{T}\} = 0 \tag{3.24}$$

By the definition of covariance, we have

$$\mathrm{Cov}(V_1, U_1|\boldsymbol{T}) = \mathrm{E}(V_1U_1|\boldsymbol{T}) - \mathrm{E}(V_1|\boldsymbol{T})\mathrm{E}(U_1|\boldsymbol{T}). \tag{3.25}$$

Substituting (3.24) into (3.25),

$$\mathrm{Cov}(V_1, U_1|\boldsymbol{T}) = \mathrm{E}\{V_1\mathrm{E}(U_1|V_1, \boldsymbol{T})|\boldsymbol{T}\} = \mathrm{E}(V_1 \cdot 0|\boldsymbol{T}) = 0.$$

Therefore,

$$\begin{aligned}\mathrm{Var}(W_1|\boldsymbol{T}) &= \mathrm{Var}(V_1 + U_1|\boldsymbol{T}) \\ &= \mathrm{Var}(V_1|\boldsymbol{T}) + \mathrm{Var}(U_1|\boldsymbol{T}) + 2\,\mathrm{Cov}(V_1, U_1|\boldsymbol{T}) \\ &= \mathrm{Var}(V_1|\boldsymbol{T}) + \mathrm{Var}(U_1|\boldsymbol{T}). \end{aligned} \tag{3.26}$$

The fact that $\mathrm{Cov}(V_1, U_1|\boldsymbol{T}) = 0$ is used to obtain (3.26). Since $(V_i, U_i)_{i=1}^n$ are i.i.d. given $\boldsymbol{T}$,

$$\mathrm{Var}(W_1|\boldsymbol{T}) = \mathrm{E}\left\{\frac{\sum_{i=1}^n (W_i - \overline{W})^2}{n-1}|\boldsymbol{T}\right\} \text{ and } \mathrm{Var}(V_1|\boldsymbol{T}) = \mathrm{E}\left\{\frac{\sum_{i=1}^n (V_i - \overline{V})^2}{n-1}|\boldsymbol{T}\right\}.$$

Hence,

$$\mathrm{E}(\widehat{\sigma}_2^2|\boldsymbol{T}) = \mathrm{E}\left\{\frac{\sum_{i=1}^n (W_i - \overline{W})^2}{n-1}|\boldsymbol{T}\right\} - \mathrm{E}\left\{\frac{\sum_{i=1}^n (V_i - \overline{V})^2}{n-1}|\boldsymbol{T}\right\} = \mathrm{Var}(W_1|\boldsymbol{T}) - \mathrm{Var}(V_1|\boldsymbol{T}). \tag{3.27}$$

Substituting (3.26) into (3.27), we obtain $\mathrm{E}(\widehat{\sigma}_2^2|\boldsymbol{T}) = \mathrm{Var}(U_1|\boldsymbol{T})$. Also, $\mathrm{E}(\widehat{\sigma}_1^2|\boldsymbol{T}) =$

$\mathrm{Var}(U_1|\boldsymbol{T})$. Thus,

$$\mathrm{E}(\widehat{\sigma}_1^2|\boldsymbol{T}) = \mathrm{E}(\widehat{\sigma}_2^2|\boldsymbol{T}) = \mathrm{Var}(U|\boldsymbol{T}).$$

Next, we continue to show $\mathrm{Var}(\widehat{\sigma}_1^2|\boldsymbol{T}) \leq \mathrm{Var}(\widehat{\sigma}_2^2|\boldsymbol{T})$.

Rewrite $\widehat{\sigma}_2^2 = \widehat{\sigma}_1^2 + \widehat{\sigma}_2^2 - \widehat{\sigma}_1^2$. Then

$$\mathrm{Var}(\widehat{\sigma}_2^2|\boldsymbol{T}) = \mathrm{Var}(\widehat{\sigma}_1^2|\boldsymbol{T}) + \mathrm{Var}(\widehat{\sigma}_2^2 - \widehat{\sigma}_1^2|\boldsymbol{T}) + 2\,\mathrm{Cov}(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2 - \widehat{\sigma}_1^2|\boldsymbol{T}).$$

It suffices to show $\mathrm{Cov}(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2 - \widehat{\sigma}_1^2|\boldsymbol{T}) \geq 0$ to prove $\mathrm{Var}(\widehat{\sigma}_1^2|\boldsymbol{T}) \leq \mathrm{Var}(\widehat{\sigma}_2^2|\boldsymbol{T})$.

Substituting $\mathrm{E}(\widehat{\sigma}_1^2|\boldsymbol{T}) = \mathrm{E}(\widehat{\sigma}_2^2|\boldsymbol{T})$ into the definition of $\mathrm{Cov}(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2 - \widehat{\sigma}_1^2|\boldsymbol{T})$ yields

$$\begin{aligned}
\mathrm{Cov}(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2 - \widehat{\sigma}_1^2|\boldsymbol{T}) &= \mathrm{E}\{\widehat{\sigma}_1^2(\widehat{\sigma}_2^2 - \widehat{\sigma}_1^2)|\boldsymbol{T}\} - \mathrm{E}(\widehat{\sigma}_1^2|\boldsymbol{T})\mathrm{E}\{(\widehat{\sigma}_2^2 - \widehat{\sigma}_1^2)|\boldsymbol{T}\} \\
&= \mathrm{E}\{\widehat{\sigma}_1^2(\widehat{\sigma}_2^2 - \widehat{\sigma}_1^2)|\boldsymbol{T}\} \\
&= \mathrm{E}\{s_U^2(\widehat{\sigma}_2^2 - \widehat{\sigma}_1^2)|\boldsymbol{T}\} \qquad (3.28) \\
& \qquad (3.29)
\end{aligned}$$

Substituting (3.23) into above (3.28),

$$\begin{aligned}
\mathrm{Cov}(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2 - \widehat{\sigma}_1^2|\boldsymbol{T}) &= \mathrm{E}[s_U^2\{2(s_{VW} - s_V^2)\}|\boldsymbol{T}] \qquad (3.30) \\
&= 2\,\mathrm{E}(s_U^2 s_{VU}|\boldsymbol{T}).
\end{aligned}$$

Equation (3.22) is used to arrive at the last step. By the condition **(A2)** of Theorem 2, $\mathrm{E}(s_U^2 s_{VU}|\boldsymbol{T}) \geq 0$, we have that $\mathrm{Cov}(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2 - \widehat{\sigma}_1^2|\boldsymbol{T}) \geq 0$ and it follows that $\mathrm{Var}(\widehat{\sigma}_1^2|\boldsymbol{T}) \leq \mathrm{Var}(\widehat{\sigma}_2^2|\boldsymbol{T})$. ■

Corrollary 1. *Under the conditions of Theorem 2,* $Var(\widehat{\sigma}_1^2) \leq Var(\widehat{\sigma}_2^2)$.

Proof. The iterated variance formula yields

$$\begin{aligned}\mathrm{Var}(\widehat{\sigma}_1^2) &= \mathrm{Var}\left\{\mathrm{E}\left(\widehat{\sigma}_1^2|\boldsymbol{T}\right)\right\} + \mathrm{E}\left\{\mathrm{Var}\left(\widehat{\sigma}_1^2|\boldsymbol{T}\right)\right\}, \text{ and} \\ \mathrm{Var}(\widehat{\sigma}_2^2) &= \mathrm{Var}\left\{\mathrm{E}\left(\widehat{\sigma}_2^2|\boldsymbol{T}\right)\right\} + \mathrm{E}\left\{\mathrm{Var}\left(\widehat{\sigma}_2^2|\boldsymbol{T}\right)\right\}.\end{aligned}$$

From Theorem 2, we have $\mathrm{E}(\widehat{\sigma}_1^2|\boldsymbol{T}) = \mathrm{E}(\widehat{\sigma}_2^2|\boldsymbol{T})$ and $\mathrm{Var}(\widehat{\sigma}_1^2|\boldsymbol{T}) \leq \mathrm{Var}(\widehat{\sigma}_2^2|\boldsymbol{T})$. Thus,

$$\mathrm{Var}\left\{\mathrm{E}\left(\widehat{\sigma}_1^2|\boldsymbol{T}\right)\right\} = \mathrm{Var}\left\{\mathrm{E}\left(\widehat{\sigma}_2^2|\boldsymbol{T}\right)\right\}, \qquad \mathrm{E}\left\{\mathrm{Var}\left(\widehat{\sigma}_1^2|\boldsymbol{T}\right)\right\} \leq \mathrm{E}\left\{\mathrm{Var}\left(\widehat{\sigma}_2^2|\boldsymbol{T}\right)\right\}.$$

Therefore, $\mathrm{Var}(\widehat{\sigma}_1^2) \leq \mathrm{Var}(\widehat{\sigma}_2^2)$. ■

In the proof of Theorem 2, $\mathrm{E}(s_U^2 s_{VU}|\boldsymbol{T}) \geq 0$ is a critical condition for showing that $\mathrm{Var}(\widehat{\sigma}_1^2|\boldsymbol{T}) \leq \mathrm{Var}(\widehat{\sigma}_2^2|\boldsymbol{T})$. The next two lemmas specify two sufficient conditions for $\mathrm{E}(s_U^2 s_{VU}|\boldsymbol{T}) \geq 0$.

Lemma 1. *Under the conditions of Theorem 2 except* **(A2)**, *if V_i and U_i are independent given $\boldsymbol{T}$, then $E(s_U^2 s_{VU}|\boldsymbol{T}) = 0$.*

Proof. First, we have

$$\begin{aligned}\mathrm{E}(s_{VU}|\boldsymbol{U},\boldsymbol{T}) &= \mathrm{E}(s_{VU}|U_1,\cdots,U_n,\boldsymbol{T}) \\ &= \mathrm{E}\left\{\frac{\sum_{i=1}^n (U_i-\overline{U})V_i}{n-1}|U_1,\cdots,U_n,\boldsymbol{T}\right\} \\ &= \sum_{i=1}^n \frac{(U_i-\overline{U})\mathrm{E}(V_i|U_1,\cdots,U_n,\boldsymbol{T})}{n-1} \\ &= \sum_{i=1}^n \frac{(U_i-\overline{U})\mathrm{E}(V_i|U_i,\boldsymbol{T})}{n-1}.\end{aligned}$$

The last step is true, because conditioning on $\boldsymbol{T}$, $\{(V_i,U_i)\}_{i=1}^n$ are i.i.d., $\mathrm{E}(V_i|U_1,\cdots,U_n,\boldsymbol{T})$ $= \mathrm{E}(V_i|U_i,\boldsymbol{T})$, for $i = 1,\cdots,n$. Because V and U are conditionally independent, we ob-

tain

$$\begin{aligned}\mathrm{E}(s_{VU}|\boldsymbol{U},\boldsymbol{T}) &= \sum_{i=1}^{n}\frac{(U_i-\overline{U})\mathrm{E}(V_i|\boldsymbol{T})}{n-1},\\ &= \sum_{i=1}^{n}\frac{(U_i-\overline{U})\mathrm{E}(V_1|\boldsymbol{T})}{n-1}\\ &= 0.\end{aligned}$$

Therefore, $\mathrm{E}(s_U^2 s_{VU}|\boldsymbol{T}) = \mathrm{E}\{\mathrm{E}(s_U^2 s_{VU}|\boldsymbol{U},\boldsymbol{T})\} = \mathrm{E}\{s_U^2\mathrm{E}(s_{VU}|\boldsymbol{U},\boldsymbol{T})|\boldsymbol{T}\} = 0.$ ■

Lemma 2. *Under the conditions of Theorem 2 except* **(A2)**, *if* $E(U_i^3|V_i,\boldsymbol{T}) = 0$, *then* $E(s_U^2 s_{VU}|\boldsymbol{T}) = 0$.

Proof. Observe that condition **(A1)** yields $\mathrm{E}(U_i) = 0$. Rewrite

$$s_U^2 s_{VU} = \frac{\sum_{i=1}^{n}\{(V_i-\overline{V})U_i s_U^2\}}{n-1} = \frac{\sum_{i=1}^{n}\{(V_i-\overline{V})A_i\}}{n-1}, \text{ where } A_i = U_i s_U^2.$$

Thus,

$$\begin{aligned}\mathrm{E}(s_U^2 s_{VU}|\boldsymbol{T}) &= \mathrm{E}\{\mathrm{E}(s_U^2 s_{VU}|V_1,\cdots,V_n,\boldsymbol{T})|\boldsymbol{T}\}\\ &= \frac{\sum_{i=1}^{n}[\mathrm{E}\{(V_i-\overline{V})\mathrm{E}(A_i|V_1,\cdots,V_n,\boldsymbol{T})|\boldsymbol{T}\}]}{n-1}.\end{aligned}$$

Furtherly decompose A_i as

$$\begin{aligned}A_i &= U_i s_U^2\\ &= U_i\frac{\sum_{j=1}^{n}(U_j-\overline{U})^2}{n-1}\\ &= \frac{U_i(\sum_{j\neq i}U_j^2)+U_i^3-nU_i\overline{U}^2}{n-1}\\ &= A_{i1}+A_{i2}-A_{i3}.\end{aligned}$$

Hence,

$$\mathrm{E}(A_i|V_1,\cdots,V_n,\boldsymbol{T}) = \mathrm{E}(A_{i1}|V_1,\cdots,V_n,\boldsymbol{T})+\mathrm{E}(A_{i2}|V_1,\cdots,V_n,\boldsymbol{T})-\mathrm{E}(A_{i3}|V_1,\cdots,V_n,\boldsymbol{T}).$$

Firstly,

$$\begin{aligned}\mathrm{E}(A_{i1}|V_1,\cdots,V_n,\boldsymbol{T}) &= \frac{\mathrm{E}[U_i(\sum_{j\neq i}U_j^2)|V_1,\cdots,V_n,\boldsymbol{T}]}{n-1}\\ &= \frac{\mathrm{E}(U_i|V_1,\cdots,V_n,\boldsymbol{T})\mathrm{E}(\sum_{j\neq i}U_j^2|V_1,\cdots,V_n,\boldsymbol{T})}{n-1}.\end{aligned}$$

Since conditioning on $\boldsymbol{T}$, $\{(V_i,U_i)\}_{i=1}^n$ are i.i.d,

$$\mathrm{E}(A_{i1}|V_1,\cdots,V_n,\boldsymbol{T}) = \frac{\mathrm{E}(U_i|V_i,\boldsymbol{T})\mathrm{E}(\sum_{j\neq i}U_j^2|V_j,\boldsymbol{T})}{n-1}.$$

Note that $\mathrm{E}(U_i|V_i,\boldsymbol{T})=0$, and thus $\mathrm{E}(A_{i1}|V_1,\cdots,V_n,\boldsymbol{T})=0$.

Secondly, by the assumption that $\mathrm{E}(U_i^3|V_i,\boldsymbol{T})=0$,

$$\begin{aligned}\mathrm{E}(A_{i2}|V_1,\cdots,V_n,\boldsymbol{T}) &= \frac{\mathrm{E}(U_i^3|V_1,\cdots,V_n,\boldsymbol{T})}{n-1}\\ &= \frac{\mathrm{E}(U_i^3|V_i,\boldsymbol{T})}{n-1},\\ &= 0.\end{aligned}$$

Finally, we show that $\mathrm{E}(A_{i3}|V_i,\cdots,V_n,\boldsymbol{T})=0$. Decompose $U_i\overline{U}^2$ as

$$\begin{aligned}U_i\overline{U}^2 &= U_i\left(\frac{U_i+\sum_{j\neq i}U_j}{n}\right)^2\\ &= \frac{U_i^3+U_i(\sum_{j\neq i}U_j)^2+2\,U_i^2(\sum_{j\neq i}U_j)}{n^2},\end{aligned}$$

Because $\mathrm{E}(U_i^3|V_1,\cdots,V_n,\boldsymbol{T})=0$,

$$\mathrm{E}\left[U_i(\sum_{j\neq i}U_j)^2|V_1,\cdots,V_n,\boldsymbol{T}\right]=\mathrm{E}(U_i|V_i,\boldsymbol{T})\mathrm{E}\left[(\sum_{j\neq i}U_j)^2|V_1,\cdots,V_n,\boldsymbol{T}\right]=0,$$

and

$$\mathrm{E}\left[U_i^2(\sum_{j\neq i}U_j)|V_1,\cdots,V_n,\boldsymbol{T}\right]=\mathrm{E}(U_i^2|V_i,\boldsymbol{T})\left[\sum_{j\neq i}\mathrm{E}(U_j|V_j,\boldsymbol{T})\right]=0,$$

we have $\mathrm{E}\{U_i\overline{U}^2|V_1,\cdots,V_n,\boldsymbol{T}\}=0$ and $\mathrm{E}(A_{i3}|V_1,\cdots,V_n,\boldsymbol{T})=0$.

Putting all together, we get $\mathrm{E}(A_i|V_1,\cdots,V_n,\boldsymbol{T})=0$ and $\mathrm{E}(s_U^2 s_{VU}|\boldsymbol{T})=0$. ■

We constructed two variance estimators $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ so that the structure of $\widehat{\sigma}_1^2$ is similar to $\widehat{V}_{VoD}$ and the structure of $\widehat{\sigma}_2^2$ is similar to $\widehat{V}_{DoV}$. We showed that under conditions specified in Lemma 1 or Lemma 2, $\widehat{\sigma}_1^2$ is more efficient than $\widehat{\sigma}_2^2$. In the next section, we are going to use this to prove that $\widehat{V}_{VoD}$ is more efficient than $\widehat{V}_{DoV}$ in the ordinary linear regression model without variable selection when the level-2 bootstrap samples are drawn by the normal linear model parametric bootstrap.

3.4 Comparison of $\widehat{V}_{DoV}$ and $\widehat{V}_{VoD}$ for Least Square Linear Regression without Variable Selection

We are not able to rigorously prove that the variance estimator $\widehat{V}_{VoD}$ is more efficient than $\widehat{V}_{DoV}$ in the situation, where a variable selection procedure is used to build the model. However, we can show this for least squares linear regression without variable selection when the level-2 samples are drawn by the normal linear model parametric bootstrap instead of residual-based bootstrap. We show that $\widehat{\gamma}_{b,1}$ and $\widehat{\gamma}_b$ satisfy $\widehat{\gamma}_{b,1}=\widehat{\gamma}_b+\widehat{e}_b$, where $\mathrm{E}(\widehat{e}_b|\widehat{\gamma}_b,\boldsymbol{Y})=0$ and $\mathrm{E}(\widehat{e}_b^3|\widehat{\gamma}_b,\boldsymbol{Y})=0$. Then by Theorem 2 and Lemma 2, it follows that $\widehat{V}_{VoD}$ is more efficient than $\widehat{V}_{DoV}$.

We assume a linear regression model with k_T predictor variables available,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.31}$$

where $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ is an $n \times 1$ vector of response values, $\boldsymbol{\beta}$ is a $(k_T+1)\times 1$ regression coefficient vector, $\boldsymbol{X}$ is a $n \times (k_T + 1)$ matrix with the ith row $(1, \boldsymbol{x}_i^T)$, and $\boldsymbol{\epsilon}$ is an $n \times 1$ random error vector with each element independently drawn from a distribution with mean 0 and variance σ^2. To make the following proof easier, throughout this section, we let the first column of $\boldsymbol{X}$ be a vector of 1's and the first element of $\boldsymbol{\beta}$ correspond to the intercept.

First let us describe how to compute $\widehat{\gamma}_b$ and $\widehat{\gamma}_{b,1}$. Let $(\boldsymbol{R}_b\boldsymbol{X}, \boldsymbol{R}_b\boldsymbol{Y})$ denote the bth random-pair bootstrap sample from the original data set $(\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{R}_b$ is the bth bootstrap selection matrix. We fit the ordinary least square, obtain an estimate for the regression coefficient vector $\boldsymbol{\beta}$ and make prediction for $\boldsymbol{x} = \boldsymbol{x}_0$, i.e., $\widehat{\boldsymbol{\beta}}_b = \left\{(\boldsymbol{R}_b\boldsymbol{X})^T\boldsymbol{R}_b\boldsymbol{X}\right\}^{-1}(\boldsymbol{R}_b\boldsymbol{X})^T\boldsymbol{R}_b\boldsymbol{Y}$ and $\widehat{\gamma}_b = \boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}_b$. Using the the normal linear model parametric bootstrap method, we generate the bth level-2 sample $(\boldsymbol{R}_b\boldsymbol{X}, \boldsymbol{Y}_{b,1})$, where $\boldsymbol{Y}_{b,1} = (\boldsymbol{R}_b\boldsymbol{X})\widehat{\boldsymbol{\beta}}_b + \widehat{\sigma}_b\boldsymbol{Z}$, $\widehat{\sigma}_b$ is the square root of the mean squared error of the least square fit to the level-1 sample $(\boldsymbol{R}_b\boldsymbol{X}, \boldsymbol{R}_b\boldsymbol{Y})$, and $\boldsymbol{Z} = (Z_1, \cdots, Z_n)^T$is an i.i.d. sample from the standard normal distribution. Then we fit the least square model to the level-2 sample $(\boldsymbol{R}_b\boldsymbol{X}, \boldsymbol{Y}_{b,1})$ and make prediction for $\boldsymbol{x} = \boldsymbol{x}_0$ as $\widehat{\gamma}_{b,1} = \boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}_{b,1}$, where $\widehat{\boldsymbol{\beta}}_{b,1} = \left\{(\boldsymbol{R}_b\boldsymbol{X})^T\boldsymbol{R}_b\boldsymbol{X}\right\}^{-1}(\boldsymbol{R}_b\boldsymbol{X})^T\boldsymbol{Y}_{b,1}$.

Next we show that in this simplified case $\widehat{V}_{VoD}$ is more efficient than $\widehat{V}_{DoV}$.

Theorem 3. *For the linear model* (3.31)*, suppose that the estimates from the bth level-1 and level-2 bootstrap samples, $\widehat{\gamma}_b$ and $\widehat{\gamma}_{b,1}$, are computed as above. Then $\widehat{\gamma}_{b,1}$ can be written as $\widehat{\gamma}_{b,1} = \widehat{\gamma}_b + \widehat{e}_b$, such that $E(\widehat{e}_b|\widehat{\gamma}_b, \boldsymbol{Y}) = 0$ and $E(\widehat{e}_b^3|\widehat{\gamma}_b, \boldsymbol{Y}) = 0$. From Lemma 2, it follows that $E(s_{\widehat{e}}^2 s_{\widehat{\gamma}\widehat{e}}|\boldsymbol{Y}) = 0$. Application of Theorem 2 yields $Var(\widehat{V}_{VoD}) \leq Var(\widehat{V}_{DoV})$.*

Proof. First rewrite

$$\begin{aligned}
\widehat{\gamma}_{b,1} &= \boldsymbol{x}_0^T \widehat{\beta}_{b,1} \\
&= \boldsymbol{x}_0^T \left\{ (\boldsymbol{R}_b \boldsymbol{X})^T \boldsymbol{R}_b \boldsymbol{X} \right\}^{-1} (\boldsymbol{R}_b \boldsymbol{X})^T \boldsymbol{Y}_{b,1} \\
&= \boldsymbol{x}_0^T \left\{ (\boldsymbol{R}_b \boldsymbol{X})^T \boldsymbol{R}_b \boldsymbol{X} \right\}^{-1} (\boldsymbol{R}_b \boldsymbol{X})^T \{ (\boldsymbol{R}_b \boldsymbol{X}) \widehat{\beta}_b + \widehat{\sigma}_b \boldsymbol{Z} \} \\
&= \boldsymbol{x}_0^T \widehat{\beta}_b + \widehat{\sigma}_b x_0^T \left\{ (\boldsymbol{R}_b \boldsymbol{X})^T \boldsymbol{R}_b \boldsymbol{X} \right\}^{-1} (\boldsymbol{R}_b \boldsymbol{X})^T \boldsymbol{Z} \\
&= \widehat{\gamma}_b + \widehat{e}_b,
\end{aligned}$$

where $\widehat{e}_b = \widehat{\sigma}_b \boldsymbol{x}_0^T \left\{ (\boldsymbol{R}_b \boldsymbol{X})^T \boldsymbol{R}_b \boldsymbol{X} \right\}^{-1} (\boldsymbol{R}_b \boldsymbol{X})^T \boldsymbol{Z}$. Let $\mathbf{a}^T = \widehat{\sigma}_b \boldsymbol{x}_0^T \left\{ (\boldsymbol{R}_b \boldsymbol{X})^T \boldsymbol{R}_b \boldsymbol{X} \right\}^{-1} (\boldsymbol{R}_b \boldsymbol{X})^T$, then $\mathbf{a}$ is an $n \times 1$ vector and $\widehat{e}_b = \mathbf{a}^T \boldsymbol{Z}$. We know that $(\widehat{e}_b | \boldsymbol{R}_b, \boldsymbol{Y}) \sim N(0, \mathbf{a}^T \mathbf{a})$. It follows that $\mathrm{E}(\widehat{e}_b | \widehat{\gamma}_b, \boldsymbol{Y}) = 0$ and $\mathrm{E}(\widehat{e}_b^3 | \widehat{\gamma}_b, \boldsymbol{Y}) = 0$.

Now, let $\widehat{\gamma}_{b,1}$, $\widehat{\gamma}_b$, $\widehat{e}_b$ and $\boldsymbol{Y}$ play the role of W_i, V_i, U_i and $\boldsymbol{T}$ in Theorem 2, respectively. Conditioning on $\boldsymbol{Y}$, for $b = 1, \cdots, B$, $(\widehat{\gamma}_{b,1}, \widehat{\gamma}_b, \widehat{e}_b)$ are i.i.d. vectors. By Lemma 2, we have that $\mathrm{E}(s_{\widehat{e}}^2 s_{\widehat{\gamma}\widehat{e}} | \boldsymbol{Y}) = 0$. Also from the definition of $\widehat{V}_{VoD}$ and $\widehat{V}_{DoV}$ we have

$$\widehat{V}_{VoD} = \frac{\sum_{b=1}^B \{ (\widehat{\gamma}_{b,1} - \widehat{\gamma}_b) - (\overline{\widehat{\gamma}}_{\cdot,1} - \overline{\widehat{\gamma}}_{\cdot}) \}^2}{B-1} = \widehat{\sigma}_1^2,$$

$$\begin{aligned}
\widehat{V}_{DoV} &= \frac{\sum_{b=1}^B (\widehat{\gamma}_{b,1} - \overline{\widehat{\gamma}}_{\cdot,1})^2}{B-1} - \left(1 - \frac{1}{B}\right) \frac{\sum_{b=1}^B (\widehat{\gamma}_b - \overline{\widehat{\gamma}}_{\cdot})^2}{B-1} \\
&\approx \frac{\sum_{b=1}^B (\widehat{\gamma}_{b,1} - \overline{\widehat{\gamma}}_{\cdot,1})^2}{B-1} - \frac{\sum_{b=1}^B (\widehat{\gamma}_b - \overline{\widehat{\gamma}}_{\cdot})^2}{B-1} \\
&= \widehat{\sigma}_2^2.
\end{aligned}$$

Since $\widehat{\sigma}_2^2 \approx \widehat{V}_{DoV}$, we have $\mathrm{Var}(\widehat{\sigma}_2) \approx \mathrm{Var}(\widehat{V}_{DoV})$ and $\mathrm{E}(\widehat{V}_{DoV}) \approx \mathrm{E}(\widehat{\sigma}_2^2)$. By Theorem 2 and its corollary in Section 3.3, it follows that $\mathrm{E}(\widehat{V}_{DoV}) \approx \mathrm{E}(\widehat{\sigma}_2^2) = \mathrm{E}(\widehat{\sigma}_1^2) = \mathrm{E}(\widehat{V}_{DoV})$, and $\mathrm{Var}(\widehat{V}_{VoD}) = \mathrm{Var}(\widehat{\sigma}_1^2) \leq \mathrm{Var}(\widehat{\sigma}_2) \approx \mathrm{Var}(\widehat{V}_{DoV})$. ■

We have shown that $\widehat{V}_{VoD}$ has smaller variance than $\widehat{V}_{DoV}$ and they have approximately the same expectation. Thus $\widehat{V}_{VoD}$ is more efficient in estimating the variance of

the bagging estimator $\widehat{\gamma}_{BAG}$ without variable selection in the linear regression model and level-2 bootstrap samples are drawn by the normal linear model parametric bootstrap. In practice the residual-based bootstrap method is used to draw level-2 samples. Also variable selection is performed to build the final prediction model. However, the efficiency result of this section in the simplified situation motivates us to estimate the variance of the bagging estimator by $\widehat{V}_{VoD}$.

Chapter 4

Simulation Study

It is difficult to theoretically analyze the performance of bootstrap-based variance estimators when variable selection is performed to build a predictive model. In this section, we investigate the parallel bootstrap DoV, the parallel bootstrap VoD, and the bootstrap-after-bootstrap (BaB) variance estimators by Monte Carlo simulation in linear regression settings with variable selection procedures: the adaptive LASSO (ALASSO) and the fast FSR (FFSR).

4.1 Simulation Design

We consider linear regression models, $\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$, ϵ_i are i.i.d. $N(0, \sigma^2)$, β_0 is the intercept, and $\boldsymbol{\beta}$ is the the $k_T \times 1$ regression coefficient vector. The design matrix is denoted by $\boldsymbol{X}$, where the ith row is $\boldsymbol{x}_i^T$, and $\mathbf{1}$ is the $n \times 1$ vector of 1s. The size of the data set is n and k_T is the total number of predictors. Thus, $Y_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$, $i = 1, \cdots, n$.

Throughout this study, we consider a fixed $\boldsymbol{X}$ version of the linear regression model satisfying

$$\mathbf{1}_{n\times 1}^T \boldsymbol{X}_{n\times k_T} = \mathbf{0}_{k_T}, \text{ and } \boldsymbol{X}^T\boldsymbol{X} = n\boldsymbol{C}_{k_T\times k_T}(\rho),$$

where $\boldsymbol{C}(\rho)$ has elements $c_{ij} = \rho^{|i-j|}$. Specifically, for each (k_T, ρ) combination, a fixed $\boldsymbol{X}$ is constructed as $\boldsymbol{X} = (\boldsymbol{Z} - n^{-1}\mathbf{1}\mathbf{1}^T\boldsymbol{Z})\boldsymbol{A}$, where the elements of $\boldsymbol{Z}$ are an i.i.d. random sample from $N(0,1)$ and $\boldsymbol{A}$ is chosen so that $\boldsymbol{X}^T\boldsymbol{X} = n\boldsymbol{C}(\rho)$. The design matrix $\boldsymbol{X}$ has an autoregressive covariance structure of the first order. For each of the different situations, $\boldsymbol{X}$ remains fixed for generating the $n_{sim} = 1000$ independent sets of responses, where n_{sim} denotes the number of Monte Carlo replications.

For the linear model $\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with fixed design matrix $\boldsymbol{X}$, theoretical R^2 is defined as

$$R^2 = \frac{\boldsymbol{\mu}^T\boldsymbol{\mu} - (\mathbf{1}^T\boldsymbol{\mu})^2/n}{\boldsymbol{\mu}^T\boldsymbol{\mu} - (\mathbf{1}^T\boldsymbol{\mu})^2/n + n\sigma^2}, \text{ where } \boldsymbol{\mu} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}. \tag{4.1}$$

For specifying $\boldsymbol{\beta}$, we choose an initial regression coefficient vector $\boldsymbol{\beta}_*$ according to different configurations of informative covariates and complexity of models, then scale $\boldsymbol{\beta}_*$ to obtain the true underlying regression coefficients $\boldsymbol{\beta} = c\boldsymbol{\beta}_*$ so that the linear model has a specific theoretical R^2 value. By the definition of the theoretical R^2 in (4.1), the scale constant c is determined by

$$c^2 = \frac{n\sigma^2R^2}{(1-R^2)\{\boldsymbol{\mu}^T\boldsymbol{\mu} - (\mathbf{1}^T\boldsymbol{\mu})^2/n\}}.$$

Using the design matrix $\boldsymbol{X}$ and the regression coefficient vector $\boldsymbol{\beta}$, we generate the response vector as $\boldsymbol{Y} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$, ϵ_i's are i.i.d. $N(0, \sigma^2)$. In our simulation, we choose $\beta_0 = 1$ and $\sigma^2 = 9$. The sample size n is equal to 150. The number of Monte Carlo replications is $n_{sim} = 1000$. Notice that for a given sample size n and R^2, β/σ is fixed. So the choice of σ^2 does not matter.

This simulation study uses a factorial design with the following factors:

1. The autoregressive covariance structure of the design matrix $\boldsymbol{X}$: $\rho = 0.2$ corresponding to a low correlation case, and $\rho = 0.6$ corresponding to a high correlation case.

2. The theoretical R^2: $R^2 = 0.3$ created a low signal-to-noise ratio, and $R^2 = 0.7$ created a high signal-to-noise ratio.

3. The initial regression coefficient vector $\boldsymbol{\beta}_*$: the structure of $\boldsymbol{\beta}_*$ is chosen according to the model complexity measured by the total number covariates k_T and number of informative covariates k_I. It also incorporates different configurations with the informative covariates concentrated at the beginning, or clumped at three places (the beginning, the middle, and the end), or evenly mixed with zeros. Different configurations represent different correlation structures of the design matrix. This factor has the following 9 initial regression vectors:

   (a) $k_T = 26$, $k_I = 6$:

   Model $M_1 : \boldsymbol{\beta}_* = (\mathbf{1}_6^T, \mathbf{0}_{20}^T)^T$.

   Model $M_2 : \boldsymbol{\beta}_* = (\mathbf{1}_2^T, \mathbf{0}_{10}^T, \mathbf{1}_2^T, \mathbf{0}_{10}^T, \mathbf{1}_2^T)^T$.

   Model $M_3 : \boldsymbol{\beta}_* = (1, \mathbf{0}_4^T, 1, \mathbf{0}_4^T, 1, \mathbf{0}_4^T, 1, \mathbf{0}_4^T, 1, \mathbf{0}_4^T, 1)^T$.

   (b) $k_T = 26$, $k_I = 12$:

   Model $M_4 : \boldsymbol{\beta}_* = (\mathbf{1}_{12}^T, \mathbf{0}_{14}^T)^T$.

   Model $M_5 : \boldsymbol{\beta}_* = (\mathbf{1}_4^T, \mathbf{0}_7^T, \mathbf{1}_4^T, \mathbf{0}_7^T, \mathbf{1}_4^T)^T$.

   Model $M_6 : \boldsymbol{\beta}_* = (1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0)^T$.

   (c) $k_T = 48$, $k_I = 8$:

   Model $M_7 : \boldsymbol{\beta}_* = (\mathbf{1}_8^T, \mathbf{0}_{40}^T)^T$.

   (d) $k_T = 48$, $k_I = 12$:

   Model $M_8 : \boldsymbol{\beta}_* = (1, \mathbf{0}_3^T, 1, \mathbf{0}_3^T, 1, \mathbf{0}_3^T, 1, \mathbf{0}_3^T, 1, \mathbf{0}_3^T, 1, \mathbf{0}_3^T, 1, \mathbf{0}_3^T, 1, \mathbf{0}_3^T, 1, \mathbf{0}_3^T, 1, \mathbf{0}_3^T, 1, \mathbf{0}_3^T, 1)^T$.

   (e) $k_T = 48$, $k_I = 18$:

   Model $M_9 : \boldsymbol{\beta}_* = (\mathbf{1}_6^T, \mathbf{0}_{15}^T, \mathbf{1}_6^T, \mathbf{0}_{15}^T, \mathbf{1}_6^T)^T$.

For each of these $2 \times 2 \times 9 = 36$ situations, we generate $n_{sim} = 1000$ independent data sets with the same design matrix $\boldsymbol{X}$ and different responses. Recall that $\boldsymbol{X}$ remains

fixed for each combination of (k_T, ρ). Actually, there are 18 different $\boldsymbol{X}$ used. For each data set, we compute the bagging estimator for a predictor based on the adaptive LASSO (ALASSO), and based on the fast FSR (FFSR). For each of these predictions, we compute the "variance-of-difference" version of the parallel bootstrap (VoD), the "difference-of-variances" version of the parallel bootstrap (DoV), and the bootstrap-after-bootstrap (BaB) variance estimators.

In this simulation study we choose to predict the first quartile of simulated true means of the sample response vector, $\boldsymbol{\mu}_{([0.25n])}$, where $\boldsymbol{\mu} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{\mu}_{(1)} \leq \boldsymbol{\mu}_{(2)} \leq \cdots \boldsymbol{\mu}_{(n)}$ are the ordered means. In other words, if we denote the associated predictor vector as $\boldsymbol{x}_{([0.25n])}$, we predict Y at $\boldsymbol{x} = \boldsymbol{x}_{([0.25n])}$. Here $[0.25n]$ means the integer part of $0.25n$. Note that $\boldsymbol{x}_{([0.25n])}$ is fixed for each $(\rho, \boldsymbol{\beta})$.

We compare these three variance estimators in terms of bias and variation. Bias is measured by the ratio of the average of standard error estimates to the Monte Carlo standard deviation of the n_{sim} bagging estimators. This ratio, called the bias ratio (BR), is

$$BR = \frac{\text{MEAN}}{S_{MC}}, \tag{4.2}$$

where MEAN$= {n_{sim}}^{-1} \sum_{i=1}^{n_{sim}} \sqrt{\widehat{\text{Var}}(\widehat{\gamma}_i)}$ and $S_{MC}^2 = (n_{sim} - 1)^{-1} \sum_{i=1}^{n_{sim}} (\widehat{\gamma}_i - \overline{\widehat{\gamma}}_{\cdot})^2$.

If $BR < 1$, then the method is underestimating the true standard deviation as estimated by the Monte Carlo sample standard deviation S_{MC}. Variation of the standard deviation estimates $\sqrt{\widehat{\text{Var}}(\widehat{\gamma})}$ is measured by the coefficient of variation (CV),

$$CV = \frac{\text{SD}}{\text{MEAN}}, \tag{4.3}$$

where MEAN is defined as above and $\text{SD}^2 = (n_{sim} - 1)^{-1} \sum_{i=1}^{n_{sim}} \left(\sqrt{\widehat{\text{Var}}(\widehat{\gamma}_i)} - \text{MEAN} \right)^2$. The CV is used instead of SD in order to compare methods having different MEAN values.

4.2 Bagging Predictor and Bootstrap Variance Estimators

The goal of our simulation is to assess the performance of three bootstrap variance estimators for a bagging predictor. The bagging predictor is the average of different predicted values computed from different bootstrap samples. In this section we discuss in detail how to compute the bagging predictor and the three bootstrap-based variance estimators.

In the following algorithms, $\boldsymbol{R}$ denotes a bootstrap selection matrix and $\overline{\widehat{\gamma}}_{\cdot}$ is the sample average of $\widehat{\gamma}_b$, $b = 1, \cdots, B$, where $\widehat{\gamma}_b$ is the predicted value based on the bth bootstrap sample. Denote the original sample data as $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{Y})$ and the bth bootstrap sample as $\mathcal{L}^{(b)}$. Using $\mathcal{L}^{(b)}$, the estimated intercept and regression coefficient vector $(\widehat{\beta}_0^{(b)}, \{\widehat{\boldsymbol{\beta}}^{(b)}\}^T)^T$ are obtained by each estimation method. Here we emphasize variable selection methods, though bagging may be used with any estimation method. In this simulation study, two variable selection methods are used. One is the fast FSR applied to the forward addition sequence (FFSR) and the other one is the adaptive LASSO (ALASSO).

Bagging algorithm for a predictor:

Step (1) Draw the bth bootstrap sample $\mathcal{L}^{(b)} = (\boldsymbol{R}_b\boldsymbol{X}, \boldsymbol{R}_b\boldsymbol{Y})$ by the random-pair bootstrap.

Step (2) Use $\mathcal{L}^{(b)}$ to obtain $(\widehat{\beta}_0^{(b)}, \{\widehat{\boldsymbol{\beta}}^{(b)}\}^T)^T$ and make a prediction with $\boldsymbol{x}_{([0.25n])}$, $\widehat{\gamma}_b = \widehat{\beta}_0^{(b)} + \boldsymbol{x}_{([0.25n])}^T\widehat{\boldsymbol{\beta}}^{(b)}$.

Step (3) Repeat Step (1) and Step (2) B times to have B predicted values and average them to result in the bagging predictor $\widehat{\gamma}_{BAG} = \sum_{b=1}^{B}\widehat{\gamma}_b/B$.

Three bootstrap-based variance estimators are used to estimate the variance of $\widehat{\gamma}_{BAG}$. They are the "variance-of-difference" version of the parallel bootstrap (VoD), the "difference-

of-variances" version of the parallel bootstrap (DoV), and the bootstrap-ater-bootstrap (BaB). Next we describe three algorithms corresponding to these three variance estimators. The first two parallel bootstrap algorithms use the same B random-pair bootstrap samples from the bagging algorithm.

Parallel bootstrap VoD algorithm:

Step (1) Using the same bth bootstrap sample $\mathcal{L}^{(b)}$ and the resulting linear predictive model as those obtained in Step (1) and Step (2) of the above bagging algorithm, compute the predicted response vector and adjusted residual vector as

$$\widehat{\boldsymbol{Y}}^{(b)} = \mathbf{1}\widehat{\beta}_0^{(b)} + (\boldsymbol{R}_b\boldsymbol{X})\widehat{\boldsymbol{\beta}}^{(b)},$$

$$\widehat{\boldsymbol{e}}^{(b)} = \frac{\boldsymbol{R}_b\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^{(b)}}{\sqrt{1 - p^{(b)}/n}}, \text{ where } p^{(b)} \text{ is the size of selected linear model with intercept.}$$

Step (2) Draw one bootstrap sample from $\mathcal{L}^{(b)}$ by the residual-based bootstrap to produce one level-2 bootstrap sample $\mathcal{L}^{(b,1)} = (\boldsymbol{X}^{(b,1)}, \boldsymbol{Y}^{(b,1)})$, where $\boldsymbol{X}^{(b,1)} = \boldsymbol{R}_b\boldsymbol{X}$ and $\boldsymbol{Y}^{(b,1)} = \widehat{\boldsymbol{Y}}^{(b)} + \boldsymbol{R}_{b,1}\widehat{\boldsymbol{e}}^{(b)}$. Use $\mathcal{L}^{(b,1)}$ to obtain $(\widehat{\beta}_0^{(b,1)}, \{\widehat{\boldsymbol{\beta}}^{(b,1)}\}^T)^T$ and calculate the prediction, $\widehat{\gamma}_{b,1} = \widehat{\beta}_0^{(b,1)} + \boldsymbol{x}_{([0.25n])}^T\widehat{\boldsymbol{\beta}}^{(b,1)}$.

Step (3) Repeat the above two steps B times to have $\widehat{\gamma}_{b,1}$, for $b = 1, \cdots, B$. The VoD variance estimator is computed as

$$\widehat{V}_{VoD} = \frac{\sum_{b=1}^{B}\{(\widehat{\gamma}_{b,1} - \widehat{\gamma}_b) - (\overline{\widehat{\gamma}}_{\cdot,1} - \overline{\widehat{\gamma}}_{\cdot})\}^2}{B-1},$$

where $\widehat{\gamma}_b$, for $b = 1, \cdots, B$, are obtained in Step (2) of the bagging algorithm.

Parallel bootstrap DoV algorithm:

Step (1) Use the B predicted values resulting from the bagging algorithm to compute

the second component of DoV as

$$\widehat{V}_{DoV,2} = \left(1 - \frac{1}{B}\right) \frac{\sum_{b=1}^{B} (\widehat{\gamma}_b - \overline{\widehat{\gamma}}_\cdot)^2}{B-1}.$$

Step (2) Use the original data $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{Y})$ to obtain $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}^T)^T$. Compute the predicted response vector and adjusted residual vector as

$$\widehat{\boldsymbol{Y}} = \mathbf{1}\widehat{\beta}_0 + \boldsymbol{X}\widehat{\boldsymbol{\beta}},$$

$$\widehat{\boldsymbol{e}} = \frac{\boldsymbol{Y} - \widehat{\boldsymbol{Y}}}{\sqrt{1 - p/n}}, \text{ where } p \text{ is the size of selected linear model including intercept.}$$

Step (3) Draw one level-1 bootstrap sample from $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{Y})$ by the residual-based bootstrap to produce $\mathcal{L}^{(i)} = (\boldsymbol{X}, \boldsymbol{Y}^{(i)})$, where $\boldsymbol{Y}^{(i)} = \widehat{\boldsymbol{Y}} + \boldsymbol{R}_i \widehat{\boldsymbol{e}}$. Draw level-2 bootstrap sample from $\mathcal{L}^{(i)}$ by the random-pair bootstrap to have $\mathcal{L}^{(i,1)} = (\boldsymbol{X}^{(i,1)}, \boldsymbol{Y}^{(i,1)})$, where $\boldsymbol{X}^{(i,1)} = \boldsymbol{R}_{i,1}\boldsymbol{X}$ and $\boldsymbol{Y}^{(i,1)} = \boldsymbol{R}_{i,1}\boldsymbol{Y}^{(i)}$. Use $\mathcal{L}^{(i,1)}$ to obtain $(\widehat{\beta}_0^{(i,1)}, \{\widehat{\boldsymbol{\beta}}^{(i,1)}\}^T)^T$ and make a prediction with $\boldsymbol{x}_{([0.25n])}$, $\widehat{\gamma}_{i,1} = \widehat{\beta}_0^{(i,1)} + \boldsymbol{x}_{([0.25n])}^T \widehat{\boldsymbol{\beta}}^{(i,1)}$.

Step (4) Repeat Step (3) B times to have $\widehat{\gamma}_{i,1}$, for $i = 1, \cdots, B$. Compute the first component of DoV as

$$\widehat{V}_{DoV,1} = \left(\frac{n}{n-1}\right) \frac{\sum_{i=1}^{B} (\widehat{\gamma}_{i,1} - \overline{\widehat{\gamma}}_{\cdot,1})^2}{B-1}.$$

Step (5) The DoV variance estimator is computed as

$$\widehat{V}_{DoV} = \begin{cases} \widehat{V}_{DoV,1} - \widehat{V}_{DoV,2}\ , & \text{if } \widehat{V}_{DoV,1} > \widehat{V}_{DoV,2}. \\ 0, & \text{otherwise.} \end{cases}$$

Bootstrap-after-bootstrap (BaB) algorithm:

Step (1) Use the original data $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{Y})$ to obtain $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}^T)^T$. Compute the predicted

response vector and adjusted residual vector as

$$\widehat{\boldsymbol{Y}} = \mathbf{1}\widehat{\beta}_0 + \boldsymbol{X}\widehat{\boldsymbol{\beta}},$$

$$\widehat{\boldsymbol{e}} = \frac{\boldsymbol{Y} - \widehat{\boldsymbol{Y}}}{\sqrt{1 - p/n}}, \text{ where } p \text{ is the size of selected linear model including intercept.}$$

Step (2) Draw one bootstrap sample from $\mathcal{L}$ by the residual-based bootstrap to produce $\mathcal{L}^{(k)} = (\boldsymbol{X}, \boldsymbol{Y}^{(k)})$, where $\boldsymbol{Y}^{(k)} = \widehat{\boldsymbol{Y}} + \boldsymbol{R}_k\widehat{\boldsymbol{e}}$.

Step (3) Draw B_2 level-2 bootstrap samples from $\mathcal{L}^{(k)}$ by the random-pair bootstrap to have $\mathcal{L}^{(k,j)} = (\boldsymbol{X}^{(k,j)}, \boldsymbol{Y}^{(k,j)})$, where $\boldsymbol{X}^{(k,j)} = \boldsymbol{R}_{k,j}\boldsymbol{X}$ and $\boldsymbol{Y}^{(k,j)} = \boldsymbol{R}_{k,j}\boldsymbol{Y}^{(k)}$, $j = 1, \cdots, B_2$. Use $\mathcal{L}^{(k,j)}$ to obtain $(\widehat{\beta}_0^{(k,j)}, \{\widehat{\boldsymbol{\beta}}^{(k,j)}\}^T)^T$ and calculate the prediction, $\boldsymbol{x}_{([0.25n])}$, $\widehat{\gamma}_{k,j} = \widehat{\beta}_0^{(k,j)} + \boldsymbol{x}_{([0.25n])}^T\widehat{\boldsymbol{\beta}}^{(k,j)}$.

Step (4) Compute $\widehat{\gamma}_{k,BAG} = \sum_{j=1}^{B_2}\widehat{\gamma}_{k,j}/B_2$.

Step (5) Repeat step (2), (3) and (4) B_1 times to have $\widehat{\gamma}_{k,BAG}$, for $k = 1, \cdots, B_1$. Compute the BaB variance estimator as

$$\widehat{V}_{BaB} = \frac{\sum_{k=1}^{B_1}(\widehat{\gamma}_{k,BAG} - \overline{\widehat{\gamma}}_{\cdot,BAG})^2}{B_1 - 1}.$$

We choose level-1 and level-2 bootstrap sample sizes so that different methods have the same computational cost to calculate a bagging prediction and estimate its variance. In this study the common computational cost is 600 function calls to variable selection procedures. Notice that in estimating the variance for $\widehat{\gamma}_{BAG}$, the parallel bootstrap methods reuse the function calls that are needed to make predictions, but the bootstrap-after-bootstrap does not. Thus we choose different bootstrap sample sizes as follows. The level-1 and level-2 bootstrap sample size for the parallel bootstrap are both 300. For the bootstrap-after-bootstrap method, 300 bootstrap samples are used to make predictions and the level-1 and level-2 bootstrap sample sizes for variance estimator are 30 and 10

respectively.

4.3 Results of Simulation

We study the performance of three bootstrap-based variance estimators for the same bagging predictor by comparing their bias and variation. The bias of variance estimators is measured by the bias ratio (BR) defined in (4.2), and the variation is measured by the coefficient of variation (CV) defined in (4.3). $BR = 1$ means that the variance estimator has no bias. $BR > 1$ means the variance estimator tends to overestimate the true variance, and $BR < 1$ means the variance estimator tends to underestimate. A large value of CV indicates large relative variation of the variance estimator. We also check if the bagging technique improves prediction accuracy. For each of the variable selection methods, if the mean squared error (MSE) of bagging predictor is less than that of non-bagging predictor in Table 4.12, we note bagging had better prediction accuracy, even though the difference of corresponding MSE's need not be statistically significant.

We summarize the comparison of variance estimators in Tables 4.3 − 4.11 according to different regression vectors $\boldsymbol{\beta}_*$ and the MSE of different predictors in Table 4.12. In each table of variance estimators, there are two blocks, respectively, summarizing the bias ratio (BR) and coefficient of variation (CV) for the standard deviation estimates for bagging FFSR and ALASSO prediction at $\boldsymbol{x} = \boldsymbol{x}_{([0.25n])}$. For each variable selection method, BIP (bagging improves prediction) with value $*$ means the MSE entry of bagging predictor was smaller than that of the corresponding non-bagging predictor in Table 4.12.

4.4 Conclusions

Among the three bootstrap-based variance estimators for the bagging predictor, the parallel bootstrap VoD and the bootstrap-after-bootstrap work reasonably well. Usually the parallel bootstrap VoD slightly overestimates the true variance, and the bootstrap-after-

bootstrap underestimates the variance of the bagging predictor by a small amount. The parallel bootstrap DoV tends to severely underestimate the variance in many situations, due to the phenomenon that the parallel bootstrap DoV produces a high proportion of negative variance estimates and these are set to 0. These three variance estimators perform similarly on FFSR and ALASSO.

For the simulation of all 36 situations, the averaged bias ratio of the parallel bootstrap VoD, the parallel bootstrap DoV and the bootstrap-after-bootstrap are summarized in Table 4.1.

Table 4.1: Averaged Bias Ratio (BR) of standard deviation estimators for all models

	ALASSO	FFSR
VoD	1.07	1.08
DoV	0.77	0.77
BaB	0.97	1.00

Our simulation also indicates that underestimation of the bootstrap-after-bootstrap usually occurs when $R^2 = 0.3$ and $\rho = 0.2$. The average bias ratio of the bootstrap-after-bootstrap over all situations where $R^2 = 0.3$ and $\rho = 0.2$ were 0.88 for ALASSO and 0.93 for FFSR.

In terms of the coefficient of variation, the parallel bootstrap VoD was the best. The bootstrap-after-bootstrap has slightly larger variation than the parallel bootstrap VoD. In our simulation, the level-1 and level-2 bootstrap sample sizes are 30 and 10. Actually, when the level-1 and level-2 bootstrap sample size increase, the CV in the bootstrap-after-bootstrap variance estimator reduces. Large variation of the parallel bootstrap DoV can be partially explained by the high proportion of 0 variance estimates resulting from the negative estimates. The averaged coefficient of variation of the three variance estimators are summarized in Table 4.2.

Table 4.12 reveals the situations where bagging used with variable selection procedures

Table 4.2: Averaged Coefficient of Variation (CV) of standard deviation estimators for all models

	ALASSO	FFSR
VoD	0.10	0.10
DoV	0.44	0.51
BaB	0.17	0.18

improved prediction accuracy. Here when the MSE of a bagging predictor is less than that of the associated non-bagging predictor, we say bagging helps, even though their difference may not be statistically significant. Generally speaking, bagging is useful when the signal-to-noise ratio is low, particularly when the predictor variables have relative strong correlation. Bagging does not help much when the signal-to-noise ratio is large. In some situations bagging increased the MSE of predictors.

When bagging improves prediction accuracy, it usually produces larger reduction in MSE for FFSR than ALASSO. This is due to the fact that FFSR is based on the forward selection sequence, and thus FFSR is not as stable as ALASSO.

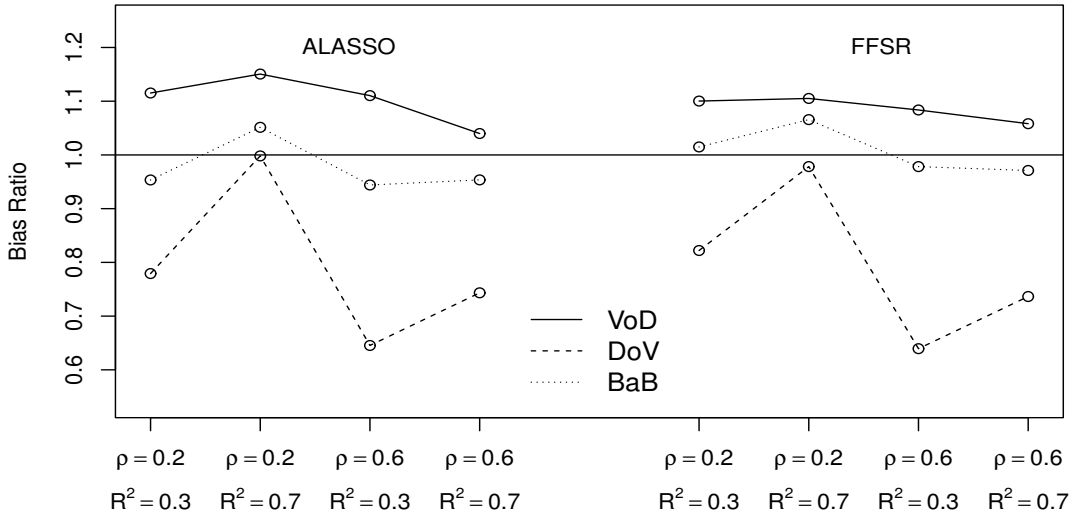

Figure 4.1: Bias Ratio (BR) plot of standard deviation estimators for bagging predictor $\widehat{\gamma}_{BAG}$ in model M1. VoD=Variance-of-Difference, DoV=Difference-of-Variances, BaB=Bootstrap-after-Bootstrap.

Table 4.3: Bias Ratio (BR) and Coefficient of Variation (CV) of standard deviation estimators for the bagging predictor $\widehat{\gamma}_{BAG}$ in model $M_1 : \boldsymbol{\beta}_* = (\mathbf{1}_6^T, \mathbf{0}_{20}^T)^T$

			ALASSO				FFSR			
	ρ	R^2	BIP	VoD	DoV	BaB	BIP	VoD	DoV	BaB
BR	0.2	0.3		1.12	0.78	0.95		1.10	0.82	1.01
	0.2	0.7		1.15	1.00	1.05		1.11	0.98	1.07
	0.6	0.3	*	1.11	0.65	0.94	*	1.08	0.64	0.98
	0.6	0.7	*	1.04	0.74	0.95	*	1.06	0.74	0.97
CV	0.2	0.3		0.10	0.43	0.16		0.11	0.45	0.16
	0.2	0.7		0.10	0.29	0.16		0.10	0.32	0.16
	0.6	0.3	*	0.12	0.54	0.17	*	0.12	0.66	0.20
	0.6	0.7	*	0.10	0.39	0.17	*	0.10	0.48	0.18

Note: the range of Monte Carlo standard errors for BR entries is [0.018, 0.026] and for CV entries is $[0.002, 0.019]$. An asterisk ($*$) in column BIP indicates that the MSE of bagging predictors in Table 4.12 is lower.

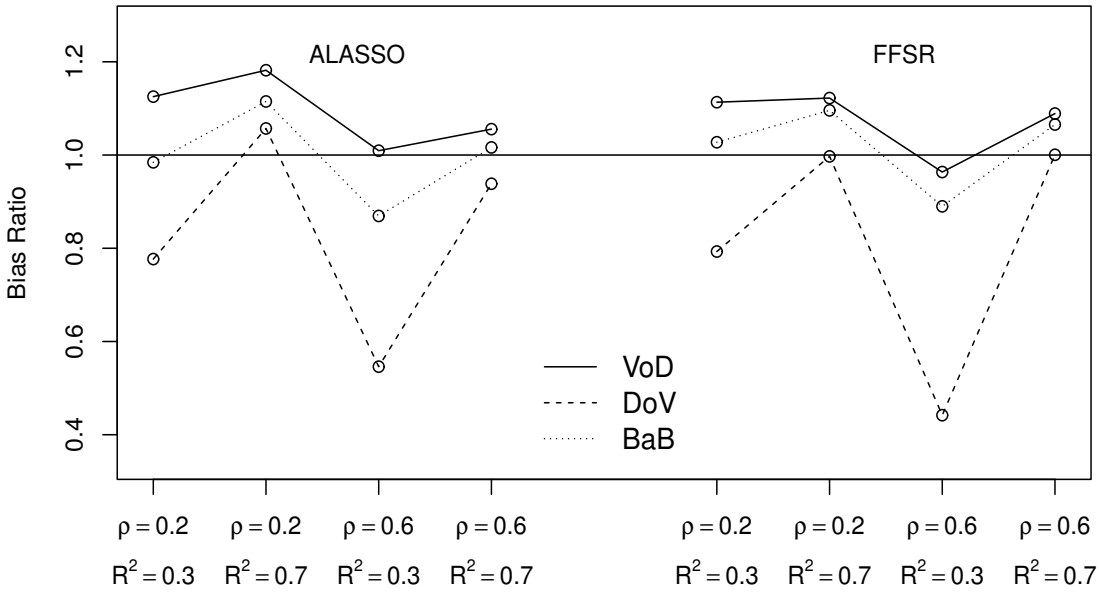

Figure 4.2: Bias Ratio (BR) plot of standard deviation estimators for bagging predictor $\widehat{\gamma}_{BAG}$ in model M2. VoD=Variance-of-Difference, DoV=Difference-of-Variances, BaB=Bootstrap-after-Bootstrap.

Table 4.4: Bias Ratio (BR) and Coefficient of Variation (CV) of standard deviation estimators for the bagging predictor $\widehat{\gamma}_{BAG}$ in model $M_2 : \boldsymbol{\beta}_* = (\mathbf{1}_2^T, \mathbf{0}_{10}^T, \mathbf{1}_2^T, \mathbf{0}_{10}^T, \mathbf{1}_2^T)^T$

			ALASSO				FFSR			
	ρ	R^2	BIP	VoD	DoV	BaB	BIP	VoD	DoV	BaB
BR	0.2	0.3		1.13	0.78	0.98	*	1.11	0.79	1.03
	0.2	0.7		1.18	1.06	1.11		1.12	1.00	1.10
	0.6	0.3	*	1.01	0.55	0.87	*	0.96	0.44	0.89
	0.6	0.7		1.06	0.94	1.02		1.09	1.00	1.07
CV	0.2	0.3		0.11	0.46	0.17	*	0.11	0.49	0.17
	0.2	0.7		0.11	0.30	0.16		0.10	0.34	0.16
	0.6	0.3	*	0.10	0.59	0.18	*	0.10	0.87	0.18
	0.6	0.7		0.08	0.24	0.16		0.10	0.25	0.16

Note: the range of Monte Carlo standard errors for BR entries is [0.014, 0.030] and for CV entries is $[0.002, 0.025]$. An asterisk ($*$) in column BIP indicates that the MSE of bagging predictors in Table 4.12 is lower.

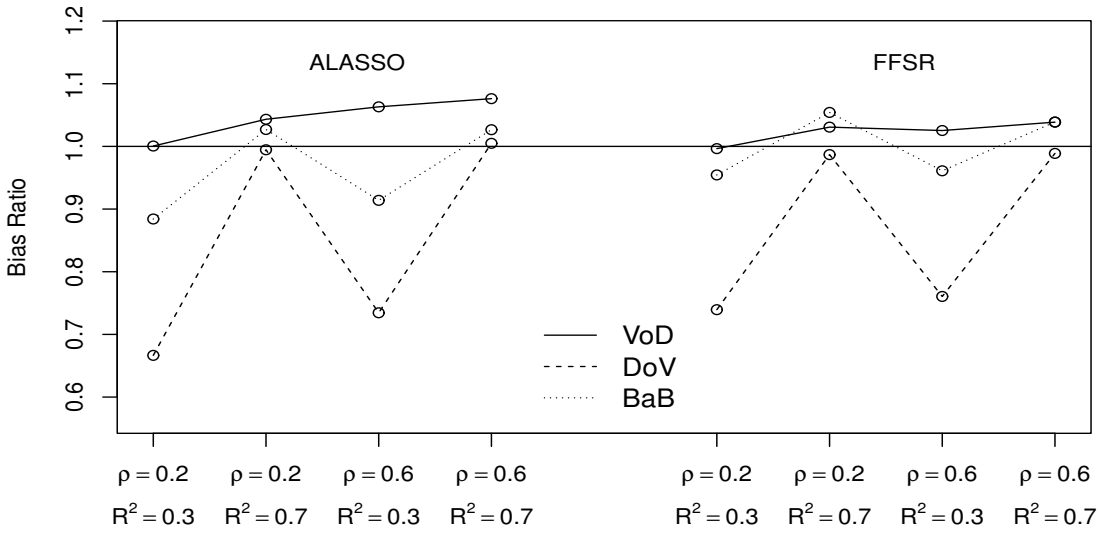

Figure 4.3: Bias Ratio (BR) plot of standard deviation estimators for bagging predictor $\widehat{\gamma}_{BAG}$ in model M3. VoD=Variance-of-Difference, DoV=Difference-of-Variances, BaB=Bootstrap-after-Bootstrap.

Table 4.5: Bias Ratio (BR) and Coefficient of Variation (CV) of standard deviation estimators for the bagging predictor $\widehat{\gamma}_{BAG}$ in model M_3 : $\boldsymbol{\beta}_* = (1, \mathbf{0}_4^T, 1, \mathbf{0}_4^T, 1, \mathbf{0}_4^T, 1, \mathbf{0}_4^T, 1, \mathbf{0}_4^T, 1)^T$

			ALASSO				FFSR			
	ρ	R^2	BIP	VoD	DoV	BaB	BIP	VoD	DoV	BaB
BR	0.2	0.3	*	1.00	0.67	0.88	*	1.00	0.74	0.95
	0.2	0.7		1.04	0.99	1.03		1.03	0.99	1.05
	0.6	0.3	*	1.06	0.73	0.91	*	1.03	0.76	0.96
	0.6	0.7		1.08	1.00	1.03		1.04	0.99	1.04
CV	0.2	0.3	*	0.10	0.49	0.18	*	0.10	0.46	0.18
	0.2	0.7		0.09	0.21	0.15		0.08	0.23	0.15
	0.6	0.3	*	0.09	0.40	0.18	*	0.09	0.40	0.17
	0.6	0.7		0.09	0.20	0.15		0.09	0.22	0.15

Note: the range of Monte Carlo standard errors for BR entries is [0.018, 0.025] and for CV entries is $[0.002, 0.016]$. An asterisk ($*$) in column BIP indicates that the MSE of bagging predictors in Table 4.12 is lower.

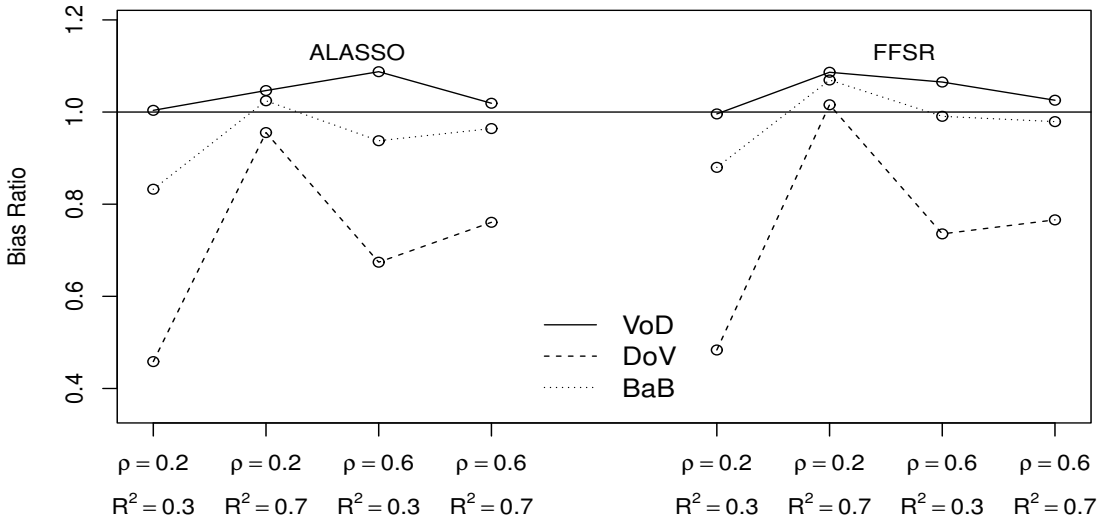

Figure 4.4: Bias Ratio (BR) plot of standard deviation estimators for bagging predictor $\widehat{\gamma}_{BAG}$ in model M4. VoD=Variance-of-Difference, DoV=Difference-of-Variances, BaB=Bootstrap-after-Bootstrap.

Table 4.6: Bias Ratio (BR) and Coefficient of Variation (CV) of standard deviation estimators for the bagging predictor $\widehat{\gamma}_{BAG}$ in model $M_4 : \boldsymbol{\beta}_* = (\mathbf{1}_{12}^T, \mathbf{0}_{14}^T)^T$

			ALASSO				FFSR			
	ρ	R^2	BIP	VoD	DoV	BaB	BIP	VoD	DoV	BaB
BR	0.2	0.3	*	1.00	0.46	0.83	*	1.00	0.48	0.88
	0.2	0.7		1.05	0.96	1.02		1.09	1.02	1.07
	0.6	0.3	*	1.09	0.67	0.94	*	1.07	0.74	0.99
	0.6	0.7	*	1.02	0.76	0.96	*	1.03	0.77	0.98
CV	0.2	0.3	*	0.13	0.79	0.24	*	0.12	0.84	0.24
	0.2	0.7		0.08	0.23	0.15		0.09	0.23	0.15
	0.6	0.3	*	0.10	0.50	0.18	*	0.11	0.50	0.19
	0.6	0.7	*	0.08	0.35	0.15	*	0.08	0.39	0.16

Note: the range of Monte Carlo standard errors for BR entries is [0.013, 0.025] and for CV entries is $[0.002, 0.024]$. An asterisk $(*)$ in column BIP indicates that the MSE of bagging predictors in Table 4.12 is lower.

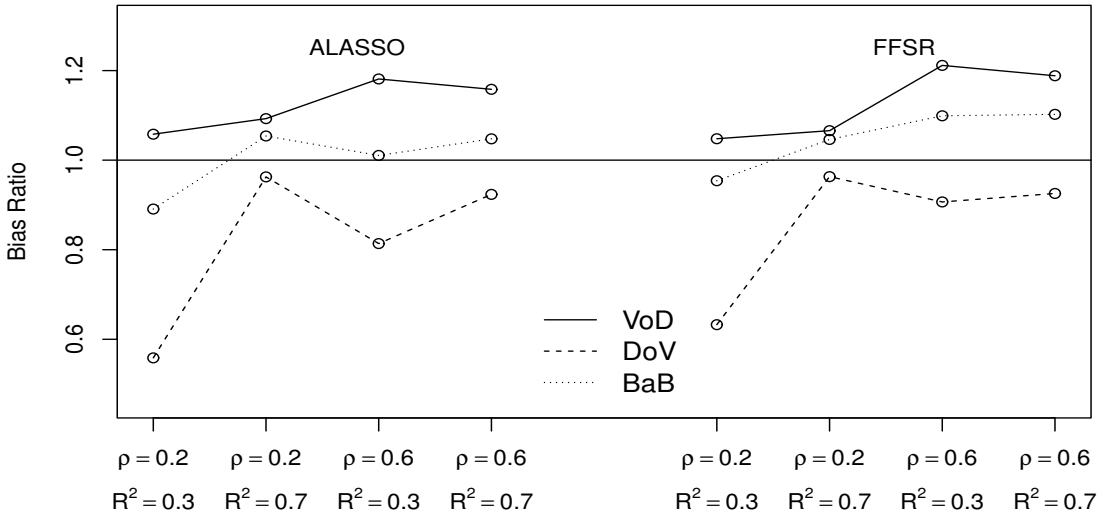

Figure 4.5: Bias Ratio (BR) plot of standard deviation estimators for bagging predictor $\widehat{\gamma}_{BAG}$ in model M5. VoD=Variance-of-Difference, DoV=Difference-of-Variances, BaB=Bootstrap-after-Bootstrap.

Table 4.7: Bias Ratio (BR) and Coefficient of Variation (CV) of standard deviation estimators for the bagging predictor $\widehat{\gamma}_{BAG}$ in model $M_5 : \boldsymbol{\beta}_* = (\mathbf{1}_4^T, \mathbf{0}_7^T, \mathbf{1}_4^T, \mathbf{0}_7^T, \mathbf{1}_4^T)^T$

			ALASSO				FFSR			
	ρ	R^2	BIP	VoD	DoV	BaB	BIP	VoD	DoV	BaB
BR	0.2	0.3	*	1.06	0.56	0.89	*	1.05	0.63	0.95
	0.2	0.7		1.09	0.96	1.05		1.07	0.96	1.05
	0.6	0.3	*	1.18	0.81	1.01	*	1.21	0.91	1.10
	0.6	0.7	*	1.16	0.92	1.05	*	1.19	0.93	1.10
CV	0.2	0.3	*	0.11	0.63	0.19	*	0.11	0.62	0.19
	0.2	0.7		0.09	0.25	0.14		0.09	0.25	0.16
	0.6	0.3	*	0.11	0.42	0.16	*	0.12	0.43	0.17
	0.6	0.7	*	0.10	0.30	0.15	*	0.10	0.36	0.17

Note: the range of Monte Carlo standard errors for BR entries is [0.018, 0.029] and for CV entries is $[0.002, 0.019]$. An asterisk $(*)$ in column BIP indicates that the MSE of bagging predictors in Table 4.12 is lower.

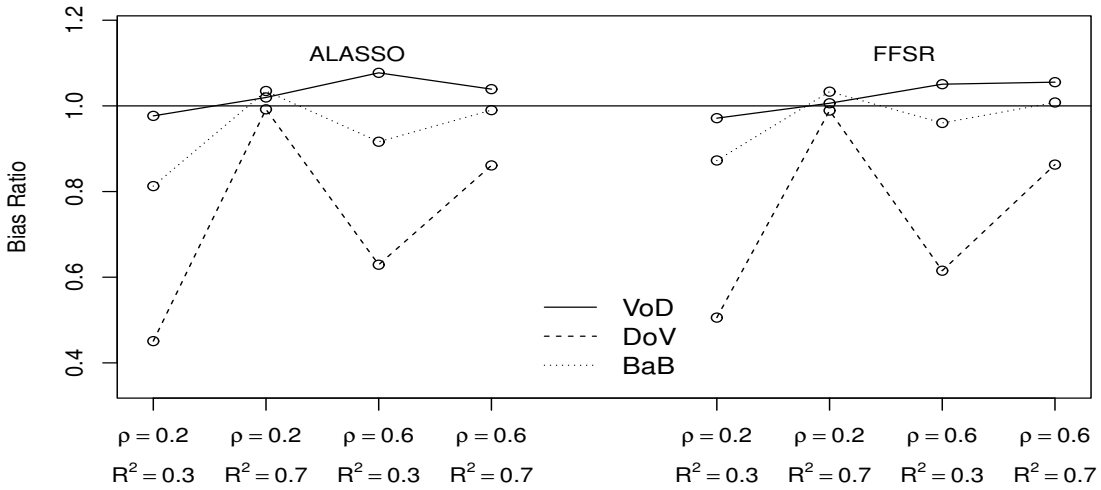

Figure 4.6: Bias Ratio (BR) plot of standard deviation estimators for bagging predictor $\widehat{\gamma}_{BAG}$ in model M6. VoD=Variance-of-Difference, DoV=Difference-of-Variances, BaB=Bootstrap-after-Bootstrap.

Table 4.8: Bias Ratio (BR) and Coefficient of Variation (CV) of standard deviation estimators for the bagging predictor $\widehat{\gamma}_{BAG}$ in model M_6 : $\boldsymbol{\beta}_* = (1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,\ 0,0,0)^T$

			ALASSO				FFSR			
	ρ	R^2	BIP	VoD	DoV	BaB	BIP	VoD	DoV	BaB
BR	0.2	0.3	*	0.98	0.45	0.81	*	0.97	0.51	0.87
	0.2	0.7		1.02	0.99	1.03		1.01	0.99	1.03
	0.6	0.3	*	1.08	0.63	0.92	*	1.05	0.62	0.96
	0.6	0.7	*	1.04	0.86	0.99	*	1.06	0.86	1.01
CV	0.2	0.3	*	0.11	0.78	0.22	*	0.11	0.79	0.23
	0.2	0.7		0.08	0.16	0.15		0.09	0.18	0.15
	0.6	0.3	*	0.09	0.53	0.18	*	0.10	0.56	0.18
	0.6	0.7	*	0.08	0.27	0.15	*	0.09	0.29	0.15

Note: the range of Monte Carlo standard errors for BR entries is [0.013, 0.024] and for CV entries is $[0.002, 0.023]$. An asterisk ($*$) in column BIP indicates that the MSE of bagging predictors in Table 4.12 is lower.

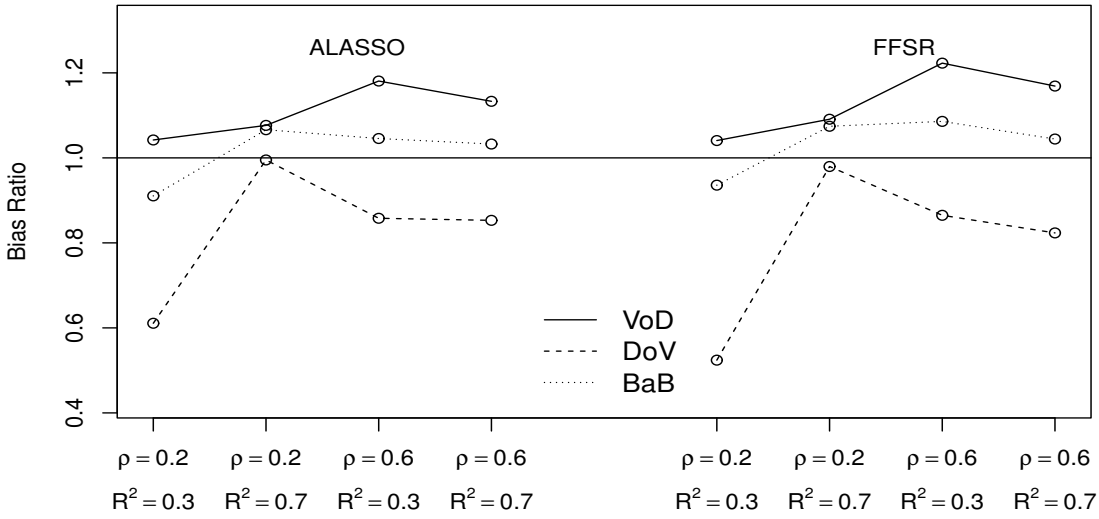

Figure 4.7: Bias Ratio (BR) plot of standard deviation estimators for bagging predictor $\widehat{\gamma}_{BAG}$ in model M7. VoD=Variance-of-Difference, DoV=Difference-of-Variances, BaB=Bootstrap-after-Bootstrap.

Table 4.9: Bias Ratio (BR) and Coefficient of Variation (CV) of standard deviation estimators for the bagging predictor $\widehat{\gamma}_{BAG}$ in model $M_7 : \boldsymbol{\beta}_* = (\mathbf{1}_8^T, \mathbf{0}_{40}^T)^T$

			ALASSO				FFSR			
	ρ	R^2	BIP	VoD	DoV	BaB	BIP	VoD	DoV	BaB
BR	0.2	0.3	*	1.04	0.61	0.91	*	1.04	0.52	0.94
	0.2	0.7		1.08	1.00	1.07		1.09	0.98	1.07
	0.6	0.3	*	1.18	0.86	1.05		1.22	0.86	1.09
	0.6	0.7		1.13	0.85	1.03	*	1.17	0.82	1.04
CV	0.2	0.3	*	0.10	0.59	0.18	*	0.11	0.79	0.19
	0.2	0.7		0.09	0.29	0.15		0.10	0.34	0.16
	0.6	0.3	*	0.11	0.45	0.16		0.13	0.55	0.17
	0.6	0.7		0.10	0.42	0.16	*	0.11	0.51	0.17

Note: the range of Monte Carlo standard errors for BR entries is [0.017, 0.029] and for CV entries is $[0.002, 0.022]$. An asterisk $(*)$ in column BIP indicates that the MSE of bagging predictors in Table 4.12 is lower.

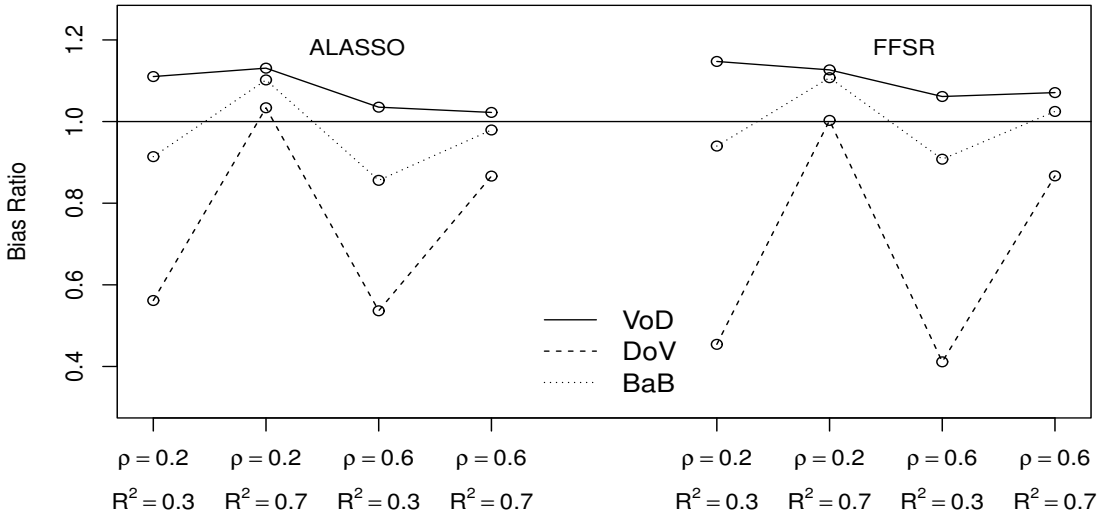

Figure 4.8: Bias Ratio (BR) plot of standard deviation estimators for bagging predictor $\widehat{\gamma}_{BAG}$ in model M8. VoD=Variance-of-Difference, DoV=Difference-of-Variances, BaB=Bootstrap-after-Bootstrap.

Table 4.10: Bias Ratio (BR) and Coefficient of Variation (CV) of standard deviation estimators for the bagging predictor $\widehat{\gamma}_{BAG}$ in model $M_8 : \boldsymbol{\beta}_* = ((1,0,0,0)_{12})^T$

			ALASSO				FFSR			
	ρ	R^2	BIP	VoD	DoV	BaB	BIP	VoD	DoV	BaB
BR	0.2	0.3	*	1.11	0.56	0.91	*	1.15	0.45	0.94
	0.2	0.7		1.13	1.03	1.10		1.13	1.00	1.11
	0.6	0.3	*	1.04	0.54	0.86	*	1.06	0.41	0.91
	0.6	0.7	*	1.02	0.87	0.98	*	1.07	0.87	1.02
CV	0.2	0.3	*	0.10	0.70	0.17	*	0.10	0.97	0.21
	0.2	0.7		0.09	0.33	0.15		0.10	0.36	0.15
	0.6	0.3	*	0.09	0.64	0.17	*	0.11	0.93	0.21
	0.6	0.7	*	0.08	0.33	0.14	*	0.10	0.34	0.15

Note: the range of Monte Carlo standard errors for BR entries is [0.015, 0.026] and for CV entries is $[0.002, 0.027]$. An asterisk $(*)$ in column BIP indicates that the MSE of bagging predictors in Table 4.12 is lower.

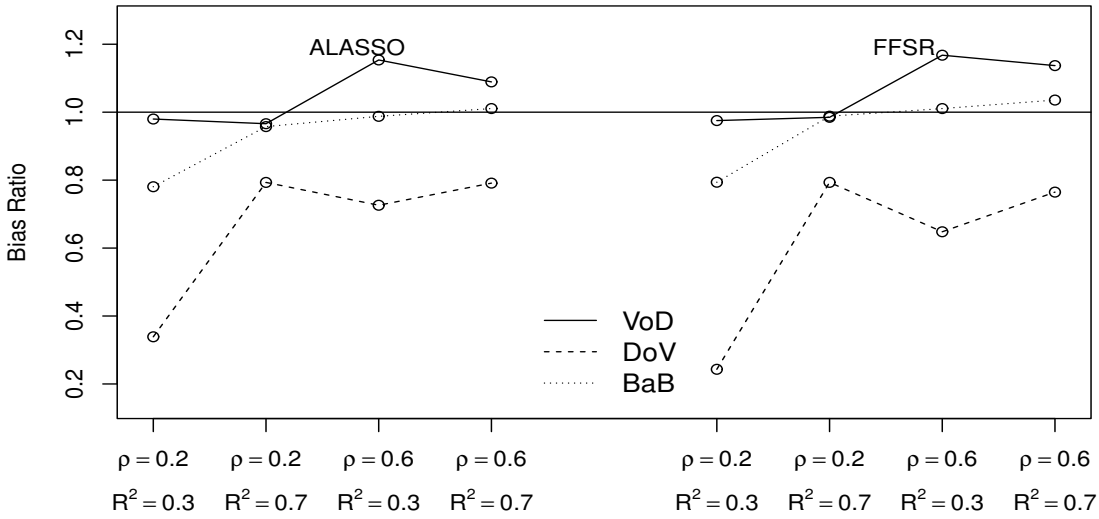

Figure 4.9: Bias Ratio (BR) plot of standard deviation estimators for bagging predictor $\widehat{\gamma}_{BAG}$ in model M9. VoD=Variance-of-Difference, DoV=Difference-of-Variances, BaB=Bootstrap-after-Bootstrap.

Table 4.11: Bias Ratio (BR) and Coefficient of Variation (CV) of standard deviation estimators for the bagging predictor $\widehat{\gamma}_{BAG}$ in model $M_9 : \boldsymbol{\beta}_* = (\mathbf{1}_6^T, \mathbf{0}_{15}^T, \mathbf{1}_6^T, \mathbf{0}_{15}^T, \mathbf{1}_6^T)^T$

			ALASSO				FFSR			
	ρ	R^2	BIP	VoD	DoV	BaB	BIP	VoD	DoV	BaB
BR	0.2	0.3	*	0.98	0.34	0.78	*	0.98	0.24	0.79
	0.2	0.7	*	0.97	0.79	0.96	*	0.98	0.79	0.99
	0.6	0.3	*	1.15	0.73	0.99	*	1.17	0.65	1.01
	0.6	0.7	*	1.09	0.79	1.01	*	1.14	0.76	1.04
CV	0.2	0.3	*	0.11	0.98	0.21	*	0.13	1.43	0.27
	0.2	0.7	*	0.08	0.37	0.15	*	0.10	0.40	0.15
	0.6	0.3	*	0.10	0.52	0.17	*	0.11	0.65	0.16
	0.6	0.7	*	0.09	0.40	0.15	*	0.09	0.48	0.15

Note: the range of Monte Carlo standard errors for BR entries is [0.011, 0.027] and for CV entries is $[0.002, 0.027]$. An asterisk ($*$) in column BIP indicates that the MSE of bagging predictors in Table 4.12 is lower.

Table 4.12: Monte Carlo estimates of the Mean Squared Error (MSE) of the bagging predictor (BAG) and non-bagging predictor (NBAG) for all models

		$\rho = 0.2$ $R^2 = 0.3$		$\rho = 0.2$ $R^2 = 0.7$		$\rho = 0.6$ $R^2 = 0.3$		$\rho = 0.6$ $R^2 = 0.7$	
		NBAG	BAG	NBAG	BAG	NBAG	BAG	NBAG	BAG
	ALASSO	0.60	0.62	0.38	0.51	0.81	0.54	1.08	0.86
M1	FFSR	0.65	0.69	0.45	0.71	1.07	0.59	1.30	0.89
	ALASSO	0.87	0.90	0.55	0.73	1.62	1.22	1.31	1.47
M2	FFSR	1.20	1.06	0.69	1.02	2.54	1.31	1.05	1.31
	ALASSO	1.41	1.25	0.95	1.05	1.52	1.11	1.04	1.09
M3	FFSR	1.97	1.46	0.93	1.11	2.16	1.38	0.97	1.16
	ALASSO	1.58	1.13	1.43	1.45	1.45	1.11	1.96	1.65
M4	FFSR	2.57	1.34	1.26	1.39	1.71	1.17	2.11	1.64
	ALASSO	0.88	0.69	0.63	0.74	0.51	0.40	0.45	0.44
M5	FFSR	1.09	0.76	0.71	0.90	0.36	0.30	0.42	0.39
	ALASSO	1.45	1.00	0.99	1.06	1.41	0.96	1.63	1.49
M6	FFSR	2.14	1.13	0.87	1.02	2.22	1.10	1.68	1.53
	ALASSO	1.76	1.54	1.34	1.58	0.71	0.83	0.79	0.97
M7	FFSR	2.83	1.68	1.12	1.60	0.63	0.63	0.96	0.89
	ALASSO	1.92	1.89	1.62	2.05	2.37	1.94	2.39	2.67
M8	FFSR	2.97	2.05	1.53	2.24	3.93	2.27	1.98	2.56
	ALASSO	2.75	2.08	3.84	3.50	1.06	0.68	1.40	1.00
M9	FFSR	4.49	2.42	4.10	3.58	1.20	0.62	1.53	0.90

Note: Based on $1,000$ Monte Carlo replications, the relative standard errors of MSE (SE/MSE) are in [0.032,0.065].

Chapter 5

Additional Simulation Results

We demonstrate by simulation that the jackknife-after-bootstrap variance estimator does not effectively estimate the variance of bagging predictors with the same computational cost. In addition, we give simulation results to compare the performance of the residual-based bootstrap and the traditional random-pair bootstrap in computing the variance of non-bagging predictors.

5.1 Jackknife-after-bootstrap

The jackknife-after-bootstrap is always biased too high and is not competitive with the proposed parallel bootstrap VoD and the bootstrap-after-bootstrap. We illustrate this in the same linear regression settings as model $M1$, $M4$ and $M8$, specified in Chapter 4. As in these models, the prediction of interest is $\boldsymbol{\mu}_{([0.25n])}$, the first quartile of true means of the response vector. The corresponding co-variate vector is $\boldsymbol{x}_{([0.25n])}$.

In Chapter 4 the total computational cost used to compute the bagging predicted value and its variance is 600 functional calls. Thus, the bootstrap sample size B is chosen to be 600 so that the total cost of jackknife after bootstrap is equal to that of the simulations in Chapter 4. The Monte Carlo replication size is $n_{sim} = 1000$. The predictive models are chosen by the variable selection procedures ALASSO and FFSR.

In the following algorithm of jackknife-after-bootstrap, $\boldsymbol{R}$ denotes a bootstrap selection matrix and $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{Y})$ denotes the original sample data. Using the bth bootstrap sample $\mathcal{L}^{(b)}$, the estimated intercept and regression coefficient vector $(\widehat{\beta}_0^{(b)}, \{\widehat{\boldsymbol{\beta}}^{(b)}\}^T)^T$ of the linear model are obtained by either FFSR or ALASSO.

Jackknife-after-bootstrap (JaB) algorithm:

Step (1) Draw the bth bootstrap sample from the original data set $\mathcal{L}$ by the random-pair bootstrap to obtain $\mathcal{L}^{(b)} = (\boldsymbol{R}_b\boldsymbol{X}, \boldsymbol{R}_b\boldsymbol{Y})$.

Step (2) Use $\mathcal{L}^{(b)}$ to obtain $(\widehat{\beta}_0^{(b)}, \{\widehat{\boldsymbol{\beta}}^{(b)}\}^T)^T$ and calculate $\widehat{\gamma}_b = \widehat{\beta}_0^{(b)} + \boldsymbol{x}^T_{([0.25n])}\widehat{\boldsymbol{\beta}}^{(b)}$.

Step (3) Repeat Step (1) and Step (2) B times to have B predicted values, $\widehat{\gamma}_1, \cdots, \widehat{\gamma}_B$, and average them to result in the bagging predictor $\widehat{\gamma}_{BAG} = \sum_{b=1}^{B} \widehat{\gamma}_b/B$.

Step (4) Compute the average of the predicted values from the bootstrap samples without the ith observation of $\mathcal{L}$ to have

$$\widehat{\gamma}_{B,[i]} = \frac{\sum_{b=1}^{B} \widehat{\gamma}_b \boldsymbol{I}(\boldsymbol{x}_i \notin \mathcal{L}^{(b)})}{\sum_{b=1}^{B} \boldsymbol{I}(\boldsymbol{x}_i \notin \mathcal{L}^{(b)})},$$

where $\boldsymbol{I}(\boldsymbol{x}_i \notin \mathcal{L}^{(b)})$ means the bth bootstrap sample does not include the ith observation.

Step (5) Repeat Step (4) for $i = 1, 2, \cdots, n$ to produce $\widehat{\gamma}_{B,[1]}, \cdots, \widehat{\gamma}_{B,[n]}$.

Step (6) For each observation in the original sample data, compute the influence function for the ith observation to obtain

$$\widetilde{u}_i = (n-1)\left(\frac{\sum_{j=1}^{n} \widehat{\gamma}_{B,[j]}}{n} - \widehat{\gamma}_{B,[i]}\right), \quad i = 1, \ldots, n.$$

Step (7) Compute the jackknife-after-bootstrap variance estimator for $\widehat{\gamma}_{BAG}$ as

$$\widehat{V}_{JaB} = \frac{\sum_{i=1}^{n} \widetilde{u}_i^2}{n(n-1)}.$$

The performance of the jackknife-after-bootstrap variance estimator in models $M1$, $M4$ and $M8$ are summarized in Table 5.1. The bias of the variance estimator is of primary interest and only the bias ratio's (BR) are included.

Table 5.1: Bias Ratio (BR) of the jackknife-after-bootstrap standard deviation estimator for the bagging predictor $\widehat{\gamma}_{BAG}$ in models M1, M4 and M8

	$\rho = 0.2$ $R^2 = 0.3$		$\rho = 0.2$ $R^2 = 0.7$		$\rho = 0.6$ $R^2 = 0.3$		$\rho = 0.6$ $R^2 = 0.7$	
	ALASSO	FFSR	ALASSO	FFSR	ALASSO	FFSR	ALASSO	FFSR
M1	1.34	1.42	1.33	1.38	1.40	1.46	1.34	1.38
M4	1.34	1.41	1.31	1.36	1.34	1.40	1.31	1.34
M8	1.50	1.61	1.44	1.48	1.36	1.47	1.34	1.40

Note: The range of Monte Carlo standard errors for BR entries is $[0.029, 0.038]$.

Comparing Table 5.1 with Tables 4.3, 4.6 and 4.10, we can see that the performance of the jackknife-after-bootstrap variance estimator is not as good as that of the parallel bootstrap VoD, or the bootstrap-after-bootstrap with the same computational cost. The jackknife-after-bootstrap tends to severely overestimate the true variance of the bagging predictor in models $M1$, $M4$ and $M8$.

5.2 Bootstrap-based Variance Estimators for a Non-Bagging Predictor

The random-pair bootstrap and the residual-based bootstrap are available to estimate the variance of a non-bagging predictor. When variable selection procedure LASSO is used to determine the predictive model, the random-pair bootstrap does not estimate the variance of a non-bagging predictor consistently, as pointed out by Knight and Fu (2000). In contrast, Chatterjee and Lahiri (2010) show that the residual-based bootstrap consistently estimates the variance of non-bagging predictor when the ALASSO is used

to determine the predictive model. When variable selection is performed by FFSR, no research has been done on the performance of these bootstrap methods in estimating the variance. This section briefly compares these two bootstrap variance estimators by simulation.

Using the same notation as in the algorithm of jackknife-after-bootstrap, the algorithms of random-pair bootstrap variance estimator and residual-based bootstrap variance estimator are given as follows.

Random-pair bootstrap variance estimator algorithm:

Step (1) Draw the bth bootstrap sample from the original data set $\mathcal{L}$ by the random-pair bootstrap to obtain $\mathcal{L}^{(b)} = (\boldsymbol{R}_b\boldsymbol{X}, \boldsymbol{R}_b\boldsymbol{Y})$.

Step (2) Use $\mathcal{L}^{(b)}$ to obtain $(\widehat{\beta}_0^{(b)}, \{\widehat{\boldsymbol{\beta}}^{(b)}\}^T)^T$ and calculate $\widehat{\gamma}_b = \widehat{\beta}_0^{(b)} + \boldsymbol{x}_{([0.25n])}^T\widehat{\boldsymbol{\beta}}^{(b)}$.

Step (3) Repeat Step (1) and Step (2) B times to have B predicted values, $\widehat{\gamma}_1, \cdots, \widehat{\gamma}_B$.

Step (4) Compute the variance of non-bagging predictor as the sample variance of $\widehat{\gamma}_1, \cdots, \widehat{\gamma}_B$,

$$\widehat{\mathrm{Var}}(\widehat{\gamma}) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\gamma}_b - \overline{\widehat{\gamma}}_{\cdot})^2.$$

Residual-based bootstrap variance estimator algorithm:

Step (1) Use the original data $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{Y})$ to obtain $(\widehat{\beta}_0, \{\widehat{\boldsymbol{\beta}}\}^T)^T$. Compute the predicted response vector and adjusted residual vector as

$$\widehat{\boldsymbol{Y}} = \mathbf{1}\widehat{\beta}_0 + \boldsymbol{X}\widehat{\boldsymbol{\beta}},$$

$$\widehat{\boldsymbol{e}} = \frac{\boldsymbol{Y} - \widehat{\boldsymbol{Y}}}{\sqrt{1-p/n}},$$ where p is the size of selected linear model including the intercept.

Step (2) Draw the bth bootstrap sample from $\mathcal{L}$ by the residual-based bootstrap to obtain $\mathcal{L}^{(b)} = (\boldsymbol{X}, \boldsymbol{Y}^{(b)})$, where $\boldsymbol{Y}^{(b)} = \widehat{\boldsymbol{Y}} + \boldsymbol{R}_b\widehat{\boldsymbol{e}}$.

Step (3) Use $\mathcal{L}^{(b)}$ to obtain $(\widehat{\beta}_0^{(b)}, \{\widehat{\boldsymbol{\beta}}^{(b)}\}^T)^T$ and calculate $\widehat{\gamma}_b = \widehat{\beta}_0^{(b)} + \boldsymbol{x}_{([0.25n])}^T \widehat{\boldsymbol{\beta}}^{(b)}$.

Step (4) Repeat Step (2) and Step (3) B times to have B predicted values, $\widehat{\gamma}_1, \cdots, \widehat{\gamma}_B$.

Step (5) Compute the variance of non-bagging predictor as the sample variance of $\widehat{\gamma}_1, \cdots, \widehat{\gamma}_B$,

$$\widehat{\mathrm{Var}}(\widehat{\gamma}) = \frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\gamma}_b - \overline{\widehat{\gamma}}_\cdot)^2.$$

The comparison of random-pair bootstrap and residual-based bootstrap in estimating the variance of a non-bagging predictor was conducted in the same settings as in $M1$, $M4$ and $M8$ specified in Chapter 4. The bootstrap sample size $B = 300$. The simulation results are summarized in Table 5.2.

Table 5.2: Bias Ratio (BR) of random-pair bootstrap (RP) and residual-based bootstrap (RB) standard deviation estimators for the non-bagging predictor $\widehat{\gamma}$ in M1, M4 and M8

		$\rho = 0.2$ $R^2 = 0.3$		$\rho = 0.2$ $R^2 = 0.7$		$\rho = 0.6$ $R^2 = 0.3$		$\rho = 0.6$ $R^2 = 0.7$	
		RP	RB	RP	RB	RP	RB	RP	RB
n=150	ALASSO	1.32	0.86	1.44	1.02	1.08	0.72	1.05	0.85
M1	FFSR	1.46	0.88	1.67	1.05	1.08	0.67	1.07	0.80
n=150	ALASSO	1.04	0.67	1.11	1.02	1.09	0.77	1.04	0.84
M4	FFSR	0.97	0.64	1.24	1.07	1.11	0.77	1.09	0.81
n=150	ALASSO	1.47	0.67	1.45	1.09	1.17	0.62	1.17	0.96
M8	FFSR	1.39	0.68	1.69	1.11	0.99	0.63	1.43	1.03
n=600	ALASSO	1.18	1.04	1.17	1.05	1.06	0.96	1.15	1.07
M8	FFSR	1.48	1.18	1.47	1.10	1.20	1.04	1.33	1.08

Note: Based on $1,000$ Monte Carlo replications, the range of standard errors for the bias ratio (BR) entries is in [0.013,0.046]. The bootstrap sample size is 300.

Contrary to expectation, it seems that our simulation does not indicate one bootstrap variance estimator is universally better than the other one for finite sample. However, when we increase the sample size from 150 to 600 (the last row of Table 5.2), the bias

ratio of the residual-based bootstrap for ALASSO estimator is much closer to 1 and is consistent with the conclusion of Chatterjee and Lahiri (2010). When the correlation among covariates is low and the signal-to-noise ratio is high, the residual-based bootstrap (RB) is always better than the random-pair bootstrap (RP). When the predictor variables have high correlation, RP has better performance than RB except for $M8$. When the predictor variables have weak correlation and the signal-to-noise ratio is small, neither bootstrap methods effectively estimate the variance of the non-bagging predictor.

Chapter 6

Application

Numerical simulation in Chapter 4 demonstrated that the parallel bootstrap VoD and the bootstrap-after-bootstrap (BaB) estimate the variance of a bagging predictor reasonably well. In this chapter we illustrate the performance of the parallel bootstrap VoD method and the BaB method in the NCAA data, taken from from Mangold, Bean, Adams (2003).

6.1 NCAA Data

The NCAA data are taken from the 1996-1999 editions of the US News *Best Colleges in America* and from the US Department of Education's Integrated Postsecondary Educational Data System. The data set consists of 94 observations, which correspond to 94 NCAA Division 1A schools. Each observation has 19 predictor variables summarizing the different characteristics of each institution. Mangold, Bean, Adams (2003) investigated the effect of the successful sports program on the graduation rate. The response variable Y and the 19 predictor variables $x_1, \cdots, x_{19}$ are defined in the following:

Y=Average 6 year graduation rate for 1996, 1997, 1998

x_1=top10: Percentage of admitted students in the top 10 percent high school

x_2=act25: The 25th percentile ACT composite score

x_3=oncampus: Percentage of students living on campus

x_4=ft_pct: Percentage of first-time freshmen among total enrollment

x_5=size: Total Enrollment in the unit of 1000

x_6=ta_pct: Percentage of courses taught by TAs

x_7=bbrank: Composite of basketball ranking

x_8=tuition: In-state tuition in \$1000

x_9=board: Room and board in \$1000

x_{10}=bbattend: Average basketball game home attendance

x_{11}=full_sal: Full professor salary

x_{12}=sf_ratio: Student to faculty ratio

x_{13}=white: Percentage of white students

x_{14}=ast_sal: Assistant professor salary

x_{15}=pop: Population of the city where located

x_{16}=prof_phd: Percentage of faculty with PHD degree

x_{17}=ad_rate: Acceptance rate

x_{18}=loan_pct: Percentage of students receiving loans

x_{19}=outstate_pct: Percentage of out of state students

We are interested in predicting the 6 year graduation rate of a NCAA 1A university, say North Carolina State University (NCSU). We use ALASSO and FFSR variable selection methods to build the predictive model with the observation of NCSU deleted. Bagging is used to enhance the prediction stability. The associated standard deviation is estimated by the parallel bootstrap VoD method and the modified bootstrap-after-bootstrap method.

The predictor vector for NCSU is $\boldsymbol{x} = (34, 21, 35, 17, 28, 7, -0.80431, 2.27, 5.65, 10800, 79.9, 15, 82, 50.6, 264000, 89, 69, 49, 8)^T$ and the average observed graduation rate= 66%. The bagging estimated 6 year graduation rate of NCSU with its standard deviation estimation is summarized in Table 6.1. The bagging bootstrap sample size is 300. The level-1 and level-2 bootstrap sample sizes for the bootstrap-after-bootstrap are 30 and 10. In

comparison, the predicted non-bagging graduation rate of NCSU with its standard error is summarized in Table 6.2. The standard errors for FFSR and ALASSO are calculated by using the residual-based bootstrap with bootstrap sample size 300. Table 6.3 summarizes the estimated coefficients when bagging is used and the estimated coefficients when bagging is not used.

Table 6.1: Bagging predicted NCSU graduation rate (observed value=66%) with the estimated standard deviation

	Bagging Prediction (%)	VoD	BaB
FFSR	62.1	2.33	1.76
ALASSO	62.1	2.52	1.92

Table 6.2: Non-bagging predicted NCSU graduation rate (observed value=66%) with estimated standard deviation from residual-based bootstrap

	Non-bagging Prediction (%)	Standard Error
Full Model	64.4	2.81
FFSR	60.4	2.03
ALASSO	61.5	1.78

Table 6.3: Estimated Coefficients for NCAA Selected Variables

	FFSR	ALASSO	Bagging FFSR	Bagging ALASSO
x_1=top10	0	0	0.02	0.02
x_2=act25	3.49	3.42	3.09	2.99
x_3=oncompus	0.24	0.24	0.20	0.21
x_4=ft_pct	0.68	0.66	0.79	0.73
x_5=size	0.27	0.19	0.21	0.19
x_6=ta_pct	0	0.07	0.10	0.10
x_7=bbrank	-2.54	-1.96	-2.20	-1.83
x_8=tuition	0	-0.24	-0.34	-0.36
x_9=board	0	1.49	1.25	1.42
x_{10}=bbattend	0	0	0	0
x_{11}=full_sal	0	0	0.07	0.08
x_{12}=sf_ratio	0	0	-0.12	-0.14
x_{13}=white	0	0	0.02	0.03
x_{14}=ast_sal	0	0	0.03	0.01
x_{15}=pop	0	0	0	0
x_{16}=prof_phd	0	0	0.04	0.03
x_{17}=ad_rate	0	0	-0.04	-0.03
x_{18}=loan_pct	0	0	-0.03	-0.03
x_{19}=outstate_pct	0	0	0.01	0.01

REFERENCES

Akaike, H. (1973), Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, **60**, 255-265.

Boos, D., Stefanski, L. and Wu, Y. (2009), Fast FSR variable selection with applications to clinical trials. *Biometrics*, **65**, 692-700.

Breiman, L. (1996a), Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24**, 2350-2383.

Breiman, L. (1996b), Bagging predictors. *Machine Learning*, **24**, 123-140.

Chatterjee, A. and Lahiri, S. (2010), Bootstrapping lasso estimators. To appear in *Journal of the American Statistical Association*.

Efron, B. (1979), Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1-26.

Efron, B. (1992), Jackknife-after-Bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society, Series B*, **54**, 83-127.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), Least angle regression. *The Annals of Statistics*, **32**, 407-499.

Efron, B., Tibshirani, R. (1993), An introduction to the bootstrap. *Chapman & Hall*.

Fan, J., and Li, Z. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

Gilmour, S. (1996), The interpretation of Mallows's C_p-statistics. *The Statistician*, **45**, 49-56.

Hurvich, C. and Tsai, C. (1990), The impact of model selection on inference in linear regression. *The American Statistician*, **44**, 214-217.

Knight, K. and Fu, W. (2000), Asymptotics for Lasso-type estimators. *The Annals of Statistics*, **28**, 1356-1378.

Leeb, H., and Pötscher, B. M (2005), Model selection and inference: facts and fiction. *Econometric Theory*, **21**, 21-59.

Mangold, W., Bean, L. and Adams, D. (2003), The impact of intercollegiate athletics on graduation rates among major NCAA Division I universities. *The Journal of Higher Education*, **74**, 540-562.

Miller, A. (2002), Subset selection in regression. *Chapmall & Hall*.

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

Sexton, J., and Laake P. (2009), Standard errors for bagged and random forest estimators. *Computational Statistics and Data Analysis*, **53**, 801-811.

Shao, J. and Tu, D. (1995), The jackknife after bootstrap, *Springer Series in Statistics*.

Shen, X., Huang, H., and Ye, J. (2004), Inference after model selection. *Journal of the American Statistical Association*, **99**, 751-762.

Stone, M.(1977), An asymptotic equivalence of choice of model by cross-validation and Akaike criterion. *Journal of the Royal Statistical Society, Series B*, **39**, 44-47.

Stone, M.(1979), Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society, Series B*, **41**, 276-278.

Tibshirani, R. (1996), Regression shrinkage and selection via the Lasso. *Journal of the*

Royal Statistical Society, Series B, **58**, 267-288.

Yu, B. and Buhlmann, P. (2002), Analyzing bagging. *The Annals of Statistics*, **30**, 927-961.

Wu, Y., Boos, D. and Stefanski, L. (2007), Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*, **102**, 235-243.

Zhang, P. (1992), Inference after variable selection in linear regression models. *Biometrika*, **79**, 741-746.

Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67**, 301-320.

Zou, H. (2006), The adaptive Lasso and its oracle propertes. *Journal of the American Statistical Association*, **101**, 1418-1429.