ON SOME ASPECTS OF MODEL SELECTION VARIABILITY

ZHAO WEI YANG

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE
STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

MASTER OF ARTS

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

AUGUST 2016

Abstract

In this thesis, we investigate the data analytic approach to integrate the model selection uncertainty into the statistical inferences of high dimensional estimators. Two closed-form formulae of covariance matrices are derived for high dimensional bagging estimators, one for the nonparametric bootstrapping and the other for the parametric bootstrapping. Two simulation studies are completed in detail for demonstrating the validity of the new formulae. Several model selection methods — the hypothesis testing, the Mallows' C_p, AIC, BIC and LASSO — are compared in terms of the effects on the accuracy of bagging estimators in the context of multivariate linear regression. The confidence region and its coverage probability are also estimated for the bagging estimators with those model selection methods.

Acknowledgments

I would first like to thank my supervisor Professor Yuehua Wu who selected this wonderful research topic for me. I have benefited a lot from her valuable guidances and generous supports. It is my pleasure to be her student.

I wish to express my appreciation to Professor Michael Chen as a member of my supervisory committee for his helpful comments and advice on thesis writing. I also wish to thank Xiaoying Sun for the software package information on the LASSO method.

My appreciation also goes to all the members of the Department of Mathematics and Statistics of York University for their assistance.

I would dedicate this thesis to my family — my wife Lixia for her constant support for my pursuing a higher level study, and my daughter Lily.

Contents

List of Notations

$\boldsymbol{y}$	$\in \mathbb{R}^p$	a p-dimensional column vector of response variables
$\boldsymbol{Y}$	$\in \mathbb{R}^{n\times p}$	a $n \times p$ matrix in which the i-th row is $\boldsymbol{y}^T$
$\boldsymbol{e}$	$\in \mathbb{R}^p$	a p-dimensional column vector of random errors
$\boldsymbol{E}$	$\in \mathbb{R}^{n\times p}$	a $n \times p$ matrix in which the i-th row is $\boldsymbol{e}^T$
$\boldsymbol{x}$	$\in \mathbb{R}^q$	a q-dimensional column vector of covariates
X	$\in \mathbb{R}^{n\times q}$	a design matrix in which the i-th row is $\boldsymbol{x}^T$
$\mathcal{L}$		$\triangleq \{(\boldsymbol{y}_i, \boldsymbol{x}_i)\}$, where, $i = 1, 2, \cdots, n$; a data sample
$\mathcal{L}^*$		a bootstrap sample
$\boldsymbol{B}$	$\in \mathbb{R}^{q\times p}$	a $q \times p$ matrix of coefficients
$\boldsymbol{Z}$	$\in \mathbb{R}^p$	a p-dimensional random vector
$\boldsymbol{z}$	$\in \mathbb{R}^p$	a p-dimensional column vector
$F(\boldsymbol{z})$	$: \mathbb{R}^p \to \mathbb{R}$	a cumulative distribution function (CDF) of $\boldsymbol{Z}$
$\boldsymbol{a}(\boldsymbol{z})$	$\in \mathbb{R}^d$	a d-dimensional vector function
$\boldsymbol{T}_F(\boldsymbol{z})$	$: \mathbb{R}^p \to \mathbb{R}^d$	a d-dimensional linear functional
$\boldsymbol{\mu}$	$\in \mathbb{R}^d$	a parameter of interest in the d-dimensional space
$\boldsymbol{\Sigma}_{(\cdot)\,(\cdot)}$		the covariance between two random elements

1 Introduction

1.1 The Variability of Model Selection

In the classical statistical theory, the response and covariates are first determined and transformed as needed on the basis of the observed data sample (ChatField, 1995). After the mathematical relation is assumed between the response and covariates, the covariates are selected once or iteratively by expert's domain knowledge, statistical hypothesis testings, etc. The determined model is then fitted through the give data sample, and statistical inferences are made from the model fitting results.

The variability of model selection is in practice ignored in making statistical inferences as if the selected model is certain, although the best model is searched from a class of possible candidates. The estimated accuracy is solely based on the pre-selected model, and is therefore over-optimistic. The estimated standard error is less than what it actually is, and the estimated confidence interval is nar-

rower(Berk *et al.*, 2013). This problem has been recognized in the statistics community for a long time, and it is well accepted that the model selection variability needs to be incorporated into the accuracy estimate (Bickel, 1984; Pötscher, 1991; Kabaila, 1998).

1.2 A Review of Some Relevant Literature

The approaches to incorporating the model selection uncertainty into the accuracy of estimators include: constructing confidence intervals irrespective of model selection procedures, making inferences from the limiting distributions of estimators that contains the model selection uncertainty, and approximating the model selection uncertainty through bootstrap samples.

For the cases that the model selection procedure is unknown or hard to specify, it is appropriate to build conservative intervals to incorporate the uncertainty of model selection. (Berk *et al.*, 2013) propose an approach to build conservative simultaneous confidence intervals (CI) on the basis of normal theory. The true parameter is guaranteed to be covered by the properly widened CIs, though the true sub-model need not be in the full set of candidate models. However, the estimated bounds are less than the Scheffé bound, as their method takes advantage of the intrinsic structures of the functionals of CI limits. The random errors are assumed to be estimated independently of the model selection procedures, hence

the accuracy estimates are completely isolated from the selected model. Their method generates valid inferences even for the misspecified models.

For the cases that the model selection procedure is known and can be specified, the properties of a model selection procedure and the model averaging strategy can be employed to improve the accuracy assessment of an estimator. On the basis of the limiting distributions of an estimator, (Hjort and Claeskens, 2003) build confidence intervals that accommodate the model selection uncertainty. The variability of a model selection procedure is unified in a framework as a disturbance to the data distribution through additional model parameters. And the true data distribution is assumed known, but the true data model does not have to be among the candidate models. Both post-model estimators and model averaging estimators are analyzed in the frequentist view.

The computational approach incorporates the model selection uncertainty into inferences of estimators through an approximation of the sampling distribution of the observed data sample. This approach is more data analytic and assumes that the true model is in the full set of candidate models. (Efron, 2014) introduces the bootstrapping methods to simulate the variability of model selection procedures. Bagging estimators are recommended for smoothing out the discontinuity in the estimates of parameter of interest, because model selectors usually oscillate abruptly among the optimal models for bootstrap samples. Formulae for the

variances of bagging estimators are derived and analyzed for both nonparametric and parametric bagging procedures, though only one dimensional estimators are investigated.

1.3 Multivariate Linear Regression (MLR)

Let $\boldsymbol{y} \triangleq [y_1, y_2, \ldots, y_p]^T$ be a vector of p (> 1) response variables, and $\boldsymbol{x} \triangleq [x_1, x_2, \ldots, x_q]^T$ be a vector of q (≥ 1) covariates. Without loss of generality, let us assume that both the response variables and the covariates are centered, so the intercept can be omitted. Let the random components of the response variables be put into a vector $\boldsymbol{e} = [e_1, e_2, \ldots, e_p]^T$, which is from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. In a multivariate linear regression (MLR) model, the response variables are linearly associated with the covariate separately as $y_j = x_1 b_{1j} + x_2 b_{2j} + \cdots + x_q b_{qj} + e_j$, where $j = 1, 2, \ldots, p$. The the vector form of MLR model is $\boldsymbol{y}^T = \boldsymbol{x}^T \boldsymbol{B} + \boldsymbol{e}^T$, where $\boldsymbol{B}$ is the coefficient matrix of which each column is a coefficient vector for the associated response variable.

Denote $\mathcal{L} \triangleq \{(\boldsymbol{y}_i, \boldsymbol{x}_i)\}$, $i = 1, 2, \ldots, n$; as a data sample of size n. Let $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n$ be independent and identically distributed on a p-dimensional random vector $\boldsymbol{y}$, while $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ are treated as fixed values. After the vectors of response variables are stacked in rows of a matrix, the MLR model can be

expressed in a matrix form as follows.

$$\boldsymbol{Y} = X\boldsymbol{B} + \boldsymbol{E}, \tag{1.1}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1p} \\ y_{21} & y_{22} & \dots & y_{2p} \\ . & . & . & . \\ y_{n1} & y_{n2} & \dots & y_{np} \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{12} & . & x_{1q} \\ x_{21} & x_{22} & . & x_{2q} \\ . & . & . & . \\ x_{n1} & x_{n2} & . & x_{nq} \end{bmatrix},$$

$$\boldsymbol{B} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ . & . & . & . \\ b_{q1} & b_{q2} & \dots & b_{qp} \end{bmatrix}, \quad \boldsymbol{E} = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1p} \\ e_{21} & e_{22} & \dots & e_{2p} \\ . & . & . & . \\ e_{n1} & e_{n2} & \dots & e_{np} \end{bmatrix}.$$

Let the error matrix $\boldsymbol{E}$ be partitioned into column vectors as follows.

$$\boldsymbol{E} = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1p} \\ e_{21} & e_{22} & \dots & e_{2p} \\ . & . & . & . \\ e_{n1} & e_{n2} & \dots & e_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_{(1)} & \boldsymbol{e}_{(2)} & \cdots & \boldsymbol{e}_{(p)} \end{bmatrix}.$$

It is assumed that $\mathrm{E}\left[\boldsymbol{e}_{(j)}\right] = \mathbf{0}$ and $\mathrm{cov}\left(\boldsymbol{e}_{(j)}, \boldsymbol{e}_{(k)}\right) = \sigma_{jk}\boldsymbol{I}_n$, where $j, k = 1, 2, \cdots, n$. This indicates that the mean for an error component of response variable is $\mathbf{0}$, and the variance or covariance between any two error components is a constant value.

The coefficient matrix $\boldsymbol{B}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the random vector $\boldsymbol{e}$ are unknown parameters, and the MLE's (Johnson and Wichern, 2007) for $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ are

$$\widehat{\boldsymbol{B}} = \left[X^T X\right]^{-1} X^T \boldsymbol{Y},$$
$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\widehat{\boldsymbol{E}}^T\widehat{\boldsymbol{E}} = \frac{1}{n}\left(\boldsymbol{Y} - X\widehat{\boldsymbol{B}}\right)^T\left(\boldsymbol{Y} - X\widehat{\boldsymbol{B}}\right)$$

respectively. The MLE's of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ are estimated from bootstrap samples for the bagging estimators in the context of MLR in the simulation study (Section 2.4).

For the purpose of hypothesis testing, let us divide the covariates into two groups $\{x_1, x_2, \cdots, x_r\}$ and $\{x_{r+1}, x_{r+2}, \cdots, x_q\}$ with the indices being reordered from the original ones when necessary. Accordingly, the design matrix X is partitioned into the left and right blocks as $\left[X_{(1)}\middle|X_{(2)}\right]$, and the coefficient matrix into the upper and lower blocks as $\left[\boldsymbol{B}_{(1)}^T\middle|\boldsymbol{B}_{(2)}^T\right]^T$. The hypothesis testing is on whether the second group of covariates makes insignificant contributions to the

responses.

$$H_0 : \boldsymbol{B}_{(2)} = 0, \quad \text{v.s.} \quad H_1 : \boldsymbol{B}_{(2)} \neq 0.$$

Under H_0, the MLR model turns into $\boldsymbol{Y} = X_{(1)}\boldsymbol{B}_{(1)} + \boldsymbol{E}$, and the MLE's of $\boldsymbol{B}$ and $\boldsymbol{E}$ are

$$\begin{aligned}
\widehat{\boldsymbol{B}}_{(1)} &= \left[X_{(1)}^T X_{(1)}\right]^{-1} X_{(1)}^T \boldsymbol{Y}, \\
\widehat{\boldsymbol{\Sigma}}_{(1)} &= \frac{1}{n}\left[(\boldsymbol{Y} - X_{(1)}\widehat{\boldsymbol{B}}_{(1)}\right]^T \left[\boldsymbol{Y} - X_{(1)}\widehat{\boldsymbol{B}}_{(1)}\right]
\end{aligned}$$

respectively.

The test statistics popular in software packages are Wilks' lambda, Pillai's trace and Hotelling-Lawley trace, to name just a few (Izenman, 2008). Small Wilks' lambda, large Pillai's trace or Hotelling-Lawley trace lead to reject the null hypothesis.

$$\begin{aligned}
\text{Wilks' lambda} &= \frac{\left|\widehat{\boldsymbol{\Sigma}}\right|}{\left|\widehat{\boldsymbol{\Sigma}}_{(1)}\right|}. \\
\text{Pillai's trace} &= \operatorname{tr}\left[(\widehat{\boldsymbol{\Sigma}}_{(1)} - \widehat{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\Sigma}}_{(1)}^{-1}\right]. \\
\text{Hotelling-Lawley trace} &= \operatorname{tr}\left[(\widehat{\boldsymbol{\Sigma}}_{(1)} - \widehat{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\Sigma}}^{-1}\right].
\end{aligned}$$

The Pillai's trace statistic is used in the simulation study for a representative of

hypothesis testing based variable selection methods.

The simulation study in Chapter 2 is based on the multivariate linear regression models. The parameter of interest is on the prediction of mean response of a multivariate linear regression model.

1.4 Model Selection Methods for the MLR Model

There are a wide variety of model selection methods, and we only review some of the commonly-used ones.

Hypothesis testing (Westfall and Young, 1993) can be applied to the cases where the candidate models are nested. Covariates are tested against the null hypothesis through the use of properly designed test statistics, and the statistically significant ones are retained in the selected model. Some of the frequently used test statistics are enumerated in section 1.3 for the MLR models.

Mallows' C_p (Mallows, 1973) is a technique to select the optimal model from candidate liner regression models. The C_p statistic is a criterion to assess how well a linear model fits the data. The optimal model is determined as the candidate model with the smallest C_p value. For the purpose of comparing the candidate linear models that differ only in the number of covariates, the C_p statistic can be

defined as

$$C_p = \text{RSS of the candidate model} \quad + \\ 2\,\widehat{\sigma^2} \times (\text{the number of unknown parameters}),$$

where RSS stands for the residual sum of squares, and $\widehat{\sigma^2}$ is the variance of random errors that are estimated from residuals of fitting the full model. For a multivariate linear regression models, the statistic (Fujikoshi and Satoh, 1997) may be updated to

$$C_p = n\,\mathrm{tr}\left[\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Sigma}}_t\right] + 2\,p\,t,$$

where n is the sample size, $\widehat{\mathbf{\Sigma}}$ is the covariance matrix of the random errors estimated from the residuals of fitting the full model with q covariates, p is the number of responses per observation, and $\widehat{\mathbf{\Sigma}}_t$ is the covariance matrix of the random errors estimated from fitting the candidate model with t covariates.

Information criteria based approaches for the model selection derive various quantities of the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) that measures the information loss in the approximation of the true distribution

with another distribution. The KL distance may be expressed as

$$KL(g; f) = \int g(x) \log \left[\frac{g(x)}{f(x)} \right] dx,$$

where $g(x)$ and $f(x)$ are the pdf functions for the unknown distribution and the approximating distribution respectively.

The Akaike's information criterion (AIC) (Akaike, 1974) takes into account the bias of the maximized log-likelihood function of a model as an estimator of the relative KL distance to the generating model. The general equation for the AIC value of a model is given by

$$\begin{aligned} \text{AIC} = -2 \times (\text{the maximum log-likelihood of candidate model}) \quad + \\ 2 \times (\text{the number of unknown parameters}). \end{aligned}$$

Accordingly, the AIC criterion for the purpose of selecting an optimal MLR model may be defined as (Fujikoshi and Satoh, 1997)

$$\text{AIC} = n \log(\left|\widehat{\mathbf{\Sigma}}_t\right|) + 2\,p\,t,$$

where n is the sample size, $\widehat{\mathbf{\Sigma}}_t$ is the covariance matrix of random errors estimated from fitting the candidate model with t covariates, and p is the number

of responses per observation. Of all the candidate models, the one with the least AIC value is deemed as the optimal model.

The Bayesian information criterion (BIC) (Schwarz, 1978), also named Schwarz criterion, has a penalty term that is dependent with the sample size. Its general equation is given by

$$\begin{aligned}\text{BIC} = -2 \times (&\text{the maximum log-likelihood of candidate model}) \quad + \\ &(\text{the number of unknown parameters}) \times \log(\text{the sample size}).\end{aligned}$$

For the model selection in an MLR context, BIC may be defined as (Kass and Raftery, 1995)

$$\text{BIC} = n \log\left|\widehat{\boldsymbol{\Sigma}}_t\right| + p\, t \log n,$$

where n is the sample size, $\widehat{\boldsymbol{\Sigma}}_t$ is the covariance matrix of random errors estimated from fitting the candidate model with t covariates, and p is the number of responses per observation. The candidate model with the least BIC value is regarded as the optimal model.

LASSO (least absolute shrinkage and selection operator) is a regression method that performs regularization and variable selection simultaneously (Tibshirani,

1996). It tends to produce some coefficients that are exactly zero instead of shrinking the coefficients altogether, thus can be employed as a model selection procedure. In the context of MLR, the solution given by the LASSO algorithm is

$$\widehat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \|\boldsymbol{Y} - X\boldsymbol{B}\|_2^2 + \lambda \left|\boldsymbol{B}\right|_1 ,$$

where $|\boldsymbol{B}| = \sum_{i=1}^{q}\sum_{j=1}^{p}|b_{ij}|$ is the L_1 norm of $\boldsymbol{B}$, and λ is the tuning parameter.

There are other model selection methods. For example, cross-validation methods (Shao, 1993) select the model with the best performance after repetitively partitioning data sample and building models. Bootstrapping methods (Shao, 1996) select the model with the best performance across the bootstrap samples.

1.5 Influence Function

Let random variables $Z, Z_1, Z_2, \ldots, Z_n$ be independently and identically distributed. Denote the CDF of Z by $F(z) = P(Z \leq z)$, $-\infty < z < \infty$. The empirical distribution function $\widehat{F}_n$ is the CDF that puts mass $\frac{1}{n}$ at each data point. Denote $T(F)$ as a statistical functional, which is any function of F. An influence function quantifies the rate of change in a statistical function upon a slight contamination in the distribution F.

Definition 1.1. (Wasserman, 2006) Let δ_z be a point mass at z, the **influence**

function $L_F(z)$ is defined as

$$L_F(z) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\delta_z) - T(F)}{\epsilon}.$$

Theorem 1.1. *(Wasserman, 2006) Let $T(F) = \int a(z)dF(z)$ be a linear functional, then the influence function of $T(F)$ has following properties.*

1. $L_F(z) = a(z) - T(F)$.

2. *For any distribution G, $T(G) = T(F) + \int L_F(z)dG(z)$.*

3. $\int L_F(z)dF(z) = 0$.

4. *If $\tau^2 \triangleq \int L_F^2(z)dF(z) < \infty$, then*

$$\sqrt{n}\left[T(F) - T(\widehat{F}_n)\right] \xrightarrow{D} \mathcal{N}(0, \tau^2).$$

5. *Let $\widehat{\tau^2} = \frac{1}{n}\sum_{i=1}^n \widehat{L}_F^2(Z_i) = \frac{1}{n}\sum_{i=1}^n \left[a(Z_i) - T(\widehat{F}_n)\right]^2$, then $\widehat{\tau}^2 \xrightarrow{P} \tau^2$ and*

$$\frac{\widehat{\tau}/\sqrt{n}}{\sqrt{\text{var}(T(\widehat{F}_n))}} \xrightarrow{P} 1.$$

6. $\frac{\sqrt{n}\left[T(F) - T(\widehat{F}_n)\right]}{\widehat{\tau}} \xrightarrow{D} \mathcal{N}(0, 1)$.

Proof. See (Wasserman, 2006) for the proofs. □

Let the p-dimensional random vectors $\boldsymbol{Z}, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_n$ be independently and identically distributed. Denote the CDF of $\boldsymbol{Z}$ by $F(\boldsymbol{z}) = P(\boldsymbol{Z} \leq \boldsymbol{z})$. Let $\boldsymbol{z} \triangleq [z_1, \cdots, z_j, \cdots, z_p]^T$, where $-\infty < z_j < \infty$, $j = 1, 2, \cdots, p$. Theorem 1.1 is extended for a high dimensional linear functional in the following theorem.

Theorem 1.2. *Let $\boldsymbol{T}(F) = \int \boldsymbol{a}(\boldsymbol{z}) dF(\boldsymbol{z})$ be a d-dimensional linear functional, and $\delta_{\boldsymbol{z}}$ be a point mass at $\boldsymbol{z}$. The influence function of $\boldsymbol{T}(F)$ is defined as*

$$\boldsymbol{L}_F(\boldsymbol{z}) = \lim_{\epsilon \to 0} \frac{\boldsymbol{T}((1-\epsilon)F + \epsilon\delta_{\boldsymbol{z}}) - \boldsymbol{T}(F)}{\epsilon},$$

which possesses the the following properties.

1. $\boldsymbol{L}_F(\boldsymbol{z}) = \boldsymbol{a}(\boldsymbol{z}) - \boldsymbol{T}(F)$.

2. *For any distribution G,* $\boldsymbol{T}(G) = \boldsymbol{T}(F) + \int \boldsymbol{L}_F(\boldsymbol{z}) dG(\boldsymbol{z})$.

3. $\int \boldsymbol{L}_F(\boldsymbol{z}) dF(\boldsymbol{z}) = \boldsymbol{0}$.

4. *If $\boldsymbol{\Sigma_{LL}} \triangleq \int \boldsymbol{L}_F(\boldsymbol{z}) \boldsymbol{L}_F^T(\boldsymbol{z}) dF(\boldsymbol{z})$ is a positive definite matrix, then*

$$\sqrt{n}\left[\boldsymbol{T}(F) - \boldsymbol{T}(\widehat{F}_n)\right] \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_{LL}}).$$

5. *Let* $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{LL}} = \frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{L}}(\boldsymbol{Z}_i)\widehat{\boldsymbol{L}}^T(\boldsymbol{Z}_i)$
$= \frac{1}{n}\sum_{i=1}^n \left[\boldsymbol{a}(\boldsymbol{Z}_i) - \boldsymbol{T}(\widehat{F}_n)\right]\left[\boldsymbol{a}(\boldsymbol{Z}_i) - \boldsymbol{T}(\widehat{F}_n)\right]^T,$

   then

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{LL}} \xrightarrow{P} \boldsymbol{\Sigma_{LL}} \quad \textit{and} \quad \frac{1}{n}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{LL}}\left[\mathrm{var}(\boldsymbol{T}(\widehat{F}_n))\right]^{-1} \xrightarrow{P} \boldsymbol{I}_d.$$

6. $n\left[\boldsymbol{T}(F)-\boldsymbol{T}(\hat{F}_n)\right]^T \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{LL}}^{-1}\left[\boldsymbol{T}(F)-\boldsymbol{T}(\hat{F}_n)\right] \xrightarrow{D} \chi_d^2$.

Proof. 1. By the definition of influence function, we have

$$\begin{aligned}
\boldsymbol{L}_F(\boldsymbol{z}) &= \lim_{\epsilon\to 0}\frac{\boldsymbol{T}((1-\epsilon)F+\epsilon\delta_{\boldsymbol{z}})-\boldsymbol{T}(F)}{\epsilon} \\
&= \lim_{\epsilon\to 0}\frac{\int \boldsymbol{a}(\boldsymbol{z})d\left[(1-\epsilon)F+\epsilon\delta_{\boldsymbol{z}}\right]-\int \boldsymbol{a}(\boldsymbol{z})dF((\boldsymbol{z}))}{\epsilon} \\
&= \lim_{\epsilon\to 0}\frac{\int \boldsymbol{a}(\boldsymbol{z})dF(\boldsymbol{z})-\epsilon\int \boldsymbol{a}(\boldsymbol{z})dF(\boldsymbol{z})+\epsilon\boldsymbol{a}(\boldsymbol{z})-\int \boldsymbol{a}(\boldsymbol{z})dF(\boldsymbol{z})}{\epsilon} \\
&= -\int \boldsymbol{a}(\boldsymbol{z})dF(\boldsymbol{z})+\boldsymbol{a}(\boldsymbol{z}) \\
&= \boldsymbol{a}(\boldsymbol{z})-\boldsymbol{T}(F).
\end{aligned}$$

2. By using the first claim, we have

$$\begin{aligned}
\int \boldsymbol{L}_F(\boldsymbol{z})dG(\boldsymbol{z}) &= \int\left[\boldsymbol{a}(\boldsymbol{z})-\boldsymbol{T}(F)\right]dG(\boldsymbol{z}) \\
&= \int \boldsymbol{a}(\boldsymbol{z})dG(\boldsymbol{z})-\int \boldsymbol{T}(F)dG(\boldsymbol{z}) \\
&= \boldsymbol{T}(G)-\boldsymbol{T}(F) \\
\implies \quad \boldsymbol{T}(G) &= \boldsymbol{T}(F)+\int \boldsymbol{L}_F(\boldsymbol{z})dG(\boldsymbol{z}).
\end{aligned}$$

3. By using the first claim, we have

$$\begin{aligned}\int \boldsymbol{L}_F(\boldsymbol{z})dF(\boldsymbol{z}) &= \int \left[\boldsymbol{a}(\boldsymbol{z}) - \boldsymbol{T}(F)\right] dF(\boldsymbol{z}) \\ &= \int \boldsymbol{a}(\boldsymbol{z})dF(\boldsymbol{z}) - \int \boldsymbol{T}(F)dF(\boldsymbol{z}) \\ &= \boldsymbol{T}(F) - \boldsymbol{T}(F) = \mathbf{0}.\end{aligned}$$

4. By using the first and third claims, we have

$$\begin{aligned}\boldsymbol{T}(\widehat{F}_n) &= \int \boldsymbol{a}(\boldsymbol{z})d\widehat{F}_n(\boldsymbol{z}) \\ &= \int \left[\boldsymbol{T}(F) + \boldsymbol{a}(\boldsymbol{z}) - \boldsymbol{T}(F)\right] d\widehat{F}_n(\boldsymbol{z}) \\ &= \int \boldsymbol{T}(F)d\widehat{F}_n(\boldsymbol{z}) + \int \boldsymbol{L}_F(\boldsymbol{z})d\widehat{F}_n(\boldsymbol{z}) \quad \text{(by using 1st claim)} \\ &= \boldsymbol{T}(F) + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{L}_F(\boldsymbol{Z}_i)\end{aligned}$$

$$\begin{aligned}\implies \mathrm{E}\left\{\sqrt{n}\left[\boldsymbol{T}(\widehat{F}_n) - \boldsymbol{T}(F)\right]\right\} &= \int \sqrt{n}\left[\boldsymbol{T}(\widehat{F}_n) - \boldsymbol{T}(F)\right] dF(\boldsymbol{z}) \\ &= \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\int \boldsymbol{L}_F(\boldsymbol{z})dF(\boldsymbol{z}) \\ &= \mathbf{0} \quad \text{(by the 3rd claim), and}\end{aligned}$$

$$\begin{aligned}
\text{var}\left\{\sqrt{n}\left[\boldsymbol{T}(\widehat{F}_n)-\boldsymbol{T}(F)\right]\right\} &= n\int\left[\boldsymbol{T}(\widehat{F}_n)-\boldsymbol{T}(F)\right]\left[\boldsymbol{T}(\widehat{F}_n)-\boldsymbol{T}(F)\right]^T dF(\boldsymbol{z}) \\
&= n\int\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{L}_F(\boldsymbol{z}_i)\right]\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{L}_F(\boldsymbol{z}_i)\right]^T dF(\boldsymbol{z}) \\
&= \frac{n}{n^2}\sum_{i=1}^{n}\int\boldsymbol{L}_F(\boldsymbol{z})\boldsymbol{L}_F^T(\boldsymbol{z})dF(\boldsymbol{z}) \\
&= \int\boldsymbol{L}_F(\boldsymbol{z})\boldsymbol{L}_F^T(\boldsymbol{z})dF(\boldsymbol{z}) = \boldsymbol{\Sigma}_{\boldsymbol{LL}}.
\end{aligned}$$

According to the central limit theory (Ferguson, 1996), we obtain

$$\sqrt{n}\left[\boldsymbol{T}(F)-\boldsymbol{T}(\widehat{F}_n)\right]\xrightarrow{D}\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}_{\boldsymbol{LL}}).$$

5. By using $\text{E}\left[\widehat{F}_n(\boldsymbol{z})\right] = F(\boldsymbol{z})$ (Wasserman, 2006) and the law of large numbers (Ferguson, 1996), we get

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{LL}} &= \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{L}}_F(\boldsymbol{Z}_i)\widehat{\boldsymbol{L}}_F^T(\boldsymbol{Z}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{a}(\boldsymbol{Z}_i)-\boldsymbol{T}(\widehat{F}_n)\right]\left[\boldsymbol{a}(\boldsymbol{Z}_i)-\boldsymbol{T}(\widehat{F}_n)\right]^T \\
&\xrightarrow{P}\int\left[\boldsymbol{a}(\boldsymbol{z})-\boldsymbol{T}(F)\right]\left[\boldsymbol{a}(\boldsymbol{z})-\boldsymbol{T}(F)\right]^T dF(\boldsymbol{z}) \\
&= \int\boldsymbol{L}_F(\boldsymbol{z})\boldsymbol{L}_F^T(\boldsymbol{z})dF(\boldsymbol{z}) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{LL}},
\end{aligned}$$

$$
\begin{aligned}
\operatorname{var}\left[\boldsymbol{T}(\widehat{F}_n)\right] &= \operatorname{var}\left[\int \boldsymbol{a}(\boldsymbol{z}) d\widehat{F}_n(\boldsymbol{z})\right] = \operatorname{var}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}(\boldsymbol{Z}_i)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}\left[\boldsymbol{a}(\boldsymbol{Z}_i)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\left[\boldsymbol{a}(\boldsymbol{Z}_i)-\mathrm{E}\,\boldsymbol{a}(\boldsymbol{Z}_i)\right]\left[\boldsymbol{a}(\boldsymbol{Z}_i)-\mathrm{E}\,\boldsymbol{a}(\boldsymbol{Z}_i)\right]^T \\
&= \frac{1}{n}\mathrm{E}\left[\boldsymbol{a}(\boldsymbol{Z})-\mathrm{E}\,\boldsymbol{a}(\boldsymbol{Z})\right]\left[\boldsymbol{a}(\boldsymbol{Z})-\mathrm{E}\,\boldsymbol{a}(\boldsymbol{Z})\right]^T \\
&= \frac{1}{n}\mathrm{E}\left[\boldsymbol{a}(\boldsymbol{Z})-\boldsymbol{T}(F)\right]\left[\boldsymbol{a}(\boldsymbol{Z})-\boldsymbol{T}(F)\right]^T \\
&= \frac{1}{n}\mathrm{E}\left[\boldsymbol{L}_F(\boldsymbol{Z})\boldsymbol{L}_F^T(\boldsymbol{Z})\right] = \frac{1}{n}\int \boldsymbol{L}_F(\boldsymbol{z})\boldsymbol{L}_F^T(\boldsymbol{z})dF(\boldsymbol{z}) \\
&= \frac{1}{n}\boldsymbol{\Sigma}_{\boldsymbol{LL}}.
\end{aligned}
$$

Therefore $\quad \frac{1}{n}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{LL}}\left[\operatorname{var}(\boldsymbol{T}(\widehat{F}_n))\right]^{-1} = \frac{1}{n}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{LL}}\left[\frac{1}{n}\boldsymbol{\Sigma}_{\boldsymbol{LL}}\right]^{-1} \xrightarrow{P} \boldsymbol{I}_d.$

6. By using the fourth claim we get

$$
n\left[\boldsymbol{T}(F)-\boldsymbol{T}(\hat{F}_n)\right]^T \boldsymbol{\Sigma}_{\boldsymbol{LL}}^{-1}\left[\boldsymbol{T}(F)-\boldsymbol{T}(\hat{F}_n)\right] \xrightarrow{D} \chi_d^2.
$$

By applying the Slutsky theorem (Ferguson, 1996) and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{LL}} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{LL}}$ from the fifth claim, we get

$$
\begin{aligned}
& n\left[\boldsymbol{T}(F)-\boldsymbol{T}(\hat{F}_n)\right]^T \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{LL}}^{-1}\left[\boldsymbol{T}(F)-\boldsymbol{T}(\hat{F}_n)\right] \\
\xrightarrow{D}\ & n\left[\boldsymbol{T}(F)-\boldsymbol{T}(\hat{F}_n)\right]^T \boldsymbol{\Sigma}_{\boldsymbol{LL}}^{-1}\left[\boldsymbol{T}(F)-\boldsymbol{T}(\hat{F}_n)\right] \\
\xrightarrow{D}\ & \chi_d^2. \qquad \square
\end{aligned}
$$

From the sixth claim of Theorem 1.2, a $(1-\alpha)100\%$ asymptotic confidence region for $\boldsymbol{T}(F)$ can be constructed as

$$\left\{\boldsymbol{T}(F): \quad n\left[\boldsymbol{T}(F)-\boldsymbol{T}(\hat{F}_n)\right]^T \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{LL}}^{-1}\left[\boldsymbol{T}(F)-\boldsymbol{T}(\hat{F}_n)\right] \leq \chi_d^2\right\},$$

where α is the significance level and $\chi_p^2(\alpha)$ denotes the upper (100α)th percentile of χ_d^2 distribution.

In Section 2.2, the covariance matrix for the nonparametric bagging estimator is constructed as $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{LL}}$ defined in Theorem 1.2. And the confidence region for the true parameter of interest is built with the nonparametric delta method, as shown above.

1.6 Purposes and Outline

The primary objective of this thesis is to investigate the model selection variability through a data analytic approach that applies model selection procedures on bootstrap samples. We will derive the covariance matrices for the high dimensional bagging estimators, and examine the performance of those estimators through simulation studies. We will also compare the effects of several model selection procedures.

In Chapter 2, the derivation of covariance matrices is detailed for both the

nonparametric and the parametric bagging estimators in the context of multivariate linear regression. Some commonly used model selection methods are applied in the simulation study for the comparison of their effects on the bagging estimators. The properties of the nonparametric bagging estimator are investigated through a polynomial regression in the MLR context. And the properties of the parametric bagging estimator are examined through an MLR model with the full set of five candidate covariates. Three covariates make contributions to the responses, however they are highly correlated with the other two that actually make no contribution to the responses.

In Chapter 3, we summarize the main results of this study and discuss the future work.

2 Bootstrap Smoothing for Multivariate Estimators

2.1 Introduction

Bootstrap (Efron, 1979) is a widely-used general method to approximate the sampling distribution of a statistic. Its theoretic foundations (Singh, 1981) prove that bootstrapping distributions approximate sampling distributions fairly well. The simplicity of bootstrap has led to its universality, as the expense of computing power has been decreasing each year. Bootstrap has been directly used in approximating the standard error, bias, and confidence interval for an estimator (Efron, 1987). Bootstrapped confidence intervals are usually asymptotically more accurate than the standard confidence intervals derived under the assumptions of normality (DiCiccio and Efron, 1996).

The data analytic approach of assessing model selection variability employs the bootstrap approximation of sampling distribution to evaluate the variability of model selection procedures (Politis, 2014). A model selection procedure may

assign different models to different samples, hence the space of samples are divided into partitions. At a border between two partitions, the estimate of a parameter of interest may change abruptly although the samples from two partitions are close enough.

Bootstrap aggregating (bagging) (Breiman, 1996; Buja and Stuetzle, 2006; Hjort, 2014) is an effective computational methods to enhance the stability of estimators. Bagging reduces the variance of estimation of testing-based linear regression because hypothesis testing performs a role of hard thresholding operation and bagging smoothes out its spurious abruptness (Bühlmann and Yu, 2002). The same spirit applies in handling the discontinuity of estimation due to uncertainty from other variable selection procedures (Gelman and Vehtari, 2014).

There are two major types of bootstrap schemes: nonparametric bootstrapping and parametric bootstrapping (Efron and Tibshirani, 1993). Nonparametric bootstrapping resamples with replacement the observed data sample by assigning usually equal probability of being drawn to each observed sample point. And parametric bootstrapping generates samples from a distribution that is usually fitted from the observed data sample with the maximum likelihood estimation.

The rest of this chapter is organized as follows: the covariance matrix for the nonparametric bagging estimator is derived in section 2.2. The covariance matrix for the parametric bagging estimator is derived in section 2.3. Two examples in the

context of multivariate linear regression, one for the nonparametric bootstrapping and the other for the parametric bootstrapping, are enumerated and analyzed in section 2.4.

2.2 Nonparametric Bootstrap Smoothing

Let $\boldsymbol{y}$ be a p-dimensional vector of response variables, and $\boldsymbol{x}$ be a q-dimensional vector of fixed values for the covariates. Denote $(\boldsymbol{y}_j, \boldsymbol{x}_j)$ as a sample point, where $j = 1, 2, \cdots, n$; n is the sample size, and $\boldsymbol{\mu}$ the parameter of interest in a d-dimensional space. The bootstrapping is to resample the points of observed data sample. The nonparametric bootstrap smoothing algorithm for the MLR is outlined as follows.

1. Resample with replacement the observed data sample $\mathcal{L} = \{(\boldsymbol{y}_j, \boldsymbol{x}_j)\}, j = 1, 2, \ldots, n$; to obtain the bootstrap samples of same size as the observed one, $\mathcal{L}_b^* = \{(\boldsymbol{y}_{bj}, \boldsymbol{x}_{bj})^*\}, b = 1, 2, \ldots, S$.

2. Count the occurrences of each observed point in each bootstrap sample, and record those counts into a vector $\boldsymbol{K}_b^* = [K_{b1}^*, K_{b2}^*, \ldots, K_{bm}^*, \ldots, K_{bn}^*]^T$, where $K_{bj}^* = \sharp\{(\boldsymbol{y}_{bm}, \boldsymbol{x}_{bm})^* = (\boldsymbol{y}_j, \boldsymbol{x}_j)\}$, $\sum_{m=1}^{n} K_{bm}^* = n, m = 1, 2, \ldots, n$.

3. Select the optimal MLR models $\mathcal{M}_b^*$ for each bootstrap sample according to a model selection procedure.

4. Fit the selected MLR models $\mathcal{M}_b^*$ to each bootstrap sample to obtain the MLE's $\widehat{\boldsymbol{B}}_b^*$, $\widehat{\boldsymbol{\Sigma}}_b^*$, and accordingly estimate the parameter of interest $\widehat{\boldsymbol{\mu}}$ with respect to each bootstrap sample.

5. Average the estimates of the parameter of interest across all the bootstrapped optimal models to obtain the bagging estimator $\widetilde{\boldsymbol{\mu}}(\boldsymbol{x}) = \frac{1}{S}\sum_{b=1}^{S} \widehat{\boldsymbol{\mu}}(\mathcal{M}_b^*(\boldsymbol{x}, \mathcal{L}_b^*))$.

We derive the covariance matrix for the smoothed multivariate parameter of interest $\widetilde{\boldsymbol{\mu}}(\boldsymbol{x})$, which extends the formula in (Efron, 2014) from the one dimensional case to the high dimensional case.

The probability of being selected is same for each point $(\boldsymbol{y}_j, \boldsymbol{x}_j)$ in the observed sample, namely $\frac{1}{n}$ since the observed sample is resampled with replacement and equal weights are assigned to the sample points. The bootstrap sample is of same size (i.e., n) as the observed sample, therefore there are n^n permutations in total. The counting vector $\boldsymbol{K}^*$ has a multinomial distribution with n trials on n categories and the n-dimensional probability vector $\boldsymbol{p}_0 = \left[\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right]^T$, which is denoted as $\boldsymbol{K}^* \sim \text{Mult}_n(n, \boldsymbol{p}_0)$. The expectation and the variance matrix of $\boldsymbol{K}^*$ are $\mathbf{1}_n \triangleq [1, 1, \ldots, 1]^T$ and $\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ respectively.

The estimator $\widehat{\boldsymbol{\mu}}$ can be viewed as a functional of a selected model object $\mathcal{M}^*$, an input vector $\boldsymbol{x}$ or a model training sample $\mathcal{L}^*$. It is usually discontinuous with

respect to the sample space $\{(\boldsymbol{y}, \boldsymbol{x})\}$. This is because model selection procedures select different models, which results in jumps in estimates at borders between partitions of a sample space. However, $\widehat{\boldsymbol{\mu}}$ can be viewed as a function of $\boldsymbol{K}^*$, and accordingly $\widetilde{\boldsymbol{\mu}}$ is the expectation of $\widehat{\boldsymbol{\mu}}$ in the ideal case in which the bootstrapping repeats n^n times.

The smoothed estimator $\widetilde{\boldsymbol{\mu}}$ is a continuous functional with respect to the probability vector. It can be viewed as a functional of $\widehat{\boldsymbol{\mu}}$, and further as a continuous functional of the probability distribution that underlies the counting vector $\boldsymbol{K}^*$. Thanks to its continuity property in the probability space, the first derivatives of $\widetilde{\boldsymbol{\mu}}$ can be evaluated by quantifying its change over a slight contamination of a probability distribution. By using the influence function of $\widetilde{\boldsymbol{\mu}}$, we derive the variance matrix of $\widetilde{\boldsymbol{\mu}}$, as shown in the following theorem.

Theorem 2.1. *Let* $\widehat{\boldsymbol{\mu}} \triangleq [\hat{\mu}_1, \cdots, \hat{\mu}_i, \cdots, \hat{\mu}_d]^T$ *and* $\widetilde{\boldsymbol{\mu}} \triangleq [\widetilde{\mu}_1, \cdots, \widetilde{\mu}_i, \cdots, \widetilde{\mu}_d]^T$, *where* $i = 1, 2, ...d$. *Denote the variance of* $\widetilde{\boldsymbol{\mu}}$ *as a* $d \times d$ *matrix* $\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}}$ *of which the* (i, l)*-th entry is* $\mathrm{cov}(\widetilde{\mu}_i, \widetilde{\mu}_l) \triangleq \mathrm{E}\,[(\widetilde{\mu}_i - \mathrm{E}\,\widetilde{\mu}_i)(\widetilde{\mu}_l - \mathrm{E}\,\widetilde{\mu}_l)]$, *where* $l = 1, 2, \cdots, d$. *Denote the covariance between* $\widehat{\boldsymbol{\mu}}$ *and* K_j *(the number of times the* j*-th observed point has been resampled in a bootstrap sample) as a* $d \times 1$ *matrix* $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}},K_j}$ *in which the* $(i, 1)$*-th entry is* $\mathrm{cov}(\hat{\mu}_i, K_j) \triangleq \mathrm{E}\,[(\hat{\mu}_i - \mathrm{E}\,\hat{\mu}_i)(K_j - \mathrm{E}\,K_j)]$, *where* $j = 1, 2, \cdots, n$.

Then the nonparametric delta-method estimate of the covariance matrix for

the smoothed estimator $\widetilde{\boldsymbol{\mu}}$ *is*

$$\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}} = \sum_{j=1}^{n} \left(\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}},K_j} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}},K_j}^T \right). \tag{2.1}$$

Proof. We first consider the ideal case, where the observed sample is resampled with replacement $S = n^n$ times. The smoothed estimator at the probability vector $\boldsymbol{p}$ can be expressed as

$$\begin{aligned} \widetilde{\boldsymbol{\mu}}(\boldsymbol{p}) &= \sum_{b=1}^{S} f(y_i^*, y_i^*, \ldots, y_n^*) \widehat{\boldsymbol{\mu}}_b^* \\ &= \sum_{b=1}^{S} p_1^{K_{b1}^*} p_2^{K_{b2}^*} \ldots p_n^{K_{bn}^*} \widehat{\boldsymbol{\mu}}_b^* \\ &= \sum_{b=1}^{S} \frac{p_1^{K_{b1}^*} p_2^{K_{b2}^*} \ldots p_n^{K_{bn}^*}}{(\frac{1}{n})^{K_{b1}^*} (\frac{1}{n})^{K_{b2}^*} \ldots (\frac{1}{n})^{K_{bn}^*}} \frac{1}{n^n} \widehat{\boldsymbol{\mu}}_b^* \\ &= \frac{1}{S} \sum_{b=1}^{S} \prod_{m=1}^{n} (np_m)^{K_{bm}^*} \widehat{\boldsymbol{\mu}}_b^*. \end{aligned} \tag{2.2}$$

Denote $w_b(\boldsymbol{p}) \triangleq \prod_{m=1}^{n} (np_m)^{K_{bm}^*}$, then we have

$$\widetilde{\boldsymbol{\mu}}(\boldsymbol{p}) = \frac{1}{S} \sum_{b=1}^{S} w_b(\boldsymbol{p}) \widehat{\boldsymbol{\mu}}_b^*. \tag{2.3}$$

Let $\boldsymbol{p}(\epsilon) \triangleq \boldsymbol{p}_0 + \epsilon(\boldsymbol{\delta}_j - \boldsymbol{p}_0)$, where $\boldsymbol{\delta}_j = [0, 0, \ldots, 1, \ldots, 0]^T$ is a vector with all components of 0 except the j-th one. Then $n\boldsymbol{p}(\epsilon) = [1 - \epsilon, \ldots, 1 + (n-1)\epsilon, \ldots, 1 - \epsilon]^T$,

and

$$
\begin{aligned}
w_b(\boldsymbol{p}(\epsilon)) &= [1+(n-1)\epsilon]^{K^*_{bj}} \prod_{\substack{m=1 \\ m\neq j}}^{n} (1-\epsilon)^{K^*_{bm}} \\
&= [1+(n-1)\epsilon]^{K^*_{bj}} (1-\epsilon)^{\sum_{m\neq j}^{n} K^*_{bm}} \\
&= [1+(n-1)\epsilon]^{K^*_{bj}} (1-\epsilon)^{n-K^*_{bj}}.
\end{aligned} \tag{2.4}
$$

Apply the Taylor's expansion of log functions

$$
\begin{aligned}
\log(1-x) &= -\sum_{k=1}^{\infty} \frac{x^k}{k} \quad \text{for } |x|<1, \\
\log(1+x) &= \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k} \quad \text{for } |x|<1
\end{aligned}
$$

to equation (2.4), then we get

$$
\begin{aligned}
\log w_b(\boldsymbol{p}(\epsilon)) &= K^*_{bj} \log[1+(n-1)\epsilon] + (n-K^*_{bj}) \log(1-\epsilon) \\
&= K^*_{bj} \sum_{k=1}^{\infty} (-1)^{k+1} \frac{[(n-1)\epsilon]^k}{k} - (n-K^*_{bj}) \sum_{k=1}^{\infty} \frac{\epsilon^k}{k} \\
&\approx K^*_{bj}(n-1)\epsilon - (n-K^*_{bj})\epsilon \\
&= n\epsilon(K^*_{bj}-1).
\end{aligned} \tag{2.5}
$$

Thus, $w_b(\boldsymbol{p}(\epsilon))$ can be approximated as

$$\begin{aligned} w_b(\boldsymbol{p}(\epsilon)) &\approx \exp\left[n\epsilon(K_{bj}^* - 1)\right] \\ &= \sum_{k=0}^{\infty} \frac{1}{k!}\left[n\epsilon(K_{bj}^* - 1)\right]^k \\ &\approx 1 + n\epsilon(K_{bj}^* - 1). \end{aligned} \tag{2.6}$$

By plugging equation (2.6) into (2.3), we get

$$\begin{aligned} \widetilde{\boldsymbol{\mu}}(\boldsymbol{p}(\epsilon)) &= \frac{1}{S}\sum_{b=1}^{S} w_b(\boldsymbol{p}(\epsilon))\widehat{\boldsymbol{\mu}}_b^* \\ &= \frac{1}{S}\sum_{b=1}^{S}\left[1 + n\epsilon(K_{bj}^* - 1)\right]\widehat{\boldsymbol{\mu}}_b^* \\ &= \frac{1}{S}\sum_{b=1}^{S}\widehat{\boldsymbol{\mu}}_b^* + n\epsilon\left[\frac{1}{S}\sum_{b=1}^{S}(K_{bj}^* - 1)\widehat{\boldsymbol{\mu}}_b^*\right] \\ &= \widetilde{\boldsymbol{\mu}}(\boldsymbol{p}_0) + n\epsilon\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}},K_j}. \end{aligned} \tag{2.7}$$

And the influence function of $\widetilde{\boldsymbol{\mu}}$ at $\boldsymbol{p}_0$ is

$$IF(\widetilde{\boldsymbol{\mu}})_j = \lim_{\epsilon\to 0}\frac{\widetilde{\boldsymbol{\mu}}(\boldsymbol{p}(\epsilon)) - \widetilde{\boldsymbol{\mu}}(\boldsymbol{p}_0)}{\epsilon} = n\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}},K_j}. \tag{2.8}$$

According to Theorem 1.2, we obtain the variance matrix for $\widetilde{\boldsymbol{\mu}}$ at $\boldsymbol{p}_0$ as

$$\begin{aligned}\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}} &= \frac{1}{n^2}\sum_{j=1}^{n}\left[IF(\widetilde{\boldsymbol{\mu}}_j)\right]\left[IF(\widetilde{\boldsymbol{\mu}}_j)\right]^T \\ &= \frac{1}{n^2}\sum_{j=1}^{n}\left[n\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}},K_j}\right]\left[n\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}},K_j}\right]^T \\ &= \sum_{j=1}^{n}\left(\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}},K_j}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}},K_j}^T\right).\end{aligned} \tag{2.9}$$

As for the practical cases in which S is less than n^n, it is appropriate to use the formula (2.9) for the variance matrix for $\widetilde{\boldsymbol{\mu}}$. □

For a one-dimensional parameter of interest, its standard deviation is usually estimated as the sample standard deviation of its estimates from the bootstrap samples ($\widehat{\sigma}$). It has been shown that the standard deviation of bagging estimator ($\widetilde{\sigma}$) is no greater than $\widehat{\sigma}$ (Efron, 2014). This result also holds true for the variance of each component of the high-dimensional parameter of interest, as illustrated in the following theorem.

Theorem 2.2. *Denote* $\widehat{\boldsymbol{\mu}} \triangleq [\widehat{\mu}_1, \cdots, \widehat{\mu}_m, \cdots, \widehat{\mu}_d]^T$, *where* $m = 1, 2, \cdots, d$. *For the* m*-th component, stack its centered estimates from all the bootstrap samples into to a column vector* $\boldsymbol{\mu}_m^* = [\mu_{m1}^* - \mu_{m\cdot}^*, \cdots, \mu_{mb}^* - \mu_{m\cdot}^*, \cdots, \mu_{mS}^* - \mu_{m\cdot}^*]^T$, *where* $\mu_{m\cdot}^* = \frac{1}{S}\sum_{b=1}^{S}\mu_{mb}^*$ *and* $b = 1, 2, \cdots, S$. *Then the variance of* $\widehat{\mu}_m$ *is estimated by* $\boldsymbol{\Sigma}_{\widehat{\mu}_m\widehat{\mu}_m} = \frac{1}{S}\sum_{b=1}^{S}\left(\mu_{mb}^* - \mu_{m\cdot}^*\right)^2$.

Let $\widetilde{\boldsymbol{\mu}} \triangleq [\widetilde{\mu}_1, \cdots, \widetilde{\mu}_m, \cdots, \widetilde{\mu}_d]^T$*, where* $m = 1, 2, \cdots, d$*; and denote the variance of the* m*-th component of* $\widetilde{\boldsymbol{\mu}}$ *by* $\boldsymbol{\Sigma}_{\widetilde{\mu}_m \widetilde{\mu}_m}$ *according to Theorem 2.1. Then we have*

$$\boldsymbol{\Sigma}_{\widetilde{\mu}_m \widetilde{\mu}_m} \leq \boldsymbol{\Sigma}_{\widehat{\mu}_m \widehat{\mu}_m} \quad \textit{for } m = 1, 2, \cdots, d. \tag{2.10}$$

Proof. Consider first the ideal case, where the observed sample is resampled with replacement $S = n^n$ times.

Define an $S \times n$ matrix A as

$$A = [K_1^* - \mathbf{1}_n, K_b^* - \mathbf{1}_n, \cdots, K_S^* - \mathbf{1}_n]^T,$$

where $[K_b^* - \mathbf{1}_n] = [K_{b1}^* - 1, K_{b2}^* - 1, \cdots, K_{bn}^* - 1]^T$, $b = 1, 2, \cdots, S$.

As $K_b^* \overset{iid}{\sim} \text{Mult}_n(n, \boldsymbol{p}_0)$, we have

$$\text{var}(K^*) = \tfrac{1}{S} A^T A = \boldsymbol{I}_n - \mathbf{1}_n \mathbf{1}_n^T$$

and $\text{rank}(A) = n - 1$.

Accordingly the singular value decomposition (SVD) of A is $A = \sqrt{S} U V^T$, where U is an $S \times (n-1)$ orthonormal matrix and V is an $(n-1) \times (n-1)$ orthonormal matrix, for the singular values are all ones.

From Theorem 2.1, the variance of $\widetilde{\mu}_m$ is

$$\begin{aligned}
\boldsymbol{\Sigma}_{\widetilde{\mu}_m\widetilde{\mu}_m} &= \sum_{j=1}^{n} \left(\boldsymbol{\Sigma}_{\widehat{\mu}_m K_j} \boldsymbol{\Sigma}_{\widehat{\mu}_m K_j}\right) \\
&= \left(\frac{1}{S} A^T \boldsymbol{\mu}_m^*\right)^T \left(\frac{1}{S} A^T \boldsymbol{\mu}_m^*\right) \\
&= \frac{\left(\boldsymbol{\mu}_m^*\right)^T A A^T \boldsymbol{\mu}_m^*}{S^2} \\
&= \left(\boldsymbol{\mu}_m^*\right)^T \frac{\sqrt{S} U V^T V U^T \sqrt{S}}{S^2} \boldsymbol{\mu}_m^* \\
&= \frac{1}{S} \left(\boldsymbol{\mu}_m^*\right)^T U U^T \boldsymbol{\mu}_m^* \\
&= \frac{1}{S} \left\| U^T \boldsymbol{\mu}_m^* \right\|^2 .
\end{aligned} \tag{2.11}$$

Hence, S times the variance of $\widetilde{\mu}_m$ can be interpreted as the squared length of the projection of vector $\boldsymbol{\mu}_m^*$ onto the $(n-1)$-dimensional space that is spanned by the column vectors of matrix U.

Since $\boldsymbol{\Sigma}_{\widehat{\mu}_m\widehat{\mu}_m} = \frac{1}{S}\sum_{b=1}^{S}\left(\mu_{mb}^* - \mu_{m\cdot}^*\right)^2 = \frac{1}{S}\left\|\boldsymbol{\mu}_m^*\right\|^2$, S times the variance of $\widehat{\mu}_m$ can be interpreted as the squared length of $\boldsymbol{\mu}_m^*$. Therefore the claim is true, and the equal sign holds if $\boldsymbol{\mu}_m^*$ is in the $(n-1)$-dimensional space spanned by the column vectors of matrix U.

For the practical cases in which S is less than n^n, the variance of K^* needs be replaced with the sample variance, and the derivation is valid too.

□

2.3 Parametric Bootstrap Smoothing

Let $\boldsymbol{y}$ denote a p-dimensional vector of response variables, and $\boldsymbol{x}$ a q-dimensional vector of fixed values for the covariates. Let $(\boldsymbol{y}_j, \boldsymbol{x}_j)$ denote a sample point, where $j = 1, 2, \cdots, n$; n is the sample size. New points are drawn from the distribution that fits the observed data sample. The parametric bootstrap smoothing algorithm for the MLR is outlined as follows.

1. Select the optimal MLR model $\mathcal{M}$ for the observed data sample $\mathcal{L} = \{(\boldsymbol{y}_j, \boldsymbol{x}_j)\}$, $j = 1, 2, \ldots, n$.

2. Fit the selected MLR model $\mathcal{M}$ to the observed data sample to obtain the MLE's $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\Sigma}}$.

3. Generate the bootstrap samples $\mathcal{L}_b^* = \{(\boldsymbol{y}_j^*, \boldsymbol{x}_j)\}$, $j = 1, 2, \ldots, n$; $b = 1, 2, \ldots, S$; where $\boldsymbol{y}_j^*$ is drawn from the multivariate normal distribution $\mathcal{N}(\boldsymbol{x}_i^T \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Sigma}})$.

4. Select the optimal MLR model $\mathcal{M}_b^*$ for each bootstrap sample according to a model selection procedure.

5. Fit the selected MLR model $\mathcal{M}_b^*$ to each bootstrap sample to obtain the MLE's $\widehat{\boldsymbol{B}}_b^*$, $\widehat{\boldsymbol{\Sigma}}_b^*$.

6. Average the estimates of the parameter of interest across all the boot-

strapped optimal models to get the bagging estimator
$\widetilde{\boldsymbol{\mu}}(\boldsymbol{x}) = \frac{1}{S}\sum_{b=1}^{S} \widehat{\boldsymbol{\mu}}(\mathcal{M}_b^*(\boldsymbol{x}, \mathcal{L}_b^*))$.

We derive the covariance matrix for the smoothed parameter of interest $\widetilde{\boldsymbol{\mu}}(\boldsymbol{x})$ in a high dimensional space, which extends the formula in (Efron, 2014) from the one dimensional case to the high dimensional case. The coefficient matrix $\boldsymbol{B}$ is first vectorized by column into a vector $\boldsymbol{\beta} = [b_{11}, \cdots, b_{q1}, b_{21}, \cdots, \cdots, b_{qp}]^T$. We can assume that the distribution of estimator $\widehat{\boldsymbol{\beta}}$ belongs to an exponential family

$$f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}}) = \exp\{\boldsymbol{\theta}^T\widehat{\boldsymbol{\beta}} - \psi(\boldsymbol{\theta})\} f_0(\widehat{\boldsymbol{\beta}}), \tag{2.12}$$

where $\boldsymbol{\theta}$ is the qp-dimensional natural parameter vector, $\psi(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$ only, and $f_0(\widehat{\boldsymbol{\beta}})$ is the "carrying density" function of $\widehat{\boldsymbol{\beta}}$ only. This assumption holds true since $\widehat{\boldsymbol{\beta}}$ follows a multivariate normal distribution according to the assumptions of MLR models. The exponential family of such a form maintains the following properties:

$$\mathrm{E}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}}) = \frac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \boldsymbol{\beta}, \tag{2.13}$$

$$\mathrm{var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}}) = \frac{\partial^2\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2} = \frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\theta}}. \tag{2.14}$$

In the case of MLR bagging, $\widehat{\boldsymbol{\beta}}$ refers to the vectorized estimates of the coefficient matrix, which is the MLE of fitting the optimal model through the observed

data sample. And the estimate of $\widehat{\boldsymbol{\beta}}_b^*$ can be viewed as sampled from a MLE distribution $f_{\widehat{\boldsymbol{\theta}}}(.)$, i.e.,

$$f_{\widehat{\boldsymbol{\theta}}}(\cdot) \xrightarrow{i.i.d.} \widehat{\boldsymbol{\beta}}_1^*, \widehat{\boldsymbol{\beta}}_2^*, ..., \widehat{\boldsymbol{\beta}}_b^*, ..., \widehat{\boldsymbol{\beta}}_S^*. \tag{2.15}$$

And the smoothed estimate of parameter of interest can be expressed as a function of $\widehat{\boldsymbol{\beta}}_b^*$, i.e.,

$$\widetilde{\boldsymbol{\mu}}(\boldsymbol{x}) = \frac{1}{S}\sum_{b=1}^{S} \widehat{\boldsymbol{\mu}}(\boldsymbol{x}, \widehat{\boldsymbol{\beta}}_b^*). \tag{2.16}$$

Theorem 2.3. *Let* $\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}} \triangleq \mathrm{E}\left[(\widetilde{\boldsymbol{\mu}} - \mathrm{E}\,\widetilde{\boldsymbol{\mu}})(\widetilde{\boldsymbol{\mu}} - \mathrm{E}\,\widetilde{\boldsymbol{\mu}})^T\right]$ *be the covariance matrix of* $\widetilde{\boldsymbol{\mu}}$. *Let* $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\mu}}} \triangleq \mathrm{E}\left[(\widehat{\boldsymbol{\beta}} - \mathrm{E}\,\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\mu}} - \mathrm{E}\,\widehat{\boldsymbol{\mu}})^T\right]$ *be the covariance matrix between* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\boldsymbol{\mu}}$. *Then the parametric delta-method estimate of the covariance matrix for the smoothed estimator* $\widetilde{\boldsymbol{\mu}}$ *is*

$$\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}} = \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\mu}}}^T \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}^{-1} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\mu}}}. \tag{2.17}$$

Proof. Consider first the ideal case, $S \to \infty$. We apply the importance sampling technique to the smoothed estimate $\widetilde{\boldsymbol{\mu}}$ with respect to the distribution parame-

terized with $\boldsymbol{\theta}$,

$$\begin{aligned}\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}} &= \mathrm{E}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\mu}}) = \int \widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{\beta}}) f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}}) d\widehat{\boldsymbol{\beta}} \\ &= \int \frac{\widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{\beta}}) f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}})}{f_{\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\beta}})} f_{\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\beta}}) d\widehat{\boldsymbol{\beta}} = \frac{\sum_{b=1}^{S} \frac{\widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{\beta}}_b^*) f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}}_b^*)}{f_{\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\beta}}_b^*)}}{\sum_{b=1}^{S} \frac{f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}}_b^*)}{f_{\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\beta}}_b^*)}}.\end{aligned} \tag{2.18}$$

From the equation(2.12),

$$\begin{aligned}\frac{f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}}_b^*)}{f_{\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\beta}}_b^*)} &= \frac{\exp\{\boldsymbol{\theta}^T \widehat{\boldsymbol{\beta}}_b^* - \psi(\boldsymbol{\theta})\} f_0(\widehat{\boldsymbol{\beta}}_b^*)}{\exp\{\widehat{\boldsymbol{\theta}}^T \widehat{\boldsymbol{\beta}}_b^* - \psi(\widehat{\boldsymbol{\theta}})\} f_0(\widehat{\boldsymbol{\beta}}_b^*)} \\ &= \frac{\exp\{\boldsymbol{\theta}^T (\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}}) + \boldsymbol{\theta}^T \widehat{\boldsymbol{\beta}} - \psi(\boldsymbol{\theta})\}}{\exp\{\widehat{\boldsymbol{\theta}}^T (\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{\theta}}^T \widehat{\boldsymbol{\beta}} - \psi(\widehat{\boldsymbol{\theta}})\}} \\ &= \exp\{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T (\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}})\} \exp\{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T \widehat{\boldsymbol{\beta}} - \psi(\boldsymbol{\theta}) + \psi(\widehat{\boldsymbol{\theta}})\}.\end{aligned} \tag{2.19}$$

Plugging equation(2.19) into equation(2.18) gives

$$\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = \frac{\sum_{b=1}^{S} \exp\{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T (\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}})\} \widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{\beta}}_b^*)}{\sum_{b=1}^{S} \exp\{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T (\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}})\}}. \tag{2.20}$$

Let $\boldsymbol{\theta} \to \widehat{\boldsymbol{\theta}}$ and apply $e^x \approx 1 + x$, we get

$$\begin{aligned}\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}} &= \frac{\sum_{b=1}^{S} [1 + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T (\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}})] \widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{\beta}}_b^*)}{\sum_{b=1}^{S} [1 + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T (\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}})]} \\ &= \frac{1}{S} \sum_{b=1}^{S} [1 + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T (\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}})] \widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{\beta}}_b^*).\end{aligned} \tag{2.21}$$

Expand $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ at $\widehat{\boldsymbol{\theta}}$ and ignore the higher order terms as

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}} &= \widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\theta}}} + \left[(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T \frac{\partial \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\beta}}\right]\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \\
&= \widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\theta}}} + \left[(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}} \frac{\partial \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right]\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \\
&= \widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\theta}}} + \left[(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T [\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}}]^{-1} \frac{\partial \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right]\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \\
&= \widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\theta}}} + (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \left[V_{\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\beta}})\right]^{-1} \frac{1}{S}\sum_{b=1}^{S}(\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}})[\widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{\beta}}_b^*)]^T \\
&= \widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\theta}}} + (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}^{-1}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\mu}}}.
\end{aligned} \tag{2.22}$$

Hence, the covariance matrix of $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ is

$$\begin{aligned}
\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}} &= \mathrm{var}\left[(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}^{-1}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\mu}}}\right] \\
&= \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\mu}}}^T \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}^{-1} \mathrm{var}\left[(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right] \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}^{-1}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\mu}}} \\
&= \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\mu}}}^T \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}^{-1}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}^{-1}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\mu}}} \\
&= \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\mu}}}^T \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}^{-1}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\mu}}}.
\end{aligned} \tag{2.23}$$

For the practical cases where S is finite, the variance matrix of $\widetilde{\boldsymbol{\mu}}$ is obtained by simply replacing the corresponding covariance matrices with sample covariance matrices. □

For a one-dimensional parameter of interest, the sample standard deviation of

its estimates across the bootstrap samples is no greater than that of the bagging estimator (Efron, 2014). This conclusion is also valid for the variance of each component of the high-dimensional parameter of interest, as described in the following theorem.

Theorem 2.4. *Denote* $\widehat{\boldsymbol{\mu}} \triangleq [\widehat{\mu}_1, \cdots, \widehat{\mu}_m, \cdots, \widehat{\mu}_d]^T$, *where* $m = 1, 2, \cdots, d$. *For the* m*-th component, denote its centered estimates from all the bootstrap samples by a vector* $\boldsymbol{\mu}_m^* = [\mu_{m1}^* - \mu_{m\cdot}^*, \cdots, \mu_{mb}^* - \mu_{m\cdot}^*, \cdots, \mu_{mS}^* - \mu_{m\cdot}^*]^T$, *where* $\mu_{m\cdot}^* = \frac{1}{S}\sum_{b=1}^{S} \mu_{mb}^*$ *and* $b = 1, 2, \cdots, S$. *Then the variance of* $\widehat{\mu}_m$ *is estimated by* $\boldsymbol{\Sigma}_{\widehat{\mu}_m\widehat{\mu}_m} = \frac{1}{S}\sum_{b=1}^{S} (\mu_{mb}^* - \mu_{m\cdot}^*)^2$.

Let $\widetilde{\boldsymbol{\mu}} \triangleq [\widetilde{\mu}_1, \cdots, \widetilde{\mu}_m, \cdots, \widetilde{\mu}_d]^T$, *where* $m = 1, 2, \cdots, d$; *and denote the variance of the* m*-th component of* $\widetilde{\boldsymbol{\mu}}$ *by* $\boldsymbol{\Sigma}_{\widetilde{\mu}_m\widetilde{\mu}_m}$ *according to Theorem 2.3. Then we have*

$$\boldsymbol{\Sigma}_{\widetilde{\mu}_m\widetilde{\mu}_m} \leq \boldsymbol{\Sigma}_{\widehat{\mu}_m\widehat{\mu}_m} \quad \text{for } m = 1, 2, \cdots, d. \tag{2.24}$$

Proof. Consider first the ideal case, $S \to \infty$.

Define an $S \times qp$ matrix A as

$$A = \left[\widehat{\boldsymbol{\beta}}_1^* - \widehat{\boldsymbol{\beta}}, \cdots, \widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}}, \cdots, \widehat{\boldsymbol{\beta}}_S^* - \widehat{\boldsymbol{\beta}}\right]^T,$$

where $\widehat{\boldsymbol{\beta}}$ is the column-wise vectorization of coefficient matrix $\widehat{\boldsymbol{B}}$ that is the MLE for $\boldsymbol{B}$ fitted through the observed data sample, and $\widehat{\boldsymbol{\beta}}_b^*$ is with respect to the

bootstrap sample, $b = 1, 2, \cdots, S$.

As $S \to \infty$, then $\frac{1}{S} A^T \mathbf{1}_S \xrightarrow{P} \mathbf{0}$, and $\frac{1}{S} \sum_{b=1}^{S} \widehat{\boldsymbol{\beta}}_b^* \xrightarrow{P} \widehat{\boldsymbol{\beta}}$.

Apply SVD to matrix A as

$$A = UDV^T,$$

where U is an $S \times qp$ orthonormal matrix, D is a $qp \times qp$ diagonal matrix with the singular values at the diagonal entries, and V is a $qp \times qp$ orthonormal matrix.

By Theorem 2.3, the variance of $\widetilde{\mu}_m$ is

$$\begin{aligned}
\boldsymbol{\Sigma}_{\widetilde{\mu}_m \widetilde{\mu}_m} &= \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}} \widehat{\mu}_m}^T \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}} \widehat{\boldsymbol{\beta}}}^{-1} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}} \widehat{\mu}_m} \\
&= \frac{1}{S} \left[\boldsymbol{\mu}_m^*\right]^T A \left[\frac{1}{S} A^T A\right]^{-1} A^T \boldsymbol{\mu}_m^* \frac{1}{S} \\
&= \frac{1}{S} \left[\boldsymbol{\mu}_m^*\right]^T A \left[A^T A\right]^{-1} A^T \boldsymbol{\mu}_m^* \\
&= \frac{1}{S} \left[\boldsymbol{\mu}_m^*\right]^T UDV^T \left[VDU^T UDV^T\right]^{-1} VDU^T \boldsymbol{\mu}_m^* \\
&= \frac{1}{S} \left[\boldsymbol{\mu}_m^*\right]^T UU^T \boldsymbol{\mu}_m^* \\
&= \frac{1}{S} \left\| U^T \boldsymbol{\mu}_m^* \right\|^2.
\end{aligned} \tag{2.25}$$

Geometrically speaking, S times the variance of $\widetilde{\mu}_m$ is the squared length of the projection of vector $\boldsymbol{\mu}_m^*$ onto the qp-dimensional space that is spanned by the column vectors of matrix U.

Since $\boldsymbol{\Sigma}_{\widehat{\mu}_m \widehat{\mu}_m} = \frac{1}{S} \sum_{b=1}^{S} \left(\mu_{mb}^* - \mu_{m\cdot}^*\right)^2 = \frac{1}{S} \left\|\boldsymbol{\mu}_m^*\right\|^2$, then S times the variance of $\widehat{\mu}_m$ is the squared length of $\boldsymbol{\mu}_m^*$. Therefore, the claim is true, and the equal

sign holds if $\boldsymbol{\mu}_m^*$ is already in the qp-dimensional space spanned by the column vectors of matrix U.

For the practical cases in which S is finite, the derivation applies too.

□

2.4 Simulation

We carry out the simulations to evaluate the performance of the bagging estimator with various model selection methods in the context of multivariate linear regression, by comparing its results with those of the standard method. Both a nonparametric example and a parametric example are investigated.

2.4.1 A Nonparametric Example

A polynomial of degree three is chosen as the linear regression equation for either response variable, and candidate polynomials of degree up to six are examined by the model selection procedures. The coefficients of (10, 1.5, 1.0, 0.1, 0, 0, 0) are preselected for the polynomial for the first response, while (20, 2.0, 1.2, 0.2, 0, 0, 0) is for the second response. The total of 200 values of the independent variable X are sampled from the uniform distribution with support of $[-3, 3]$, and are treated as fixed values within the parameter estimation procedure. The additive random errors in response variables are sampled from a bivariate normal distribution that

is implemented in the R function mvnorm. Both error components are with the variance of 6.25, while the correlation between them is 0.5.

Several model selection methods are employed, i.e., AIC, Mallows' C_p, BIC, hypothesis testing, and LASSO. The informatic scores of AIC, Mallows' C_p and BIC are used to determine the highest degree of the polynomials. The degree of the polynomial is determined as the one that gives the highest score. The significance level of 5% is used in the hypothesis testings, and variables with p-value less than 0.05 are selected into the model. The LASSO algorithm in the R package glmnet is used to select the subset of variables that result in the maximum deviance ratio.

The simulated data sample is resampled with replacement 4000 times. Each bootstrap sample is of the same size as the original data sample, and is the training data set for a model selection procedure to select the optimal model. Then the coefficients of the selected model are estimated before the parameters of interest are computed, e.g., the fitted center and its covariance matrix, the confidence region, etc.

With each of the aforementioned model selection procedures being applied, the performances of the standard method and the bagging method are compared in terms of the fitted value, the variance of the estimate, and the coverage probability. Twenty-one values of X that evenly cover the interval $[-2.95, 2.95]$ are chosen

for the comparison purpose. The fitted values with respect to those X values are shown in Figure 2.1. The standard deviations of the estimates are plotted in Figure 2.2. We consider the 95% confidence regions for both methods, and Figure 2.3 is for the coverage probabilities of the confidence regions constructed from the results of both estimators. There are slight differences in the fitted centers from these two estimators, as the standard estimator fits the optimal model through the observed data sample while the bagging estimator averages over the variants of the data sample. The standard deviations of the bagging estimates are always less than or equal to those of the standard estimates. This is because the variance of the bagging estimator is a component of the sample variance of the estimates across the bootstrap samples, while the variance of a standard estimator is simply the sample variance of the estimates across all the bootstrap samples. The coverage probability of the bagging confidence region is equivalent with that of the standard confidence region. The model selection procedures appear to cause noticeable differences in the fitted values and the estimated standard deviations, as the model selection procedures tend to favor different subsets of variables.

As an example, we look into the results for the x value of -0.295, which is close to the minimum x value of -3.0 for this numerical study. The results are listed in Table 2.1 for each model selection method. It is interesting to note that BIC,

Mallows' C_p and hypothesis testing give same fitted values since they pick the same polynomial degree for the simulated data sample. However, the standard deviations are different because the variations in the bootstrap samples come into play and disturb the decisions of these three model selection procedures.

2.4.2 A Parametric Example

There are two response variables and five candidate covariates in the regression model. The true coefficients $[1, 0, 1, 0, 1]^T$ are for the first response variable, and $[1.1, 0, 1.3, 0, 0.9]^T$ for the second. The design matrix is drawn from the multivariate normal distribution with the mean of $[0, 0, 0, 0, 0]^T$ and the covariance matrix

$$\begin{bmatrix} 1.00 & 0.93 & 0.66 & 0.90 & 0.76 \\ 0.93 & 1.00 & 0.81 & 0.99 & 0.91 \\ 0.66 & 0.81 & 1.00 & 0.88 & 0.63 \\ 0.90 & 0.99 & 0.88 & 1.00 & 0.88 \\ 0.76 & 0.91 & 0.63 & 0.88 & 1.00 \end{bmatrix}.$$

A sample of 200 points is drawn from this distribution as the given (or the observed) data sample. The additive random errors are draw from the bivariate normal distribution with the mean of $[0, 0]^T$, the variance of 6 for either component, and the correlation coefficient of 0.6 between these two components. Accordingly,

the given data sample (or the "observed" data sample) is constructed by adding up the true means (multiplying the design matrix with the true coefficient matrix) and the random errors.

The model selection methods that are based on AIC, Mallows' C_p, BIC, hypothesis testing, and LASSO, are employed to choose covariates for the optimal models. For the informatic model selection criteria AIC, Mallows' C_p, BIC, all non-empty subsets of the five candidate covariates are examined, and the subset with the best score is chosen as the variables in the optimal model. Therefore, 31 subsets of covariates are checked for each data sample. The significance level of 5% is applied in the hypothesis testings, and variables with p-value less than 0.05 are selected into the model. The LASSO algorithm in the package glmnet is used. The R functions output 100 candidate sets of coefficients together with deviance. The BIC scores are thus computed from the output deviance. The set of coefficients that results in the minimum value for the BIC score is selected.

The full model with all the five candidate covariates is fitted through the given data sample for the fitted mean of each sample point and the fitted covariance matrix. A bootstrap sample is obtained by drawing points from the bivariate normal distributions with the previously fitted mean values and covariance matrix. The optimal subset of candidate covariates is determined by the model selection methods, and the values for the chosen variables are estimated thereafter. A total

of 4000 bootstrap samples are drawn and inputted into model selection procedures. And then the parameters of interest are computed, e.g., the fitted center and its covariance matrix, the confidence region, etc.

At point $\boldsymbol{x}_0 \triangleq [2.377, 2.377, 2.377, 2.377, 2.377]^T$, the estimates for the mean response, the standard deviations for the estimates, and the coverage probability are calculated by using the bagging method and the standard bootstrap estimator. The component of $\boldsymbol{x}_0$ is close to the maximum value of the observed sample being used in the simulation study. At points close to the minimum convex hull that contains the whole sample points of covariates, the model selection uncertainty is usually larger than that at the points close to the center of the hull. This phenomenon is also noticed in (Efron, 2014), as a point close to the minimum covariate value is chosen to demonstrate the fluctuation of selected polynomial model. Similarly, a 6-dimensional point with all components being -2.5 is chosen for checking the fitted mean response and its variance in (Wang, Sherwood, and Li, 2014), where the values for each covariate are drawn independently from the standard normal distribution.

The estimation results at point $\boldsymbol{x}_0$ are listed in Table 2.2. There is no significant difference in the fitted centers among those model selection methods. However, the standard deviations estimated with AIC are larger than those with other model selection methods. This is because the AIC method tends to select

more variables and thus overfit the bootstrap samples in this high-correlation case, while other model selection methods tend to select parsimonious models with less severity of collinearity. Table 2.3 lists the percentage of each candidate covariate being selected across all the bootstrap samples.

Table 2.1: Comparison of the Estimation Results from the Standard Estimator with those from the Nonparametric Bagging Estimator under Various Model Selection Methods

Model Selection	Estimation Method	Fitted Center	Standard Deviation	Coverage Probability
AIC	bagging	(12.50, 18.77)	(0.91, 0.80)	0.91
	standard	(12.67, 18.91)	(1.00, 0.92)	0.95
BIC	bagging	(12.16, 19.13)	(0.75, 0.84)	0.90
	standard	(11.88, 18.86)	(0.87, 0.96)	0.93
C_p	bagging	(12.05, 19.40)	(0.62, 0.95)	0.92
	standard	(11.88, 18.86)	(0.72, 1.09)	0.95
Hypothesis	bagging	(12.40, 18.89)	(0.85, 0.78)	0.90
Testing	standard	(11.88, 18.86)	(0.95, 0.89)	0.92
LASSO	bagging	(12.41, 18.48)	(1.00, 0.93)	0.93
	standard	(12.40, 18.89)	(1.05, 0.99)	0.94

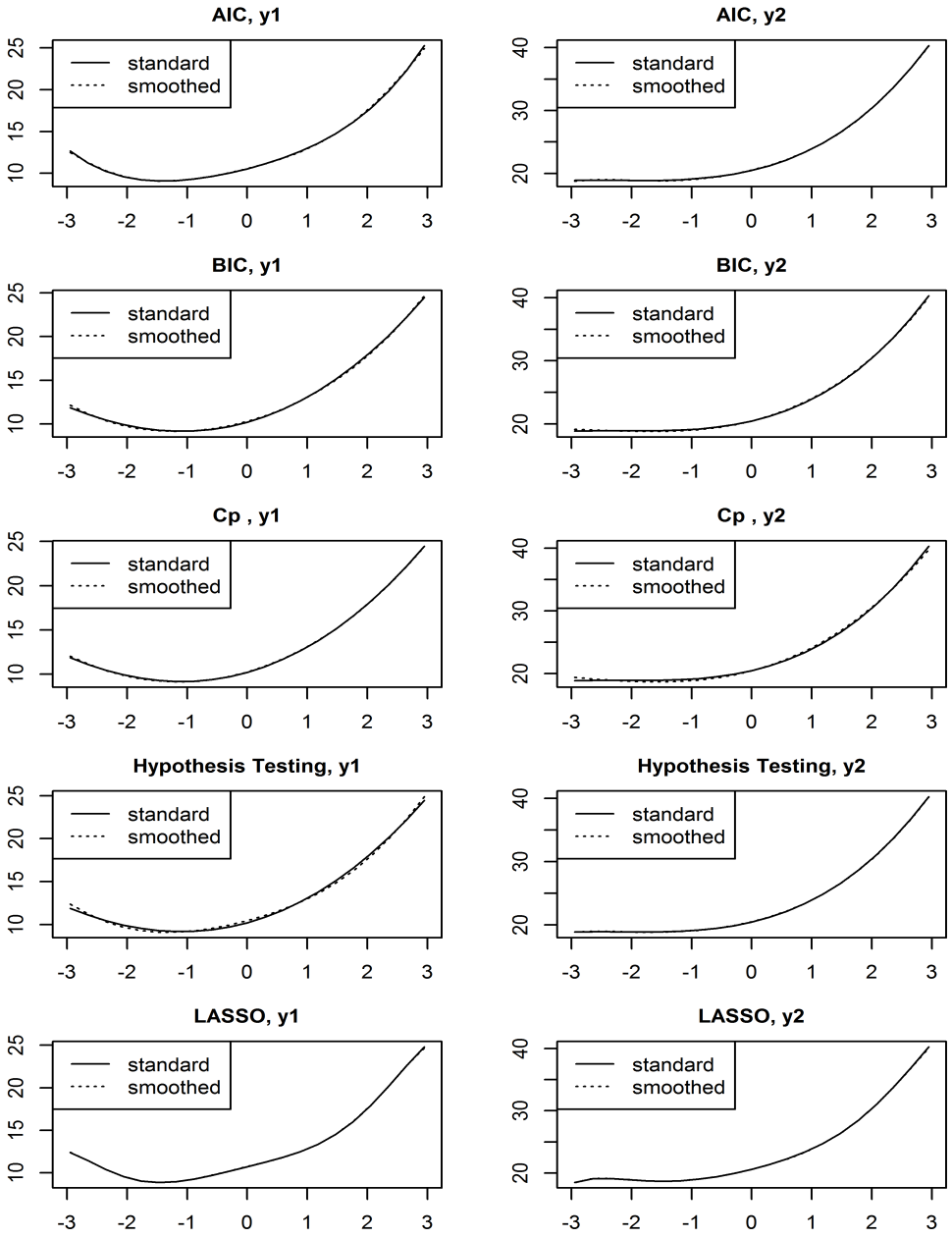

Figure 2.1: Comparison of the Centers Estimated by the Standard Estimator and those by the Nonparametric Bagging Estimator

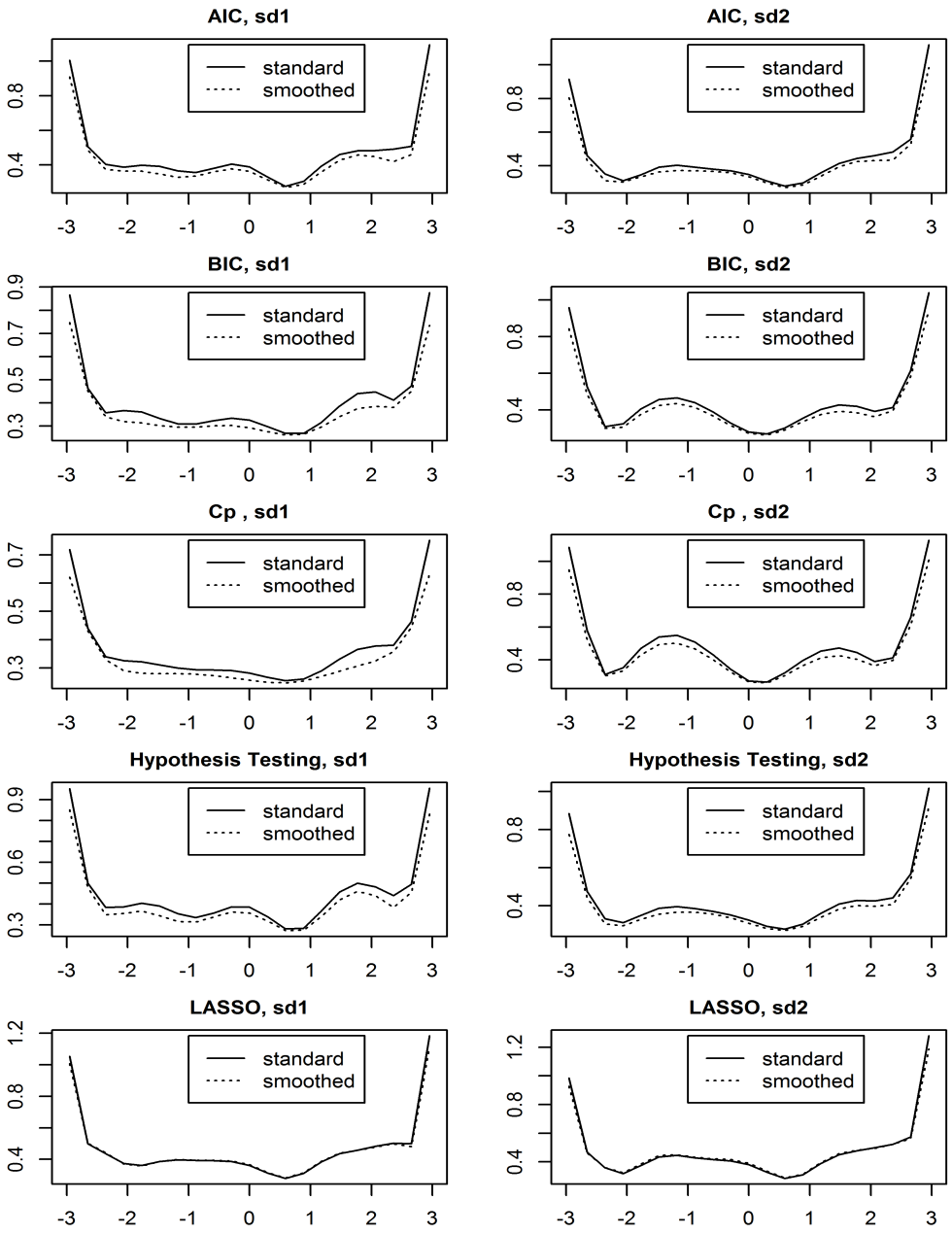

Figure 2.2: Comparison of the Standard Deviations Estimated by the Standard Estimator and those by the Nonparametric Bagging Estimator

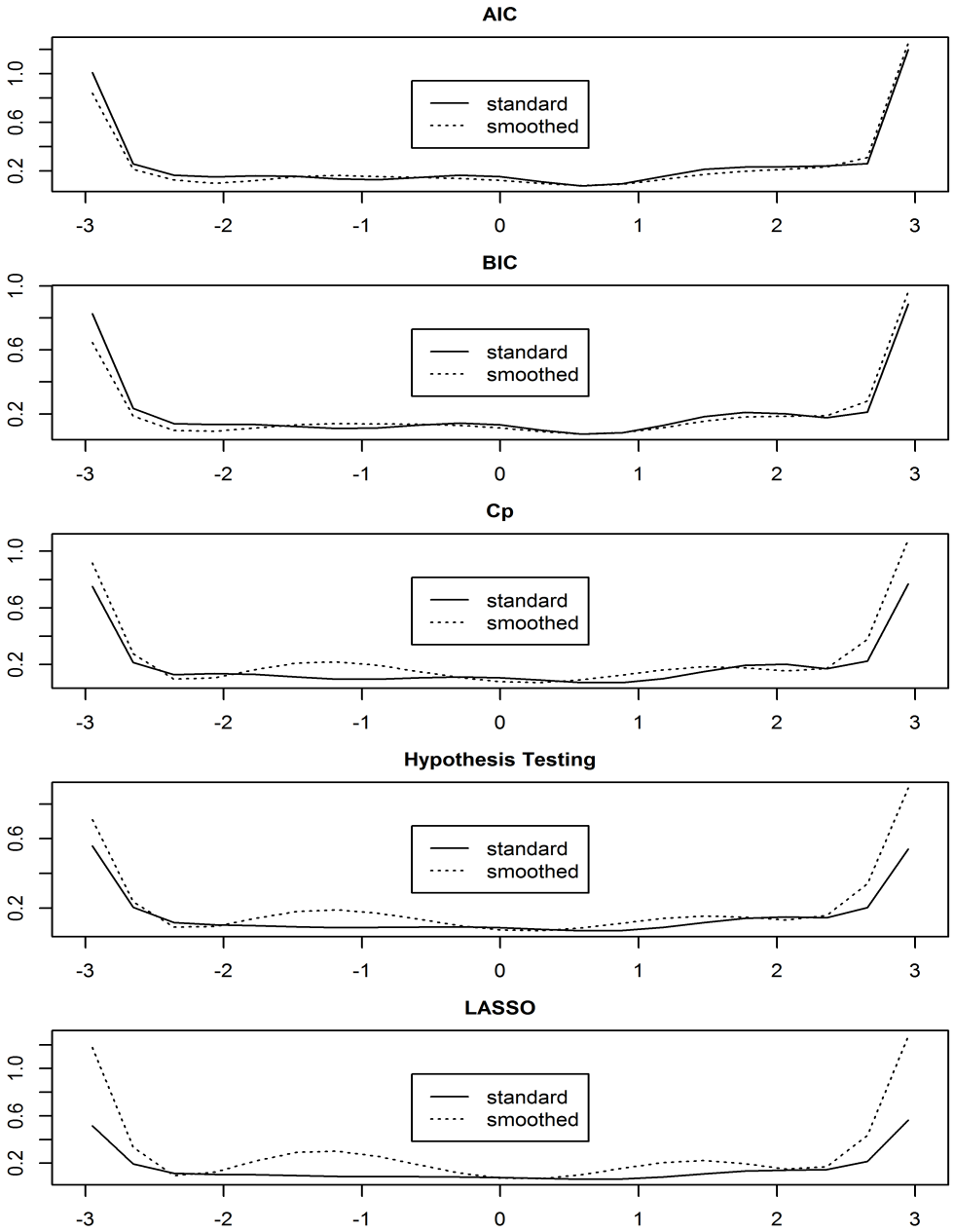

Figure 2.3: Comparison of the Coverage Probabilities Estimated by the Standard Estimator and those by the Nonparametric Bagging Estimator

Table 2.2: Comparison of the Estimation Results from the Standard Estimator and those from the Parametric Bagging Estimator with Various Model Selection Methods

Model Selection	Estimation Method	Fitted Center	Standard Deviation	Coverage Probability
AIC	bagging	(6.25, 6.69)	(0.97,1.09)	0.90
	standard	(6.06, 7.06)	(0.98,1.31)	0.91
BIC	bagging	(6.61, 7.30)	(0.53,0.58)	0.95
	standard	(6.65, 7.24)	(0.66,0.58)	0.97
C_p	bagging	(7.01, 7.54)	(0.52,0.47)	0.96
	standard	(6.92, 7.42)	(0.62,0.65)	0.97
Hypothesis	bagging	(6.78, 7.37)	(0.59,0.60)	0.93
Testing	standard	(6.55, 7.19)	(0.86,0.74)	0.95
LASSO	bagging	(6.47, 7.07)	(0.53,0.50)	0.93
	standard	(6.47, 7.07)	(0.53,0.54)	0.95

Table 2.3: Percentage of Candidate Covariates being Selected during the Parametric Bootstrapping Example (Only Covariates V1,V3,V5 are Included in the True Model)

Model Selection	V1	V2	V3	V4	V5
AIC	43	81	40	44	47
BIC	18	65	18	20	27
C_p	15	45	22	63	20
Hypothesis Testing	100	100	24	24	25
LASSO	0	100	100	0	100

3 Discussion

3.1 Summary

In this research, the data analytic approach has been investigated for integrating model selection uncertainty into the statistical inferences of an estimator. For the high dimensional bagging estimator, the formulae for covariance matrices are derived, and the confidence regions are constructed and evaluated.

In Chapter 1, the consequences of ignoring model selection uncertainty have been explained. Some primary approaches to tackling this problem have been reviewed. The background theoretical results of multivariate linear regression, model selection criteria, and the nonparametric delta method have been outlined. Properties of the influence function have been derived for the high dimensional linear functionals.

In Chapter 2, the derivation of covariance matrices for the high dimensional bagging estimators are elucidated for both the nonparametric and the parametric

cases. The performances of bagging estimators are analyzed empirically though simulation studies in the context of multivariate linear regression. The effects of some model selection methods have bee compared. The simulations have shown that the model selection may influence on the accuracy of bagging estimators. The derived covariance matrices perform better than the standard one under all the the model selection procedures considered in this research.

3.2 Future Work

Bootstrap smoothing can be potentially applied in the context of other regression models. For example, (Wang, Sherwood, and Li, 2014) have applied it to Poisson regression and nonparametric regression with spline basis functions. (Shang and Cavanaugh, 2008) investigate variants of bootstrapping schemes in the context of mixed models. It is possible to apply the bootstrap smoothing to log-linear models, linear mixed models, etc.

The new method of estimating covariance matrix may be applied to a bagging M-estimator. (Hu, 2001) analyzed the properties of a resampling M-estimator in the linear models. Bagging M-estimators in the context of MLR models may improve the accuracy and robustness of regression models against outliers and/or high-level noises in a data sample.

It is interesting to compare the bootstrap smoothing approach with the asymp-

totic approach (Hjort and Claeskens, 2003) in the context of multivariate linear regression, as both take advantage of the intrinsic properties of model selection procedures.

New model selection methods can also be combined with bootstrap smoothing, for example, SCAD (Wang, Sherwood, and Li, 2014), ALASSO (Gupta and Lahiri, 2014), etc.

Bibliography

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723.

[2] R. Berk *et al.* Valid Post-selection inference. *The Annals of Statistics* 41.2 (2013), pp. 802–837.

[3] P. Bühlmann and B. Yu. Analyzing Bagging. *The Annals of Statistics* 30.4 (2002), pp. 927–961.

[4] P. Bickel. Parametric robustness: small biases can be worthwhile. *The Annals of Statistics* 12.3 (1984), pp. 864–879.

[5] L. Breiman. Bagging Predictors. *Machine Learning* 24.2 (1996), pp. 123–140.

[6] A. Buja and W. Stuetzle. Observations on bagging. *Statistica Sinica* 16 (2006), pp. 323–351.

[7] C. ChatField. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 158.3 (1995), pp. 419–466.

[8] T. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical Science* 11.3 (1996), pp. 189–228.

[9] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 7.1 (1979), pp. 1–26.

[10] B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association* 82.397 (1987), pp. 171–185.

[11] B. Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association* 109.507 (2014), pp. 991–1007.

[12] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Boca Raton, Florida, USA: Chapman and Hall CRC Press, 1993.

[13] T. Ferguson. *A course in large sample theory*. Boca Raton, Florida, USA: Chapman and Hall CRC Press, 1996.

[14] Y. Fujikoshi and K. Satoh. Modified AIC and Cp in multivariate linear regression. *Biometrika* 84.3 (1997), pp. 707–716.

[15] A. Gelman and A. Vehtari. Comment on Estimation and accuracy after model selection. *Journal of the American Statistical Association* 109.507 (2014), pp. 1015–1016.

[16] S. Gupta and S. Lahiri. Comment on Estimation and accuracy after model selection. *Journal of the American Statistical Association* 109.507 (2014), pp. 1013–1015.

[17] N. Hjort. Comment on Estimation and accuracy after model selection. *Journal of the American Statistical Association* 109.507 (2014), pp. 1017–1020.

[18] N. L. Hjort and G. Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association* 98.464 (2003), pp. 879–899.

[19] F. Hu. Efficiency and robustness of a resampling M-estimator in the linear model. *Journal of Multivariate Analysis* 78 (2001), pp. 252–271.

[20] A. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. 233 Spring Street, New York, NY 10013, USA: Springer Science+Business Media, 2008.

[21] R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. New Jersey, USA: Pearson Prentice Hall, 2007.

[22] P. Kabaila. Valid confidence intervals in regression after variable selection. *Econometric Theory* 14.4 (1998), pp. 463–482.

[23] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795.

[24] S. Kullback and R. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86.

[25] C. L. Mallows. Some Comments on Cp. *Technometrics* 15.4 (1973), pp. 661–675.

[26] D. Politis. Comment on Estimation and accuracy after model selection. *Journal of the American Statistical Association* 109.507 (2014), pp. 1010–1013.

[27] B. Pötscher. Effects of model selection on inference. *Econometric Theory* 7.2 (1991), pp. 163–185.

[28] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics* 6.2 (1978), pp. 461–464.

[29] J. Shang and J. E. Cavanaugh. An assumption for the development of bootstrap variants of the Akaike information criterion in mixed models. *Statistics & Probability Letters* 78.12 (2008), pp. 1422–1429.

[30] J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association* 88.422 (1993), pp. 486–494.

[31] J. Shao. Bootstrap model selection. *Journal of the American Statistical Association* 91.434 (1996), pp. 655–665.

[32] K. Singh. On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics* 9.6 (1981), pp. 1187–1195.

[33] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society* 58.1 (1996), pp. 267–288.

[34] L. Wang, B. Sherwood, and R. Li. Comment on Estimation and accuracy after model selection. *Journal of the American Statistical Association* 109.507 (2014), pp. 1007–1010.

[35] L. Wasserman. *All of nonparametric statistics*. 233 Spring Street, New York, NY 10013, USA: Springer Science+Business Media, 2006.

[36] P. Westfall and S. Young. *Resampling-based multiple testing: examples and methods for p-value adjustment*. USA: Wiley-Interscience, 1993.