

Bootstrap Aggregating (Bagging) with Ranked Set Sampling

by

Madhawa K. B. Hulangamuwa

A Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Statistics
University of Manitoba
Winnipeg

Copyright © 2026 by Madhawa K.B. Hulangamuwa

Abstract

Regression models can exhibit considerable variability in their predictions when fitted using limited training data, leading to unstable predictions. This study combines Ranked Set Sampling (RSS) with bootstrap aggregating (bagging) to improve the predictive power of regression models trained on limited data. Three strategies are examined: (i) classical bagging under Simple Random Sampling (SSB), (ii) rank-wise bagging with RSS training data (RSB), and (iii) retrospective SRS-RSS bagging via bootstrap (SRB). Theoretical results establish the unbiasedness of the proposed predictors and derive closed-form expressions for their conditional and unconditional variances. Variance comparisons explicitly demonstrate that using rank information leads to provable variance reduction of bagged models relative to their SSB counterparts, while the proposed SRB framework closely approximates the efficiency of RSB without requiring an RSS training sample. Proposed methods are evaluated using extensive numerical analyses across multiple linear and nonlinear associations between the response and covariates, under paired and residual bootstrap settings. Linear mixed models, LASSO regularization, ordinary linear regression, and spline regression are employed as base learners within the proposed rank-enriched bagging framework.

Acknowledgment Page

Foremost, I would like to express my sincere gratitude to my supervisor, Dr. Mohammad Jafari Jozani, for his unwavering guidance and support. I am deeply thankful for his encouragement and will always value the knowledge I have gained while working with him over the past two years.

I would also like to thank Dr. Alexandre Leblanc and Dr. Lei Ding for their valuable feedback. I greatly appreciate the time and effort they devoted to reviewing this thesis.

Lastly, I would like to extend my thanks to everyone in the Department of Statistics who contributed to making my time in the Master's program so meaningful. Whether through teaching or organizing student events, the experiences and memories I have gained will remain with me for years to come.

Use of Generative Artificial Intelligence. Generative AI tools were used during the preparation of this dissertation for editorial assistance, including grammar checking, text restructuring, and L^AT_EX formatting. All mathematical content, proofs, algorithmic designs, experimental results, and scientific interpretations are the original work of the author. The author reviewed and takes full responsibility for all content in this dissertation.

Dedication Page

To my family and loved ones, who stood by me through every challenge and celebrated each success.

Contents

Contents	iii
List of Tables	vii
List of Figures	viii
1 Introduction	1
1.1 Motivation	1
1.2 Ensemble Learning	6
1.3 Bootstrapping	7
1.3.1 Paired Bootstrapping	9
1.3.2 Residual Bootstrapping	9
1.4 Bagging	10
1.5 Bagging with SRS (SSB)	15
1.6 Regression Estimation with SSB	19
1.6.1 Linear Functional Representation of Regression Estimators under SRS	19
1.6.2 Linear Smoother SSB Regression Estimator	23
1.7 Research Statement and Thesis Outline	24

2	RSS Bagging	26
2.1	Introduction	26
2.2	Ranked Set Sampling	27
2.3	RSS for Estimating Population Means	30
2.4	Ranking Mechanisms	31
2.5	Estimation of Means using RSS	33
2.6	Bagging with Ranked Set Sampling (RSB)	36
2.7	Regression Estimation with RSB	41
2.7.1	Linear Functional Representation of Regression Estimators under RSS	42
2.7.2	Linear Smoother RSB Regression Estimator	43
2.8	Modeling Approach	46
2.8.1	Weighted Least Squares with Adjusted Response Variable using RSB	49
2.8.2	Multilevel Modeling Approach toward Regression Analysis with RSB	52
2.9	Summary and Conclusion	55
3	Rank-Informed Bagging	57
3.1	Introduction	57
3.2	Bagging under SRS-RSS (SRB)	58
3.3	Regression Estimation with SRB	73
3.3.1	Linear Smoother SRB Regression Estimator	73
3.4	Summary and Conclusion	76

4	Numerical Study	78
4.1	Data Generation	78
4.1.1	Single Ranker RSS Construction	79
4.1.2	Weighted RSS Construction for Multiple Ranks	79
4.2	Bootstrapping on RSS	81
4.2.1	RSS Paired Bootstrapping	81
4.2.2	RSS Residual Bootstrapping	82
4.3	RSS Test Data Generation	85
4.4	SRS-RSS Construction with Paired Bootstrapping	88
4.5	SRS-RSS Construction with Residual Bootstrapping	89
4.5.1	RSS Residual Bootstrapping Extension to SRS-RSS	89
4.5.2	SRS-RSS Rank-Informed Residual Bootstrapping	89
4.6	Estimation and Prediction using Linear Models	92
4.6.1	Single Covariate with Single Ranker	92
4.6.2	Single Covariate with Multiple Rankers	96
4.7	Estimation and Prediction using Spline Models	99
4.7.1	Single Covariate Spline Regression with Multiple Rankers	104
4.7.2	Additive Spline Regression with Multiple Rankers	109
4.8	Real Data: Body Fat	111
4.9	Summary and Conclusion	119

5 Conclusion and Future Work	121
5.1 Summary	121
5.2 Future Work	123
Bibliography	126

List of Tables

4.1	Prediction relative efficiencies (RE_{pred}) of fitted simple linear regression models using the paired bootstrap in the single-ranker setting	94
4.2	Prediction relative efficiencies (RE_{pred}) of fitted simple linear regression models using the residual bootstrap in the single-ranker setting	95
4.3	Prediction relative efficiencies (RE_{pred}) of fitted simple linear regression models in the multiple-rankers setting	98
4.4	Prediction relative efficiencies (RE_{pred}) of fitted single covariate spline regression models using the paired bootstrap in the single-ranker setting	103
4.5	Prediction relative efficiencies (RE_{pred}) of fitted single covariate spline regression models using the residual bootstrap in the single-ranker setting	104
4.6	Prediction relative efficiencies (RE_{pred}) of fitted single covariate spline regression models in the multiple-rankers setting	108
4.7	Prediction relative efficiencies (RE_{pred}) of LASSO-regularized additive fitted spline regression models in the multiple-rankers setting	110
4.8	Relative efficiencies (RE) of the LASSO-regularized fitted multiple linear regression models in estimating body fat percentage (%BF)	116
4.9	Prediction relative efficiencies (RE_{pred}) of the LASSO-regularized fitted multiple linear regression models in estimating body fat percentage (%BF)	118

List of Figures

1.1	Illustration of the standard bagging procedure	11
2.1	Illustration of the RSB procedure for RSS data with set size k under paired bootstrapping.	45
3.1	Illustration of the SRB procedure for SRS data with set size 3 under paired bootstrapping.	75
4.1	Illustration of the RSB procedure for RSS data with set size k under residual bootstrapping.	85
4.2	Illustration of the SRB procedure for SRS data with size $n = mk$ under rank-informed residual bootstrapping.	91
4.3	Single covariate fitted spline regression models obtained using the paired bootstrap framework for the single-ranker setting	101
4.4	Single covariate fitted spline regression models obtained using the residual bootstrap framework for the single-ranker setting	102
4.5	Single covariate fitted spline regression models obtained using the paired bootstrap framework for the multiple-ranker setting	106
4.6	Single covariate fitted spline regression models obtained using the residual bootstrap framework for the multiple-ranker setting	107

4.7	Distribution of body fat (%)	115
-----	--	-----

Chapter 1

Introduction

This chapter introduces the motivation and objectives of the thesis. It gives an overview of ensemble learning and its key theoretical properties. Particular attention is given to bootstrap aggregating (bagging), including its theoretical foundation and practical significance. We study the unbiasedness and variance of bagged predictors in both conditional and unconditional settings. These results are first established for mean-type functionals and are subsequently extended to regression functions of the form $m(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$ at a fixed point $\mathbf{x}$. The chapter concludes with an outline of the thesis.

1.1 Motivation

Machine learning techniques are often data-hungry and require substantial training data to make accurate predictions and capture meaningful patterns (Adadi, 2021). Without sufficient data to represent the underlying distribution, these models suffer from overfitting, in which they may memorize the limited training data rather than identify underlying patterns (Ying, 2019). Data scarcity can also lead to high variance, where small changes in the training dataset cause significant fluctuations in the model's predictions (Alzubaidi et al., 2023). Moreover, insufficient data may lead to fitted

models that generalize poorly to unseen test samples (Aliferis and Simon, 2024). These issues tend to produce unreliable results and degrade performance (Aliferis and Simon, 2024). However, obtaining large datasets is often expensive, time-consuming, or even impractical, particularly in domains such as medicine with rare diseases (Shaikhina and Khovanova, 2017). Clinical trials are costly and slow, and ethical constraints often limit data sharing due to concerns about patient privacy. In fields like agriculture, experiments often require substantial time and resources, and samples show high natural variability due to weather patterns and soil conditions (Le et al., 2025). Financial datasets are strictly regulated and sensitive, further restricting sample availability (Wang and Tsai, 2022). In ecological studies, long-term observations may take decades to generate datasets of meaningful size (Le et al., 2025). The development of new algorithms and accurate prediction models for such problems represents significant methodological challenges across many scientific and industrial domains.

Modern statistical learning strategies address these issues by employing or developing models tailored to limited data, including ensemble learning techniques, which are particularly effective at mitigating high variance and improving predictive stability. Ensemble learning performs prediction by fusing the outputs of multiple base classifiers (Dietterich, 2000). In small sample scenarios, individual models often fail to capture the underlying data-generating structure. The core principle behind ensemble learning is the “wisdom of the crowd”, the idea that aggregating decisions from a diverse set of models leads to better performance than relying on a single model (Hastie et al., 2009). This principle is particularly important in settings where individual complex models often exhibit high variance. In such contexts, constructing a diversified ensemble of simple base models is often more effective than using a single complex model (Wang and Yao, 2009).

One of the most influential and classical ensemble techniques is bagging (bootstrap

aggregating), introduced by Breiman (Breiman, 1996). Bagging constructs multiple base learners by training them on different bootstrap samples of an available training dataset. The final prediction is obtained by averaging over the outputs of the base learners. Bagging is particularly effective for high-variance, unstable learners, because it substantially reduces variance without increasing bias (Breiman, 1996). However, the reliability and performance of bagged models depend on the variability among the base learners and the degree to which bootstrap samples are representative of the underlying population (Bühlmann and Yu, 2002). Although bagging has shown promising results on large datasets, its performance on small datasets poses unique challenges (Bühlmann and Yu, 2002). For example, if bootstrap samples fail to capture sufficient variation or structure in the original data, the ensemble may exhibit poor generalization. This limitation motivates our thesis question: whether we can make the sampled data itself more informative before applying bagging. Our proposed framework in this thesis provides a solution to this problem.

Bootstrap sampling, introduced by Efron in 1979 (Efron and Tibshirani, 1994), has become a foundational resampling technique that helps researchers approximate the sampling distribution of estimators of population parameters. However, on average, only about 63.2% of the original observations are included in a given bootstrap sample, while the remaining 36.8% are excluded. As a result, each bootstrap sample reflects only a partial view of the complete dataset, which may limit the representativeness of learners trained on it; this issue is exacerbated when the dataset is small (Tibshirani and Efron, 1993).

On the other hand, there are well-known advanced sampling methods that have been developed to address the limitations of Simple Random Sampling (SRS), such as stratified sampling (Neyman, 1992), multi-stage sampling (Kuno, 1976), systematic sampling (Madow and Madow, 1944), and the quasi-monte carlo method (Hastings, 1970). Among these methods, Ranked Set Sampling (RSS) (McIntyre, 1952) is particularly appealing

when measuring the response variable is costly, time-consuming, or resource-intensive, while ranking units (via expert judgment, visual inspection, or a concomitant variable) is inexpensive. In such scenarios, RSS yields substantially more informative samples than SRS. RSS has been shown to outperform SRS in population mean estimation (Wolfe, 2004), kernel density estimation (Chen, 1999), distribution function estimation (Stokes and Sager, 1988), and hypothesis testing (Öztürk, 2002). Moreover, RSS data lead to higher Fisher information (Hatefi and Jafari Jozani, 2013) than their SRS counterparts. In this thesis, we build on these advantages by developing RSS-based bagging frameworks that integrate rank information directly into the resampling and model training stages.

In many machine learning applications, the potential value of domain expertise is often underutilized. Integrating such expert knowledge through judgment-based ranking, or consideration of concomitant variables, can yield more representative samples and enhance model performance, particularly when training data are scarce (Zhang, 2023). In practice, such ranking information can arise from many sources, like previously developed predictive models in the literature, lagged observations in longitudinal or time-dependent data, or expert judgment in fields such as medicine, cancer diagnosis, and other domain-specific applications. RSS naturally incorporates this kind of expert knowledge into sampling through a ranking process. To obtain an RSS sample, units are first randomly drawn from the population and partitioned into sets of equal size. The units within each set are then ranked using the available auxiliary information or expert judgment, after which one unit from each rank is selected for full measurement. This process is repeated over multiple cycles to produce a balanced RSS sample. A detailed description of the RSS sampling procedure is provided in Chapter 2.

Although RSS has been used to improve the quality of training data, comparatively few studies have explicitly examined how to incorporate it into model training. Faraji et al. (2022) demonstrates one such case, showing how multilevel regression models can be adapted to the structure of RSS data. A primary practical challenge in RSS involves

efficiently integrating rank information into regression models, particularly when the auxiliary rank information is imperfect. This gap motivates the proposed framework in this thesis, which asks how ranking information can be used when RSS training data are unavailable and we only have access to SRS training data.

Regression analysis under an RSS framework was first introduced by [Muttalak \(1995\)](#), who investigated simple linear regression using ranked set samples. He showed that the resulting parameter estimates are unbiased under both perfect and imperfect ranking, and that the estimator variance is smaller than that of SRS-based estimators when ranking quality is reasonably good. [Barreto and Barnett \(1999\)](#) further derived best linear unbiased estimators (BLUE) under simple linear regression and showed that RSS yields substantial efficiency gains under normality. [Muttalak \(1996\)](#) extended these results to multiple linear regression, again showing that RSS improves the efficiency of regression estimators relative to SRS. These findings demonstrate that RSS enhances regression estimation by leveraging additional information embedded in the ranking process, thereby extracting more information from each sampled unit

In our proposed RSS bagging framework, we embed rank information directly into resampling and model training using RSS training data to improve predictive accuracy and stability, particularly in small-sample settings where traditional SRS-based bagging often performs poorly. By incorporating a ranking structure at each stage, this method is expected to enhance both the efficiency and robustness of the resulting regression models. The second framework addresses the more common practical limitation in which RSS training data is unavailable. Here, rather than requiring RSS training data from the outset, we ask whether partially available ranking knowledge can be integrated directly into SRS training data to create a more informative learning process. We focus on settings where some form of domain expert knowledge exists. Our approach leverages this by introducing one or more rankers that transform SRS data into a more informative RSS structure via judgment post-stratification ([MacEachern et al., 2004](#)),

thereby enriching the SRS sample with ranking information to improve representativeness and learning efficiency. Together, these frameworks form the core thesis contribution: they use ranking information to strengthen the learning process whenever it is available, in full or in part. Unlike classical RSS, this framework is considerably more practical for real-world applications, as the original sample remains SRS.

1.2 Ensemble Learning

Ensemble learning methods train multiple base learners on a training dataset and combine their predictions to produce a final model. These base learners are constructed using a base learning algorithm, such as a regression model, a decision tree, a neural network, or another machine learning technique. When all base learners are generated using the same learning algorithm, the resulting ensemble is known as a homogeneous ensemble. Conversely, when different learning algorithms are employed to create the base learners, the ensemble is referred to as a heterogeneous ensemble. Ensemble methods typically exhibit stronger generalization capabilities than single-model approaches because they combine the strengths of multiple learners. A major motivation behind ensemble learning is the ability to transform a collection of weak learners, models that perform only slightly better than random guessing, into a highly accurate predictive model.

The development of modern ensemble learning rests on three major paradigms: combining classifiers, ensembles of weak learners, and mixtures of experts (Zhou, 2025). The practice of combining classifiers, also known as “classifier ensembles,” has evolved into a significant discipline (Kuncheva, 2014). This paradigm originated primarily within the pattern recognition community, where researchers integrated the outputs of multiple classifiers to enhance predictive accuracy and robustness (Zhou, 2025). At a broad level, classifier ensembles are located at the intersection of engineering, computing, mathematics, and statistics. At a more detailed level, their development is driven

by advances in pattern recognition, machine learning, and data science. This line of research led to the development of influential ensemble methods, including Bagging and AdaBoost, and offered theoretical insights into the relationship between weak and strong learnability (Breiman, 1996; Freund and Schapire, 1997). The mixtures-of-experts paradigm was developed largely within the neural network community and is based on a divide-and-conquer strategy, in which multiple expert models are trained jointly and a gating mechanism determines how their outputs should be combined to solve a problem effectively (Jacobs et al., 1991; Jordan and Jacobs, 1994).

Hansen and Salamon (1990) demonstrated that combining predictions from multiple classifiers often yields greater accuracy than relying on the best single classifier. Schapire (1990) established that weak learners can be boosted to form strong learners. Because strong learners are highly desirable but challenging to construct, while weak learners are readily available, this finding has opened a promising avenue for generating strong learners through ensemble techniques. Depending on how base learners are generated, there are two approaches to ensemble methods: sequential ensembles, in which learners are created sequentially, and parallel ensembles, in which learners are created in parallel. The primary motivation for sequential methods is to leverage dependencies among base learners, as the overall performance can be boosted by reducing residuals. In the parallel setting, those methods exploit independence among the base learners, since the error can be drastically minimized by aggregating them. In this thesis, we primarily focus on bagging, a parallel ensemble method. As the name suggests, the two main components are bootstrap and aggregating.

1.3 Bootstrapping

Bootstrapping belongs to a broader class of resampling methods that generate repeated samples from the observed data to assess the variability of statistical estimators. Resam-

pling techniques have a long history in statistics. The jackknife method was introduced by [Quenouille \(1949\)](#), while permutation methods originated from the work of Fisher and Pitman in the 1930s ([Fisher, 1949](#); [Pitman, 1937](#)).

The bootstrap was formally introduced by [Efron \(1979\)](#), who unified these earlier ideas into a general and practical framework for statistical inference. Efron’s nonparametric bootstrap generates repeated samples of the same size as the original dataset by sampling with replacement from the observed data, thereby allowing one to approximate the sampling distribution of a statistic without relying on strong distributional assumptions. In contrast to classical inferential methods, which often depend on asymptotic theory or analytical variance formulas such as the delta method, the bootstrap approximates the sampling distribution directly from the data through resampling. For each bootstrap sample, the statistic or predictive model of interest is recalculated, producing an empirical distribution that can be used to estimate standard errors, bias, confidence intervals, and prediction error ([Tibshirani and Efron, 1993](#)). As a result, bootstrapping has become a standard tool in regression analysis, classification, machine learning, survival analysis, and other predictive modeling applications ([Davison and Hinkley, 1997](#)).

For predictive modeling with samples consisting of covariates $\mathbf{x}$ and a response variable y , several bootstrap methods have been developed, each based on different assumptions about the underlying data-generating process. The choice of bootstrap procedure depends largely on the statistical model and the objectives of the analysis. Since this research mainly focuses on predictive modeling, we consider the two commonly used approaches discussed in [Freedman \(1981\)](#), paired and residual bootstrapping.

1.3.1 Paired Bootstrapping

Consider a predictive modeling problem based on a training sample $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$.

The paired bootstrap proceeds by drawing bootstrap samples

$$(\mathbf{x}_i^{(b)}, y_i^{(b)}) \stackrel{\text{i.i.d.}}{\sim} \mathcal{DU}\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}, \quad i = 1, \dots, n,$$

where $\mathcal{DU}\{\cdot\}$ denotes the discrete uniform distribution over the original sample. Since each bootstrap pair is selected with equal probability (i.e., with replacement) from the original sample, it is easy to see that:

$$\mathbb{P}\left((\mathbf{x}_i^{(b)}, y_i^{(b)}) = (\mathbf{x}_j, y_j)\right) = \frac{1}{n}, \quad j = 1, \dots, n.$$

Then we refit the model on the bootstrap sample $\{(\mathbf{x}_i^{(b)}, y_i^{(b)})\}_{i=1}^n$ to obtain $\hat{f}^{(b)}$, and compute the bootstrap fitted values $\hat{y}_i^{(b)} = \hat{f}^{(b)}(\mathbf{x}_i)$, $i = 1, \dots, n$.

1.3.2 Residual Bootstrapping

The residual bootstrap is based on the fixed-design assumption of the regression model by holding the covariates $\mathbf{x}_i$ fixed and resampling only from the residuals. We first fit the model on the original sample $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ to obtain fitted values

$$\hat{y}_i = \hat{f}(\mathbf{x}_i), \quad i = 1, \dots, n,$$

and compute the corresponding residuals

$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n.$$

For each bootstrap replicate $b = 1, \dots, B$, we then proceed as follows:

- (i) Draw bootstrap residuals $e_i^{(b)} \stackrel{\text{i.i.d.}}{\sim} \mathcal{DU}\{e_1, \dots, e_n\}$, $i = 1, \dots, n$, and construct bootstrap responses $y_i^{(b)} = \hat{y}_i + e_i^{(b)}$, $i = 1, \dots, n$.

(ii) Refit the model on the bootstrap sample $\{(\mathbf{x}_i, y_i^{(b)})\}_{i=1}^n$ to obtain $\hat{f}^{(b)}$, and compute

$$\hat{y}_i^{(b)} = \hat{f}^{(b)}(\mathbf{x}_i).$$

1.4 Bagging

The term "Bagging" is derived from the abbreviation for Bootstrap AGGREGating. Bagging utilizes bootstrapping to generate diverse base learners. Specifically, it applies bootstrap sampling (Tibshirani and Efron, 1993) to obtain data subsets for training the base learners. Figure 1.1 provides a visual representation of the bagging procedure. Starting from a population $\mathcal{D} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^N$, a training dataset $\mathcal{D}_S = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ is obtained using SRS. If a single predictive model is fitted directly to the training sample $\mathcal{D}_S$, the resulting estimator may exhibit high variance and poor generalization, particularly when the underlying learning algorithm is sensitive to changes in the training data.

To address this issue, bagging generates B bootstrap samples, $\mathcal{D}^{*(1)}, \mathcal{D}^{*(2)}, \dots, \mathcal{D}^{*(B)}$, by sampling with replacement from the original training dataset $\mathcal{D}_S$. Each bootstrap sample has the same size as the original training sample, but some observations may appear multiple times while others may be omitted. A separate model is then trained on each bootstrap sample, producing fitted functions $\hat{f}_1^*(\mathbf{X}), \hat{f}_2^*(\mathbf{X}), \dots, \hat{f}_B^*(\mathbf{X})$. For regression problems, the final bagged predictor is obtained by averaging these fitted functions

$$\hat{f}_{\text{bag}}(\mathbf{X}) = \frac{1}{B} \sum_{b=1}^B \hat{f}_b^*(\mathbf{X}). \quad (1.1)$$

Hence, for a new observation $\mathbf{X}_{i,\text{new}}$, b -th model produces the prediction $\hat{Y}_i^{*b} = \hat{f}_b^*(\mathbf{X}_{i,\text{new}})$ and, the final bagged prediction is $\hat{Y}_{i,\text{bag}} = \frac{1}{B} \sum_{b=1}^B \hat{Y}_i^{*b}$.

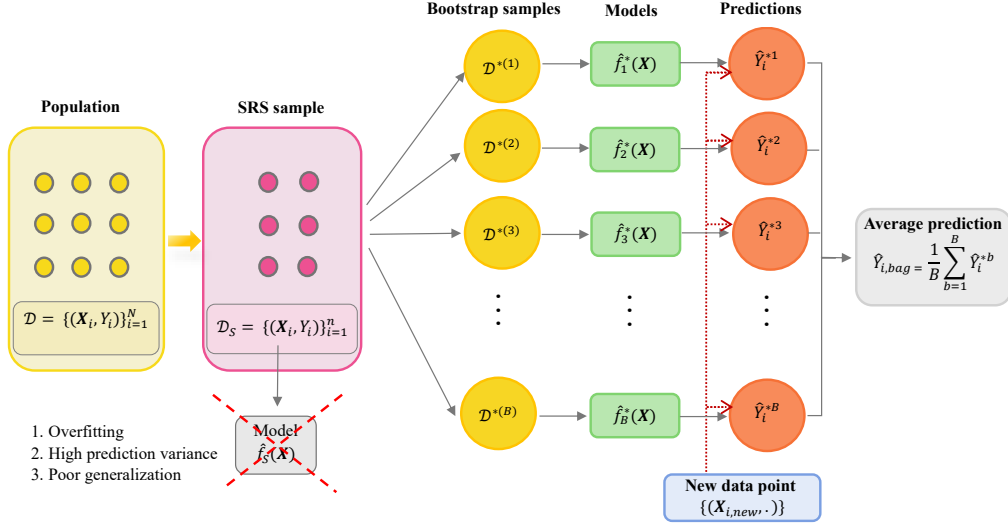


Figure 1.1: Illustration of the standard bagging procedure. Multiple bootstrap samples are generated from the original SRS training dataset; separate models are fit to each bootstrap sample, and the resulting predictions are averaged to obtain a more stable and accurate predictor.

Theorem 1.4.1 (Variance of bagged predictors). Let $\hat{f}_1^*(\mathbf{x}), \hat{f}_2^*(\mathbf{x}), \dots, \hat{f}_B^*(\mathbf{x})$ denote predictions obtained from B bootstrap trained base models at a fixed input $\mathbf{x}$. Suppose that each predictor has variance σ^2 and that the average pairwise correlation between any two distinct predictors is ρ . Then the variance of the bagged predictor (1.1) is given by

$$\text{Var}(\hat{f}_{bag}(\mathbf{x})) = \frac{\sigma^2}{B} [1 + (B-1)\rho].$$

Proof. Assuming that $\text{Var}(\hat{f}_b^*(\mathbf{x})) = \sigma^2$ for all b and that $\text{Cov}(\hat{f}_b^*(\mathbf{x}), \hat{f}_j^*(\mathbf{x})) = \rho\sigma^2$ for $b \neq j$, it is easy to see that

$$\begin{aligned} \text{Var}\left(\sum_{b=1}^B \hat{f}_b^*(\mathbf{x})\right) &= \sum_{b=1}^B \text{Var}(\hat{f}_b^*(\mathbf{x})) + \sum_{b \neq j} \text{Cov}(\hat{f}_b^*(\mathbf{x}), \hat{f}_j^*(\mathbf{x})) \\ &= B\sigma^2 + B(B-1)\rho\sigma^2. \end{aligned}$$

Accordingly

$$\mathbb{V}ar\left(\hat{f}_{\text{bag}}(\mathbf{x})\right) = \frac{1}{B^2} [B\sigma^2 + B(B-1)\rho\sigma^2] = \frac{\sigma^2}{B} [1 + (B-1)\rho].$$

□

In the case where bootstrap predictors are uncorrelated (i.e., $\rho = 0$), the variance of the bagged estimator would be $\mathbb{V}ar(\hat{f}_{\text{bag}}(\mathbf{x})) = \frac{\sigma^2}{B}$. However, bootstrap predictors are typically positively correlated ($\rho > 0$), in which case $\mathbb{V}ar(\hat{f}_{\text{bag}}(\mathbf{x}))$ exceeds $\frac{\sigma^2}{B}$. The magnitude of the variance reduction depends on the number of bootstrap replicates B , the variance of the base learner, and, critically, the correlation among the bootstrap predictors. As B increases, the variance reduction becomes more significant, and the smaller the correlation among the bootstrap predictors, the greater the gain achieved through bagging. In fact, as $B \rightarrow \infty$ one can see that $\mathbb{V}ar(\hat{f}_{\text{bag}}(\mathbf{x})) \rightarrow \rho\sigma^2 < \sigma^2$. Bagging works well for regression problems under squared error loss, as averaging reduces variance while leaving bias unchanged.

Theorem 1.4.2 (Expected prediction error of bagging under mean squared error (MSE)). Let $\hat{f}_1^*(\mathbf{x}_0), \hat{f}_2^*(\mathbf{x}_0), \dots, \hat{f}_B^*(\mathbf{x}_0)$ denote B predictions using base learners trained on bootstrap resamples of a training dataset $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, evaluated at a fixed input $\mathbf{x}_0$, where the observations are generated i.i.d. from a joint distribution $F_{\mathbf{X},Y}$. Under squared error loss, the expected prediction error of $\hat{f}_{\text{bag}}$ is always less than that of each $\hat{f}_b$ for any $b = 1, \dots, B$, that is

$$\mathbb{E}_{\mathbf{X},Y} \left[(y_0 - \hat{f}_{\text{bag}}(\mathbf{x}_0))^2 \right] \leq \mathbb{E}_{\mathbf{X},Y} \left[(y_0 - \hat{f}_b^*(\mathbf{x}_0))^2 \right].$$

Also,

$$\mathbb{E}_{\mathbf{X},Y} \left[\frac{1}{B} \sum_{b=1}^B (\hat{f}_b^*(\mathbf{x}_0) - \hat{f}_{\text{bag}}(\mathbf{x}_0))^2 \right] = \sigma^2 \frac{B-1}{B} (1 - \rho),$$

where σ^2 is the variance of an individual bootstrap predictor and ρ is the pairwise correlation between distinct predictors.

Proof. For a new observation $(\mathbf{x}_0, y_0)$, the prediction error under squared error loss is

$$\Delta(\hat{f}_{\text{bag}}(\mathbf{x}_0)) = \frac{1}{B} \sum_{b=1}^B (\hat{f}_b^*(\mathbf{x}_0) - \hat{f}_{\text{bag}}(\mathbf{x}_0))^2 = \frac{1}{B} \sum_{b=1}^B (y_0 - \hat{f}_b^*(\mathbf{x}_0))^2 - (y_0 - \hat{f}_{\text{bag}}(\mathbf{x}_0))^2.$$

Taking expectations,

$$\mathbb{E}_{\mathbf{X}, Y}[\Delta(\hat{f}_{\text{bag}}(\mathbf{x}_0))] = \frac{1}{B} \sum_{b=1}^B \mathbb{E}_{\mathbf{X}, Y}[(y_0 - \hat{f}_b^*(\mathbf{x}_0))^2] - \mathbb{E}_{\mathbf{X}, Y}[(y_0 - \hat{f}_{\text{bag}}(\mathbf{x}_0))^2].$$

Since all bootstrap models are identically distributed:

$$\mathbb{E}_{\mathbf{X}, Y}[(y_0 - \hat{f}_b^*(\mathbf{x}_0))^2] = \mathbb{E}_{\mathbf{X}, Y}[(y_0 - \hat{f}_{b'}^*(\mathbf{x}_0))^2] \quad \forall b, b' \in \{1, \dots, B\},$$

and $\Delta(\hat{f}_{\text{bag}}(\mathbf{x}_0)) \geq 0$, we conclude:

$$\mathbb{E}_{\mathbf{X}, Y}[(y_0 - \hat{f}_{\text{bag}}(\mathbf{x}_0))^2] \leq \mathbb{E}_{\mathbf{X}, Y}[(y_0 - \hat{f}_b^*(\mathbf{x}_0))^2]. \quad (1.2)$$

Without loss of generality, let $\mathbb{E}[\hat{f}_b^*(\mathbf{x}_0)] = 0$. Since

$$\text{Var}(\hat{f}_b^*(\mathbf{x}_0)) = \sigma^2, \text{ and } \text{Cov}(\hat{f}_b^*(\mathbf{x}_0), \hat{f}_j^*(\mathbf{x}_0)) = \begin{cases} \sigma^2, & j = b, \\ \rho\sigma^2, & j \neq b, \end{cases}$$

the expected prediction error decomposes as

$$\mathbb{E}_{\mathbf{X}, Y}[\Delta(\hat{f}_{\text{bag}}(\mathbf{x}_0))] = \mathbb{E}_{\mathbf{X}, Y}[\hat{f}_b^*(\mathbf{x}_0)^2] + \mathbb{E}_{\mathbf{X}, Y}[\hat{f}_{\text{bag}}(\mathbf{x}_0)^2] - 2\mathbb{E}_{\mathbf{X}, Y}[\hat{f}_b^*(\mathbf{x}_0)\hat{f}_{\text{bag}}(\mathbf{x}_0)]. \quad (1.3)$$

Note that

$$\text{Var}(\hat{f}_{\text{bag}}(\mathbf{x}_0)) = \mathbb{E}_{\mathbf{X}, Y} \left[\left(\hat{f}_{\text{bag}}(\mathbf{x}_0) - 0 \right)^2 \right] = \mathbb{E}_{\mathbf{X}, Y}[\hat{f}_{\text{bag}}(\mathbf{x}_0)^2] = \frac{\sigma^2}{B} (1 + (B-1)\rho).$$

Also,

$$\mathbb{E}_{\mathbf{X},Y}[\Delta(\hat{f}_{\text{bag}}(\mathbf{x}_0))] = \sigma^2 - \text{Var}(\hat{f}_{\text{bag}}(\mathbf{x}_0)) = \sigma^2 \frac{B-1}{B} (1-\rho).$$

Using these, one can easily see that

$$\begin{aligned} \mathbb{E}_{\mathbf{X},Y}[\hat{f}^{*b}(\mathbf{x}_0)\hat{f}_{\text{bag}}(\mathbf{x}_0)] &= \frac{1}{B} \sum_{b=1}^B \mathbb{E}_{\mathbf{X},Y} [\hat{f}^{*b}(\mathbf{x}_0)\hat{f}^{*j}(\mathbf{x}_0)] \\ &= \frac{\sigma^2 [1 + (B-1)\rho]}{B}. \end{aligned}$$

Substituting these into (1.3) gives the desired result:

$$\mathbb{E}_{\mathbf{X},Y}[\Delta(\hat{f}_{\text{bag}}(\mathbf{x}_0))] = \sigma^2 + \frac{1 + (B-1)\rho}{B} \sigma^2 - 2 \cdot \frac{1 + (B-1)\rho}{B} \sigma^2 = \sigma^2 \frac{B-1}{B} (1-\rho).$$

□

Under squared-error loss, bagging never increases the expected prediction error; for unstable learners like regression trees, it typically reduces the MSE. The improvement comes from reducing variance while preserving bias (as shown in the previous proof). It is clear that for large numbers of bootstrap samples, $B \rightarrow \infty$, the factor $(B-1)/B \rightarrow 1$ maximizes variance reduction. Unstable base learners, characterized by high variance σ^2 , benefit the most from bagging under MSE. Furthermore, diverse bootstrap samples that induce low or even negative correlation ρ among base learners yield greater improvements.

The expected mean squared error can be decomposed into irreducible error, the bias in the fitted value, and the variance of the fitted value as

$$\begin{aligned} \mathbb{E}_{\mathbf{X},Y} \left[(Y_0 - \hat{f}(\mathbf{x}_0))^2 \mid \mathbf{X} = \mathbf{x}_0 \right] &= \sigma_\epsilon^2 + \left(\mathbb{E}_{\mathbf{X},Y}[\hat{f}(\mathbf{x}_0)] - f(\mathbf{x}_0) \right)^2 \\ &\quad + \mathbb{E}_{\mathbf{X},Y} \left[\left(\hat{f}(\mathbf{x}_0) - \mathbb{E}_{\mathbf{X},Y}[\hat{f}(\mathbf{x}_0)] \right)^2 \right]. \end{aligned}$$

The irreducible error (σ_ε^2) denotes the variance of Y_0 around its true conditional mean at $\mathbf{x}_0$, which cannot be reduced, as it is an inherent part of the data-generating system. In more complex models, a higher number of degrees of freedom typically results in lower squared bias $\left(\mathbb{E}[\hat{f}(\mathbf{x}_0)] - f(\mathbf{x}_0)\right)^2$ and likely results in greater variance $\mathbb{E}\left[\left(\hat{f}(\mathbf{x}_0) - \mathbb{E}[\hat{f}(\mathbf{x}_0)]\right)^2\right]$ (Berk et al., 2008). Bagging effectively addresses the trade-off between bias and variance under squared loss. By averaging across multiple bootstrap samples, bagging yields a more stable set of fitted values than those derived from a single sample, for any given level of bias. This approach substantially reduces variance without increasing bias, thereby yielding more reliable and consistent predictions.

Remark 1.4.1. Equation (1.2) implies that under squared-error loss, the bagged predictor cannot, on average, perform worse than an individual bootstrap predictor. However, the result in (1.2) does not generally extend to other loss functions, such as the 0-1 loss used in classification problems. Unlike squared-error loss, the 0-1 loss does not admit an additive bias-variance decomposition. Therefore, in classification settings, bagging does not necessarily guarantee improved performance. For classification, bagging a good classifier can improve it, but bagging a bad classifier can worsen it.

In the next part of this chapter, we focus on the unbiasedness of the bagged SRS mean estimator and derive the corresponding expression for its variance.

1.5 Bagging with SRS (SSB)

We first state the problem in a general setting for estimating means and then show how these result extend to regression settings. Suppose $\mathcal{D}_S = \{Y_1, \dots, Y_n\}$, is an SRS training data of size n from F with the empirical distribution function (EDF) $F_n(y) = \frac{1}{n} \sum_{j=1}^n \mathbb{I}\{Y_j \leq y\}$. Our objective is to estimate $\mu_h = \mathbb{E}_F[h(Y)]$, where $h(\cdot)$ is a

suitable function of the response variable. This formulation satisfies many commonly used population parameters.

- (i) If $h(y) = y$, then $\mu_h = \mathbb{E}_F(Y)$ is the population mean.
- (ii) If $h(y) = y^2$, then $\mu_h = \mathbb{E}_F(Y^2)$ is the second population moment.
- (iii) If $h(y) = \mathbb{I}\{y \leq t\}$, where $t \in \mathbb{R}$ is fixed, then $\mu_h = \mathbb{E}_F[\mathbb{I}\{Y \leq t\}] = F(t)$ is the CDF evaluated at t . The corresponding SRS estimator is the EDF,

$$\hat{\mu}_{h,S} = F_n(t) = \frac{1}{n} \sum_{j=1}^n \mathbb{I}\{Y_j \leq t\}.$$

- (iv) If $h(y) = (y - \mu)^2$, where $\mu = \mathbb{E}_F(Y)$, then $\mu_h = \mathbb{E}_F[(Y - \mu)^2] = \text{Var}(Y)$ is the population variance.

The SRS estimator of μ_h is the sample mean of $h(Y)$,

$$\hat{\mu}_{h,S} = \frac{1}{n} \sum_{j=1}^n h(Y_j).$$

Now we generate B bootstrap samples $\mathcal{D}^{*(1)}, \dots, \mathcal{D}^{*(B)}$, each of size n , by sampling with replacement from $\mathcal{D}_S$. Denote the i -th observation in the b -th bootstrap sample by Y_i^{*b} . For the b -th bootstrap, define the bootstrap sample mean as

$$\hat{\mu}_{h,S}^{*b} = \frac{1}{n} \sum_{i=1}^n h(Y_i^{*b}). \tag{1.4}$$

Accordingly, the bagged estimator of μ_h is

$$\hat{\mu}_{h,\text{SSB}} = \frac{1}{B} \sum_{b=1}^B \hat{\mu}_{h,S}^{*b} = \frac{1}{Bn} \sum_{b=1}^B \sum_{i=1}^n h(Y_i^{*b}). \tag{1.5}$$

Conditional on $\mathcal{D}_S$, each bootstrap draw takes the observed value Y_j with probability $1/n$, so for any fixed b and i ,

$$\mathbb{P}(Y_i^{*b} = Y_j \mid \mathcal{D}_S) = \frac{1}{n}, \quad j = 1, \dots, n.$$

Proposition 1.5.1 (Unbiasedness of the SSB mean estimator). Conditioning on $\mathcal{D}_S$,

$$\mathbb{E}_{F_n}[h(Y_i^{*b})] = \hat{\mu}_{h,S},$$

where $\mathbb{E}_{F_n}$ is conditioned on $\mathcal{D}_S$. Also $\mathbb{E}_F[\hat{\mu}_{h,S}] = \mu_h$.

Proof. By considering the definition of $\hat{\mu}_{h,S}^{*b}$ in (1.4),

$$\mathbb{E}_{F_n}[\hat{\mu}_{h,S}^{*b}] = \mathbb{E}_{F_n} \left[\frac{1}{n} \sum_{i=1}^n h(Y_i^{*b}) \right] = \hat{\mu}_{h,S}.$$

Similarly, using the definition of $\hat{\mu}_{h,SSB}$ in (1.5), it is easy to see that,

$$\mathbb{E}_{F_n}[\hat{\mu}_{h,SSB}] = \hat{\mu}_{h,S}. \tag{1.6}$$

Finally, we have

$$\mathbb{E}_F[\mathbb{E}_{F_n}[\hat{\mu}_{h,SSB}]] = \mathbb{E}_F[\hat{\mu}_{h,S}] = \mu_h. \tag{1.7}$$

□

Proposition 1.5.1 establishes the unbiasedness of the SSB estimator, both conditionally on the observed $\mathcal{D}_S$ and unconditionally with respect to the underlying population distribution.

Theorem 1.5.1 (Variance of SSB mean estimator). Let $\hat{\mu}_{h,SSB}$ be the bagged estimator defined in (1.5). The conditional variance of $\hat{\mu}_{h,SSB}$ given $\mathcal{D}_S$ is

$$\text{Var}_{F_n}(\hat{\mu}_{h,SSB} \mid \mathcal{D}_S) = \frac{n-1}{Bn^2} s_{n,h}^2, \tag{1.8}$$

where $s_{n,h}^2 = \frac{1}{n-1} \sum_{j=1}^n (h(Y_j) - \hat{\mu}_{h,S})^2$.

Proof. Since the bootstrap samples are independent given $\mathcal{D}_S$, the variance of the bagged estimator is

$$\begin{aligned}\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) &= \frac{1}{B^2} \sum_{b=1}^B \mathbb{V}ar_{F_n}(\hat{\mu}_{h,S}^{*b} \mid \mathcal{D}_S) \\ &= \frac{1}{Bn^2} \sum_{i=1}^n \mathbb{V}ar_{F_n}(h(Y_i^{*b}) \mid \mathcal{D}_S).\end{aligned}$$

As $Y_i^{*b} \sim F_n$ given $\mathcal{D}_S$, one can easily show that

$$\begin{aligned}\mathbb{V}ar_{F_n}(h(Y_i^{*b}) \mid \mathcal{D}_S) &= \mathbb{E}_{F_n}[(h(Y_i^{*b}) - \mathbb{E}_{F_n}[h(Y_i^{*b})])^2 \mid \mathcal{D}_S] \\ &= \frac{1}{n} \sum_{j=1}^n (h(Y_j) - \hat{\mu}_{h,S})^2 = \frac{n-1}{n} s_{n,h}^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) &= \frac{1}{Bn^2} \sum_{i=1}^n \left(\frac{1}{n} \sum_{j=1}^n (h(Y_j) - \hat{\mu}_{h,S})^2 \right) \\ &= \frac{1}{B} \cdot \frac{n-1}{n^2} s_{n,h}^2.\end{aligned}$$

For the unconditional variance, applying the law of total variance and obtain

$$\begin{aligned}\mathbb{V}ar_F(\hat{\mu}_{h,\text{SSB}}) &= \mathbb{E}_F[\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S)] + \mathbb{V}ar_F(\mathbb{E}_{F_n}[\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S]) \\ &= \mathbb{E}_F\left[\frac{1}{B} \cdot \frac{n-1}{n^2} s_{n,h}^2\right] + \mathbb{V}ar_F(\hat{\mu}_{h,S}) \\ &= \frac{1}{B} \cdot \frac{n-1}{n^2} \sigma_h^2 + \frac{\sigma_h^2}{n} \\ &= \sigma_h^2 \left(\frac{Bn + n - 1}{Bn^2} \right),\end{aligned}$$

which is (1.8), where $\mathbb{E}_F[s_{n,h}^2] = \sigma_h^2 = \mathbb{V}ar_F(h(Y))$ and $\mathbb{V}ar_F(\hat{\mu}_{h,S}) = \frac{\sigma_h^2}{n}$. $\square$

This concludes the analysis of the conditional and unconditional variance of the SSB mean estimator.

1.6 Regression Estimation with SSB

In Section 1.5, we developed the bagging methodology for estimating a broad class of mean-type functionals and established its unbiasedness and variance properties. Now the objective is to estimate a conditional mean function rather than a population mean.

Let $(\mathbf{X}, Y)$ be a pair of random variables with joint distribution F , where $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathbb{R}$. We are interested in the regression function

$$m(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}),$$

at a fixed prediction point $\mathbf{x} \in \mathcal{X}$. In this section, the effect of the bagging is studied under SRS training data $\mathcal{D}_S = \{(\mathbf{X}_i, Y_i) : i = 1, \dots, n\}$. The extension from mean estimation to regression is based on the observation that many regression procedures admit a *linear smoother* representation.

1.6.1 Linear Functional Representation of Regression Estimators under SRS

We view a regression procedure as a map that, given training data, returns a fitted regression function. Under SRS, for example, we start with $\mathcal{D}_S$, and the procedure produces a random function

$$\hat{m}_S(\cdot) : \mathcal{X} \rightarrow \mathbb{R},$$

which, when evaluated at a new point $\mathbf{x} \in \mathcal{X}$, yields the prediction $\hat{m}_S(\mathbf{x})$.

In many important cases, the fitted regression function is linear in the responses. Specifically, for each fixed prediction point $\mathbf{x}$, $\hat{m}_S(\mathbf{x})$ can be expressed as a finite linear combination of the observed responses with data-dependent weights. This motivates the following definition.

Definition 1.6.1 (Linear smoother representation). Fix $\mathbf{x} \in \mathcal{X}$. Under SRS training data $\mathcal{D}_S$, we say that the regression estimator is a *linear smoother* if,

$$\hat{m}_S(\mathbf{x}) = \sum_{i=1}^n W_{ni}(\mathbf{x}) Y_i, \quad (1.9)$$

for some (random) weight functions

$$W_{ni}(\cdot) : \mathcal{X} \rightarrow \mathbb{R}, \quad i = 1, \dots, n,$$

which are measurable functions of $\mathbf{x}$ and depend on the training data $\mathcal{D}_S$. Equivalently, defining pseudo-responses

$$Z_i(\mathbf{x}) = n W_{ni}(\mathbf{x}) Y_i, \quad (1.10)$$

we can write

$$\hat{m}_S(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n Z_i(\mathbf{x}). \quad (1.11)$$

The above expression is analogous to the SRS estimator of $\hat{\mu}_{h,S} = \frac{1}{n} \sum_{j=1}^n h(Y_j)$, where the pseudo-response $Z_i(\mathbf{x})$ serves as $h(Y_j)$.

In this setting, the family of functions $\mathbf{x} \rightarrow \{Z_i(\mathbf{x})\}_{i=1}^n$ encode the fitted regression function through simple averaging representations. This formulation emphasizes that the regression procedure first constructs a fitted function $\mathbf{x} \rightarrow \hat{m}(\mathbf{x})$, and only then is a prediction obtained by evaluating this function at a new point $\mathbf{x}$.

In bagging, we generate B bootstrap samples, fit B regression functions $\hat{m}_1(\cdot), \dots, \hat{m}_B(\cdot)$, and define the bagged estimator

$$\hat{m}_{bag}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{m}_b(\mathbf{x}).$$

If each $\hat{m}_b(\mathbf{x})$ admits a representation of the form (1.9), then $\hat{m}_{bag}(\mathbf{x})$ also admits such a representation, with weights averaged over b . In all cases, the plug-in prediction at $\mathbf{x}$ is obtained by evaluating the fitted function at $\mathbf{x}$.

Example 1.6.1. Consider SRS training data $\mathcal{D}_S$ and the Nadaraya–Watson estimator at $\mathbf{x}$,

$$\hat{m}_{\text{NW}}(\mathbf{x}) = \frac{\sum_{i=1}^n K_h(\mathbf{X}_i - \mathbf{x}) Y_i}{\sum_{i=1}^n K_h(\mathbf{X}_i - \mathbf{x})},$$

where $K_h(\cdot) = K(\cdot/h)/h^p$ is a kernel with bandwidth $h > 0$ on $\mathbb{R}^p$. Define the normalized kernel weights

$$W_{ni}(\mathbf{x}) = \frac{K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^n K_h(\mathbf{X}_i - \mathbf{x})}, \quad i = 1, \dots, n.$$

Then

$$\hat{m}_{\text{NW}}(\mathbf{x}) = \sum_{i=1}^n W_{ni}(\mathbf{x}) Y_i = \frac{1}{n} \sum_{i=1}^n Z_i(\mathbf{x}), \text{ with } Z_i(\mathbf{x}) := n W_{ni}(\mathbf{x}) Y_i.$$

Thus, equations (1.9) – (1.11) hold with these weights. Here, $W_{ni}(\cdot)$ is a function of $\mathbf{x}$ constructed from the observed covariates $\mathbf{X}_1, \dots, \mathbf{X}_n$.

Example 1.6.2. Suppose we fit a linear model

$$Y_i = \mathbf{X}_i^\top \beta + \varepsilon_i, \quad i = 1, \dots, n,$$

with $\mathbf{X}_i \in \mathbb{R}^p$, by ordinary least squares (OLS). Let $\mathbb{X} \in \mathbb{R}^{n \times p}$ be the design matrix whose i -th row is $\mathbf{X}_i^\top$, and let $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$. The OLS estimator is

$$\hat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{Y},$$

and the fitted regression function is

$$\hat{m}_{\text{OLS}}(\mathbf{x}) = \mathbf{x}^\top \hat{\beta} = \mathbf{x}^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{Y} = \sum_{i=1}^n W_{ni}^{\text{OLS}}(\mathbf{x}) Y_i,$$

where

$$W_{ni}^{\text{OLS}}(\mathbf{x}) = \left(\mathbf{x}^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X} \right)_i, \quad i = 1, \dots, n.$$

Defining $Z_i(\mathbf{x}) := n W_{ni}^{\text{OLS}}(\mathbf{x}) Y_i$, gives

$$\hat{m}_{\text{OLS}}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n Z_i(\mathbf{x}),$$

so OLS fits into the linear smoother framework.

Example 1.6.3. Consider a regression tree fitted to SRS training data $\mathcal{D}_S$ by some recursive partitioning algorithm (e.g., CART). Let the terminal nodes (leaves) be $\mathcal{L}_1, \dots, \mathcal{L}_M \subset \mathcal{X}$, and let $L(\mathbf{x})$ denote the index of the leaf containing $\mathbf{x}$:

$$L(\mathbf{x}) \in \{1, \dots, M\}, \quad \mathbf{x} \in \mathbf{X}.$$

The usual regression tree prediction at $\mathbf{x}$ is

$$\hat{m}_{\text{tree}}(\mathbf{x}) = \frac{1}{N_{L(\mathbf{x})}} \sum_{i: \mathbf{X}_i \in \mathcal{L}_{L(\mathbf{x})}} Y_i, \text{ with } N_{L(\mathbf{x})} = \#\{i : \mathbf{X}_i \in \mathcal{L}_{L(\mathbf{x})}\}.$$

Define the weights

$$W_{ni}^{\text{tree}}(\mathbf{x}) = \begin{cases} \frac{1}{N_{L(\mathbf{x})}} & \text{if } \mathbf{X}_i \in \mathcal{L}_{L(\mathbf{x})}, \\ 0 & \text{otherwise,} \end{cases}$$

and the pseudo-responses $Z_i(\mathbf{x}) := n W_{ni}^{\text{tree}}(\mathbf{x}) Y_i$. Then

$$\hat{m}_{\text{tree}}(\mathbf{x}) = \sum_{i=1}^n W_{ni}^{\text{tree}}(\mathbf{x}) Y_i = \frac{1}{n} \sum_{i=1}^n Z_i(\mathbf{x}),$$

so (1.9) – (1.11) again hold. Here, the functions $\mathbf{x} \mapsto W_{ni}^{\text{tree}}(\mathbf{x})$ are determined by the fitted partition built from the full sample $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$.

Example 1.6.4. Consider a random forest constructed from B regression trees $\hat{m}_{\text{tree}}^{(1)}, \dots, \hat{m}_{\text{tree}}^{(B)}$, each fitted on a bootstrap (or subsampled) version of $\mathcal{D}_S$ with random feature splits.

The forest prediction at $\mathbf{x}$ is

$$\hat{m}_{\text{RF}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{m}_{\text{tree}}^{(b)}(\mathbf{x}).$$

Each individual tree admits a representation

$$\hat{m}_{\text{tree}}^{(b)}(\mathbf{x}) = \sum_{i=1}^n W_{ni}^{(b)}(\mathbf{x}) Y_i,$$

with weights $W_{ni}^{(b)}(\mathbf{x})$ defined as in the previous example, but for tree b . Averaging over trees yields the overall forest weights

$$W_{ni}^{\text{RF}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B W_{ni}^{(b)}(\mathbf{x}),$$

so that

$$\hat{m}_{\text{RF}}(\mathbf{x}) = \sum_{i=1}^n W_{ni}^{\text{RF}}(\mathbf{x}) Y_i = \frac{1}{n} \sum_{i=1}^n Z_i(\mathbf{x}), \text{ with } Z_i(\mathbf{x}) := n W_{ni}^{\text{RF}}(\mathbf{x}) Y_i.$$

Thus, random forests also satisfy the linear smoother representation (1.9) – (1.11), with $\mathbf{x}$ entering only as an argument of the weight functions.

The examples above show that the representations (1.9) – (1.11) cover a broad class of regression estimators. In all such cases, our SSB bootstrap constructions and the associated variance expressions apply pointwise in $\mathbf{x}$ via the pseudo-responses $Z_i(\mathbf{x})$.

1.6.2 Linear Smoother SSB Regression Estimator

We now extend the SSB results in Section 1.5 to the regression setting by applying the corresponding resampling scheme. Throughout, we assume that the regression estimator admits the linear smoother representation introduced in Section 1.6.1.

Definition 1.6.2 (SSB Regression Estimator). Fix $\mathbf{x} \in \mathbf{X}$ and let $Z_i(\mathbf{x})$ be as in (1.11). For $b = 1, \dots, B$:

- (a) Draw bootstrap indices $I_1^{*b}, \dots, I_n^{*b}$ i.i.d. from $\{1, \dots, n\}$ with replacement, corresponding to an SRS bootstrap sample of size n from $\mathcal{D}_S$.
- (b) Define the bootstrap pseudo-responses $Z_i^{*b}(\mathbf{x}) := Z_{I_i^{*b}}(\mathbf{x})$, $i = 1, \dots, n$, and the bootstrap regression estimator at $\mathbf{x}$ by $\hat{m}_S^{*b}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n Z_i^{*b}(\mathbf{x})$.

The B -replicate bagged SRS bootstrap estimator is

$$\hat{m}_{\text{SSB}}^{(B)}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{m}_S^{*b}(\mathbf{x}). \quad (1.12)$$

In terms of the pseudo-responses $Z_i(\mathbf{x})$, this construction is algebraically identical to the bagged bootstrap estimator for mean-function estimation in (1.5), with $h(Y_i)$ replaced by $Z_i(\mathbf{x})$. This establishes a direct connection between the theory developed for SSB in Section 1.5 and the regression setting.

The expectation and variance results derived for the mean-functional estimators in Section 1.5 depend only on the averaging structure of the estimator and not on the specific form of the transformation $h(\cdot)$, therefore these results carry over immediately upon replacing $h(Y_i)$ by $Z_i(\mathbf{x})$ in the SSB setting. In particular, the unbiasedness properties, conditional and unconditional variance expressions, and variance-comparison results established for SSB remain valid for any regression estimator that admits a linear smoother representation.

1.7 Research Statement and Thesis Outline

Bagging is a widely used ensemble learning technique that enhances prediction accuracy and stability by aggregating multiple bootstrap-based estimators. Conventional bagging methods, however, primarily rely on SRS and do not utilize auxiliary ranking information that is often available in practical applications. RSS has demonstrated greater efficiency

than SRS in various statistical contexts, yet its integration into ensemble learning and bagging methodologies remains largely unexplored. This thesis addresses this gap by developing novel RSS-based bagging frameworks for both mean-type functionals and regression estimation. The theoretical properties of the proposed estimators, including unbiasedness and variance behavior, are established. The methodologies are further extended to weighted least squares and mixed-effects modeling frameworks, and the performance of the developed RSS-based bagged regression models is evaluated through extensive simulation studies and real-data applications.

The structure of this thesis is as follows. Chapter 2 introduces the fundamentals of RSS and develops the RSS Bagging (RSB) methodology for both mean-type functionals and regression estimation. Key theoretical results, including unbiasedness and variance properties, are established, and the framework is extended to Weighted Least Squares (WLS) and Linear Mixed Models (LMMs). Chapter 3 addresses practical limitations of RSB by proposing a novel rank-informed bagging framework constructed from SRS training data, along with its theoretical properties and comparisons to existing bagging approaches. Chapter 4 presents an extensive simulation study and real-data application to evaluate the performance of the proposed methods under various ranking structures, bootstrap schemes, and regression models. Chapter 5 summarizes the thesis's main contributions and discusses potential directions for future research.

Chapter 2

RSS Bagging

This chapter presents a comprehensive overview of ranked set sampling (RSS), its basic properties, and its applications in the statistical literature. We then develop the RSS bagging methodology, first for general mean-type functionals and then for estimating the regression function $m(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$ at a fixed point $\mathbf{x}$. Theoretical results for the proposed estimators include unbiasedness, expressions for the conditional and unconditional variances, and comparisons of variances with classical bagging procedures based on simple random sampling (SRS). We develop RSS-based predictive models using weighted least squares regression and multilevel modeling techniques.

2.1 Introduction

The effectiveness of bagging depends primarily on the quality of the training data and the representativeness of the training sample of the underlying population. Because bagging generates bootstrap samples directly from the observed training data, any deficiencies in the original sample are propagated through the resampling process. Consequently, the resulting ensemble is constrained by the information in the initial training sample.

This issue is especially noticeable in small-sample settings. When training data are limited, bootstrap samples contain many repeated observations and are often highly

correlated. As a result, diversity among bootstrap learners may be reduced. For complex predictive models, this lack of diversity can lead to overfitting, increased instability, and reduced predictive performance. Although bagging generally reduces variance, its ability to improve prediction accuracy is limited when the original training sample does not adequately capture the target population.

Ranked set sampling (RSS) offers a promising approach to mitigating these challenges. By incorporating ranking information into the sampling process, RSS produces samples that are often more representative of the population than those obtained through simple random sampling (SRS) alone. In this chapter, we study bagging predictive models using RSS training data.

2.2 Ranked Set Sampling

RSS was first introduced by McIntyre (1952) as a more efficient alternative to SRS for estimating mean pasture yields. Measuring the yield of pasture plots is labor-intensive and time-consuming, as it involves cutting, moving, and weighing hay from each sampled unit. However, McIntyre observed that an experienced observer could rank a set of plots by their approximate yield without physically harvesting them, which forms the conceptual foundation of RSS.

RSS emerged as a direct response to the limitations of SRS, particularly in small-sample settings where obtaining large samples is impractical due to time-consuming and/or costly restrictions. SRS often disregards the underlying population structure and, especially in small samples, may yield observations clustered in a single region of the distribution, leaving other regions underrepresented. RSS addresses this by introducing a two-stage data collection mechanism that improves the representativeness of the sample, reflecting the true distribution of the population. The basic idea of RSS is that a small collection of sampling units drawn from the population can be ranked,

visually or otherwise, with respect to the variable of interest without actually measuring them.

To obtain an RSS sample of size $n = mk$, we proceed as follows. First, a simple random sample of k units is drawn from the population, and these units are ranked with respect to the variable of interest, Y , without measuring it. This can be done using auxiliary information, expert knowledge, or any other inexpensive way. From this ranked set, only the unit assigned rank 1 is retained for final measurement, while the remaining $k - 1$ units are discarded. This selected observation, denoted by $Y_{[1]}$, corresponds to an observation from the smallest judgment-ranked unit in the set. A second independent simple random sample of k units is then drawn and ranked in the same manner. This time, only the unit assigned rank 2 is measured, and the remaining units are discarded. This now corresponds to the second-smallest judgment-ranked unit in the set, and the resulting observation is denoted by $Y_{[2]}$. The process continues until the k -th sample, at which point the unit assigned rank k is selected for measurement. This yields the observations $Y_{[r]}$ for $r = 1, \dots, k$, and forms a balanced RSS data of size k , containing exactly one observation from each rank. Repeating this cycle m times yields a balanced RSS of size $n = mk$, denoted by $Y_{[r]j}$, $r = 1, \dots, k$, $j = 1, \dots, n$. More generally, the number of observations selected from each rank need not be equal. Let n_r denote the number of observations measured from rank r , where $r = 1, \dots, k$. When the values $n_1, n_2, \dots, n_k$ are not all equal, the resulting design is called *unbalanced RSS*, denoted by $Y_{[r]j}$, $r = 1, \dots, k$, $j = 1, \dots, n_r$. The total sample size is then $n = \sum_{r=1}^k n_r$. Balanced RSS is obtained as the special case where $n_1 = n_2 = \dots = n_k = m$, whereas unbalanced RSS allows certain ranks to be sampled more heavily than others, providing greater flexibility in study design and estimation.

RSS has been widely used in the statistical literature to address a variety of estimation and inference problems. Early contributions focused primarily on the estimation of the population mean (McIntyre, 1952; Wolfe, 2004). Subsequently, the RSS methodology

was extended to the estimation of distribution functions (Stokes and Sager, 1988), the development of nonparametric testing procedures (Bohn and Wolfe, 1992, 1994), and the construction of statistical quality control charts (Muttalak and Al-Sabah, 2003; Jafari Jozani and Mirkamali, 2010, 2011; Asghari et al., 2018). In contrast, relatively little attention has been paid to regression analysis using RSS data. Among the early contributions in this area, Muttalak (1995) considered the estimation of regression parameters in a simple linear regression model using median and extreme RSS designs. Later, Barreto and Barnett (1999) developed best linear unbiased estimators for the parameters of simple linear regression under the RSS. Further development along this direction of research was carried out by Öztürk (2002) and Murff and Sager (2006). More flexible modeling approaches, including spline and polynomial regression models fitted to RSS data, were investigated by Algarni (2016) and Al Kadiri (2017).

Despite these advances, most existing approaches treat RSS primarily as a data-collection strategy and do not explicitly incorporate the ranking information of RSS data into the parameter estimation process. Faraji et al. (2022) developed a new method for regression modeling using RSS data through weighted least squares (WLS) with adjusted responses and multilevel modeling techniques, and provided efficient methods for incorporating rank information of RSS data into model fitting. More recently, RSS has been incorporated into ensemble classification methods using bagging (Wang et al., 2023). Although this represents an important step toward integrating RSS with modern predictive modeling techniques, RSS’s role remains largely confined to the sampling stage. The ranking information itself is not directly incorporated into the model estimation or prediction procedures. This limitation provides the key motivation for this chapter. Specifically, we extend the weighted least squares and multilevel modeling approaches proposed by Faraji et al. (2022) to ensemble learning via the bagging framework, allowing the ranking information inherent in RSS to contribute not only to sample selection but also to model fitting.

2.3 RSS for Estimating Population Means

Let $Y_1, Y_2, \dots, Y_k$ be i.i.d. continuous random variables drawn from a population with CDF $F(\cdot)$ and PDF $f(\cdot)$, respectively. Suppose ranking is perfect and $Y_{(r)}$ is the r -th order statistic of the sample. The PDF of $Y_{(r)}$ denoted by $f_{(r)}(y)$ is given by,

$$f_{(r)}(y) = \frac{k!}{(r-1)!(k-r)!} f(y) [F(y)]^{r-1} [1 - F(y)]^{k-r}, \quad (2.1)$$

with CDF

$$F_{(r)}(y) = \sum_{j=r}^k \binom{k}{j} [F(y)]^j [1 - F(y)]^{k-j}. \quad (2.2)$$

Theorem 2.3.1. The PDF $f(\cdot)$ and CDF $F(\cdot)$ of a continuous random variable can be represented as an equally weighted mixture of the PDF and CDF of order statistics:

$$\begin{aligned} \text{(i)} \quad f(y) &= \frac{1}{k} \sum_{r=1}^k f_{(r)}(y), \\ \text{(ii)} \quad F(y) &= \frac{1}{k} \sum_{r=1}^k F_{(r)}(y), \end{aligned} \quad (2.3)$$

where $f_{(r)}(y)$ and $F_{(r)}(y)$ are the PDF and CDF of the r -th order statistic of a sample size k from a population with PDF $f(y)$ and CDF $F(y)$, respectively.

Proof. Using (2.1), it is easy to see that:

$$\frac{1}{k} \sum_{r=1}^k f_{(r)}(y) = f(y) [F(y) + (1 - F(y))]^{k-1} = f(y).$$

(ii) is trivial by integration. □

The empirical distribution function (EDF) for an RSS sample of size $n = mk$ is,

$$\begin{aligned}
F_{n,R}(t) &= \frac{1}{mk} \sum_{j=1}^m \sum_{r=1}^k \mathbb{1}(Y_{(r)j} \leq t) \\
&= \frac{1}{k} \sum_{r=1}^k \left\{ \frac{1}{m} \sum_{j=1}^m \mathbb{1}(Y_{(r)j} \leq t) \right\} \\
&= \frac{1}{k} \sum_{r=1}^k F_{m,(r)}(t),
\end{aligned}$$

where $F_{m,(r)}(t) = \frac{1}{m} \sum_{j=1}^m \mathbb{1}(Y_{(r)j} \leq t)$ is the EDF for the r -th order statistic using m observations $\{Y_{(r)1}, \dots, Y_{(r)m}\}$ obtained through m cycles of the RSS design.

2.4 Ranking Mechanisms

In RSS, the ranking mechanism is said to be *perfect* if the unit assigned judgment rank r corresponds exactly to the true r -th order statistic for all $r = 1, \dots, k$. When an error exists between the assigned rank and the true rank, the ranking is said to be *imperfect*. While perfect ranking is rarely achieved in practice, many results established under perfect ranking continue to hold under imperfect ranking, provided that the ranking mechanism is *consistent*.

Definition 2.4.1 (Consistent Ranking Mechanism). A ranking mechanism is said to be consistent if

$$F(y) = \frac{1}{k} \sum_{r=1}^k F_{[r]}(y), \text{ for all } y \in \mathbb{R},$$

where $F_{[r]}(\cdot)$ denotes the CDF of $Y_{[r]}$, r -th judgment rank order statistic.

A widely used model for generating imperfect rankings in simulation studies was proposed by [Dell and Clutter \(1972\)](#). Let $R = Y + \varepsilon$, where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ is independent

of Y . Ranking is performed using R rather than Y . The correlation between Y and the ranking variable R is

$$\rho = \frac{\sigma_Y}{\sqrt{\sigma_Y^2 + \sigma_\varepsilon^2}} = \frac{1}{\sqrt{1 + \sigma_\varepsilon^2/\sigma_Y^2}},$$

which decreases as σ_ε^2 increases. Consequently, larger ranking errors result in poorer ranking quality. In the simulation study presented in this thesis, the [Dell and Clutter \(1972\)](#) model is used to generate rankers with varying levels of correlation between Y and the ranking variable R . Note that this ranking model is only useful for simulation studies.

Under perfect ranking, $Y_{[r]} = Y_{(r)}$, where $Y_{(r)}$ denotes the r -th order statistic from a SRS of size k . Perfect ranking satisfies the consistency condition. Several important imperfect ranking mechanisms are also consistent, as shown below.

(i) Imperfect ranking based on the variable of interest: [Bohn and Wolfe \(1994\)](#)

considered a ranking mechanism in which p_{rs} denotes the probability of true rank r is judged as rank s , for $r, s = 1, \dots, k$. Under this model, the distribution of the judgment order statistic $Y_{[s]}$ is given by

$$F_{[s]}(y) = \sum_{r=1}^k p_{rs} F_{(r)}(y).$$

Ranking mechanism described above is consistent whenever $\sum_{r=1}^k p_{rs} = \sum_{s=1}^k p_{rs} = 1$.

(ii) Ranking based on a concomitant variable: In many applications, ranking is performed using a concomitant variable X , and measurements are obtained only for the selected units on Y . Let $X_{(r)}$ denote the r th order statistic of X and let $Y_{[r]}$ denote the corresponding induced order statistic of Y . If $f_{Y|X_{(r)}}(y|x)$ denotes

the conditional density of Y given $X_{(r)} = x$ and $f_{(r)}(x)$ denotes the density of $X_{(r)}$, then

$$f_{[r]}(y) = \int f_{Y|X_{(r)}}(y|x) f_{(r)}(x) dx.$$

Ranking based on a concomitant variable produces a consistent ranking mechanism.

- (iii) **Ranking based on multiple concomitant variables:** When multiple concomitant variables are available, ranking can be performed using any function of these variables. Such procedures often improve ranking accuracy and consequently increase the efficiency of RSS estimators. Under suitable regularity conditions, these ranking mechanisms also satisfy the consistency property. Additional details regarding the ranking based on multiple concomitant variables can be found in Chapter 6 of [Chen et al. \(2003\)](#).

2.5 Estimation of Means using RSS

Consider a measurable function $h(\cdot)$ with finite mean. Let $\mu_h = \mathbb{E}_F[h(Y)]$ be the target parameter of interest. Suppose that $\{Y_{(r)j} : r = 1, \dots, k, j = 1, \dots, m\}$ represents a RSS of size mk .

Theorem 2.5.1. The RSS estimator $\hat{\mu}_{h,R} = \frac{1}{mk} \sum_{r=1}^k \sum_{j=1}^m h(Y_{(r)j})$ is an unbiased estimator of the population mean $\mu_h = \mathbb{E}_F[h(Y)]$ with variance

$$\mathbb{V}ar_F(\hat{\mu}_{h,R}) = \frac{1}{mk^2} \sum_{r=1}^k \mathbb{V}ar_F(h(Y_{(r)})) = \mathbb{V}ar_F(\hat{\mu}_{h,S}) - \frac{1}{mk^2} \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2, \quad (2.4)$$

where $\mu_{h,(r)} = \mathbb{E}_F[h(Y_{(r)})]$, and $\hat{\mu}_{h,S}$ is an estimator of μ_h based on an SRS of size $n = mk$ from the underlying population.

Proof. As observations obtained from m cycles are identically distributed, we have

$$\begin{aligned}\mathbb{E}_F[\hat{\mu}_{h,R}] &= \frac{1}{k} \sum_{r=1}^k \mathbb{E}_F[h(Y_{(r)})] \\ &= \int h(y) \left[\frac{1}{k} \sum_{r=1}^k f_{(r)}(y) \right] dy \\ &= \int h(y) f(y) dy = \mu_h.\end{aligned}$$

Since $Y_{(r)j}$'s are independent across cycles, and observations from different cycles are identically distributed:

$$\begin{aligned}\mathbb{V}ar_F(\hat{\mu}_{h,R}) &= \frac{1}{mk^2} \sum_{r=1}^k \mathbb{V}ar_F(h(Y_{(r)})) \\ &= \frac{1}{mk^2} \sum_{r=1}^k (\mathbb{E}_F[h^2(Y_{(r)})] - \mathbb{E}_F^2[h(Y_{(r)})]) \\ &= \frac{1}{mk} \left(\frac{1}{k} \sum_{r=1}^k \mathbb{E}_F[h^2(Y_{(r)})] - \frac{1}{k} \sum_{r=1}^k \mathbb{E}_F^2[h(Y_{(r)})] \right).\end{aligned}$$

Note that $\frac{1}{k} \sum_{r=1}^k \mathbb{E}_F[h^2(Y_{(r)})] = \mathbb{E}_F[h^2(Y)] = m_{h,2}$, the second moment of $h(Y)$. Let $\mathbb{E}_F[h(Y_{(r)})] = \mu_{h,(r)}$, and note that $\frac{1}{k} \sum_{r=1}^k \mu_{h,(r)} = \mu_h$. Then

$$\begin{aligned}\mathbb{V}ar_F(\hat{\mu}_{h,R}) &= \frac{\sigma_h^2}{mk} - \frac{1}{mk^2} \left(\sum_{r=1}^k \mu_{h,(r)}^2 - k\mu_h^2 \right) \\ &= \frac{\sigma_h^2}{mk} - \frac{1}{mk^2} \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2.\end{aligned}$$

□

As $\frac{\sigma_h^2}{mk}$ is the variance of $\hat{\mu}_{h,S}$ the moment estimator of μ_h based on a SRS of size mk , Theorem (2.5.1) implies that $\hat{\mu}_{h,R}$ has a lower variance than $\hat{\mu}_{h,S}$ given equal sample sizes.

Theorem 2.5.2. Let $\hat{\mu}_{h,R} = \frac{1}{mk} \sum_{r=1}^k \sum_{j=1}^m h(Y_{(r)j})$ and $s_{h,R}^2 = \frac{1}{mk-1} \sum_{r=1}^k \sum_{i=1}^m (h(Y_{(r)i}) - \hat{\mu}_{h,R})^2$ respectively. Then

$$\mathbb{E}_F [s_{h,R}^2] = \sigma_h^2 + \frac{1}{k(mk-1)} \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2.$$

Proof. Using the definition of sample variance, we have

$$s_{h,R}^2 = \frac{1}{mk-1} \left[\sum_{r=1}^k \sum_{i=1}^m h^2(Y_{(r)i}) - mk \hat{\mu}_{h,R}^2 \right].$$

One can easily show that

$$\mathbb{E}_F [s_{h,R}^2] = \frac{1}{mk-1} \left\{ \sum_{r=1}^k \sum_{i=1}^m \mathbb{E}_F (h^2(Y_{(r)i})) - mk \mathbb{E}_F (\hat{\mu}_{h,R}^2) \right\}.$$

Since $\mathbb{E}_F [h(Y_{(r)i})] = \mu_{h,(r)}$ and $\mathbb{V}ar_F (h(Y_{(r)i})) = \sigma_{h,(r)}^2$ then $\mathbb{E}_F [h^2(Y_{(r)i})] = \sigma_{h,(r)}^2 + \mu_{h,(r)}^2$,

$$\sum_{r=1}^k \sum_{i=1}^m \mathbb{E}_F (h^2(Y_{(r)i})) = m \sum_{r=1}^k (\sigma_{h,(r)}^2 + \mu_{h,(r)}^2).$$

Also $\mathbb{E}_F (\hat{\mu}_{h,R}^2) = \mathbb{V}ar_F (\hat{\mu}_{h,R}) + \mu_h^2 = \frac{1}{mk^2} \sum_{r=1}^k \sigma_{h,(r)}^2 + \mu_h^2$. Substituting these expressions gives

$$\mathbb{E}_F [s_{h,R}^2] = \frac{1}{mk-1} \left[\left(m - \frac{1}{k} \right) \sum_{r=1}^k \sigma_{h,(r)}^2 + m \sum_{r=1}^k \mu_{h,(r)}^2 - mk \mu_h^2 \right].$$

Since $\sigma_h^2 = \frac{1}{k} \sum_{r=1}^k \sigma_{h,(r)}^2 + \frac{1}{k} \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2$, then $\sum_{r=1}^k \sigma_{h,(r)}^2 = k\sigma_h^2 - \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2$, and $\sum_{r=1}^k \mu_{h,(r)}^2 - k\mu_h^2 = \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2$. Therefore

$$\begin{aligned} \mathbb{E}_F [s_{h,R}^2] &= \frac{1}{mk-1} \left[(mk-1)\sigma_h^2 + \frac{1}{k} \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2 \right] \\ &= \sigma_h^2 + \frac{1}{k(mk-1)} \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2. \end{aligned}$$

□

2.6 Bagging with Ranked Set Sampling (RSB)

Suppose we have an RSS training sample of size $n = mk$ denoted by $\mathcal{D}_R = \{Y_{(r)j} : r = 1, \dots, k, j = 1, \dots, m\}$. RSS data can be written as a collection of samples from rank strata $\mathcal{D}_{(r)}$, where $\mathcal{D}_{(r)} = \{Y_{(r)1}, \dots, Y_{(r)m}\}$, $r = 1, \dots, k$.

We saw that $F_{n,R}(y) = \frac{1}{k} \sum_{r=1}^k F_{m,(r)}(y)$ where the EDF for rank r data $\mathcal{D}_{(r)}$ is given by $F_{m,(r)}(y) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{Y_{(r)i} \leq y\}$, and the rank-wise sample mean given by $\hat{\mu}_{h,(r)} = \frac{1}{m} \sum_{i=1}^m h(Y_{(r)i})$. The RSS estimator of μ is then the average of observations across all ranks and is given by

$$\hat{\mu}_{h,R} = \frac{1}{k} \sum_{r=1}^k \hat{\mu}_{h,(r)}. \quad (2.5)$$

Having access to $\mathcal{D}_{(r)}$'s, one can easily perform bootstrap resampling within each rank stratum. This is the main idea of the proposed RSS bagging methodology. Rather than resampling from the pooled sample as in conventional SRS-based bagging, resampling is performed separately within each rank, thereby preserving the rank structure inherent in the RSS. Consequently, when we construct RSS bootstrap samples from the fully observed $\mathcal{D}_R$ data, the imperfection of the original ranking only affects the distribution of $\mathcal{D}_R$, not the bootstrap data. Taking B bootstrap samples by resampling independently within each rank, for each bootstrap $b = 1, \dots, B$, for each rank $r = 1, \dots, k$, and for each cycle $j = 1, \dots, m$, we draw $Y_{(r)}^{**(b,j)} \stackrel{\text{i.i.d.}}{\sim} F_{m,(r)}$. The bootstrap sample mean for the b -th bootstrap replicate is given by

$$\hat{\mu}_{h,R}^{*b} = \frac{1}{mk} \sum_{r=1}^k \sum_{j=1}^m h(Y_{(r)}^{**(b,j)}). \quad (2.6)$$

Averaging over all B bootstrap replicates yields the proposed bagged RSS estimator of

the mean-type functional, defined as

$$\hat{\mu}_{h,\text{RSB}} = \frac{1}{B} \sum_{b=1}^B \hat{\mu}_{h,\text{R}}^{*b} = \frac{1}{Bmk} \sum_{b=1}^B \sum_{r=1}^k \sum_{j=1}^m h(Y_{(r)}^{**(b,j)}). \quad (2.7)$$

Now we outline the main theoretical properties of the proposed RSS bagging estimator. In particular, we focus on the unbiasedness of the RSB mean estimator and the corresponding expression for its variance.

We note that conditional expectation given the observed RSS data $\mathcal{D}_R$, for any fixed rank r , bootstrap b and cycle j , the resampled observation equals one of the observed rank- r values with probability $1/m$, i.e.,

$$\mathbb{P}(Y_{(r)}^{**(b,j)} = Y_{(r)i} \mid \mathcal{D}_R) = \frac{1}{m}, \quad i = 1, \dots, m.$$

Lemma 2.6.1 (Unbiasedness of the RSB mean estimator). Conditioning on $\mathcal{D}_R$,

$$\begin{aligned} \mathbb{E}_{F_{n,\text{R}}} [h(Y_{(r)}^{**(b,j)})] &= \sum_{i=1}^m h(Y_{(r)i}) \cdot \mathbb{P}(Y_{(r)}^{**(b,j)} = Y_{(r)i} \mid \mathcal{D}_R) \\ &= \frac{1}{m} \sum_{i=1}^m h(Y_{(r)i}) = \hat{\mu}_{h,(r)}. \end{aligned}$$

Considering the definition of $\hat{\mu}_{h,\text{R}}^{*b}$ in (2.6), the expectation of the b -th bootstrap RSS

$$\begin{aligned} \mathbb{E}_{F_{m,(r)}} [\hat{\mu}_{h,\text{R}}^{*b}] &= \frac{1}{mk} \sum_{r=1}^k \sum_{j=1}^m \mathbb{E}_{F_{m,(r)}} [h(Y_{(r)}^{**(b,j)})] \\ &= \frac{1}{mk} \sum_{r=1}^k \sum_{j=1}^m \hat{\mu}_{h,(r)} \\ &= \frac{1}{k} \sum_{r=1}^k \hat{\mu}_{h,(r)} = \hat{\mu}_{h,\text{R}}. \end{aligned}$$

As $\mathbb{E}_{F_{m,(r)}} [\hat{\mu}_{h,\text{R}}^{*b}]$ does not depend on b , taking the average over $b = 1, \dots, B$ and using the expectation using the definition of (2.7), results in

$$\mathbb{E}_{F_{n,R}}[\hat{\mu}_{h,\text{RSB}}] = \hat{\mu}_{h,R}. \quad (2.8)$$

Finally, we have,

$$\mathbb{E}_F[\mathbb{E}_{F_{n,R}}[\hat{\mu}_{h,\text{RSB}}]] = \mathbb{E}_F[\hat{\mu}_{h,R}] = \mu_h. \quad (2.9)$$

Lemma 2.6.1 establishes the unbiasedness of the proposed RSB estimator, both conditionally on the observed $\mathcal{D}_R$ and unconditionally with respect to the underlying population distribution.

Theorem 2.6.1 (Variance of the RSB mean estimator). Let $\hat{\mu}_{h,\text{RSB}}$ be the RSB estimator defined in (2.7).

(i) The conditional variance of $\hat{\mu}_{h,\text{RSB}}$ given $\mathcal{D}_R$ is

$$\mathbb{V}ar_{F_{n,R}}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) = \frac{m-1}{Bm^2k^2} \sum_{r=1}^k s_{h(r)}^2, \quad (2.10)$$

where $s_{h(r)}^2 = \frac{1}{m-1} \sum_{i=1}^m (h(Y_{(r)i}) - \hat{\mu}_{h,(r)})^2$.

(ii) The unconditional variance of $\hat{\mu}_{h,\text{RSB}}$ is given by

$$\mathbb{V}ar_F(\hat{\mu}_{h,\text{RSB}}) = \frac{\sigma_h^2}{mk} \left(\frac{Bmk + mk - 1}{Bmk} \right) - \frac{1}{mk^2} \left(1 - \frac{1}{Bmk} \right) \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2. \quad (2.11)$$

Proof. (i) The variance of the RSB estimator is

$$\begin{aligned} \mathbb{V}ar_{F_{n,R}}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) &= \frac{1}{B^2} \sum_{b=1}^B \mathbb{V}ar_{F_{n,R}}(\hat{\mu}_{h,R}^{*b} \mid \mathcal{D}_R) \\ &= \frac{1}{B^2mk^2} \sum_{b=1}^B \sum_{r=1}^k \mathbb{V}ar_{F_{n,R}}(h(Y_{(r)}^{**b,j}) \mid \mathcal{D}_R), \end{aligned}$$

because of the independence across bootstrap samples. Also, it is easy to see that

$$\mathbb{V}ar_{F_{n,R}}(h(Y_{(r)}^{** (b,j)}) \mid \mathcal{D}_R) = \frac{1}{m} \sum_{i=1}^m (h(Y_{(r)i}) - \hat{\mu}_{h,(r)})^2 = \frac{m-1}{m} s_{h(r)}^2.$$

Therefore, the conditional variance of the RSB estimator is obtained as follows

$$\mathbb{V}ar_{F_{n,R}}(\hat{\mu}_{h,RSB} \mid \mathcal{D}_R) = \frac{m-1}{Bm^2k^2} \sum_{r=1}^k s_{h(r)}^2.$$

(ii) Using the Theorem 2.5.2, the unconditional variance of the RSB estimator is

$$\begin{aligned} \mathbb{V}ar_F(\hat{\mu}_{h,RSB}) &= \mathbb{E}_F[\mathbb{V}ar_{F_{n,R}}(\hat{\mu}_{h,RSB} \mid \mathcal{D}_R)] + \mathbb{V}ar_F(\mathbb{E}_{F_{n,R}}[\hat{\mu}_{h,RSB} \mid \mathcal{D}_R]) \\ &= \mathbb{E}_F\left[\frac{m-1}{Bm^2k^2} \sum_{r=1}^k s_{h(r)}^2\right] + \mathbb{V}ar_F(\hat{\mu}_{h,R}) \\ &= \frac{mk-1}{Bm^2k^2} \mathbb{E}_F[s_{h,R}^2] + \left(\frac{\sigma_h^2}{mk} - \frac{1}{mk^2} \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2\right) \\ &= \frac{\sigma_h^2}{mk} \left(\frac{Bmk + mk - 1}{Bmk}\right) - \frac{1}{mk^2} \left(1 - \frac{1}{Bmk}\right) \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2, \end{aligned}$$

where $\mathbb{E}_F[s_{h,R}^2] = \sigma_h^2 + \frac{1}{k(mk-1)} \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2$. □

Once the conditional and unconditional variances of the RSB estimator are obtained, its variance is compared to that of the conventional SSB estimator. The subsequent proposition illustrates the variance reduction achieved by the RSB estimator relative to the SSB estimator.

Proposition 2.6.1 (Variance comparison between SSB and RSB estimators). For the SSB and RSB mean estimators with B bootstrap replicates, set size k , and cycle size m :

(i) The conditional variances satisfy

$$\mathbb{V}ar_{F_{n,R}}(\hat{\mu}_{h,SSB} \mid \mathcal{D}_S) = \mathbb{V}ar_{F_{n,R}}(\hat{\mu}_{h,RSB} \mid \mathcal{D}_R) + \frac{1}{Bmk^2} \sum_{r=1}^k (\hat{\mu}_{h,(r)} - \hat{\mu}_{h,S})^2, \quad (2.12)$$

and thus

$$\mathbb{V}ar_{F_{n,R}}(\hat{\mu}_{h,SSB} \mid \mathcal{D}_S) \geq \mathbb{V}ar_{F_{n,R}}(\hat{\mu}_{h,RSB} \mid \mathcal{D}_R). \quad (2.13)$$

(ii) The unconditional variances satisfy

$$\mathbb{V}ar_F(\hat{\mu}_{h,SSB}) = \mathbb{V}ar_F(\hat{\mu}_{h,RSB}) + \frac{1}{mk^2} \left(1 - \frac{1}{Bmk}\right) \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2, \quad (2.14)$$

and hence

$$\mathbb{V}ar_F(\hat{\mu}_{h,SSB}) \geq \mathbb{V}ar_F(\hat{\mu}_{h,RSB}). \quad (2.15)$$

Proof. (i) From the sample variance definition and by rearranging the data into a matrix of size $n = mk$, we have

$$\begin{aligned} s_{n,h}^2 &= \frac{1}{mk-1} \left[\sum_{r=1}^k \sum_{i=1}^m (h(Y_{(r)i}) - \hat{\mu}_{h,S})^2 \right] \\ &= \frac{1}{mk-1} \left[\sum_{r=1}^k \sum_{i=1}^m (h(Y_{(r)i}) - \hat{\mu}_{h,(r)} + \hat{\mu}_{h,(r)} - \hat{\mu}_{h,S})^2 \right] \quad \text{where} \quad \hat{\mu}_{h,(r)} = \frac{1}{m} \sum_{i=1}^m h(Y_{(r)i}) \\ &= \frac{m-1}{mk-1} \sum_{r=1}^k s_{h(r)}^2 + \frac{m}{mk-1} \sum_{r=1}^k (\hat{\mu}_{h,(r)} - \hat{\mu}_{h,S})^2, \end{aligned}$$

where the cross product term is zero due to $\sum_{i=1}^m (h(Y_{(r)i}) - \hat{\mu}_{h,(r)}) = 0$. Also, under the usual SRS bagging approach we have

$$\begin{aligned} \mathbb{V}ar_{F_n}(\hat{\mu}_{h,SSB} \mid \mathcal{D}_S) &= \frac{n-1}{Bn^2} s_{n,h}^2 \\ &= \frac{mk-1}{Bm^2k^2} \left[\frac{m-1}{mk-1} \sum_{r=1}^k s_{h(r)}^2 + \frac{m}{mk-1} \sum_{r=1}^k (\hat{\mu}_{h,(r)} - \hat{\mu}_{h,S})^2 \right] \\ &= \frac{m-1}{Bm^2k^2} \sum_{r=1}^k s_{h(r)}^2 + \frac{1}{Bmk^2} \sum_{r=1}^k (\hat{\mu}_{h,(r)} - \hat{\mu}_{h,S})^2. \end{aligned}$$

From (2.10), we have

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) = \frac{m-1}{Bm^2k^2} \sum_{r=1}^k s_{h(r)}^2.$$

Therefore,

$$\begin{aligned} \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) &= \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) + \frac{1}{Bmk^2} \sum_{r=1}^k (\hat{\mu}_{h,(r)} - \hat{\mu}_{h,S})^2 \\ &\geq \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) \end{aligned}$$

Proof of part (ii) follows from (2.11). Note that

$$\frac{\sigma_h^2}{mk} \left(\frac{Bmk + mk - 1}{Bmk} \right) = \mathbb{V}ar_F(\hat{\mu}_{h,\text{RSB}}) + \frac{1}{mk^2} \left(1 - \frac{1}{Bmk} \right) \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2.$$

Hence

$$\begin{aligned} \mathbb{V}ar_F(\hat{\mu}_{h,\text{SSB}}) &= \mathbb{V}ar_F(\hat{\mu}_{h,\text{RSB}}) + \frac{1}{mk^2} \left(1 - \frac{1}{Bmk} \right) \sum_{r=1}^k (\mu_{h,(r)} - \mu_h)^2 \\ &\geq \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R), \end{aligned}$$

and this completes the proof. $\square$

Proposition 2.6.1 demonstrates that the RSB estimator achieves greater variance reduction compared to the SSB estimator. This result highlights the potential advantage of incorporating RSS into the bagging procedure to improve estimator stability.

2.7 Regression Estimation with RSB

In Section 2.6, we developed a bagging methodology for general mean-type functionals for RSS data, similar to what we did in Chapter 1, Section 1.5, for SRS training

data, we extend our results to regression estimation under RSS with imperfect ranking. Specifically, our goal is to study the effect of bagging on estimating the regression function $m(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$, at a fixed prediction point $\mathbf{x}$.

Let $(\mathbf{X}, Y)$ be a pair of random variables with joint distribution F , where $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathbb{R}$. Suppose the training data are obtained using RSS with imperfect ranking, $\mathcal{D}_R = \{(\mathbf{X}_{[r]j}, Y_{[r]j}) : r = 1, \dots, k, j = 1, \dots, m\}$, where $n = mk$, and $(\mathbf{X}_{[r]j}, Y_{[r]j})$ denotes the observation selected with nominal rank r in cycle j . Thus, the RSS sample consists of k rank strata, each containing m observations. Moreover, when we construct RSS bootstrap replicates, any imperfections in the original $\mathcal{D}_R$ do not affect the bootstrap ranking itself.

2.7.1 Linear Functional Representation of Regression Estimators under RSS

We view a regression procedure as a map that, given RSS training data, returns a fitted regression function. Under RSS, we start with $\mathcal{D}_R$ and the procedure produces a random function

$$\hat{m}_R(\cdot) : \mathcal{X} \rightarrow \mathbb{R},$$

which, when evaluated at a prediction point $\mathbf{x} \in \mathcal{X}$, yields the fitted value $\hat{m}_R(\mathbf{x})$. This construction is analogous to the SRS case in Section 1.6.1. Specifically, for each fixed prediction point $\mathbf{x}$, the fitted value $\hat{m}_R(\mathbf{x})$ can be expressed as a finite linear combination of the observed responses. This representation motivates the following definition.

Definition 2.7.1 (Linear smoother representation). Fix $\mathbf{x} \in \mathcal{X}$. Under RSS, training data $\mathcal{D}_R$ with $n = mk$ observed pairs. We say that the regression estimator is a *linear smoother* under RSS if, for each $\mathbf{x}$,

$$\hat{m}_R(\mathbf{x}) = \frac{1}{k} \sum_{r=1}^k \hat{m}_{[r]}(\mathbf{x}) \quad \text{with} \quad \hat{m}_{[r]}(\mathbf{x}) = \sum_{j=1}^m W_{[r]j}(\mathbf{x}) Y_{[r]j} \quad (2.16)$$

for (random) weight functions $W_{[r]j}(\cdot)$ depending on $\mathcal{D}_R$. Equivalently, defining

$$Z_{[r]j}(\mathbf{x}) = mW_{[r]j}(\mathbf{x})Y_{[r]j}, \quad (2.17)$$

we obtain

$$\hat{m}_R(\mathbf{x}) = \frac{1}{k} \sum_{r=1}^k \hat{m}_{[r]}(\mathbf{x}) = \frac{1}{k} \sum_{r=1}^k \frac{1}{m} \sum_{j=1}^m Z_{[r]j}(\mathbf{x}). \quad (2.18)$$

Under a consistent ranking mechanism, the results established for perfect ranking extend naturally to the imperfect-ranking. This representation is identical in form to the RSS estimator of a mean functional Equation (2.5),

$$\hat{\mu}_{h,R} = \frac{1}{k} \sum_{r=1}^k \hat{\mu}_{h,[r]} = \frac{1}{k} \sum_{r=1}^k \frac{1}{m} \sum_{j=1}^m h(Y_{[r]j}),$$

with the pseudo-responses $Z_{[r]j}(\mathbf{x})$ playing the role of $h(Y_{[r]j})$.

In the RSS setting, the collection of functions $\{Z_{[r]j}(\mathbf{x})\}_{r,j}$ determines the fitted regression function through a linear combination of the observed pseudo responses. This representation emphasizes that the regression procedure first constructs the fitted regression function $\hat{m}_R(\mathbf{x})$ and predictions are then obtained by evaluating this function at the desired prediction point $\mathbf{x}$. The examples in Section 1.6.1 illustrate that the linear functional representation applies to a broad class of regression estimators. Since the same representation holds under RSS, the RSB bootstrap construction and the associated variance expressions extend directly and can be applied pointwise in $\mathbf{x}$ through the pseudo-responses $Z_{[r]j}(\mathbf{x})$.

2.7.2 Linear Smoother RSB Regression Estimator

We now extend the finite-sample RSB bootstrap theory developed in Section 2.6 to regression estimators admitting the linear representations described above. We work

point-wise in $\mathbf{x}$, treating the corresponding pseudo-responses $Z_{[r]j}(\mathbf{x})$ as the observations to which the bootstrap procedure is applied.

Definition 2.7.2 (RSB Regression Estimator). Fix $\mathbf{x} \in \mathcal{X}$ and let $Z_{[r]j}(\mathbf{x})$ be as in (2.17). For each rank $r = 1, \dots, k$ and bootstrap replicate $b = 1, \dots, B$:

- (a) Draw rank-wise bootstrap indices $J_{[r]1}^{*b}, \dots, J_{[r]m}^{*b}$ i.i.d. from $\{1, \dots, n\}$ with replacement, corresponding to an SRS bootstrap sample of size m within rank r .
- (b) Define the rank-wise bootstrap pseudo-responses $Z_{[r]j}^{*b}(\mathbf{x})$, $j = 1, \dots, m$, and the rank-wise bootstrap estimator $\hat{m}_{[r]}^{*b}(\mathbf{x}) = \frac{1}{m} \sum_{j=1}^m Z_{[r]j}^{*b}(\mathbf{x})$.

The full RSS bootstrap estimator at $\mathbf{x}$ for replicate b is

$$\hat{m}_R^{*b}(\mathbf{x}) = \frac{1}{k} \sum_{r=1}^k \hat{m}_{[r]}^{*b}(\mathbf{x}),$$

and the B -replicate RSB estimator is

$$\hat{m}_{\text{RSB}}^{(B)}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{m}_R^{*b}(\mathbf{x}) = \frac{1}{Bk} \sum_{b=1}^B \sum_{r=1}^k \hat{m}_{[r]}^{*b}(\mathbf{x}). \quad (2.19)$$

In terms of the pseudo-responses $Z_{[r]j}(\mathbf{x})$, this construction is algebraically equivalent to the RSB estimator for mean estimation described in Section 2.6, with $h(Y_{[r]j})$ replaced by $Z_{[r]j}(\mathbf{x})$. Under this representation, the theoretical results established in this chapter for mean functionals are not restricted to the sample mean setting. The unbiasedness properties, the conditional and unconditional variance expressions, and the variance-comparison results derived for the RSB estimator can be transferred directly to the regression framework through an appropriate linear representation of the fitted regression function.

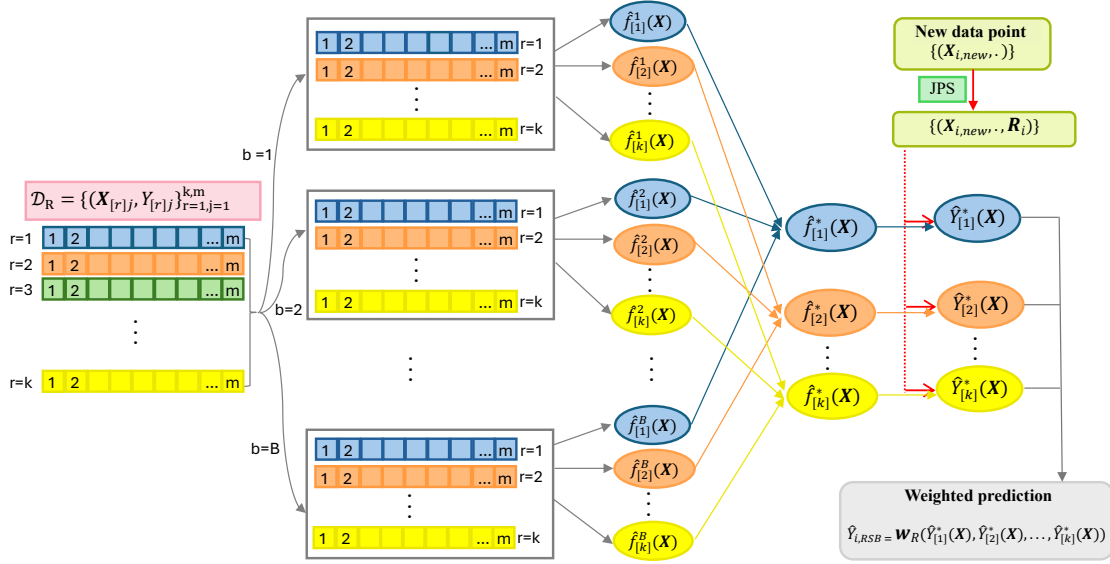


Figure 2.1: RSB resampling and prediction procedure for RSS data with set size k under paired bootstrapping.

Depending on the type of bootstrap used within the RSB procedure, the overall resampling and prediction workflow can be viewed from two main perspectives. To aid understanding of the RSB resampling and prediction procedure, a visual illustration of paired bootstrapping is provided. Figure 2.1 illustrates the complete workflow of the RSB prediction procedure under paired bootstrapping. Starting from the RSS training sample, bootstrap samples are generated independently within each rank stratum by sampling with replacement from the m observations belonging to that rank. For each bootstrap sample, we fit a single regression model and use it to obtain k rank-specific predictions at the target point $\mathbf{x}$. We then aggregate these predictions across the B bootstrap replicates to form bagged rank-specific predictors. For a new SRS test observation, the judgment post-stratification (JPS) procedure estimates its rank distribution, and we then combine the rank-specific predictions through a weighted aggregation scheme to obtain the final RSB prediction.

A more comprehensive algorithmic description covering both paired and residual bootstrapping is provided in Chapter 4.

2.8 Modeling Approach

In this section, we extend the approaches presented in [Faraji et al. \(2022\)](#) to estimate the relationship between $\mathbf{X}$ and Y using RSS data by incorporating bagging. To this end, we first provide a brief overview of the results in [Faraji et al. \(2022\)](#). Consider RSS data $\mathcal{D}_R = \{(\mathbf{X}_{[r]j}, Y_{[r]j}) : r = 1, \dots, k, j = 1, \dots, m\}$ from the underlying population, with the model $Y_{[r]j} = \mathbb{X}_{[r]j}^\top \boldsymbol{\beta} + \varepsilon_{[r]j}$, where $r = 1, \dots, k$ and $j = 1, \dots, m$, which can be represented in matrix form as:

$$\mathbf{Y}_{[RSS]} = \mathbb{X}_{[RSS]} \boldsymbol{\beta} + \boldsymbol{\varepsilon}_{[RSS]}, \quad (2.20)$$

where

$$\mathbf{Y}_{[RSS]} = \begin{bmatrix} \mathbf{Y}_{[1]} \\ \mathbf{Y}_{[2]} \\ \vdots \\ \mathbf{Y}_{[k]} \end{bmatrix}_{mk \times 1}, \quad \mathbb{X}_{[RSS]} = \begin{bmatrix} \mathbf{X}_{[1]} \\ \mathbf{X}_{[2]} \\ \vdots \\ \mathbf{X}_{[k]} \end{bmatrix}_{mk \times (p+1)}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1) \times 1}, \quad \boldsymbol{\varepsilon}_{[RSS]} = \begin{bmatrix} \boldsymbol{\varepsilon}_{[1]} \\ \boldsymbol{\varepsilon}_{[2]} \\ \vdots \\ \boldsymbol{\varepsilon}_{[k]} \end{bmatrix}_{mk \times 1}.$$

Here $\mathbf{Y}_{[r]} = (y_{[r]1}, \dots, y_{[r]m})^\top$ and $\boldsymbol{\varepsilon}_{[r]} = (\varepsilon_{[r]1}, \dots, \varepsilon_{[r]m})^\top$. The $\mathbf{X}_{[r]}$ matrix is constructed by stacking the $m \times (p+1)$ covariate rows corresponding to rank r ([Faraji et al., 2022](#)).

The true error structure of $y_{[r]j}$ is modeled assuming that, for fixed r ([Tiku and Akkaya, 2004](#)),

$$\varepsilon_{[r]j} \stackrel{\text{i.i.d.}}{\sim} f_{\varepsilon_{[r]}}(t) = \sum_{i=1}^k p_{ri} f_{\varepsilon_{(i)}}(t), \quad (2.21)$$

where p_{ri} denotes the ranking error probability matrix and $f_{\varepsilon_{(i)}}(t)$ is the pdf of the i -th order statistic in an SRS of size k from $f_{\varepsilon}(\cdot)$:

$$f_{\varepsilon_{(i)}}(t) = \frac{k!}{(i-1)!(k-i)!} f_{\varepsilon}(t) F_{\varepsilon}(t)^{i-1} [1 - F_{\varepsilon}(t)]^{k-i}, \quad \text{for } i = 1, \dots, k. \quad (2.22)$$

In many practical applications, ranking information may be available from multiple concomitant variables. To estimate the ranking accuracy for each ranking variable, we

adapt the bootstrap procedure of [Nourmohammadi et al. \(2015\)](#) to the multiple-rankers setting. Suppose that L ranking variables $R^{(1)}, R^{(2)}, \dots, R^{(L)}$ are available. For each ranking variable, a separate rank-confusion matrix is estimated as follows.

- (i) Fit a regression model to the training data and obtain the fitted values and residuals. If response-based ranking is used, this step is omitted.
- (ii) Draw a bootstrap sample of size k with replacement from the available data.
- (iii) For the selected observations, determine the “true” ranks using either the response values Y or the residuals from the fitted regression model.
- (iv) Independently rank the same observations according to the ℓ -th ranking variable $R^{(\ell)}$.
- (v) Compare the ranks from step (3) with those from step (4) and record the corresponding rank assignments in a $k \times k$ confusion matrix.
- (vi) Repeat steps (2)–(5) a total of B times and compute the average confusion matrix. The resulting matrix, $\mathbb{P}_k^{(\ell)} = \left(p_{ri}^{(\ell)} \right)$, provides an estimate of the ranking probabilities for the ℓ -th ranking variable. The $p_{ri}^{(\ell)} = \mathbb{P}\left(\varepsilon_{[r]j}^{(\ell)} = \varepsilon_{(i)j} \right)$ represents the probability that an observation assigned rank r according to the ℓ -th ranking variable belongs to the i -th true rank. Here, $0 \leq p_{ri}^{(\ell)} \leq 1$, with $\sum_{i=1}^k p_{ri}^{(\ell)} = 1$, $r = 1, \dots, k$.

After obtaining $\mathbb{P}_k^{(1)}, \mathbb{P}_k^{(2)}, \dots, \mathbb{P}_k^{(L)}$, the individual matrices are combined to form an overall rank-confusion matrix $\mathbb{P}_k$.

To incorporate information from multiple ranking variables, weights are assigned based on the strength of association between each ranking variable and the response

variable. An aggregated ranking error probability matrix is then constructed as:

$$\mathbb{P}_k = \sum_{\ell=1}^L w_\ell \mathbb{P}_k^{(\ell)}, \quad (2.23)$$

where $w_\ell \geq 0$ and $\sum_{\ell=1}^L w_\ell = 1$. In this study w_ℓ is calculated as;

$$w_\ell = \frac{|\rho_{\ell Y}|}{\sum_{j=1}^L |\rho_{jY}|}, \quad \ell = 1, \dots, L, \quad (2.24)$$

where $\rho_{\ell Y}$ denotes the correlation between the ℓ -th ranking variable and the response variable Y (Hatefi and Jafari Jozani, 2017).

The resulting matrix summarizes the ranking uncertainty across all ranking variables. The key limitation of the residual-based approach is that the estimated $\mathbb{P}_k$ matrix depends on the regression model used to obtain the residuals. In contrast, the response-based approach does not depend on model fitting and is computationally simpler. Empirically, both approaches often yield comparable relative efficiencies, although response-based $\mathbb{P}_k$ matrices may introduce a small bias in parameter estimation.

Furthermore, when the set size k is large or complex machine-learning models are employed, residual-based rankings can become unstable, leading to less reliable estimates of the ranking probabilities. For this reason, response-based $\mathbb{P}_k$ estimation is often preferred in applications involving high-dimensional predictors or flexible prediction models.

With the use of (2.21) the expectation and variance of $\varepsilon_{[r]j}$ can be calculated as scaled mean $a[r] = \frac{1}{\sigma} \mathbb{E}(\varepsilon[r]_j)$ and scaled variance $d[r] = \frac{1}{\sigma^2} \text{Var}(\varepsilon[r]_j)$. It is easy to show that $\text{Cov}(\boldsymbol{\varepsilon}^{[RSS]}) = \sigma^2 \mathbb{C}$ and $\mathbb{E}(\boldsymbol{\varepsilon}^{[RSS]}) = \sigma \mathbb{A}$, with scaled mean matrix:

$$\mathbb{A} = (a_{[1]} \mathbb{1}_m^\top, \dots, a_{[k]} \mathbb{1}_m^\top)^\top \quad (2.25)$$

and covariance matrix:

$$\mathbb{C} = \text{diag}(d[1] \mathbf{I}_m, d[2] \mathbf{I}_m, \dots, d[k] \mathbf{I}_m) \quad (2.26)$$

The inverse matrix $\mathbb{C}^{-1}$ provides the weights used in the weighted estimation procedures. Given $\mathbb{P}_k$, we compute the mean shift parameters $a_{[r]}$ and variance inflation factors $d_{[r]}$ as:

$$a[r] = \sum_{i=1}^k \hat{p}_{ri} \int_0^1 \Phi^{-1}(w) \text{Beta}_{i,k-i+1}(w) dw, \quad (2.27)$$

$$d[r] = \sum_{i=1}^k \hat{p}_{ri} \int_0^1 [\Phi^{-1}(w)]^2 \text{Beta}_{i,k-i+1}(w) dw - a[r]^2, \quad (2.28)$$

where $\Phi^{-1}(\cdot)$ is the cumulative distribution function of the standard normal and $\text{Beta}_{i,k-i+1}$ is a beta distribution with parameters $(i, k - i + 1)$ (Faraji et al., 2022).

Equations (2.27) and (2.28) indicate that the error terms associated with ranked set samples generally do not share a common variance, and their means are not necessarily equal to zero. Consequently, standard regression procedures suggested by Muttalak (1995), may fail to fully account for the distributional characteristics induced by the ranks. To address this issue, we extended two main approaches presented in Faraji et al. (2022), incorporating these features into the estimation of the relationship between Y and $\mathbf{X}$ in the bagging framework. The first approach is based on adjusting the response variable to account for ranking effects, whereas the second employs a multilevel modeling framework to explicitly accommodate the rank-specific error structure.

2.8.1 Weighted Least Squares with Adjusted Response Variable using RSB

Building upon the weighted least squares (WLS) approach to RSS proposed by Faraji et al. (2022), this method is extended to support bagging. The proposed framework repeatedly fits the RSS regression model on bootstrap samples and combines the resulting estimates, thereby reducing estimator variance while preserving the efficiency gains

associated with ranked set sampling. The main idea is to apply the bias-adjusted WLS estimation procedure to each bootstrap replicate of the RSS dataset and then aggregate the results.

Following the single RSS framework presented in [Faraji et al. \(2022\)](#), the underlying model is first adjusted to have zero-mean error terms for each order statistic $y_{[r]j}$ and its concomitants. We correct for the bias in $\boldsymbol{\varepsilon}_{[RSS]}$, assuming σ is known (later estimated via a plug-in method). By adding and subtracting $\sigma\mathbb{A}$ to the model (2.20), we obtain

$$\mathbf{Y}_{[RSS]}^* = \mathbb{X}_{[RSS]}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{[RSS]}^*, \quad (2.29)$$

where $\boldsymbol{\varepsilon}_{[RSS]}^* = \boldsymbol{\varepsilon}_{[RSS]} - \sigma\mathbb{A}$, $\mathbb{E}(\boldsymbol{\varepsilon}_{[RSS]}^*) = 0$, $\text{Cov}(\boldsymbol{\varepsilon}_{[RSS]}^*) = \sigma^2\mathbb{C}$, and $\mathbf{Y}_{[RSS]}^* = \mathbf{Y}_{[RSS]} - \sigma\mathbb{A}$. The weighted least squares estimator for a given dataset is then

$$\hat{\boldsymbol{\beta}}(\sigma) = (\mathbb{X}_{[RSS]}^\top \mathbb{C}^{-1} \mathbb{X}_{[RSS]})^{-1} \mathbb{X}_{[RSS]}^\top \mathbb{C}^{-1} \mathbf{Y}_{[RSS]}^*. \quad (2.30)$$

Now, consider a bootstrap RSS sample $\mathcal{D}_R^{(b)}$, $b = 1, \dots, B$, drawn from the original RSS dataset. For each bootstrap sample $b \in \{1, \dots, B\}$, we apply the WLS estimation above, producing

$$\hat{\boldsymbol{\beta}}^{(b)}(\sigma) = \left(\mathbb{X}_{[RSS]}^{(b)\top} \mathbb{C}^{-1} \mathbb{X}_{[RSS]}^{(b)} \right)^{-1} \mathbb{X}_{[RSS]}^{(b)\top} \mathbb{C}^{-1} \mathbf{Y}_{[RSS]}^{*(b)}. \quad (2.31)$$

When the number of predictors is large or variable selection is desired, the WLS estimator is replaced by a weighted LASSO estimator. Specifically, the coefficient vector is obtained by minimizing

$$\hat{\boldsymbol{\beta}}^{(b)} = \arg \min_{\boldsymbol{\beta}} \left\{ \left(\mathbf{Y}_{[RSS]}^{*(b)} - \mathbb{X}_{[RSS]}^{(b)} \boldsymbol{\beta} \right)^\top \mathbb{C}^{-1} \left(\mathbf{Y}_{[RSS]}^{*(b)} - \mathbb{X}_{[RSS]}^{(b)} \boldsymbol{\beta} \right) + \lambda \sum_{j=1}^p |\beta_j| \right\}, \quad (2.32)$$

where λ is selected using cross-validation. In the implementation, a five-fold cross-validated LASSO procedure is employed to determine the optimal penalty parameter.

Applying either WLS or weighted LASSO, this yields B bootstrap-specific parameter estimates $\{\hat{\beta}^{(1)}, \dots, \hat{\beta}^{(B)}\}$, each accompanied by order-specific bias corrections. Finally, the bagged estimator is then obtained by averaging the bootstrap-specific estimates:

$$\hat{\beta}_{\text{bag}} = \frac{1}{B} \sum_{b=1}^B \hat{\beta}^{(b)}. \quad (2.33)$$

Similarly, rank-specific parameter estimates (indexed by r) are aggregated across bootstraps, resulting in r parallel regression models that share the same slope but have distinct intercepts and bias adjustments for each order statistic:

$$\hat{\beta}_{\text{bag},[r]} = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_{[r]}^{(b)}, \quad r = 1, \dots, k. \quad (2.34)$$

As in the single-RSS case, when σ is unknown, we follow the plug-in approach of [Tiku and Akkaya \(2004\)](#), replacing σ with its estimate $\hat{\sigma}$. Let $\mathbf{Y}_{[RSS]}^{**{(b)}} = \mathbf{Y}_{[RSS]}^{*{(b)}} - \hat{\sigma}^{(b)}\mathbb{A}$, and obtain the parameter estimates

$$\hat{\beta}^{(b)} = \left(\mathbb{X}_{[RSS]}^{(b)\top} \mathbb{C}^{-1} \mathbb{X}_{[RSS]}^{(b)} \right)^{-1} \mathbb{X}_{[RSS]}^{(b)\top} \mathbb{C}^{-1} \mathbf{Y}_{[RSS]}^{**{(b)}}, \quad (2.35)$$

for each bootstrap replicate b . The bias-corrected fitted values for replicate b are then

$$\hat{\mathbf{Y}}_{[RSS]}^{(b)} = \mathbb{X}_{[RSS]}^{(b)} \hat{\beta}^{(b)} + \hat{\sigma}^{(b)} \mathbb{A}. \quad (2.36)$$

To estimate σ , we use the method of [Stokes \(1980\)](#),

$$\hat{\sigma}^2 = \frac{1}{mk-1} \sum_{j=1}^m \sum_{r=1}^k (y_{[r]j} - \hat{\mu})^2, \quad \hat{\mu} = \frac{1}{mk} \sum_{j=1}^m \sum_{r=1}^k y_{[r]j}, \quad (2.37)$$

which is asymptotically unbiased.

Remark 2.8.1. The estimation procedure described above applies to both bootstrap schemes. For paired bootstrapping, $\mathbb{X}_{[RSS]}^{(b)}$ and $\mathbf{Y}_{[RSS]}^{(b)}$ are obtained by resampling observations from the RSS dataset. For residual bootstrapping, $\mathbb{X}_{[RSS]}^{(b)} = \mathbb{X}_{[RSS]}$ remains fixed and only the response vector $\mathbf{Y}_{[RSS]}^{(b)}$ is replaced by the bootstrap generated responses.

2.8.2 Multilevel Modeling Approach toward Regression Analysis with RSB

RSS data inherently display a clustering structure due to the way observations are collected. Observations within the same rank are typically more similar to each other than to those from different ranks. Therefore, observations can be considered as clustered by rank. To address this dependence structure, the linear mixed model (LMM) framework proposed by Faraji et al. (2022) is extended by incorporating bagging. In this approach, the total variability was decomposed into between-rank and within-rank variance components.

Following Faraji et al. (2022), we use the random-intercept LMM

$$\mathbf{Y}_{[RSS]} = \mathbb{X}_{[RSS]}\boldsymbol{\beta} + \mathbb{Z}\mathbf{u} + \boldsymbol{\varepsilon}^{\star}, \quad (2.38)$$

with $\mathbb{E} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon}^{\star} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$ and $\text{Cov} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon}^{\star} \end{bmatrix} = \begin{bmatrix} \mathbb{G} & \mathbf{0} \\ \mathbf{0} & \mathbb{R} \end{bmatrix}$ where $\mathbf{u}_{k \times 1} = (U_1, \dots, U_k)^{\top}$, $\mathbb{R}_{n \times n} = \sigma_{\varepsilon^{\star}}^{(2)} \mathbb{I}_n$, $\mathbb{Z}_{mk \times k}$ contains the rank-specific random effects, $\mathbf{u} \sim N(\mathbf{0}, \mathbb{G})$, and $\boldsymbol{\varepsilon}^{\star} \sim N(\mathbf{0}, \mathbb{R})$. For the j -th observation at rank r , the error

$$\varepsilon_{[r]j} = u_{[r]} + \varepsilon_{rj}^{\star},$$

splits into a rank-specific random effect $u_{[r]}$ and pure residual $\varepsilon_{rj}^{\star}$. $\mathbb{G}$ represents the variance-covariance structure per rank group, which can be written in the form

$$\mathbb{G}_{k \times k} = \text{diag}(\sigma_{[1]}^2, \sigma_{[2]}^2, \dots, \sigma_{[k]}^2),$$

and $\mathbb{Z}$ is the random effects design matrix in the form

$$\mathbb{Z}_{mk \times k} = \text{diag}(\mathbf{1}_{m \times 1}, \mathbf{1}_{m \times 1}, \dots, \mathbf{1}_{m \times 1}).$$

For the random-intercept model

$$y_{[r]j} = \mathbb{X}_{[r]j}^\top \beta + u_{[r]} + \varepsilon_{[r]j}^*, \quad r = 1, \dots, k, \quad j = 1, \dots, m.$$

Consequently, $\mathbb{E}(y_{[r]j} \mid u_{[r]}) = \mathbb{X}_{[r]j}^\top \beta + u_{[r]}$ and $\text{Var}(y_{[r]j}) = \text{Var}(u_{[r]}) + \text{Var}(\varepsilon_{[r]j}^*) = \sigma_{[r]}^2 + \sigma_\varepsilon^2$. The covariance matrix of the response is $\mathbb{V} \equiv \text{Cov}(y) = \mathbb{Z}\mathbb{G}\mathbb{Z}^\top + \mathbb{R}$. Suppose that $\mathbb{G}$ and $\mathbb{R}$ are known, to estimate the β we can use

$$\beta(\mathbb{V}) = (\mathbb{X}_{[RSS]}^\top \mathbb{V}^{-1} \mathbb{X}_{[RSS]})^{-1} \mathbb{X}_{[RSS]}^\top \mathbb{V}^{-1} \mathbf{Y}_{[RSS]}. \quad (2.39)$$

To predict the random effects $\mathbf{u}$ we can use

$$\hat{\mathbf{u}}(\mathbb{V}) = \mathbb{G}\mathbb{Z}^\top \mathbb{V}^{-1} \left(\mathbf{Y}_{[RSS]} - \mathbb{X}_{[RSS]} \hat{\beta}(\mathbb{V}) \right). \quad (2.40)$$

Since $\mathbb{G}$ and $\mathbb{R}$ are unknown in practice, their parameters must be estimated from the data. Following [Ruppert et al. \(2003\)](#), this can be done using either Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) estimation. Once estimates $\mathbb{G}$ and $\mathbb{R}$ are obtained, they are substituted into equations (2.39) and (2.40), yielding feasible estimators for the fixed effects β and the random effects $\mathbf{u}$. Unlike the adjusted-response WLS approach, the LMM framework directly models rank-specific variability via random effects and therefore does not require estimating rank error probability matrices $\mathbb{P}_k$.

Let $\mathcal{D}_R^{(b)} = (\mathbb{X}^{(b)}, \mathbf{Y}^{(b)})$ denote the b -th bootstrap replicate. For paired bootstrapping, both predictors and responses are resampled from the observed RSS data. For residual bootstrapping, the predictor matrix remains fixed, and bootstrap responses are generated as $\mathbf{Y}_i^{*(b)} = \hat{\mathbf{Y}}_i + \varepsilon_i^{*(b)}$, where $\varepsilon_i^{*(b)}$ is sampled with replacement from the fitted residuals. When the number of predictors is large, variable selection is performed before fitting the

mixed model. For each bootstrap replicate, a LASSO model is fitted using the bootstrap response vector,

$$\hat{\boldsymbol{\beta}}^{(b)} = \arg \min_{\boldsymbol{\beta}} \left\{ \left(\mathbf{Y}_{[RSS]}^{*(b)} - \mathbb{X}_{[RSS]}^{(b)} \boldsymbol{\beta} \right)^\top \left(\mathbf{Y}_{[RSS]}^{*(b)} - \mathbb{X}_{[RSS]}^{(b)} \boldsymbol{\beta} \right) + \lambda \sum_{j=1}^p |\beta_j| \right\}, \quad (2.41)$$

where the tuning parameter λ is selected using five-fold cross-validation. Only predictors corresponding to non-zero LASSO coefficients are retained for subsequent mixed-model estimation. For each bootstrap replicate, the selected predictors are used to fit the random-intercept LMM

$$\mathbf{Y}_{[RSS]}^{*(b)} = \mathbb{X}_{[RSS]}^{(b)} \boldsymbol{\beta}^{(b)} + \mathbb{Z} \mathbf{u}^{(b)} + \boldsymbol{\varepsilon}^{*(b)}. \quad (2.42)$$

Similarly, for each bootstrap sample, rank specific random effect estimates are obtained

$$\hat{\mathbf{u}}^{(b)}(\mathbb{V}) = \mathbb{G}^{(b)} \mathbb{Z}^\top \mathbb{V}^{(b)-1} \left(\mathbf{Y}_{[RSS]}^{*(b)} - \mathbb{X}_{[RSS]}^{(b)} \hat{\boldsymbol{\beta}}^{(b)}(\mathbb{V}) \right). \quad (2.43)$$

The bootstrap LMM procedure produces $\{\hat{\boldsymbol{\beta}}^{(1)}, \dots, \hat{\boldsymbol{\beta}}^{(B)}\}$ together with the corresponding random-effect estimates $\{\hat{\mathbf{u}}^{(1)}, \dots, \hat{\mathbf{u}}^{(B)}\}$. The bagged fixed-effect estimator is

$$\hat{\boldsymbol{\beta}}_{\text{bag}} = \frac{1}{B} \sum_{b=1}^B \hat{\boldsymbol{\beta}}^{(b)}, \quad (2.44)$$

while the bagged rank-specific random effects are

$$\hat{u}_{\text{bag},[r]} = \frac{1}{B} \sum_{b=1}^B \hat{u}_{[r]}^{(b)}, \quad r = 1, \dots, k. \quad (2.45)$$

The resulting prediction for an observation belonging to rank r is

$$\hat{g}_{[r]j} = \mathbb{X}_{[r]j}^\top \hat{\boldsymbol{\beta}}_{\text{bag}} + \hat{u}_{\text{bag},[r]}.$$

In practice, LMM can be fitted to RSS data using the `nlme` package in R and the `statsmodels` package in Python. The primary functions, `lme` in `nlme` and `mixedlm` in `statsmodels`, support random intercept models and, allowing the residual variance to differ across rank groups, directly accommodate the heteroskedasticity that arises from the order-statistic structure of the sample.

The performance of the proposed RSB methodology is evaluated in Chapter 4 through extensive simulation studies conducted under both linear and nonlinear regression settings, considering a range of ranking accuracies, set sizes, and bootstrap schemes. In addition, the practical applicability of the proposed framework is demonstrated through a real-data analysis using the Body Fat dataset. Numerical results for the proposed RSB regression framework are presented in Chapter 4.

2.9 Summary and Conclusion

In this chapter, we first discussed the limitations of classical bagging, which relies on SRS for bootstrap resampling. To overcome these limitations, we introduced ranked RSS and reviewed its fundamental properties before developing the RSB framework, which extends classical bagging to the RSS setting. By generating bootstrap samples independently within each rank stratum, the proposed resampling scheme preserves the data's ranked structure while exploiting the greater efficiency of RSS.

The theoretical properties of the RSB estimator were first established for general mean-type functionals and subsequently extended to the estimation of the regression function $m(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$. It was shown that the proposed estimators are unbiased, and explicit expressions were derived for both their conditional and unconditional variances. Furthermore, by expressing regression estimators in terms of the linear smoother representation, these theoretical results were shown to hold for a broad class of regression estimators, thereby establishing the corresponding unbiasedness and variance

properties of the RSB regression estimators.

The chapter then extended the RSB framework to predictive modeling through weighted least squares regression and multilevel modeling. To enhance its practical applicability, the ranking mechanism was further generalized to accommodate multiple ranking variables. Finally, the RSB prediction methodology was presented under the paired bootstrapping framework to illustrate the complete resampling, model fitting, and prediction process. Overall, the proposed RSB framework provides a theoretically rigorous and practically effective approach to bagging that leverages the efficiency of RSS when complete ranking information is directly available from the population.

Chapter 3

Rank-Informed Bagging

This chapter addresses the practical limitations associated with the RSS bagging (RSB) methodology developed in Chapter 2. We propose a novel bagging framework based on rank-informed samples generated from the simple random sample (SRS) training data. The proposed methodology is first established for general mean-type functionals and then extended to estimate the regression function $m(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$ at a fixed prediction point $\mathbf{x}$. Theoretical results are derived for the proposed estimators, including unbiasedness, conditional and unconditional variance expressions, and comparisons of variances with both classical SRS bagging (SSB) and RSB. In addition, the proposed framework is evaluated within the weighted least squares (WLS) and linear mixed model (LMM) settings presented in Chapter 2.

3.1 Introduction

The theoretical developments in the previous chapter and the numerical results in Chapter 4 demonstrate that RSB can achieve substantially greater variance reduction than SSB, highlighting the potential benefits of using ranking information within the bagging framework. However, these advantages come with limitations. The practical

implementation of RSS relies on the availability of RSS training data. In many real-world applications, training datasets are typically collected using SRS, and RSS is rarely available. Constructing an RSS requires either direct ranking by domain experts or accurate concomitant variables or predictive models that reliably rank units without measuring the response. Despite more information being contained in RSS data, working with RSS data is limited to situations where rank-informed models are available, and this limits application of RSS data.

This limitation motivates the development of an alternative strategy. Rather than replacing the original SRS design entirely and starting with RSS training data, we work with SRS data and aim to enrich the observed SRS training data by incorporating ranking information when available. The objective is not to reconstruct a full RSS design, but rather to utilize partially available ranking knowledge to improve the quality of the resulting bootstrap samples. This ranking information can come from several sources, such as a previously developed predictive models in the literature, lagged observations in longitudinal or time-dependent data, or expert judgment in areas such as medical problems and other domain-specific applications. The central idea is that even partial or imperfect ranking information can provide valuable structural insight into the data and should therefore be exploited to strengthen the learning process.

These considerations motivate the rank-informed bagging framework proposed in this chapter. The methodology combines the practical simplicity of SRS with the efficiency gains associated with ranking information, thereby providing a flexible and implementable alternative to both SRS bagging and proposed RSS bagging.

3.2 Bagging under SRS-RSS (SRB)

Suppose $\mathcal{D}_S = \{Y_1, \dots, Y_n\}$ is an SRS training data of size $n = mk$ with EDF $F_n(y) = \frac{1}{n} \sum_{j=1}^n \mathbb{1}\{Y_j \leq y\}$. Here, we draw an RSS bootstrap sample of size n with set size k

and cycle size m from $\mathcal{D}_S$. For the b -th bootstrap and j -th cycle, we have k observations

$$Y_1^{*(b,j)}, \dots, Y_k^{*(b,j)},$$

where $Y_1^{*(b,j)}, \dots, Y_k^{*(b,j)}$ are i.i.d. from F_n . After ranking them, we get

$$Y_{(1:k)}^{*(b,j)} \leq Y_{(2:k)}^{*(b,j)} \leq \dots \leq Y_{(k:k)}^{*(b,j)}.$$

Here, ranking is perfect across all observed $Y_i^{*(b,j)}$. One can also use ranking variables to obtain imperfect bootstrap RSS data. However, we restrict our result to perfect ranking. We retain the r -th order statistic $Y_{(r:k)}^{*(b,j)}$ as the judgment observation for rank r in that b -th bootstrap and j -th cycle.

For b -th bootstrap and j -th cycle, let $P_{k,n}^{b,j}(r, s)$ be the probability that the r -th observation of RSS bootstrap sample of size k equals the s -th order statistic of the $\mathcal{D}_S$ of size n . In other words:

$$P_{k,n}^{b,j}(r, s) = \mathbb{P}(Y_{(r:k)}^{*(b,j)} = Y_{(s:n)} \mid \mathcal{D}_S), \quad r = 1, \dots, k, \quad s = 1, \dots, n.$$

One can easily show that $P_{k,n}^{b,j}(r, s)$ can be expressed as

$$P_{k,n}^{b,j}(r, s) = I_{\frac{s}{n}}(r, k - r + 1) - I_{\frac{s-1}{n}}(r, k - r + 1),$$

where $I_x(a, b)$ is the incomplete Beta function given by

$$I_x(a, b) = \frac{\int_0^x t^{a-1}(1-t)^{b-1} dt}{B(a, b)} \text{ with } B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Some basic properties of the $P_{k,n}^{b,j}(r, s)$ are stated below.

Lemma 3.2.1. (i) For each fixed $r = 1, \dots, k$ in b^{th} bootstrap and j^{th} cycle

$$\sum_{s=1}^n P_{k,n}^{b,j}(r, s) = 1. \tag{3.1}$$

(ii) For each fixed $s = 1, \dots, n$ in b^{th} bootstrap and j^{th} cycle

$$\sum_{r=1}^k P_{k,n}^{b,j}(r, s) = \frac{k}{n}. \quad (3.2)$$

Proof. (i) Using the telescopian properties of $P_{k,n}^{b,j}(r, s)$, we have

$$\begin{aligned} \sum_{s=1}^n P_{k,n}^{b,j}(r, s) &= \sum_{s=1}^n \left[I_{\frac{s}{n}}(r, k - r + 1) - I_{\frac{s-1}{n}}(r, k - r + 1) \right] \\ &= I_1(r, k - r + 1) - I_0(r, k - r + 1) \\ &= 1 \end{aligned}$$

(ii) Let M_s be the number of times $Y_{(s:n)}$ appears in a bootstrap sample of size k . Since each bootstrap draw picks $Y_{(s:n)}$ with probability $1/n$,

$$M_s \mid \mathcal{D}_S \sim \text{Bin} \left(k, \frac{1}{n} \right).$$

Let

$$\mathbb{I}_{(Y_{(r:k)}=Y_{(s:n)})} = \begin{cases} 1 & \text{if } Y_{(r:k)} = Y_{(s:n)}, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\sum_{r=1}^k \mathbb{I}_{(Y_{(r:k)}=Y_{(s:n)})} = M_s,$$

since each appearance of $Y_{(s:n)}$ in the bootstrap sample occupies exactly one rank position in the bootstrap RSS sample. Taking conditional expectation

$$\sum_{r=1}^k P_{k,n}^{b,j}(r, s) = \mathbb{E}[M_s \mid \mathcal{D}_S] = k \cdot \frac{1}{n} = \frac{k}{n}. \quad (3.3)$$

□

The bagged predictor under the SRB setting is defined as

$$\hat{\mu}_{h,\text{SRB}} = \frac{1}{Bmk} \sum_{b=1}^B \sum_{j=1}^m \sum_{r=1}^k h \left(Y_{(r:k)}^{*(b,j)} \right). \quad (3.4)$$

We now investigate the unbiasedness of the SRB estimator. The following theorem shows that the SRB estimator is unbiased both conditionally on the observed data $\mathcal{D}_S$ and unconditionally with respect to the sampling design.

Theorem 3.2.1 (Unbiasedness of the SRB estimator). Let $\hat{\mu}_{h,\text{SRB}}$ be the SRB estimator defined by (3.4). Then $\hat{\mu}_{h,\text{SRB}}$ is an unbiased estimator of μ_h ;

$$\mathbb{E}_F [\hat{\mu}_{h,\text{SRB}}] = \mu_h.$$

Proof. Taking the expectation

$$\mathbb{E}_{F_n} [\hat{\mu}_{h,\text{SRB}}] = \frac{1}{Bmk} \sum_{b=1}^B \sum_{j=1}^m \sum_{r=1}^k \mathbb{E}_{F_n} [h \left(Y_{(r:k)}^{*(b,j)} \right)].$$

But for fixed indices b, j, r

$$\begin{aligned} \mathbb{E}_{F_n} [h(Y_{(r:k)}^{*(b,j)})] &= \sum_{s=1}^n h(Y_{(s:n)}) \mathbb{P}(Y_{(r:k)}^{*(b,j)} = Y_{(s:n)} \mid \mathcal{D}_S) \\ &= \sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s). \end{aligned}$$

By substituting:

$$\begin{aligned}
\mathbb{E}_{F_n} [\hat{\mu}_{h,\text{SRB}}] &= \frac{1}{Bmk} \sum_{b=1}^B \sum_{j=1}^m \sum_{r=1}^k \sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \\
&= \frac{1}{Bmk} \sum_{b=1}^B \sum_{j=1}^m \sum_{r=1}^k \sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \quad (\text{since independent across } b) \\
&= \frac{1}{mk} \sum_{j=1}^m \sum_{s=1}^n h(Y_{(s:n)}) \left[\sum_{r=1}^k P_{k,n}^{b,j}(r, s) \right] \\
&= \frac{1}{m} \sum_{j=1}^m \left[\frac{1}{n} \sum_{s=1}^n h(Y_{(s:n)}) \right].
\end{aligned}$$

Note that $\frac{1}{n} \sum_{s=1}^n h(Y_{(s:n)}) = \frac{1}{n} \sum_{j=1}^n h(Y_j) = \hat{\mu}_{h,S}$. We get

$$\mathbb{E}_{F_n} [\hat{\mu}_{h,\text{SRB}}] = \frac{1}{m} \sum_{j=1}^m \hat{\mu}_{h,S} = \hat{\mu}_{h,S}.$$

Finally, we have

$$\mathbb{E}_F [\mathbb{E}_{F_n} [\hat{\mu}_{h,\text{SRB}}]] = \mathbb{E}_F [\hat{\mu}_{h,S}] = \mu_h.$$

□

Theorem 3.2.2 (Variance of the SRB mean estimator). Let $\hat{\mu}_{h,\text{SRB}}$ be the bagged SRB estimator $\hat{\mu}_{h,\text{SRB}}$ defined in (3.4).

(i) The conditional variance of $\hat{\mu}_{h,\text{SRB}}$, given $\mathcal{D}_S$ is

$$\begin{aligned}
\text{Var}_{F_n} (\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) &= \frac{1}{Bm^3k^2} \sum_{s=1}^n h(Y_{(s:n)})^2 - \frac{1}{Bm^2k^2} \sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2.
\end{aligned} \tag{3.5}$$

(ii) The unconditional variance of $\hat{\mu}_{h,\text{SRB}}$ is given by

$$\begin{aligned} \mathbb{V}ar_F(\hat{\mu}_{h,\text{SRB}}) &= \frac{1}{Bm^2k} (\sigma_h^2 + \mu_h^2) \\ &\quad - \frac{1}{Bm^2k^2} \sum_{r=1}^k \sum_{s=1}^n \sum_{s'=1}^n P_{k,n}^{b,j}(r, s) P_{k,n}^{b,j}(r, s') \mathbb{E}_F [h(Y_{(s:n)}) h(Y_{(s':n)})] + \frac{\sigma_h^2}{n}, \end{aligned} \quad (3.6)$$

where

$$\mathbb{E}[h(Y_{(s:n)}) h(Y_{(s':n)})] = \begin{cases} \mathbb{V}ar(h(Y_{(s:n)})) + \mathbb{E}_F^2[h(Y_{(s:n)})] = (\sigma_h^2 + \mu_h^2) & s = s', \\ \mathbb{C}ov(h(Y_{(s:n)}), h(Y_{(s':n)})) + \mathbb{E}_F[h(Y_{(s:n)})] \mathbb{E}_F[h(Y_{(s':n)})] & s \neq s'. \end{cases}$$

Proof. (i) Conditional on $\mathcal{D}_S$

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) = \mathbb{V}ar_{F_n} \left(\frac{1}{Bmk} \sum_{b=1}^B \sum_{j=1}^m \sum_{r=1}^k h(Y_{(r:k)}^{*(b,j)}) \mid \mathcal{D}_S \right).$$

As bootstrap samples are independent across $b = 1, \dots, B$ and cycles are independent across $j = 1, \dots, m$, we have

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) = \frac{1}{Bm^2k^2} \sum_{r=1}^k \mathbb{V}ar_{F_n} \left(h(Y_{(r:k)}^{*(b,j)}) \mid \mathcal{D}_S \right).$$

Then, the variance for a fixed rank r is given by

$$\begin{aligned} \mathbb{V}ar_{F_n} \left(h(Y_{(r:k)}^{*(b,j)}) \mid \mathcal{D}_S \right) &= \mathbb{E}_{F_n} \left[h^2(Y_{(r:k)}^{*(b,j)}) \mid \mathcal{D}_S \right] - \mathbb{E}_{F_n}^2 \left[h(Y_{(r:k)}^{*(b,j)}) \mid \mathcal{D}_S \right] \\ &= \sum_{s=1}^n h^2(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) - \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2. \end{aligned}$$

Therefore,

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) = \frac{1}{Bm^2k^2} \sum_{r=1}^k \left[\sum_{s=1}^n h^2(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) - \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 \right].$$

Then using Lemma 3.2.1, we have

$$\sum_{r=1}^k \sum_{s=1}^n h(Y_{(s:n)})^2 P_{k,n}^{b,j}(r, s) = \sum_{s=1}^n h(Y_{(s:n)})^2 \sum_{r=1}^k P_{k,n}^{b,j}(r, s) = \frac{k}{n} \sum_{s=1}^n h(Y_{(s:n)})^2.$$

Finally

$$\begin{aligned} \text{Var}_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) &= \frac{1}{Bm^2k^2} \left[\frac{k}{n} \sum_{s=1}^n h^2(Y_{(s:n)}) - \sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 \right] \\ &= \frac{1}{Bm^3k^2} \sum_{s=1}^n h^2(Y_{(s:n)}) - \frac{1}{Bm^2k^2} \sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2. \end{aligned}$$

(ii) To obtain the unconditional variance $\hat{\mu}_{h,\text{SRB}}$, we use the total law of variance to get

$$\begin{aligned} \text{Var}_F(\hat{\mu}_{h,\text{SRB}}) &= \mathbb{E}_F[\text{Var}_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S)] + \text{Var}_F(\mathbb{E}_{F_n}[\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S]) \\ &= \mathbb{E}_F \left[\frac{1}{Bm^3k^2} \sum_{s=1}^n h^2(Y_{(s:n)}) - \frac{1}{Bm^2k^2} \sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 \right] + \frac{\sigma_h^2}{n} \\ &= \frac{1}{Bm^2k} (\sigma_h^2 + \mu_h^2) - \frac{1}{Bm^2k^2} \sum_{r=1}^k \mathbb{E}_F \left[\left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 \right] + \frac{\sigma_h^2}{n} \\ &= \frac{1}{Bm^2k} (\sigma_h^2 + \mu_h^2) + \frac{\sigma_h^2}{n} \\ &\quad - \frac{1}{Bm^2k^2} \sum_{r=1}^k \sum_{s=1}^n \sum_{s'=1}^n P_{k,n}^{b,j}(r, s) P_{k,n}^{b,j}(r, s') \mathbb{E}_F [h(Y_{(s:n)}) h(Y_{(s':n)})]. \end{aligned}$$

□

Having established the variance properties of the SRB estimator, we now investigate the variance gain achieved through bagging. Specifically, we compare the variance of the bagged predictor based on the ranked-informed SRS training data with that obtained from the conventional SRS and RSS training data.

Proposition 3.2.1 (Conditional variance comparison between SSB and SRB estimators).

Conditionally on the SRS training data $\mathcal{D}_S$, the variance of the SRB estimator satisfies

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) \leq \frac{1}{m} \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S), \quad (3.7)$$

or equivalently,

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) \geq m \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S). \quad (3.8)$$

Proof. From SSB estimator (1.8),

$$\frac{1}{Bm^2k^2} \sum_{s=1}^n h^2(Y_{(s:n)}) = \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) + \frac{1}{Bmk} \hat{\mu}_{h,S}^2.$$

Substituting (3.5),

$$\begin{aligned} & \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) \\ &= \frac{1}{m} \left(\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) + \frac{1}{Bmk} \hat{\mu}_{h,S}^2 \right) - \frac{1}{Bm^2k^2} \sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 \\ &= \frac{1}{m} \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) - \frac{1}{Bm^2k^2} \left[\sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 - k \hat{\mu}_{h,S}^2 \right]. \end{aligned}$$

However, using Lemma 3.2.1, we have

$$\begin{aligned} \hat{\mu}_{h,S} &= \frac{1}{n} \sum_{s=1}^n h(Y_{(s:n)}) \\ &= \frac{1}{k} \sum_{s=1}^n h(Y_{(s:n)}) \sum_{r=1}^k P_{k,n}^{b,j}(r, s) \\ &= \frac{1}{k} \sum_{r=1}^k \sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s). \end{aligned}$$

Finally, using Jensen's inequality, we get

$$\sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 - k \hat{\mu}_{h,S}^2 \geq 0,$$

and this complete the proof. $\square$

Proposition 3.2.1 shows that the SRB estimator achieves at least an m -fold reduction in variance relative to the SSB estimator, where m denotes the cycle size. Consequently, the bootstrap estimators generated by SRB are considerably more stable, resulting in substantially lower variability than conventional bagging.

Proposition 3.2.2 (Conditional variance comparison between RSB and SRB estimators).
Conditionally on the training data,

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) \leq \frac{1}{m}(\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) + \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S)) \quad (3.9)$$

Proof. To show the result, using (2.10)

$$\begin{aligned} \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) &= \frac{m-1}{Bm^2k^2} \sum_{r=1}^k s_{h(r)}^2 \\ &= \frac{1}{Bm^2k^2} \sum_{r=1}^k \sum_{i=1}^m (h(Y_{(r)i}) - \hat{\mu}_{h,(r)})^2 \\ &= \frac{1}{Bm^2k^2} \left[\sum_{r=1}^k \sum_{i=1}^m h^2(Y_{(r)i}) \right] - \frac{1}{Bmk^2} \left[\sum_{r=1}^k \hat{\mu}_{h,(r)}^2 \right]. \end{aligned}$$

By changing the terms,

$$\frac{1}{Bm^2k^2} \left[\sum_{r=1}^k \sum_{i=1}^m h^2(Y_{(r)i}) \right] = \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) + \frac{1}{Bmk^2} \left[\sum_{r=1}^k \hat{\mu}_{h,(r)}^2 \right]. \quad (3.10)$$

Summing over the same set of data points,

$$\sum_{r=1}^k \sum_{i=1}^m h(Y_{(r)i}) = \sum_{s=1}^n h(Y_{(s:n)}) = \sum_{i=1}^n h(Y_i) \text{ and } \sum_{r=1}^k \sum_{i=1}^m h^2(Y_{(r)i}) = \sum_{s=1}^n h^2(Y_{(s:n)}) = \sum_{i=1}^n h^2(Y_i).$$

These result in:

$$\begin{aligned}
\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) &= \frac{1}{Bm^2k^2} \sum_{r=1}^k \left[\sum_{s=1}^n h^2(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) - \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 \right] \\
&= \frac{1}{Bm^3k^2} \sum_{r=1}^k \sum_{i=1}^m h^2(Y_{(r)i}) - \frac{1}{Bm^2k^2} \sum_{r=1}^k \left(\sum_{i=1}^m h(Y_{(r)i}) P_{k,n}^{b,j}(r, s) \right)^2 \\
&= \frac{1}{Bm^3k^2} \sum_{r=1}^k \sum_{i=1}^m h^2(Y_{(r)i}) - \frac{1}{Bm^2k^2} \sum_{r=1}^k \left(\sum_{i=1}^m h(Y_{(r)i}) \cdot \frac{k}{n} \right)^2 \\
&= \frac{1}{Bm^3k^2} \sum_{r=1}^k \sum_{i=1}^m h^2(Y_{(r)i}) - \frac{1}{Bm^2k^2} \sum_{r=1}^k \left(\sum_{i=1}^m h(Y_{(r)i}) \right)^2 \quad (n = mk) \\
&= \frac{1}{m} \left(\frac{1}{Bm^2k^2} \sum_{r=1}^k \sum_{i=1}^m h^2(Y_{(r)i}) \right) - \frac{1}{Bm^2k^2} \sum_{r=1}^k \left(\sum_{i=1}^m h(Y_{(r)i}) \right)^2.
\end{aligned}$$

By substitution from (3.10), we get

$$\begin{aligned}
&\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) \\
&= \frac{1}{m} \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) + \frac{1}{Bm^2k^2} \left[\sum_{r=1}^k \hat{\mu}_{h,(r)}^2 - \sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 \right].
\end{aligned} \tag{3.11}$$

Let

$$Bm^2k \cdot \Delta_{1,2} = \frac{1}{k} \sum_{r=1}^k \hat{\mu}_{h,(r)}^2 - \frac{1}{k} \sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2. \tag{3.12}$$

Using Jensen's inequality one again,

$$\begin{aligned}
\frac{1}{k} \sum_{r=1}^k \hat{\mu}_{h,(r)}^2 &= \frac{1}{k} \sum_{r=1}^k \left(\frac{1}{m} \sum_{j=1}^m h(Y_{(r)j}) \right)^2 \leq \frac{1}{k} \sum_{r=1}^k \frac{1}{m} \sum_{j=1}^m h^2(Y_{(r)j}) \\
&= \frac{1}{mk} \sum_{r=1}^k \sum_{j=1}^m h^2(Y_{(r)j}) = \frac{1}{n} \sum_{i=1}^n h^2(Y_i).
\end{aligned}$$

Also,

$$\begin{aligned} \frac{1}{k} \sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 &\geq \left(\frac{1}{k} \sum_{r=1}^k \sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 \\ &= \left(\frac{1}{n} \sum_{s=1}^n h(Y_{(s:n)}) \right)^2 = \hat{\mu}_{h,S}^2. \end{aligned}$$

So, from (3.12),

$$Bm^2k \cdot \Delta_{1,2} \geq \frac{1}{n} \sum_{i=1}^n h^2(Y_i) - \hat{\mu}_{h,S}^2 = \frac{n-1}{n} s_{n,h}^2.$$

Therefore,

$$\begin{aligned} \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) &\leq \frac{1}{m} \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}}) + \frac{n-1}{Bn^2m} s_{n,h}^2 \\ &= \frac{1}{m} (\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) + \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S)). \end{aligned}$$

□

Remark 3.2.1. From Proposition 2.6.1, we have

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) \geq \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R).$$

Consequently, Proposition 3.2.2 yields the bound

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) \geq \frac{m}{2} \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S).$$

However, Proposition 3.2.1 provides a sharper result,

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) \geq m \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S).$$

The following corollary combines the conditional variance bounds established in Propositions 3.2.1 and 3.2.2, thereby summarizing the variance relationships among the SSB, RSB, and SRB estimators.

Corollary 3.2.1 (Conditional variance comparison between SSB, RSB and SRB estimators.). The three bootstrap estimators satisfy

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) \geq \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R), \quad (3.13)$$

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S) \geq m \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S), \quad (3.14)$$

$$\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) \leq \frac{1}{m} (\mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) + \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S)) \quad (3.15)$$

$$\leq \frac{2}{m} \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}}) \quad (3.16)$$

Consequently,

$$\min \{ \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R), m \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) \} \leq \mathbb{V}ar_{F_n}(\hat{\mu}_{h,\text{SSB}} \mid \mathcal{D}_S). \quad (3.17)$$

Inequalities (3.13) and (3.14) establish that SSB has the largest conditional variance among the competing methods, confirming that both RSB and SRB provide improvements over conventional SRS bagging. In addition, the conditional variance of the SRB estimator is guaranteed to be at least a factor of m smaller than that of SSB, where m denotes the cycle size.

Inequality (3.15) provides an important link between SRB and RSB. Unlike RSB, which uses rank-based training data using the RSS design, SRB constructs rank groups using only the available SRS training data. Despite relying on this weaker source of ranking information, the variance of SRB remains bounded by $1/m$ times the combined conditional variance of RSB and SSB. This result indicates that SRB captures a substantial portion of the variance-reduction benefits associated with RSS-based resampling, while avoiding the practical difficulties of working with RSS training data.

From a practical perspective, this finding is particularly important because the availability of accurate ranking information is often the main obstacle to implementing RSS procedures. The proposed SRB framework demonstrates that meaningful variance reductions can still be achieved using only an initial SRS training sample. Consequently,

SRB provides a balance between the simplicity of SSB and the efficiency of RSB, offering considerable gains in estimator stability without requiring a fully ranked-set sampling design.

Proposition 3.2.3 (Useful deterministic bounds between RSB and SRB). The following inequalities hold,

$$(i) \quad k\hat{\mu}_{h,S}^2 \leq \sum_{r=1}^k \hat{\mu}_{h,(r)}^2 \leq \frac{1}{m} \sum_{s=1}^n h^2(Y_{(s:n)}), \quad (3.18)$$

$$(ii) \quad \frac{k^2}{m} \hat{\mu}_{h,S}^2 \leq \sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 \leq \frac{1}{m} \sum_{s=1}^n h^2(Y_{(s:n)}). \quad (3.19)$$

Proof. To show (i) one can easily see that

$$\begin{aligned} \sum_{r=1}^k \hat{\mu}_{h,(r)}^2 &= \sum_{r=1}^k \left(\frac{1}{m} \sum_{j=1}^m h(Y_{(r)j}) \right)^2 \\ &\leq \frac{1}{m} \sum_{r=1}^k \sum_{j=1}^m h^2(Y_{(r)j}) \quad (\text{by Jensen's inequality}) \\ &= \frac{1}{m} \sum_{s=1}^n h^2(Y_{(s:n)}). \end{aligned}$$

Also,

$$\begin{aligned} \frac{1}{k} \sum_{r=1}^k \hat{\mu}_{h,(r)}^2 &\geq \left(\frac{1}{k} \sum_{r=1}^k \hat{\mu}_{h,(r)} \right)^2 \quad (\text{by Jensen's inequality}) \\ &= \left(\frac{1}{mk} \sum_{r=1}^k \sum_{j=1}^m h(Y_{(r)j}) \right)^2 \\ &= \left(\frac{1}{mk} \sum_{s=1}^n h(Y_{(s:n)}) \right)^2 \\ &= \hat{\mu}_{h,S}^2. \end{aligned}$$

Thus we have, $k\hat{\mu}_{h,S}^2 \leq \sum_{r=1}^k \hat{\mu}_{h,(r)}^2 \leq \frac{1}{m} \sum_{s=1}^n h^2(Y_{(s:n)})$.

(ii) Using Jensen's inequality we have

$$\begin{aligned} \sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 &\leq \sum_{r=1}^k \sum_{s=1}^n h^2(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \\ &= \sum_{s=1}^n h^2(Y_{(s:n)}) \sum_{r=1}^k P_{k,n}^{b,j}(r, s) \\ &= \frac{1}{m} \sum_{s=1}^n h^2(Y_{(s:n)}). \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{1}{m} \sum_{r=1}^k \left(\sum_{s=1}^n h(Y_{(s:n)}) P_{k,n}^{b,j}(r, s) \right)^2 &\geq \left(\frac{1}{m} \sum_{s=1}^n h(Y_{(s:n)}) \sum_{r=1}^k P_{k,n}^{b,j}(r, s) \right)^2 \\ &= \frac{1}{m^4} \left(\sum_{s=1}^n h(Y_{(s:n)}) \right)^2 \\ &= \frac{k^2}{m^2} \hat{\mu}_{h,S}^2, \end{aligned}$$

which completes the proof. □

Remark 3.2.2. The following deterministic variance bounds hold for the SRB and RSB procedures.

$$\begin{aligned} \text{Var}_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) &\leq \frac{1}{m} \text{Var}_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) + \frac{1}{Bm^2k^2} \left[\frac{1}{m} \sum_{s=1}^n h^2(Y_{(s:n)}) - \frac{k^2}{m} \hat{\mu}_{h,S}^2 \right]. \\ \text{Var}_{F_n}(\hat{\mu}_{h,\text{SRB}} \mid \mathcal{D}_S) &\geq \frac{1}{m} \text{Var}_{F_n}(\hat{\mu}_{h,\text{RSB}} \mid \mathcal{D}_R) + \frac{1}{Bm^2k^2} \left[km^2 \hat{\mu}_{h,S}^2 - \frac{1}{m} \sum_{s=1}^n h^2(Y_{(s:n)}) \right]. \end{aligned}$$

Proposition 3.2.4 extends the conditional variance comparisons to the population level by establishing corresponding inequalities for the unconditional variances.

Proposition 3.2.4 (Variance comparison between SSB, RSB and SRB estimators). For the SSB, RSB, and SRB mean estimators with B bootstrap replicates, set size k , and cycle size m ,

(i)

$$\mathbb{V}ar_F(\hat{\mu}_{h,\text{SSB}}) \geq \mathbb{V}ar_F(\hat{\mu}_{h,\text{RSB}}). \quad (3.20)$$

(ii)

$$\mathbb{V}ar_F(\hat{\mu}_{h,\text{SSB}}) \geq \mathbb{V}ar_F(\hat{\mu}_{h,\text{SRB}}). \quad (3.21)$$

Proof. (i) and (ii) follow directly from Proposition 3.2.1, Corollary 3.2.1, Proposition 1.5.1, Lemma 2.6.1, Theorem 3.2.1, and the law of total expectation. $\square$

Both RSB and SRB possess smaller unconditional variances than the SSB estimator. All three SSB, RSB, and SRB estimators are unbiased for estimating the population means. Therefore, the reductions observed in the conditional bootstrap variances are directly attributable to reducing overall sampling variability.

The proposition establishes that the variance reduction achieved by RSB and SRB extends beyond the bootstrap resampling stage and reflects an actual improvement in the estimator's efficiency. RSB attains this gain through an RSS design; SRB achieves a similar advantage using only information from the original SRS training sample. This makes SRB particularly attractive in applications where direct ranking information is partially available or costly to obtain, while still providing a substantial reduction in overall estimator variability relative to standard bagging.

3.3 Regression Estimation with SRB

In Section 3.2, we established the unbiasedness and derived the unconditional and conditional variance expressions for the estimation of general mean-type functionals under the SRB framework.

We now extend this framework to the estimation of the regression function, $m(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$. This extension closely follows from the development in Chapter 2, particularly Section 2.7.1, where regression estimation under RSS was formulated using the linear smoother representation. Although SRB originates from an initial SRS, the rank-informed sample constructed in Section 3.2 possesses the same structural form as an RSS sample. Consequently, the regression framework developed for RSS can be applied directly to the SRB setting. The only modification required is to replace the original RSS training sample with the bootstrap samples generated from the SRS data.

3.3.1 Linear Smoother SRB Regression Estimator

We consider an RSS bootstrap applied to SRS training data, mimicking an RSS design within each bootstrap replicate. Here, the ranking is performed on Y , while the regression structure enters through the associated pseudo-responses $Z_i(\mathbf{x})$.

Definition 3.3.1 (SRB Regression Estimator). Suppose $\mathcal{D}_S = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ is an SRS sample with $n = mk$, and let $Z_i(\mathbf{x})$ be as in (1.10). For each bootstrap replicate $b = 1, \dots, B$:

- (a) For each cycle $j = 1, \dots, m$, draw a set of k candidate indices $I_{1j}^{*b}, \dots, I_{kj}^{*b}$ i.i.d. from $\{1, \dots, n\}$ with replacement, corresponding to candidate pairs

$$\left\{ (Y_{I_{1j}^{*b}}, Z_{I_{1j}^{*b}}(\mathbf{x})), \dots, (Y_{I_{kj}^{*b}}, Z_{I_{kj}^{*b}}(\mathbf{x})) \right\}.$$

- (b) Rank the k candidates within cycle j according to their Y -values, assuming perfect ranking. This yields the order statistics $Y_{[1]j}^{*b} \leq \dots \leq Y_{[k]j}^{*b}$, together with the associated pseudo-responses $Z_{[1]j}^{*b}(\mathbf{x}), \dots, Z_{[k]j}^{*b}(\mathbf{x})$, obtained by permuting the $Z_{I_{kj}^{*b}}(\mathbf{x})$'s according to the order of the $Y_{I_{kj}^{*b}}$'s.
- (c) Define the single-replicate SRB estimator at $\mathbf{x}$ by

$$\hat{m}_{\text{SRB}}^{*b}(\mathbf{x}) = \frac{1}{mk} \sum_{j=1}^m \sum_{r=1}^k Z_{[r]j}^{*b}(\mathbf{x}). \quad (3.22)$$

The B -replicate bagged SRB estimator is

$$\hat{m}_{\text{SRB}}^{(B)}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{m}_{\text{SRB}}^{*b}(\mathbf{x}). \quad (3.23)$$

If we view $\{Z_i(\mathbf{x})\}_{i=1}^n$ as the observations and rank them according to the associated Y_i 's, then (3.22) is structurally identical to the SRB mean estimator in Section 3.2, with $h(Y_i)$ replaced by the pseudo-response $Z_i(\mathbf{x})$.

As in the RSB framework, many regression estimators admit a linear smoother representation, allowing the fitted regression function to be expressed as an average of rank-specific pseudo-responses. Therefore, the theoretical developments established for mean-type functionals naturally carry over to the regression setting by replacing the original observations with the corresponding pseudo-responses. In particular, the unbiasedness properties, conditional and unconditional variance expressions, and variance-comparison results derived for the SRB estimators in Section 3.2 remain valid for regression estimators that admit a linear smoother representation.

Therefore, the variance reduction advantages established for SRB in the mean-functional setting extend directly to the estimation and prediction of regression functions. An important feature of this result is that the variance reduction gains are achieved

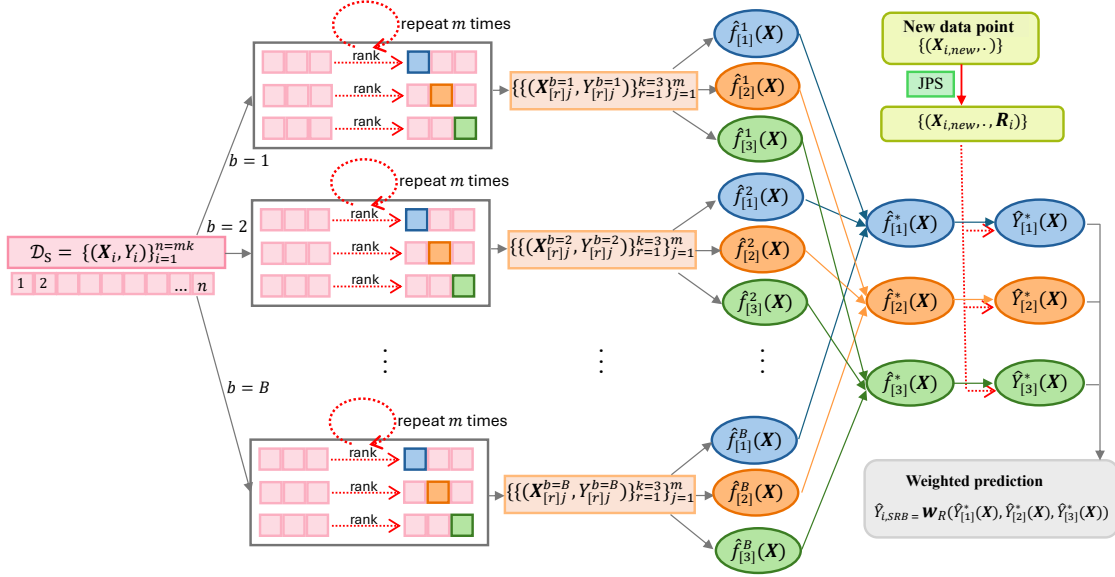


Figure 3.1: SRB resampling and prediction procedure for SRS data with set size 3 under paired bootstrapping.

without requiring the population-level ranking information assumed in the RSS framework. Instead, SRB uses ranking information extracted from available SRS training data to construct rank-informed samples, capturing the benefits of RSS-based bagging while retaining the practical simplicity and accessibility of an SRS design. This provides a practically attractive alternative to RSS in situations where reliable population-level ranking information is unavailable.

In this thesis, we consider two bootstrapping approaches within the proposed SRB framework, namely paired and residual bootstrapping. For a better understanding of the overall SRB methodology, Figure 3.1 illustrates the complete workflow for SRB resampling and prediction with rank $k = 3$ under paired bootstrapping. Starting from the SRS training sample incorporating ranking information, independent subsets of size k are repeatedly drawn from the SRS data. Then, within each subset, the observations are ranked using auxiliary information or expert judgment. From each subset, only the observation corresponding to the k -th rank is retained. This ranking-and-selection process is repeated for m cycles, yielding m observations at each rank level. Consequently,

each bootstrap RSS dataset has a total sample size of mk . For each bootstrap sample, we fit a single regression model and use it to obtain k rank-specific predictions at the target point $\mathbf{x}$. We then aggregate these predictions across the B bootstrap replicates to form bagged rank-specific predictors. For a new SRS test observation, the JPS procedure estimates its rank distribution, and we then combine the rank-specific predictions through a weighted aggregation scheme to obtain the final SRB prediction. A detailed overview of SRB resampling and prediction for both paired and residual bootstrapping is presented in Chapter 4.

3.4 Summary and Conclusion

This chapter introduced the proposed SRB framework as a practical alternative to the RSB procedure. Unlike RSB, which assumes that RSS are directly available from the population, the proposed SRB framework constructs RSS from an initial SRS training sample by incorporating ranking information into each bootstrap replicate, thereby preserving the practical simplicity of SRS while leveraging the efficiency gains of RSS.

Theoretical properties of the SRB estimator were established for general mean-type functionals. It was shown that the proposed estimator is unbiased, and expressions were derived for both the conditional and unconditional variances. These results were further used to compare the variance of the SRB with SSB and the RSB procedure. The theoretical analysis demonstrated that SRB consistently achieves smaller variances than SSB while closely approximating the variance reduction obtained by RSB.

The chapter then extended the SRB framework to regression estimation, focusing on linear smoother regression estimators. By representing regression estimators through pseudo-responses, the theoretical developments obtained for mean-type functionals were shown to carry over directly to regression estimation. The unbiasedness, variance expressions, and variance comparison results established for SRB remain valid for a broad class of regression estimators admitting a linear smoother representation.

Finally, the SRB prediction methodology was presented for both paired and residual bootstrapping settings. A visual illustration of the paired bootstrapping procedure was provided to clarify the complete resampling, model fitting, and prediction workflow, while the corresponding algorithmic implementations for both bootstrapping approaches are presented in the numerical study. Overall, the proposed SRB framework provides a computationally practical and theoretically justified approach that combines the accessibility of SRS with the variance reduction benefits of RSS-based bagging, making it particularly attractive for predictive modeling when direct population-level ranking information is unavailable.

Chapter 4

Numerical Study

In this chapter, we conduct simulation studies and a real-data analysis to evaluate the performance of RSS-based bagged regression predictors under both linear and nonlinear regression models. We first study bagging linear and nonlinear models using RSS with different ranking structures. To this end, we consider single-covariate models in which RSS is performed using a single ranking variable and extend them to models with multiple ranking variables in the simulation study. We then further extend the simulation study to multiple-covariate models with multiple ranking variables to assess the performance of the proposed methods under more complex modeling scenarios. Then, we apply the proposed RSB and SRB methods to multiple linear regression models with multiple ranking variables, using a real data set to predict body fat in men and women.

4.1 Data Generation

This chapter begins by describing the simulation setting and the procedure used to generate RSS. The RSS samples are generated under a balanced design. We first describe the construction of RSS using a single ranking variable and then extend the procedure to multiple ranking variables by incorporating the weighting scheme in Equation (2.24). Next, we present the bootstrap strategies applied directly to RSS data.

Since the proposed SRB approach relies on constructing RSS from an SRS through a conversion process, we subsequently present the SRS-RSS conversion procedure and its corresponding bootstrap methodology.

4.1.1 Single Ranker RSS Construction

Let the population consist of N units, each characterized by a predictor vector $\mathbf{X}_i$, a response Y_i , and a ranking variable R_i , $i = 1, \dots, N$. Note that in practice R_i needs to be known only for the sampling units selected in each set of the RSS design and not necessarily for the whole population. An RSS sample of total size $n = mk$ is obtained by repeating the following cycle m times:

- (i) Draw k SRS samples of size k without replacement from the population.
- (ii) Rank the k sampled units in each set in ascending order using the ranking variable R .
- (iii) From the r -th ordered set, retain only the unit with rank r , that is, the r -th order statistic with respect to R , for $r = 1, \dots, k$.
- (iv) Record the predictor vector $\mathbf{X}$, response Y , and ranking variable R of the selected observations.

Repeating steps (i) to (iv) m times results in a balanced RSS of size $n = mk$.

4.1.2 Weighted RSS Construction for Multiple Ranks

In many practical applications, multiple auxiliary ranking variables are available. Let $R_1, \dots, R_L$ denote $L > 1$ such ranking variables. Rather than using only a single ranking variable to construct the RSS sample, one can combine the information from all available ranking variables to produce an RSS. In this section, we propose a weighted

RSS construction procedure that integrates multiple ranking variables into the sample selection process.

For each cycle, an SRS of size k is first selected from the population. Let $r_{i\ell}$ denote the rank assigned to unit i according to the ℓ -th ranking variable R_ℓ , $\ell = 1, \dots, L$ within the selected set. Let $\mathbf{w} = (w_1, \dots, w_L)^\top$ be a vector of weights as in (2.24) satisfying $\sum_{\ell=1}^L w_\ell = 1$. Then the weighted rank score for unit i is computed as

$$M_i = \sum_{\ell=1}^L w_\ell r_{i\ell}, \quad i = 1, \dots, k. \quad (4.1)$$

Finally, order the sampled units by their weighted rank scores. As noted above, one can use the weighted rank score M_i instead of the single rank R_i to obtain the weighted RSS sample considering ℓ different rankers. Repeating this procedure over m cycles yields a weighted RSS of size $n = mk$. The complete procedure is summarized in Algorithm 1.

Remark 4.1.1. The single-rank construction is a special case of the weighted RSS construction corresponding to $L = 1$, in which the rank reduces to r_i , and the associated rank weight is $w_1 = 1$. Therefore, Algorithm 1 can be easily applied directly to the single-rank RSS construction.

Algorithm 1 Weighted RSS construction (multiple ranks)

Require: Population $\mathcal{P} = \{(\mathbf{X}_i, Y_i, R_{i1}, \dots, R_{iL})\}_{i=1}^N$; set size k ; number of cycles m ; weights $\mathbf{w} = (w_1, \dots, w_L)^\top$, where $\sum_{\ell=1}^L w_\ell = 1$; seed s .

Ensure: Weighted RSS sample $\mathcal{D}_{wR}$ of size $n = mk$.

```
1: for  $j = 1, \dots, m$  do
2:   for  $r = 1, \dots, k$  do
3:     Draw a simple random sample of size  $k$  from  $\mathcal{P}$ .
4:     for  $\ell = 1, \dots, L$  do
5:       Rank the sampled units in set  $r$  according to ranking variable  $R_\ell$ .
6:       Let  $r_{i\ell}$  denote the rank of unit  $i$  according to  $R_\ell$ .
7:     end for
8:     for each sampled unit  $i = 1, \dots, k$  do
9:       Compute the weighted rank score
```

$$M_i = \sum_{\ell=1}^L w_\ell r_{i\ell}.$$

```
10:    end for
11:    Sort the sampled units in set  $r$  in ascending order of  $M_i$ .
12:    Select the unit having the  $r$ -th smallest weighted rank score.
13:    Store its original predictor vector  $\mathbf{X}_{[r]j}$ , response  $Y_{[r]j}$ , and ranking variables.
14:  end for
15: end for
16: return  $\mathcal{D}_{wR} = \{(\mathbf{X}_{[r]j}, Y_{[r]j}) : r = 1, \dots, k, j = 1, \dots, m\}$ .
```

4.2 Bootstrapping on RSS

Once RSS (or weighted RSS) training data is constructed, bootstrap samples are generated by resampling independently within each rank stratum $r \in \{1, \dots, k\}$ using the paired or residual bootstrap. This stratified resampling preserves the balanced RSS structure across sampling cycles and maintains the order-statistic interpretation of each rank stratum throughout the bootstrap replications.

4.2.1 RSS Paired Bootstrapping

The paired bootstrap of RSS data resamples the observed predictor-response pairs, $(\mathbf{X}_{[r]j}, Y_{[r]j})$, for $r = 1, \dots, k$ and $j = 1, \dots, m$, with replacement within each rank

Algorithm 2 RSS paired bootstrapping

Require: RSS data $\mathcal{D}_R$; number of bootstrap replications B ; seed s ;

Ensure: Bootstrap replicates $\{\mathcal{D}_R^{*(b)}\}_{b=1}^B$.

- 1: Partition $\mathcal{D}_R$ into rank strata: $\mathcal{D}_r \leftarrow \{(m, \mathbf{X}_{[r]j}, Y_{[r]j})\}_{r=1, j=1}^{k, m}$
- 2: **for** $b = 1$ **to** B **do** ▷ Parallelize over b
- 3: **for** $r = 1$ **to** k **do**
- 4: Draw m pairs with replacement from the stratum $\mathcal{D}_r$ using seed $s + b$:

$$\mathcal{D}_r^{*(b)} \leftarrow \{SRS\}_{\text{wr}}(\mathcal{D}_r, m).$$

- 5: **end for**
 - 6: Combine rank strata to create a bootstrap replicate: $\{\mathcal{D}_R^{*(b)}\} \leftarrow \bigcup_{r=1}^k \mathcal{D}_r^{*(b)}$.
 - 7: **end for**
 - 8: **return** $\{\mathcal{D}_R^{*(b)}\}_{b=1}^B$.
-

stratum. Jointly resampling the predictor vectors and responses preserves the dependence structure between $\mathbf{X}$ and Y and imposes no parametric assumptions on the underlying error distribution. The complete RSS paired bootstrap procedure is described in Algorithm 2, while a graphical illustration of the resampling scheme is provided in Section 2.7.2.

4.2.2 RSS Residual Bootstrapping

The residual bootstrap keeps the predictor values $\mathbf{X}_{[r]j}$ fixed, while resampling residuals with replacement within each rank stratum from a rank-informed fitted model. This model accounts for the order-statistic structure of the RSS through either the WLS approach described in Section 2.8.1 or the LMM approach described in Section 2.8.2, depending on the estimation method employed. The residual bootstrap procedure is presented in two main steps.

Step 1: Rank-informed model fitting: A single regression model is fitted to the $\mathcal{D}_R$ data, yielding a coefficient matrix $\beta \in \mathbb{R}^p$, and a rank-specific correction term $\delta_{[r]}$

for each stratum r , defined as

$$\boldsymbol{\delta}_{[r]} = \begin{cases} \hat{\boldsymbol{\alpha}}_{[r]} & \text{bias-correction (WLS),} \\ \hat{\boldsymbol{u}}_{[r]} & \text{random effect (LMM),} \end{cases} \quad (4.2)$$

where $\hat{\boldsymbol{\alpha}}_{[r]}$ is an estimated bias-correction for the r -th order statistic and $\hat{\boldsymbol{u}}_{[r]}$ is a predicted random effect for stratum r . The fitted value and residual for observation (r, j) are then

$$\hat{\mathbf{Y}}_{[r]j} = \mathbf{X}_{[r]j}^\top \hat{\boldsymbol{\beta}} + \boldsymbol{\delta}_{[r]}, \quad \hat{\mathbf{e}}_{[r]j} = \mathbf{Y}_{[r]j} - \hat{\mathbf{Y}}_{[r]j}. \quad (4.3)$$

These will be used to perform the residual bootstrap.

Step 2: Bootstrap resampling of residuals: For each stratum k , resample the m residuals $\{\hat{\mathbf{e}}_{[r]j}\}_{j=1}^m$ with replacement to form bootstrap responses, while keeping $\hat{\mathbf{Y}}_{[r]j}$ fixed for each resampling cycle. The Algorithm 3 describes the procedure for generating RSS residual bootstraps.

Figure 4.1 illustrates the complete RSB prediction procedure under residual bootstrapping for set size k . Starting from the RSS training sample, a regression model $\hat{f}_R(\mathbf{X})$, is fitted using either the WLS or the LMM approach. The corresponding residuals, $\hat{\mathbf{e}}_{[r]j}$, are then computed as described in Step 1, and residual bootstrap samples are generated within each rank stratum following Step 2. Following Algorithm 3, B RSS bootstrap replicates are constructed. For each bootstrap sample, we fit a single regression model and use it to obtain k rank-specific predictions at the target point $\mathbf{x}$. We then aggregate these predictions across the B bootstrap replicates to form bagged rank-specific predictors. For a new SRS test observation, the JPS procedure estimates its rank distribution, and we then combine the rank-specific predictions through a weighted aggregation scheme to obtain the final RSB prediction.

Algorithm 3 RSS residual bootstrapping

Require: RSS data $\mathcal{D}_R$; design matrix β with method $\in \{\text{WLS}, \text{LMM}\}$; number of bootstrap replications B ; seed s .

Ensure: Bootstrap replicates $\{\mathcal{D}_R^{*(b)}\}_{b=1}^B$.

- 1: Fit model on $\mathcal{D}_R$ to obtain β , and $\{\delta_{[r]}\}_{r=1}^k$ as in (4.2).
- 2: **for** $r = 1$ **to** k **do**
- 3: Compute fitted values and residuals for all m :

$$\hat{\mathbf{Y}}_{[r]j} \leftarrow \mathbb{X}_{[r]j}^\top \hat{\beta} + \delta_{[r]}, \quad \hat{\mathbf{e}}_{[r]j} \leftarrow \mathbf{Y}_{[r]j} - \hat{\mathbf{Y}}_{[r]j}.$$

- 4: Store $\mathbf{a}_{[r]} \leftarrow [\hat{\mathbf{e}}_{[r]1}, \dots, \hat{\mathbf{e}}_{[r]m}]^\top$.
- 5: **end for**
- 6: **for** $b = 1$ **to** B **do** ▷ Parallelize over b
- 7: Initialize bootstrap data frame as copy of $\mathcal{D}_R$.
- 8: **for** $r = 1$ **to** k **do**
- 9: Resample m residuals with replacement using seed $s + b$:

$$\mathbf{a}_{[r]}^{*(b)} \leftarrow \text{SRS}_{\text{wr}}(\mathbf{a}_{[r]}, m).$$

- 10: Construct bootstrap responses for stratum r :

$$\mathbf{Y}_{[r]j}^{*(b)} \leftarrow \hat{\mathbf{Y}}_{[r]j} + \mathbf{a}_{[r]j}^{*(b)}, \quad j = 1, \dots, m.$$

- 11: Replace $\mathbf{Y}_{[r]j}$ with $\mathbf{Y}_{[r]j}^{*(b)}$ in the bootstrap data frame.
 - 12: **end for**
 - 13: Compute bootstrap replicate: $\{\mathcal{D}_R^{*(b)}\}$.
 - 14: **end for**
 - 15: **return** $\{\mathcal{D}_R^{*(b)}\}_{b=1}^B$.
-

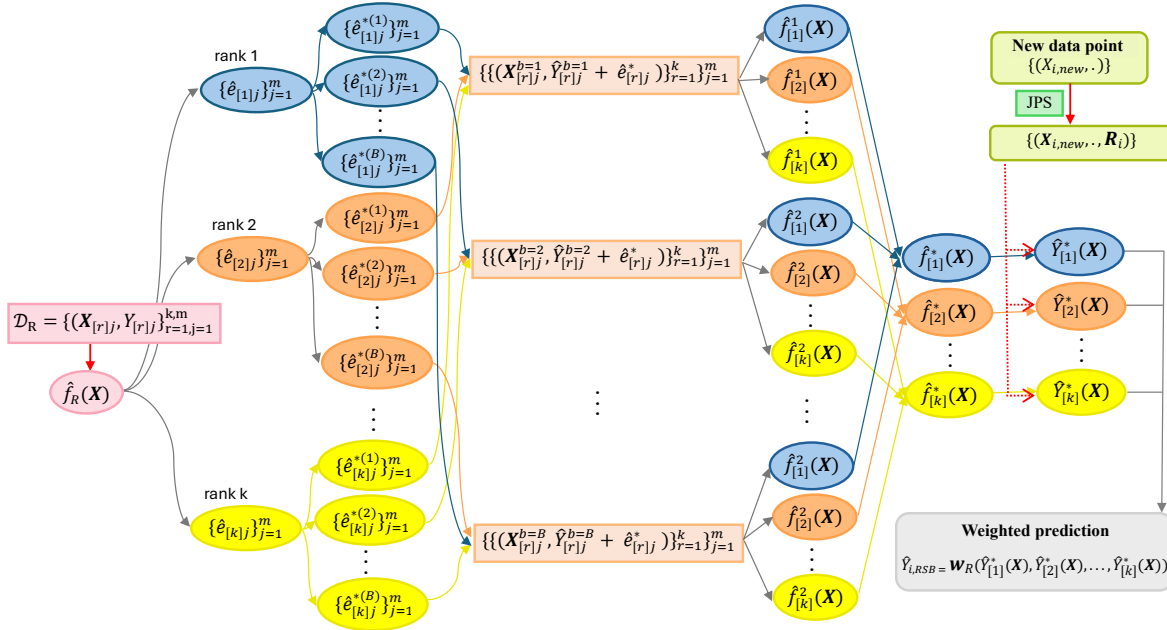


Figure 4.1: RSB resampling and prediction procedure for RSS data with set size k under residual bootstrapping.

4.3 RSS Test Data Generation

To evaluate the predictive performance of the SSB, RSB, and SRB methods, draw an independent SRS test sample, distinct from the training sample, from the population, and assign ranks to it using a judgment post-stratification (JPS) approach (MacEachern et al., 2004).

The idea of the JPS procedure is similar to the RSS design described earlier. First, an SRS test sample of size n is selected, and the response variable is measured for each observation. For each test observation i , $i = 1, \dots, n$, a comparison set of $k - 1$ auxiliary observations is randomly selected from the population without replacement; the test observation and the comparison units form a set of size k . Using the available ranking variables, together with judgment or auxiliary information, the units within the set are ranked from 1 to k . Unlike RSS, where the selected unit is predetermined by its rank, JPS assigns a rank to the observed test unit after measurements are obtained.

Algorithm 4 RSS test data generation

Require: Population $\mathcal{D}$ of size N with $\{R_\ell\}_{\ell=1}^L$ with weights $\{w_\ell\}_{\ell=1}^L$; SRS test data $\mathcal{D}_S^{\text{test}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$; number of replications B ; set size k ; seed s .

Ensure: Estimated rank probabilities $\hat{p}_{ir}$, $i = 1, \dots, n$; $r = 1, \dots, k$.

- 1: Initialize $p_{ir}^{(b)} \leftarrow 0$ for all $b = 1, \dots, B$, $i = 1, \dots, n$, $r = 1, \dots, k$.
- 2: **for** $b = 1, \dots, B$ **do** $\triangleright$ parallelizable over b
- 3: Initialize random number generator with seed $s + b$
- 4: **for** $i = 1, \dots, n$ **do**
- 5: Draw $k - 1$ comparison units $S_i^{(b)}$ without replacement from $\mathcal{D}$.
- 6: **for** each ranking variable R_ℓ , $\ell = 1, \dots, L$, with weight w_ℓ **do**.
- 7: $R_i \leftarrow$ value of R_ℓ for test observation i .
- 8: $\{R_j\}_{j \in S_i^{(b)}} \leftarrow$ values of R_ℓ for the sampled comparison units.
- 9: $C \leftarrow \text{sort}(\{R_i\} \cup \{R_j\}_{j \in S_i^{(b)}})$.
- 10: $r_i^{(\ell, b)} \leftarrow$ position of R_i in C .
- 11: **if** $r_i^{(\ell)}$ exists **then**
- 12: $p_{i, r_i^{(\ell, b)}}^{(b)} \leftarrow p_{i, r_i^{(\ell, b)}}^{(b)} + w_\ell$
- 13: **end if**
- 14: **end for**
- 15: **end for**
- 16: **end for**
- 17: $\hat{p}_{ir} \leftarrow \frac{1}{B} \sum_{b=1}^B p_{ir}^{(b)}$, for all $i = 1, \dots, n$, $r = 1, \dots, k$
- 18: **for** $i = 1, \dots, n$ **do**
- 19: **Maximum rank approach:** $\hat{r}_i \leftarrow \arg \max_r \hat{p}_{ir}$
- 20: **Full probability distribution approach:** retain $(\hat{p}_{i1}, \dots, \hat{p}_{ik})$
- 21: **end for**
- 22: **return** $\{\hat{r}_i\}_{i=1}^n, \{\hat{p}_{i1}, \dots, \hat{p}_{ik}\}_{i=1}^n$

The probability that test observation i belongs to rank stratum r , denoted by $\hat{p}_{ir}$ for $r = 1, \dots, k$, is estimated by repeatedly applying the JPS procedure described above over B replications. The estimated rank probabilities are then obtained by aggregating the rank assignments across these replications. Throughout this study, the JPS procedure is implemented using $B = 10$ replications. Two rank representations are considered for prediction.

1. **Full probability distribution approach:** For each test observation i , the

estimated rank probability is

$$\hat{\mathbf{p}}_i = (\hat{p}_{i1}, \dots, \hat{p}_{ik}),$$

where $\hat{p}_{ir} = \frac{1}{B} \sum_{b=1}^B p_{ir}^{(b)}$, $r = 1, \dots, k$. When multiple ranking variables are available, the bootstrap-specific rank probability contribution is given by

$$p_{ir}^{(b)} = \sum_{\ell=1}^L w_{\ell} \mathbb{1}\left\{r = r_i^{(\ell,b)}\right\},$$

where $r_i^{(\ell,b)}$ denotes the rank assigned to test observation i using the ℓ -th ranking variable in bootstrap replication b , and w_{ℓ} is the weight assigned to ranking variable R_{ℓ} , satisfying $\sum_{\ell=1}^L w_{\ell} = 1$. Consequently, $\sum_{r=1}^k p_{ir}^{(b)} = 1$, and averaging these probability vectors over the B replications yields the estimated rank probability vector $\hat{\mathbf{p}}_i$. The estimated probability vector is then retained and used directly in the prediction procedure, allowing each test observation to contribute to every rank stratum according to its estimated probabilities, thereby accounting for uncertainty in the rank assignment.

2. **Maximum rank approach:** As a simplified alternative, the estimated probability vector $\hat{\mathbf{p}}_i$ is first computed as described above. The test observation is then assigned to the rank stratum r corresponding to the largest estimated probability, where $r = 1, \dots, k$:

$$\hat{r}_i = \arg \max_r \hat{p}_{ir}.$$

Thus, the assigned rank is determined from the aggregated weighted rank probabilities across all bootstrap replications rather than from the most probable rank within an individual bootstrap replication.

The complete procedure for generating the RSS test data using the JPS approach is presented in Algorithm 4.

Algorithm 5 SRS-RSS construction with paired bootstrapping

Require: SRS data $\mathcal{D}_S$, number of bootstrap replications B , set size k , cycles m , seed s

Ensure: Bootstrap replicates $\{\mathcal{D}_R^{*(b)}\}_{b=1}^B$.

- 1: **for** $b = 1$ to B **do**
 - 2: Set random seed to $s + b$
 - 3: Generate an RSS sample $\mathcal{D}_R^{*(b)}$ of size mk from $\mathcal{D}_S$ using sampling with replacement.
 - 4: **end for**
 - 5: **return** $\{\mathcal{D}_R^{*(b)}\}_{b=1}^B$.
-

In the next part of the thesis, we discuss one of the key components of the proposed SRB methodology, the procedure for converting the observed SRS training data into an RSS training dataset using the limited ranking information available. This conversion follows one of two distinct approaches, depending on the bootstrap scheme employed.

4.4 SRS-RSS Construction with Paired Bootstrapping

Under the paired bootstrap construction of rank-informed bootstraps from SRS training data, it closely follows the standard RSS sampling procedure described in Section 4.1.2. The key difference is that the original population is no longer available; instead, the observed SRS training sample is treated as a finite population from which the RSS sample is constructed.

For each bootstrap replication $b = 1, \dots, B$, apply the RSS sampling procedure independently to the observed SRS training data. Although each RSS bootstrap sample is drawn from the same SRS training sample, the random formation of sets and selection of ranked units produce a distinct rank-informed bootstrap sample in each replication. The complete SRS-RSS paired bootstrap procedure is presented in Algorithm 5.

Next, we consider the SRS-RSS conversion in the residual bootstrapping scheme.

4.5 SRS-RSS Construction with Residual Bootstrapping

In this section, we propose two approaches to generate residual bootstrap samples by converting the SRS training sample into an RSS sample.

4.5.1 RSS Residual Bootstrapping Extension to SRS-RSS

Let $\mathcal{D}_S$ denote an SRS training sample of size n . For each bootstrap replicate $b = 1, \dots, B$, a balanced RSS sample of size $n = mk$ is first constructed from $\mathcal{D}_S$ using the ranking procedure described in Section 4.1.2. Thus, a new RSS is generated for each bootstrap replicate.

Once the RSS sample has been obtained, apply the residual bootstrapping procedure developed for RSS in Section 4.2.2. Fit a separate single regression model to obtain the fitted values and rank-specific residuals. Then resample residuals with replacement within each rank, and generate new bootstrap responses by adding the resampled residuals to the corresponding fitted values.

4.5.2 SRS-RSS Rank-Informed Residual Bootstrapping

The residual bootstrap procedure described above preserves the ranked structure of the predictor variables by first constructing an RSS sample and then resampling fitted residuals independently within each rank stratum. The residuals are used only to generate the bootstrap responses and play no role in ranking or in assigning observations to ranks.

In the current approach, we propose an alternative residual bootstrapping strategy that explicitly incorporates the fitted residuals into the SRS-RSS construction. Rather than resampling residuals separately within each rank, we first sample residuals from

Algorithm 6 SRS-RSS rank-informed residual bootstrapping

Require: SRS data $\mathcal{D}_S$, number of bootstrap replications B , set size k , cycles m , seed s

Ensure: Bootstrap replicates $\{\mathcal{D}_R^{*(b)}\}_{b=1}^B$.

- 1: Fit the regression model to $\mathcal{D}_S$ and obtain fitted values $\{\hat{Y}_i\}_{i=1}^n$ and residuals $\{\hat{e}_i\}_{i=1}^n$.
- 2: **for** $b = 1$ to B **do**
- 3: Set random seed to $s + b$.
- 4: Randomly partition the observations into k rank groups, each containing m observations.
- 5: **for** $r = 1$ to k **do**
- 6: **for** $j = 1$ to m **do**
- 7: Draw k residuals with replacement from $\{\hat{e}_i\}_{i=1}^n$.
- 8: Sort the sampled residuals in ascending order:

$$\hat{e}_{[1]j}^{*(b)} \leq \hat{e}_{[2]j}^{*(b)} \leq \dots \leq \hat{e}_{[k]j}^{*(b)}.$$

- 9: Assign the r -th ordered residual $\hat{e}_{[r]j}^{*(b)}$ to the j -th observation in rank r .
- 10: Construct the bootstrap response

$$Y_{[r]j}^{*(b)} = \hat{Y}_{[r]j} + \hat{e}_{[r]j}^{*(b)}.$$

- 11: **end for**
 - 12: **end for**
 - 13: Form the residual bootstrap RSS sample $\mathcal{D}_R^{*(b)}$.
 - 14: **end for**
 - 15: **return** $\{\mathcal{D}_R^{*(b)}\}_{b=1}^B$.
-

the fitted model, order them, and assign them to the corresponding rank positions. We then form the bootstrap responses by adding these rank-specific residuals to the fitted values. As a result, the ranking information is reflected not only through the sampled observations but also through the residual component of the bootstrap responses, yielding residual bootstrap samples whose error structure is consistent with the RSS framework. The proposed SRS-RSS rank-informed residual bootstrap is presented in Algorithm 6.

In this study, all SRS-RSS residual bootstraps are performed using the SRS-RSS rank-informed residual bootstrap method.

Figure 4.2 illustrates the complete workflow of the SRB procedure under rank-informed residual bootstrapping. Starting from the SRS training data $\mathcal{D}_S = \{(\mathbf{X}_i, Y_i)\}_{i=1}^{n=mk}$, a regression model $\hat{f}_S(\mathbf{X})$ is first fit to obtain the fitted values and residuals $\{\hat{e}_i\}_{i=1}^{n=mk}$. For

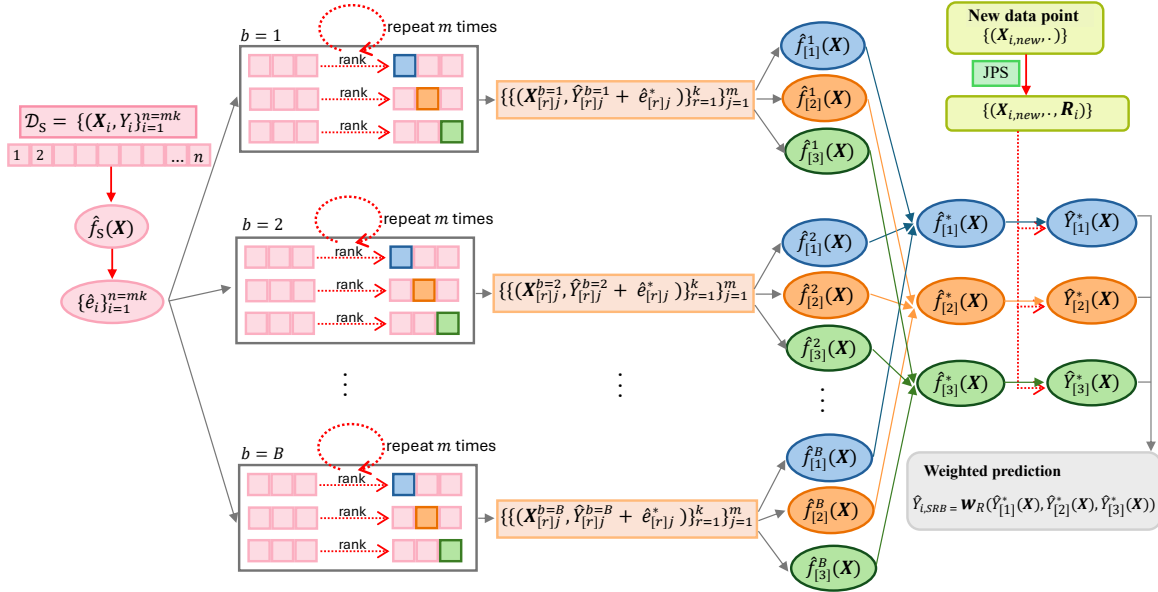


Figure 4.2: SRB resampling and prediction procedure for SRS data with set size 3 under rank-informed residual bootstrapping.

each bootstrap replication $b = 1, \dots, B$, the observations are randomly partitioned into k rank groups, and, over m repeated cycles, residuals are resampled with replacement, ordered, and assigned by rank following the rank-informed procedure in Algorithm 6. Adding each assigned residual to its corresponding fitted value yields the residual bootstrap RSS sample. For each bootstrap sample, we fit a single regression model and use it to obtain k rank-specific predictions at the target point $\mathbf{x}$. We then aggregate these predictions across the B bootstrap replicates to form bagged rank-specific predictors. For a new SRS test observation, the JPS procedure estimates its rank distribution, and we then combine the rank-specific predictions through a weighted aggregation scheme to obtain the final SRB prediction.

This concludes the description of the training and test data generation procedures and the resampling techniques employed by the proposed RSB and SRB methods throughout the simulation study.

4.6 Estimation and Prediction using Linear Models

In this section, we present simulation results for different covariate and ranking structures. These settings evaluate the estimation and predictive performance of the proposed rank-enriched bagging methods, RSB and SRB, and compare them with the standard SSB approach.

4.6.1 Single Covariate with Single Ranker

We consider a simple linear regression model of the form

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (4.4)$$

where the random errors satisfy $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, so that $\mathbb{E}(\varepsilon_i) = 0$, and $\text{Var}(\varepsilon_i) = \sigma^2$. The regression coefficients are set to $\beta_0 = -1.121$ and $\beta_1 = 0.424$, corresponding to the simple linear regression model in (4.4) derived from a bivariate normal distribution $\mathcal{N}_2(\mu_X = 1, \mu_Y = 5, \sigma_X^2 = 1, \sigma_Y^2 = 2, \rho_{XY} = 0.6)$ (Faraji et al., 2022). Following Dell and Clutter (1972), a ranking variable R_ℓ is constructed such that its correlation with the response variable Y is $\rho_{R_\ell, Y} \in \{0.85, 0.75, 0.50\}$ representing strong, moderate, and weak ranking respectively.

We considered a population of size $N = 100,000$. From this population, an SRS training sample $\mathcal{D}_S$ of size $n = 120$ and an independent SRS test sample $\mathcal{D}_S^{\text{test}}$ of size 10,000 are selected. Balanced RSS training samples $\mathcal{D}_R$ of size $n = 120$ are also constructed using set sizes $k \in \{2, 4, 6\}$, following the procedure described in Section 4.1.2.

The SRS test sample is converted into an RSS representation using the JPS procedure described in Algorithm 4. Two prediction strategies are considered, described in Section 4: (i) assigning each observation to its maximum-probability rank and (ii) using

the full estimated rank probability distribution. For the adjusted-response method, the correction terms **a** and **d** are computed using Equations (2.27) and (2.28), respectively. For the linear mixed model approach, model parameters are estimated using the **BFGS** optimizer in the **statsmodels** Python package, with **L-BFGS** and **CG** employed as fallback optimization methods when necessary. Throughout this simulation study, it is assumed that regression errors follow a normal distribution.

For each simulation setting, $B = 1000$ paired and residual bootstrap samples are generated to construct the corresponding bagged predictors. The entire experiment is then independently repeated six times. The results are summarized by the average relative efficiencies for prediction,

$$\text{RE}(\text{method}) = \frac{\widehat{\text{MSE}}(\text{SSB})}{\widehat{\text{MSE}}(\text{method})}, \quad \text{RE}_{\text{pred}}(\text{method}) = \frac{\widehat{\text{MSPE}}(\text{SSB})}{\widehat{\text{MSPE}}(\text{method})}, \quad (4.5)$$

where $\text{method} \in \{\text{RSB}, \text{SRB}\}$.

Tables 4.1 and 4.2 present the predictive relative efficiencies under paired and residual bootstrapping, respectively. Across all simulation settings, both rank-informed bagging procedures generally achieve relative efficiencies above 100%, indicating that incorporating ranking information into the bagging process yields more accurate predictions than SSB.

Under the paired bootstrap, predictive performance improves as ranking quality increases and declines as it deteriorates, confirming that the benefit of ranked bagging depends directly on the quality of the available ranking information. The predictive performance of RSB and SRB is similar under paired bootstrapping. Although RSB uses rank information directly from the population, whereas SRB constructs RSS retrospectively from an SRS sample, both procedures yield highly comparable prediction accuracies once sufficiently informative ranked bootstrap samples have been generated.

The residual bootstrap exhibits patterns similar to those observed under paired

bootstrap; predictive efficiency increases. Both RSB and SRB consistently outperform SSB. The adjusted-response approach provides stable gains across all correlation levels. Under the LMM framework, RSB demonstrates particularly strong performance, with predictive efficiencies generally increasing with set size and reaching levels comparable

Table 4.1: Relative efficiencies (RE_{pred}) of the fitted simple linear regression model using the proposed RSB and SRB procedures, relative to the SSB, in the single-ranker setting. Results are based on a test sample of size $n = 120$ under paired bootstrapping, with correlation levels $\rho \in \{0.85, 0.75, 0.5\}$, set sizes $k \in \{2, 4, 6\}$, and $B = 1000$ bootstrap replicates. “Adjusted” denotes weighted least squares with adjusted response, while “LMM” refers to the linear mixed model approach. Reported values are the (mean $\pm$ sd) relative efficiencies over six independent replications.

Paired Bootstrap						
Method	$\rho_{R_\ell, Y}$	Set Size	Adjusted		LMM	
			Rank	Prob.	Rank	Prob.
RSB	0.85	2	138.51 $\pm$ 2.63	140.59 $\pm$ 4.98	132.45 $\pm$ 5.27	131.05 $\pm$ 8.10
		4	155.36 $\pm$ 6.08	153.34 $\pm$ 6.35	166.45 $\pm$ 4.79	172.45 $\pm$ 10.19
		6	151.86 $\pm$ 2.12	149.38 $\pm$ 1.91	179.85 $\pm$ 9.04	190.94 $\pm$ 7.53
	0.75	2	126.38 $\pm$ 1.43	127.41 $\pm$ 2.97	120.36 $\pm$ 4.90	119.17 $\pm$ 7.26
		4	134.91 $\pm$ 4.06	133.22 $\pm$ 4.43	143.54 $\pm$ 2.28	146.85 $\pm$ 4.24
		6	130.96 $\pm$ 3.09	129.25 $\pm$ 2.86	146.83 $\pm$ 6.17	152.19 $\pm$ 9.33
	0.50	2	110.97 $\pm$ 1.16	111.28 $\pm$ 1.52	108.12 $\pm$ 2.24	107.96 $\pm$ 3.73
		4	111.36 $\pm$ 1.53	111.09 $\pm$ 1.58	111.60 $\pm$ 2.40	112.33 $\pm$ 2.97
		6	108.24 $\pm$ 3.59	107.82 $\pm$ 3.56	112.15 $\pm$ 5.27	114.74 $\pm$ 6.41
SRB	0.85	2	136.32 $\pm$ 1.86	138.80 $\pm$ 6.07	133.71 $\pm$ 3.35	132.65 $\pm$ 3.27
		4	154.67 $\pm$ 6.27	153.63 $\pm$ 8.77	166.84 $\pm$ 5.27	171.51 $\pm$ 7.35
		6	151.14 $\pm$ 7.94	148.71 $\pm$ 9.35	183.92 $\pm$ 6.75	191.41 $\pm$ 9.76
	0.75	2	124.85 $\pm$ 2.00	126.09 $\pm$ 4.80	123.53 $\pm$ 2.39	122.64 $\pm$ 2.43
		4	133.03 $\pm$ 4.37	131.78 $\pm$ 5.38	141.98 $\pm$ 3.71	143.80 $\pm$ 4.68
		6	129.17 $\pm$ 4.43	127.45 $\pm$ 5.00	150.05 $\pm$ 4.13	152.64 $\pm$ 5.83
	0.50	2	108.92 $\pm$ 1.23	109.44 $\pm$ 2.26	108.46 $\pm$ 1.00	108.00 $\pm$ 1.23
		4	109.84 $\pm$ 1.78	109.50 $\pm$ 1.99	112.60 $\pm$ 1.16	112.99 $\pm$ 1.90
		6	107.58 $\pm$ 1.72	107.10 $\pm$ 1.83	114.21 $\pm$ 1.58	114.47 $\pm$ 2.14

to those obtained under paired bootstrapping.

Under residual bootstrapping for SRB, the probability-based prediction approach consistently outperforms the maximum-rank approach, particularly under the LMM framework. These findings show that the probability-based approach better captures

Table 4.2: Relative efficiencies (RE_{pred}) of the fitted simple linear regression model using the proposed RSB and SRB procedures, relative to the SSB, in the single-ranker setting. Results are based on a test sample of size $n = 120$ under residual bootstrapping, with correlation levels $\rho \in \{0.85, 0.75, 0.5\}$, set sizes $k \in \{2, 4, 6\}$, and $B = 1000$ bootstrap replicates. “Adjusted” denotes weighted least squares with adjusted response, while “LMM” refers to the linear mixed model approach. Reported values are the (mean $\pm$ sd) relative efficiencies over six independent replications.

Residual Bootstrap						
Method	$\rho_{R_\ell, Y}$	Set Size	Adjusted		LMM	
			Rank	Prob.	Rank	Prob.
RSB	0.85	2	138.85 $\pm$ 3.35	140.22 $\pm$ 5.05	132.08 $\pm$ 5.62	130.65 $\pm$ 8.47
		4	155.90 $\pm$ 5.53	154.63 $\pm$ 5.63	165.69 $\pm$ 5.74	171.70 $\pm$ 11.38
		6	151.16 $\pm$ 3.91	149.44 $\pm$ 3.01	178.19 $\pm$ 9.44	188.89 $\pm$ 8.25
	0.75	2	126.74 $\pm$ 2.11	127.35 $\pm$ 3.19	119.82 $\pm$ 5.39	118.65 $\pm$ 7.70
		4	135.13 $\pm$ 3.47	133.83 $\pm$ 3.74	142.96 $\pm$ 2.80	146.15 $\pm$ 4.90
		6	130.62 $\pm$ 3.20	129.22 $\pm$ 2.86	145.10 $\pm$ 7.24	150.27 $\pm$ 10.42
	0.50	2	111.09 $\pm$ 1.31	111.34 $\pm$ 1.64	107.88 $\pm$ 2.37	107.75 $\pm$ 3.90
		4	111.25 $\pm$ 1.72	111.02 $\pm$ 1.79	111.19 $\pm$ 2.47	111.93 $\pm$ 3.06
		6	108.24 $\pm$ 2.79	107.85 $\pm$ 2.78	112.01 $\pm$ 5.25	114.53 $\pm$ 6.49
SRB	0.85	2	130.03 $\pm$ 6.78	133.85 $\pm$ 3.80	131.21 $\pm$ 8.93	140.26 $\pm$ 8.10
		4	144.69 $\pm$ 8.02	145.49 $\pm$ 6.61	146.22 $\pm$ 16.38	167.30 $\pm$ 18.29
		6	143.16 $\pm$ 5.50	141.88 $\pm$ 4.80	150.80 $\pm$ 19.86	172.16 $\pm$ 22.51
	0.75	2	121.45 $\pm$ 3.89	123.38 $\pm$ 2.62	121.58 $\pm$ 6.40	131.64 $\pm$ 6.15
		4	129.10 $\pm$ 4.50	128.60 $\pm$ 4.07	127.56 $\pm$ 11.19	145.94 $\pm$ 12.14
		6	126.46 $\pm$ 2.89	125.13 $\pm$ 2.92	128.74 $\pm$ 12.90	146.02 $\pm$ 14.11
	0.50	2	108.18 $\pm$ 1.49	108.81 $\pm$ 1.67	102.50 $\pm$ 3.07	113.92 $\pm$ 2.85
		4	109.44 $\pm$ 1.83	109.16 $\pm$ 1.93	96.30 $\pm$ 4.72	111.04 $\pm$ 4.60
		6	107.28 $\pm$ 1.16	106.83 $\pm$ 1.32	93.82 $\pm$ 5.15	106.16 $\pm$ 5.11

uncertainty in retrospectively constructed ranks and yields more reliable out-of-sample predictions, even in weaker-ranking scenarios where assigning each observation to a single estimated rank may lead to modest efficiency losses, especially for larger set sizes. These findings suggest that the probability-based approach better captures the uncertainty associated with retrospectively constructed ranks and leads to more reliable out-of-sample predictions.

The relatively small standard deviations across the six independent replications indicate that the proposed procedures are stable and produce consistent predictive performance. Overall, both RSB and SRB provide substantial improvements over conventional SSB, with the largest gains occurring when accurate ranking information is available. The results also highlight the practical advantage of combining the LMM framework with probability-based prediction, particularly for SRB under residual bootstrapping, where accounting for rank uncertainty leads to the most reliable predictive performance.

4.6.2 Single Covariate with Multiple Rankers

This setting follows the same simple linear regression model as in (4.4), with the same population, training sample, test sample, ranking variables, and simulation specifications as in the previous subsection. The key difference is that the available rank information of R_1 , R_2 , and R_3 is combined using the weighting mechanism described in Section 4.1.2 to construct weighted RSS training samples of size $n = 120$.

The SRS test sample is similarly converted into an RSS representation using the JPS procedure in Algorithm 4, where the rank probabilities of the three ranking variables are aggregated using the proposed weighting scheme. As in the single-ranker setting, both the maximum-rank and full-probability-distribution approaches are considered for prediction. For each simulation configuration, $B = 1000$ paired and residual bootstrap samples are generated, and the experiment is repeated over the same six random seeds

used in the previous subsection. In the adjusted response method, the corresponding aggregated ranking error probability matrix $\mathbb{P}_k$ is obtained as in (2.23). The predictive performances of the proposed RSB and SRB procedures are summarized using the relative efficiencies defined in Equation (4.5). Table 4.3 presents the corresponding results under paired and residual bootstrapping.

The proposed RSB and SRB estimators outperform the SSB, with RE_{pred} exceeding 100% in all simulation settings. Compared with the corresponding single-ranker results in Table 4.2, using multiple ranking variables generally results in higher predictive efficiency, highlighting the potential benefits of incorporating multiple rankers in the proposed bagging frameworks.

Under paired bootstrapping, RSB and SRB exhibit very similar predictive performance. The adjusted-response approach favors the maximum-rank prediction strategy, with efficiencies generally increasing as the set size increases. In the LMM framework, substantially larger gains are observed, particularly for larger set sizes.

Residual bootstrap shows trends similar to those under paired bootstrap. For RSB, predictive efficiency remains stable in the two resampling schemes. The adjusted-response approach again favors the maximum-rank prediction strategy, whereas under the LMM framework, the probability-based prediction becomes increasingly competitive as set size increases. For SRB, the adjusted-response method follows the same overall pattern as in the paired bootstrap. Under the LMM framework, the probability-based prediction strategy consistently outperforms the maximum-rank approach across all set sizes.

Table 4.3: Relative efficiencies (RE_{pred}) of the fitted simple linear regression models using the proposed RSB and SRB procedures relative to the SSB, using three rankers with $\rho_{R_1,Y} = 0.85$, $\rho_{R_2,Y} = 0.75$, and $\rho_{R_3,Y} = 0.50$. Results are based on a test sample of size $n = 120$ under paired and residual bootstrapping with set sizes $k \in \{2, 3, 4, 5, 6\}$ and $B = 1000$ bootstrap replicates. “Adjusted” denotes weighted least squares with adjusted response, while “LMM” refers to the linear mixed model approach. Reported values are the (mean $\pm$ sd) relative efficiencies over six independent replications.

Paired Bootstrap					
Method	Set Size	Adjusted		LMM	
		Rank	Prob.	Rank	Prob.
RSB	2	139.94 $\pm$ 3.77	129.86 $\pm$ 4.68	136.14 $\pm$ 7.49	126.36 $\pm$ 7.10
	3	147.21 $\pm$ 6.72	133.40 $\pm$ 5.84	159.13 $\pm$ 10.57	146.70 $\pm$ 12.76
	4	148.90 $\pm$ 3.37	135.10 $\pm$ 2.74	175.62 $\pm$ 8.11	167.08 $\pm$ 13.30
	5	150.06 $\pm$ 4.14	136.58 $\pm$ 4.11	188.13 $\pm$ 10.83	198.54 $\pm$ 21.05
	6	143.20 $\pm$ 4.25	131.22 $\pm$ 4.51	189.39 $\pm$ 8.56	190.08 $\pm$ 14.40
SRB	2	137.39 $\pm$ 2.91	128.10 $\pm$ 4.69	138.47 $\pm$ 3.94	128.70 $\pm$ 2.50
	3	148.02 $\pm$ 4.63	135.14 $\pm$ 5.43	162.24 $\pm$ 4.37	150.29 $\pm$ 4.68
	4	147.70 $\pm$ 5.28	134.58 $\pm$ 5.58	177.37 $\pm$ 5.95	166.62 $\pm$ 7.00
	5	146.01 $\pm$ 5.79	133.48 $\pm$ 6.00	186.95 $\pm$ 6.29	179.14 $\pm$ 8.14
	6	142.48 $\pm$ 5.64	130.96 $\pm$ 5.80	193.79 $\pm$ 7.43	188.87 $\pm$ 9.79
Residual Bootstrap					
Method	Set Size	Adjusted		LMM	
		Rank	Prob.	Rank	Prob.
RSB	2	140.19 $\pm$ 4.21	129.86 $\pm$ 5.02	135.60 $\pm$ 7.88	125.98 $\pm$ 7.48
	3	146.86 $\pm$ 6.36	133.85 $\pm$ 5.49	158.50 $\pm$ 10.56	146.42 $\pm$ 12.99
	4	148.32 $\pm$ 3.38	135.82 $\pm$ 2.29	174.39 $\pm$ 9.18	166.34 $\pm$ 14.02
	5	146.78 $\pm$ 4.62	135.98 $\pm$ 4.03	186.79 $\pm$ 11.18	197.12 $\pm$ 22.23
	6	141.90 $\pm$ 3.62	131.59 $\pm$ 2.96	187.59 $\pm$ 8.53	188.29 $\pm$ 14.73
SRB	2	132.18 $\pm$ 5.18	125.29 $\pm$ 2.31	137.41 $\pm$ 10.30	137.51 $\pm$ 7.17
	3	140.68 $\pm$ 5.68	131.36 $\pm$ 3.22	148.26 $\pm$ 14.37	158.24 $\pm$ 13.05
	4	140.98 $\pm$ 5.95	131.16 $\pm$ 3.34	150.73 $\pm$ 17.73	170.17 $\pm$ 17.19
	5	139.60 $\pm$ 4.50	130.10 $\pm$ 2.41	150.33 $\pm$ 19.02	177.09 $\pm$ 20.11
	6	137.22 $\pm$ 4.27	128.13 $\pm$ 2.27	149.93 $\pm$ 21.23	182.06 $\pm$ 22.78

In the multiple-ranker setting, predictive performance remains comparable to or better than that of the single-ranker setting, despite the inclusion of one relatively weak ranker ($\rho_{R_\ell, Y} = 0.50$). This is because the stronger rankers continue to provide reliable ranking information. In the single-ranker setting, predictive performance depends entirely on the quality of the available ranking variable and deteriorates noticeably when the ranking correlation decreases.

4.7 Estimation and Prediction using Spline Models

In many applications, more flexible models than linear regression are required to approximate nonlinear relationships, improve predictive performance, and obtain more accurate estimates. In recent years, cubic smoothing spline regression models have become increasingly popular in statistical modeling. A cubic spline is a continuous piecewise polynomial function constructed by joining cubic polynomials over intervals defined by a set of knots. Its first and second derivatives, as well as the function itself, must remain continuous across the entire range of the covariate.

In general, a cubic spline regression model with Q knots $\{\xi_1, \dots, \xi_Q\}$ can be expressed as

$$Y_i = \beta_0 + \sum_{d=1}^3 \beta_d X_i^d + \sum_{q=1}^Q \beta_{3q} (X_i - \xi_q)_+^3 + \varepsilon_i, \quad i = 1, \dots, n, \quad (4.6)$$

where $(X_i - \xi_q)_+^3 = (X_i - \xi_q)^3 \mathbb{I}(X_i \geq \xi_q)$ denotes a truncated cubic polynomial. For an SRS data $\mathcal{D}_S = \{(\mathbf{X}_i, Y_i) : i = 1, \dots, n\}$, the cubic spline model can be written in matrix form as $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbb{X}$ is the design matrix whose i -th row is given by $\mathbb{X}_i^\top = \{1, X_i, X_i^2, X_i^3, (X_i - \xi_1)_+^3, \dots, (X_i - \xi_Q)_+^3\}$, and $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_{D1}, \dots, \beta_{DQ})$. Under the standard assumptions for SRS data, the error vector satisfies $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$.

In previous studies, [Algarni \(2016\)](#) and [Al Kadiri \(2017\)](#) explored spline regression modeling using RSS. However, these approaches followed the framework introduced by [Muttlak \(1995\)](#) and did not fully incorporate the available rank information in model estimation. In this simulation study, we consider a cubic spline regression model with three knots with degree three. For an RSS sample $\mathcal{D}_R$ of size $n = mk$, represented as $\{(X_{[r]j}, Y_{[r]j}) : r = 1, 2, 3, j = 1, \dots, m\}$ the corresponding cubic spline model can be written as

$$Y_{[r]j} = \beta_0 + \sum_{d=1}^3 \beta_d X_{[r]j}^d + \sum_{q=1}^3 \beta_{3q} (X_{[r]j} - \xi_q)_+^3 + \varepsilon_{[r]j}. \quad (4.7)$$

In the nonlinear setting, the data generation mechanism of [Faraji et al. \(2022\)](#) is followed. Data are generated from the nonlinear model

$$Y = 2 \sin(X) e^{-X^2} + \varepsilon, \quad (4.8)$$

where $X \sim \text{Uniform}(-2, 2)$ and $\varepsilon \sim \mathcal{N}(0, 0.12^2)$ ([Algarni, 2016](#)). The ranking variable R_ℓ is generated following the model of [Dell and Clutter \(1972\)](#) such that $\rho_{R_\ell, Y} \in \{0.85, 0.75, 0.50\}$, representing strong, moderate, and weak ranking quality, respectively. For the adjusted-response approach, the imperfect-ranking probabilities are estimated from the SRS training sample of size $n = 120$.

The simulation settings are identical to those used in the linear regression study, except that the number of bootstrap replications is reduced to $B = 500$ due to the increased computational cost of fitting spline models. Figures [4.3](#) and [4.4](#) present the fitted spline regression curves obtained using the baseline SSB method together with the proposed RSB and SRB methods under the paired and residual bootstrap schemes, respectively. Although model performance was evaluated using an independent test sample of size $n = 10,000$ and 500 bootstrap replicates, for visual clarity only a randomly selected subset of 240 test observations and a maximum of 200 bootstrap curves are

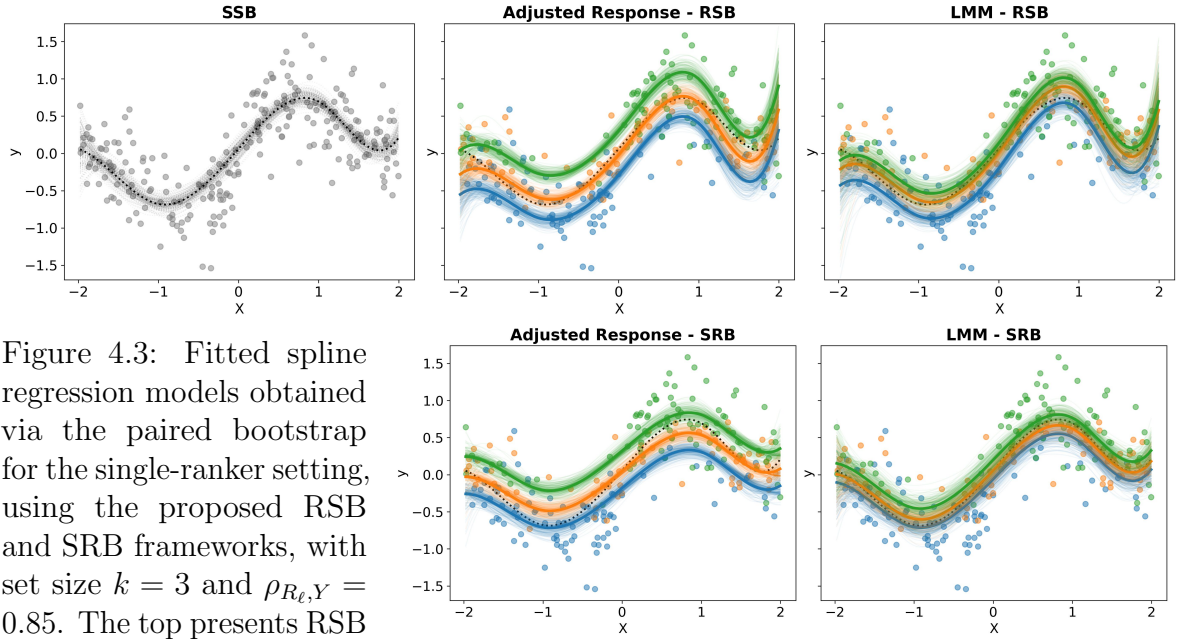


Figure 4.3: Fitted spline regression models obtained via the paired bootstrap for the single-ranker setting, using the proposed RSB and SRB frameworks, with set size $k = 3$ and $\rho_{R_\ell, Y} = 0.85$. The top presents RSB

results, where the left, middle, and right panels correspond to the SSB estimator obtained from SRS test data, the RSB estimator based on the adjusted-response approach, and the RSB estimator based on the LMM approach, respectively. The bottom row presents the corresponding SRB estimators based on the adjusted-response and LMM approaches. Grey points denote the original SRS test observations, whereas colored points indicate the observations assigned to each rank: blue ($k = 1$), orange ($k = 2$), and green ($k = 3$). The thick curves represent the corresponding bagged spline estimators, while the faint curves show a bootstrap spline fit used to construct the final bagged estimators. The black dotted curve in each panel represents the SSB estimator.

displayed for set size $k = 3$, where the ranking variable is generated using the Dell and Clutter model with $\rho_{R_\ell, Y} = 0.85$.

The RE_{pred} are computed using Equation (4.5), and the corresponding efficiencies from paired and residual bootstrapping are summarized in Tables 4.4 and 4.5, respectively, for comparison. Non-linear RE_{pred} followed the same pattern observed for linear single- and multi-rankers settings: all relative efficiencies exceeded 100%, demonstrating that incorporating ranking information into the bagging procedure consistently improves predictive performance over the SSB approach, even in the more challenging non-linear setting.

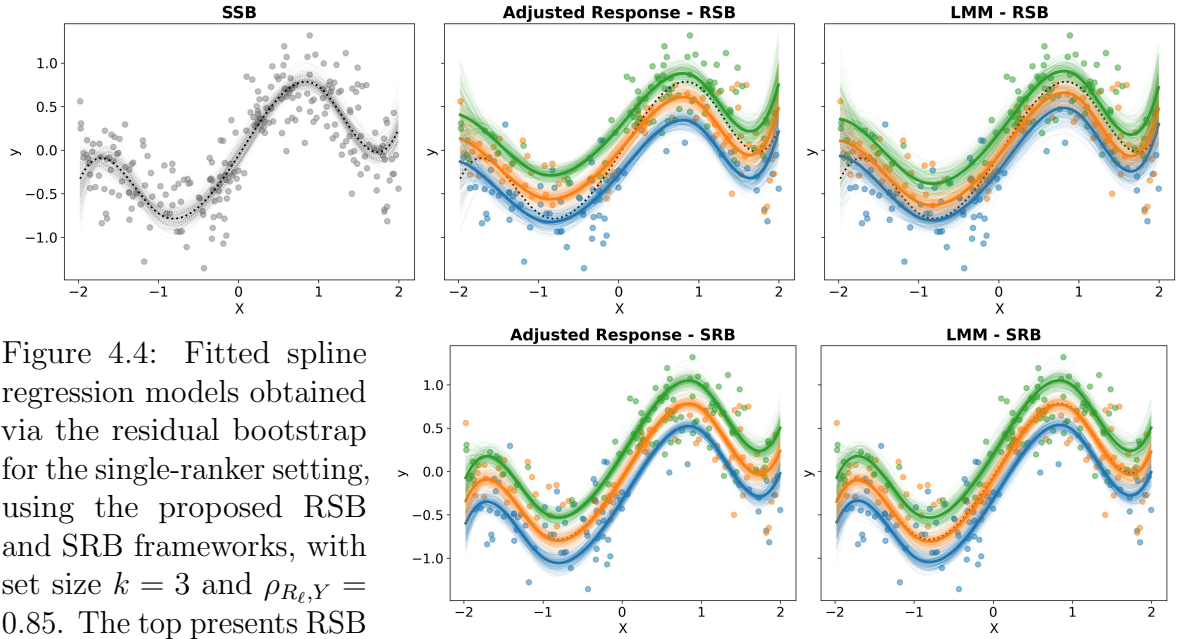


Figure 4.4: Fitted spline regression models obtained via the residual bootstrap for the single-ranker setting, using the proposed RSB and SRB frameworks, with set size $k = 3$ and $\rho_{R_\ell, Y} = 0.85$. The top presents RSB

results, where the left, middle, and right panels correspond to the SSB estimator obtained from SRS test data, the RSB estimator based on the adjusted-response approach, and the RSB estimator based on the LMM approach, respectively. The bottom row presents the corresponding SRB estimators based on the adjusted-response and LMM approaches. Grey points denote the original SRS test observations, whereas colored points indicate the observations assigned to each rank: blue ($k = 1$), orange ($k = 2$), and green ($k = 3$). The thick curves represent the corresponding bagged spline estimators, while the faint curves show a bootstrap spline fit used to construct the final bagged estimators. The black dotted curve in each panel represents the SSB estimator.

Predictive performance increases with ranking quality and generally with larger set sizes in paired bootstrapping. For both RSB and SRB, the strongest improvements are observed when $\rho_{R_\ell, Y} = 0.85$; as ranker quality deteriorates, efficiencies become more modest. As in the linear regression setting, the adjusted-response and LMM approaches yield comparable results, whereas the probability-based prediction strategy outperforms the maximum-rank approach. RSB and SRB exhibit very similar predictive performance, indicating that the retrospective construction of ranked bootstrap samples in SRB effectively recovers most of the efficiency achieved when RSS training samples are directly available.

The residual bootstrap exhibits similar overall trends, although its relative efficiency is generally lower than that of paired bootstrapping. In SRB, the two prediction strategies

are clearly distinguished, and the probability-based prediction approach consistently outperforms the maximum-rank approach across all ranking qualities and set sizes, particularly under the LMM framework. Overall, the single-covariate spline model with single-rank results reinforces the conclusions from the linear regression study.

Table 4.4: Relative efficiencies (RE_{pred}) of the fitted spline model using the proposed RSB and SRB procedures, relative to the SSB, in the single-ranker setting. Results are based on a test sample of size $n = 120$ under paired bootstrapping, with correlation levels $\rho \in \{0.85, 0.75, 0.5\}$, set sizes $k \in \{2, 4, 6\}$, and $B = 500$ bootstrap replicates. “Adjusted” denotes weighted least squares with adjusted response, while “LMM” refers to the linear mixed model approach. Reported values are the (mean $\pm$ sd) relative efficiencies over six independent replications.

Paired Bootstrap						
Method	$\rho_{R_\ell, Y}$	Set Size	Adjusted		LMM	
			Rank	Prob.	Rank	Prob.
RSB	0.85	2	116.78 $\pm$ 9.91	128.27 $\pm$ 8.14	111.98 $\pm$ 7.05	111.44 $\pm$ 7.22
		4	141.44 $\pm$ 9.19	152.31 $\pm$ 8.91	135.76 $\pm$ 7.74	139.76 $\pm$ 9.70
		6	146.37 $\pm$ 6.53	147.30 $\pm$ 6.44	144.91 $\pm$ 10.00	150.30 $\pm$ 10.08
	0.75	2	111.93 $\pm$ 10.58	118.86 $\pm$ 9.65	108.34 $\pm$ 6.78	107.83 $\pm$ 6.81
		4	125.33 $\pm$ 7.16	129.95 $\pm$ 6.65	121.46 $\pm$ 7.43	124.96 $\pm$ 10.97
		6	130.80 $\pm$ 4.86	130.34 $\pm$ 4.68	128.90 $\pm$ 7.56	133.65 $\pm$ 6.69
	0.50	2	103.04 $\pm$ 8.71	105.88 $\pm$ 8.30	100.04 $\pm$ 5.61	100.16 $\pm$ 5.89
		4	107.20 $\pm$ 6.31	108.02 $\pm$ 6.98	104.05 $\pm$ 4.83	105.87 $\pm$ 7.61
		6	109.83 $\pm$ 4.57	109.94 $\pm$ 4.66	105.59 $\pm$ 5.53	107.67 $\pm$ 4.95
SRB	0.85	2	122.41 $\pm$ 4.80	135.69 $\pm$ 6.11	116.50 $\pm$ 2.89	115.80 $\pm$ 3.25
		4	143.13 $\pm$ 5.62	150.69 $\pm$ 7.86	134.20 $\pm$ 6.70	135.56 $\pm$ 8.16
		6	146.68 $\pm$ 7.01	147.97 $\pm$ 8.06	144.00 $\pm$ 8.87	146.66 $\pm$ 11.26
	0.75	2	116.15 $\pm$ 4.31	124.40 $\pm$ 4.91	111.81 $\pm$ 2.88	111.11 $\pm$ 3.17
		4	127.16 $\pm$ 3.98	130.29 $\pm$ 5.03	121.26 $\pm$ 5.65	121.95 $\pm$ 6.88
		6	127.63 $\pm$ 4.43	127.35 $\pm$ 4.62	126.21 $\pm$ 7.58	127.03 $\pm$ 8.78
	0.50	2	106.19 $\pm$ 2.56	109.08 $\pm$ 2.70	104.29 $\pm$ 1.63	104.12 $\pm$ 1.87
		4	109.73 $\pm$ 2.85	110.26 $\pm$ 2.70	106.70 $\pm$ 3.36	106.74 $\pm$ 3.57
		6	109.00 $\pm$ 2.82	109.03 $\pm$ 2.80	107.52 $\pm$ 3.66	107.73 $\pm$ 4.02

Table 4.5: Relative efficiencies (RE_{pred}) of the fitted spline model using the proposed RSB and SRB procedures, relative to the SSB, in the single-ranker setting. Results are based on a test sample of size $n = 120$ under residual bootstrapping, with correlation levels $\rho \in \{0.85, 0.75, 0.5\}$, set sizes $k \in \{2, 4, 6\}$, and $B = 500$ bootstrap replicates. “Adjusted” denotes weighted least squares with adjusted response, while “LMM” refers to the linear mixed model approach. Reported values are the (mean $\pm$ sd) relative efficiencies over six independent replications.

Residual Bootstrap						
Method	$\rho_{R_\ell, Y}$	Set Size	Adjusted		LMM	
			Rank	Prob.	Rank	Prob.
RSB	0.85	2	118.17 $\pm$ 4.36	126.74 $\pm$ 7.38	112.48 $\pm$ 3.48	111.47 $\pm$ 3.73
		4	137.72 $\pm$ 6.00	145.38 $\pm$ 8.10	128.15 $\pm$ 8.63	131.61 $\pm$ 10.23
		6	139.29 $\pm$ 5.70	143.62 $\pm$ 6.57	139.68 $\pm$ 3.87	150.22 $\pm$ 8.63
	0.75	2	112.96 $\pm$ 2.65	118.30 $\pm$ 4.41	107.74 $\pm$ 2.77	106.82 $\pm$ 3.00
		4	123.54 $\pm$ 3.40	126.82 $\pm$ 3.72	118.87 $\pm$ 4.56	121.76 $\pm$ 6.64
		6	122.34 $\pm$ 8.50	123.82 $\pm$ 8.40	122.46 $\pm$ 5.27	128.86 $\pm$ 6.32
	0.50	2	105.33 $\pm$ 2.87	107.77 $\pm$ 2.58	102.60 $\pm$ 2.08	102.17 $\pm$ 2.08
		4	106.28 $\pm$ 7.90	106.99 $\pm$ 8.19	104.09 $\pm$ 8.20	105.17 $\pm$ 8.92
		6	103.22 $\pm$ 4.94	103.63 $\pm$ 5.05	102.58 $\pm$ 4.46	104.82 $\pm$ 5.18
SRB	0.85	2	104.11 $\pm$ 6.44	121.67 $\pm$ 5.52	109.06 $\pm$ 4.90	120.28 $\pm$ 4.96
		4	114.54 $\pm$ 7.62	125.43 $\pm$ 7.85	109.43 $\pm$ 7.87	124.29 $\pm$ 8.40
		6	120.95 $\pm$ 8.06	126.89 $\pm$ 7.79	108.72 $\pm$ 8.52	122.68 $\pm$ 9.50
	0.75	2	105.19 $\pm$ 5.38	116.06 $\pm$ 3.43	105.15 $\pm$ 3.60	116.24 $\pm$ 3.88
		4	113.70 $\pm$ 4.92	118.79 $\pm$ 4.93	102.85 $\pm$ 6.00	116.83 $\pm$ 6.44
		6	116.99 $\pm$ 5.03	119.25 $\pm$ 4.61	101.45 $\pm$ 6.52	113.81 $\pm$ 7.15
	0.50	2	102.72 $\pm$ 3.02	106.57 $\pm$ 1.27	95.60 $\pm$ 2.19	106.75 $\pm$ 1.85
		4	105.84 $\pm$ 1.67	106.77 $\pm$ 1.85	88.21 $\pm$ 3.89	100.56 $\pm$ 3.30
		6	106.33 $\pm$ 1.61	106.92 $\pm$ 1.47	85.18 $\pm$ 4.13	95.60 $\pm$ 4.08

4.7.1 Single Covariate Spline Regression with Multiple Rankers

In the previous section, spline regression models were investigated in the single-ranking setting, considering ranking variables with strong, moderate, and weak ranking abilities.

We now extend this framework to the multiple-ranking setting by generating three ranking variables, R_1 , R_2 , and R_3 , such that $\rho_{R_\ell, Y} \in \{0.85, 0.75, 0.50\}$, $\ell = 1, 2, 3$, and using them all in constructing RSS data. Balanced RSS samples are constructed using set sizes $k \in \{2, 3, 4, 5, 6\}$ while maintaining a fixed training sample size of $n = 120$.

The remaining simulation settings are identical to those used in the previous spline regression study. In particular, we used $B = 500$ bootstrap replicates and repeated the experiment with the same six independent random seeds used in the linear regression simulations. Figures 4.5 and 4.6 present the fitted spline regression curves obtained using the baseline SSB method together with the proposed RSB and SRB methods under the paired and residual bootstrap schemes, respectively. Although model performance was evaluated using an independent test sample of size $n = 10,000$ and 500 bootstrap replicates, only a randomly selected subset of 240 test observations and a maximum of 200 bootstrap curves are displayed for visual clarity. The results correspond to a set size of $k = 3$. Table 4.6 summarizes the average predictive relative efficiencies, calculated using Equation (4.5), for the proposed RSB and SRB methods relative to SSB using paired and residual bootstrapping.

Despite incorporating a relatively weak ranking variable, both RSB and SRB achieve higher predictive gains under paired bootstrapping across all set sizes. The adjusted-response and LMM approaches produce very similar trends, and the differences between the maximum-rank and probability-based prediction strategies remain relatively small. The two methods also perform almost identically throughout, even in this more complex non-linear setting. Under residual bootstrapping, the overall efficiencies are lower than those obtained under paired bootstrapping but remain consistently above 100%. For RSB, predictive performance improves steadily as set size increases under both the adjusted-response and LMM approaches. By contrast, SRB again benefits substantially from the probability-based prediction strategy, particularly under the LMM framework.

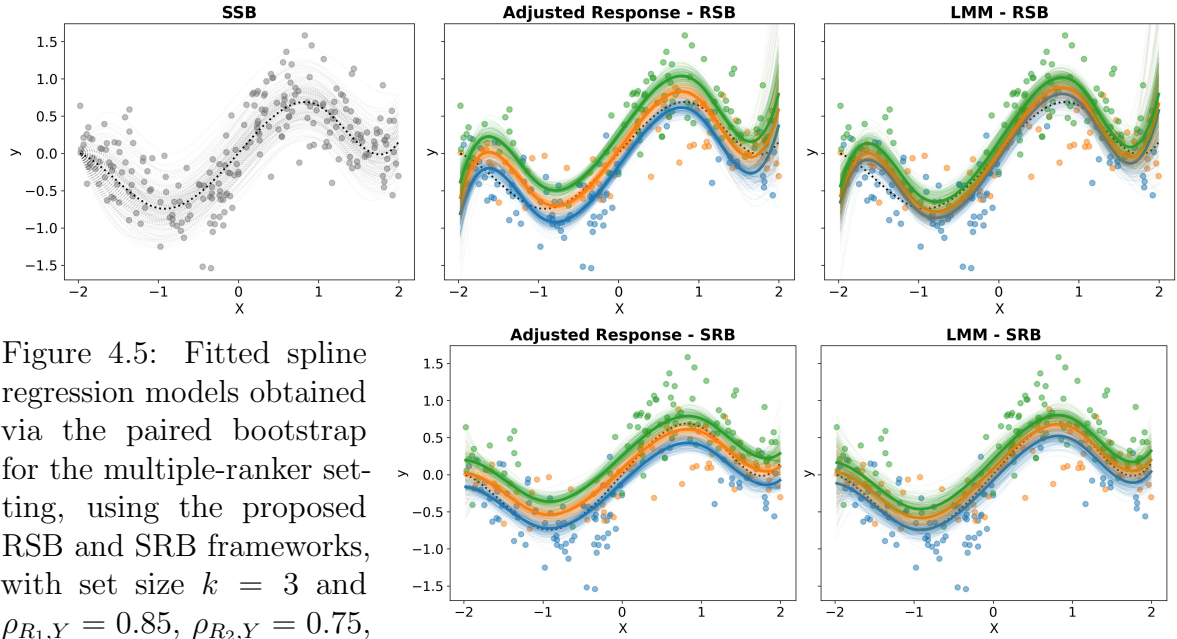


Figure 4.5: Fitted spline regression models obtained via the paired bootstrap for the multiple-ranker setting, using the proposed RSB and SRB frameworks, with set size $k = 3$ and $\rho_{R_1,Y} = 0.85$, $\rho_{R_2,Y} = 0.75$, and $\rho_{R_3,Y} = 0.50$. The top

presents RSB results, where the left, middle, and right panels correspond to the SSB estimator obtained from SRS test data, the RSB estimator based on the adjusted-response approach, and the RSB estimator based on the LMM approach, respectively. The bottom row presents the corresponding SRB estimators based on the adjusted-response and LMM approaches. Grey points denote the original SRS test observations, whereas colored points indicate the observations assigned to each rank: blue ($k = 1$), orange ($k = 2$), and green ($k = 3$). The thick curves represent the corresponding bagged spline estimators, while the faint curves show a bootstrap spline fit used to construct the final bagged estimators. The black dotted curve in each panel represents the SSB estimator.

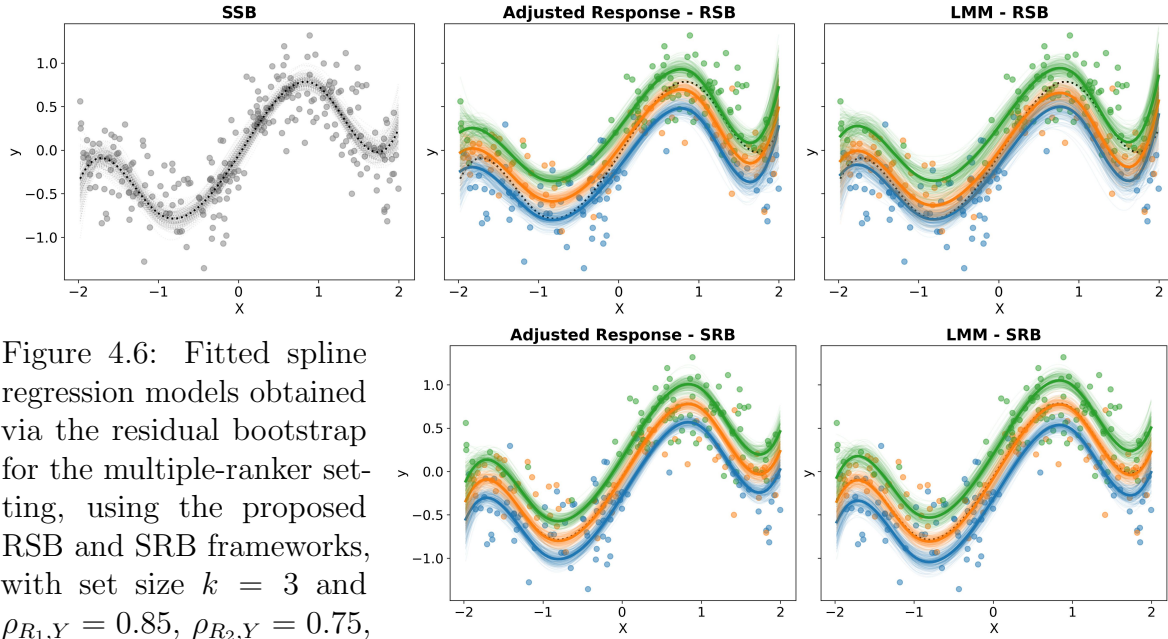


Figure 4.6: Fitted spline regression models obtained via the residual bootstrap for the multiple-ranker setting, using the proposed RSB and SRB frameworks, with set size $k = 3$ and $\rho_{R_1,Y} = 0.85$, $\rho_{R_2,Y} = 0.75$, and $\rho_{R_3,Y} = 0.50$. The top

presents RSB results, where the left, middle, and right panels correspond to the SSB estimator obtained from SRS test data, the RSB estimator based on the adjusted-response approach, and the RSB estimator based on the LMM approach, respectively. The bottom row presents the corresponding SRB estimators based on the adjusted-response and LMM approaches. Grey points denote the original SRS test observations, whereas colored points indicate the observations assigned to each rank: blue ($k = 1$), orange ($k = 2$), and green ($k = 3$). The thick curves represent the corresponding bagged spline estimators, while the faint curves show a bootstrap spline fit used to construct the final bagged estimators. The black dotted curve in each panel represents the SSB estimator.

Table 4.6: Relative efficiencies (RE_{pred}) of the fitted spline regression models using the proposed RSB and SRB procedures relative to the SSB, using three rankers with $\rho_{R_1,Y} = 0.85$, $\rho_{R_2,Y} = 0.75$, and $\rho_{R_3,Y} = 0.50$. Results are based on a test sample of size $n = 120$ under paired and residual bootstrapping with set sizes $k \in \{2, 3, 4, 5, 6\}$ and $B = 500$ bootstrap replicates. “Adjusted” denotes weighted least squares with adjusted response, while “LMM” refers to the linear mixed model approach. Reported values are the (mean $\pm$ sd) relative efficiencies over six independent replications.

Paired Bootstrap					
Method	Set Size	Adjusted		LMM	
		Rank	Prob.	Rank	Prob.
RSB	2	220.87 $\pm$ 101.43	218.30 $\pm$ 95.91	213.23 $\pm$ 103.87	204.78 $\pm$ 97.46
	3	239.21 $\pm$ 107.46	236.18 $\pm$ 104.66	227.55 $\pm$ 105.99	223.85 $\pm$ 99.44
	4	246.50 $\pm$ 116.80	238.50 $\pm$ 113.87	244.28 $\pm$ 111.48	244.02 $\pm$ 104.36
	5	256.02 $\pm$ 118.78	244.26 $\pm$ 114.02	266.74 $\pm$ 119.54	268.29 $\pm$ 114.66
	6	244.87 $\pm$ 109.63	236.00 $\pm$ 108.14	253.75 $\pm$ 109.64	257.91 $\pm$ 112.55
SRB	2	227.84 $\pm$ 105.53	229.29 $\pm$ 109.42	212.78 $\pm$ 99.84	206.23 $\pm$ 97.61
	3	247.75 $\pm$ 116.89	243.08 $\pm$ 119.45	235.77 $\pm$ 111.23	228.04 $\pm$ 109.95
	4	254.53 $\pm$ 122.98	243.94 $\pm$ 120.28	252.21 $\pm$ 120.39	244.94 $\pm$ 119.68
	5	255.18 $\pm$ 122.09	241.69 $\pm$ 116.88	262.79 $\pm$ 124.70	257.65 $\pm$ 126.35
	6	251.00 $\pm$ 121.56	236.31 $\pm$ 114.89	270.75 $\pm$ 128.37	268.14 $\pm$ 132.72
Residual Bootstrap					
Method	Set Size	Adjusted		LMM	
		Rank	Prob.	Rank	Prob.
RSB	2	122.54 $\pm$ 6.65	119.68 $\pm$ 4.47	113.58 $\pm$ 6.97	110.33 $\pm$ 6.32
	3	127.45 $\pm$ 10.32	125.00 $\pm$ 8.30	122.66 $\pm$ 9.29	120.88 $\pm$ 9.16
	4	134.90 $\pm$ 9.27	130.49 $\pm$ 8.63	132.88 $\pm$ 11.28	129.60 $\pm$ 9.79
	5	137.84 $\pm$ 4.78	131.11 $\pm$ 3.86	144.60 $\pm$ 7.83	142.34 $\pm$ 10.55
	6	128.41 $\pm$ 6.50	124.29 $\pm$ 5.55	141.97 $\pm$ 5.04	144.26 $\pm$ 9.10
SRB	2	111.72 $\pm$ 7.01	119.23 $\pm$ 3.11	111.71 $\pm$ 5.74	119.64 $\pm$ 4.58
	3	117.53 $\pm$ 8.74	122.97 $\pm$ 3.87	112.09 $\pm$ 8.40	126.81 $\pm$ 7.11
	4	120.51 $\pm$ 7.67	123.20 $\pm$ 4.95	109.54 $\pm$ 8.73	128.48 $\pm$ 8.31
	5	122.02 $\pm$ 6.83	122.77 $\pm$ 4.35	107.17 $\pm$ 9.57	129.30 $\pm$ 9.09
	6	123.32 $\pm$ 5.80	122.33 $\pm$ 3.87	105.95 $\pm$ 8.54	129.89 $\pm$ 9.39

Comparing the single and multiple-ranker settings reveals an important practical finding. Although the multiple-ranker framework incorporates a weak-ranking variable with $\rho_{R_\ell, Y} = 0.50$, its predictive performance is superior to that of the single ranker case. This shows that the proposed weighted RSS mechanism combines several ranking variables to integrate complementary information while reducing the influence of any individual weak ranker. As a result, the collective ranking signal provides a more accurate representation of the underlying ordering than any single variable alone, yielding substantially larger gains in predictive efficiency. Collectively, these results confirm that combining multiple ranking variables is more beneficial than relying on a single ranker, even when some of the individual ranks are only moderately or weakly informative.

4.7.2 Additive Spline Regression with Multiple Rankers

We now extend the nonlinear regression framework to the multiple-covariate, multiple-ranker setting. Eight independent covariates, $X_1, \dots, X_8 \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}(-2, 2)$, are generated, and the response variable is defined by

$$Y = \sum_{j=1}^8 2 \sin(X_j) e^{-X_j^2} + \varepsilon, \quad (4.9)$$

where $\varepsilon \sim \mathcal{N}(0, 0.12^2)$. The remaining simulation settings, including the ranking mechanism, sampling design, bootstrap procedures, and prediction strategies, are identical to those described in Section 4.7.1.

Unlike the single-covariate setting, this scenario involves a high-dimensional nonlinear regression problem in which the additive spline basis yields a large number of predictors relative to the sample size. Each covariate is modeled using a cubic B-spline basis of degree 3 with 3 interior knots. The spline basis is constructed to include an intercept, resulting in a single global intercept in the final additive spline regression model. To

Table 4.7: Relative efficiencies (RE_{pred}) of the LASSO-regularized additive fitted spline regression models using the proposed RSB and SRB procedures relative to the SSB, using three rankers with $\rho_{R_1,Y} = 0.85$, $\rho_{R_2,Y} = 0.75$, and $\rho_{R_3,Y} = 0.50$. Results are based on a test sample of size $n = 120$ under paired and residual bootstrapping with set sizes $k \in \{2, 4, 6\}$ and $B = 500$ bootstrap replicates. “Adjusted” denotes weighted least squares with adjusted response, while “LMM” refers to the linear mixed model approach. Reported values are the (mean $\pm$ sd) relative efficiencies over six independent replications.

Paired Bootstrap					
Method	Set Size	Adjusted		LMM	
		Rank	Prob.	Rank	Prob.
RSB	2	85.71 $\pm$ 9.76	102.19 $\pm$ 12.05	103.26 $\pm$ 7.62	102.78 $\pm$ 7.44
	4	107.22 $\pm$ 14.19	117.01 $\pm$ 18.47	109.77 $\pm$ 20.00	109.46 $\pm$ 20.45
	6	117.64 $\pm$ 10.82	117.90 $\pm$ 12.64	113.95 $\pm$ 12.85	114.16 $\pm$ 14.10
SRB	2	92.64 $\pm$ 9.65	109.73 $\pm$ 3.42	100.52 $\pm$ 4.21	99.95 $\pm$ 4.06
	4	108.42 $\pm$ 10.17	117.30 $\pm$ 2.53	102.56 $\pm$ 3.94	101.69 $\pm$ 3.68
	6	116.04 $\pm$ 6.05	118.31 $\pm$ 3.48	102.62 $\pm$ 4.42	101.74 $\pm$ 4.14
Residual Bootstrap					
Method	Set Size	Adjusted		LMM	
		Rank	Prob.	Rank	Prob.
RSB	2	80.12 $\pm$ 15.33	80.40 $\pm$ 17.39	106.42 $\pm$ 10.11	106.09 $\pm$ 10.13
	4	108.87 $\pm$ 25.52	107.23 $\pm$ 29.13	109.84 $\pm$ 24.16	109.72 $\pm$ 24.53
	6	116.37 $\pm$ 14.79	112.36 $\pm$ 16.11	113.38 $\pm$ 21.16	114.30 $\pm$ 22.82
SRB	2	93.30 $\pm$ 19.00	120.66 $\pm$ 11.80	112.64 $\pm$ 10.35	114.26 $\pm$ 8.66
	4	101.77 $\pm$ 25.34	121.67 $\pm$ 18.73	113.69 $\pm$ 14.15	120.60 $\pm$ 13.11
	6	110.83 $\pm$ 22.84	122.15 $\pm$ 17.33	112.53 $\pm$ 15.36	121.80 $\pm$ 15.10

accommodate this setting, we employ LASSO-regularized WLS for the adjusted response and LMM, as described in Chapter 2. The Prediction performance is evaluated using the relative efficiency measure RE_{pred} defined in Equation (4.5). The corresponding results under paired and residual bootstrapping are summarized in Table 4.7.

The adjusted-response approach with paired bootstrap, combined with the probability-

based prediction strategy, consistently provides the strongest performance for both RSB and SRB. For both methods, RE_{pred} increases steadily with the set size. In contrast, the maximum-rank prediction strategy occasionally performs worse than SSB for smaller set sizes, in a complex nonlinear setting. LMM approach exhibits different behavior under paired bootstrapping. RSB produces modest but consistent improvements over SSB, whereas SRB yields efficiencies close to 100%.

Under residual bootstrapping, the differences between RSB and SRB become more apparent. The probability-based prediction strategy substantially improves SRB performance across all set sizes, while RSB produces moderate gains. Compared with the linear regression results, the efficiency gains are smaller, and the variability across replications is larger. This is expected because additive spline models estimate substantially more parameters from the same sample size, making the prediction problem inherently more difficult. Nevertheless, the proposed RSS-based bagging methods continue to outperform classical SSB in most settings.

4.8 Real Data: Body Fat

To conclude Chapter 4, we illustrate the proposed ranked-enriched bagging methods using a real dataset to predict the percentage of body fat (BF) in adults, an important indicator of obesity. Accurate measurement of body fat percentage often requires specialized techniques, such as hydrodensitometry (underwater weighing), which rely on sophisticated equipment typically available only in clinical or research settings [Johnson \(1996\)](#). Developing reliable predictive models using readily available covariates, together with a limited number of body fat measurements, is of considerable practical importance [Johnson \(1996\)](#).

In this setting, the proposed RSB and SRB frameworks are expected to yield more accurate and stable bagged predictions than SSB based on SRS. By incorporating

ranking information into the resampling and prediction processes, the proposed method leverages the statistical advantages, yielding more informative bootstrap samples and improved prediction accuracy for body fat percentage.

The dataset used in this analysis is publicly available through Kaggle ([Johnson and Zappatini, 2023](#)). It combines data from two independent studies on body fat estimation, comprising 252 male participants [Johnson \(1996\)](#) and 184 female participants [Johnson \(2021\)](#), for a total of 436 individuals. In both studies, body fat percentage was estimated from body density measured by hydrodensitometry and converted using the equation proposed by [Siri \(1956\)](#), a widely accepted and accurate method for assessing body fat. The dataset contains 15 variables, including body fat percentage (%BF), sex (M/F), weight (kg), age (years), height (m), and 10 anthropometric circumference measurements (cm). Although direct measurement of body fat is relatively costly and requires specialized equipment, the remaining variables can be obtained quickly and inexpensively using standard equipment such as a scale and a tape measure.

The body fat dataset has been widely used in the statistical literature as a real-data benchmark for evaluating regression and resampling methodologies. [Rabbi et al. \(2021\)](#) used the dataset to illustrate a paired bootstrap model selection criterion, while [Norouzirad et al. \(2018\)](#) applied it to weighted least absolute deviations regression models. In addition, [Terpstra and Wang \(2008\)](#) employed the dataset to investigate confidence intervals for a single population proportion under RSS, using body mass index (BMI), computed from an individual's height and weight, as the ranking variable.

In this section, we treat the BF dataset as the finite population. Our objective is to develop predictive models to estimate body fat percentage from available covariates. As discussed previously, ranking information may be obtained from various sources. In this analysis, we use five empirical equations reported in the literature to estimate %BF from BMI, age, weight, and gender. The resulting estimated %BF serves as a ranking

variables.

- (i) [Deurenberg et al. \(1991\)](#) developed the following equation using data from a Dutch cohort of 521 men and 708 women aged 7–83 years

$$\%BF = 1.20 \text{ BMI} + 0.23 \text{ age} - 10.8 \text{ gender} - 5.4,$$

where gender = 1 for men and 0 for women.

- (ii) [Martarelli et al. \(2008\)](#) proposed separate equations for men and women based on an Italian cohort of 422 men and 341 women aged 11–80 years

$$\text{Men: } \%BF = \frac{1.407 \text{ BMI} - 21.389}{\text{weight}} \times 100,$$

$$\text{Women: } \%BF = \frac{1.9337 \text{ BMI} - 26.422}{\text{weight}} \times 100.$$

- (iii) The following body fat prediction equation, attributed to Gallagher et al. and reported by [Bennouar et al. \(2023\)](#), was derived from data collected on an American cohort of 214 men and 290 women aged 20–94 years

$$\%BF = 1.506 \text{ BMI} + 0.133 \text{ age} - 11.481 \text{ gender} - 11.52,$$

where gender = 1 for men and 0 for women.

- (iv) [Jackson et al. \(1980\)](#) derived another equation from an American cohort of 402 men and 283 women aged 19–43 years

$$\%BF = 1.61 \text{ BMI} + 0.13 \text{ age} - 12.1 \text{ gender} - 13.9,$$

where gender = 1 for men and 0 for women.

(v) [Gómez-Ambrosi et al. \(2012\)](#) proposed the following equation using a sample of 6,510 white individuals (2,154 men and 4,356 women) aged 18–80 years

$$\begin{aligned}\%BF = & -44.988 + 0.503 \text{ age} + 10.689 \text{ sex} + 3.172 \text{ BMI} - 0.026 \text{ BMI}^2 \\ & + 0.181 \text{ BMI} \times \text{sex} - 0.020 \text{ BMI} \times \text{age} \\ & - 0.005 \text{ BMI}^2 \times \text{sex} + 0.00021 \text{ BMI}^2 \times \text{age},\end{aligned}$$

where $\text{sex} = 0$ for men and 1 for female.

As each above equation was derived from data corresponding to a specific age range, the appropriate equation is applied only within its valid age interval to generate ranks. The ranking variables were generated from the participants' age, sex, weight, height, and BMI; these variables were not included as covariates in the predictive models. The remaining 10 anthropometric measurements, neck, chest, abdomen, hip, thigh, knee, ankle, bicep, forearm, and wrist circumferences, were used as predictor variables, while body fat percentage %BF served as the response variable. The weighting mechanism presented in (2.24) was used to assign weights to each rank based on the correlation between the ranking variable and %BF. The distributions of %BF for male and female participants are presented in Figure 4.7.

Few observations were identified as data-entry errors and removed before analysis. For example, based on the BMI distribution, two extreme outliers corresponding to observations 39 and 42 were excluded [Loewen and Jafari Jozani \(2026\)](#). Similarly, observation 372 was removed because it contained implausible values, with the participant's age recorded as 1 year and height recorded as 1.64 m. Outliers were identified using the interquartile range (IQR) criterion, considering the covariates listed above, and removed before analysis.

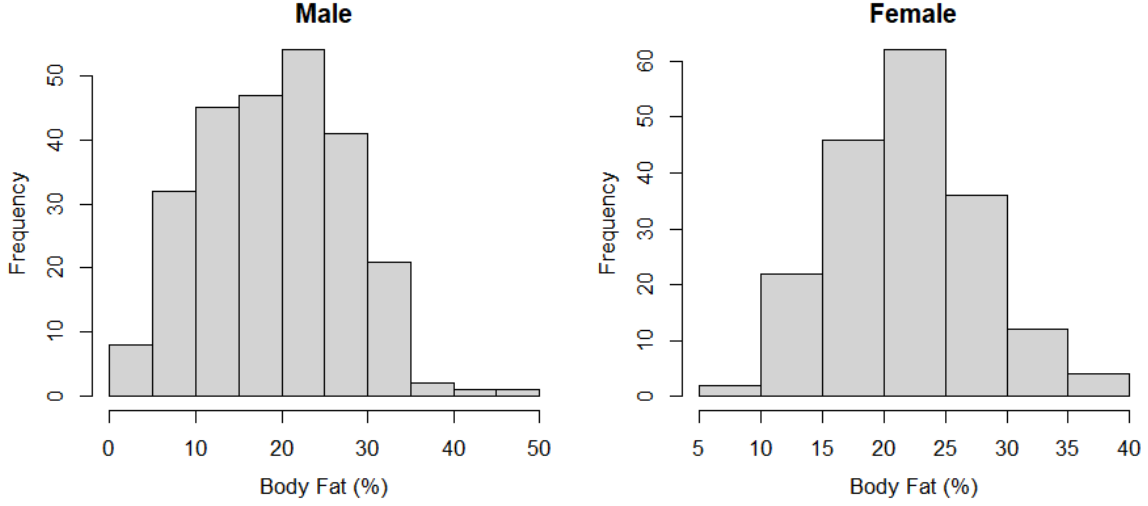


Figure 4.7: Distribution of body fat (%) for 252 males (left) and 184 females (right).

We consider a multiple linear regression model and randomly generate SRS and RSS training samples of size $n = 120$, together with an independent test sample of the same size. For the RSS design, balanced ranked set samples are constructed using set sizes $k \in \{2, 3, 4, 5, 6\}$ with cycle size $m = n/k$. For each setting, $B = 1000$ bootstrap samples are generated under both paired and residual bootstrapping, and the corresponding bagged models are fitted. The entire experiment is then independently repeated six times.

LASSO-regularized fitted models are used to predict %BF for the test observations using the two prediction strategies introduced in Section 4.3: (i) the maximum-rank, and (ii) the probability-based approach. Model performance is then evaluated using the relative efficiencies, RE and RE_{pred} , of the proposed RSB and SRB methods relative to the classical SSB. For the adjusted-response approach, the average confusion matrix $\mathbb{P}_k^{(\ell)}$, described in Section 2.8, is estimated from the initial SRS training sample used to construct the SRB procedure.

Table 4.8 summarizes the average estimation efficiencies obtained from the training samples over the six replications. Whenever out-of-bag (OOB) observations are available,

they are used to compute the training error, providing an unbiased assessment of estimation performance. Under paired bootstrapping, both RSB and SRB consistently outperform the classical SSB procedure, with nearly all relative efficiencies exceeding 100%. The SRB procedure consistently achieves larger efficiency gains than RSB, particularly under the adjusted-response approach, while the LMM framework yields the highest efficiencies for both methods. Although the improvement does not increase monotonically with set size under imperfect ranking, the overall results demonstrate that the proposed methods remain effective even when the ranking information is subject to error.

Table 4.8: Relative efficiencies (RE) of the LASSO-regularized fitted multiple linear regression models for estimating %BF using the proposed RSB and SRB procedures relative to the SSB. Results are based on training samples of size $n = 120$ under paired and residual bootstrapping with set sizes $k \in \{2, 3, 4, 5, 6\}$ and $B = 1000$ bootstrap replicates. “Adjusted” denotes weighted least squares with adjusted-response, while “LMM” refers to the linear mixed model approach. Reported values are relative efficiencies (mean $\pm$ sd) over six independent replications.

Bootstrap	Set size	RSB		SRB	
		Adjusted	LMM	Adjusted	LMM
Paired	2	118.02 $\pm$ 13.88	120.47 $\pm$ 17.57	127.17 $\pm$ 6.70	141.25 $\pm$ 10.52
	3	120.68 $\pm$ 19.04	124.80 $\pm$ 26.65	132.91 $\pm$ 7.83	146.81 $\pm$ 11.46
	4	123.38 $\pm$ 14.57	129.04 $\pm$ 20.31	134.61 $\pm$ 7.75	151.47 $\pm$ 12.26
	5	130.43 $\pm$ 24.49	139.11 $\pm$ 23.52	134.67 $\pm$ 7.34	154.62 $\pm$ 12.22
	6	111.27 $\pm$ 10.21	120.56 $\pm$ 14.58	133.76 $\pm$ 7.47	157.30 $\pm$ 12.99
Residual	2	103.89 $\pm$ 5.68	126.00 $\pm$ 16.11	72.14 $\pm$ 3.51	78.26 $\pm$ 2.18
	3	101.61 $\pm$ 12.65	132.97 $\pm$ 23.74	74.01 $\pm$ 3.60	69.86 $\pm$ 2.06
	4	100.16 $\pm$ 13.90	138.27 $\pm$ 19.90	75.45 $\pm$ 3.64	65.67 $\pm$ 2.09
	5	104.46 $\pm$ 19.92	151.64 $\pm$ 22.00	76.74 $\pm$ 3.51	63.48 $\pm$ 2.06
	6	90.17 $\pm$ 9.04	131.63 $\pm$ 15.64	77.78 $\pm$ 3.20	61.52 $\pm$ 2.07

However, residual bootstrap exhibits a different training pattern. For the residual bootstrapping adjusted-response method, RSB provides modest but consistent improve-

ments over SSB, while SRB exhibits efficiencies below 100% for training data, indicating that estimating the ranking information introduces additional variability during model fitting. However, this reduction in training efficiency does not carry over to the prediction results. In contrast, the LMM approach substantially improves performance for both RSB and SRB, with higher efficiency gains.

Table 4.9 reports the average prediction efficiencies evaluated on the independent test samples across the six replications. Previous studies identified approximately 2.5% of the BF observations as influential according to Cook’s distance [Loewen and Jafari Jozani \(2026\)](#), suggesting that the dataset presents a challenging prediction problem for conventional OLS regression.

With paired bootstrapping, both RSB and SRB outperform the SSB across all set sizes. The adjusted-response method provides consistent improvements over the SSB, while the LMM framework produces the largest gains for both RSB and SRB. LLM efficiency gains generally increase with set size. The probability-based prediction strategy performs comparably to the maximum-rank approach, although neither strategy completely dominates across all settings. Overall, these results demonstrate that both prediction strategies provide reliable improvements over the SSB.

Table 4.9: Relative efficiencies (RE_{pred}) of the LASSO-regularized fitted multiple linear regression models for estimating %BF using the proposed RSB and SRB procedures relative to the SSB. Results are based on a test sample of size $n = 120$ under paired and residual bootstrapping with set sizes $k \in \{2, 3, 4, 5, 6\}$ and $B = 1000$ bootstrap replicates. “Adjusted” denotes weighted least squares with adjusted response, while “LMM” refers to the linear mixed model approach. Reported values are the (mean $\pm$ sd) relative efficiencies over six independent replications.

Paired Bootstrap					
Method	Set Size	Adjusted		LMM	
		Rank	Prob.	Rank	Prob.
RSB	2	112.86 $\pm$ 15.82	113.22 $\pm$ 16.00	114.72 $\pm$ 20.53	112.59 $\pm$ 19.05
	3	122.35 $\pm$ 14.12	119.70 $\pm$ 14.99	127.92 $\pm$ 20.77	126.95 $\pm$ 20.94
	4	118.07 $\pm$ 11.39	115.06 $\pm$ 12.39	121.99 $\pm$ 20.19	120.14 $\pm$ 22.61
	5	115.61 $\pm$ 9.94	113.24 $\pm$ 9.59	119.82 $\pm$ 18.99	117.54 $\pm$ 18.85
	6	117.97 $\pm$ 16.37	115.18 $\pm$ 16.27	123.65 $\pm$ 17.05	122.09 $\pm$ 18.64
SRB	2	114.04 $\pm$ 7.82	115.57 $\pm$ 2.65	119.64 $\pm$ 16.61	116.84 $\pm$ 14.72
	3	119.02 $\pm$ 6.15	117.09 $\pm$ 3.14	126.64 $\pm$ 15.09	122.78 $\pm$ 14.16
	4	118.22 $\pm$ 3.93	115.44 $\pm$ 2.46	127.68 $\pm$ 13.08	125.38 $\pm$ 14.35
	5	116.92 $\pm$ 4.00	115.04 $\pm$ 3.40	130.32 $\pm$ 15.03	128.51 $\pm$ 14.15
	6	115.96 $\pm$ 3.68	113.48 $\pm$ 2.02	135.57 $\pm$ 18.99	131.06 $\pm$ 15.40
Residual Bootstrap					
Method	Set Size	Adjusted		LMM	
		Rank	Prob.	Rank	Prob.
RSB	2	127.02 $\pm$ 34.39	127.87 $\pm$ 34.65	143.50 $\pm$ 41.84	140.99 $\pm$ 40.13
	3	127.86 $\pm$ 22.90	126.07 $\pm$ 21.84	158.45 $\pm$ 39.54	157.22 $\pm$ 38.16
	4	122.86 $\pm$ 14.41	119.77 $\pm$ 13.97	150.79 $\pm$ 39.62	148.78 $\pm$ 41.87
	5	123.13 $\pm$ 15.95	121.09 $\pm$ 16.54	146.51 $\pm$ 36.44	144.15 $\pm$ 36.07
	6	124.43 $\pm$ 21.43	122.28 $\pm$ 21.03	151.07 $\pm$ 36.10	149.87 $\pm$ 36.81
SRB	2	100.44 $\pm$ 9.62	103.87 $\pm$ 4.77	122.20 $\pm$ 17.17	136.49 $\pm$ 11.33
	3	104.02 $\pm$ 8.68	104.56 $\pm$ 6.46	120.53 $\pm$ 15.71	137.48 $\pm$ 12.69
	4	105.20 $\pm$ 5.79	104.04 $\pm$ 4.80	115.59 $\pm$ 8.20	132.59 $\pm$ 10.30
	5	104.75 $\pm$ 5.91	104.31 $\pm$ 4.44	113.56 $\pm$ 11.40	132.28 $\pm$ 9.90
	6	105.38 $\pm$ 6.12	104.44 $\pm$ 4.31	113.84 $\pm$ 12.73	131.92 $\pm$ 9.92

The performance of RSB under residual bootstrapping remains consistently strong across all set sizes. Both adjusted-response and LMM approaches achieve efficiencies well above 100%. LMM framework provides the largest improvements, demonstrating that residual-based ranked bagging remains highly effective when the ranking information is available for multi-level modeling. SRB procedure also performs favorably under residual bootstrapping. Unlike the training results, both the adjusted-response and LMM approaches now achieve prediction efficiencies exceeding the SSB across all set sizes. Although the adjusted-response method yields more modest gains, the LMM framework consistently achieves higher gains. The probability strategy generally performs similarly to the maximum-rank for the adjusted-response method, while providing noticeable improvements under the LMM. Despite the presence of influential points, both the RSB and SRB procedures continue to improve predictive performance over SSB, with LMM providing the most considerable improvement. This result suggests that incorporating ranking information into the bagging process remains beneficial for prediction even under challenging real-data conditions.

4.9 Summary and Conclusion

This chapter presented the simulation framework and resampling procedures used to evaluate the proposed rank-enriched bagging methods. We first described the construction of balanced RSS samples for the single-ranker setting, then extended the framework to accommodate multiple ranking variables through the proposed weighted RSS sampling scheme. Building on these sampling designs, we presented the paired and residual bootstrap procedures for generating bootstrap samples from RSS data.

The chapter then introduced the proposed SRS-RSS construction, which retrospectively converts an SRS training sample into an RSS representation. For paired bootstrapping, we described the direct SRS-RSS bootstrap procedure. For residual bootstrapping, we presented two approaches: a direct extension of the conventional

RSS residual bootstrap and the proposed rank-informed residual bootstrap, which incorporates ordered residuals to construct rank-informed residual bootstrap samples.

Finally, we described the overall simulation design used throughout this thesis, including the linear and nonlinear regression models, the different covariate and ranking structures, and the performance measures used to compare the proposed RSB and SRB methods with the conventional SSB approach. To demonstrate the practical applicability of the proposed methods, the simulation study was complemented by a real-data prediction analysis using the BF dataset. Overall, these findings demonstrate the effectiveness of incorporating ranking information into the bagging framework. Although both RSB and SRB outperform the classical SSB procedure, SRB yields more stable performance across set sizes and, in several cases, matches or surpasses the efficiencies achieved by RSB.

Chapter 5

Conclusion and Future Work

5.1 Summary

This thesis investigated the integration of RSS into bagging for regression prediction. Classical bagging relies on SRS, but it does not use the auxiliary ranking information often available in practice. Therefore, the primary objective of this research was to develop rank-enriched bagging procedures that improve predictive accuracy while retaining the desirable theoretical properties of classical bagging.

Chapter 1 established the theoretical foundation of the thesis by reviewing ensemble learning, bootstrap resampling, and the paired and residual bootstrapping schemes used throughout the study. The classical SSB was then formulated for general mean-type functionals and extended to regression estimation through the regression function $m(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$. The unbiasedness and variance properties of SSB provided the theoretical benchmark for developing and evaluating the proposed methods.

Chapter 2 introduced the proposed RSB framework. Unlike conventional bagging, RSB generates bootstrap samples independently within each rank stratum, thereby preserving the ranked structure of the training data while exploiting the efficiency gains associated with RSS. Theoretical properties of RSB were first established for general mean-type functionals and then extended to regression estimators via the linear smoother

framework. Explicit expressions for the conditional and unconditional variances were derived, and the methodology was further extended to weighted least squares and linear mixed model estimation. To improve practical applicability, a weighted RSS construction was proposed to accommodate multiple ranking variables.

Although RSB provides substantial efficiency gains, its practical implementation requires RSS training samples, which are often unavailable. To address this limitation, Chapter 3 proposed the SRB framework. Rather than assuming an RSS training sample, SRB constructs RSS retrospectively from an initial SRS training sample by incorporating ranking information during resampling. Theoretical results showed that SRB remains unbiased and has lower variance than conventional SSB while closely approximating the efficiency of RSB. These results were extended to regression estimation through linear smoother representation, and practical prediction algorithms were developed for paired and residual bootstrapping schemes.

Chapter 4 described the simulation framework used to evaluate the proposed methodologies. Balanced RSS and weighted RSS sampling schemes were developed for single- and multiple-ranker settings. Corresponding paired and residual bootstrap algorithms were presented for RSB and SRB, including a new rank-informed residual bootstrap procedure for retrospectively constructed RSS samples. The simulation study considered a wide range of linear and nonlinear regression models, different ranking qualities, multiple covariate structures, and real-data applications based on the BF dataset.

The numerical studies demonstrated that both RSB and SRB consistently outperform the conventional SSB approach in terms of predictive accuracy. The largest improvements were observed when informative ranking variables were available, although both methods continued to provide meaningful gains under moderate and weak ranking conditions. Across a wide range of simulation settings, SRB achieved prediction accuracies comparable to RSB's despite requiring only an initial SRS training sample, highlighting its practical value. Furthermore, incorporating multiple ranking variables through the

proposed weighted RSS framework improved robustness by reducing the influence of weaker rankers. The probability-based prediction strategy also proved particularly effective under residual bootstrapping, especially when combined with the linear mixed model framework.

Overall, this thesis contributes both theoretical and methodological advances to the integration of RSS and ensemble learning. It establishes a unified theoretical framework for rank-enriched bagging, develops practical algorithms for settings with or without direct RSS samples, and demonstrates, through extensive simulation and real-data studies, that incorporating ranking information into the bagging process can substantially improve predictive performance compared with conventional SRS-based bagging. These results provide a foundation for future extensions of rank-enriched bagging to broader regression settings and other predictive learning problems, including classification and modern machine learning methods.

5.2 Future Work

As discussed in Chapter 3, one primary limitation of RSS is its requirement for ranking information. Although the proposed SRB approach can use partially available ranking information from an SRS training sample, there are many practical situations in which even partial ranking information is unavailable. Therefore, to overcome this limitation, future research will focus on developing data-driven ranking mechanisms that infer ranking from training data. Such approaches would broaden the applicability of the proposed methodology by eliminating dependence on external ranking sources.

This thesis primarily focuses on regression problems. However, many real-world applications involve classification rather than continuous response variables. Accordingly, an important direction for future research is to extend the proposed RSB and SRB frameworks to binary and multiclass classification problems by incorporating the structural properties of RSS directly into the classification modeling process. Although

recent studies have explored RSS in classification, they have largely treated it as a sampling strategy to obtain more informative and representative samples (Wang et al., 2023). In contrast, following the regression framework developed in this thesis, we plan to integrate rank information throughout the classification modeling and ensemble learning procedures, rather than viewing RSS merely as a data collection strategy. This extension has the potential to further improve classification accuracy and predictive performance.

Beyond traditional statistical modeling, the proposed RSS framework has the potential to benefit several emerging areas of artificial intelligence and large language model (LLM) training. Future research will explore applications of the proposed methods in the following directions:

- (i) **Reinforcement Learning from Human Feedback (RLHF):** Human preference data are naturally expressed through rankings. Embedding ranking information into the sampling and estimation procedures used in policy optimization may reduce estimation variance and lead to more stable learning of the reward model.
- (ii) **Retrieval-Augmented Generation (RAG):** Retrieved documents naturally possess varying degrees of relevance and reliability. Extending the core idea of embedding ranking information into resampling and estimation procedures may provide more effective weighting of retrieved contexts, potentially improving retrieval quality, enhancing prediction accuracy, and reducing hallucinations in large language models.
- (iii) **Parameter-Efficient Fine-Tuning (PEFT), including Low-Rank Adaptation (LoRA):** Fine-tuning large language models is often constrained by limited labeled data. By retrospectively ranking training examples and applying the proposed SRB framework, it may be possible to extract more informative training

signals from small datasets, thereby reducing the variance of parameter-efficient fine-tuning methods such as LoRA and improving model generalization.

Overall, these directions demonstrate that the proposed ranking-based bagging framework has the potential to extend well beyond classical regression problems, offering promising opportunities in modern statistical learning, machine learning, and large-scale foundation model training.

Bibliography

- Adadi, A. (2021). A survey on data-efficient algorithms in big data era. *Journal of Big Data* 8(1), 24. [1](#)
- Al Kadiri, M. (2017). Linear penalized spline model estimation using ranked set sampling technique. *Hacettepe Journal of Mathematics and Statistics* 46(4), 669–683. [29](#), [100](#)
- Algarni, A. (2016). Median and extreme ranked set sampling for penalized spline estimation. *Applied Mathematics & Information Sciences* 10(1), 243. [29](#), [100](#)
- Aliferis, C. and G. Simon (2024). Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and ai. *Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls*, 477–524. [2](#)
- Alzubaidi, L., J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-Dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy, et al. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data* 10(1), 46. [1](#)
- Asghari, S., B. Sadeghpour Gildeh, J. Ahmadi, and G. Mohtashami Borzadaran (2018). Sign control chart based on ranked set sampling. *Quality Technology & Quantitative Management* 15(5), 568–588. [29](#)

- Barreto, M. C. M. and V. Barnett (1999). Best linear unbiased estimators for the simple linear regression model using ranked set sampling. *Environmental and Ecological Statistics* 6(2), 119–133. 5, 29
- Bennouar, S., A. B. Cherif, H. M. Hani, A. Kerrouche, and S. Abdi (2023). Prediction of body fat percentage: Development and validation of new anthropometric equations. *Clinical nutrition ESPEN* 57, 510–518. 113
- Berk, R. A. et al. (2008). *Statistical learning from a regression perspective*, Volume 14. Springer. 15
- Bohn, L. L. and D. A. Wolfe (1992). Nonparametric two-sample procedures for ranked-set samples data. *Journal of the American Statistical Association* 87(418), 552–561. 29
- Bohn, L. L. and D. A. Wolfe (1994). The effect of imperfect judgment rankings on properties of procedures based on the ranked-set samples analog of the mann-whitney-wilcoxon statistic. *Journal of the American Statistical Association* 89(425), 168–176. 29, 32
- Breiman, L. (1996). Bagging predictors. *Machine learning* 24(2), 123–140. 3, 7
- Bühlmann, P. and B. Yu (2002). Analyzing bagging. *The annals of Statistics* 30(4), 927–961. 3
- Chen, Z. (1999). Density estimation using ranked-set sampling data. *Environmental and Ecological Statistics* 6(2), 135–146. 4
- Chen, Z., B. Zhidong, and B. K. Sinha (2003). *Ranked set sampling: theory and applications*, Volume 176. Springer. 33
- Davison, A. C. and D. V. Hinkley (1997). *Bootstrap methods and their application*. Number 1. Cambridge university press. 8

- Dell, T. R. and J. Clutter (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 545–555. [31](#), [32](#), [92](#), [100](#)
- Deurenberg, P., J. A. Weststrate, and J. C. Seidell (1991). Body mass index as a measure of body fatness: age-and sex-specific prediction formulas. *British journal of nutrition* 65(2), 105–114. [113](#)
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer. [2](#)
- Efron, B. (1979, January). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* 7(1), 1–26. [8](#)
- Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC. [3](#)
- Faraji, N., M. Jafari Jozani, and N. Nematollahi (2022). Another look at regression analysis using ranked set samples with application to an osteoporosis study. *Biometrics* 78(4), 1489–1502. [4](#), [29](#), [46](#), [49](#), [50](#), [52](#), [92](#), [100](#)
- Fisher, R. A. (1949). The design of experiments. [8](#)
- Freedman, D. A. (1981). Bootstrapping regression models. *The annals of statistics*, 1218–1228. [8](#)
- Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55(1), 119–139. [7](#)
- Gómez-Ambrosi, J., C. Silva, V. Catalán, A. Rodríguez, J. C. Galofré, J. Escalada, V. Valentí, F. Rotellar, S. Romero, B. Ramírez, et al. (2012). Clinical usefulness of a new equation for estimating body fat. *Diabetes care* 35(2), 383–388. [113](#)

- Hansen, L. K. and P. Salamon (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12(10), 993–1001. 7
- Hastie, T., R. Tibshirani, J. Friedman, et al. (2009). The elements of statistical learning. 2
- Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika* 57(1), 97–109. 3
- Hatefi, A. and M. Jafari Jozani (2013). Fisher information in different types of perfect and imperfect ranked set samples from finite mixture models. *Journal of Multivariate Analysis* 119, 16–31. 4
- Hatefi, A. and M. Jafari Jozani (2017). An improved procedure for estimation of malignant breast cancer prevalence using partially rank ordered set samples with multiple concomitants. *Statistical methods in medical research* 26(6), 2552–2566. 48
- Jackson, A. S., M. L. Pollock, and A. Ward (1980). Generalized equations for predicting body density of women. *Medicine and science in sports and exercise* 12(3), 175–181. 113
- Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G. E. Hinton (1991). Adaptive mixtures of local experts. *Neural computation* 3(1), 79–87. 7
- Jafari Jozani, M. and S. J. Mirkamali (2010). Improved attribute acceptance sampling plans based on maxima nomination sampling. *Journal of Statistical Planning and Inference* 140(9), 2448–2460. 29
- Jafari Jozani, M. and S. J. Mirkamali (2011). Control charts for attributes with maxima nominated samples. *Journal of Statistical Planning and Inference* 141(7), 2386–2398. 29

- Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education* 4(1). 111, 112
- Johnson, R. W. (2021). Fitting percentage of body fat to simple body measurements: College women. *Journal of Statistics and Data Science Education* 29(3), 304–316. 112
- Johnson, R. W. and S. Zappatini (2023). Body fat extended dataset. Accessed: 2026-06-29. 112
- Jordan, M. I. and R. A. Jacobs (1994). Hierarchical mixtures of experts and the em algorithm. *Neural computation* 6(2), 181–214. 7
- Kuncheva, L. I. (2014). *Combining Pattern Classifiers: Methods and Algorithms* (2 ed.). Hoboken, NJ: John Wiley & Sons. 6
- Kuno, E. (1976). Multi-stage sampling for population estimation. *Population Ecology* 18(1), 39–56. 3
- Le, V. H., S. L. Collins, and R. Vargas (2025). Long-term ecological studies must continue: insights from a dryland transition zone. *Oecologia* 207(11), 1–12. 2
- Loewen, N. and M. Jafari Jozani (2026). Leveraging rank information for robust regression analysis: A nomination sampling approach. *Statistics in Medicine* 45(1-2), e70362. 114, 117
- MacEachern, S. N., E. A. Stasny, and D. A. Wolfe (2004). Judgement post-stratification with imprecise rankings. *Biometrics* 60(1), 207–215. 5, 85
- Madow, W. G. and L. H. Madow (1944). On the theory of systematic sampling, i. *The Annals of Mathematical Statistics* 15(1), 1–24. 3

- Martarelli, D., B. Martarelli, and P. Pompei (2008). Body composition obtained from the body mass index: an italian study. *European Journal of Nutrition* 47(8), 409–416. 113
- McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets. *Australian journal of agricultural research* 3(4), 385–390. 3, 27, 28
- Murff, E. J. T. and T. W. Sager (2006). The relative efficiency of ranked set sampling in ordinary least squares regression. *Environmental and Ecological Statistics* 13(1), 41–51. 29
- Muttlak, H. and W. Al-Sabah (2003). Statistical quality control based on ranked set sampling. *Journal of Applied Statistics* 30(9), 1055–1078. 29
- Muttlak, H. A. (1995). Parameters estimation in a simple linear regression using rank set sampling. *Biometrical Journal* 37(7), 799–810. 5, 29, 49, 100
- Muttlak, H. A. (1996). Estimation of parameters in a multiple regression model using rank set sampling. *Journal of Information and Optimization Sciences* 17(3), 521–533. 5
- Neyman, J. (1992). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Breakthroughs in statistics: Methodology and distribution*, pp. 123–150. Springer. 3
- Norouzirad, M., S. Hossain, and M. Arashi (2018). Shrinkage and penalized estimators in weighted least absolute deviations regression models. *Journal of Statistical Computation and Simulation* 88(8), 1557–1575. 112
- Nourmohammadi, M., M. Jafari Jozani, and B. C. Johnson (2015). Distribution-free tolerance intervals with nomination samples: Applications to mercury contamination in fish. *Statistical Methodology* 26, 16–33. 47

- Öztürk, Ö. (2002). Ranked set sample inference under a symmetry restriction. *Journal of statistical planning and inference* 102(2), 317–336. 4, 29
- Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society* 4(1), 119–130. 8
- Quenouille, M. H. (1949). Problems in plane sampling. *The annals of mathematical statistics*, 355–375. 8
- Rabbi, F., S. Khan, A. Khalil, W. K. Mashwani, M. Shafiq, P. Göktaş, and Y. A. Unvan (2021). Model selection in linear regression using paired bootstrap. *Communications in Statistics-Theory and Methods* 50(7), 1629–1639. 112
- Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric regression*. Number 12. Cambridge university press. 53
- Schapire, R. E. (1990). The strength of weak learnability. *Machine learning* 5(2), 197–227. 7
- Shaikhina, T. and N. A. Khovanova (2017). Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial intelligence in medicine* 75, 51–63. 2
- Siri, W. E. (1956). The gross composition of the body. In *Advances in biological and medical physics*, Volume 4, pp. 239–280. Elsevier. 112
- Stokes, S. L. (1980). Estimation of variance using judgment ordered ranked set samples. *Biometrics*, 35–42. 51
- Stokes, S. L. and T. W. Sager (1988). Characterization of a ranked-set sample with application to estimating distribution functions. *Journal of the American Statistical Association* 83(402), 374–381. 4, 29

- Terpstra, J. T. and P. Wang (2008). Confidence intervals for a population proportion based on a ranked set sample. *Journal of Statistical Computation and Simulation* 78(4), 351–366. [112](#)
- Tibshirani, R. J. and B. Efron (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability* 57(1), 1–436. [3](#), [8](#), [10](#)
- Tiku, M. L. and A. D. Akkaya (2004). *Robust estimation and hypothesis testing*. New Age International. [46](#), [51](#)
- Wang, J., F. Li, J. Li, C. Hou, Y. Qian, and J. Liang (2023). Rss-bagging: Improving generalization through the fisher information of training data. *IEEE Transactions on Neural Networks and Learning Systems* 36(2), 1974–1988. [29](#), [124](#)
- Wang, S. and X. Yao (2009). Diversity analysis on imbalanced data sets by using ensemble models. In *2009 IEEE symposium on computational intelligence and data mining*, pp. 324–331. IEEE. [2](#)
- Wang, Y.-R. and Y.-C. Tsai (2022). The protection of data sharing for privacy in financial vision. *Applied Sciences* 12(15), 7408. [2](#)
- Wolfe, D. A. (2004, November). Ranked set sampling: An approach to more efficient data collection. *Statistical Science* 19(4), 636–643. [4](#), [28](#)
- Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of physics: Conference series*, Volume 1168, pp. 022022. IOP Publishing. [1](#)
- Zhang, C. (2023). *Incorporating Expert Knowledge into Rule Learning via Reinforcement Learning*. University of Washington. [4](#)
- Zhou, Z.-H. (2025). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC. [6](#)