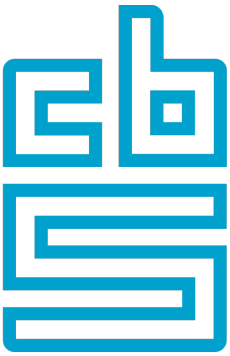
Regularizing Probability Sample Estimates Through an Angle-Based Similarity Approach



Jacob Westlund (s2135248)
s2135248@vuw.leidenuniv.nl

Thesis advisors CBS:
Prof. T. de Waal
Dr. S. Scholtus

Thesis advisor Universiteit Leiden:
Prof. E. McCormick

MASTER THESIS
Defended on August 1st, 2024
STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Abstract

Recent decades have seen the cost of acquiring probability survey samples (PSs) increase in tandem with a general decline in public survey participation. As a consequence, National Statistical Institutes have become increasingly interested in using register data (among others) to substitute or supplement their current PSs, since register data is both much cheaper and comes in large quantities. However, register data can be a type of non-probability sample (NPS), which means it is often selective and has an unknown sampling design. As a result, estimates derived from such samples tend to be biased, often preventing their direct integration into official statistics. Already, much research has gone into trying to either de-bias NPS estimates or find ways to leverage their information by combining them with estimates from a smaller PS. These solutions have however only been moderately successful, seeing either limited use, limited improvement, or relying on unreasonable assumptions.

The main objective of this thesis was to apply the Angel-Based Transfer Learning Estimator (ABTLE), a type of penalized regression estimator that leverages the angle similarity between estimates from two related data sources. The hope was to apply this estimator to a small unbiased PS and a much larger biased NPS, to leverage the stability of the NPS, whilst maintaining the unbiased characteristic of the PS, to improve the overall accuracy of the estimates. The ABTLE was applied to simulated data with varying population and sample characteristics. It was evaluated through a comparison of the average root mean squared error and the mean of average bias against the maximum likelihood estimator using just the PS, and other related estimators.

The analysis found that the ABTLE can significantly reduce the average root mean squared error of the estimates compared to just using the PS, while only incorporating a fraction of the bias of the NPS estimates. The relative usefulness of ABTLE was higher as the accuracy of the PS estimates decreased. If the PS estimates were already quite accurate, the marginal improvements were small and even sometimes resulted in deterioration due to a poor estimation of the penalty parameters. The findings imply that the ABTLE can be an efficient way of incorporating an NPS into official statistics. However, more research is needed on how to better estimate the penalty parameters to fully utilize its potential, and it should not be applied without due consideration to the available sample size

Keywords: Data integration, Penalized regression, Ridge regression, Probability samples, Non-probability samples.

Acknowledgements

I would like to express my appreciation to my supervisors, Prof. T. de Waal and Dr. S. Scholtus from the Centraal Bureau voor de Statistiek (CBS), for their expert guidance and constructive feedback throughout the development of this thesis. Their knowledge and experience have helped me greatly, always making time for weekly meetings or just during the day whenever I was stuck and needed help. Their suggestions were instrumental in the design of this thesis, and I am very happy to have been allowed to work on one of their projects.

I am also thankful to Prof. E. McCormick from Leiden University for his valuable insights and academic advice, which have been vital in refining this thesis. Throughout every stage of the thesis, he has given me important pointers on how to improve, both in terms of the subject matter and the writing style. Moreover, Prof. McCormick's encouragement for me to start writing early is in hindsight also greatly appreciated as I initially, strongly underestimated the complexity of writing a thesis. Had he not told me to start early with the actual writing process, I fear that the final weeks of this thesis would have been incredibly stressful.

Contents

List of Figures

1. Introduction

1.1 Background

Official statistics is not a modern concept, with countries worldwide having utilized statistics to monitor their subjects through various bureaucratic institutions since the emergence of the centralized state (Bethlehem, 2009a). Although the concept of official statistics is ancient, the methods by which it is conducted are not. Especially the shape and size of the survey sample have undergone a radical transformation following the development of sampling theory as a modern field of study and the increased needs of society.

For most of the history of official statistics, estimating population characteristics from a sample of the population was not a known method. Thus, the only method available to infer anything from the population was through a complete census. The problem with that, however, is that although censuses are a very comprehensive and accurate method of creating statistics, they come with the significant downside of being extremely costly and time-consuming. Therefore, relying only on censuses prevented a frequent flow of information and limited the breadth of topics that could be examined at any one time. Nevertheless, lacking a better method and a demand for a wider breadth of statistics, sporadic yet comprehensive censuses were the method utilized for centuries. The one exception to this was a few flawed attempts at so-called partial investigations, which were selective uses of early sampling theory to attempt to infer population characteristics from parts of a population (Bethlehem, 2009a).

It was not until 1895 that the dominance of censuses in official statistics started to be contested. In what can be regarded as the emergence of modern sampling theory, the Norwegian statistician Anders Kiaer published the theory of the representative method. He had devised a method by which a small sample, specifically designed to be representative of the wider population, could be used to accurately estimate population characteristics, negating the need for a full census. In 1926 Arthur Bowley built on this idea of sampling, proposing the idea of probability sampling where every case in the population has a random but equal probability of being included in a sample. He showed that if this is the case, estimates from a random probability sample (PS) would be approximately unbiased. The estimates would also come with their own valid uncertainty estimates, something that could not be done using the representative method. These two methods existed side by side until 1934 when Jerzy Neyman compared the two methods and showed that representative sampling performed worse than probability sampling and that it would only be effective under very restrictive conditions (Seng, 1951; Smith, 1976). This solidified probability sampling as the superior method over representative sampling, which was promptly discarded. Further work by Daniel Horwitz and Donovan Thompson in 1952 would show that for probability sampling equal inclusion probabilities are not a necessity, but that as long as accurate inclusion probabilities exist for the entire population, an unbiased estimate can always be created from a PS (Bethlehem, 2009b).

Despite the development and refinement of estimation using PSs, fundamentally using sam-

ples rather than censuses was initially resisted. Within national statistical institutes (NSIs), it was only applied limitedly as they feared that estimates would not be accurate enough, and that probability sampling would be too difficult to implement in practice (Bethlehem, 2009b). It was only after the modernization of society which saw a significant increase in the demand for statistics that estimation using PSs began to be seriously implemented by NSIs. Resource-constrained, they recognized the infeasibility of only continuing to rely on census-based statistics, which had become far too costly and slow to apply widely. Instead, they recognized the significantly lowered cost, time, and increased potential of probability sampling and quickly began shifting towards working primarily with PSs, despite the increased uncertainty they brought with them (Bethlehem, 2009b).

1.2 Problem statement

Today, estimation using PSs is ubiquitous within NSIs and the method has developed into the new gold standard due to its strong theoretical properties and relative ease of application. However, the increase in demand for more and faster statistics which led to the initial proliferation of estimation using PSs never stopped, and for resource-constrained NSIs, collecting quality PSs has now in turn become a relatively costly method of acquiring data (Bakker et al., 2014; van den Brakel, 2019). This has been further exacerbated in recent decades by survey response rates steadily declining, also affecting NSIs who now find it harder to contact and convince individuals to participate in their surveys (de Leeuw & de Heer, 2002; Luiten et al., 2020). For NSIs, such a decline has a twofold impact. First, non-response threatens the estimates' validity by increasing the risk of bias and potentially altering the selection mechanism outside the researcher's control (Bethlehem, 2009a). Second, and more practically, non-response also increases the work needed to reach the same sample sizes, which means that either the observed sample sizes will have to be reduced, or more resources need to be allocated to each probability survey sample (Luiten et al., 2020).

Because of this, there has been a growing interest within NSIs in replacing or supplementing the PSs with non-probability samples (NPSs), which are samples obtained outside the probability sampling framework, through for example register data from the tax office or other administrative authorities. This type of register-based NPS is very appealing since it is a very cheap method for acquiring large amounts of data, with no additional burden or requirements for respondents (van den Brakel, 2019). The hope is that by incorporating these new types of sources, NSIs can both improve the quality of their estimates and reduce their costs. However, like traditional PS estimates, these new NPS estimates are no silver bullet as they come with their own flaws. Most problematic is the fact that NPS comes without a known sampling design or inclusion probabilities, and that they tend to suffer from a combination of selectivity and coverage issues. These two factors combined mean that NPS estimates tend to be biased and, lacking a sampling frame, there is no way for NSIs to correct this bias. This is highly problematic for NSIs who place a large value on the unbiasedness of estimates, often preventing NPS estimates from being used directly in official statistics (Bakker et al., 2014).

Therefore, NSIs are currently facing a problem. Both statistically and financially, relying

purely on traditional PSs is becoming a prohibitively expensive and troublesome approach. However, transitioning to the alternative NPSs risks producing biased estimates, which is highly problematic for NSIs whose main purpose is to produce accurate descriptions of a country. Nevertheless, given the high potential of NPSs, the question is raised if there is not some sort of method by which the bias of NPS estimates could be mitigated, directly or indirectly, allowing it to be utilized in official statistics like probability sample estimates were less than a century ago.

The rest of the thesis is organized as follows: Section 2 begins with an overview of the literature on addressing selectivity and bias in NPS estimates, highlighting their theoretical contributions and situating this thesis within the broader academic context. Section 3 provides a general introduction to ridge regression, followed by a detailed description of the Angle-Based Transfer Learning Estimator and its benefits. This section concludes with a brief discussion on estimating penalty parameters and the importance of methods like cross-validation in this endeavor. Section 4 focuses on the practical process of simulation. It starts with an explanation of data simulation and sampling procedures. It then moves on to the analysis by describing the target model, the estimation process, and the evaluation method. Section 5 presents the results of the analysis, focusing on average root mean square error and mean of absolute bias. Section 6 discusses the results in relation to the research question and highlights the use and limitations of the Angle-Based Transfer Learning Estimator. Finally, section 7 outlines research limitations, suggests potential areas for future research, and briefly summarizes the thesis.

2. Literature review

Given the potential and thus increased importance of NPSs, several methods have already been developed attempting to deal with the selectivity problem present in most NPSs, to make its estimates acceptable for official statistics. These methods can broadly be classified into either a correction type approach or integration type approach. Here the defining separating characteristic is whether the method is used to correct the bias present in the NPS estimates, or if the method is used to integrate the biased NPS estimates with unbiased PS estimates, as to improve the PS estimates in some way.

2.1 Correction methods

Valliant (2020) and Valliant et al. (2018) outline three traditional correction methods, quasi-randomization (also known as propensity score adjustment), superpopulation modelling, and doubly robust estimation. These are broad methods that can vary in their applications via various sub-methods depending on the context. As such, this review will only briefly touch upon these methods and only go into detail if relevant to the research topic.

Quasi-randomization is a design-based method where lacking real inclusion probabilities, a PS is used as a surrogate to in some way estimate pseudo-inclusion probabilities for the NPS. Using a common set of auxiliary variables the PS can be used to estimate the assumed existing but unknown inclusion probability in the NPS. Alternatively, the PS can be used to statistically match each NPS observation to a PS observation, which has an attached inclusion probability that can then be "donated". These new pseudo-inclusion probabilities can then be used to re-weight the NPS estimates to reduce or even remove the bias. (Elliott & Valliant, 2017; Valliant et al., 2018). Superpopulation modeling breaks with the design-based approach and rather treats the NPS as just a sample from a theoretically infinite "superpopulation", where the outcome of interest follows some unknown probability distribution. The goal of superpopulation modeling is then, given a set of auxiliary variables that explain the selectivity, to use the NPS to model the relationship between the auxiliary variables and estimates of interest. Coefficients can then be extracted and applied to a wider population for population-level statistics that should account for the sample selectivity (Elliott & Valliant, 2017; Valliant et al., 2018). Finally, there is the doubly robust method, which is a combination of quasi-randomization and superpopulation modeling. First inclusion probabilities are estimated similarly to quasi-randomization, reweighing the NPS. The outcome of interest is then also modeled similarly to superpopulation modeling. The two methods are then integrated into one combined estimator by augmenting both components with their counterpart and adding them together. The assumptions of each component do not change, but the benefit of the doubly robust method is that as long as at least one of its components is correctly specified and its assumptions are not violated, the results should be unbiased. This holds true even if the other component is wrong. It is thus a more robust method than either of its parts but requires more data and is more complex to implement (Valliant, 2020).

These three correction methods can work (see for example Chen et al. (2020), Lee and

Valliant (2009), and Pan et al. (2022)) however, they do rely on similar and often practically problematic assumptions. As highlighted by Cornesse et al. (2020) and Valliant (2020) both quasi-randomization and superpopulation rely on modeling at some stage to correct the NPS selectivity. Although this is not problematic per se, the modeling stage requires that the NPS contains all relevant variables explaining its selectivity, something which is practically never the case in official statistics (van den Brakel, 2019). This opens up risks for model misspecification which in turn can reduce or even nullify any bias reduction of the NPS estimates. This reliance on modeling also means that neither method is possible should the NPS be "not missing at random", since then both the pseudo-inclusion probabilities and the estimates will be biased in turn (Elliott & Valliant, 2017; Valliant et al., 2018).

2.2 Integration methods

In response to these limitations, other researchers have taken a different approach to incorporating NPSs. Rather than utilizing the NPS estimates alone, several authors have attempted to harness the information from an NPS to improve the estimates of a related PS instead. Although the methods differ, the general idea is to maintain approximate unbiasedness of the combined estimates through the PS, whilst leveraging the potentially larger sample size of the NPS to reduce their variances.

One of the first attempts at this came from Elliott and Haviland (2007) looking at integrating web-based opt-in surveys into a probability sample estimate, through a composite estimator. Their method is based on mean squared error (MSE) where the relative contribution of the NPS to the estimate is determined by the ratio of the MSE from the probability sample to the total MSE from both probability and non-probability samples. The composite estimator was later refined by Villalobos Aliste (2022), who relaxed the assumption that the bias in the NPS estimates needs to be known beforehand, allowing for it to be estimated from the NPS and PS. Although effective in reducing MSE, the composite estimator is limited by the fact that it is only useful in scenarios where the NPS bias is quite small, and the PS is sizeable, two requirements that reduce its usefulness in practice.

Disogra et al. (2011) propose an estimation method called "blended calibration" where a calibrated PS is combined with an uncalibrated NPS. The combined sample is then calibrated again using differentiator variables from the PS alone, resulting in a final estimate from the combined sample. They showcase that this method can achieve a pretty substantial accuracy improvement over just the PS estimates with only limited bias. However, it should be noted that these results were achieved using a relatively large PS compared to the NPS. To the author's knowledge, no other research has attempted to apply their method on more unbalanced samples, limiting the generalizability of their findings.

Finally, Wiśniowski et al. (2020) takes a Bayesian approach to the incorporation problem, using an NPS to construct priors which are used to estimate a posterior distribution in combination with the PS. Conducting a simulation study they found that a "conjugate" prior performed the best since it led to a significant reduction in variance of posterior estimates and was most robust against bias stemming from the NPS selectivity. Nevertheless, the authors

do caution that regardless of the type of prior used, if the NPS estimates are sufficiently biased, the posterior distribution will be skewed. If the PS estimates are already quite accurate then any bias in the NPS prior could make the posterior estimates deteriorate relative to the PS estimates in terms of MSE.

2.3 Contribution to the literature

The above methods do have their use cases, yet none is designed for producing robust results in a scenario with an unbalanced sample distribution (large NPS and small PS), significant NPS selectivity, and a limited number of auxiliary variables. Correction methods are limited by the lack of auxiliary variables to correct for the NPS selectivity, whilst integration methods on the other hand are more widely applicable, but their results are mixed given very biased NPS estimates and a small PS.

Seeing the larger potential in using integration methods, this thesis seeks to contribute to the literature by proposing an alternative type of integration estimator to the composite, blended calibration, and Bayesian approach. The estimator draws from the wider literature on penalized regression, which has seen earlier success in integrating estimates from heterogeneous data sources (see C. Li et al. (2014), Liang et al. (2020), and Tian and Feng (2023)). The general idea of penalized regression for data integration is to use regression with additional penalties. These penalties constrain a set of target estimates of interest towards a set of auxiliary estimates. While we do not care about the auxiliary estimates, they share some similarities with the underlying target estimands. The hope is that this will increase the accuracy of the target estimates by leveraging information about the magnitude and direction of the target estimands of interest from the auxiliary estimates, reducing their variance for only a marginal increase in bias. This type of estimation method tends to stem from biostatistics and has to the author's knowledge never been applied in a setting such as the one described above. It is true that one type of estimator proposed by Wiśniowski et al. (2020) is the Bayesian interpretation of the ridge regression estimator (a type of penalized regression), however, it was not a major part of their discussion or research.

Therefore, given its earlier success in general data integration but unknown utility in official statistics, the overarching goal will be to apply a specific type of penalized regression in a context more relevant to NSIs and to answer the question of: **How can penalized regression be used to incorporate non-probability samples into official statistics?** To answer this question, this thesis will seek to specifically apply the Angle-Based Transfer Learning Estimator (ABTLE) which is an extension of the traditional ridge regression estimator proposed by Gu et al. (2022). The method is similar to superpopulation modeling and the Bayesian approach of Wiśniowski et al. (2020) in that it is also model-based. However, rather than directly correcting the selectivity, or using the NPS to construct a prior, the ABTLE uses the estimates of the NPS as part of a penalty, seeking to constrain the PS estimates by rewarding them for aligning angle-wise to the estimates of the NPS. This allows for a correction of the PS estimates should the two differ, borrowing the stability of the NPS estimates whilst still through the PS estimates ensuring some protection against the potential bias of the NPS estimates. The ABTLE has three main advantages over the

aforementioned methods mentioned in earlier sections: **1**. It is data-cheap, meaning that it only requires estimates directly relevant to the target estimates from the NPS. No actual microdata is required nor are any additional covariates to explain the selectivity of the NPS needed. **2**. Given sufficient angle similarity, the degree of bias in the NPS is less impactful on the estimator's quality, making it robust against a very biased NPS, even with a small PS. **3**. There is no theoretical risk of negative transfer, meaning the estimate should never be worse than just using the PS estimates in terms of MSE, as long as sufficiently correct penalty parameters are applied.

The goal of this thesis is twofold. Firstly, it will apply and evaluate the ABTLE through a simulation study on a wide range of practical scenarios with differing sample sizes, degrees of bias, multicollinearity, sample correlation, and residual variance. This allows for the evaluation of the estimation method in general but also highlights in what context the method is most or least useful. Although the primary metric of interest is the average root mean squared error (ARMSE), the mean of average bias of the estimator in various scenarios is also of interest given the importance of unbiased estimates in official statistics. The second supplementary goal will be to compare this estimation method to other benchmarking estimators such as the Bayesian approach proposed by Wiśniowski et al. (2020) and the ridge regression estimator, to highlight if and when the proposed ABTLE is the appropriate choice.

3. Method

3.1 Ridge regression

3.1.1 Background

Consider the normal multivariate linear regression model

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{3.1}$$

In this equation, $\mathbf{y}$ is the response vector, an $n \times 1$ column vector representing the observed values of the dependent variable for each observation. The matrix $\boldsymbol{X}$ is the design matrix, an $n \times p$ matrix where each row represents an observation and each column represents an independent variable. The vector $\boldsymbol{\beta}$ is the vector of coefficients, a $p \times 1$ column vector that represents the effect of each independent variable. Finally, $\boldsymbol{\epsilon}$ is the error term, an $n \times 1$ column vector that captures the normally-distributed residuals with mean 0 and variance σ^2 ($\boldsymbol{\epsilon} \sim N(0, \sigma^2)$) of each observation. Although we observe both $\mathbf{y}$ and $\boldsymbol{X}$, the estimand $\boldsymbol{\beta}$ is unknown and has to be estimated from the observed data. Since we tend to only have access to a fraction of the population (a sample) we rely on statistical estimation methods to infer the value of $\boldsymbol{\beta}$, where the goal is to arrive at as unbiased and certain estimates of $\boldsymbol{\beta}$ as possible.

One, if not the most common estimation method to do this is the maximum likelihood estimator (MLE), which provides efficient and unbiased parameter estimates under the assumption of normally distributed errors. The MLE does this by maximizing the likelihood function which in turn is done by finding the values of $\boldsymbol{\beta}$ that maximizes the probability of observing the given data. Under no violated model assumptions, following the Gauss-Markov theorem, this solution results in the best linear unbiased estimator possible, meaning that no other estimator is likewise unbiased with less variance. Another neat property of the MLE is that for a normal multivariate linear regression, the best estimates for $\boldsymbol{\beta}$ can be found analytically through the closed-form solution $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbf{y}$ (Hastie et al., 2009).

However, despite its status as the best linear unbiased estimator, the MLE faces challenges in high-dimensional settings, specifically in scenarios where $p > n$. In such a setting, the rank of the square $p \times p$ matrix composed of two $n \times p$ dimensional design matrices $\boldsymbol{X}$ is at most n, meaning that $\boldsymbol{X}^\top \boldsymbol{X}$ has at least $p - n$ zero eigenvalues. Because of this, the MLE's closed-form solution $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbf{y}$ is undefined since $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ does not exist in a setting with one or more zero eigenvalues. This can better be shown by rewriting $\boldsymbol{X}^\top \boldsymbol{X}$ as $\boldsymbol{A}$ and applying the spectral decomposition on $\boldsymbol{A}$: $\boldsymbol{A} = \sum_{j=1}^{p} \nu_j \mathbf{v}_j \mathbf{v}_j^\top$ where ν_j is a vector of eigenvalues and $\mathbf{v}_j$ their corresponding eigenvectors. The inverse of $\boldsymbol{A}$ can then be written as $\boldsymbol{A}^{-1} = \sum_{j=1}^{p} \nu_j^{-1} \mathbf{v}_j \mathbf{v}_j^\top$ which is clearly undefined if any $\nu_j = 0$ (van Wieringen, 2023). Even when $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ is defined, if it is near singular (i.e., at least one eigenvalue is very close to zero), the variance of the MLE estimates can become severely inflated. This issue is particularly pronounced in high-dimensional settings, where the number of predictors is close to the number of observations. In such cases, even minor fluctuations in the data can

lead to substantial changes in the estimated coefficients, resulting in MLE estimates with much higher variance than expected (Groß, 2003).

The ridge regression estimator $\hat{\boldsymbol{\beta}}_\lambda$ (henceforth known as the RRE) was initially developed as an ad-hoc solution to the problem of estimating said MLE in a high-dimensional setting. Its straightforward solution was simply to add a constant term to the diagonal of $\boldsymbol{X}^\top\boldsymbol{X}$, turning the MLE solution $\hat{\boldsymbol{\beta}}$ into the RRE with the updated closed-form solution of:

$$\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\mathbf{y}. \tag{3.2}$$

The RRE has the effect of adding a positive constant λ where $\lambda \in [0, \infty)$ to each eigenvalue, turning them non-zero and making estimation possible, since $\boldsymbol{A}^{-1}$ is no longer undefined (by turning the combined matrix positive definite) (van Wieringen, 2023). However, the initial solution was explicitly designed to solve the problem of an undefined matrix and not much thought was given to the fact that each distinct value of λ affects the estimation of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$, leading to varying estimates that follow along the so-called regularization path. This regularization path is important since it tends towards 0 as λ increases and on which $\hat{\boldsymbol{\beta}}$ is by definition not on (Hastie et al., 2009).

With an infinite number of λ values, there exists an infinite number of estimates along the regularization path, where each distinct value of λ has a direct impact on the performance and fit of $\hat{\boldsymbol{\beta}}_\lambda$. Rather than picking an arbitrarily small value of λ, it would be better if the choice could be informed so that we can maximize its performance. However, to understand what value of λ to use and how the RRE can have further utility outside of just solving a high-dimensional problem, one needs to further study the first two moments of $\hat{\boldsymbol{\beta}}_\lambda$. From there one can best see the trade-off of increasing λ on the bias, variance, and how their composite changes affect the accuracy (MSE) of $\hat{\boldsymbol{\beta}}_\lambda$.

3.1.2 Bias

Evaluating the first moment of $\hat{\boldsymbol{\beta}}_\lambda$, we can show that the $\mathbb{E}[\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}]$ is not equal to $\mathbb{E}[\hat{\boldsymbol{\beta}}]$ as long as $\lambda > 0$

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}] &= \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}\right] \\
&= \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbb{E}[\mathbf{y}] \\
&= \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\
&\neq \mathbb{E}[\hat{\boldsymbol{\beta}}] \\
&= \mathbb{E}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\right] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}[\mathbf{y}] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}.
\end{aligned} \tag{3.3}$$

As such, it is possible to show that $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ is always a biased estimate of the estimand $\boldsymbol{\beta}$ and by extension a more biased estimator than $\hat{\boldsymbol{\beta}}$ (since it is unbiased). As λ increases, the bias of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ will also increase since as $\lambda \to \infty$ the $\mathbb{E}[\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}] \to \mathbf{0}$ (away from $\boldsymbol{\beta}$) with the

denominator $(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}$ increasing vis-à-vis the fixed numerator $(\mathbf{X}^\top\mathbf{y})$. $\hat{\boldsymbol{\beta}}_\lambda$ can never be exactly $\mathbf{0}$. It can however be reduced to a vector of very small numbers that are practically indistinguishable from 0 within the context of bias. These two facts combined result in the upper bound of $\hat{\boldsymbol{\beta}}_\lambda$ bias being $\approx |\boldsymbol{\beta}|$, with the performance of $\hat{\boldsymbol{\beta}}_\lambda$ in terms of bias diminishing as λ increases (Hastie et al., 2009).

3.1.3 Variance

Before looking at the effect of increasing λ on $\text{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})$ it is useful to first introduce $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ as the product of $\hat{\boldsymbol{\beta}}$ and the matrix $\boldsymbol{W}_\lambda$ (van Wieringen, 2023),

$$\begin{aligned}\hat{\boldsymbol{\beta}}_\lambda &= (\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}\\ &= (\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\\ &= \mathbf{W}_\lambda(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\\ &= \mathbf{W}_\lambda\hat{\boldsymbol{\beta}}.\end{aligned} \tag{3.4}$$

This is because it is then possible to re-express $\text{Var}(\hat{\boldsymbol{\beta}}_\lambda)$ as $\text{Var}(\mathbf{W}_\lambda\hat{\boldsymbol{\beta}})$, which simplifies the relationship between $\text{Var}(\hat{\boldsymbol{\beta}}_\lambda)$ and $\text{Var}(\hat{\boldsymbol{\beta}})$, since we can the express the variance of $\hat{\boldsymbol{\beta}}_\lambda$ by the covariance matrix:

$$\text{Var}(\hat{\boldsymbol{\beta}}_\lambda) = \sigma^2\mathbf{W}_\lambda(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\mathbf{W}_\lambda^\top, \tag{3.5}$$

and for $\hat{\boldsymbol{\beta}}$ by the covariance matrix:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{3.6}$$

In a multivariate case, $\text{Var}(\hat{\boldsymbol{\beta}}_\lambda)$ can be said to be smaller than $\text{Var}(\hat{\boldsymbol{\beta}})$ as long as $\text{Var}(\boldsymbol{\alpha}\hat{\boldsymbol{\beta}}_\lambda) \leq \text{Var}(\boldsymbol{\alpha}\hat{\boldsymbol{\beta}})$, which is equivalent to $\boldsymbol{\alpha}[\text{Var}(\hat{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}}_\lambda)]\boldsymbol{\alpha}^\top \geq 0$, where $\boldsymbol{\alpha}$ is a vector with the non-zero value α. This inequality will only always hold true as long as the difference between the two covariance matrices results in a positive definite matrix (van Wieringen, 2023). Expressing the difference by using the expression from (3.5) and (3.6), after some matrix wrangling we end up with the difference expression of (for a full derivation see equation (A.1) in the appendix):

$$\text{Var}[\hat{\boldsymbol{\beta}}] - \text{Var}[\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}] = \sigma^2(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\left[2\mathbf{I}+\lambda^2(\mathbf{X}^\top\mathbf{X})^{-1}\right][(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}]^\top. \tag{3.7}$$

The final matrix expression is positive definite for any value of $\lambda > 0$ since every other component in the expression is at least non-negative definite (van Wieringen, 2023). This means that $\text{Var}(\hat{\boldsymbol{\beta}}) \succeq \text{Var}(\hat{\boldsymbol{\beta}}_\lambda)$ for any $\lambda > 0$.

From equation (3.5) we can also derive a general trend in $\text{Var}(\hat{\boldsymbol{\beta}}_\lambda)$ as $\lambda \to \infty$, by expanding the expression and applying the eigendecomposition on the matrices $(\mathbf{X}^\top\mathbf{X})$ and $(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}$, we can simply express the effect of increasing the value of λ as increasing a constant that shrinks the eigenvalues (λ_i) of the covariance matrix (for a full derivation and explanation see equation (A.2) in the appendix)

$$\text{Var}[\boldsymbol{\beta}_\lambda] = \sigma^2\mathbf{Q}\left[\text{diag}\left(\frac{\lambda_i}{(\lambda_i+\lambda)^2}\right)\right]\mathbf{Q}^\top. \tag{3.8}$$

As λ increases, the variance decreases and eventually shrinks towards zero. This is because in the expression $\text{diag}\left(\frac{\lambda_i}{(\lambda_i+\lambda)^2}\right)$, the denominator increases while the numerator remains constant, causing the overall fraction to get smaller as λ increases. Thus, contrary to the effect on the bias, the relative performance of $\hat{\boldsymbol{\beta}}_\lambda$ in terms of variance then actually improves as λ increases (Saleh et al., 2019).

3.1.4 MSE

Given both a formulation and consistent behavior of $\hat{\boldsymbol{\beta}}_\lambda$ in terms of bias and variance for an increase in the value of λ, it is now possible to evaluate the effect of increasing λ on overall estimation accuracy, formalized through the MSE. The MSE is the second moment of the error, meaning that it incorporates both the variance and the bias. It is formalized for $\hat{\boldsymbol{\beta}}_\lambda$ as the trace of the covariance matrix and the squared bias (for a full derivation see equation (A.3) in the appendix)

$$\text{MSE}(\hat{\boldsymbol{\beta}}_\lambda) = \sigma^2 \text{tr}[\mathbf{W}_\lambda(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{W}_\lambda^\top] + \boldsymbol{\beta}^\top(\mathbf{W}_\lambda - \mathbf{I})^\top(\mathbf{W}_\lambda - \mathbf{I})\boldsymbol{\beta}, \tag{3.9}$$

and for $\hat{\boldsymbol{\beta}}$ (since it is an unbiased estimator) as just the trace of the covariance matrix

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = \sigma^2 \text{tr}[(\mathbf{X}^\top\mathbf{X})^{-1}]. \tag{3.10}$$

Because of the above-discovered trade-off between decreased variance for increased bias, the relative performance of $\hat{\boldsymbol{\beta}}_\lambda$ against $\hat{\boldsymbol{\beta}}$ is not constant in terms of MSE. There exists only a range of λ values for which the reduction in variance is larger than the increase in squared bias, and where $MSE[\hat{\boldsymbol{\beta}}_\lambda] < MSE[\hat{\boldsymbol{\beta}}]$. For any value of λ larger than this range, the marginal increase in squared bias is larger than the marginal reduction in variance, resulting in an increase in the MSE (van Wieringen, 2023).

Likewise as with the variance, in a multivariate case the MSE of $\hat{\boldsymbol{\beta}}_\lambda$ can only be said to be smaller than $\hat{\boldsymbol{\beta}}$ if the difference between their two corresponding second-order moment error matrices ($\text{M}(\hat{\boldsymbol{\beta}})$ and $\text{M}(\hat{\boldsymbol{\beta}}_\lambda)$) results in a positive definite matrix (van Wieringen, 2023). Important to note is that this is not equivalent to (3.9) and (3.10) but rather it is formalized for $\hat{\boldsymbol{\beta}}_\lambda$ as:

$$\text{M}(\hat{\boldsymbol{\beta}}_\lambda) = \sigma^2\mathbf{W}_\lambda(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{W}_\lambda^\top + (\mathbf{W}_\lambda - \mathbf{I})\boldsymbol{\beta}\boldsymbol{\beta}^\top(\mathbf{W}_\lambda - \mathbf{I})^\top, \tag{3.11}$$

and for $\hat{\boldsymbol{\beta}}$ as:

$$\text{M}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}. \tag{3.12}$$

The first (and only part) is the covariance matrix already seen in the previous section. the second part is the squared bias reformulated. We can take the difference between equation (3.11) and equation (3.12) and after doing some matrix manipulation we can derive a final expression for this difference as a function of λ (for a full derivation see equation (A.4) in the appendix)

$$\text{M}(\hat{\boldsymbol{\beta}}) - \text{M}(\hat{\boldsymbol{\beta}}_\lambda) = \lambda(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}[2\sigma^2\mathbf{I} + \lambda\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} - \lambda\boldsymbol{\beta}\boldsymbol{\beta}^\top][(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}]^\top. \tag{3.13}$$

The final difference matrix is positive definite (and thus a range exists) only as long as $2\sigma^2\mathbf{I} + \lambda\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} - \lambda\boldsymbol{\beta}\boldsymbol{\beta}^\top \succ 0$. This is in turn always true as long as at least $2\sigma^2\mathbf{I} - \lambda\boldsymbol{\beta}\boldsymbol{\beta}^\top \succ$

0, which holds for any value of lambda where $2\sigma^2(\boldsymbol{\beta}^\top\boldsymbol{\beta})^{-1} > \lambda$ (Theobald, 1974). Since both σ^2 and $(\boldsymbol{\beta}^\top\boldsymbol{\beta})^{-1}$ are always positive values (assuming at least one non-zero β), a positive range of λ values always exists for which $\mathrm{M}(\hat{\boldsymbol{\beta}}) - \mathrm{M}(\hat{\boldsymbol{\beta}}_\lambda) \succ 0$ and by extension $\mathrm{MSE}[\hat{\boldsymbol{\beta}}_\lambda] < \mathrm{MSE}[\hat{\boldsymbol{\beta}}]$. This means that for every feasible scenario, theoretically, $\hat{\boldsymbol{\beta}}_\lambda$ can be estimated as a more accurate estimator than $\hat{\boldsymbol{\beta}}$ in terms of MSE.Gu et al. (2022) further outlines how given the parameters' true values, the optimal value within the above range for λ which minimizes the MSE is:

$$\lambda^\dagger = \frac{\frac{p}{n}\sigma^2}{\|\boldsymbol{\beta}\|_2^2}. \tag{3.14}$$

However, given that the optimal range and point depend on both $\boldsymbol{\beta}$ and σ^2, two unknown population parameters, the expressions are not useful in finding the actual values or range. Instead, it is more useful in highlighting that a range and point always exist and that they both increase as the ratio of residual variance to the squared magnitude of the coefficients increases.

3.1.5 Penalized regression

With the knowledge of the existence of a theoretically more accurate estimator, $\hat{\boldsymbol{\beta}}_\lambda$ can thusly be justified not just as a method to solve a high-dimensional estimation problem, but also as a method to improve estimation accuracy through minimizing the MSE. Because of this, it is useful to reformulate the RRE as a more structural method of improving an estimator's accuracy. This is best done by instead expressing ridge regression as a type of penalized regression where the RRE is the solution to the following loss function:

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2. \tag{3.15}$$

The RRE is the values of $\boldsymbol{\beta}$ which minimizes the residual sum of squares $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ given an additional "ridge penalty" of $\lambda\|\boldsymbol{\beta}\|_2^2$. Here λ is interpreted as a scalar (labeled as the penalty parameter) to the additional ridge penalty (Saleh et al., 2019). Intuitively, if λ is set to 0, then the ridge penalty is nullified, and what is left is simply the minimization of the residual sum of squares, which is minimized by the MLE $\hat{\boldsymbol{\beta}}$. As $\lambda \to \infty$ however, the ridge penalty will increasingly dominate the ridge loss function. In such a scenario the best way to minimize the ridge loss function is to shrink $\boldsymbol{\beta}$ towards 0 as to attempt to nullify the large ridge penalty. This causes the bias to increase since the estimated coefficients are pushed closer to zero, deviating from their true values. However, as $\boldsymbol{\beta}$ approaches $\mathbf{0}$, the variance decreases because the coefficients become more stable and less sensitive to fluctuations in the data (Hastie et al., 2009).

Taking the derivative of the ridge loss function with regards to $\boldsymbol{\beta}$ results in the following expression:

$$\frac{\partial}{\partial\boldsymbol{\beta}}(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2) = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta}. \tag{3.16}$$

It is then straightforward to solve for 0 and to show that the solution that minimizes the ridge loss function in equation (3.15) is the RRE closed form solution already introduced in

equation (3.2)

$$\begin{aligned} -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} &= 0 \\ (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} &= \mathbf{X}^\top\mathbf{y} \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}. \end{aligned} \tag{3.17}$$

By calculating the second derivative of both components in the ridge loss function with regard to $\boldsymbol{\beta}$, we can demonstrate that the identified solution is not only a local minimum but also a global minimum. This is because the component $\mathbf{X}^\top\mathbf{X}$ is positive semi-definite (assuming no high-dimensionality), and for a non-zero λ, the component $\lambda\mathbf{I}$ is positive definite. Since the sum of a positive semi-definite matrix and a positive definite matrix is positive definite by definition, this ensures that the Hessian matrix of the ridge loss function is positive definite. Consequently, the ridge loss function is strictly convex, guaranteeing that the identified minimum ($\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$) is indeed the global minimum (Saleh et al., 2019).

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2) = 2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}) \tag{3.18}$$

3.2 Extended penalized regression

3.2.1 Background

Seeing the utility of penalized regression as an estimation framework, there has in recent years been substantial development of innovative penalization schemes that seek to extend the idea of ridge regression to penalize estimate differences, rather than estimate magnitudes. The hope is to further improve the estimator accuracy by leveraging additional information from an auxiliary estimator $\hat{\mathbf{w}}$ that comes from an auxiliary data source. Both of these terms are quite flexible where for example the auxiliary source can refer to a different sample from the same population (Such as an NPS) or a sample from a different population. Similarly, $\hat{\mathbf{w}}$ can relate to a different estimand $\mathbf{w}$, which is assumed to share some sort of similarity with the estimand of interest $\boldsymbol{\beta}$, or just directly to $\boldsymbol{\beta}$.

A common extension is to utilize a distance-based penalty between $\boldsymbol{\beta}$ and $\hat{\mathbf{w}}$, as was done by Tian and Feng (2023). They proposed to replace $\lambda\|\boldsymbol{\beta}\|_2^2$ with $\lambda\|\boldsymbol{\beta} - \hat{\mathbf{w}}\|_2$, which would encourage $\boldsymbol{\beta}$ to approach $\hat{\mathbf{w}}$ in terms of Euclidean distance. S. Li et al. (2022) also proposed a similar penalization scheme using a distance-based penalty, but they utilized Manhattan distance instead. These distance-based penalization schemes can work. However, they do not account for common directionality between $\boldsymbol{\beta}$ and $\hat{\mathbf{w}}$, something which can become problematic when they are highly concordant, but with a large distance between them. In such a scenario, the distance-based penalties become sub-optimal since they throw away relevant directional information, potentially reducing the accuracy improvements possible. In the worst case scenario, when $\hat{\mathbf{w}}$ highly correlated with $\boldsymbol{\beta}$ and the distance is larger, the distance penalties can even lead to so-called negative transfer, where the incorporation of additional information results in deteriorating estimates as measured by MSE (Gu et al., 2022).

A more general alternative to a distance-based penalty would be to instead leverage the direction $\boldsymbol{\beta}$ and $\hat{\mathbf{w}}$ through some sort of angle-based penalty. This could be done through the hypothetical penalty of $\lambda\sqrt{1-\left(\frac{\hat{\mathbf{w}}\cdot\boldsymbol{\beta}}{\|\hat{\mathbf{w}}\|_2\|\boldsymbol{\beta}\|_2}\right)^2}$, or a bivariate ridge regression where each target coefficient is assumed to share some degree of correlation with its auxiliary counterpart, as proposed by C. Li et al. (2014). Such solutions are preferred over a distance penalty since a short distance also implies a small angle while the reverse need not be true. However, the hypothetical penalization scheme does not result in a simple closed-form solution (Gu et al., 2022), while the bivariate ridge regression has only been applied using two predictors and requires the estimation of 4 penalty parameters. This in turn risks inducing unreliable results in small samples due to penalty parameter estimation variance (Riley et al., 2021).

3.2.2 The Angle-based Transfer Learning Estimator

Incorporating the earlier research on extended penalized regression Gu et al. (2022) instead propose their alternative estimator labeled the "Angle-Based Transfer Learning Estimator" (henceforth known as the ABTLE or $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$) which is formalized as the solution to the following loss function

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 - 2\eta\hat{\mathbf{w}}^\top\boldsymbol{\beta}. \tag{3.19}$$

It can be interpreted as an extension to the RRE which in addition to the residual sum of squares $\frac{1}{n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$ (adapted to account for sample size) and the normal penalty term on the squared magnitude $\|\boldsymbol{\beta}\|_2^2$, also incorporates a reward term of $-2\eta\hat{\mathbf{w}}^\top\boldsymbol{\beta}$. This additional term helps leverage the angle similarity between $\boldsymbol{\beta}$ and $\hat{\mathbf{w}}$ since $\boldsymbol{\beta}^\top\hat{\mathbf{w}} = \|\boldsymbol{\beta}\|_2\|\hat{\mathbf{w}}\|_2\cos\Theta$, encouraging $\boldsymbol{\beta}$ to be more concordant with $\hat{\mathbf{w}}$ depending on the value of the new additional independent penalty parameter η. It is thus also a type of angle-based penalization scheme, but a lot simpler than the aforementioned ones. As a consequence, it lends itself better to a straightforward closed-form solution, as well as only requiring two penalty parameters and only estimates from the auxiliary dataset and no actual micro-data.

Taking the derivative of the loss function in (3.19) with regards to $\boldsymbol{\beta}$, we end up with the expression:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{1}{n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 - 2\eta\hat{\mathbf{w}}^\top\boldsymbol{\beta}\right) = -\frac{2}{n}\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} - 2\eta\hat{\mathbf{w}}. \tag{3.20}$$

From there, by reshuffling the terms, a closed-form solution for $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ can be found

$$\begin{aligned}
-\frac{2}{n}\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} - 2\eta\hat{\mathbf{w}} &= 0 \\
\frac{1}{n}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta} &= \frac{1}{n}\mathbf{X}^\top\mathbf{y} + \eta\hat{\mathbf{w}} \\
\left(\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I}\right)\boldsymbol{\beta} &= \mathbf{X}^\top\mathbf{y} + n\eta\hat{\mathbf{w}} \\
\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} &= \left(\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^\top\mathbf{y} + n\eta\hat{\mathbf{w}}\right).
\end{aligned} \tag{3.21}$$

The solution $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} = (\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I})^{-1}(\mathbf{X}^\top\mathbf{y} + n\eta\hat{\mathbf{w}})$ is quite similar to $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ but with the additional term $n\eta\hat{\mathbf{w}}$. Essentially $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ becomes a combination of the RRE and the auxiliary

estimator $\hat{\mathbf{w}}$, where λ again controls the shrinkage of $\boldsymbol{\beta}$ and η adjusts the influence of $\hat{\mathbf{w}}$. The ratio between the two penalty parameters then balances the fit to the data by shrinking $\boldsymbol{\beta}$ to some extent and adding weighted contribution from the auxiliary estimates $\hat{\mathbf{w}}$. If both penalty parameters are very close to zero then neither penalty has a significant effect and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} \approx \hat{\boldsymbol{\beta}}$. In the other extreme scenario where both λ and η are very large (when $\boldsymbol{\beta}$ and $\hat{\mathbf{w}}$ are highly correlated), then $\boldsymbol{\beta}$ will be shrunk to essentially nothing and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} \approx c\hat{\mathbf{w}}$, where is a c is a constant that rescales $\hat{\mathbf{w}}$ by some factor depending on the ratio between η and λ (Gu et al., 2022).

Under a scenario with no-high dimensionality and a $\hat{\mathbf{w}}$ with very low to no estimation error, the optimal penalty parameters which when applied minimize (3.19) are approximately:

$$\lambda^* \approx \frac{\frac{p}{n}\sigma^2}{\|\boldsymbol{\beta}\|_2^2(1-\rho^2)} \quad \text{and} \quad \eta^* \approx \lambda^* \rho \frac{\|\boldsymbol{\beta}\|_2}{\|\mathbf{w}\|_2}. \tag{3.22}$$

Here ρ refers to the correlation between $\boldsymbol{\beta}$ and $\hat{\mathbf{w}}$ and σ^2 the residual variance of the target sample (PS) (Gu et al., 2022). Just as with the RRE, we see that the optimal value of λ depends again on both the residual variance and squared magnitude of the coefficients but here also on the squared correlation. η in turn is just a scaled version of λ, depending on the ratio of magnitudes between $\boldsymbol{\beta}$ and $\hat{\mathbf{w}}$ and their correlation. As with the RRE, all of these are unknown parameters meaning that this is a more theoretical solution than practically useful.

One key advantage of this specific penalization scheme utilizing two independent penalty parameters is that it incorporates both the RRE and the so-called Distace-Based Transfer Learning Estimator (henceforth known as the DBTLE or $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\lambda}}$) as special cases of itself with a fixed value for η. The incorporation of RRE is pretty straightforward. Simply by fixing $\eta = 0$ the additional penalty term in equation (3.19) becomes completely negated. This leaves only $\frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$, which is just the RRE. To see how the DBTLE is incorporated the combined penalty terms must first be reexpressed as:

$$\begin{aligned}
\lambda\|\boldsymbol{\beta}\|_2^2 - 2\eta\boldsymbol{\beta}^\top\hat{\mathbf{w}} &= \lambda(\boldsymbol{\beta}^\top\boldsymbol{\beta}) - 2\eta\boldsymbol{\beta}^\top\hat{\mathbf{w}} \\
&= \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta}^\top\hat{\mathbf{w}} + \lambda\hat{\mathbf{w}}^\top\hat{\mathbf{w}} - 2\eta\boldsymbol{\beta}^\top\hat{\mathbf{w}} + 2\lambda\boldsymbol{\beta}^\top\hat{\mathbf{w}} - \lambda\hat{\mathbf{w}}^\top\hat{\mathbf{w}} \\
&= \lambda\|\boldsymbol{\beta} - \hat{\mathbf{w}}\|_2^2 - 2(\eta - \lambda)\boldsymbol{\beta}^\top\hat{\mathbf{w}} - \lambda\|\hat{\mathbf{w}}\|_2^2.
\end{aligned} \tag{3.23}$$

Using the reexpressed penalty terms it again is pretty straightforward to show how the ABTLE incorporates the DBTLE. Simply fixing $\eta = \lambda$, the middle terms become completely negated, leaving only the distance penalty, which is equivalent to turning $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ into the DBTLE. It is important to note that the DBTLE is not the same as the one proposed by Tian and Feng (2023) but that it is a slightly altered version of it incorporating the squared l2 norm and the additional but not relevant term $\|\hat{\mathbf{w}}\|_2^2$.

The fact that both the DBTLE as well as the RRE can be interpreted as constrained versions of the ABTLE combined with an approximate distribution of the optimal penalty parameters, aids in clarifying the scenarios where the ABTLE is expected to outperform its reduced counterparts. Because the DBTLE, is achieved by fixing $\eta = \lambda$, it is possible to see that the ABTLE will always outperform or be equivalent to the DBTLE. This is because if $\eta = \lambda$ the

optimal value of η (η^*) will only intersect with the optimal value of lambda (λ^*) at points where $\rho\frac{\|\boldsymbol{\beta}\|_2}{\|\mathbf{w}\|_2} = 1$, at which its performance is equivalent to that of the ABTLE. For all other scenarios, by fixing $\eta = \lambda$ the DBTLE will either over or under-align depending on the true correlation and magnitude ratio. In the worst-case scenario, this can even lead to negative transfer, where because of a too-large value of η, the DBTLE over-aligns $\boldsymbol{\beta}$ with $\hat{\mathbf{w}}$ resulting in a worse estimator than just using $\hat{\boldsymbol{\beta}}$.

For the RRE, it is also possible to show that by fixing $\eta = 0$, the RRE will only be equally good to the ABTLE when $\boldsymbol{\beta}$ and $\hat{\mathbf{w}}$ are completely orthogonal since only then is $\lambda^\dagger = \frac{\frac{p}{n}\sigma^2}{\|\boldsymbol{\beta}\|_2^2} = \frac{\frac{p}{n}\sigma^2}{\|\boldsymbol{\beta}\|_2^2(1-\rho^2)} = \lambda^*$ and by extension $\lambda^*\rho\frac{\|\boldsymbol{\beta}\|_2}{\|\boldsymbol{w}\|_2} = 0$. For all other situations, the RRE under-penalizes and under-aligns, resulting in a performance loss. It is important to note that, unlike the DBTLE, the RRE will never over-align since it does not actually incorporate any auxiliary information, meaning that it will never be worse than just $\hat{\boldsymbol{\beta}}$ as seen in section 3.1.4. Gu et al. (2022) also provide a more technical explanation for this, utilizing the fact that the minimal loss can be expressed for both estimators as $\frac{\sigma^2}{\lambda v(-\lambda)}$, which is a monotonically decreasing function where $v(\lambda)$ is the Stieltjes transformation of the spectral distribution F_Σ, of the matrix $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$, dependent on the values of $\lambda^\dagger$ and λ^*. Because the function is monotonically decreasing, and we know that for any degree of correlation $\lambda^\dagger \leq \lambda^*$, the loss of the ABTLE will always be smaller than for the RRE. Further proofs and derivations are outside of the knowledge scope of the author, so for further information see the supplementary material of Gu et al. (2022). Nevertheless, this has an important practical implication for the ABTLE since it means that given that λ and η are correctly specified, the ABTLE is guaranteed to never be worse than both the RRE by extension the MLE. The ABTLE is thus immune against negative transfer, a problem which previous attempts at integrating a PS and NPS have struggled with occurring for various scenarios (see Elliott and Haviland (2007) and Wiśniowski et al. (2020)).

To conclude this section, we demonstrate that the identified minimum loss from $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ is a global minimum by calculating the second derivative of the loss function in equation (3.19). The second derivative is equivalent to that of (3.18) but with the inclusion of the positive scalar n (sample size). Since the sample size is always positive, the full expression remains a positive definite matrix, just like in (3.18). Therefore, the ABTLE loss function is strictly convex, ensuring that the identified minimum is indeed the global minimum.

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\left(\frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 - 2\eta\hat{\mathbf{w}}^\top\boldsymbol{\beta}\right) = \frac{2}{n}(\mathbf{X}^\top\mathbf{X} + \lambda n\mathbf{I}) + 0 \tag{3.24}$$

3.3 Cross-validation

Throughout the above sections, the penalty parameters λ and η have been treated as known values which are assumed to be optimal for minimizing their corresponding loss functions seen in equation (3.15) and equation (3.19). However, because the optimal values of λ and η depend directly on the unknown population parameters of $\boldsymbol{\beta}$, σ^2 (and $\hat{\mathbf{w}}$), when actually applying the ABTLE, λ and η have to be empirically estimated beforehand. The challenge of estimating the best penalty parameters for ABTLE arises because typically only a small

sample of the population of interest is available. Relying on this limited sample to both fit the ABTLE and determine the optimal λ and η is risky, often leading to overfitting, where a λ and η are selected that fit the available sample exceptionally well, but that fails to generalize to the full population. The primary issue is that the selected λ and η are tailored to the specific characteristics of the sample, which can vary significantly from other samples and the wider population, resulting in a sub-optimal performance when applied to new data (Hastie et al., 2009). While the most straightforward approach is to use two distinct sets of samples, one for determining penalty parameters and another for estimating $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$, this is often impractical due to the high cost of data collection and limited resources. Consequently, an alternative widely adopted method to generalize the estimation process using only one dataset, for any pair of penalty parameters ($\boldsymbol{\theta}$), with any fit function $f(X, \boldsymbol{\theta})$, is k-fold cross-validation.

In k-fold cross-validation, the available data is partitioned into k evenly sized, non-overlapping folds. The fit function is then estimated on $k-1$ folds using any predetermined pair of values for $\boldsymbol{\theta}$, resulting in a fold and $\boldsymbol{\theta}$ specific fit function

$$\hat{f}_{-k} = f(\mathbf{y}_{-k}, \mathbf{X}_{-k}, \boldsymbol{\theta}). \tag{3.25}$$

$\hat{f}_{-k}$ is then evaluated by applying it on the remaining k fold through a cross-validation error function, resulting in some cross-validation error based on how well $\hat{f}_{-k}$ generalizes to new unseen data

$$CV(\hat{f}_{-k}, \boldsymbol{\theta}) = L(\hat{f}_{-k}(\mathbf{y}_k, \mathbf{X}_k)). \tag{3.26}$$

This process is repeated k times, each time with a different fold being used for evaluation, ensuring that each data point is used for both estimation and evaluation. We can then sum the cross-validation error for all k folds resulting in a total cross-validation error for some values of $\boldsymbol{\theta}$

$$CV(f, \boldsymbol{\theta}) = \sum_{k=1}^{K} L(\hat{f}_{-k}(\mathbf{y}_k, \mathbf{X}_k)). \tag{3.27}$$

Cross-validation is then performed for every possible combination of values of $\boldsymbol{\theta}$ along a predetermined set range of values where we expect to find the pair $\boldsymbol{\theta}^*$ that are the true optimal values of $\boldsymbol{\theta}$. Whichever $\boldsymbol{\theta}$ results in the smallest total cross-validation error is then selected as the assumed optimal vector of penalty parameters which are used for a final estimation on the full sample.

$$f(\mathbf{y}, \mathbf{X}, \hat{\boldsymbol{\theta}}) = \arg\min_{\boldsymbol{\theta}} \sum_{k=1}^{K} L(\hat{f}_{-k}(\mathbf{y}_k, \mathbf{X}_k)) \tag{3.28}$$

Important to note is that this does not mean that the determined upon $\hat{\boldsymbol{\theta}}$ is necessarily the optimal $\boldsymbol{\theta}^*$, but it is merely the value which generalized best across several "samples". Because of this, it is still well possible that the cross-validation process – due to randomness in specific folds or even the full sample – results in a suboptimal value of $\hat{\boldsymbol{\theta}}$ on the population level. Nevertheless, given the practical constraints, cross-validation mimics a scenario with k correlated albeit different samples and has been found to select penalty parameter values that generalize better than unpartitioned data (Hastie et al., 2009; James et al., 2023).

4. Simulation

4.1 Data simulation

The goal of this thesis is to evaluate the potential of using penalized regression to incorporate NPSs into official statistics, specifically through the ABTLE. ABTLE will therefore be evaluated on simulated data as a proof of concept study, with a series of varying parameters designed to emulate different scenarios in official statistics where the performance of the ABTLE is expected to vary (Morris et al., 2019). The process of estimation is designed to emulate the context faced in an NSI, meaning that the simulation process is limited to only altering either the population distribution, the sampling mechanism, or the target model. Because of this, rather than combining two independent sets of data, both the target data and the auxiliary estimates come from the same underlying population, but in the form of a PS and NPS respectively. To ensure consistency with the framework developed by Gu et al. (2022) and to avoid notational overlap, estimates from the auxiliary NPS are labeled using $\hat{\mathbf{w}}$ even though they related directly to the estimand $\boldsymbol{\beta}$. Another thing to note is that since the study is simulated, we have access to the true population parameter. Because of this, rather than evaluating the estimation methods based on how well they estimate the target outcome ($\mathbf{y}$), we will instead focus on how well they estimate the target model coefficients ($\boldsymbol{\beta}$).

Based on the paper by Gu et al. (2022) as well as the general literature on the topic, five parameters have been identified as important for the ABTLE's performance. These are: **1**. The correlation between the individual independent variables of interest (γ), **2**. The degree of bias (f) in the NPS estimates $\hat{\mathbf{w}}$, **3**. The expected correlation (the non-geometric interpretation of angle distance) between $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{w}}$ (ρ), **4**. The degree of residual variance (σ^2), and **5**. The sample size of the PS (n_{ps}). Because there likely are interaction effects between parameters, every combination of relevant parameter values is tested, resulting in **54** unique population scenarios. For each scenario four differently sized probability samples are collected resulting in a total of **216** simulated scenarios.

4.1.1 Population generation

Each simulation scenario begins by generating a finite population of $N = 100{,}000$ units where each unit consists of $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}\}$ and an additional variable labelled $\mathbf{z}$. $\boldsymbol{X}$ contains the three observed variables of interest whose coefficients $\boldsymbol{\beta} = \{\beta_1, \beta_2, \beta_3\}$ we are interested in estimating as accurately as possible. $\boldsymbol{X}$ is generated following:

$$\boldsymbol{X} \sim \text{MVN}\left(\mathbf{0}, \begin{pmatrix} 1 & \gamma & \gamma \\ \gamma & 1 & \gamma \\ \gamma & \gamma & 1 \end{pmatrix}\right), \tag{4.1}$$

as three multivariate standard normally distributed variables with an evenly distributed covariance of γ between and across all three variables. For this study, we consider $\gamma \in \{0.38, 0.64, 0.87\}$. These values result in a coefficient of multiple correlations of approxi-

mately 0.2, 0.5, and 0.8 for this specific setup, reflecting three scenarios of limited, moderate, and severe multicollinearity respectively.

The variable $\mathbf{z}$ is an additionally simulated latent variable present in each unit, where $\mathbf{z} \sim N(0,1)$. $\mathbf{z}$ thus follows a similar standard normal distribution as any $\boldsymbol{x_k}$ but is generated completely independent of them. It forms the basis of the inclusion probability in the NPS and is a covariate partially explaining the outcome of interest. However, this study treats $\mathbf{z}$ as a hidden latent variable, meaning that despite it contributing to the outcome (and NPS inclusion probability), it is not observed within either the PS or the NPS.

Each unit also contains a continuous zero-centered outcome variable $\mathbf{y}$ distributed along:

$$\mathbf{y} \sim \boldsymbol{X\beta} + \boldsymbol{X}\mathbf{z}\mathbf{I}\boldsymbol{\phi} + \boldsymbol{\epsilon}. \tag{4.2}$$

Here matrix $\boldsymbol{X}$ and vector $\mathbf{z}$ are the aforementioned variables, $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are coefficient vectors (of which we are only interested in $\boldsymbol{\beta}$), and $\boldsymbol{\epsilon}$ the residual, distributed following $\boldsymbol{\epsilon}. \sim N(0, \sigma^2)$. In this study $\sigma^2 \in \{2, 10\}$ where a value of 2 is a relatively small residual variance corresponding to (in a scenario of only $\mathbf{y} \sim \boldsymbol{X\beta} + \boldsymbol{\epsilon}$) an R^2 of approximately 0.7. Opposite, a value of 10 can be seen as a very large residual variance, corresponding to (in a scenario of only $\mathbf{y} \sim \boldsymbol{X\beta} + \boldsymbol{\epsilon}$) an R^2 of approximately 0.08.

To ensure cross-scenario comparability since the magnitude of $\boldsymbol{\beta}$ affects the estimation process, rather than randomly drawing $\boldsymbol{\beta} \sim N(\mu, \sigma_\beta^2)$ for some μ and σ_β^2, for each scenario, we always assume $\beta_1 = 2$, $\beta_2 = 1$, and $\beta_3 = 0.5$. These values are arbitrary other than they are taken from the simulation process of Wiśniowski et al. (2020) whose estimation method is used for comparison. $\boldsymbol{\phi} = \{\phi_1, \phi_2, \phi_3\}$ is generated as a random vector of values where $\boldsymbol{\phi} \sim U(0, 3.5)$ ($3.5 = \sum_{k=1}^{3} \beta_k$). $\boldsymbol{\phi}$ is then rescaled so that $\sum_{k=1}^{3} \phi_k = f * \sum_{k=1}^{3} \beta_k$, and rotated so that $Cor(\boldsymbol{\beta}, \boldsymbol{\beta} + \boldsymbol{\phi}) = \rho$. In this study $\rho \in \{0.5, 0.7, 0.9\}$ where the values of ρ simply reflect a range of potential correlations from moderate to very strong, loosely based on the classification of Evans (1996). f is a bias factor scaling the magnitude of $\boldsymbol{\phi}$ where $f \in \{-0.5, 1, 3\}$. f is a relative value, where a larger value reflects a larger degree of bias, but where the absolute degree of bias depends on the underlying scale of the estimands (Wiśniowski et al., 2020). For reasons that will become apparent in the analysis section, a higher absolute value of f will directly translate into more bias of the NPS estimates $\hat{\mathbf{w}}$ whilst the value of ρ will determine the expected correlation between the estimates of the PS ($\hat{\boldsymbol{\beta}}$) and NPS ($\hat{\mathbf{w}}$).

4.1.2 Sampling

For each population generated, a series of samples of various sizes are drawn. For the PS we consider $n_{ps} \in \{25, 50, 75, 100\}$ where each sample is drawn using simple random sampling. For the NPS only $n_{nps} = 10^5$ is considered. The NPS is drawn with unequal inclusion probabilities generated by (4.3) following the method applied by Smit (2021) and Villalobos Aliste (2022), but slightly adjusted to better suit the current context (increased inflection point from 0.75 to 1).

The individual inclusion probabilities into the NPS (p_i) are created by transforming the individual selectivity variable value z_i using a logistic function with a growth rate of 2

and an inflection point of 1. In practice, this means that an individual with a z value of 1 has a 50% chance of being included in the NPS (before normalization). The vector of inclusion probabilities $\mathbf{p}$ is then normalized to sum up to the desired sample size of 10^5, after which the NPS is then sampled using systematic random sampling with unequal inclusion probabilities

$$p_i = \frac{1}{1 + \exp(-2(z_i - 1))}. \tag{4.3}$$

4.2 Analysis

4.2.1 Target model

As mentioned in the previous section, $\mathbf{z}$ is a latent variable whose values "cannot" be observed in either sample. As such, the estimation of all estimators is done on a reduced model (4.4). It contains only the target variables of interest where the interaction term $\boldsymbol{X}\mathbf{z}\mathbf{I}\boldsymbol{\phi}$ have been subsumed into the error term $\boldsymbol{\epsilon'} = \boldsymbol{\epsilon} + \boldsymbol{X}\mathbf{z}\mathbf{I}\boldsymbol{\phi}$.

$$\mathbf{y} \sim \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon'} \tag{4.4}$$

Using the reduced model and employing the two sampling schemes detailed above has a direct twofold impact on just using the MLE for both samples. Given just the PS, when the sample is drawn using simple random sampling $\mathbb{E}[\mathbf{z}] = 0$. Consequently, $\mathbb{E}[\boldsymbol{\epsilon'}] = \mathbb{E}[\boldsymbol{\epsilon}] + (\mathbb{E}[\boldsymbol{X}\boldsymbol{\phi}] \odot \mathbb{E}[\mathbf{z}]) = 0$. This means that the PS MLE $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ despite the omission of $\boldsymbol{X}\mathbf{z}\mathbf{I}\boldsymbol{\phi}$, as seen here:

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] \\
&= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon'})] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon'}] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}.
\end{aligned} \tag{4.5}$$

It should be noted that even though $\hat{\boldsymbol{\beta}}$ is still unbiased, the introduction of the latent variable $\mathbf{z}$ does nonetheless still affect the estimation process. This is because $\boldsymbol{\phi}$ still contributes to the error term $\boldsymbol{\epsilon'}$, increasing it, and by extension also the variance of $\hat{\boldsymbol{\beta}}$. This means that as the bias factor (f) increases, the residual variance of the PS model does as well.

Due to the unequal inclusion probabilities created in equation (4.3), given just the NPS the $\mathbb{E}[\mathbf{z}] = 1$, meaning that $\mathbb{E}[\boldsymbol{\epsilon'}] = \mathbb{E}[\boldsymbol{\epsilon}] + (\mathbb{E}[\boldsymbol{X}\boldsymbol{\phi}] \odot \mathbb{E}[\mathbf{z}]) = \boldsymbol{X}\boldsymbol{\phi}$. The expected value of ϵ' is thus no longer zero, resulting in the introduction of a systematic component correlated with $\boldsymbol{X}$ into the error term. This correlation violates the assumption that the error term is independent of the regressors, leading to the NPS MLE $\hat{\mathbf{w}}$ being a biased estimate of $\boldsymbol{\beta}$ as

seen below:

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{w}}] &= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] \\ &= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}')] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}'] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\phi} \\ &= \boldsymbol{\beta}+\boldsymbol{\phi}.\end{aligned} \tag{4.6}$$

Given that $\sum_{k=1}^{3}\phi_k = f * \sum_{k=1}^{3}\beta_k$, the expected total bias introduced across $\hat{\mathbf{w}}$ should be $\boldsymbol{\beta} - (\boldsymbol{\beta} + f * \boldsymbol{\beta}) = f * \boldsymbol{\beta}$ and the expected correlation with $\hat{\boldsymbol{\beta}}$ $Cor(\hat{\boldsymbol{\beta}}, \hat{\mathbf{w}}) = Cor(\boldsymbol{\beta}, \boldsymbol{\beta}+\boldsymbol{\phi}) = \rho$.

4.2.2 Estimation

Once the population has been generated and a PS and NPS have been sampled, the process of estimating $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ can begin. The process of estimation is done in three stages, First, utilizing only the NPS, the auxiliary estimates $\hat{\mathbf{w}}$ are estimated. Secondly utilizing the estimated $\hat{\mathbf{w}}$ and the PS, cross-validation is done to estimate the penalty parameters $\hat{\lambda}$ and $\hat{\eta}$. Finally utilizing all aforementioned components, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ is estimated from the PS, with the aid of $\hat{\mathbf{w}}$, $\hat{\lambda}$ and $\hat{\eta}$.

Because of inherent variability in the sampling process, small random changes in the simulation process can have an unknown yet significant impact on the one-time performance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$. Thus estimation of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ must be done for a series of random samples to ascertain reliable and stable results (Morris et al., 2019). Given prior similar research (Smit, 2021; Villalobos Aliste, 2022), the decision was made to utilize 1000 iterations for each of the 216 total scenarios.

Estimation of auxiliary estimates

To estimate $\hat{\mathbf{w}}$ from the auxiliary NPS, we simply utilize normal maximum likelihood estimation, which yields a straightforward closed-form solution. Specifically, given the NPS design matrix $\boldsymbol{X}_{\boldsymbol{nps}}$ and the NPS response vector $\boldsymbol{y}_{\boldsymbol{nps}}$, the estimator $\hat{\mathbf{w}}$ can be computed using the formula $\hat{\mathbf{w}} = (\boldsymbol{X}_{\boldsymbol{nps}}^{\top}\boldsymbol{X}_{\boldsymbol{nps}})^{-1}(\boldsymbol{X}_{\boldsymbol{nps}}^{\top}\boldsymbol{y}_{\boldsymbol{nps}})$.

Estimation of penalty parameters

Once $\hat{\mathbf{w}}$ has been estimated, the following step is to estimate the penalty parameters $\hat{\lambda}$ and $\hat{\eta}$ through k-fold cross-validation. The goal is to minimize (3.28) given a PS and $\hat{\mathbf{w}}$, for some range of λ and η values. In this research $k = 10$, since that is a conventional choice for the number of folds which finds a middle ground between validation bias vis-à-vis variance, and only results in a moderate computational burden (James et al., 2023). Regarding the range of penalty parameters, after attempting the cross-validation a few times it was found that both λ and η vary quite a bit depending on the scenario, but that the most common values were also very small or very large. Thus, the range decided upon was: 150 evenly distributed values from 0.0005 to 1, 100 evenly distributed values from 1 to 5, 6 to 100 in increments of

1, and the very large value of 10^5. The first range contains by far the most common optimal values for the penalty parameters, and thus it is the most detailed range. The second and third ranges are coarser and exist to accommodate outlier samples. The final value of 10^5 is there for the situation where $\hat{\mathbf{w}}$ and $\hat{\boldsymbol{\beta}}$ are highly correlated and where the residual sum of squares is minimized by just re-scaling $\hat{\mathbf{w}}$ by some constant c.

However, employing such a fine-grain but also wide-net for a grid search of two variables becomes problematic for two reasons. Firstly, given that λ and η can take two very large values. At high values, it then becomes impossible to find an accurate ratio between the two without also extending the range to an infeasible range. Secondly, every additional value added to the range rapidly increases the number of inversions needed, quickly leading to a very high computational burden. Therefore, rather than conducting a grid search across all of the pairwise values of the range, here we utilize an alternative stage-wise estimation process for both penalty parameters.

First, we formalize the cross-validation error function as:

$$CV(\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k}, \lambda, \eta) = \|\mathbf{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k}\|_2^2, \tag{4.7}$$

where the function $\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k}$ (fit function) is the solution from equation (3.21) applied on all but the kth fold

$$\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k} = (\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}(\mathbf{X}_{-k}^\top\mathbf{y}_{-k} + n\eta\hat{\mathbf{w}}). \tag{4.8}$$

We can then utilize the fact that if we input the closed-form solution (4.8) into equation (4.7) and take the derivative with regard to η, there exists a best estimate of $\hat{\eta}_k$ for a given k, n and λ value in the form of (see (A.10) for a full derivation):

$$\hat{\eta}_k = \frac{\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}\right]^\top(\mathbf{X}_k^\top\mathbf{y}_k) - \left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}(\mathbf{X}_{-k}^\top\mathbf{y}_{-k})\right]^\top(\mathbf{X}_k^\top\mathbf{X}_k)(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}}{n\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}\right]^\top(\mathbf{X}_k^\top\mathbf{X}_k)\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}\right]}. \tag{4.9}$$

This means that rather than searching through every value of η for every k-fold, to find the best value of λ we just need to find $\hat{\eta}_k$ in every fold, since any other pair will result in a higher fold specific cross-validation error. Plugging in the $\hat{\eta}_k$ value for every value of λ, we can then find $\hat{\lambda}$ by minimizing the total cross-validation error through:

$$\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\eta}}_{\boldsymbol{k}}} = \arg\min_{\lambda} \sum_{k=1}^{K} \|\mathbf{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{\lambda,\hat{\eta}_k,-k}\|_2^2. \tag{4.10}$$

Once we find $\hat{\lambda}$, we find $\hat{\eta}$ by minimizing the total cross-validation error given $\hat{\lambda}$. The process is exactly the same as how to arrive at (4.9) only that rather than minimizing cross-validation error per fold, we do it for all folds at the same time, resulting in:

$$\hat{\eta} = \frac{\sum_{k=1}^{K}\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\hat{\lambda}\mathbf{I})^{-1}\hat{\mathbf{w}}\right]^\top(\mathbf{X}_k^\top\mathbf{y}_k) - \sum_{k=1}^{K}\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\hat{\lambda}\mathbf{I})^{-1}(\mathbf{X}_{-k}^\top\mathbf{y}_{-k})\right]^\top(\mathbf{X}_k^\top\mathbf{X}_k)(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\hat{\lambda}\mathbf{I})^{-1}\hat{\mathbf{w}}}{\sum_{k=1}^{K} n\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\hat{\lambda}\mathbf{I})^{-1}\hat{\mathbf{w}}\right]^\top(\mathbf{X}_k^\top\mathbf{X}_k)\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\hat{\lambda}\mathbf{I})^{-1}\hat{\mathbf{w}}\right]}. \tag{4.11}$$

This alternative process of finding $\hat{\lambda}$ and $\hat{\eta}$ is not perfect, but it does significantly reduce the computational burden. This in turn does have a real practical benefit, since it allows for a larger and more detailed grid of penalty parameters to be searched through. The one downside is that for tiny values of $\hat{\lambda}$, sometimes $\hat{\eta}_k$ and $\hat{\eta}$ are found to be similarly small negative values. In such a situation, the decision was made to artificially constrain η, setting it to 0.

Estimation of target estimates

Once $\hat{\mathbf{w}}$ and penalty parameters $\hat{\lambda}$ and $\hat{\eta}$ have been estimated, we estimate the ABTLE on the PS via the closed form solution $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} = (\mathbf{X}^\top\mathbf{X} + n\hat{\lambda}\mathbf{I})^{-1}(\mathbf{X}^\top\mathbf{y} + n\hat{\eta}\hat{\mathbf{w}})$ which was found in (3.21).

In addition to $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ four extra estimators are estimated, for which the relative performance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ is interesting to compare. First is the PS MLE $\hat{\boldsymbol{\beta}}$. This is the baseline case against which the other estimators' accuracy is benchmarked against. The second and third additional estimators are the reduced versions of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$, namely the RRE $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ and the DBTLE $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\lambda}}$. They are estimated using the same process as $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ but where we set $\eta = 0$ and $\eta = \lambda$ respectively. Finally, we also estimate $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ which is an estimator developed by Wiśniowski et al. (2020) using a conjugate prior based on the NPS. The idea of the conjugate prior is to check if the difference between $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{w}}$ is significant using Hotelling T^2 test. If they are found to be different the prior variance is set to be more uninformative, resulting in posterior estimates less influenced by the NPS. If they are not found to be different then the prior variance is set to be very informative, resulting in the posterior estimates being more influenced by the NPS. For a more technical explanation as well as the code for the estimation process of $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$, see Wiśniowski et al. (2020).

4.2.3 Evaluation

For every scenario and sample size, each estimation method is evaluated through the average root mean squared error (ARMSE). First, the root mean squared error (RMSE) is calculated by taking the squared difference between the estimated parameter $\hat{\beta}_{k,*}$ from any estimation method and its true counterpart β_k, averaging it across the 1000 iterations and then taking the square root of it. Then, because we are interested in the complete performance of $\hat{\boldsymbol{\beta}}_*$ we then average the RMSE across all k coefficients, resulting in the ARMSE:

$$\text{ARMSE}(\hat{\boldsymbol{\beta}}_*) = \frac{1}{k}\sum_{k=1}^{k}\sqrt{\frac{1}{R}\sum_{i=1}^{R}(\hat{\beta}_{k,*} - \beta_k)^2}. \tag{4.12}$$

Purely for presentation purposes, given that in these simulated conditions, the actual values are on an arbitrary scale, the final ARMSE results are converted to be relative to the ARMSE of $\hat{\boldsymbol{\beta}}$. To achieve this, every estimator is transformed and normalized via the formula:

$$\text{Relative ARMSE}(\hat{\boldsymbol{\beta}}_*) = \frac{\text{ARMSE}(\hat{\boldsymbol{\beta}}_*) - \text{ARMSE}(\hat{\boldsymbol{\beta}})}{\text{ARMSE}(\hat{\boldsymbol{\beta}})}. \tag{4.13}$$

Also of interest is the mean of absolute bias (MoAB) which is the average absolute bias across all iterations averaged across all coefficients. It is important to note that it is not the same as the mean absolute error, which refers to the average absolute error of each coefficient, but rather this refers to

$$\text{MoAB}(\hat{\boldsymbol{\beta}}_*) = \frac{1}{k}\sum_{k=1}^{k}\left|\frac{1}{R}\sum_{i=1}^{R}(\hat{\beta}_{k,*} - \beta_k)\right|. \tag{4.14}$$

This is done so that the bias across coefficients does not cancel out and gives the impression that the model performs better than it does. Similar to the ARMSE, for presentation purposes this is also presented in a relative manner, benchmarked against the MoAB of $\hat{\mathbf{w}}$ through the formula:

$$\text{Relative MoAB}(\hat{\boldsymbol{\beta}}_*) = \frac{\text{MoAB}(\hat{\boldsymbol{\beta}}_*) - \text{MoAB}(\hat{\mathbf{w}})}{\text{MoAB}(\hat{\mathbf{w}})} \tag{4.15}$$

4.2.4 Iterations

1000 iterations is a relatively large number of iterations, meaning that the results for each sample in each scenario should have converged before the maximum number of iterations is reached. However, to be sure that it is in fact sufficient, a post-hoc check on the convergence of AMSE (average non-square rooted MSE across all coefficients) values across iterations was conducted to check whether the results for all the estimators had converged.

The check was done by extracting the AMSE value for every estimator for every single iteration from a scenario. The convergence is then simply calculated as the difference in average AMSE for every additionally added iteration vis-à-vis using one less iteration (so using 3 vs 2 or 1000 vs 999). This check was only conducted on the population generated, with maximum multicollinearity (0.8), highest bias factor (3), lowest coefficient correlation (0.5), and highest residual variance (10). From this population scenario, only the PS of size 25 was considered. The reason for doing this was that given the most extreme scenario parameters, it was expected that the estimation process would vary the most across samples within this scenario, and thus be the slowest scenario to converge.

As shown in Figure 4.1, the AMSE difference diminishes with each additional iteration. Early increases significantly enhance the stability of the estimates, while later increases have only a marginal impact on the AMSE difference. Although there is no clear cutoff point, already before 1000 iterations the relative change has decreased significantly, and at 1000 iterations, the difference in AMSE for an additional iteration is smaller than 0.01 for every single estimator. Based on these results, it was decided that 1000 iterations was sufficient.

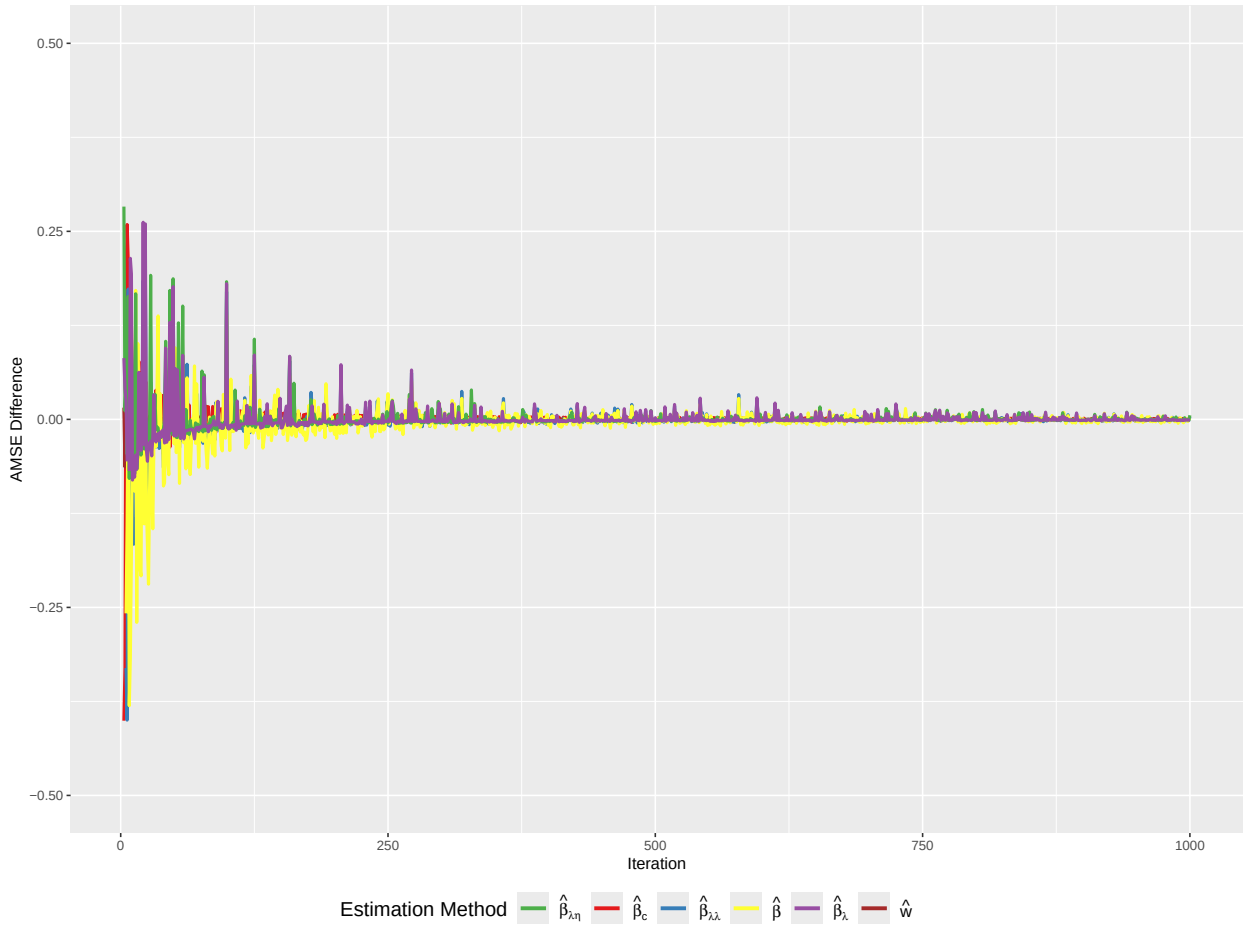


Figure 4.1: The change in average MSE for an increase in the number of iterations for each estimator. It is calculated as the difference in AMSE between using n and using n-1 iteration starting at 3 iterations and ending at 1000. The test was done using a population with a bias factor of 3, multicollinearity of 0.8, correlation of 0.5, residual variance of 10, and a PS size of 25. The initial AMSE difference is for some models a lot higher but to prevent scale distortion the AMSE difference visualized here is limited to ±0.5.

4.3 Software

The analysis was conducted using Rstudio version 2023.06.02 + 561, running R Windows version 4.2.3 (R Core Team, 2023). For general data wrangling "Tidyverse" was used (Wickham et al., 2019). "MASS" version 7.3-60 was used for simulating data while the library "sampling" version 2.10 was used for sampling the NPS (Tillé & Matei, 2023; Venables & Ripley, 2002). Analysis-wise, cross-validation was conducted with the help of the libraries "e1071" and "caret" (Kuhn, 2008; Meyer et al., 2023). The process was parallelized to speed up the analysis using the libraries "parallel" and "doParallel" (Microsoft Corporation & Weston, 2022; R Core Team, 2023). Finally, all visualizations were done using the libraries "ggplot2", "ggally", "cowplot", and "ggforce" (Pedersen, 2024; Schloerke et al., 2024; Wickham, 2016; Wilke, 2024). The full code for every step of the analysis process as well as the

data for the results presented can be found on the following GitHub page.

5. Results

5.1 Relative ARMSE

Figure 5.1 presents the main results of this research, which is the relative ARMSE to $\hat{\boldsymbol{\beta}}$ for the four estimators of interest. The results are relative, meaning that any value below 0 indicates that the estimator outperformed the $\hat{\boldsymbol{\beta}}$ by a certain percentage, whilst a positive value means that the estimator was worse than $\hat{\boldsymbol{\beta}}$. As the relative ARMSE approaches -1 it indicates an increasingly perfect estimate with 0 ARMSE. An important thing to note is that any change in relative performance does not necessarily mean that the estimator got better or worse, but could likewise be caused by a marginally larger improvement or deterioration of $\hat{\boldsymbol{\beta}}$. To see the absolute ARMSE estimates for each estimator see figure B.1 in Appendix B.

In terms of general relative ARMSE performance, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ performs well, although from figure 5.1 we can see that this performance is not even across scenarios, with there being pretty large differences depending on the population and sample parameters. Specifically for the scenarios involving $\sigma^2 = 2$ (top row of panels), $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ is only guaranteed to outperform all other estimators across sample sizes and provides a relative ARMSE decrease against $\hat{\boldsymbol{\beta}}$ as long as the correlation between $\hat{\mathbf{w}}$ and $\boldsymbol{\beta}$ is 0.9 (top right-most panel). If the correlation is lower than 0.9, then the $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ only really becomes a useful estimator in scenarios with high multicollinearity (≥ 0.5) and high bias ($f \geq 3$) (Bottom row and middle right in each panel). For all other scenarios, the $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ performs equivalent or worse than the other estimators. Furthermore, if neither factor is sufficiently large and $n_{ps} > 25$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ can even suffer from negative transfer. As can be seen in the top middle and top left panel, for larger samples $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ does sometimes showcase a positive relative ARMSE value, indicating that $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ is actually deteriorating the estimates relative to $\hat{\boldsymbol{\beta}}$. This likewise occurs for $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ with only $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ consistently outperforming $\hat{\boldsymbol{\beta}}$. However, looking at figure B.1 in the appendix, in all these scenarios the difference in absolute ARMSE values for the estimators and $\hat{\boldsymbol{\beta}}$ is practically indistinguishable. The relative performance decline as well as negative transfer seem to both not be as significant, and it seems to be largely driven by the fact that $\hat{\boldsymbol{\beta}}$ is already an accurate estimator. Nevertheless, these results are quite surprising and contrary to prior beliefs. As already mentioned in section 3.2, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ is a theoretical lower bound of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$'s ARMSE, meaning that it should never perform worse than it. The fact that this occurs indicates that the estimation of the penalty parameters was not accurate enough and that $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ probably over-penalizes.

For scenarios involving $\sigma^2 = 10$ (bottom row of panels), we see a more consistent good performance of the $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ with none of the scenarios resulting in a deteriorating performance of the estimator. Furthermore, for scenarios with a bias factor 1 or larger (middle and bottom row in each panel), $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ consistently performs best across all remaining parameters. When the bias is smaller than that, even though $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ outperforms its reduced versions, for small sample sizes, $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ is actually the best estimator. However, as sample sizes increase its relative performance decreases much faster than that of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$, meaning that at sample sizes of 75 and

100 $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ always outperforms the other estimators.

An increase in the sample size does tend to deteriorate the relative performance of the $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$, however the extent to which an increased sample does so is affected by the bias and multicollinearity, with lower values for each parameter expediting the relative decline. However again looking at the original ARMSE estimates in figure B.1, this is almost exclusively driven by an improvement of $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ performing consistently across sample sizes and values of bias and multicollinearity.

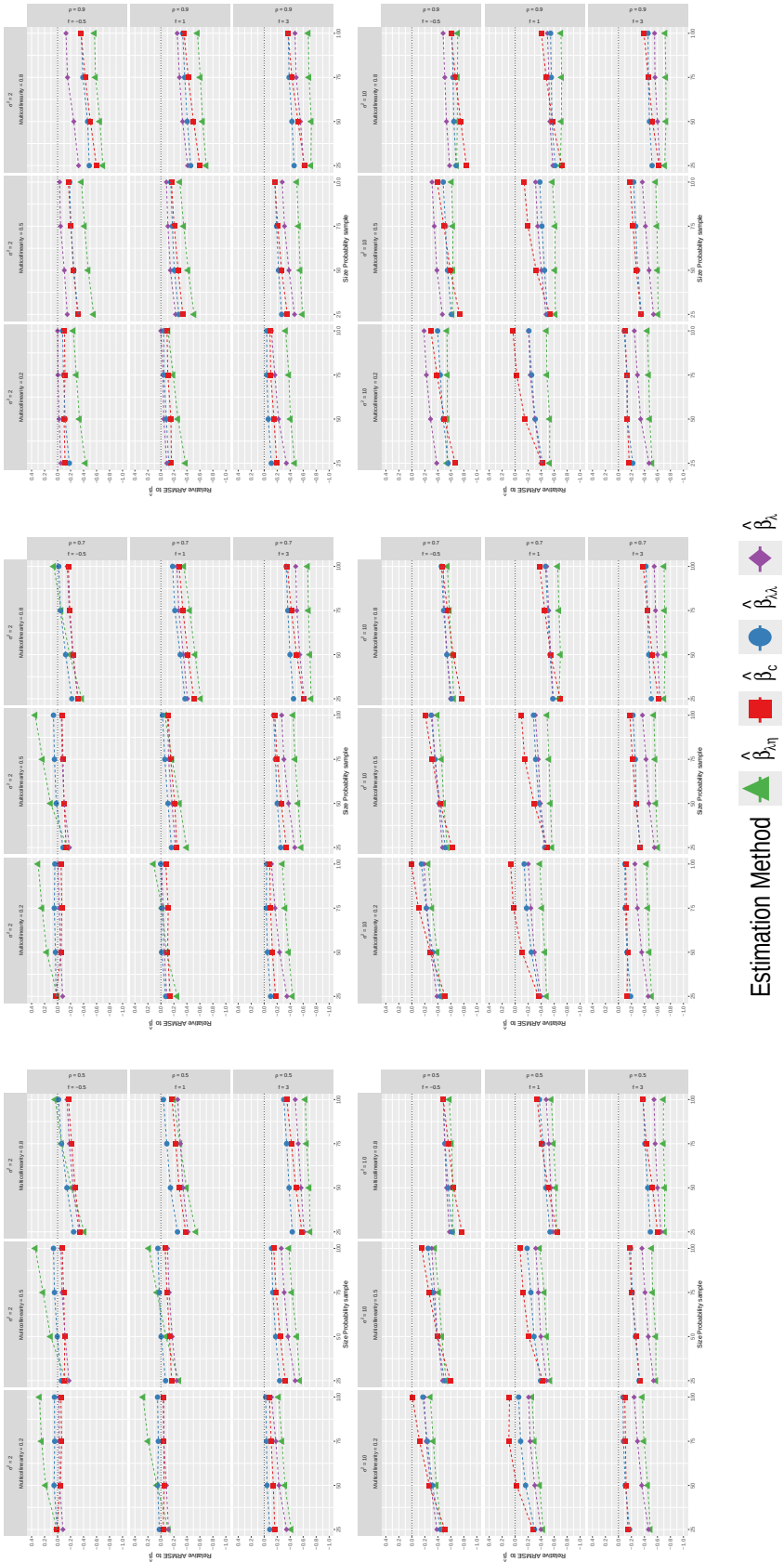


Figure 5.1: Average root mean squared error relative to $\hat{\boldsymbol{\beta}}$ for every type of estimator across every simulated scenario. Every scenario was replicated 1000 times by method of resampling.

5.2 Relative MoAB

The bias of the $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ is also important within official statistics, so figure 5.2 plots the relative MoAB of each estimator to that of $\hat{\mathbf{w}}$ to see to what degree its bias is incorporated. Again, the results are relative with a negative value indicating an improvement and a positive value a deterioration. As the MoAB approaches -1 it indicates a completely unbiased estimator. For the absolute MoAB estimates see figure B.2

For scenarios involving $\sigma^2 = 2$ (top row of panels) we see that for a correlation of 0.9 (top right panel), the degree of bias incorporated is very stable regardless of other parameters, incorporating only between 10-25% of the bias. This is however the case for all estimators in these scenarios, meaning that $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$, is rarely the best estimator but rather interchangeable or equal. If the correlation is lower than that (> 0.9) and the bias factor is smaller (< 3) the relative MoAB of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ actually tends to be the highest out of all the estimators (top row of the middle and leftmost panels on the top row), although outlier cases against this exist for when the sample size $n_{ps} = 25$. As the bias of the NPS estimator $\hat{\mathbf{w}}$ increases, the relative MoAB of all estimators tends to converge at a very low relative MoAB (bottom row of all top panels). However looking at the absolute values of MoAB, as with the ARMSE, this change seems to be largely driven by the fact the $\hat{\mathbf{w}}$ gets much worse rather than the other estimators actually getting better. Overall, the MoAB tends to stay consistent across sample sizes, but increasing the sample size can lead to a slight reduction in MoAB for scenarios with lower bias and multicollinearity.

For scenarios involving $\sigma^2 = 10$ (bottom row of panels), overall the trends in results are quite similar albeit sometimes the performance is (relatively) marginally better. For a correlation of 0.9 (bottom right panel), $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ tends to at lower values of bias result in a lower relative MoAB than the other estimators. However, as the bias of the NPS estimator $\hat{\mathbf{w}}$ increases, the results for all estimators converge at a very low relative MoAB (bottom row of all bottom panels). For a correlation lower than 0.9 and a lower bias factor, the results are mixed. Sometimes the $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ outperforms some, all, or no other estimators and it is very hard to discern a general trend. It is also important to note that the relative improvement against the other estimators is mainly driven by them reporting higher MoAB values than the MoAB value of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ actually decreasing (see figure B.2). A final thing to note from these results is also that the results are more stable across sample sizes, no longer showcasing a discernible decrease in relative MoAB for an increase in sample size.

These results are very encouraging since we see that the relative MoAB remains stable or decreases even as the bias factor increases, showcasing a degree of robustness against bias in the NPS. Looking at the absolute MoAB estimates in figure B.2 these results are further supported. Not only does the MoAB of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ stay pretty consistent for higher values of the bias factor, but as the correlation increases between $\hat{\mathbf{w}}$ and $\boldsymbol{\beta}$, the MoAB decreases.

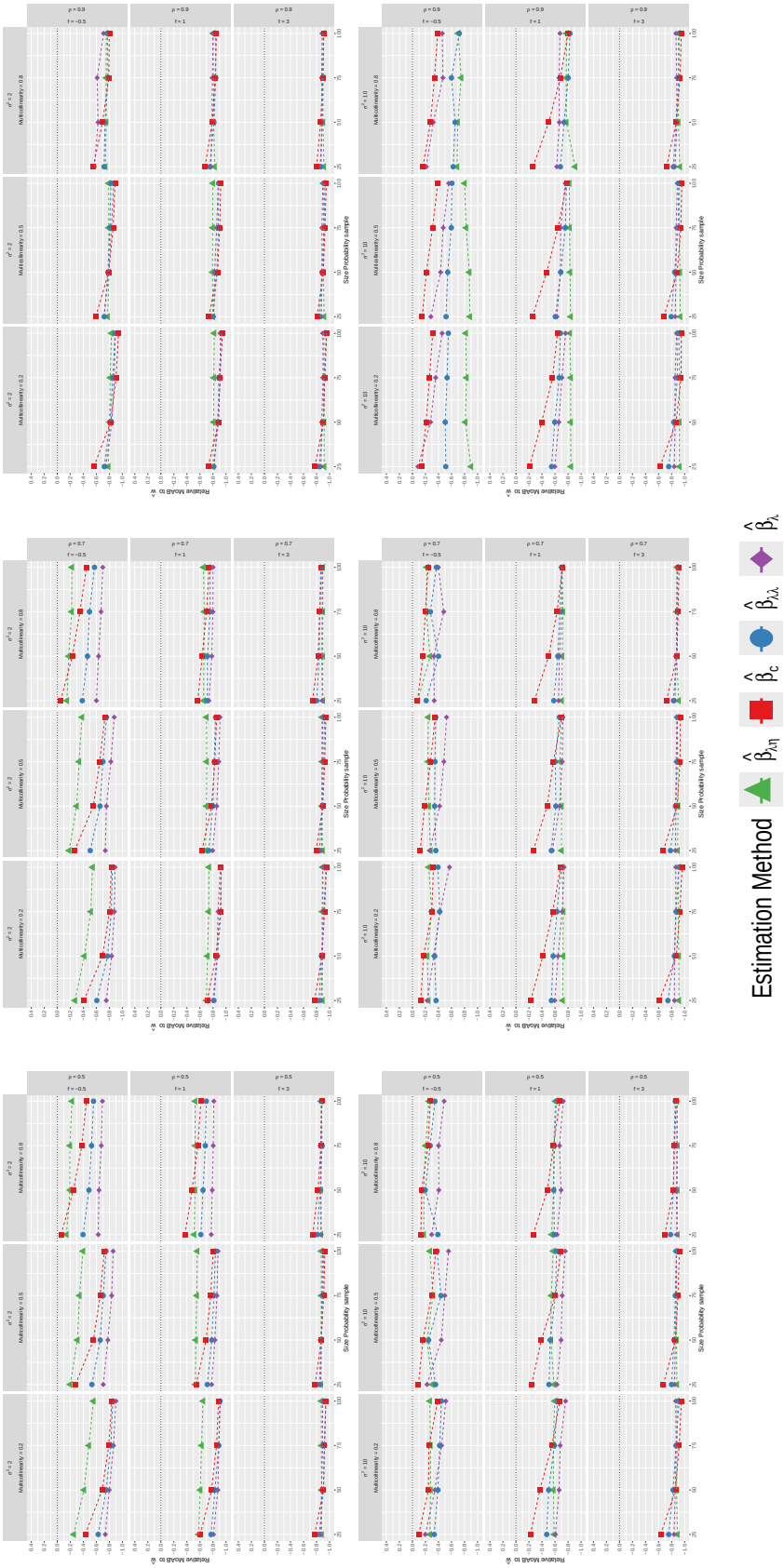


Figure 5.2: Mean of average bias relative to $\hat{\mathbf{w}}$ for every type of estimator across every simulated scenario. Every scenario was replicated 1000 times by method of resampling.

5.3 Penalty parameter estimation

Contrary to expectations, the results showcased a deteriorating relative ARMSE performance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ to $\hat{\boldsymbol{\beta}}$. Therfore, a short auxiliary analysis is also conducted on the estimation process of the penalty parameters since that is of empirical relevance for the estimation of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$. Because the optimal η^* depends directly on λ^*, showcasing both is unnecessary and the analysis limits itself to only looking at λ.

The occurrence of deteriorating relative ARMSE performance indicates that over-penalization is occurring, so figure 5.3 plots the proportion of times that $\hat{\lambda} > \lambda^*$ and $\hat{\lambda} = 10^5$. Here λ^* is the theoretically optimal value for lambda, following formula (3.22). From the figure, we can see that cross-validation tends to strongly and systematically overestimate $\hat{\lambda}$ relative to λ^* with every scenario seeing somewhere between 75 and 100 percent of its estimated λ values being higher than the optimal value. On average across all scenarios $\hat{\lambda}$ was larger than λ^* for 91% (910) of the iterations, as indicated by the black dashed line. Furthermore, we also see that for a large section of the scenarios, this systematic overestimation is partially or almost completely driven by the cross-validation estimating $\hat{\lambda} = 10^5$, which is the maximum value that transforms $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}} \approx c\hat{\mathbf{w}}$. The small cyclical downward spikes refer to an increase in sample size across every four scenarios (so 25-100), the larger thicker spikes are the groups of scenarios where the bias factor is the highest.

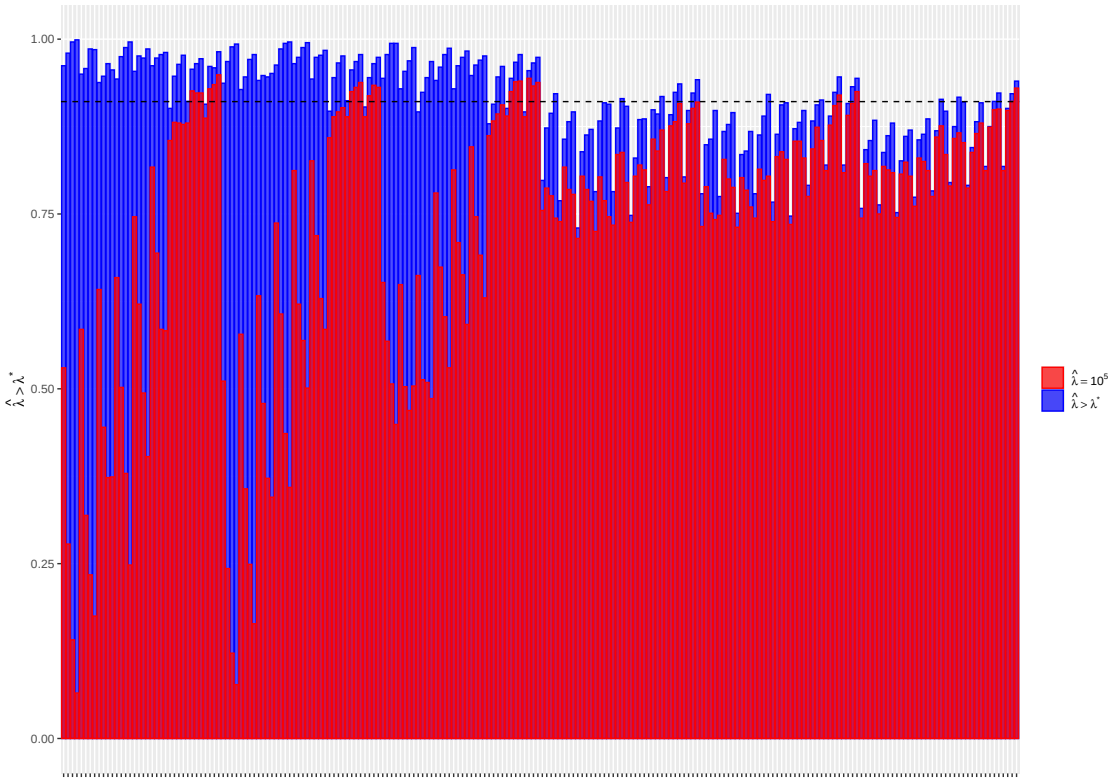


Figure 5.3: Proportion of iterations per scenario (X-axis) where $\hat{\lambda} > \lambda*$. The blue bar is the proportion of times $\hat{\lambda} > \lambda^*$ by any value. The red bar is the proportion of times where $\hat{\lambda} = 10^5$. The dashed line at 91% is the average proportion of $\hat{\lambda} > \lambda*$ across all scenarios.

The overpenalization does not have to be as problematic as long as it is relatively small or if the marginal loss is low for an increase in λ and η. Thus, to check the impact of over-penalization on the actual performance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$, figure 5.4 displays the relative ARMSE of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}^*,\boldsymbol{\eta}^*}$, and compares it to its empirical counterpart $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$, and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ (same as in figure 5.1). $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}^*,\boldsymbol{\eta}^*}$ is the theoretical ABLTE whose penalty parameters are calculated based on the scenario parameters from the formula (3.22), rather than estimated.

As seen in the figure, the loss of performance caused by estimating $\hat{\lambda}$ and $\hat{\eta}$ varies quite significantly across scenarios. Nonetheless, with the exception of the scenarios with $\sigma^2 = 2$, low bias, and high correlation, the difference is not insignificant with the theoretical estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}*,\boldsymbol{\eta}*}$ outperforming its empirical counterpart $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$ with as much as 25 percentage points. The inability to properly estimate the optimal penalty parameters from the sample itself can thus result in a major loss of performance for $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\eta}}$. Furthermore, we can also see that previously encountered negative transfer also could have been completely prevented, with $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}*,\boldsymbol{\eta}*}$ never resulting in a worse estimator than $\hat{\boldsymbol{\beta}}$ or $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$. This is a further indication that the practical task of estimating the optimal or at least near-optimal penalty parameters is not always possible using conventional cross-validation with it systematically over-penalizing, sometimes even to the detriment of the estimator.

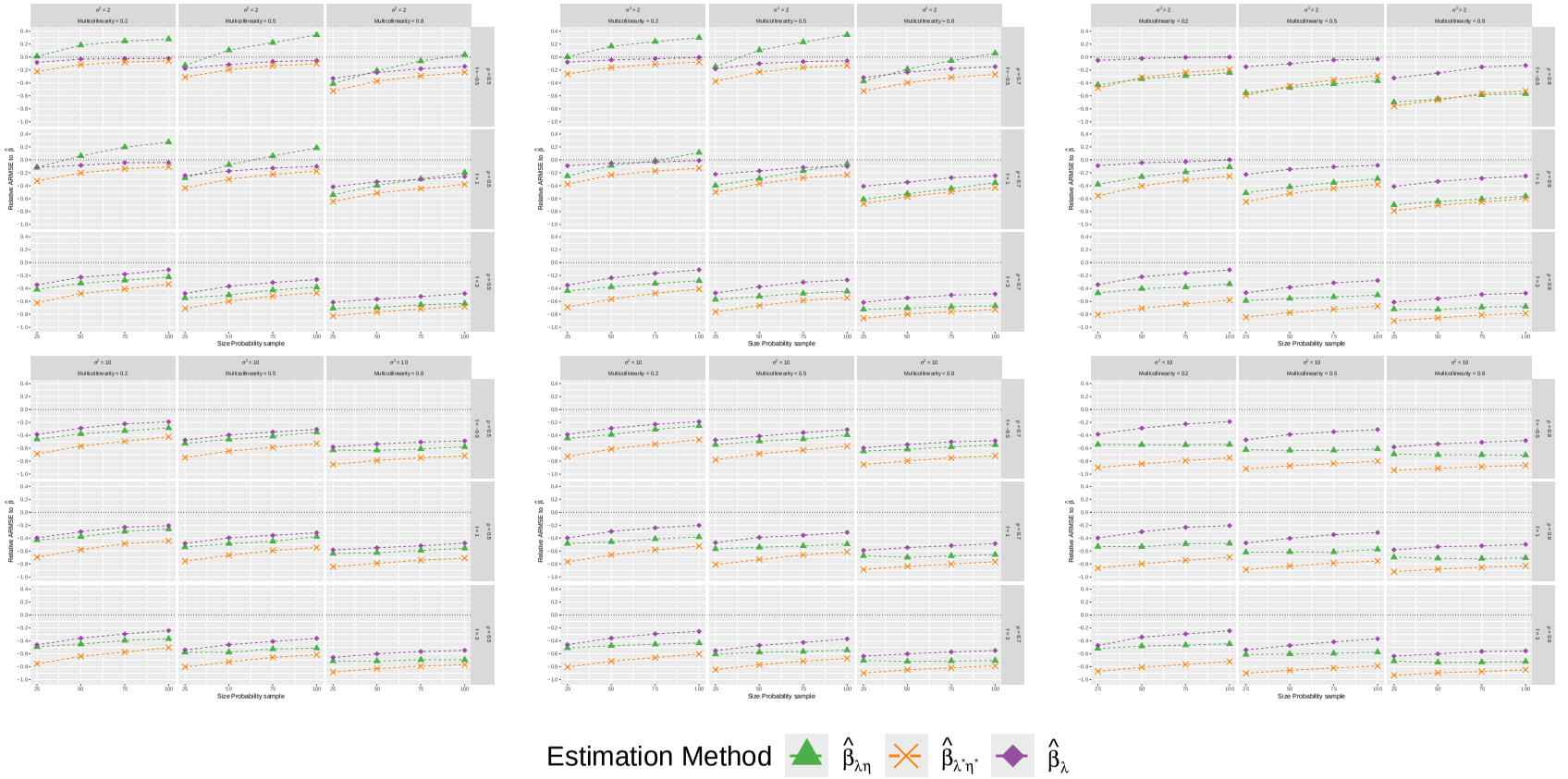


Figure 5.4: Average root mean squared error relative to $\hat{\beta}$ for the theoretical the ABTLE, empirical ABTLE, and RRE. Every scenario was replicated 1000 times by method of resampling.

6. Discussion

Summarizing the results of the analysis, the findings suggest that the ABTLE can aid in improving the accuracy of ARMSE relative to the PS estimates. The actual improvement however depended a lot on population and sample parameters. Generally, the ABTLE tended to perform better (relatively) in scenarios where the baseline PS estimates were less accurate (higher multicollinearity and larger residual variance). However, this was largely driven by the fact that the PS MLE performed worse with the actual performance of the ABTLE itself staying consistent. Part of this can be explained by the fact that $\hat{\lambda}$ tended to be quite large meaning that quite often the ABLTE simply turned into a rescaled version of the NPS MLE, regardless of sample size. However, this was not the case for all scenarios, and even when $\hat{\lambda}$ were much smaller the ARMSE results tended to be stable. Regarding the bias, the ABTLE never really outperformed the other estimators, consistently showcasing a higher or merely similar bias as the benchmarking methods. Nevertheless, this is not necessarily as problematic since this was most prevalent at smaller levels of bias. As the bias in the NPS estimates increased, the relative bias of all other estimators tended to converge at a really low relative bias to the NPS estimates, showcasing that the ABTLE is also robust when incorporating very biased NPS. The ABTLE only really outperformed the other estimators in scenarios of high correlation but even then the bias was comparable with other methods.

As anticipated, and consistent with the findings of Gu et al. (2022), the results indicate that significant performance improvements in relation to the RRE can be achieved by expanding the penalization scheme to not only penalize magnitudes but also reward aligning it angle-wise with auxiliary estimates. Unless the accuracy of the PS MLE was already accurate, then even with a flawed penalty parameter estimation this was shown to be true. The results also reveal that a distance-based penalization scheme lacks flexibility as is evident from the fact that the DBTLE tended to underperform compared to the ABTLE. It should be noted that, unlike the results of Gu et al. (2022), it remains unclear from this research whether this issue stems from over or under-penalization of the DBTLE. Finally, relating the results to the research of Wiśniowski et al. (2020), they are in line with their findings. We can see from the relative performance that although the Bayesian estimator works relatively well (sometimes even better than the ABTLE), it is vulnerable to bias in the NPS estimates. As the bias increases to the maximum, the Bayesian approach is never better than the ABTLE, offering further evidence of the limitations of their approach in specific scenarios.

Contrary to the results of Gu et al. (2022), this research also found that the ABTLE can result in negative transfer. Specifically, incorporating the NPS sometimes led to a deterioration in ARMSE relative to the RRE and PS MLE when the accuracy of the PS MLE was already high. As seen by the application of the theoretically optimal ABTLE, this was caused by over-penalization which was not reported as a potential problem in the research of Gu et al. (2022). Looking at the formula for the optimal value of λ (3.22), the likely cause behind both these differences lies in the different simulation designs employed here and by Gu et al. (2022). Fundamentally, there is a difference in the simulation design of the two datasets where for Gu et al. (2022) they are generated as two different populations with very small

residual variance and a set correlation between their corresponding estimands ($\boldsymbol{\beta}$ and $\mathbf{w}$). In this research, however, the two datasets were samples, each with an expected correlation and residual variance. Given that the PSs were quite small, both the empirical correlation and the empirical residual variance could differ substantially in practice from underlying true values. This difference in research design is significant because, as seen in the formulas (3.22), the optimal value of λ (and by extension also η) depends on both the correlation and residual variance, so when they shift even slightly, it can result in vastly different penalty parameters. The results for the type of simulation design employed in this thesis are thus much more unstable. By chance, it is possible (and was observed) that a sample (or a fold during cross-validation) exhibited a very high correlation or residual variance, inflating the model's perception of the optimal value of the penalty parameters. This in turn can lead to over-penalization and the occurrences of negative transfer, which does not occur if the theoretically optimal values are used. As outlined by Riley et al. (2021) there is always some variability for this type of penalized regression since the penalty parameters do have to be estimated in some way. If the sample is small, this can result in sub-optimal penalty parameters. The problem highlighted by this analysis is that, unlike the normal RRE, the ABTLE is a lot more sensitive to small changes in the data due to the inclusion of the additional correlation term. One should thus be careful in applying the ABTLE with a small sample size since it could adversely lead to a deteriorating estimator.

Tying the discussion back to the research question of **How can penalized regression be used to incorporate non-probability samples into official statistics?** the results do highlight a potential type of scenario where the ABTLE does fit. Given its relative success, the ABTLE seems to be a good alternative to the correction methods outlined in section 2.1 in scenarios where there are few to no additional covariates to explain the selectivity in the NPS (as here with $\boldsymbol{z}$) and where the bias of the NPS estimates are large. There is no such thing as the best estimator for every case (Hastie et al., 2009). However, given this type of scenario, correction methods cannot be used, the Bayesian approach and the composite estimator have been shown to be vulnerable to bias. Limited data, high bias scenarios can therefore be seen as a type of scenario where the ABLTE is most likely to be the best choice. Another more practical scenario where the ABTLE could be a good alternative is in situations where the auxiliary data is sensitive, and sharing micro-level data is not feasible. In such cases, sharing estimates might be more viable which benefits the ABTLE that unlike all other comparable methods only requires auxiliary estimates from the NPS.

However, in its current form, there are still clearly some scenarios where ABTLE is to be avoided. With small samples, the estimation of penalty parameters becomes increasingly uncertain, and it is not guaranteed to result in a performance increase. In such scenarios, it might then be preferable to simply use ridge regression since it is more stable, and unlike all other estimators was the only method that never led to deteriorating estimates. However one must then keep in mind that no actual data integration is then taking place and the outcome is just a regularized PS. Finally, it also needs to be noted that if uncertainty estimates are also required, the ABTLE should never really be applied, regardless of performance, since it (and penalized regression in general) does not come with standard errors directly. For the RRE, it is possible, e.g., by bootstrapping or the Edgeworth Expansion to approximate standard errors. However, these methods have never been applied to the AB-

TLE (Firinguetti & Bobadilla, 2011; Revan Özkale & Altuner, 2023). Even still, as argued by Goeman et al. (2022) even if possible, normal uncertainty estimates (standard errors, confidence intervals) reflect the uncertainty in estimates due to variance alone, making the assumption that the estimator is unbiased. This means that even if it is possible to produce uncertainty estimates for the ABTLE, they are of little use since they would provide an incomplete and potentially misleading picture of estimator precision, not accounting for the additional bias-related inaccuracies.

7. Conclusion

This thesis has sought to broaden our understanding of methods to integrate non-probability samples into official statistics by answering the question **How can penalized regression be used to incorporate non-probability samples into official statistics?** It did this by applying and evaluating the Angle-Based Transfer Learning Estimator (ABTLE) as an estimation method to incorporate information from an NPS in an informative way, with the goal of increasing the estimation accuracy without incorporating much of the NPS bias. It was done for various scenarios with different degrees of bias, correlation, residual variance, sample sizes, and multicollinearity. The relative improvement of the ABTLE was assessed against the PS maximum likelihood estimator and compared to its related estimators and the Bayesian method developed by Wiśniowski et al. (2020). The result suggests that the ABTLE can be a useful method to incorporate an NPS to increase the accuracy of an estimator, whilst still being robust against also incorporating too much of the bias from the NPS. However, the general relative performance differed with the ABTLE performing better in scenarios where the PS maximum likelihood estimator performed worse. With larger sample sizes in scenarios that favor PS estimate accuracy, the benefits of the ABTLE were marginal at best, or even sometimes detrimental because of sub-optimal penalty parameter estimation.

Before finishing the limitations of this thesis need to be discussed. First, the proposed method assumes that the probability sample is unbiased, or at least has a low bias compared to the non-probability sample. However, as already mentioned in Section 1., this assumption may not always apply to real-life PS due to non-response. Particularly when estimating on sensitive topics there is a risk that the non-response is selective and may result in biased PS estimates. If the PS is also biased then the protection against bias from the PS is no longer there and the ABTLE would likely result in a lot more biased estimates. A more general limitation in line with this is that the method was only evaluated on simulated data as a proof of concept. As such, the simulated conditions are quite limited and stylized, reducing the generalizability of the results. Applying the ABTLE to a real dataset could support the results by verifying that the estimator works in a more noisy environment with mild model violations (non-linearity, heteroscedasticity, etc).

The research and its results open up three direct avenues of research which relate to the limitations of this thesis, but also the wider theoretical framework of penalized regression. Firstly, it is already known that the normal RRE can be expressed in a Bayesian framework by assuming the regression coefficients have independent Gaussian priors with mean zero and a common variance, which acts as a regularization parameter. Future research should try to do the same for the ABTLE. Incorporating the ABTLE in a Bayesian framework could be useful since not only does it open up for alternative uncertainty measures such as credibility intervals, but it also could aid in finding a more stable method of estimating penalty parameters through formulating them as priors. The author is aware that this might not be a trivial task, and if not possible, then an alternative avenue of research would be to attempt to apply alternative methods of estimating the penalty parameters, such as AIC or BIC. However, they also rely on the likelihood function which likewise might not be easy to

derive. Neither method solves the underlying instability caused by potentially bad samples but they remove the sub-sampling step which exacerbates any sample issues. Secondly, the focus of this research was on a type of penalized regression based on ridge penalty. There is to the authors' knowledge no current angle-based Lasso-type penalty framework but Liang et al. (2020) have developed transLASSO which does incorporate distance information between $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{w}}$. Future research should try to apply this alternative penalization scheme and explore the possibility of extending the angle-based penalties to a Lasso-type penalization framework. Finally, as also outlined by Gu et al. (2022) the ABTLE can be extended to incorporate several sets of auxiliary estimators. This research limited itself to only utilizing one, but it would be interesting to see what would happen if you apply several sets with auxiliary estimators of varying correlation and degree of bias.

In conclusion, this thesis performed a simulation-based test of the Angle-Based Transfer Learning Estimator, focusing on issues of estimation in the context of official statistics to solve the problem of integrating NPSs into official statistics by combining it with a smaller PS. While the results only partially conformed to the expectations, this new information nevertheless allows us to move forward on NPS sample integration, especially in a context where few variables are available to explain the selectivity in the NPS and its bias is high. Ultimately, however, the estimation method provided here is no panacea and official statistics will continue to face problems in integrating NPSs and PSs, especially when the PS is small. Nevertheless, the research contained here offers an important next step to address these challenges long-term by introducing penalized regression as an alternative estimation approach that can be developed further.

Bibliography

Bakker, B. F., van Rooijen, J., & van Toor, L. (2014). The System of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, *30*(4), 411–424. https://doi.org/10.3233/SJI-140803

Bethlehem, J. (2009a). *Applied Survey Methods: A Statistical Perspective*. John Wiley & Sons.

Bethlehem, J. (2009b). *The rise of survey sampling*, Statistics Netherlands.

Chen, Y., Li, P., & Wu, C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, *115*(532), 2011–2021. https://doi.org/10.1080/01621459.2019.1677241

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. https://doi.org/10.1093/jssam/smz041

de Leeuw, E., & de Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse* (pp. 41–54). Wiley.

Disogra, C., Cobb, C., Chan, E., & Dennis, J. M. (2011). Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics. *Section on Survey Reseach Methods – JSM Proceedings*, 4501–4515.

Elliott, M., & Haviland, A. (2007). Use of a Web-Based Convenience Sample to Supplement a Probability Sample. *Survey Methodology*, *33*(2), 211–215.

Elliott, M., & Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science*, *32*(2), 249–264. https://doi.org/10.1214/16-STS598

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences.* Thomson Brooks/Cole Publishing Co.

Firinguetti, L., & Bobadilla, G. (2011). Asymptotic confidence intervals in ridge regression based on the Edgeworth expansion. *Statistical Papers*, *52*(2), 287–307. https://doi.org/10.1007/s00362-009-0229-5

Goeman, J., Meijer, R., & Chaturvedi, N. (2022). L1 and L2 Penalized Regression Models. https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf

Groß, J. (2003). *Linear regression*. Springer.

Gu, T., Han, Y., & Duan, R. (2022). Robust angle-based transfer learning in high dimensions. https://arxiv.org/abs/2210.12759

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Elements of Statistical Learning Data Mining, Inference, and Prediction* (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer. https://doi.org/https://doi.org/10.1007/978-1-0716-1418-1

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods and Research*, *37*(3), 319–343. https://doi.org/10.1177/0049124108329643

Li, C., Yang, C., Gelernter, J., & Zhao, H. (2014). Improving genetic risk prediction by leveraging pleiotropy. *Human Genetics*, *133*(5), 639–650. https://doi.org/10.1007/s00439-013-1401-5

Li, S., Cai, T. T., & Li, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *84*(1), 149–173. https://doi.org/10.1111/rssb.12479

Liang, M., Park, J., Lu, Q., & Zhong, X. (2020). Robust and flexible learning of a high-dimensional classification rule using auxiliary outcomes. http://arxiv.org/abs/2011.05493

Luiten, A., Hox, J., & de Leeuw, E. (2020). Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys. *Journal of Official Statistics*, *36*(3), 469–487. https://doi.org/10.2478/jos-2020-0025

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2023). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. https://CRAN.R-project.org/package=e1071

Microsoft Corporation & Weston, S. (2022). doParallel: Foreach Parallel Adaptor for the 'parallel' Package. https://CRAN.R-project.org/package=doParallel

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Pan, Y., Cai, W., & Liu, Z. (2022). Inference for non-probability samples under high-dimensional covariate-adjusted superpopulation model. *Statistical Methods and Applications*, *31*(4), 955–979. https://doi.org/10.1007/s10260-021-00619-w

Pedersen, T. L. (2024). ggforce: Accelerating 'ggplot2'. https://CRAN.R-project.org/package=ggforce

R Core Team. (2023). R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Revan Özkale, M., & Altuner, H. (2023). Bootstrap confidence interval of ridge regression in linear regression model: A comparative study via a simulation study. *Communications in Statistics - Theory and Methods*, *52*(20), 7405–7441. https://doi.org/10.1080/03610926.2022.2045024

Riley, R. D., Snell, K. I., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., & Collins, G. S. (2021). Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *Journal of Clinical Epidemiology*, *132*, 88–96. https://doi.org/10.1016/j.jclinepi.2020.12.005

Saleh, A. K. M. E., Kibria, B. M. G., Arashi, M., & Kibria, G. (2019). *Theory of ridge regression estimation with applications*. John Wiley & Sons.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2024). GGally: Extension to 'ggplot2'. https://CRAN.R-project.org/package=GGally

Seng, Y. P. (1951). Historical Survey of the Development of Sampling Theories and Practice. *Journal of the Royal Statistical Society. Series A (General)*, *114*(2), 214–231. https://doi.org/10.2307/2980977

Smit, V. (2021). *Correcting Selectivity in Datasets with Pseudo-Weights: a Simulation Study*. [Master Thesis, Universiteit Leiden].

Smith, T. M. F. (1976). The Foundations of Survey Sampling: A Review. *Journal of the Royal Statistical Society. Series A (General)*, *139*(2), 183. https://doi.org/10.2307/2345174

Theobald, C. M. (1974). Generalizations of Mean Square Error Applied to Ridge Regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(1), 103–106. https://about.jstor.org/terms

Tian, Y., & Feng, Y. (2023). Transfer Learning Under High-Dimensional Generalized Linear Models. *Journal of the American Statistical Association*, *118*(544), 2684–2697. https://doi.org/10.1080/01621459.2022.2071278

Tillé, Y., & Matei, A. (2023). sampling: Survey Sampling. https://CRAN.R-project.org/package=sampling

Valliant, R. (2020). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*, *8*(2), 231–263. https://doi.org/10.1093/jssam/smz003

Valliant, R., Dever, J. A., & Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples* (2nd ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-93632-1

van den Brakel, J. (2019). *New data sources and inference methods for statistics*, Statistics Netherlands. https://www.cbs.nl/en-gb/background/2019/27/new-data-sources-and-inference-methods-for-statistics

van Wieringen, W. N. (2023). Lecture notes on ridge regression. https://doi.org/https://doi.org/10.48550/arXiv.1509.09169

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Villalobos Aliste, S. (2022). *Combining Probability and Nonprobability Samples on an Aggregated Level*. [Master Thesis, Utrecht University].

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilke, C. O. (2024). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. https://CRAN.R-project.org/package=cowplot

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., & Blom, A. G. (2020). Integrating Probability and Nonprobability Samples for Survey Inference. *Journal of Survey Statistics and Methodology*, *8*(1), 120–147. https://doi.org/10.1093/jssam/smz051

A. Appendix derivations

Full derivation of $Var[\hat{\boldsymbol{\beta}}] - Var[\hat{\boldsymbol{\beta}}_\lambda]$ Adapted from van Wieringen (2023).

$$
\begin{aligned}
&Var[\hat{\boldsymbol{\beta}}] - Var[\hat{\boldsymbol{\beta}}_\lambda] \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} - \sigma^2\mathbf{W}_\lambda(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{W}_\lambda^\top \\
&= \sigma^2\mathbf{W}_\lambda\left\{[\mathbf{I}+\lambda(\mathbf{X}^\top\mathbf{X})^{-1}](\mathbf{X}^\top\mathbf{X})^{-1}[\mathbf{I}+\lambda(\mathbf{X}^\top\mathbf{X})^{-1}]^\top - (\mathbf{X}^\top\mathbf{X})^{-1}\right\}\mathbf{W}_\lambda^\top \\
&= \sigma^2\mathbf{W}_\lambda\left[2\lambda(\mathbf{X}^\top\mathbf{X})^{-2} + \lambda^2(\mathbf{X}^\top\mathbf{X})^{-3}\right]\mathbf{W}_\lambda \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\left[2\lambda\mathbf{I} + \lambda^2(\mathbf{X}^\top\mathbf{X})^{-1}\right](\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1^\top}
\end{aligned}
\tag{A.1}
$$

Full derivation of expressing $Var[\hat{\boldsymbol{\beta}}_\lambda]$ through eigendecomposition, Adapted from Saleh et al. (2019).

The third step in the derivation process relies on the eigendecomposition where $\mathbf{X}^\top\mathbf{X}$ and $(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}$ are expressed as $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top$ and $\mathbf{Q}(\boldsymbol{\Lambda}+\lambda\mathbf{I})^{-1}\mathbf{Q}^\top$ respectively. Here $\mathbf{Q}$ is a $p \times p$ matrix containing the eigenvectors whilst $\boldsymbol{\Lambda}$ is a diagonal matrix containing the corresponding eigenvalues λ_i. Since $\boldsymbol{\Lambda}$ is just a diagonal matrix it can simply be rewritten as $\text{diag}(\lambda_i)$ and its inverse as $\text{diag}(1/\lambda_i)$ (or $\text{diag}(1/(\lambda_i+\lambda))$, which allows for further simplification.

$$
\begin{aligned}
\text{Var}[\boldsymbol{\beta}_\lambda] &= \text{Var}(\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) = \mathbf{W}_\lambda\text{Var}(\hat{\boldsymbol{\beta}})\mathbf{W}_\lambda^\top \\
&= \sigma^2\mathbf{W}_\lambda(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{W}_\lambda^\top \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1} \\
&= \sigma^2\left(\mathbf{Q}(\boldsymbol{\Lambda}+\lambda\mathbf{I})^{-1}\mathbf{Q}^\top\right)\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top\left(\mathbf{Q}(\boldsymbol{\Lambda}+\lambda\mathbf{I})^{-1}\mathbf{Q}^\top\right) \\
&= \sigma^2\mathbf{Q}\left[\text{diag}\left(\frac{\lambda_i}{(\lambda_i+\lambda)^2}\right)\right]\mathbf{Q}^\top
\end{aligned}
\tag{A.2}
$$

Full derivation of $MSE[\hat{\boldsymbol{\beta}}_\lambda]$ Adapted from van Wieringen (2023).

$$
\begin{aligned}
&MSE[\hat{\boldsymbol{\beta}}_\lambda] = \mathbb{E}[(\mathbf{W}_\lambda\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2] \\
=&\mathbb{E}(\hat{\boldsymbol{\beta}}^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) - \mathbb{E}(\boldsymbol{\beta}^\top\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) - \mathbb{E}(\hat{\boldsymbol{\beta}}^\top\mathbf{W}_\lambda^\top\boldsymbol{\beta}) + \mathbb{E}(\boldsymbol{\beta}^\top\boldsymbol{\beta}) \\
=&\mathbb{E}(\hat{\boldsymbol{\beta}}^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) - \mathbb{E}(\boldsymbol{\beta}^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) - \mathbb{E}(\hat{\boldsymbol{\beta}}^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda\boldsymbol{\beta}) + \mathbb{E}(\boldsymbol{\beta}^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda\boldsymbol{\beta}) \\
&- \mathbb{E}(\boldsymbol{\beta}^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda\boldsymbol{\beta}) + \mathbb{E}(\boldsymbol{\beta}^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) + \mathbb{E}(\hat{\boldsymbol{\beta}}^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda\boldsymbol{\beta}) \\
&- \mathbb{E}(\boldsymbol{\beta}^\top\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) - \mathbb{E}(\hat{\boldsymbol{\beta}}^\top\mathbf{W}_\lambda^\top\boldsymbol{\beta}) + \mathbb{E}(\boldsymbol{\beta}^\top\boldsymbol{\beta}) \\
=&\mathbb{E}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})] \\
&- \boldsymbol{\beta}^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda\boldsymbol{\beta} \\
&- \boldsymbol{\beta}^\top\mathbf{W}_\lambda\boldsymbol{\beta} - \boldsymbol{\beta}^\top\mathbf{W}_\lambda^\top\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{\beta} \\
=&\mathbb{E}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\mathbf{W}_\lambda^\top\mathbf{W}_\lambda(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})] + \boldsymbol{\beta}^\top(\mathbf{W}_\lambda-\mathbf{I})^\top(\mathbf{W}_\lambda-\mathbf{I})\boldsymbol{\beta} \\
=&\sigma^2\,\text{tr}[\mathbf{W}_\lambda(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{W}_\lambda^\top] + \boldsymbol{\beta}^\top(\mathbf{W}_\lambda-\mathbf{I})^\top(\mathbf{W}_\lambda-\mathbf{I})\boldsymbol{\beta}
\end{aligned}
\tag{A.3}
$$

Full derivation of $\mathrm{M}(\hat{\boldsymbol{\beta}}) - \mathrm{M}(\hat{\boldsymbol{\beta}}_\lambda)$ Adapted from van Wieringen (2023).

$$\begin{aligned}
&\mathrm{M}(\hat{\boldsymbol{\beta}}) - \mathrm{M}(\hat{\boldsymbol{\beta}}_\lambda) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} - \sigma^2\mathbf{W}_\lambda(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{W}_\lambda^\top - (\mathbf{W}_\lambda \\
&\quad - \mathbf{I})\boldsymbol{\beta}\boldsymbol{\beta}^\top(\mathbf{W}_\lambda - \mathbf{I})^\top \\
=&\sigma^2\mathbf{W}_\lambda(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{W}_\lambda^\top \\
&\quad - \sigma^2\mathbf{W}_\lambda(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{W}_\lambda^\top - (\mathbf{W}_\lambda - \mathbf{I})\boldsymbol{\beta}\boldsymbol{\beta}^\top(\mathbf{W}_\lambda - \mathbf{I})^\top \\
=&\sigma^2\mathbf{W}_\lambda[2\lambda(\mathbf{X}^\top\mathbf{X})^{-2} + \lambda^2(\mathbf{X}^\top\mathbf{X})^{-3}]\mathbf{W}_\lambda^\top \\
&\quad - \lambda^2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}\boldsymbol{\beta}^\top[(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}]^\top \\
=&\sigma^2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}[2\lambda\mathbf{I} + \lambda^2(\mathbf{X}^\top\mathbf{X})^{-1}][(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}]^\top \\
&\quad - \lambda^2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}\boldsymbol{\beta}^\top[(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}]^\top \\
=&\lambda(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}[2\sigma^2\mathbf{I} + \lambda\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} - \lambda\boldsymbol{\beta}\boldsymbol{\beta}^\top][(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}]^\top
\end{aligned} \tag{A.4}$$

Full derivation of $\hat{\eta}$ and η^*

Here we are assuming a cross-validation loss function of:

$$CV(\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k}, \lambda, \eta) = \|\mathbf{Y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k}\|_2^2 \tag{A.5}$$

and a fit function of:

$$\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k} = (\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}(\mathbf{X}_{-k}^\top\mathbf{Y}_{-k} + n\eta\hat{\mathbf{w}}) \tag{A.6}$$

To find $\hat{\eta}$ we first expand equation (A.5) and simplify its terms.

$$\|\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k}\|_2^2 = \mathbf{Y}_k^T\mathbf{Y}_k - 2\mathbf{Y}_k^T\mathbf{X}_k\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k} + \hat{\boldsymbol{\beta}}_{\lambda,\eta,-k}^T\mathbf{X}_k^T\mathbf{X}_k\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k} \tag{A.7}$$

We then substitute $\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k}$ in equation (A.7) for its closed form solution in equation (A.6) and re-express (A.5) as a function of a series of scalars and η:

$$\begin{aligned}
\|\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k}\|_2^2 =& \mathbf{Y}_k^T\mathbf{Y}_k - 2\mathbf{Y}_k^T\mathbf{X}_k\left((\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}(\mathbf{X}_{-k}^\top\mathbf{Y}_{-k} + n\eta\hat{\mathbf{w}})\right) \\
&+ \left((\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}(\mathbf{X}_{-k}^\top\mathbf{Y}_{-k} + n\eta\hat{\mathbf{w}})\right)^\top\mathbf{X}_k^\top\mathbf{X}_k \\
&\left((\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}(\mathbf{X}_{-k}^\top\mathbf{Y}_{-k} + n\eta\hat{\mathbf{w}})\right) \\
=&\mathbf{Y}_k^T\mathbf{Y}_k - 2\mathbf{Y}_k^T\mathbf{X}_k(\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}\mathbf{X}_{-k}^\top\mathbf{Y}_{-k} \\
&+ \left((\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}\mathbf{X}_{-k}^\top\mathbf{Y}_{-k}\right)^T\mathbf{X}_k^T\mathbf{X}_k\left((\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}\mathbf{X}_{-k}^\top\mathbf{Y}_{-k}\right) \\
&- 2\eta\mathbf{Y}_k^T\mathbf{X}_k(\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}n\hat{\mathbf{w}} \\
&+ 2\eta\left((\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}\mathbf{X}_{-k}^\top\mathbf{Y}_{-k}\right)^T\mathbf{X}_k^T\mathbf{X}_k(\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}n\hat{\mathbf{w}} \\
&+ \eta^2\left(n^2\hat{\mathbf{w}}^T(\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}\mathbf{X}_k^T\mathbf{X}_k(\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}\right)
\end{aligned} \tag{A.8}$$

We then take the derivative of equation (A.8) with respect to η.

$$\begin{aligned}\frac{\partial}{\partial\eta}\|\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\beta}}_{\lambda,\eta,-k}\|_2^2 = \\ &- 2\mathbf{Y}_k^T\mathbf{X}_k(\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}n\hat{\mathbf{w}} \\ &+ 2\left((\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}\mathbf{X}_{-k}^\top\mathbf{Y}_{-k}\right)^T\mathbf{X}_k^T\mathbf{X}_k(\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}n\hat{\mathbf{w}} \\ &+ 2\eta\left(n^2\hat{\mathbf{w}}^T(\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}\mathbf{X}_k^T\mathbf{X}_k(\mathbf{X}_{-k}^\top\mathbf{X}_{-k} + n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}\right)\end{aligned} \tag{A.9}$$

Solving for η it is then possible to find an expression for $\hat{\eta}$

$$\hat{\eta}=\frac{\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}\right]^\top(\mathbf{X}_k^\top\mathbf{Y}_k)-\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}(\mathbf{X}_{-k}^\top\mathbf{Y}_{-k})\right]^\top(\mathbf{X}_k^\top\mathbf{X}_k)(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}}{n\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}\right]^\top(\mathbf{X}_k^\top\mathbf{X}_k)\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}\right]} \tag{A.10}$$

To get η^* the same process is conducted although across all k-folds rather than for a k-fold. Thus showing the steps is redundant since it is the same as above $\hat{\eta}$ but with sums across every term. Skipping through the process, the final expression for η^* is:

$$\eta^*=\frac{\sum_{k=1}^K\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}\right]^\top(\mathbf{X}_k^\top\mathbf{Y}_k)-\sum_{k=1}^K\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}(\mathbf{X}_{-k}^\top\mathbf{Y}_{-k})\right]^\top(\mathbf{X}_k^\top\mathbf{X}_k)(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}}{\sum_{k=1}^K n\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}\right]^\top(\mathbf{X}_k^\top\mathbf{X}_k)\left[(\mathbf{X}_{-k}^\top\mathbf{X}_{-k}+n\lambda\mathbf{I})^{-1}\hat{\mathbf{w}}\right]} \tag{A.11}$$

B. Appendix results

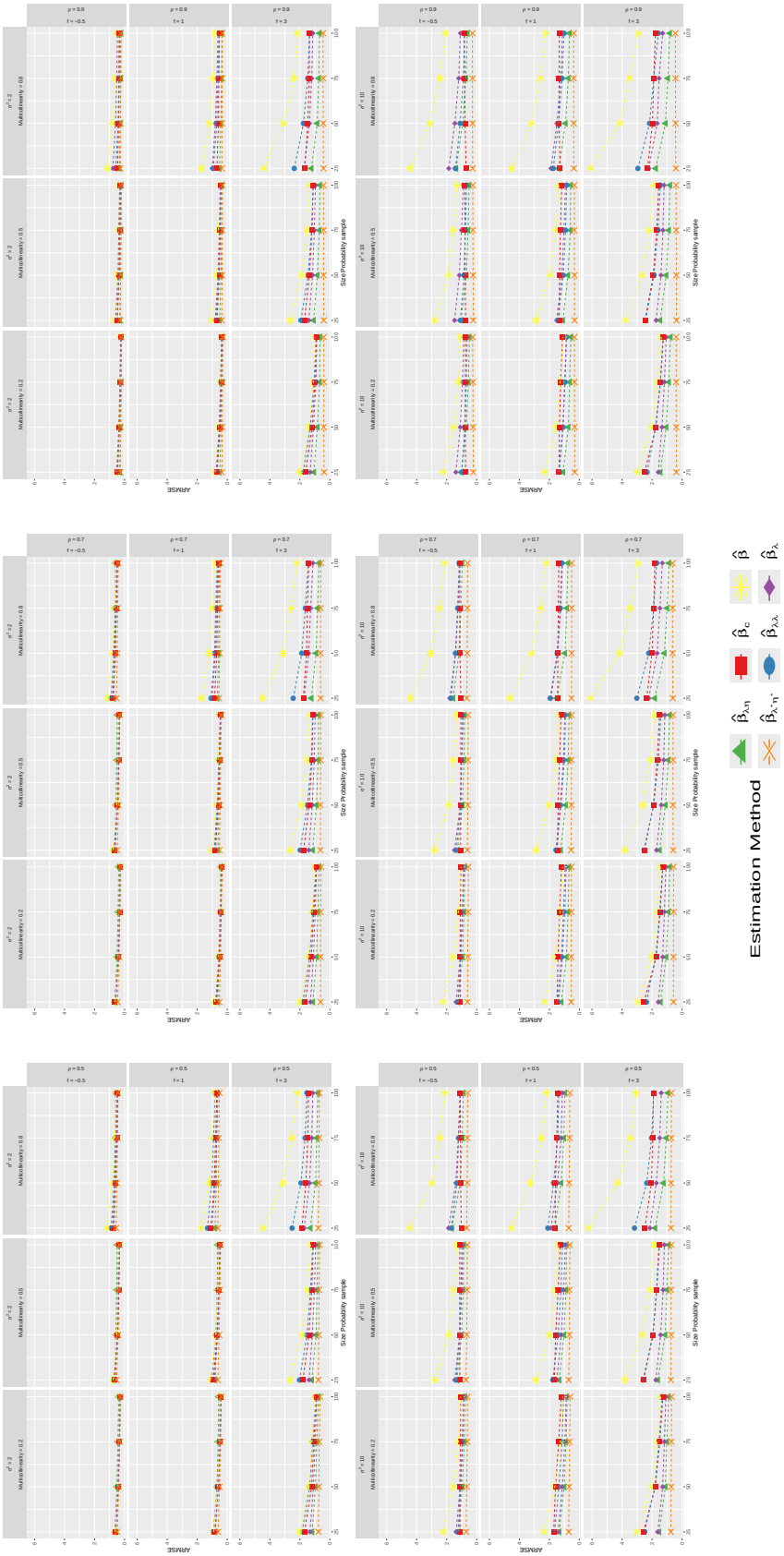


Figure B.1: Average Root Mean Squared Error for every type of estimator across every simulated scenario. Every scenario was replicated 1000 times by method of resampling

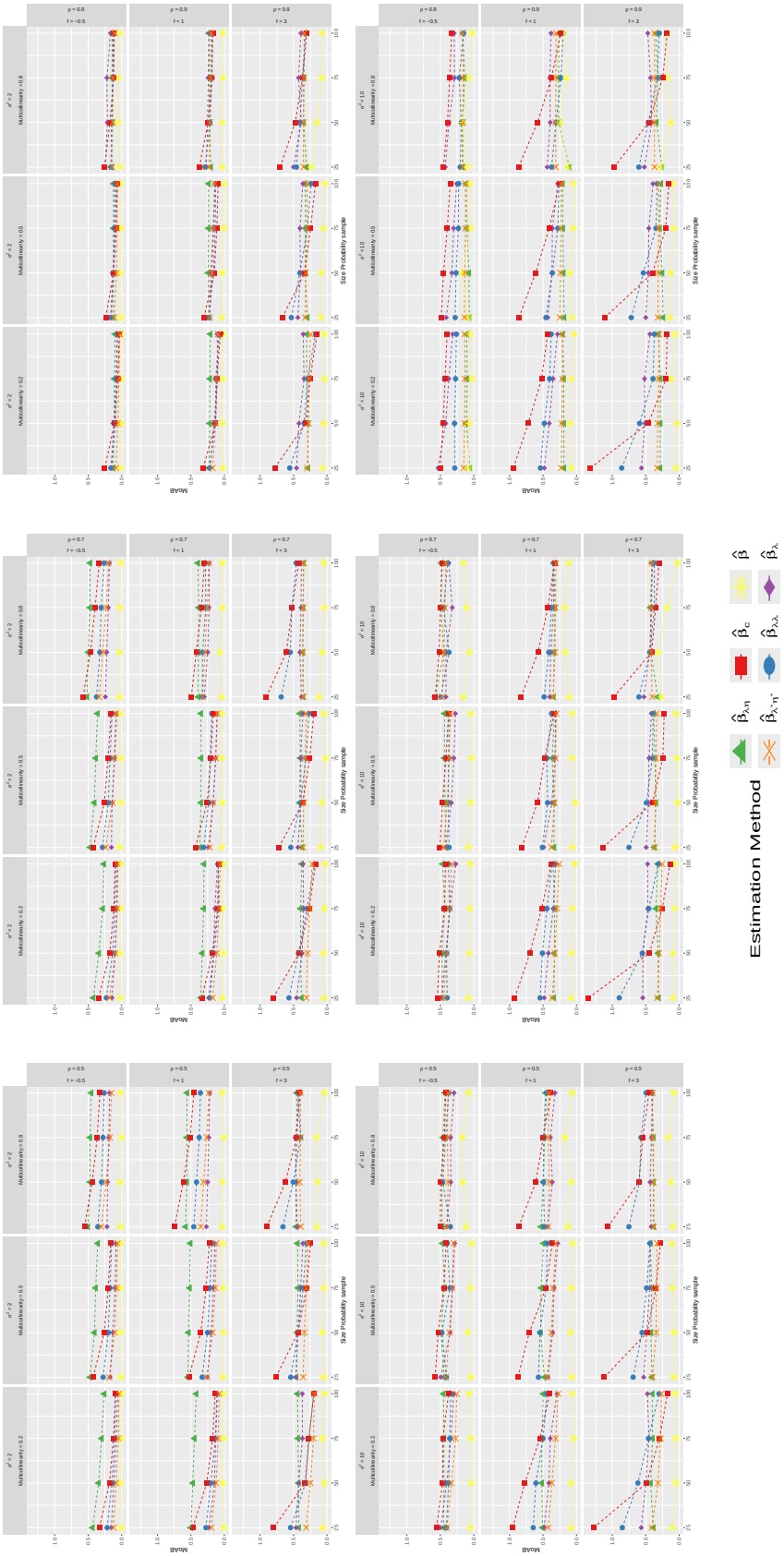


Figure B.2: Mean of average bias for every type of estimator across every simulated scenario. Every scenario was replicated 1000 times by method of resampling