UNIVERSITY OF SOUTHAMPTON

Faculty of Social Sciences
School of Economic, Social and Political Sciences

Three Essays in High Dimensional Linear Regression in Economics

by

Richard John Thorburn
BSc Economics & Actuarial Science (University of Southampton)
MSc Social Research Methods (University of Southampton)

A thesis for the degree of
Doctor of Philosophy

October 22, 2024

University of Southampton

Abstract

Faculty of Social Sciences
School of Economic, Social and Political Sciences

Doctor of Philosophy

title

by Richard John Thorburn

This thesis provides an overview of methods used for forecasting economic variables with high dimensional data sets. Based on these findings, from here, it proposes new approaches by modifying existing methods with justification based on statistical theory and tests with simulated and real-world data.

The first essay explores how the issue of high dimensional data sets is consistently increasing the challenges in both a logistical and analytical sense for any field relying on quantitative analysis. In economics this is primarily in the form of least-squares based approaches being unreliable or infeasible for linear regression frameworks in addition to the difficulties already faced by economists, such as significant predictor correlation and temporal dependence in time series covariates. Through finely tuned simulation experiments, this essay compares the prediction accuracy of existing high dimensional linear regression methods in a forecasting setting. As more correlation is induced between the covariates the results are very close with Principal Components and Ridge Regression, with the former appearing to hold a slight edge. While in the temporal dependence setting, Random Projections from the machine learning literature clearly dominates as the variables approach the unit root mark.

The second essay addresses how high dimensional data sets are becoming increasingly present to econometricians combined with predictor sets that are characterized by significant correlation amongst the covariates. While the famous Ridge Regression of Hoerl and Kennard (1970) provides a neat feasible alternative to Ordinary Least Squares, parameter estimates can suffer through their bias when the sign and magnitude of true coefficients vary significantly. To overcome such a setback, this paper proposes the Partial Ridge and Hybrid Estimation Procedure approach that vary which predictor coefficients face penalisation allowing a more desirable bias-variance tradeoff to be achieved for prediction purposes. Through theoretical analysis, a Monte Carlo Simulation study and an application to Iranian residential property price data, it is shown that while Partial Ridge alone is unable to universally dominate Ridge Regression, combining the two estimation procedures can lead to improved predictive accuracy that can outperform Ridge alone.

The final essay focuses on a method known as Random Projections and is seen as a computationally faster alternative to the widely used Principal Components Analysis for dimension reduction. However, while there is extensive evidence of work focusing on creating dependent variable predictions and forecasts using Random Projections, there is very little focus on its use for constructing estimates of the individual coefficient values themselves. Despite this, there is broad area within the statistics literature that involves finding the non-parametric bootstrap distribution of the coefficients. However, the distribution used in Random Projections is a parametric one, and allows the utilisation of additional key theoretical results to assist with estimation accuracy. Through the use of a Frisch-Waugh style approach, the so-called Partial Random Projections is proposed as a way to obtain individual parameter estimates in high dimensional settings whilst remaining in the spirit of Random Projections, which is seen to perform very well in the first essay. The theoretical analysis shows how its bias can often improve upon many other techniques when the sum of all other parameters except the one of interest is small and its associated covariate has a small amount of correlation with the others. Finally, a Monte Carlo Simulation study replicating a causal inference study demonstrates how this approach can practically provide more accurate estimates than other competing models.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signed:..... Date:.....

Acknowledgements

I would like to thank a number of individuals and groups of people for providing me with both academic and mental support throughout this extraordinary experience. Firstly, my supervisor, Professor Jean-Yves Pitarakis for his: suggesting of ideas, teaching me where to focus my efforts and encouragement with my work as well as many other methods of support. Throughout my masters and all chapters of this thesis his support has improved my work as an academic to new heights that I never thought were possible and the completion of this thesis would not have been possible without him. I would also like to thank other members of staff who provided valuable feedback at my progression reviews allowing for me to develop further. These people include: Professor Jose Olmo, Professor Anastasios Magdalinos, Professor Carmine Ornaghi and Dr. Maria Kyriacou whose contributions and ideas have helped me produce this thesis. I also mention all my friends and fellow PhD students in the office who have made my studentship such an incredible experience through a feeling of fellowship and have given me the motivation and dedication required to carry out all the important tasks of my research. I also, state my gratitude for the financial and academic support of the South Coast DTP who made it possible for me to begin on the road to this accomplishment due to the maintenance funding. Finally, I thank my family for their never ending moral support throughout my student life, they have always been there to support me emotionally during challenging times and have given me the willpower to keep giving my best which I am forever grateful for.

This is for my mum and dad, thank you for always being so supportive throughout my education and allowing me the opportunity to explore my passions.

Chapter 1

Introduction

Due to the advances in technology concerned with methods of data collection in the last 2 decades, the topic of big data analysis has become increasingly significant across all disciplines that rely on quantitative analysis for the formation of their conclusions. In addition, the growing demand for more information from both the public and private sectors in order to improve their services has lead to an ever-increasing availability of data that one can access. To provide a more clear characterisation of "big data", consider the three-Vs used commonly throughout the literature (Ng (2016), Laney (2001)), defined as: Volume, Variety and Velocity. Volume refers to the size of data sets which has increased due to advances in digital storage allowing larger amounts of data to be contained in a given place. Variety is concerned with the many different sources of data and collection methods now available leading to greater availability. It is also associated with the many different formats that data can come in (Dash et al. (2019)), for example, a large retailer may have transaction level data for in-store purchases as well click data for its online platform or possibly even audio data from interviews as part of its market research. Finally, Velocity refers to the speed at which data can be collected which, through computational advances, has also increased on a huge scale. For example, Google data on the searches and clicks of users is constantly being streamed into a database. Another example is financial markets where, co-called, tick-data is constantly updated as the prices of financial stocks and derivatives evolve.

Examples of such applications where one is faced with big data are in abundance. For example, in health care, providers use Electronic Health Records (EHRs) containing a vast range of characteristics for all patients including demographics, medical history and test results. In addition to these records containing a rich variety of information for each patient, the sheer number of patients make them huge, for example, the NHS has records for over 65 million patients (Kollewe (2019)). In biomedical research, it is estimated that a human gene compares 3.2 billion base pairs of DNA and 80% of rare

diseases are believed to be genetic (Peplow (2016)). This has lead to huge data collection projects such as the 100,000 Genomes Project and the Human Genome Project. The latter collects sequencing data and links this to diseases in order to improve future diagnosis procedures. This has also lead to the assembling of a large number of variables related to individuals, such as symptoms and biochemical test results (Kalina (2018)). In the field of Psychology research, the World Values Survey contains data from over 340,000 participants with 1377 variables leading to many researchers not using the data set in its entirety (Cheung and Jak (2016)). From a more technical perspective, online activity monitoring is also a key source of data in Psychology, for example, in Klinkenberg et al. (2011) the response of 3648 children to 3.5 million computer arithmetic tasks were monitored leading to a huge data set to draw conclusions from. In economics there are a wide variety of applications involving huge data sets that have arisen as well. For example, the Nielsen data collects retailer weekly retail scanner data from over 35,000 grocery stores across the US with over 3 million UPCs (Ng (2016)). All data is numeric and is mostly associated with the price and quantity of products sold making itself of particular interest to both micro and macro-economists. In finance based settings the use of text data employing software to analyse social media posts, news articles and company reports have been used extensively to improve predictions of stock market variables. For example, Chung (2014) develop a system called BizPro to compute business intelligence factors based on thousands of sentences from news sources with such tools becoming increasingly common as an alternative to standard historical data sets produced for analysts.

While each discipline faces a variety of issues related to big data sets, such as lack of digital storage space or security issues, this thesis focuses on one specific class of problems. This is where the data set is large to the extent that it is impossible to apply the standard statistical tools used throughout economics in order to carry out the desired analysis. In particular, one is faced with cross sectional or longitudinal data in the form of a design matrix with n cross-sectional or time series observations and p different features or individual series. Shapiro (2017) condenses this into 3 main issues characterised as follows:

- There are a huge number of observation (large n) such that the design matrix is too large to be loaded by software packages.
- The number of variables is larger than the observations ($p > n$), preventing many standard statistical models from being feasible.
- The number of covariates increases as the number of observations does.

While the majority of literature throughout Machine Learning is devoted to the first issue, this thesis focuses on the second issue and is characterised by the term "High Dimensionality". In economics and finance, this is particularly an issue of big interest due to how pivotal the Ordinary-Least-Squares (OLS) framework is and how it becomes infeasible since the design covariance matrix is no longer invertible. Even when $p < n$, if p is still close to n then the parameter estimates will have a considerable degree of uncertainty, rendering them unfit for purpose.

There are a vast number of areas throughout the literature where this is the case as detailed in Fan et al. (2011). Firstly, they mention how the widely used vector autoregressive (VAR) model of Sims (1980) results in the the number of parameters increasingly significantly with the number of covariate time series used and leads to high dimensionality becoming increasingly common. This is especially the case in recent times as mentioned earlier where more series are collected and available to the econometrician, for example, Yousuf and Ng (2021) use the FRED-MD database and are faced with 128 macroeconomic time series. Another reason for this large pool of covariates comes from how there are multiple indices available to measure similar phenomena, for example, for prices there are dozens of price indices that differ on the way they are computed despite measuring more or less the same underlying feature of the economy. Another source of scenarios where high dimensionality is common is the nature of how many standard dependent variables such as GDP are usually sampled at quarterly frequency, whereas many of the covariates will be sampled monthly or even more frequently (especially financial data which can be sampled daily or even more regularly). This may lead to one aggregating the higher frequency down to the lower frequency dependent variable, however, this leads to a loss of information. Therefore, Foroni et al. (2015) proposed creating separate predictors for each lag of a given variable. For example, if using a monthly inflation indicator to forecast quarterly GDP then one could have 3 separate covariates for each inflation observation corresponding to the 1 GDP observation for the quarter. Consequently, one can see with this structure that it would only take a small number of high frequency covariates in order to create a high dimensional setting. Shapiro (2017) also mentions how a high dimensional set of instrumental variables is a common occurrence that causes issues. For example, Belloni et al. (2012) investigate how eminent domain policies determine housing market features and use characteristics of the judges allocated to appeals cases as instruments for specific eminent domain policies. This involves a large number of variables associated with the personal characteristics of the judges such as gender, race and education. As a result, a dimension reduction technique is required here in order to allow two-stage least squares to be feasible.

As one can see, there are a whole host of settings throughout economics and finance

where high dimensionality is likely to occur. This thesis investigates the models used by economists as alternatives to OLS for parameter estimation in linear regression when the number of predictors is large. Through comparing estimation and predictive performance of the widely used methods along with the proposal of new estimation procedures, the contribution of this thesis is somewhat broad with multiple aspects of high dimensionality in economics covered. The following sections of this chapter summarise the contributions and approaches taken by the 3 main chapters of this thesis. The second chapter acts as a general comparison of the methods used with respect to predictive performance in settings with economic data. The third, proposes a new estimation procedure based on the widely used Ridge Regression of Hoerl and Kennard (1970) with improvements to side-step bias issues. The fourth chapter focuses on a method called Random Projections from the machine learning literature. This is very new to economics, and even more niche with regards to individual parameter estimation, making the contribution somewhat ambitious. A parameter estimation procedure is proposed to act as an attractive option for when one is interested in a treatment variable coefficient with a large pool of potential control variables.

1.1 An Overview of Modern Methods for Forecasting Economic Times Series in High Dimensional Settings

This essay begins by summarising the categories of approaches used for high dimensional linear regression, particularly in economics where often the data possess certain characteristics worth bearing in mind. One of these features is covariate multicollinearity (Farrar and Glauber (1967)) whereby, for various reasons, multiple predictors in the design matrix share a linear relationship leading to potential issues with identifying which predictors are truly driving changes in the dependent variable. Such a feature is even more common in high dimensional setting due to how the sheer number of covariates makes it increasingly likely that variables will share common trends by chance. Another frequently occurring characteristic, especially in financial time series data, is the issue of time series persistence. For example, measures of financial market volatility (Hansen and Lunde (2014)) and even US GDP has been argued to behave as a non-stationary process (Campbell and Mankiw (1987)). Aside from the issue of high dimensionality, these 2 features pose serious challenges for OLS. Multicollinearity leads to a large degree of uncertainty in parameter estimates while non-stationarity prevents standard asymptotic properties from no longer holding.

Regarding the most common methods used to handle regressions with a high dimensional set of covariates, a handful of broad categories can be used to summarise. Firstly, penalised least squares regressions overcome the invertibility issue of the

covariance matrix while providing parameter shrinkage and possibly a model selection element in certain cases. Examples of this include: Ridge Regression (Hoerl and Kennard (1970)), Least Absolute Shrinkage and Selection Operator (Tibshirani (1996)), Smoothly Clipped Absolute Deviation (Fan and Li (2001)) and Minimax Concave Penalty (Zhang (2010)). While all methods are based upon the minimisation of a sum of squares function with a penalty term, the way in which parameter estimates are penalised varies resulting in differing estimator properties with respect to bias, standard errors and variable selection consistency. Such approaches are computationally simple but even these methods struggle as the degree of high dimensionality increases too much ($p >> n$).

The other main class of models used throughout economics is factor models. These seek to incorporate all relevant variables in a small number of observable driving forces known as factors. Such a technique sits well with economic theory with examples of factors being present in a wide number of applications, such as the Arbitrage Pricing Theory of Ross (1976) and the Capital Asset Pricing Model of Sharpe (1964). While previously, one could apply the Kalman filter to obtain these factors (Ghysles and Marcellinio (2018, p. 504)), when the number of covariates available is large then this becomes computationally burdensome. This lead to the work of Stock and Watson (2002a, 2002b) proposing the use of Diffusion Indices computed by Principal Components Analysis (PCA). Such an approach is non-parametric based on the eigenvectors of the predictor covariance matrix allowing significant information to be maintained in a small number of factors. Work such as Forni and Lippi (1997) and Forni and Reichlin (1998) as well as others show that, for many macroeconomic variable forecasts, 2 or 3 factors provide optimal accuracy. While other methods for factor estimation have been considered and compared throughout the literature (Boivin and Ng (2005)), this chapter focuses on Principal Components as the most central methodology for macroeconomic forecasting.

While one can see that there are a wide variety of procedures one can use when faced with high dimensional linear regression problems, it may seem challenging to determine which approach is most suitable for the application at hand. As mentioned earlier, economic data can often be associated with significant multicollinearity or persistence amongst time series and, as a result, this should be accounted for when choosing an econometric approach. This chapter seeks to use detailed Monte Carlo Simulation experiments that replicate forecasting settings incorporating these two data features. By finely tuning the level of predictor correlation, temporal dependence as well as sparsity and signal-to-noise, the forecasting performance of numerous methods detailed above are compared. For that reason, this chapter aims to contribute by providing a practical overview of the forecasting performance of several popular

predictive approaches. The aim being to characterise which model should be used under various DGP conditions in order to provide guidance for predictive accuracy maximisation to econometricians.

1.2 A Ridge Regression Modification for Handling High Dimensional Economic Data

The third chapter focuses specifically on one of the methods used in the experiments of the second chapter. This method is the penalised least squares Ridge Regression of Hoerl and Kennard (1970). This approach not only is feasible when $p > n$ but is shown to always provide a lower parameter estimate variance than OLS. Although this comes at the expense of inducing some bias, Theobald (1974) showed that, under certain conditions of the penalty parameter, Ridge will provide a lower mean-squared-error value for the parameter estimates.

One property that makes Ridge appealing to economists is how the penalty term allows Ridge to handle correlated predictors successfully by shrinking their coefficient estimators towards each other as demonstrated in Kidwell and Brown (1982), Gunst et al. (1974) and Mason and Brown (1975). While other work shows that, when concerned with prediction, Ridge performs relatively well (Frank and Friedman (1993) and Dhillon et al. (2013)) and can even be used successfully for variable selection (Shao and Deng (2012)).

However, these promising attributes do not come without drawbacks. Unlike OLS, Ridge has non-zero bias when the penalty parameter is non-zero. Moreover, this along with its variance being dependant on the true value of the coefficients (Smith and Campbell (1980)) makes it challenging for one to carry out inference with empirical data. In addition, work such as O'Neill and Buttimer (1972) demonstrate how when there is a significant mix of signs in the true coefficients or covariate correlations, the bias of Ridge results in it providing unreliable estimates characterising it as a poor choice of model for settings such as demand function estimation.

This chapter proposes a new approach that seeks to maintain the benefits of l_2 norm penalisation while overcoming the bias issue that Ridge faces under certain DGPs. This is attempted by considering the Frisch-Waugh Lovell Theorem of Frisch and Waugh (1933) as a means of obtaining single OLS estimates by partialling out all other covariates. Applying l_2 norm penalisation to this creates a situation where one is

estimating a single coefficient by penalising all covariates apart from the one of interest. Not only can this act as a more focused approach when one is only interested in the estimates for a small number of coefficients, but also has implications for the bias and variance. The aim being to provide a more favourable tradeoff to produce a lower MSE than that of Ridge alone.

This approach is called Partial Ridge and this chapter investigates how the statistical properties of this approach vary in comparison to Ridge alone. Through theoretical analysis, Monte Carlo Simulation experiments and an empirical application of house price prediction, this chapter discusses when Partial Ridge can be used to overcome the problems that Ridge faces and combining the two methods to define a competitive approach for predicting economic variables with high dimensional data.

1.3 Partial Random Projection, A Novel Approach to High Dimensional Linear Regression in Economics

The fourth chapter turns attention to the machine learning literature which has had growing influence over multiple disciplines affected by big data in the last decade. Specifically, a method called Random Projections (RP) is considered for high dimensional regression problems. RP works in similar fashion to that of the widely used Principal Components Analysis whereby an auxiliary covariate matrix is created by taking p linear combinations of each row, where p is the number of predictors. While Principal Components computes these linear combinations based on the covariance matrix eigenvalues, RP constructs these based on weightings simulated from a symmetrically distributed random variable. One may question how such an approach can preserve enough of the information contained in the original data matrix in order to be useful, however, the Johsnon-Lindenstrauss Lemma of Johnson and Lindenstrauss (1984) demonstrates how the pairwise distances between points can be maintained from the original to auxiliary data matrix with high probability.

This result concerning the pairwise distances between points has made RP extremely useful in a wide range of applications that require elements of dimension reduction on large data sets. Historically, the most common examples would involve classification in areas such as human behaviour monitoring and genetics (Nabil (2017)). However, more recent uses of the RP transformation on data matrices has involved the initial processing of text and image data (Bingham and Mannila (2001)) due to the development of modern algorithms for formalising word frequency in documents and pixel brightness in image data. Often RP is preferred over other dimension reduction techniques due to its relative computational simplicity but also because the

randomised nature of the transformed data set makes it perfect for handling sensitive data from a security perspective.

Work applying RP to high dimensional linear regression is less present but predictive risk bounds have have been studied in Boot and Nibbering (2019), Kaban (2014) and Thanei et al. (2017). While these papers show promising signs for RP in forecasting and prediction settings, there is no literature yet that focus on the use of RP dimension reduction when one is focused on the specific coefficients of a small number of variables. The field of economics leans significantly upon causal inference for a wide variety of applications where one is concentrated on the estimate of the treatment variable coefficient with control variables included to improve the accuracy of the treatment effect estimation. Therefore, this chapter specialises on using the RP mechanism to construct an approach ideal for estimating a single coefficient (the treatment effect) with a high dimensional set of control variables.

This essay seeks to propose a new approach motivated by the methodology discussed in Galbraith and Zinde-Walsh (2020). Here they estimate the treatment effect by computing Principal Components from only the set of control variables and running OLS with the treatment variable and a predetermined number of Principal Components on the right-hand side. Accordingly, this essay considers the case where one estimates the treatment effect by applying RP to the controls and is defined as the Partial Random Projections (PRP) approach. While RP alone can return a set of coefficient estimates of the same scale and dimension as the original data set, this requires the estimated vector to be recompressed following the compression of the original covariate matrix imposing bias. Therefore, by removing the treatment variable from the RP transformation, one can avoid complications leading to the parameter estimation containing too much error.

This chapter formally defines the PRP procedure and analyses the theoretical aspects of it by comparing its bias and variance behaviour to that of other widely used methods. All of this will be considered with respect to fundamentals of the true DGP such as the sign and magnitude of the true coefficients, the correlation amongst predictors and signal-to-noise ratio. Finally, a Monte Carlo Simulation study will be constructed to replicate a broad range of environments that economists will likely face with estimation accuracy compared to other methods used for high dimensional linear regression analysis.

Chapter 2

An Overview of Modern Methods for Forecasting Economic Time Series in High Dimensional Settings

2.1 Introduction

In recent years, the improved capacity of data storage on modern computers combined with the increased amount of data collected by institutions for research-based purposes has led to so called "big data sets" becoming ever more common. Whilst this has provided greater opportunities for analysts to gain deeper insight into their relevant disciplines, it has also provided challenges in the form of existing statistical models failing to work as effectively when faced with huge volumes of data. In linear regression settings, it is well known for instance that Ordinary Least Squares (OLS) has an excess risk that deteriorates rapidly as the number of predictors, p, approaches the sample size, T, and in other situations, such as when $p > T$, OLS is unfeasible. To overcome these difficulties, the statistics and machine learning literature have proposed regularization techniques designed to accommodate scenarios where $p \sim T$ or even $p > T$ through parameter penalisation approaches. Popular regularisation methods include Ridge Regression of Hoerl and Kennard (1970) and the Least Absolute Shrinkage and Selection Operator (Lasso) of Tibshirani (1996) as well as numerous variants. Under suitable assumptions these techniques have been shown tot provide favourable bias-variance trade-offs while remaining feasible in high-dimensional settings. Other approaches exploit the fact that many of the predictors available come from a small handful of unobservable underlying driving forces leading to the rise of Principal Components and Factor Analysis in work such as Stock and Watson (2002a, 2011). The same motivation also lead to the recent development of an approach known as Random Projections (Kaban

(2014), Thanei et al. (2017)) whereby the original predictor set is compressed to a smaller one via random linear combinations.

Many of the theoretical arguments supporting the use of these methods are based on strong assumptions such as the absence of any strong temporal dependence in the predictors under consideration with independence and identical distributions for each observation being a common theme. Also, a large proportion of the current work only allows for a limited degree of correlation amongst the covariates with results being unreliable outside of this scenario. Such restrictions are unlikely to hold in economics and financial data sets where it is well known that a significant number of financial time series are highly persistent unit-root or near unit-root processes. Similarly, in addition to the economic reasons that induce correlations between time series, it is well shown that persistence induces correlation between multiple time series, also known as spurious correlation as shown in work such as Ferson et al. (2003) and Kruse et al. (2017). This raises many questions on how these attributes of economic and financial data sets will influence the functionality and statistical properties of regularisation and dimensionality reduction methods. This paper looks to investigate how reliable these methods are when used for computing economic predictions in these high dimensional environments.

Another recent source of need for models that can handle a large number of covariates relative to the sample size is how the increased availability of variables has resulted in a greater diversity of sample frequencies amongst time series regressors. For example, when considering economic indicators the the frequency in which the observations are sampled at (monthly, quarterly or annually) may vary significantly throughout the predictor set making it difficult for one to format the data in order to be suitable for linear regression. One could consider averaging in such a way that condenses the high frequency data down to that of the lowest frequency but this discards the information contained through the timing of such observations. An early class of models used to handle this is the Mixed Data Sampling (MIDAS) class of models of Ghysels et al. (2004, 2007), however, this often results in more overparameterised specifications. For example, if one was using monthly sampled inflation to forecast quarterly GDP then there would be 2 additional variables created as there are 3 lagged values of inflation corresponding to every single GDP observation. As one can see, this has the potential to drastically increase the number of predictors in addition to the already large predictor set for reasons already discussed.

This paper seeks to remedy the uncertainty surrounding how these popular methods perform relative to one another in an economic forecasting setting that blends high dimensionality and persistence or strong correlatedness through a Monte Carlo

simulation study. A small handful of methods is chosen but is still representative of the broad categories of approaches used to handle a large set of predictors in linear regression settings such as: Model Selection, Model Averaging, Shrinkage, Regularisation and Factor Models. The rest of the paper is organised as follows; Section 2.2 details the methods employed for comparison with intuitive justification for their inclusion. Section 2.3 describes the simulation experiment design and how the data generating processes (DGPs) are finely tuned to be fit for purpose with Section 2.4 presenting the results and interpretation. Section 2.5 concludes with consideration given to the implications of the results and how one may wish to proceed with further research in light of this study.

2.2 Models and Estimation Methods

This section begins by detailing the regularisation and dimensionality reduction methods to be compared in this study with respect given to the intuition behind their use as well as the main theoretical justification of their benefits. The variety of methods here is by no means exhaustive of the approaches used throughout the economics, statistics and machine learning literature for forecasting with high dimensional data sets. However, the list does include the most widely used models with a diverse representation of the classes of approaches mentioned in Section 2.1. It is important to first establish the key notation whereby the predictor matrix is denoted by X and takes the form of a $T \times p$ matrix with each individual column representing an individual covariate time series. The dependent variable is denoted by a $T \times 1$ vector, Y, which is what one wishes to forecast by using a linear regression framework as shown in 2.1 with the case of a single observation given as follows:

$$Y = X\beta + \epsilon\,, \tag{2.1}$$

$$y_t = x_t'\beta + \epsilon_t\,,$$

where the time index runs from $t = 0, \ldots, T$ and x_t is a $p \times 1$ vector containing the observation at time t for each of the predictors. As is the case with all linear regression frameworks, one is faced with the task of estimating the $p \times 1$ vector of slopes β which rely on certain assumptions of the error term, ϵ. As the focus of this paper is on isolating the influence of persistence and strong correlatedness on high dimensional methods, the potential for complications to arise from phenomena such as serial correlation or heteroskedasticity in the ϵ_t components is removed. Therefore, the DGPs are assumed to behave such that $E[\epsilon_t|x_t] = 0$ and $E[\epsilon_t^2|x_t] = \sigma_\epsilon^2$ for all $t = 0, \ldots, T$.

Standard out-of-sample forecasting will involve the analyst estimating the $p \times 1$ vector β using the in-sample data available. Here it is assumed that the in-sample period is defined from $t = 0, \ldots, T-1$ with the observation y_T being predicted using the equation below. It is important to note that the the slope estimates, $\hat{\beta}$ are computed using only the observations of y_t for $t = 0, \ldots, T-1$ and x_t for $t = 0, \ldots, T$ since it is assumed that the analyst observes the covariates at time T in order to compute a forecast of y_t with the following:

$$\hat{y}_{T|T-1} = x_T' \hat{\beta} \,. \tag{2.2}$$

From here one can assess the accuracy of the various approaches used to compute $\hat{\beta}$ by considering the forecasting errors defined as $\hat{e}_{T|T-1} = y_T - \hat{y}_{T|T-1}$ with the smaller values corresponding to the more reliable method.

Typically, one would wish to use OLS to estimate the best linear unbiased estimator, $\hat{\beta}_{OLS}$, of β. However, when T<p the key OLS assumption of the predictor matrix being of full rank is violated resulting in it not being possible to obtain an estimate for β in this way. Even when T>p, when p is relatively large then there is a greater degree of predictive risk and seen in 2.3 which shows hows how the predictive risk of OLS estimates depends on p,

$$\frac{1}{n} E[\|X\beta - X\hat{\beta}_{OLS}\|_2^2] = \frac{\sigma^2 p}{n} \,. \tag{2.3}$$

As a result, new methods have been created to resolve these issues and the specific ones used in this paper are detailed in the following subsections. Multicollinearity amongst the regressors can reduce the accuracy of 2.2 through a variety of ways; firstly, Taboga (2021) discusses how computing the inverse of $X'X$ becomes more error prone resulting in inaccurate coefficients being used. Secondly, it can be challenging for the model to identify the specific impact of each individual predictor when the covariates share common trend features. Finally, the standard errors of the coefficients increase under high correlation causing forecasts to become more uncertain and, hence, are less reliable to the forecaster.

Lastly, the performance of OLS forecasts suffer greatly when the regressors exhibit strong temporal dependence as the data is associated with unit roots and non-stationarity. This leads to the parameters no longer converging to their true values asymptotically and spurious regression occurs with more discussion seen in Granger and Newbold (1974). As a result of the coefficients being estimated incorrectly, the forecasts in turn are untrustworthy meaning that other means are required for forecasting purposes.

2.2.1 Ridge Regression

The Ridge Regression approach of Hoerl and Kennard (1970) has been very popular in chemical engineering and takes a form similar to that of OLS whereby the coefficients are computed by minimising the following form of penalised-least-squares

$$\hat{\beta}_{Ridge} = argmin_{\beta}\{(Y - X\beta)'(Y - X\beta) + \lambda \sum_{i=1}^{p} \beta_i^2\}\,, \tag{2.4}$$

where it can be seen that this has a similar set up to the OLS estimator with the addition of a penalty term being the sum of the squared coefficients multiplied by a penalty parameter, λ. This allows one to decide the level of trade-off between in-sample fit and parsimony as required, although for forecasting this would typically be chosen with the aim of minimising the mean-squared-forecasting-error (MSFE) or some other similar loss function. The minimisation problem in 2.4 has a closed form solution making it easy to analyse the theoretical properties of the Ridge estimator and is detailed below for a fixed design setting,

$$\hat{\beta}_{Ridge} = (X'X + \lambda I)^{-1}X'Y\,, \tag{2.5}$$

$$Bias(\hat{\beta}_{Ridge}) = -\lambda(X'X + \lambda I)^{-1}\beta\,, \tag{2.6}$$

$$Var(\hat{\beta}_{Ridge}) = \sigma_{\epsilon}^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1}\,. \tag{2.7}$$

These features of Ridge are demonstrated in Taboga (2017) where there are also proofs that the variance in 2.7 is always smaller than that of OLS for when $\lambda > 0$. Also, that the bias is non-zero when $\lambda > 0$. Therefore, it is clear that the Ridge regression approach seeks to overcome the issue of a large variance by sacrificing a small amount of bias to reduce the uncertainty of the parameter estimates, otherwise known as managing the bias-variance trade-off. The specific condition on the penalty parameter to ensure that Ridge provides a lower mean-squared-error (MSE) than that of OLS is derived in Theobald (1974) and is as follows,

$$\lambda < \frac{2\sigma_{\epsilon}^2}{\beta'\beta}\,. \tag{2.8}$$

One final point worth making, as it is most relevant to this paper and applications in economics, is that one can see how the addition of the λI term ensures that the term to be inverted is positive definite and hence invertible. Essentially, this results in correlated predictors having their coefficient magnitudes shrunk towards each other with the consequences of this to be seen in the simulation study of this paper. One would expect that this feature would work in favour of Ridge when faced with highly correlated predictors as the penalty parameter can adjusted to overcome the issues of inversion. However, when faced with temporally dependent predictors it may not

experience as much success. This is because Ridge shares some similarities with OLS in that requires stationarity in the predictors in order for the Central Limit Theorem to hold when concerning the convergence of the $\hat{\beta}$ values. Without this consistency, the coefficients can be deemed as unreliable making forecasts significantly inaccurate.

2.2.2 Least Absolute Shrinkage Selection Operator (Lasso)

One noticeable drawback of Ridge Regression is that, despite its ability to shrink parameter estimates, it will never set any of the coefficients exactly to zero. this is undesirable in sparse settings when one is faced with a large number of predictors but can be fairly certain that only a few are active, and the others have no relationship with the dependent variable at all. To rectify this, Tibshirani (1996) proposed the Lasso model which still has a penalised-least-squares set up, only the l_1 norm of β is used as a penalty as seen in 2.9,

$$\hat{\beta}_{lasso} = argmin_{\beta}\{(Y - X\beta)'(Y - X\beta) + \lambda \sum_{i=1}^{p} |\beta_i|\}\,. \tag{2.9}$$

Unlike with Ridge, the Lasso minimisation problem cannot be solved analytically, however, various algorithms are available to obtain the coefficient estimates such as the LARS algorithm of Efron et al. (2004) have been shown to be effective. The fact that the Lasso sets certain coefficients directly to zero means that it enjoys lower coefficient variances compared to Ridge, moreover, Knight and Fu (2000) show that, under certain conditions on the penalty parameter, the Lasso is consistent in estimation of β. However, unlike Ridge, when the Lasso is faced with highly correlated predictors, it selects one and discards the rest (setting their coefficient directly to zero) due the penalty term being convex, but not strictly convex, with respect to β. It is because of this that the Lasso is not always consistent in selecting the right predictors with respect to the true DGP as outlined in Zhao and Yu (2006), where they derive the Irrepresentable Condition to determine when Lasso will select the true model consistently. This condition is defined below:

$$|C_{21}^n (C_{11}^n)^{-1} sign(\beta_{(1)}^n)| < 1\,, \tag{2.10}$$

where $\beta_{(1)}^n$ represents the coefficients of the q active predictors according to the true DGP. Regarding the C^n terms, it is important to first define that $X_n(1)$ represents the q columns of the predictor matrix that correspond to the predictors significant in influencing the dependent variable while $X_n(2)$ represents the $p - q$ columns of the predictor matrix that correspond to the inactive regressors. The C^n terms refer to the

variance and covariance matrices involving $X_n(1)$ and $X_n(2)$ and are as follows:

$$\begin{aligned} C_{11}^n &= \tfrac{1}{T} X_n(1)' X_n(1) \\ C_{12}^n &= \tfrac{1}{T} X_n(1)' X_n(2) \\ C_{21}^n &= \tfrac{1}{T} X_n(2)' X_n(1) \\ C_{22}^n &= \tfrac{1}{T} X_n(2)' X_n(2)\,. \end{aligned} \tag{2.11}$$

So one can see that by nature, higher covariances between active and inactive the the predictors results in a higher C_{21} and the left-hand side of 2.10 being larger and, therefore, the likelihood of the Irrepresentable condition holding being much less and model selection being inconsistent. This means that, generally speaking, while the Lasso will likely outperform Ridge in sparse environments, one would expect Ridge to improve as contemporaneous correlation amongst the predictors is more common with this study aiming to learn about how these features affect their relative forecasting performance in applications faced within economics. When faced with persistent predictors, the Lasso relies on similar asymptotic features as Ridge making the shrinkage element of Lasso suffer as regressors approach the unit root mark. In addition, the trend nature of non-stationary data causes spurious correlation amongst the predictors which, if strong enough, reaches the same problem discussed earlier whereby the Irrepresentable Condition no longer holds and the wrong variables are kept for forecasting.

2.2.3 Adaptive Lasso

As mentioned previously, the Lasso is inconsistent in selection when too much correlation is present leading to Zou (2006) to propose the Adaptive Lasso approach. The minimisation problem takes the following form:

$$\hat{\beta}_{ADlasso} = argmin_{\beta}\{(Y - X\beta)'(Y - X\beta) + \lambda \sum_{i=1}^{p} w_i |\beta_i|\}\,, \tag{2.12}$$

$$where\, w_i = \frac{1}{|\hat{\beta}_i|^{\gamma}} for\, \gamma > 0\,.$$

One can see that the weighing scheme (w_i for $i = 1, \ldots, p$) regulates the level of shrinkage applied to each of the coefficients. The $\hat{\beta}$ part of the weight will typically be the estimate obtained from running OLS on the full data set, however, Zou (2006) states that it can be any consistent estimator of β. The reasoning for this is that coefficients with a larger magnitude will receive a smaller weight resulting in less penalisation. This contributes to the Adaptive Lasso being consistent in selection as this means that noisy irrelevant variables will face a relatively larger penalty and are more likely to be correctly discarded.

2.2.4 Elastic Net

The Elastic Net Regression first proposed by Zou and Hastie (2005) seeks to combine the merits of Ridge and Lasso, covering more scenarios that those 2 methods suffer in. The following expression defines the Elastic Net estimator:

$$\hat{\beta}_{EN} = argmin_{\beta}\{(Y - X\beta)'(Y - X\beta) + \lambda\Big((1-\alpha)\sum_{i=1}^{p}\beta_i^2 + \alpha\sum_{i=1}^{p}|\beta_i|\Big)\}, \tag{2.13}$$

where α controls how much of the penalty is the of the l_1 norm and how much is penalised by the l_2 norm, where $0 < \alpha < 1$. One can see that as α increases, the penalty term works increasingly like the Lasso introducing a greater element of variable selection over shrinkage. This combination of the Ridge and Lasso penalties results in the Elastic Net enjoying some useful benefits, firstly the Ridge element contributes to the penalty term being strictly convex. This causes correlated predictors to be shrunk collectively as opposed to discarding most of them. While this may or may not positively influence predictive accuracy, studies that rely on parameter interpretability can benefit from this, for example in Hastie et al. (2000) a study is carried on DNA data searching for correlated genes. Another advantage that the Elastic Net boasts over the Lasso is its ability to maintain its predictive accuracy performance when p>T, where the Lasso saturates as soon as T predictors are chosen as noted in Rosset et al. (2004).

2.2.5 OLS Post-Selection

Much of the early literature surrounding linear regressions with a large number of variables involved selecting a subset of predictors based on minimising some criteria, usually involving the sum of squared errors. The most famous work regarding this includes the Mallows Information Criteria of Mallows (1973), Akaike Information Criteria (AIC) of Akaike (1969) and Bayesian Information Criteria (BIC) of Schwarz (1978) which fit all possible sub models with the predictors available and grade each model based on the value of the proposed criteria which is composed of a combination of the sum of squared errors, for prediction accuracy purposes, and the number of regressors included for the interest of parsimony and variance reduction. However, model selection can fall short in certain areas, especially when the number of candidates available is so large as one is faced with the task of estimating 2^p models. For small p, this is feasible, for example when there are 10 candidate predictors there are 1024 models to compute the chosen criterion for but as p increases it becomes very apparent that it is computationally infeasible to estimate all the models and store the criterion value from each one for comparison later. Therefore, in recent years, the approach of using soft-thresholding methods, such as the Lasso, for selection, discarding the variables set to zero then running OLS has gained significant popularity. Intuitively, provided that the Lasso is consistent in model selection, this

results in the oracle model being estimated as OLS is run on only the variables that are active according the true DGP as discussed in Belloni and Chernozhukov (2013). They also show that this approach has a faster rate of convergence to the true β values than Lasso alone and still experiences lower bias when the Lasso fails to include relevant predictors.

Other studies that favour this approach include Shi et al. (2020) and Böheim and Stöllinger (2020) finding significant improvements in identifying treatment effects in the gender wage gap in the latter. As mentioned before, when there are many correlated variables, the Lasso generally picks one and discards the rest, therefore, it is not obvious how only using the selection element of Lasso and Elastic Net followed by OLS will perform although one would imagine that the Irrepresentable Condition holding would be crucucial to the success of such approaches. In this paper, OLS post selection is considered for the standard Lasso, Adaptive Lasso and Elastic Net as these have the ability to set coefficients directly to zero. Each of these 3 methods is run followed by the discarding of the variables that are set to 0. OLS is then used to generate a forecast in a similar to fashion to that discussed previously.

2.2.6 Principal Components Regression

Factor models are very common throughout a vast range of applications in economics and assume that, despite there being many different variables available, most of them will originate from a smaller set of unobserved driving forces, defined by theoretical arguments. These can then be used to forecast the variable concerned using OLS in the way discussed before. In essence, this attempts to use all the relevant information from the data set available but with a smaller set of predictors making OLS a more plausible approach. This is formally defined by the following system of equations:

$$y_{t+1} = F_t'\Gamma + \epsilon_{t+1}\,, \tag{2.14}$$

$$X_{i:t} = F_t'\gamma_i + \mu_{i:t}\,. \tag{2.15}$$

In the above expressions, F_t represents an r×1 vector of latent factor values (where r$\ll$p) at time t and Γ the coefficient matrix which is to be estimated by the forecaster in order to create out-of-sample forecasts. The expression 2.15 shows how all the variables observed in $i = 1, \ldots, p$ can be written as a linear combination of the latent factors with ϵ_{t+1} and $\mu_{i:t}$ representing idiosyncratic error terms. The next question from here is centred around how to estimate the factors and what are the theoretical and empirical success implications of these methods.

Principal Components (PC) has been the most widely used method throughout economics to construct the factors based on the co-movements and variation amongst predictors. More specifically the eigenvectors corresponding to the r largest eigenvalues of the covariance matrix of the regressors is used where r represents the number of factors to be used in regression 2.14 above. Connor and Korajczky (1988,1993) were one of the first to use PC to estimate approximate factor models in asset return prediction. However, prior to this, other more theory and intuition-based justifications were used to estimate factors depending on the application concerned, for example, work such as Lehmann and Modest (1988) consider factor estimation for the Arbitrage Pricing Theory Model of Ross (1976). Also, the famous Capital Asset Pricing Model (CAPM) of Sharpe (1964) as well as others use a similar idea where the latent market return is estimated as factor that influences all asset returns with financial theory to justify this. PC can be viewed as a way to estimate factors in the absence of sufficient prior knowledge and is successful in reducing the regressor dimension without directly discarding any information by a means that is computationally feasible. Stock and Watson (2002a) show that PCs are consistent estimators of latent factors and confirm that this holds under the case where there is variation in the coefficients of the underlying DGP over time. Stock and Watson (2002b) use Diffusion Indexes constructed by PC to forecast various macroeconomic variables across a range of horizons experiencing great success in a root-mean-squared-error (RMSE) sense relative to standard benchmark models. More noticeably, PC has been used to estimate factors in the so-called "Dynamic Factor Models" (DFMs) proposed originally by Geweke (1977) and Sargent and Sims (1977) which allow for the weighting of variables in the construction of the factors to vary over time making PC an ideal approach.

While Principal Components based methods are intuitive and have some nice properties, especially when p>T, Ng (2013) points out that there are no statistical assumptions made making the way dimensionality reduction is carried out questionable due to there being no supervision. However, one of the main drawbacks of PC and factor augmented approaches is the fact that the factors are constructed in a way that gives no consideration to the variable that is being forecasted. To remedy this Reduced Rank Regressions were proposed by Rao (1964) whereby factors are estimated by minimizing an R squared based statistic. One of the other most popular methods to surpass this is the Partial Least Squares (PLS) estimation strategy of Wold and Lyttkens (1969). However, in this paper, the focus is purely on the forecasting accuracy performance of PC whereby 2.15 is used to estimate the factors, then OLS is run on 2.14 to obtain coefficient estimates. This procedure is then repeated when the out-of-sample data becomes available, and the coefficients estimated previously are incorporated to create an out-of-sample forecast y_{t+1}.

Since PC constructs factors based on co-movements in the regressors, it can be shown that PC, by construction, is successful in preserving much of the signal from the numerous correlated predictors to a small handful of factors. Unfortunately, the same does not occur when faced with temporally dependent regressors. Gonzalo and Pitarakis (2020) show that the spurious correlation present amongst near unit root predictors results in factors to be incorrectly constructed due to the artificial relationship visible between variables. As a result, the factors are ineffective for forecasting making PC an unreliable model in this scenario.

2.2.7 Random Projection Regression

Finally, this paper considers one of the approaches originating in the machine learning literature which, as a field, has become increasingly influential on other disciplines facing high dimensional data sets due to the increasing size of databases available making storage requirement and processing time more pivotal in the decision-making process of modern data analysts. In this paper, the focus is on the Random Projections (RP) regression used in Boot and Nibbering (2019) as a possible means of gaining an advantage over standard forecasting procedures in the presence of correlation or persistence. The RP procedure is detailed as follows, a random matrix, R, of dimension (p×k) is simulated such that k$\ll$p and every element of the matrix is generated from a standard normal distribution as follows:

$$R_{p\times k} : R_{ij} \sim N(0,1)\,. \tag{2.16}$$

The RP regression in then run on the training sample data using OLS as seen below:

$$Y = XR\beta + \epsilon\,. \tag{2.17}$$

Out-of-sample forecasts are then created using the estimated coefficient vector as follows:

$$\hat{y}_{t+1} = x_t R\hat{\beta}\,. \tag{2.18}$$

This is repeated for a pre-determined number of random matrices, call it h, resulting in h different forecasts which are then averaged with equal weighting to produce a single forecast as seen:

$$\hat{y}_{t+1}^{final} = E_h[\hat{y}_{t+1}^{h}]\,. \tag{2.19}$$

A discussion on the choice of h is available in Boot and Nibbering (2019) but in this paper it chosen such that increasing it further would not improve the final forecast significantly. Johnson and Lindenstrauss (1984) show that this method of averaging over forecasts from random weight matrices is very successful in preserving the pairwise distances of the row observations in the predictor matrix which contributes to its forecasting success. Part of the motivation for using this for comparison in this

paper comes from how this process will likely be less influenced by highly correlated predictors and persistence due to coefficient biases cancelling each other out across many projections as well as reduced variance similar to that of the forecast combination approach discussed in Timmermann (2006).

2.3 Experimental Design and Performance Evaluation Implementation

In this simulation study, 2 experiments are carried out in order to test the out-of-sample forecasting performance of Ridge, Lasso, Adaptive Lasso, Elastic Net, Random Projections and Principal Components as well as OLS used after the selection element of a few of these methods has been implemented. These experiments involve varying the levels of correlation between the predictors as well as persistence approaching the unit root mark in experiments 1 and 2 respectively in order to replicate scenarios commonly faced in finance and macroeconomics. The signal-to-noise ratio in these experiments is controlled through the error term variance in the underlying DGP to keep the simulated data in line with what is likely to be faced by real-world forecasters and is often very low as discussed in Hastie et al. (2020).

The sample size is set at $T = 200$ for all experiments whereby the first 199 observations are used as the training sample for parameter estimation. These parameters are then used to forecast the 200th observation of Y using the 200th row of the predictor matrix. As mentioned in the previous section, this implies that the predictors are contemporaneous to the predictand, however, such a feature is less significant when one is focused on the impact of correlation and temporal dependence. It is important to mention that each variable in the predictor matrix is standardised before parameter estimation occurs and also standardised again once the 200th row becomes available in order to compute forecasts with predictor values that centred and scaled with respect to the full sample. To formalise this more thoroughly, the first 199 observations of each column of the predictor matrix are standardised before parameter estimation occurs. Once the parameters are estimated it is then assumed that the 200th observation of each predictor becomes available to the forecaster so now all 200 observations for each predictor are standardised again. Finally, the forecast is computed using the in-sample parameter estimates and the 200th row of the standardised predictor matrix. The forecasting variable y_t is not standardised, however, this is not an issue as one is only interested in the forecasts themselves and not the estimates of β. The error is then obtained by differencing this forecast from the true value in the Y matrix. 1000 simulations in total are run resulting in 1000 forecasting errors which are squared then averaged to obtain the

mean-square-forecasting error (MSFE) which will be the main metric for comparing forecasting performance in this paper. Separate experiments under each design are run each with a different number of candidate predictors, specifically $p \in \{20, 100, 200, 400\}$.

For each experiment 9 different forecasting methods are used, 6 of these include Ridge, Lasso, Adaptive-Lasso, Elastic Net, PC and RP and the other 3 involve running OLS on the predictor set left after discarding variables that have their coefficient set to zero by Lasso, Adaptive-Lasso and Elastic Net. Firstly, for Ridge the method detailed in 2.4 above is applied to the training sample to obtain an estimate of the coefficient matrix, which is then multiplied by the regressor vector in the testing sample to obtain out-of-sample forecasts. For each experiment this procedure is repeated over a grid of penalty parameters as $\lambda \in \{0.5, 1, 2, 3, 5, 10, 20, 50\}$. Lasso, Adaptive-Lasso and Elastic Net are done very similarly using the processes detailed in 2.9, 2.12 and 2.13 respectively over a grid of penalty parameters as $\lambda \in \{0.01, 0.1, 1, 2, 3, 5, 10, 20\}$. Specifically, for the Adaptive Lasso, the coefficients used to construct w_i in 2.12 are estimated by OLS for when $p = 20, 100$ and Ridge when $p = 200, 400$. Another, increasingly popular idea considered here is the so-called OLS-post-Lasso where one runs a lasso regression as a means of variable selection to allow the forecaster to discard the variables that have their slope set to 0 then run OLS on the remaining set of predictors. In this paper, this is done with not only the Lasso but Adaptive-Lasso and Elastic Net as they also have the potential to set coefficients directly to zero.

It is worth mentioning that, due to the number of variables and penalty parameters being considered, if it such that the regularisation method concerned chooses to discard all the candidate variables then a regression is run with a constant on the right-hand-side, representing a case where the out-of-sample forecast, $\hat{y}_{200}$, is simply the mean of y from the in-sample period. For larger values of p, it can sometimes be the case that the regularisation method fails to discard enough variables such that there are more predictors left than number of observations, for example, the in-sample data here has 199 observations so if the Lasso sets 199 or more observations to not be zero (resulting in the predictors being included) then OLS-post-Lasso could not be run for reasons discussed previously. To overcome this, it is such that when this happens then only the first $(n-1)-1$ selected variables are included, starting from the left-hand-side of the predictor matrix. Whilst this seems arbitrary it should not matter significantly given how sparse the true DGP is meaning that this is often the result of the penalty parameter chosen being too low resulting in too few variables being discarded.

For PC, Principal Components Analysis is applied on the training sample to obtain p factors. Various regressions of these factors on Y are run and vary by the number of factors included on the right hand side. The coefficients are estimated and stored to then be applied to the factors calculated from the full-sample regressor matrix (the training and testing sample resulting in 200 observations). Finally, the out-of-sample forecasts are computed by first converting the estimated coefficient vector to one that is appropriate for the original standardised data set. This is done by multiplying the estimated covariates by the eigenvectors corresponding to the factors used as detailed in 10.2 of Severn (2023). This new coefficient vector is then multiplied by the 200th row of the standardised original covariate matrix to give a single forecast. For each experiment, the number of factors included are varied over a grid, r where $r \in \{1, 2, 3, 4, 5, 10, 15, 20\}$ for $p = 20$ and $r \in \{1, 2, 3, 4, 5, 10, 20, 30\}$ for all other values of p.

For RP, out-of-sample forecasts are generated as follows; first, a random weights matrix of dimension $t \times k$ is generated as detailed in equation 2.16 where t represents the number of training sample observations (199 in this case). k represents the subspace dimension such that $k < p$ and is varied across a range of different values. These k grids are as follows: for $p = 20$, $k \in \{2, 3, 4, 5, 10, 15\}$, for $p = 100$, $k \in \{5, 10, 20, 30, 40, 50\}$ and finally, when p is equal to 200 or 400 then $k \in \{5, 10, 20, 30, 50, 100\}$. Coefficients are estimated from the in-sample regression represented in equation 2.17 and are then used to generate a single out-of-sample forecast by using the same projection matrix and coefficient vector on the out-of-sample regressor vector as shown in 2.18. Finally, this procedure is repeated h times with the forecasts being averaged to obtain a single forecast. The choice of h is discussed in Boot and Nibbering (2019) as one needs to have sufficient draws in order to restrain forecast variance and not require too many such that the model becomes computationally burdensome. Figures 2.A1 and 2.A2 in the appendix show how in a small simulation study where 100 artificial data sets are created, with more details outlined in the appendix. The forecasting performance of RP relative to the oracle under the DGP used in design 1 with 400 candidate predictors does not appear to improve convincingly when more than 100 draws are used. Therefore, in this paper, 100 projections are used as forecasts beyond this see a negligible improvement in accuracy.

A final subsection is committed to a small discussion on metrics for performance evaluation. Such a feature is important to establish as this paper focuses on forecasting performance, therefore, the main feature of interest is $X\hat{\beta}$ as opposed causal effect investigations where the researcher is typically only interested in the size, magnitude and significance of $\hat{\beta}$.

2.3.1 Design 1: Correlation

This first class of simulation experiments uses predictors simulated from a multivariate normal distribution with covariance matrix that allows one to vary the level of correlation amongst the independent variables as is done in Elliott et al. (2013). The predictor matrix of the DGP is defined by the following:

$$X_{T\times p} \sim N(0, \Sigma_X)\,, \tag{2.20}$$

$$\Sigma_X = \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \\ \rho & \dots & \dots & \rho & 1 \end{bmatrix}.$$

5 Separate experiments are carried out, each with a different value of ρ used in order to vary the level of correlation amongst the candidate predictors. ρ is varied over the following values $\{0, 0.25, 0.5, 0.8, 0.95\}$ for each given value of p. To generate the dependent variable, the following linear expression is used:

$$Y = X\beta + \epsilon \tag{2.21}$$

and for a single observation this can be written as

$$y_t = x_t'\beta + \epsilon_t\,.$$

The sparsity setting considered for β is one where the coefficient matrix β is created such that only the first 5% of the variables have a coefficient of 1 and the rest 0. The idea of this is to investigate the relative performance of the models concerned in a realistic setting whereby the forecaster has many predictors available but suspects that the vast majority have no significant signal on the variable being forecasted. This very common when forecasting financial market variables such as stock prices where, in theory, it should not be possible to forecast successfully the future path of a stock price with past information. One would expect that the Lasso would perform well relative to Ridge when there are many inactive predictors, however, under high correlation this is not the case so this design allows one to see the relative importance of these 2 features of the DGP. It is worth noting that the fact that the *first* few are active is irrelevant due to the fact that the variables are equally correlated with one another. Finally, the disturbance component is simulated such that $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ where σ_ϵ^2 is computed based on the desired signal-to-noise ratio (SNR) with the following relationship:

$$SNR = \frac{\beta'\Sigma_X\beta}{\sigma_\epsilon^2}\,. \tag{2.22}$$

Throughout this study, the SNR is set to a value of 2 corresponding to an R squared value of 67% as this provides a good balance between settings that are inherently noisy and those that have a denser true DGP. This is significant for the interpretability of the results with regards to the empirical use of such models. For example, Hastie et al. (2020) point out that many financial market applications experience very low SNRs in their true DGPs due the presence of a lot of noise whereas data faced in macroeconomics is likely to experience this to a much a lesser extent. One can show that, by construction, 2.22 is such that as p increases and as ρ increases then σ_ϵ^2 must increase as well in order to keep the SNR fixed at 2. This has implications on the performance metrics used and will be discussed shortly in Section 2.3.3.

2.3.2 Design 2: Persistence

For this second experiment, the aim is to gain an insight into how the methods mentioned previously perform relatively when the candidate predictors exhibit varied levels of persistence. The level of persistence is adjusted throughout approaching the unit root mark in a similar setting to that used by Gonzalo and Pitarakis (2020). Specifically, the underlying DGP takes the following form:

$$y_t = x_t'\beta_0 + z_t'\gamma_0 + \epsilon_t\,. \tag{2.23}$$

Which can also be written as $Y = X\beta + \epsilon$ letting $X_t = \left(x_t', z_t'\right)$ and $\beta = \left(\beta_0, \gamma_0\right)$ with X_t representing row t of X. Where x_t is a vector with p_1 rows and z_t a vector with p_2 rows along with their coefficient vectors β_0 and γ_0 respectively. As usual, ϵ_t is an idiosyncratic error term with mean 0 and variance σ_ϵ^2 which is decided in the same way as detailed in the previous section with expression 2.22 linking this to the SNR. The variables in z_t are simulated in a way that means that they do not experience temporal dependence, and will be discussed shortly, whereas the x_t variables are simulated as AR(1) processes. For $t = 1, ..., T$ the x_t values are generated in the following way:

$$x_t = \left(I_{p_1} - \frac{C}{T}\right)x_{t-1} + v_t\,, \tag{2.24}$$

where $C = diag(c_1, ..., c_{p_1})$ and for this study all values of c are the same and varied over the grid $c \in \{1, 10, 20, 50, 100\}$. Here, one can see that this results in the x_t variables following AR(1) processes with the case of c=1 being very close to the unit root process of a random walk. This can be seen by considering a single column of x_t as being equal to $\left(1 - \frac{c}{200}\right)x_t$ resulting in an AR(1) process with coefficient 0.995 when c is equal to 1. The noise component v_t is generated in coalition with the z_t variables in order to allow for some correlation between the persistent and non-persistent

predictors and is as follows; first let $G = (V, Z)$ and is represented visually below:

$$G_{T\times p} = \begin{bmatrix} v_{11} & v_{21} & \dots & v_{p_1 1} & z_{11} & \dots & z_{p_2 1} \\ v_{12} & v_{22} & \dots & v_{p_1 2} & z_{12} & \dots & z_{p_2 2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{1T} & v_{2T} & \dots & v_{p_1 T} & z_{1T} & \dots & z_{p_2 T} \end{bmatrix}, \tag{2.25}$$

where G is simulated in the exact same way as X is in equation 2.20 with ρ set to 0.3. This allows for a small amount of correlation amongst the purely stationary predictors but also between persistent and non-persistent predictors. It is also worth noting as a small detail that the simulation occurs such that x_0 is a row vector of zeros but is then discarded from the final predictor matrix once $x_1, ..., x_T$ are simulated.

Regarding the coefficients of the true process determining y_t, it is important to mention first that for this particular experiment, p_1 and p_2 are set as being equal making the ratio of persistent variables to purely stationary ones 1:1. This implies that for 20 candidate predictors there are 10 persistent and 10 non-persistent, for 100 candidates there are 50 of each type of variable and this pattern continues for 200 and 400 candidates. The coefficient matrices β_0 and γ_0 are such that the first 10% of the predictors from both x_t and z_t have a coefficient of 1 and the rest 0. For example, where there are 20 candidate predictors (10 persistent and 10 purely stationary) 1 of the variables in x_t and 1 of the variables in z_t have a coefficient of 1. Similar to the previous experiment, the order the variables in the matrix does not have any implications with regard to their properties meaning that the first 10% of variables from the left can be chosen with no difference from the case where a random 10% are chosen.

2.3.3 Performance Evaluation Metrics

This short subsection outlines the metrics used to evaluate the performance of each model relative to its competitors. It is important to remember that the main objective of this exercise is to establish which models are more useful with certain data sets for forecasting values of the dependent variable. Therefore, the general feature of interest is $x_t'\hat{\beta}$ relative to y_t as opposed to simply the sign, magnitude and significance of the $\hat{\beta}$ values, as is the case in studies investigating causal effects. Metrics from the paper by Hastie et al. (2020) are used with the first being the relative test error which is defined as:

$$RTE = \frac{E\left(y_{200} - x_{200}'\hat{\beta}\right)^2}{\sigma_\epsilon^2}, \tag{2.26}$$

where y_{200} is the true value of the dependent variable from the test sample. This ratio represents the variance of the prediction error relative to the variance of y_t according to the underlying DGP. Due to how y_t is constructed from a signal component $(X\beta)$ and a stochastic error term (ϵ), this variance is simply given by the variance of the error term, σ_ϵ^2. The perfect score is 1 in the case where the coefficients are estimated in a way as good as how as they appear in the true DGP but otherwise this shows the relative magnitude of variance of the errors arising from the estimated model to the variance of noise in the underlying process. Therefore, this is not affected by how the error variance naturally increases as p and ρ increase allowing one to compare the performance of a given method across different experiments. The expectation is carried out over the 1000 single values of y_{200} and $x'_{200}\hat{\beta}$.

Secondly, the widely used mean-squared-forecast-error (MSFE) is used which is a simpler case of the RTE defined by:

$$MSFE = E\left(y_{200} - \hat{y}\right)^2 = E\left(y_{200} - x'_{200}\hat{\beta}\right)^2. \tag{2.27}$$

This is simply a measure of pure prediction accuracy making it the most important statistic with regards to the purpose concerned here which is to forecast a certain economic variable as precisely as possible. Unlike the case with the RTE, this metric does not consider the magnitude of the error variance, therefore, as the number of variables increases the MSFEs will naturally be higher regardless of how successful the model is at handling the increased dimensionality of the data. As a result this statistic can only be used to compare the performance of multiple models across the same experiment and not across different experiments unlike the RTE.

Finally, for the Lasso based methods, the number of zero entries in the estimated predictor matrix is reported. This helps to explain the relative performance of these methods and whether it comes from the shrinkage or model selection element.

2.4 Simulation Results

In this section the results of the Monte Carlo experiment detailed in the previous section are discussed with the prime focus being on the mean squared forecasting errors as the purpose of this study is to determine which methods are superior for forecasting economic variables. The tables below present the MSFE for each method (by row) in each experiment (by column) across the range of numbers of candidate predictors with the full sample fixed at 200 throughout. The oracle MSFE is computed by running OLS on only the set predictors that have a non-zero coefficient under the true DGP, therefore, should not be viewed as an approach in its own right but a best

case scenario to provide perspective on the MSFEs of other approaches. However, as this still involves the use of OLS, which can be suboptimal under certain conditions of the true DGP, there is potential for other approaches to improve upon this. It is worth noting that, due to the coding of these experiments being split into 3 parts (PC, RP and regularisation methods) the oracle and OLS quantities are averaged across the 3 code blocks, however, this does not impact the main conclusions. Finally, for each simulation, the number of simulated data sets where the IC fails to hold for the Lasso is recorded in the row labelled Irrepresentable Condition with a value of 1000 implying that for every simulated data set, the IC does not hold.

2.4.1 Design 1 Results

Tables 2.1-2.4 show the MSFEs of each approach for each experiment with a different amount of correlation amongst the covariates. The first feature noticeable from the 4 results tables below is how Ridge performs relatively at its best when the correlation is high for all values of p. For example, in Table 2.4 it provides the second or third lowest MSFE for experiments where $\rho \geq 0.25$ and in Table 2.3 it provides the lowest MSFE when $\rho = 0.95$ but ranks much worse when the ρ is equal to 0 or 0.25. This highlights the benefits of what was discussed in section 2 regarding how Ridge shrinks correlated predictors towards each other simultaneously as this would happen to the greatest extent in settings with a higher ρ compared to the Lasso which discards most of the correlated predictors and this is seen for the larger 3 values of p in Table 2.A9. However, it also loses its ability to select the correct variables consistently as seen by how across all values of p that the number of times that the IC condition fails as ρ increases. The Adaptive Lasso overcomes this issue through its adjustable weighting scheme and, hence, does slightly better than ordinary Lasso for almost all experiments with the exception of when $p = 400$.

The methods involving OLS after regularisation enjoy some success when p is low and or ρ is low also. Specifically, OLS post Elastic Net is in the top 2 most accurate methods when $\rho = 0$ for all p with OLS post Adaptive Lasso performing similarly. The reason for this is likely down to how when there is little correlation between the predictors, choosing the correct ones becomes more important compared to the case whereby there is significant dependence amongst the variables meaning that even if a truly active variable is discarded then some of its signal will be captured through another variable that is included. This combined with how the IC holds in more cases when the correlation is low justifies the success of these selection methods that rely on the IC holding. This does raise questions as to why the OLS post Lasso approach does not experience quite the same success compared to that of OLS post Adaptive Lasso and Elastic Net which could intuitively be justified by the increased flexibility of

penalty weighting and allowance of shrinkage respectively.

The sparse setting means that PC struggles due to the factors being constructed without distinguishing between signal and noise variables leading to low predictive accuracy of the factors. Despite this, when $p = 200$ and $p = 400$ PC does very well compared to other models, especially for a high ρ. For example, in Table 2.4 PC provides the lowest MSFEs for ρ equal to 0.5, 0.8 and 0.95 with it being 6.27%, 4.78% and 4.01% lower than the next best method for each experiment respectively. This is likely due to it being immune to the consequences of $p \gg T$ that methods like the Lasso face along with its ability to exploit the correlation between active and inactive predictors resulting in factors that are genuinely informative with respect to y. RP follows a similar pattern of success whereby it is no better than the other models for low values of p but when $p = 400$ it is clearly superior to most. Table 2.4 shows that only PC is superior when $\rho \geq 0.5$ and RP provides the lowest MSFE for $\rho = 0.25$. This because, similar to PC, the estimation issues that arise when $p \gg T$ are not relevant for RP due to how it reduces the dimension of the original data set before any least-squares based analysis. Unlike PC, RP does not suffer as badly with respect to prediction accuracy for the lower values of p and is often in the same realm as the regularisation methods making it a reliable technique for a wide variety of sparse scenarios.

Finally, It is clear to see that as the level of correlation increases then so does the error variance used by construction. This results in naturally higher MSFEs meaning that to distinguish between the effect on forecasting performance of the higher error variance and the predictor correlation itself, the RTEs of Tables 2.1-2.4 can be used as they remove the effect of the error variance. These tables still show similar attributes to that discussed previously whereby Ridge is very successful when the correlation is high and even as p increases beyond T. For example, Tables 2.A3 and 2.A4 show Ridge providing the lowest RTE when $\rho = 0.95$. Principal components is still relatively successful when p and ρ are both high with it providing the lowest RTE when $\rho = 0.8$ for both cases with $p > T$. Regarding the Lasso and similar models the RTEs show nothing decisive in addition to that discussed when concerning the MSFEs and remain very competitive across the majority of the experiments, especially when ρ is small.

TABLE 2.1: Design 1 MSFEs when $p = 20$

MSFE	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
Oracle	0.4917	0.5073	0.4831	0.5185	0.5006
OLS	0.5504	0.5609	0.5374	0.5640	0.5500
Ridge	0.6057	0.6144	0.6923	0.6035	0.5426
Lasso	0.4948	0.5147	0.5660	0.5212	0.5174
Adaptive Lasso	0.4910	0.5119	0.5548	0.5085	0.5113
Elastic Net	0.5049	0.5181	0.5742	0.5264	0.5217
Principal Components	0.6186	0.6083	0.5727	0.5482	0.5429
Random Projections	0.6065	0.5911	0.5237	0.5568	0.5338
OLS post Lasso	0.5093	0.5176	0.5627	0.5122	0.5106
OLS post Adaptive Lasso	0.4883	0.5122	0.5499	0.5056	0.5037
OLS post Elastic Net	0.4878	0.5088	0.5490	0.5138	0.5116
Irrepresentable Condition	0	0	0	0	10
Average error variance	0.5	0.5	0.5	0.5	0.5

TABLE 2.2: Design 1 MSFEs when $p = 100$

MSFE	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
Oracle	2.5377	5.1732	7.8471	10.5051	12.2374
OLS	5.0082	9.9651	15.3562	21.0368	24.2611
Ridge	4.1990	7.3793	8.6951	12.0244	12.3091
Lasso	2.9392	6.2707	8.5997	12.4149	12.4996
Adaptive Lasso	2.7538	5.7973	8.3245	12.2054	12.9215
Elastic Net	3.3041	6.5609	8.2711	11.9380	12.5311
Principal Components	5.8547	7.7784	9.7711	11.1438	11.8313
Random Projections	3.8719	6.6672	9.7898	11.1232	11.8584
OLS post Lasso	3.7090	5.5348	7.9016	11.7033	12.3767
OLS post Adaptive Lasso	2.6591	5.7584	8.3212	12.2954	12.9251
OLS post Elastic Net	2.6025	5.6992	7.8831	11.7872	12.5396
Irrepresentable Condition	0	11	449	966	1000
Average error variance	2.5	5	7.5	10.5	12

TABLE 2.3: Design 1 MSFEs when $p = 200$

MSFE					
	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
Oracle	5.4021	17.0962	29.3973	42.9812	49.8578
Ridge	10.6940	22.4328	35.0879	43.9825	44.5307
Lasso	7.6125	22.7511	36.6475	44.9498	45.9177
Adaptive Lasso	6.7834	19.9242	36.0227	43.6921	44.7391
Elastic Net	9.1778	20.5402	35.4861	44.0525	44.8980
Principal Components	14.4134	23.9136	32.8153	42.8320	45.2997
Random Projections	10.6958	21.4803	32.4008	44.2836	47.1806
OLS post Lasso	11.0260	20.2669	35.6277	44.0108	44.5608
OLS post Adaptive Lasso	6.6480	20.9354	36.1093	44.8028	44.9562
OLS post Elastic Net	6.3573	20.2994	35.5833	44.0538	44.7159
Irrepresentable Condition	5	924	1000	1000	1000
Average error variance	5	16.25	27.5	41	47.75

TABLE 2.4: Design 1 MSFEs when $p = 400$

MSFE					
	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
Oracle	10.7400	62.9182	120.1296	178.3564	213.4914
Ridge	22.8872	74.4894	124.7402	181.4494	193.9326
Lasso	18.2605	77.7497	129.8990	184.8912	198.1597
Adaptive Lasso	16.3675	80.6752	134.7272	190.5042	199.7811
Elastic Net	19.2641	77.0968	128.7885	183.9489	197.1652
Principal Components	29.2629	70.9639	110.4807	169.7838	186.1484
Random Projections	25.7437	69.7068	117.8771	178.3143	207.6980
OLS post Lasso	21.4845	79.2908	132.9492	181.7997	198.7956
OLS post Adaptive Lasso	18.3191	83.5807	141.9415	190.5860	205.9468
OLS post Elastic Net	17.8980	80.3486	133.0098	182.9584	200.0020
Irrepresentable Condition	630	1000	1000	1000	1000
Average error variance	10	57.5	105	162	190.5

In summary, while it appears to be very close in terms of predictive reliability of each of these methods, in the high dimensional settings Principal Components has and edge once the significant correlation is present. However, Ridge appears to be the best alternative when looking overall at the the experiments involving a high p and ρ. One could argue that these results are less apparent for the case where $\rho = 0.95$, however, this is a more unrealistic scenario as a means to observing how the methods handle the extreme case and it is argued that more attention should be given to the case where

the correlation is between 0.5 and 0.8 where the conclusions are much more concrete.

2.4.2 Design 2 Results

Here, the role of persistence in forecasting performance is considered with Tables 2.5-2.8 showing the MSFEs of each method across all experiments with varied temporal dependence in the predictors. While for $p = 20$ and lower persistence the results show methods such as OLS and OLS post selection performing well, it is very clear from the tables below that RP is very dominant in terms of forecasting accuracy and this is amplified as the level of temporal dependence approaches the unit root mark. The effects of persistence on PC are discussed in Gonzalo and Pitarakis (2020) where they detail how factors are created spuriously due to how the trend nature of unit root variables creates artificial correlation and hence misguided parameter estimates. Despite this, PC still performs relatively well when $p = 200$ and $p = 400$ for persistence that is not too close to the unit root case which is likely to be down to its ability to handle the ultra-high dimension of the predictor matrix more successfully than the regularisation approaches. For example, when $p = 400$ PC always provides the second lowest MSFE for all experiments (only 6.16% greater for $c = 100$).

To highlight how it is the persistence that is the key determinant in these stand-out results, one can look at the RTEs in tables 2.A5-2.A8 which eliminates the effects of the increased error variance as p increases, leaving only the affect of c. While this naturally makes the MSFEs higher, this is not the case for the RTEs, therefore, the appendix tables show how the persistence increasing worsens the performance of all methods with the exception of RP, and can be arguably even be viewed as improving as the unit root mark is approached in some cases. For some cases, such as in Table 2.A6 where $p = 100$ and $c = 100$, the RP RTE is 1.33029 which is noticeably higher than the other methods showing how a small amount of persistence is not a problem for the other approaches in a relative sense. However, this swings in favour of RP very quickly as p increases and c decreases.

As the degree of temporal dependence increases, the methods based on minimising a certain form of the sum of squared residuals suffer in similar fashion to how one can show that unit roots cause the variances of the OLS coefficient estimates to diverge providing undesirable parameter estimates, hence, adversely affecting forecasting accuracy. One can also see that the IC fails to hold in the vast majority of cases when $p > T$, and more so as the amount of persistence increases. This explains why the selection methods also struggle in this setting as they also rely on the l_1 norm penalty of the Lasso. RP side steps this issue due to the nature of how the predictor matrix is

multiplied by a matrix of random weights multiple times and is then averaged resulting in the persistence, and spurious correlation, diluting. This feature in unison with how the pairwise distances between points are maintained by RP result in a situation almost as good as the hypothetical case where forecasts are being computed from a predictor matrix which is equally as informative in a signalling context but without the persistence imposing huge elements of unreliability.

TABLE 2.5: Design 2 MSFEs when $p = 20$

MSFE

	$c = 100$	$c = 50$	$c = 20$	$c = 10$	$c = 1$
Oracle	1.4910	1.9438	3.0585	5.0570	12.9060
OLS	1.6602	2.1783	3.3656	5.7125	14.7498
Ridge	1.8497	2.7073	5.4035	11.4003	88.1457
Lasso	1.6312	2.4744	5.0016	10.8768	87.3726
Adaptive Lasso	1.5975	2.4147	4.9324	10.7198	87.1193
Elastic Net	1.6556	2.4981	5.0290	10.9323	87.6209
Principal Components	1.6721	2.0742	4.0338	7.7128	47.0529
Random Projections	1.6777	2.2497	3.2666	5.7366	14.2155
OLS post Lasso	1.5676	2.2735	3.7049	7.3485	48.8829
OLS post Adaptive Lasso	1.5257	2.1932	3.5295	7.0681	48.8884
OLS post Elastic Net	1.5266	2.2003	3.5324	7.0730	49.4638
Irrepresentable Condition	0	0	2	72	690
Average error variance	1.4562	1.9023	3.1145	4.8464	12.1766

TABLE 2.6: Design 2 MSFEs when $p = 100$

MSFE

	$c = 100$	$c = 50$	$c = 20$	$c = 10$	$c = 1$
Oracle	22.0015	25.9000	40.3075	61.2762	150.4699
OLS	42.2371	54.8054	88.8557	141.8617	359.7707
Ridge	26.1475	33.6745	66.8063	125.0188	894.3338
Lasso	26.9773	35.6799	68.0647	123.4371	893.0332
Adaptive Lasso	26.0415	33.5810	67.4054	124.7783	910.0687
Elastic Net	25.0057	32.7401	64.3754	120.8191	892.4175
Principal Components	25.4562	33.4590	52.9072	90.4496	545.8734
Random Projections	26.8310	31.3168	48.8571	70.3160	159.1336
OLS post Lasso	24.6546	30.0488	50.5893	90.2625	548.8515
OLS post Adaptive Lasso	26.6404	32.7472	56.6149	96.9567	574.5897
OLS post Elastic Net	24.5550	30.0380	50.1273	89.8740	551.1874
Irrepresentable Condition	27	549	943	991	999
Average error variance	20.1113	24.9204	38.8803	57.2359	138.8406

TABLE 2.7: Design 2 MSFEs when $p = 200$

MSFE	$c = 100$	$c = 50$	$c = 20$	$c = 10$	$c = 1$
Oracle	80.9748	98.5419	150.1243	217.4690	567.7982
Ridge	86.0975	117.6042	235.7314	404.7910	3221.1010
Lasso	87.9541	122.8703	240.2854	411.5388	3226.1190
Adaptive Lasso	91.2503	118.5824	234.9709	418.4911	3113.0450
Elastic Net	87.6826	119.9602	240.1583	410.6078	3222.3180
Principal Components	88.1009	116.4923	174.8450	300.4104	1967.6070
Random Projections	86.2094	109.2127	162.7634	233.2920	541.2370
OLS post Lasso	88.6760	114.6339	195.9120	340.1472	1989.2080
OLS post Adaptive Lasso	94.2684	120.6898	208.3282	368.5898	2104.0410
OLS post Elastic Net	88.6713	113.2021	197.8948	339.0118	2000.7920
Irrepresentable Condition	776	998	1000	1000	1000
Average error variance	72.6319	88.7505	134.4669	194.4776	464.9777

TABLE 2.8: Design 2 MSFEs when $p = 400$

MSFE	$c = 100$	$c = 50$	$c = 20$	$c = 10$	$c = 1$
Oracle	331.8040	389.3076	642.6516	919.2129	2255.6050
Ridge	318.5808	423.3456	816.2487	1506.6080	11145.9900
Lasso	340.7655	459.6387	830.1582	1495.3370	11012.2000
Adaptive Lasso	353.6919	472.3033	901.7242	1630.3720	11303.3300
Elastic Net	332.5640	453.6294	828.7903	1504.8340	11025.9000
Principal Components	298.9232	385.5316	644.0417	1085.2610	6683.9200
Random Projections	281.5807	335.4002	558.8483	810.9149	1865.9430
OLS post Lasso	345.7049	434.8807	726.6423	1184.1160	6832.3460
OLS post Adaptive Lasso	397.3862	509.3047	1023.0480	1683.5330	9402.9090
OLS post Elastic Net	343.1010	433.7461	741.3599	1216.7780	6826.7320
Irrepresentable Condition	1000	1000	1000	1000	1000
Average error variance	275.1854	332.6399	496.8369	714.0774	1699.3830

In summary, Random Projections clearly dominates all other methods across the majority of experiments for the reasons mentioned previously, even when first differencing was used the general theme throughout the results remains the same. While the other methods remain competitive for smaller values of p and less persistence, especially the OLS-post-slection methods, one could easily make the case for the use of these methods when the covariates used are not too close to unit roots (or there are less of them compared to stationary processes). Although as the dimensioanlity increases, the selection element of these methods face new challenges

as discussed previously which is why RP dominates as p increases.

2.5 Discussion

This paper has created a platform for the comparison of commonly used methods to forecast with a large number of predictors available with the focus being on data sets typically faced by economists. To do this, a detailed Monte Carlo simulation study was carried out with the DGP fine tuned to replicate characteristics of typical data sets associated with economics. These features include: noticeable correlation amongst the predictors, temporal dependence amongst the individual regressors, a sparse DGP and a high level of noise relative to the signal from the true predictors. Through keeping the degree of sparsity and the SNR constant across experiments the findings show some interesting results concerning relative performance across varied persistence and correlation. Firstly, as expected, Ridge improves relatively as the level of correlation increases and the ratio $\frac{p}{T}$ as well. This is likely down to how the pivotal irrepresentable condition of the Lasso fails more frequently as ρ increases causing inconsistent model selection and, in turn, more uncertain forecasts. However, as it starts become that $p \gg T$ then RP and PC begin to triumph over the regularisation based approaches due to their dimension reduction ability proving more valuable than the shrinkage and selection features of the other methods which appears to be more fruitful for a smaller $\frac{p}{T}$.

For persistence, the results are a lot more conclusive, strongly in favour of RP which appears to be immune to the breakdown of standard assumptions necessary for OLS asymptotic properties arising from predictor temporal dependence approaching the unit root mark. While for lower values of p there is some positive signs for the OLS post regularisation methods when the persistence is somewhat less, it is very clear that RP is the only method not being severely impacted by the serial correlation issues and, hence, can deemed as the most suitable approach when forecasting with a set of predictors that are notoriously persistent.

While this paper does not analyse in great depth the theoretical features of all these methods, the practical focus of the DGP design has provided a useful guide for economic forecasters who are faced with a large number of variables and wish to use the most appropriate model to maximise forecasting accuracy. The success of RP under persistent predictors demonstrated in this paper also provides justification to the commitment of further research into the use of RP for economic or financial forecasting as a means of improving more accurate and reliable predictions.

2.A Appendix

2.A.1 RP Subspace Dimension and Number of Random Draws

FIGURE 2.A1: Relative MSFEs of RP forecasts across a variety of subspace dimensions for a varying number of draws of random weights matrices under no correlation in the predictor matrix

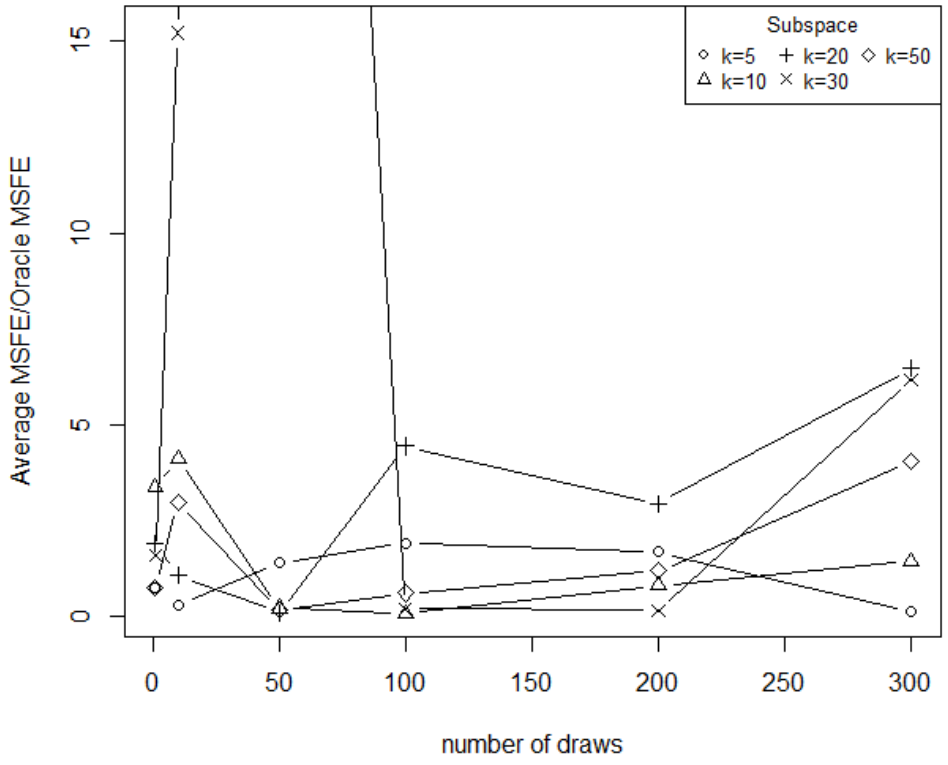
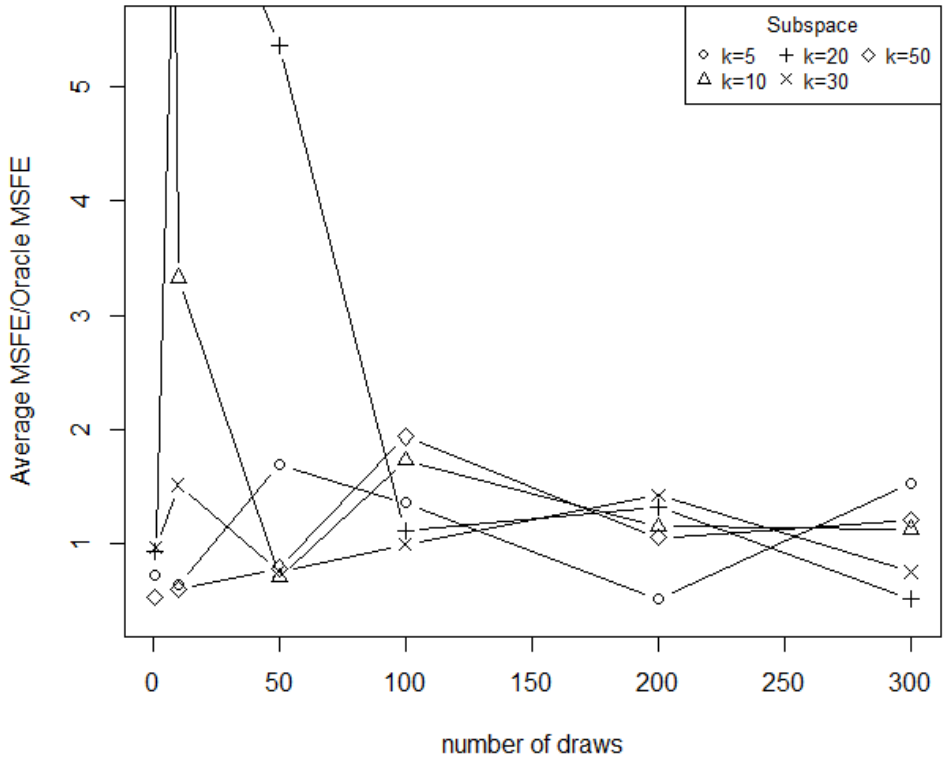


FIGURE 2.A2: Relative MSFEs of RP forecasts across a variety of subspace dimensions for a varying number of draws of random weights matrices under high correlation in the predictor matrix $\rho = 0.8$ in the design 1 DGP



For Figures 2.A1 and 2.A2, a forecasting experiment is carried out in exactly the same fashion as that of the design 1 experiments. Only instead of comparing different methods, the RP procedure is run multiple times with varying subspace (k) and varying number of draws of the R matrix (h). Each line represents the MSFE (relative to the oracle MSFE) associated for a given subspace dimension when using a different number of draws.

2.A.2 Design 1 Relative Test Errors

TABLE 2.A1: Design 1 relative test errors for $p = 20$

RTE

	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
Ridge	1.21144	1.22879	1.38464	1.20694	1.08519
Lasso	0.98961	1.02949	1.13198	1.04247	1.03488
Adaptive Lasso	0.98196	1.02372	1.10957	1.01696	1.02260
Elastic Net	1.00972	1.03622	1.14833	1.05272	1.04342
Principal Components	1.23716	1.21649	1.14544	1.09644	1.08574
Random Projections	1.21297	1.18227	1.04731	1.11351	1.06766
OLS post Lasso	1.01866	1.03512	1.12531	1.02442	1.02112
OLS post Adaptive Lasso	0.97666	1.02435	1.09979	1.01116	1.00747
OLS post Elastic Net	0.97552	1.01759	1.09799	1.02751	1.02313

TABLE 2.A2: Design 1 relative test errors for $p = 100$

RTE

	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
Ridge	1.67961	1.47586	1.15935	1.14518	1.02576
Lasso	1.17568	1.25414	1.14663	1.18237	1.04163
Adaptive Lasso	1.10150	1.15947	1.10994	1.16242	1.07679
Elastic Net	1.32165	1.31217	1.10282	1.13695	1.04426
Principal Components	2.34187	1.55567	1.30282	1.06132	0.98594
Random Projections	1.54876	1.33344	1.30530	1.05935	0.98820
OLS post Lasso	1.48358	1.10696	1.05354	1.11460	1.03140
OLS post Adaptive Lasso	1.10244	1.15168	1.10950	1.17099	1.07709
OLS post Elastic Net	1.04101	1.13985	1.05108	1.12259	1.04497

TABLE 2.A3: Design 1 relative test errors for $p = 200$

RTE

	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
Ridge	2.13879	1.38048	1.27593	1.07274	0.93258
Lasso	1.52249	1.40007	1.33264	1.09634	0.96163
Adaptive Lasso	1.35668	1.22611	1.30992	1.06566	0.93694
Elastic Net	1.83556	1.26401	1.29041	1.07445	0.94027
Principal Components	2.88269	1.47160	1.19328	1.04468	0.94868
Random Projections	2.13916	1.32186	1.17821	1.08009	0.98807
OLS post Lasso	2.20520	1.24720	1.29555	1.07343	0.93321
OLS post Adaptive Lasso	1.32960	1.28833	1.31306	1.09275	0.97629
OLS post Elastic Net	1.27145	1.24919	1.29394	1.07448	0.93646

TABLE 2.A4: Design 1 relative test errors for $p = 400$

RTE

	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
Ridge	2.28872	1.29547	1.18800	1.12006	1.01802
Lasso	1.82605	1.35217	1.23713	1.14130	1.04021
Adaptive Lasso	1.63675	1.40305	1.28312	1.17595	1.04872
Elastic Net	1.92641	1.34081	1.22656	1.13549	1.03499
Principal Components	2.92629	1.25374	1.05220	1.04804	0.97716
Random Projections	2.57437	1.21229	1.12264	1.10071	1.09028
OLS post Lasso	2.14845	1.37897	1.26825	1.12222	1.04355
OLS post Adaptive Lasso	1.83191	1.45358	1.35182	1.17646	1.08109
OLS post Elastic Net	1.78980	1.39737	1.26676	1.12937	1.04988

2.A.3 Design 2 Relative Test Errors

TABLE 2.A5: Design 2 relative test errors for $p = 20$

RTE

	$c = 100$	$c = 50$	$c = 20$	$c = 10$	$c = 1$
Ridge	1.25197	1.44016	1.75372	2.32151	6.74858
Lasso	1.10212	1.31947	1.62197	2.19727	6.63989
Adaptive Lasso	1.07880	1.28787	1.59926	2.15805	6.60965
Elastic Net	1.11883	1.33021	1.62828	2.20978	6.66113
Principal Components	1.14519	1.09529	1.28389	1.65805	3.87888
Random Projections	1.14998	1.17805	1.05061	1.26426	1.16363
OLS post Lasso	1.06727	1.22653	1.20093	1.54220	4.09601
OLS post Adaptive Lasso	1.03279	1.17655	1.15603	1.47104	4.16753
OLS post Elastic Net	1.03343	1.17915	1.15068	1.47070	4.18033

TABLE 2.A6: Design 2 relative test errors for $p = 100$

RTE

	$c = 100$	$c = 50$	$c = 20$	$c = 10$	$c = 1$
Ridge	1.30502	1.33450	1.74869	2.22638	6.22512
Lasso	1.34511	1.40607	1.78570	2.20323	6.16961
Adaptive Lasso	1.29856	1.32826	1.75057	2.21476	6.25435
Elastic Net	1.24777	1.30049	1.68062	2.14373	6.15485
Principal Components	1.26354	1.33794	1.39636	1.64701	4.23520
Random Projections	1.33029	1.26679	1.27068	1.26565	1.17911
OLS post Lasso	1.22702	1.19546	1.29818	1.60546	4.14456
OLS post Adaptive Lasso	1.32654	1.29664	1.47663	1.73070	4.28493
OLS post Elastic Net	1.22191	1.19446	1.30334	1.59842	4.15562

TABLE 2.A7: Design 2 relative test errors for $p = 200$

RTE

	$c = 100$	$c = 50$	$c = 20$	$c = 10$	$c = 1$
Ridge	1.20311	1.33776	1.77125	2.13400	6.55693
Lasso	1.22960	1.39467	1.79742	2.15950	6.53533
Adaptive Lasso	1.27321	1.34829	1.77445	2.19136	6.38223
Elastic Net	1.22591	1.36232	1.80075	2.15604	6.53074
Principal Components	1.20968	1.30360	1.29693	1.59204	4.30362
Random Projections	1.19298	1.22601	1.19774	1.16640	1.16795
OLS post Lasso	1.23920	1.31191	1.47665	1.73908	4.19604
OLS post Adaptive Lasso	1.31698	1.38076	1.57213	1.89712	4.50108
OLS post Elastic Net	1.23890	1.29375	1.49119	1.76499	4.20724

TABLE 2.A8: Design 2 relative test errors for $p = 400$

RTE

	$c = 100$	$c = 50$	$c = 20$	$c = 10$	$c = 1$
Ridge	1.16182	1.24271	1.65842	2.18159	6.39346
Lasso	1.24546	1.35277	1.69448	2.18200	6.28238
Adaptive Lasso	1.29567	1.39797	1.84492	2.34312	6.48069
Elastic Net	1.21471	1.33514	1.68843	2.18631	6.29970
Principal Components	1.08604	1.18006	1.29450	1.55531	4.17219
Random Projections	1.02071	1.00846	1.14050	1.14750	1.11135
OLS post Lasso	1.26456	1.28756	1.48100	1.74226	4.31791
OLS post Adaptive Lasso	1.45385	1.50442	2.07258	2.35237	5.64906
OLS post Elastic Net	1.25514	1.28339	1.51411	1.78488	4.28352

2.A.4 Zero Coefficients from Lasso-based Methods

TABLE 2.A9: Number of zeros on average in the predictor matrix set by the relevant methods for design 1

$p = 20$	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
Lasso	18.127	18.023	17.655	17.172	16.625
Adaptive Lasso	18.543	18.676	18.197	18.649	18.660
Elastic Net	13.033	13.762	14.744	14.704	12.744
$p = 100$					
Lasso	65.609	57.628	57.354	68.780	88.732
Adaptive Lasso	88.731	85.639	88.347	90.124	93.926
Elastic Net	40.467	86.343	82.153	80.725	79.764
$p = 200$					
Lasso	120.110	176.568	174.632	181.395	189.397
Adaptive Lasso	169.157	158.940	168.462	168.116	188.305
Elastic Net	82.278	166.774	168.475	176.394	181.428
$p = 400$					
Lasso	263.239	353.271	362.544	377.251	387.821
Adaptive Lasso	328.033	362.284	378.220	385.139	392.422
Elastic Net	361.659	348.246	358.981	372.480	375.981

TABLE 2.A10: Number of zeros on average in the predictor matrix set by the relevant methods for design 2

$p = 20$	$c = 100$	$c = 50$	$c = 20$	$c = 10$	$c = 1$
Lasso	14.144	13.384	11.983	10.656	8.863
Adaptive Lasso	16.344	16.132	15.511	14.779	16.429
Elastic Net	9.172	8.224	7.174	6.230	5.162
$p = 100$					
Lasso	34.317	33.978	79.934	80.009	79.174
Adaptive Lasso	78.848	72.438	75.703	75.504	78.709
Elastic Net	73.504	73.267	72.005	70.088	76.136
$p = 200$					
Lasso	161.488	161.633	161.644	159.990	151.311
Adaptive Lasso	159.851	145.187	153.051	163.055	124.865
Elastic Net	157.984	140.341	135.854	156.268	147.403
$p = 400$					
Lasso	350.498	324.984	351.057	348.650	370.719
Adaptive Lasso	361.661	356.422	344.748	333.483	304.611
Elastic Net	341.771	340.264	335.111	349.785	362.539

Chapter 3

A Ridge Regression Modification for Handling High Dimensional Economic Data

3.1 Introduction

As time has progressed forecasters and analysts within the field of econometrics (and other fields) have enjoyed increased accessibility to candidate predictors for what were previously somewhat simple regression models. While, intuitively, this should be seen as an advantage from the perspective of generating more accurate time series forecasts or increased precision of conclusions from causal inference analysis, it has also brought new challenges as incumbent statistical models based upon standard Ordinary Least Squares (OLS) are increasingly ill equipped for the tasks they face. More specifically, one is concerned with a linear model setting where the econometrician has a response variable, y, and p candidate predictor variables stored in a matrix, X, where one suspects that only a certain subset of the p variables has influence on y according to the true data-generating process (DGP). This is detailed below,

$$y = X\beta + \epsilon\,, \tag{3.1}$$

where y is a $n \times 1$ vector of the response variable observations and X is a $n \times p$ matrix of the candidate predictors. ϵ represents the idiosyncratic error term with mean 0 and variance $\sigma_\epsilon^2 I_n$ and β represents the $p \times 1$ vector of coefficients that the analyst wishes to estimate. This could be for many reasons, for example, they may wish to measure the magnitude of a causal effect with many controls present meaning that the estimation accuracy of a given element of β will be the main parameter of interest. Another purpose could be using the in-sample $X\hat{\beta}$ to construct out-of-sample forecasts

for the dependent variable.

Whatever the purpose may be, one must construct an estimate for the coefficient vector with the highest degree of accuracy possible facing a host of challenges from the data set concerned. While OLS is the first intuitive option due to its unbiased nature, as the number of predictors relative to the sample size ($\frac{p}{n}$ increases) OLS results in highly uncertain $\hat{\beta}$ estimates and when $p > n$ OLS is no longer feasible. To remedy this, a wide range of approaches have been practised throughout a diverse range of fields such as regularization methods including Ridge Regression of Hoerl and Kennard (1970) and the Lasso Model of Tibshirani (1996) which adopt penalised least-squares based methods in order to manage the bias-variance tradeoff. Creating a small number of factors from the large predictor has also been a very common approach, especially in economics, using methods such as Principal Components and applied to economic settings in work such as Connor and Korajczky (1988) and Stock and Watson (2002). Finally, Model Selection approaches in work such as Akaike (1969) and Schwarz (1978) have seeked to use OLS on the best subset of candidate predictors while other literature, such as Timmermann (2006) seeks to combine results from many simpler and smaller sub-models.

This paper focuses on models that are well suited to handle the common issue of high dimensionality as well as significant multicollinearity amongst predictors, with both issues being faced by economists more regularly as the ability to collect data improves. This is due to the nature of how data is gathered as well as the way that underlying human behaviour impacts a variety of key economic indicators simultaneously. More specifically, this paper focuses on Ridge Regression due to its ability to collectively shrink the coefficients of correlated predictors towards each other making it a useful tool for economic data. In contrast to methods such as the Lasso or other model selection techniques which often give reduced consideration to the collinearity between predictors, focusing mostly on the relationship of the each covariate with the dependent variable. To formalise, Ridge seeks to estimate the coefficient vector using the following minimization problem in similar fashion to OLS.

$$\hat{\beta}_\lambda = argmin_\beta \frac{1}{n}||y - X\beta||^2 + \lambda||\beta||^2\,, \tag{3.2}$$

which yields the following closed form estimator,

$$\hat{\beta}_\lambda = \frac{1}{n}\left(\frac{X'X}{n} + \lambda I_p\right)^{-1} X'y\,. \tag{3.3}$$

As p moves closer towards n the degree of predictor multicollinearity typically increases due to spurious correlations between variables (be this by chance or due to

similarities in the variable construction process). As a result, OLS estimation suffers from the covariance matrix ($X'X$) becoming increasingly ill-conditioned resulting in highly sensitive parameter estimates whereas Ridge overcomes this with the addition of the λI_p term casuing correlated predictors to have their coefficients shrunk towards each other. Some of the statistical properties of the Ridge estimator under a fixed design are given as follows:

$$Bias(\hat{\beta}_\lambda) = -\lambda \left(\frac{X'X}{n} + \lambda I_p \right)^{-1} \beta \,, \tag{3.4}$$

$$Var(\hat{\beta}_\lambda) = \frac{\sigma_\epsilon^2}{n} \left(\frac{X'X}{n} + \lambda I_p \right)^{-1} \frac{X'X}{n} \left(\frac{X'X}{n} + \lambda I_p \right)^{-1} , \tag{3.5}$$

and it can be shown that, for a $\lambda > 0$, the bias of Ridge is non-zero (greater than OLS) but the variance is lower than that of OLS. Using the relationship between the bias, variance and mean-square-error (MSE), Theobald (1974) derived the following condition on the penalty term to characterise when Ridge parameter estimates will provide a lower excess risk than that of OLS,

$$\lambda < \frac{2\sigma_\epsilon^2}{\beta'\beta} \,. \tag{3.6}$$

The condition shows that, as long as at least one true coefficient is not equal to 0 then there will always exist a $\lambda \neq 0$ which will result in Ridge having a lower excess risk than OLS. Since OLS is unbiased, one can see that this implies that there is always an optimal tradeoff between bias and variance through this form of regularisation to outperform OLS when concerned with prediction accuracy.

Regarding the background of Ridge, the approach first originated in the field of chemical engineering proposed by Hoerl and Kennard (1970) as a means of overcoming the issue of multicollinearity. As mentioned before, this causes problems in regression analysis as it leads to undesirably high standard errors of coefficient estimates which not only raises questions surrounding the magnitude, and sometimes sign, of coefficients but frequently results in predictors, that theoretically are believed as being active, being statistically insignificant. Such an issue is very common in economics (Farrar and Glauber (1967)) where multiple variables are influenced by exogenous shocks in similar ways or variables are aggregated across a large number of cross sectional units, such as firms, households or geographical regions resulting in highly correlated covariates as discussed in Brown and Nawas (1973).

To demonstrate this, Simeon and Olaiya (2021) use Ridge to investigate the magnitude of which various sectors of the Nigerian economy influence GDP growth. There are 9 predictors including indicators of production from agriculture, construction and

services. A correlation coefficient matrix is reported and shows high positive values resulting in OLS showing some false negative coefficients and insignificant predictors. However, much of this is rectified when applying Ridge resulting in more realistic contribution percentages of each sector to GDP. In a different setting, Inoue and Kilian (2008) use a set of 30 predictors from categories including production, labour markets and financial markets to carry out out-of-sample monthly forecasts of the US Consumer Price Index (CPI). Once again, the predictors are heavily correlated and Ridge along with a vast range of other high-dimensional models are used to create forecasts. Despite the large variety of competing models, Ridge is one of the most successful in terms of prediction accuracy highlighting the importance of being able to handle the correlation amongst the predictors, even when forecasting.

In a production function setting, Brown and Beattie (1975) argue in favour of using Ridge by showing how the bias (and MSE) of Ridge is lower when all the true β components are of the same sign and this is also argued by Newhouse and Oman (1971). In addition, for a given coefficient, β_j, they show that it is beneficial for the bias if β_j is similar in magnitude to the average of all other β values. Such features are ideal for production functions as one would expect nearly all coefficients to be positive as the input would contribute to producing a greater output, Also, the input variables are likely to be positively correlated as was the case in the empirical application of Brown and Beattie (1975) which measures the effect of highly correlated agricultural variables on water irrigation. However, Coniffe et al. (1976) argue that while it is true that most of the coefficients in a production function will be positive, quadratic versions on certain variables will be included having a negative sign to allow for diminishing marginal returns. While this need not be the case with a Cobb-Douglas production function of the following form:

$$Y = \alpha X_1^{\beta_1} X_2^{\beta_2} \dots X_p^{\beta_p},$$

where Y is the output variable with input covariates $X_1, \dots, X_p$, it can be seen that diminishing returns can hold for $\beta_i < 1$ allowing all coefficients to be positive. Although, this is not always an appropriate model meaning that situations where these ideal properties do not hold are likely to occur. Other functions, such as demand functions would also prove to be a challenge for Ridge due to the nature of the analyst wishing to include variables concerning complements and substitutes to the product of interest resulting in a mix of signs in the true β components. For example, O'Neill and Buttimer (1972) attempted to use Ridge to approximate the demand function for Irish beef and found themselves estimating the coefficient of the beef price with the wrong sign, highlighting the importance of the conditions discussed previously for the bias to be restricted.

Therefore, while Ridge can be a useful tool to construct reliable forecasts and predictions with correlated predictors, under certain profiles of the true β values, it will become unreliable provoking the need for an alternative procedure. However, one should not be so quick to completely discard the idea of l_2 norm penalisation as this has many strong attributes for handling the correlation and high dimensional nature of the data. Therefore, one might wish to use an estimation procedure that uses Ridge penalisation, but side-steps high bias issues that arises from strong variance in the sign and magnitude of the true coefficients. This paper proposes a new estimation procedure called Partial Ridge that allows flexibility through estimating the coefficients individually in an attempt to improve prediction accuracy of the $\hat{\beta}$ vector. Following the proposition, theoretical analysis is carried out in order to understand how the newly proposed estimation procedure behaves relative to that of Ridge under various DGPs both in terms of the estimates for individual coefficients themselves and then for estimates of the dependent variable, y. The findings show that while Partial can dominate Full Ridge in terms of the MSEs of some individual coefficients, this is not possible for all β estimates nor for predictive accuracy ($X\hat{\beta}$). Therefore a hybrid approach combining Full and Partial Ridge is proposed with noticeable gains in predictive risk over Ridge alone found and acts as a neat alternative for situations when Ridge by itself suffers through its notorious bias induction.

The rest of this paper is organized as follows; Section 3.2 outlines the proposed Partial Ridge approach with Section 3.3 investigating its theoretical properties compared to those of the ordinary Ridge Regression. Section 3.4 compares the relative performance of Full Ridge and the new estimation procedure in a Monte Carlo Simulation setting with Section 3.5 applying these approaches to an empirical application. Section 3.6 concludes and provides an overview of what has been studied throughout the paper.

3.2 Proposed Estimation Procedure

While Ridge has many advantages over OLS when the data is of a high dimension and or the degree of correlation amongst the covariates is strong, there are situations where its bias can make the tradeoff of variance unfavourable. The following small example will illustrate this where the true DGP is given by 3.1 above and there are 2 standardized predictors, x_1 and x_2 with correlation ρ between them resulting in it being such that $x_1'x_1 = x_2'x_2 = n$ and $x_1'x_2 = n\rho$. If one chooses to estimate the coefficients β_1 and β_2 with the Ridge estimator in 3.3 then one can show that the estimator of a given coefficient is given by the following:

$$\hat{\beta}_i(\lambda) = \frac{(1+\lambda)x_i'y - \rho x_j'y}{n\left((1+\lambda)^2 - \rho^2\right)},$$

where $i \in (1,2)$ and $j \neq i$ one can see that the Ridge estimator approaches 0 as $\lambda \to \infty$ but will never be set to exactly 0 unlike in certain cases for Lasso and others. So the penalisation parameter acts as a means of shrinking the magnitude of the estimate. Taking this further, one can use 3.4 to obtain the following bias expressions for each estimate,

$$Bias(\hat{\beta}_1(\lambda)) = \frac{\rho\lambda\beta_2 - \beta_1\lambda(1+\lambda)}{(1+\lambda)^2 - \rho^2},$$

$$Bias(\hat{\beta}_2(\lambda)) = \frac{\rho\lambda\beta_1 - \beta_2\lambda(1+\lambda)}{(1+\lambda)^2 - \rho^2},$$

where β_1 and β_2 represents the value of the coefficients under the true DGP. By squaring these 2 biases and summing them, one obtains the following:

$$||Bias(\hat{\beta}_\lambda)||^2 = \frac{\lambda^2\left[(\rho^2 + (1+\lambda)^2)(\beta_1^2 + \beta_2^2) - 4\rho(1+\lambda)\beta_1\beta_2\right]}{((1+\lambda)^2 - \rho^2)^2}. \tag{3.7}$$

By looking at the numerator one can see that the bias can increase significantly with the magnitude of $\beta_1$and β_2 when the true coefficients are opposite signs and $\rho > 0$ or the true coefficients having the same sign and $\rho < 0$ due to the first part of the expression always being positive. Therefore, one can see how under certain DGPs the bias of Ridge may reach levels that make the tradeoff undesirable. Such a feature is a common theme throughout the literature, with work such as Zhang and Politis (2022) proposing new approaches to overcome this.

Moreover, the nature of how the penalty parameter is fixed for the estimation of all parameters can be somewhat limiting when there are many coefficients to estimate provoking the consideration of a more flexible approach. However, one may not wish to completely discard the l_2 norm penalisation of Ridge due to its remarkable ability to collectively shrink coefficients that are highly correlated with one another. Intuitively, one may consider it strange to penalise all variables equally when it is highly likely that some covariates will be active under the true DGP, needing less penalisation, while others are more likely to be inactive, needing their coefficient shrunk more significantly.

The 2 issues mentioned make one question if it is possible to construct a new estimation procedure that can avoid that bias issues that Ridge faces while also estimating each coefficient individually to allow for a unique penalty parameter to be utilized. While staying in the realm of Ridge based estimation, consider the following estimation set up,

$$\hat{\beta}_{\lambda,i}^{PR} = argmin_\beta\Big(\frac{1}{n}(y - X\beta)'(y - X\beta) + \lambda_i(S_i\beta)'(S_i\beta)\Big), \tag{3.8}$$

where S_i is a selector matrix represented as

$$S_i = I_p - e_i e_i' .$$

Here e_i is a $p \times 1$ vector with all entries equal to 0 except for the ith entry entry being equal to 1, resulting in the idempotent property for S_i ($S_i' S_i = S_i$). Where i refers to the column number of the predictor matrix for which coefficient is being estimated, for example, when estimating β_3 the above minimization problem would have $S_{33} = 0$. This is almost identical to the ordinary Ridge estimator, only with the ith variable not facing penalisation. Following through with 3.8 leads to the following full coefficient vector,

$$\hat{\beta}_{\lambda,i}^{PR} = \frac{1}{n} \left(\frac{X'X}{n} + \lambda_i S_i \right)^{-1} X'y . \tag{3.9}$$

From here, one takes the ith row of the $\hat{\beta}_{\lambda,i}^{PR}$ vector, defined as $\hat{\beta}_i(\lambda_i, S_i) = e_i'(X'X + \lambda_i S_i)^{-1} X'y$ as the estimate for β_i with this process repeated for all $i \in (1, \ldots, p)$ to build a complete profile of coefficient estimates. This complete vector of the β estimates is called the Partial Ridge estimator due to how not quite all variables are penalised with the l_2 norm like ordinary Ridge and is defined below,

$$\hat{\beta}^{PR} = \begin{bmatrix} \hat{\beta}_1(\lambda_1, S_1) \\ \vdots \\ \hat{\beta}_p(\lambda_p, S_p) \end{bmatrix} . \tag{3.10}$$

In order to see this more closely, one can consider the alternative formulation of 3.1 as follows, where x_i represents the variable corresponding to the coefficient of interest and X_{-i} represents the predictor matrix with the ith predictor column removed,

$$y = x_i \beta_i + X_{-i} \beta_{-i} + \epsilon . \tag{3.11}$$

From here, one can rewrite 3.9 as partitioned matrices using this narrative to obtain the following:

$$\begin{bmatrix} \hat{\beta}_i(\lambda_i, S_i) \\ \hat{\beta}_{-i} \end{bmatrix} = \begin{bmatrix} \frac{x_i' x_i}{n} & \frac{x_i' X_{-i}}{n} \\ \frac{X_{-i}' x_i}{n} & \left(\frac{X_{-i}' X_{-i}}{n} + \lambda_i I_{p-1} \right) \end{bmatrix}^{-1} \begin{bmatrix} \frac{x_i' y}{n} \\ \frac{X_{-i}' y}{n} \end{bmatrix} . \tag{3.12}$$

So here, it is seen that one is penalising β_{-i} while not penalising β_i. Now, the standard partitioned inverse formula can be applied with the individual $\hat{\beta}_i$ component being shown by the following expression where $x_i^* = x_i - X_{-i} (\frac{X_{-i}' X_{-i}}{n} + \lambda_i I_{p-1})^{-1} \frac{X_{-i}' x_i}{n}$,

$$\hat{\beta}_i(\lambda_i, S_i) = \frac{x_i^{*'} y}{x_i^{*'} x_i} . \tag{3.13}$$

The same can also be carried out for Full Ridge to obtain the following:

$$\begin{bmatrix} \hat{\beta}_i(\lambda^{FR}) \\ \hat{\beta}_{-i}(\lambda^{FR}) \end{bmatrix} = \begin{bmatrix} \frac{x_i'x_i}{n} + \lambda^{FR} & \frac{x_i'X_{-i}}{n} \\ \frac{X_{-i}'x_i}{n} & \left(\frac{X_{-i}'X_{-i}}{n} + \lambda^{FR} I_{p-1}\right) \end{bmatrix}^{-1} \begin{bmatrix} \frac{x_i'y}{n} \\ \frac{X_{-i}'y}{n} \end{bmatrix}. \tag{3.14}$$

From here, using the standard partitioned inverse formula leads to an easily comparable result to that of the Partial Ridge case, only this time $x_i^* = x_i - X_{-i}(\frac{X_{-i}'X_{-i}}{n} + \lambda^{FR} I_{p-1})^{-1}\frac{X_{-i}'x_i}{n}$,

$$\hat{\beta}_i(\lambda^{FR}) = \frac{x_i^{*'} y}{x_i^{*'} x_i + n\lambda^{FR}}. \tag{3.15}$$

From here, one can carry out a singular value decomposition of $\frac{X_{-i}}{\sqrt{n}}$ to obtain general expressions for the bias, variance and mean-squared-error of both the Partial and Full Ridge estimators to facilitate comparison between the 2 approaches. Firstly, the singular value decomposition (SVD) is defined as follows:

$$\frac{X_{-i}}{\sqrt{n}} = USV', \tag{3.16}$$

where U is an $n \times (p-1)$ matrix, S is a $(p-1) \times (p-1)$ matrix with the singular values in descending order on its diagonal and V is $(p-1) \times (p-1)$ matrix. It is also important to mention that, by the definition of SVD, it is such that $U'U = I_{p-1}$, $V'V = VV' = I_{p-1}$, which is important for simplification.

Proposition 3.1: The mean-squared-error of Full Ridge and Partial Ridge are given by the following expressions:

$$MSE(\hat{\beta}_i(\lambda^{FR})) = \frac{\left(\lambda^{FR}\tilde{x}_i S(S^2 + \lambda^{FR} I_{p-1})^{-1}\beta_{-i}^* - \lambda^{FR}\beta_i\right)^2}{\left(1 - \tilde{x}_i' S^2 (S^2 + \lambda^{FR} I_{p-1})^{-1}\tilde{x}_i + \lambda^{FR}\right)^2} + \frac{\frac{\sigma_\epsilon^2}{n}\left(1 - \tilde{x}_i' S^2 (S^2 + 2\lambda^{FR} I_{p-1})(S^2 + \lambda^{FR} I_{p-1})^{-2}\tilde{x}_i\right)}{\left(1 - \tilde{x}_i' S^2 (S^2 + \lambda^{FR} I_{p-1})^{-1}\tilde{x}_i + \lambda^{FR}\right)^2}, \tag{3.17}$$

$$MSE(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\lambda_i^2\left(\tilde{x}_i S(S^2 + \lambda_i I_{p-1})^{-1}\beta_{-i}^*\right)^2}{\left(1 - \tilde{x}_i' S^2 (S^2 + \lambda_i I_{p-1})^{-1}\tilde{x}_i\right)^2} + \frac{\frac{\sigma_\epsilon^2}{n}\left(1 - \tilde{x}_i' S^2 (S^2 + 2\lambda_i I_{p-1})(S^2 + \lambda_i I_{p-1})^{-2}\tilde{x}_i\right)}{\left(1 - \tilde{x}_i' S^2 (S^2 + \lambda_i I_{p-1})^{-1}\tilde{x}_i\right)^2}, \tag{3.18}$$

where

$$\tilde{x}_i = U'\frac{x_i}{\sqrt{n}}$$

and

$$\beta^*_{-i} = V'\beta_{-i}\,.$$

A proof is provided in Appendix 3.A.1.

Therefore, the main motivation for Partial Ridge as an estimator for β_i is an alternative bias and variance decomposition that could me more favourable than the unbiased OLS and highly efficient Ridge estimator. It also allows a different λ to be used for each estimate which is advantageous as the next section shows how the optimal MSE and its corresponding λ for a single coefficient varies depending the the true coefficients, correlation and signal-to-noise ratio (SNR). However, the main focus of this paper is the alternative bias-variance tradeoff with the potentially flexible penalty parameter left to future work. The next section investigates the theoretical properties of this Partial Ridge estimator compared to Ridge where all variables are penalised with the same λ (Full Ridge).

3.3 Theoretical Results

3.3.1 Assumptions and Definitions

In this section, the theoretical properties of Partial Ridge are analysed in greater detail to provide insight into how the estimation and prediction accuracy of the Partial Ridge estimator varies relative to that of ordinary Ridge. This is done using a standard model setting with the true DGP of a fixed design detailed below,

$$y = x_1\beta_1 + \cdots + x_p\beta_p + \epsilon\,, \tag{3.19}$$

where y is an $n \times 1$ vector representing the dependent variable observations and x_i are $n \times 1$ vectors of observations for each of the p individual covariates with ϵ being the $n \times 1$ disturbance term vector. In matrix form this is shown as follows:

$$y = X\beta + \epsilon\,,$$

where β is a $p \times 1$ vector of the true coefficients and X is a $n \times p$ matrix. For simplicity, the analysis in this first subsection focuses entirely on the estimation of a single coefficient β_1 due to the individualized estimation nature of Partial Ridge, however, one can easily generalize the conclusions to the case where one is estimating the full set of coefficients. Firstly, the following assumptions are made for the fixed design setting before progressing:

A1 The error term components are independently and identically distributed with 0 mean and homoskedastic variance. $\epsilon \sim IID(0, \sigma^2_\epsilon I_n)$.

A2 The predictor covariance matrix $X'X$ is positive semi-definite.

A3 All covariates are standardized with 0 mean and unit variance resulting in it being that $x_j'x_j = n$ for all $j = 1, \ldots, p$ and $x_j'x_h = n\rho_{jh}$ for all $j \neq h$.

It is worth noting that A2 is such that $X'X$ can be non-invertible. Such a feature ensures that in a high-dimensional setting ($p > n$) then OLS as well as many other estimation procedures are no longer feasible. As seen in 3.3, Ridge overcomes this with the addition of the λI_p term to allow invertibility.

3.3.2 Toy Model: Equicorrelation

In this toy model setting the correlation between predictors ρ_{jh} are set to all be equal for simplicity and without loss of generality. This results in the following predictor covariance matrix:

$$\frac{X'X}{n} = \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix}.$$

Here, it is worth noting that, in order for this covariance matrix to be positive definite then it must be such that $\rho > -\frac{1}{p-1}$ as one can show that all covariances being lower would result in a negative determinant.

As mentioned in the previous section, the Full Ridge estimator for the complete coefficient vector is given as follows:

$$\hat{\beta}(\lambda^{FR}) = \frac{1}{n}\left(\frac{X'X}{n} + \lambda^{FR} I_p\right)^{-1} X'y\,.$$

For Partial Ridge, the estimator for the ith coefficient $\hat{\beta}_i(\lambda_i, S_i)$ is given by the ith coefficient of the following vector:

$$\hat{\beta}_i(\lambda_i, S_i) = \frac{1}{n}\left(\frac{X'X}{n} + \lambda_i S_i\right)^{-1} X'y\,.$$

Proposition 3.2: The individual estimators for a given coefficient under this fixed design are as follows:

$$\hat{\beta}_i(\lambda^{FR}) = \frac{(1 + \lambda^{FR} + (p-2)\rho)x_i'y - \rho\sum_{j\neq i}^{p} x_j'y}{n(1 + \lambda^{FR}n - \rho)(1 + \lambda^{FR} + (p-1)\rho)}\,, \tag{3.20}$$

$$\hat{\beta}_i(\lambda_i, S_i) = \frac{(1+\lambda_i+(p-2)\rho)x_i'y - \rho\sum_{j\neq i}^{p} x_j'y}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)}. \tag{3.21}$$

A proof is provided in Appendix 3.A.2.

This leads into evaluating the properties of the estimators detailed above with the following proposition.

Proposition 3.3: The bias, variance and MSE of the Partial Ridge and Full Ridge estimators for a single coefficient under equicorrelation are as follows:

$$Bias(\hat{\beta}_i(\lambda^{FR})) = \frac{\rho\lambda^{FR}\sum_{j\neq i}^{p}\beta_j - \beta_i\lambda^{FR}(1+\lambda^{FR}+(p-2)\rho)}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}, \tag{3.22}$$

$$Bias(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\rho\lambda_i\sum_{j\neq i}^{p}\beta_j}{1+\lambda_i+(p-2)\rho-(p-1)\rho^2}, \tag{3.23}$$

$$Var(\hat{\beta}_i(\lambda^{FR})) = \frac{\sigma_\epsilon^2\Big((\lambda^{FR}-(\rho-1)(1+(p-1)\rho))^2-(\rho-1)(p-1)\rho^2(1+(p-1)\rho)\Big)}{n(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+(p-1)\rho)^2}, \tag{3.24}$$

$$Var(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\sigma_\epsilon^2\Big((\lambda_i-(\rho-1)(1+(p-1)\rho))^2-(\rho-1)(p-1)\rho^2(1+(p-1)\rho)\Big)}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)^2}. \tag{3.25}$$

A proof is provided in Appendix 3.A.3 with the individuals component in 3.A.3.1, 3.A.3.2, 3.A.3.3 and 3.A.3.4 respectively.

Finally, using the relationship of squared bias and variance with MSE, $MSE(\hat{\beta}_j) = [Bias(\hat{\beta}_j)]^2 + Var(\hat{\beta}_j)$, the following expressions can be obtained:

$$MSE(\hat{\beta}_i(\lambda^{FR})) = \frac{\big(\rho\lambda^{FR}\sum_{j\neq i}^{p}\beta_j - \beta_i\lambda^{FR}(1+\lambda^{FR}+(p-2)\rho)\big)^2}{(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+(p-1)\rho)^2} + \frac{\frac{\sigma_\epsilon^2}{n}\Big((\lambda^{FR}-(\rho-1)(1+(p-1)\rho))^2-(\rho-1)(p-1)\rho^2(1+(p-1)\rho)\Big)}{(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+(p-1)\rho)^2}, \tag{3.26}$$

$$MSE(\hat{\beta}_i(\lambda_i, S_i)) = \frac{(\rho\lambda_i \sum_{j\neq i}^{p} \beta_j)^2}{(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)^2} + \frac{\frac{\sigma_\epsilon^2}{n}\Big((\lambda_i-(\rho-1)(1+(p-1)\rho))^2-(\rho-1)(p-1)\rho^2(1+(p-1)\rho)\Big)}{(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)^2}. \quad (3.27)$$

Where one can see that to ensure that the variance terms are positive then it can be such that $\rho > -\frac{1}{p-1}$, which is equivalent to the condition that the covariance matrix above is positive-semidefinite. One can also show that the denominator of the Full Ridge variance will always be greater than that of the Partial Ridge one making the Full Ridge always have a lower variance. Consider the expansion of the following expression in the denominator of 3.24:

$$(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho) = 1+\lambda^{FR}-(p-1)\rho^2+(p-2)\rho + \lambda^{FR}(1+\lambda^{FR}+(p-2)\rho).$$

So it is shown that the Full Ridge variance denominator includes the Partial Ridge Variance denominator squared component plus an additional term which is positive by the condition for the covariance matrix being positive semidefinite along with $\lambda^{FR} > 0$. Therefore for all $\lambda^{FR} = \lambda_i$ the variance of Full Ridge is lower than that of Partial Ridge. Although this will have reduced relevance when considering the full set of predictors as Full Ridge has less flexibility with the penalty parameter (due to it being fixed across all parameters) the superior order of magnitude in it's variance denominator, $O(\lambda_{FR}^4 + p^2\lambda_{FR}^2)$ compared to $O(\lambda_i^2 + p^2)$ of Partial Ridge, leaves it in a very strong position as the dimension increases.

Where Partial Ridge does have potential to outperform comes from the bias term where already one can see that when $\sum_{j\neq i}^{p} \beta_j = 0$ there is zero bias, even when $\beta_i \neq 0$ unlike with Full Ridge. One can see how the squared biased terms behave as the penalty parameter increases by expanding the squared expressions of 3.22 and 3.23 and rearranging the numerator and denominator as a polynomial of λ^{FR} and λ_i. This results in the following expressions where only the top 2 highest order of magnitude terms are considered where it is important to remember that $\lambda > 0$,

$$Bias(\hat{\beta}_i(\lambda^{FR}))^2 = \frac{\beta_i^2\lambda_{FR}^4 + 2(\beta_i^2(1+(p-2)\rho)-\rho\beta_i\sum_{j\neq i}^{p}\beta_j)\lambda_{FR}^3 + O(\lambda_{FR}^2)}{\lambda_{FR}^4 + 2(2+(p-2)\rho)\lambda_{FR}^3 + O(\lambda_{FR}^2)},$$

$$Bias(\hat{\beta}_i(\lambda_i))^2 = \frac{\lambda_i^2\rho^2(\sum_{j\neq i}^{p}\beta_j)^2}{\lambda_i^2 + 2(1-\rho)(1+(p-1)\rho)\lambda_i + O(1)}.$$

Firstly, looking at the Full Ridge squared bias term, one can see that the numerator increases at a faster rate than the denominator in λ^{FR} when $\beta_i^2 > 1$ with this relative rate of expansion increasing in β_i^2. From looking at the λ_{FR}^3 coefficients, again, a larger

β_i^2 results in faster expansion of the numerator but this can also be assisted by the term, $\rho\beta_i \sum_{j\neq i}^{p} \beta_j$, being negative which can be achieved under a positive ρ with β_i and $\sum_{j\neq i}^{p} \beta_j$ being opposite signs.

For Partial Ridge, a large β_i^2 does not impact the bias and one can see that when $\rho^2(\sum_{j\neq 1}^{p} \beta_j)^2 < 1$ then the bias is decreasing as λ_i increases. Therefore, in terms of bias, Partial Ridge can see noticeable gains over Full Ridge in the estimation of a single coefficient when the true magnitude of the given β_i is large and β_i and $\sum_{j\neq i}^{p} \beta_j$ are of opposite signs. Finally, it is worth pointing out that these arguments are relevant for when λ^{FR} and λ_i are greater than 1 with the opposite conclusions being true when the penalty parameters are less than 1. However, this may be undesirable from a variance reduction perspective as choosing $\lambda^{FR} < 1$ reduces the models ability to limit the variance under a high dimension and multicollinearity.

3.3.3 MSE Comparison

To identify situations where Partial Ridge will outperform Full Ridge in a prediction accuracy sense, the MSE expressions of both estimators from 3.26 and 3.27 are considered with the analysis concentrating on how the bias and variance behave to understand these estimators more thoroughly. Firstly, consider the case where there is no correlation amongst the covariates ($\rho = 0$) then the following the expressions of the MSEs are obtained:

$$MSE(\hat{\beta}_i(\lambda^{FR})) = \frac{n\beta_i^2\lambda_{FR}^2 + \sigma_\epsilon^2}{n(1+\lambda^{FR})^2}\,, \tag{3.28}$$

$$MSE(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\sigma_\epsilon^2}{n}\,. \tag{3.29}$$

In the nature of trying to find a condition where Partial Ridge has lower MSE for all λ, the case is considered where $\lambda = \lambda^{FR} = \lambda_i$. Now evaluate the expression for the difference in MSE values,

$$MSE(\hat{\beta}_i(\lambda^{FR})) - MSE(\hat{\beta}_i(\lambda_i, S_i)) = \frac{n\beta_i^2\lambda^2 + \sigma_\epsilon^2 - \sigma_\epsilon^2(1+\lambda)^2}{n(1+\lambda)^2}\,. \tag{3.30}$$

From here, by looking closer at the numerator, one can show that in order for the above term to be positive (the Partial Ridge MSE to be lower), the following condition must be satisfied:

$$\frac{\beta_i^2}{\sigma_\epsilon^2} > \frac{\lambda+2}{n\lambda}\,, \tag{3.31}$$

where one can see that as λ and n increase, very quickly it becomes such that the ratio of β_i^2 and σ_ϵ^2 (a form of signal to noise ratio) does not have to be very large in order for Partial Ridge to dominate. While for a very small λ Full Ridge has an increased chance of having a lower MSE, the fact that Partial Ridge estimates each $\hat{\beta}_i(\lambda_i, S_i)$

individually and Full Ridge does not becomes a real selling point here. This because that, although for $\hat{\beta}_i$ Full Ridge can provide a lower MSE when using a very low penalty parameter, this parameter must then be used for all other predictors, so it will unlikely be desirable for Full Ridge to use such a small penalty parameter for β_i estimation, as this will unlikely be optimal overall when considering the other $p-1$ predictors. Partial Ridge, on the other hand, can afford to pick the best λ_i for β_i estimation as the rest of the coefficient estimates are not bound by this. Therefore, already one can see promising signs for Partial Ridge over Full Ridge in environments of little correlation amongst the candidate predictors.

Moving to a more general case, the expressions in 3.20 and 3.21 are considered for scenarios of varied correlation, dimension ($\frac{p}{n}$), SNR and relative coefficient magnitude ($\frac{\beta_i}{\sum_{j\neq i}^{p}\beta_j}$). By setting $n = 100$, $\sigma_\epsilon^2 = 1$ and varying everything else the minimum values of 3.26 and 3.27 are reported in the 3 tables below along with the corresponding penalty parameter for each approach in each scenario. With the sample size fixed, p is varied over $(100, 150, 200)$ with each dimension having the level of correlation varied over $\rho = (0.1, 0.5, 0.8)$ and for each ρ and p, the SNR and relative coefficients magnitudes are also modified. More specifically, defining the SNR here as $\frac{\beta_i+\sum_{j\neq i}^{p}\beta_j}{\sigma_\epsilon^2} = \beta_i + \sum_{j\neq i}^{p}\beta_j$, the SNR is varied over $(1, 4, 7)$ with each case having 3 separate structures of β_i and $\sum_{j\neq i}^{p}\beta_j$. One where β_i is large relative to $\sum_{j\neq i}^{p}\beta_j$, one β_i is small compared to $\sum_{j\neq i}^{p}\beta_j$ and one where they have equal magnitude.

TABLE 3.1: MSE values provided by PR and FR along with their optimal penalty parameter values when p=100

ρ	β profile	PR ($\times 10^{-2}$)	λ_i	FR ($\times 10^{-2}$)	λ^{FR}
0.1	$\beta_1 = 0.9, \sum_{j\neq 1}^{p} \beta_j = 0.1$	1.0353	720.354	1.0859	1.248
	$\beta_i = 0.1, \sum_{j\neq i}^{p} \beta_j = 0.9$	1.0897	119.123	0.4775	112.467
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.0726	316.142	1.0529	4.106
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.0726	316.142	1.0999	0.083
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.1001	8.141	1.0473	4.608
	$\beta_i = 2, \sum_{j\neq i}^{p} \beta_j = 2$	1.0984	23.755	1.0978	0.260
	$\beta_i = 6, \sum_{j\neq i}^{p} \beta_j = 1$	1.0917	97.422	1.1006	0.030
	$\beta_i = 1, \sum_{j\neq i}^{p} \beta_j = 6$	1.1007	1.656	1.0873	1.131
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.1001	8.141	1.0999	0.084
0.5	$\beta_i = 0.9, \sum_{j\neq i}^{p} \beta_j = 0.1$	1.2795	2980.22	1.9319	1.250
	$\beta_i = 0.1, \sum_{j\neq i}^{p} \beta_j = 0.9$	1.9349	121.99	0.5780	119.686
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.8475	381.428	1.8294	4.122
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.8475	381.428	1.9769	0.084
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.9771	8.085	1.8111	4.670
	$\beta_i = 2, \sum_{j\neq i}^{p} \beta_j = 2$	1.9707	24.364	1.9701	0.259
	$\beta_i = 6, \sum_{j\neq i}^{p} \beta_j = 1$	1.9432	98.900	1.9791	0.028
	$\beta_i = 1, \sum_{j\neq i}^{p} \beta_j = 6$	1.9791	2.764	1.9360	1.143
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.9771	8.085	1.9769	0.085
0.8	$\beta_i = 0.9, \sum_{j\neq i}^{p} \beta_j = 0.1$	1.6255	4402.42	4.6590	1.250
	$\beta_i = 0.1, \sum_{j\neq i}^{p} \beta_j = 0.9$	4.6704	121.863	0.6985	118.412
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	4.1680	394.855	4.1041	4.123
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	4.1680	394.855	4.9297	0.083
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	4.9303	8.081	4.0121	4.676
	$\beta_i = 2, \sum_{j\neq i}^{p} \beta_j = 2$	4.8901	24.752	4.8872	0.258
	$\beta_i = 6, \sum_{j\neq i}^{p} \beta_j = 1$	4.7205	98.859	4.9432	0.028
	$\beta_i = 1, \sum_{j\neq i}^{p} \beta_j = 6$	4.9434	2.709	4.6823	1.144
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	4.9303	8.081	4.9294	0.084

TABLE 3.2: MSE values provided by PR and FR along with their optimal penalty parameter values when p=150

ρ	β profile	PR ($\times 10^{-2}$)	λ_i	FR ($\times 10^{-2}$)	λ^{FR}
0.1	$\beta_1 = 0.9, \sum_{j\neq 1}^{p} \beta_j = 0.1$	1.0368	1036.660	1.0891	1.244
	$\beta_i = 0.1, \sum_{j\neq i}^{p} \beta_j = 0.9$	1.0923	178.346	0.4919	108.577
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.0743	466.743	1.0563	4.072
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.0743	466.743	1.1031	0.083
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.1033	12.335	1.0526	4.400
	$\beta_i = 2, \sum_{j\neq i}^{p} \beta_j = 2$	1.1015	37.269	1.1010	0.258
	$\beta_i = 6, \sum_{j\neq i}^{p} \beta_j = 1$	1.0943	146.046	1.1038	0.030
	$\beta_i = 1, \sum_{j\neq i}^{p} \beta_j = 6$	1.1039	1.687	1.0910	1.087
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.1033	12.335	1.1031	0.084
0.5	$\beta_i = 0.9, \sum_{j\neq i}^{p} \beta_j = 0.1$	1.3853	2625.960	1.9385	1.245
	$\beta_i = 0.1, \sum_{j\neq i}^{p} \beta_j = 0.9$	1.9409	183.516	0.6069	112.936
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.8524	466.743	1.8369	4.081
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.8524	572.763	1.9835	0.083
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.9836	12.164	1.8251	4.430
	$\beta_i = 2, \sum_{j\neq i}^{p} \beta_j = 2$	1.9771	36.557	1.9767	0.256
	$\beta_i = 6, \sum_{j\neq i}^{p} \beta_j = 1$	1.9493	148.800	1.9856	0.028
	$\beta_i = 1, \sum_{j\neq i}^{p} \beta_j = 6$	1.9857	4.165	1.9443	1.092
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.9836	12.164	1.9835	0.084
0.8	$\beta_i = 0.9, \sum_{j\neq i}^{p} \beta_j = 0.1$	1.7423	4611.590	4.6757	1.245
	$\beta_i = 0.1, \sum_{j\neq i}^{p} \beta_j = 0.9$	4.6848	183.054	0.7422	112.105
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	4.1787	593.676	4.1250	4.081
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	4.1787	593.676	4.9464	0.082
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	4.9468	12.152	4.0657	4.433
	$\beta_i = 2, \sum_{j\neq i}^{p} \beta_j = 2$	4.9062	37.251	4.9042	0.255
	$\beta_i = 6, \sum_{j\neq i}^{p} \beta_j = 1$	4.7352	148.607	4.9598	0.028
	$\beta_i = 1, \sum_{j\neq i}^{p} \beta_j = 6$	4.9599	4.142	4.7094	1.093
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	4.9468	12.152	4.9461	0.083

TABLE 3.3: MSE values provided by PR and FR along with their optimal penalty parameter values when p=200

ρ	β profile	PR ($\times 10^{-2}$)	λ_i	FR ($\times 10^{-2}$)	λ^{FR}
0.1	$\beta_1 = 0.9, \sum_{j\neq 1}^{p} \beta_j = 0.1$	1.0377	1345.270	1.0908	1.242
	$\beta_i = 0.1, \sum_{j\neq i}^{p} \beta_j = 0.9$	1.0936	237.266	0.4998	106.365
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.0753	615.217	1.0581	4.055
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.0753	615.217	1.1048	0.083
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.1056	1.783	1.0554	4.299
	$\beta_i = 2, \sum_{j\neq i}^{p} \beta_j = 2$	1.1031	49.769	1.1027	0.257
	$\beta_i = 6, \sum_{j\neq i}^{p} \beta_j = 1$	1.0957	194.480	1.1055	0.030
	$\beta_i = 1, \sum_{j\neq i}^{p} \beta_j = 6$	1.1056	1.639	1.0929	1.065
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.1049	16.806	1.1048	0.083
0.5	$\beta_i = 0.9, \sum_{j\neq i}^{p} \beta_j = 0.1$	1.3872	3483.690	1.9418	1.242
	$\beta_i = 0.1, \sum_{j\neq i}^{p} \beta_j = 0.9$	1.9439	245.007	0.6218	105.949
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.8548	763.828	1.8406	4.060
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	1.8548	763.828	1.9868	0.083
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.9869	16.259	1.8319	4.317
	$\beta_i = 2, \sum_{j\neq i}^{p} \beta_j = 2$	1.9803	49.767	1.9800	0.255
	$\beta_i = 6, \sum_{j\neq i}^{p} \beta_j = 1$	1.9523	198.675	1.9889	0.028
	$\beta_i = 1, \sum_{j\neq i}^{p} \beta_j = 6$	1.9890	5.572	1.9484	1.1068
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	1.9869	16.259	1.9868	0.084
0.8	$\beta_i = 0.9, \sum_{j\neq i}^{p} \beta_j = 0.1$	1.7460	6106.60	4.6841	1.242
	$\beta_i = 0.1, \sum_{j\neq i}^{p} \beta_j = 0.9$	4.6920	244.270	0.7645	108.942
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	4.1840	792.473	4.1354	4.061
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 0.5$	4.1840	792.473	4.9547	0.082
	$\beta_i = 0.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	4.9550	16.207	4.0917	4.318
	$\beta_i = 2, \sum_{j\neq i}^{p} \beta_j = 2$	4.9143	48.626	4.9127	0.254
	$\beta_i = 6, \sum_{j\neq i}^{p} \beta_j = 1$	4.7426	198.303	4.9681	0.028
	$\beta_i = 1, \sum_{j\neq i}^{p} \beta_j = 6$	4.9682	5.531	4.7228	1.068
	$\beta_i = 3.5, \sum_{j\neq i}^{p} \beta_j = 3.5$	4.9550	16.207	4.9545	0.083

Here it can be seen that, when β_i is large relative to $\sum_{j\neq i}^{p} \beta_j$, Partial Ridge enjoys a lower MSE but the opposite is true when the coefficient magnitude relativity is reversed with there being little difference in MSEs when the β_i and $\sum_{j\neq i}^{p} \beta_j$ are equal in size. This is likely down to how the bias term of Partial Ridge in 3.23 is not dependent on β_i unlike that of Full Ridge in 3.22 where a higher β_i can cause the bias to escalate quite rapidly due to how it is multiplied by p and λ^{FR} unlike $\sum_{j\neq i}^{p} \beta_j$. It is for a similar reason that Full Ridge performs better when β_i is relatively small. One can show that as $\sum_{j\neq i}^{p} \beta_j$ increases relative to β_i then the numerator of 3.22 becomes closer to that of

3.23. This combined with the discussion below 3.25 works in favour of Full Ridge as the denominator of 3.22 will be lower than that of 3.23 meaning that there will minimal or even no gains in bias for Partial Ridge over Full Ridge. Combine this with the variance discussion previously and it arises that there is no room for Partial Ridge to outperform Full Ridge which is why Full Ridge does better in this scenario.

One can also see that these gains for either method in either scenario are amplified as the level of correlation increases but become less apparent as the SNR increases. However, for determining which method provides the lower MSE, these features do not play a key role compared to that of the relative coefficient magnitude. It can be argued that the simple nature of this toy model setting ρ to be equal for all predictors eliminates the likely importance of the correlation between x_i and all other variables compared to the correlation amongst the other $p-1$ covariates themselves but this is left for future research.

Therefore, from this first stage of theoretical analysis, one can see that when concerned with the estimation of a single coefficient, Partial Ridge can bring benefit if the true value of the given coefficient is large relative to the sum of all other true β values. Although, by definition, it is not possible for this to be the case for all predictors under a given DGP, in situations where there is a strong degree of sparsity or an even balance of positive and negative true coefficients with similar magnitudes, such as product demand function applications, it can be such that for many of the β_j's, $\sum_{h\neq j}^{p}\beta_h << \beta_j$ making Partial Ridge a worthy competitor of Full Ridge.

3.3.4 Considering All Predictors

As mentioned at the end of the previous section, while Partial Ridge can see gains in bias covert into MSE gains for single predictors that have a true magnitude large relative to the sum of all other β components, this can not be the case for all predictors. Moreover, while the estimation of the β_i values is important, more often the end goal is to maximise the prediction accuracy making $X\hat{\beta}$ a more relevant method of assessment. Specifically, the excess prediction risk (EPR) defined in Bach (2022) is used in this paper and is formally defined as follows:

$$E[\mathbb{R}(\hat{\beta})] - \mathbb{R}^* = E||\hat{\beta}-\beta||^2_{\hat{\Sigma}} = E\left[(\hat{\beta}-\beta)'\frac{X'X}{n}(\hat{\beta}-\beta)\right], \tag{3.32}$$

where R^* is the predictive risk of the oracle model defined as $R^* = \frac{1}{n}E||y-X\beta||^2 = \sigma^2_\epsilon$, $\hat{\Sigma}$ represents the predictor covariance matrix, $\hat{\beta}$ is the estimated coefficients of the chosen estimation method and β represents the coefficients under the true DGP. One can also apply the bias-variance decomposition

to 3.32 to obtain the following:

$$E[\mathbb{R}(\hat{\beta})] - \mathbb{R}^* = ||E(\hat{\beta}) - \beta||^2_{\hat{\Sigma}} + E||\hat{\beta} - E(\hat{\beta})||^2_{\hat{\Sigma}}\,. \tag{3.33}$$

With this formal definition of EPR, this sections seeks to evaluate the above expressions under the toy model from the previous section for both Full Ridge and Partial Ridge.

Proposition 3.4: The expression for the EPR of both Full Ridge and Partial Ridge are respectively given as follows:

$$\sum_{k=1}^{p}\Bigg(MSE(\hat{\beta}_k(\lambda^{FR})) + \rho\sum_{i\neq k}^{p} Bias(\hat{\beta}_k(\lambda^{FR}))Bias(\hat{\beta}_i(\lambda^{FR})) + Cov\big(\bar{\beta}_k(\lambda^{FR}), \bar{\beta}_i(\lambda^{FR})\big)\Bigg)\,, \tag{3.34}$$

$$\sum_{k=1}^{p}\Bigg(MSE(\hat{\beta}_k(\lambda_k, S_k)) + \rho\sum_{i\neq k}^{p} Bias(\hat{\beta}_k(\lambda_k, S_k))Bias(\hat{\beta}_i(\lambda_i, S_i)) + Cov\big(\bar{\beta}_k(\lambda_k, S_k), \bar{\beta}_i(\lambda_i, S_i)\big)\Bigg)\,. \tag{3.35}$$

Where $\bar{\beta}_k(\lambda^{FR})$ and $\bar{\beta}_k(\lambda_k, S_k)$ are the individual $\hat{\beta}_k$ estimates centred around their mean ($\hat{\beta}_k - E[\hat{\beta}_k]$). The discussion from the previous section showed how the MSE component can only be dominated by Partial Ridge when β_k is larger relative to $\sum_{j\neq k}^{p}\beta_j$, but this can not be the case for all predictors. While there is little difference between the 2 procedures when $\beta_i \approx \sum_{j\neq i}^{p}\beta_j$ leaving a very small number of scenarios where Partial Ridge can beat Full Ridge over the sum of individual MSEs. However, one can see from the expressions above that for a sufficiently large $\rho > 0$ the excess prediction risk can be improved when the cross product of bias terms as well as centred $\hat{\beta}_k$ estimates are negatively correlated leading to a new branch of investigation.

Similar to the previous section, expressions are 3.34 and 3.35 are evaluated under various DGP designs to see where Partial Ridge may outperform Full Ridge with the optimal λ of Partial Ridge chosen by minimising each individual MSE whereas the single λ for all the Full Ridge coefficients is chosen by running the sum of individual MSEs over a grid of penalty parameters and using the one that gives the smallest sum of MSEs.

Using a DGP based on that of 3.1, expressions 3.34 and 3.35 are evaluated under a fixed design with varied ρ and p with the sample size fixed at $n = 100$ and $\sigma_\epsilon^2 = 10$. These are varied in 3 separate designs for the structure of β_i values which are detailed as follows.

Design 1: Sparsity with variation in sign and magnitude amongst the active predictors
In this setting, the true β vector is defined as follows: $\beta_i = i(-1)^{i+1}$ for $i = 1, \ldots, 7$ and $\beta_i = 0$ for $i > 7$. This scenario would be expected to favour Full Ridge as one can see that the majority of coefficients are such that $\beta_i < \sum_{j \neq i}^{p} \beta_j$. However, one can also see that this results in many of the active coefficients being such that β_i and $\sum_{j \neq i}^{p} \beta_j$ are opposite signs which can inflate the bias of Full Ridge. In addition the varied in coefficient magnitude makes the adjustable λ_i property of Partial Ridge a key advantage over Full Ridge.

Design 2: Sparsity with active predictors summing to 0
This is similar to the previous design only here the active β_i values are as follows. $\beta_1 = 10$ and $\beta_i = -1$ for $i = 2, \ldots, 11$ with all other β_i values being equal to 0. On closer inspection one can see that for all β_i it is such that $|\beta_i| = |\sum_{j \neq i}^{p} \beta_j|$ which Tables 3.1-3.3 show that this results in almost identical MSEs for Full and Partial Ridge. However, unlike in the previous setting in Section 3.2, Full Ridge is committed to a single λ^{FR} value for all predictors leaving the possibility for Partial Ridge to dominate.

Design 3: Equal magnitude with alternation sign
Here it is such that $\beta_i = (-1)^{i+1}$ leading to every coefficient being equal to either 1 or -1. An interesting feature here is that when p is an even number then $\sum_{j \neq i}^{p} \beta_j$ is equal to 1 or -1 achieving a similar situation as Design 2. However when p is an odd number then when $\beta_i = 1$, $\sum_{j \neq i}^{p} \beta_j = 0$ and when $\beta_i = -1$, $\sum_{j \neq i}^{p} \beta_j = 2$ so there is an almost even mix of when $|\beta_i| > |\sum_{j \neq i}^{p} \beta_j|$ and $|\beta_i| < |\sum_{j \neq i}^{p} \beta_j|$.

For each setting, ρ is varied over $(0.1, 0.4, 0.7)$ with p taking a value of 100 or 200 except for Design 3 where $p \in (101, 201)$ is considered in order to achieve the desired ratios of $\frac{\beta_i}{\sum_{j \neq i}^{p} \beta_j}$. The tables below show the values of 3.34 and 3.35 under each of these designs as well as the sum of individual coefficient MSEs.

TABLE 3.4: EPR for Full and Partial Ridge across each of the designs

	p=100			p=200		
Design 1	$\rho = 0.1$	$\rho = 0.4$	$\rho = 0.7$	$\rho = 0.1$	$\rho = 0.4$	$\rho = 0.7$
Full Ridge	9.277	8.955	8.110	17.284	16.187	13.600
Partial Ridge	10.607	15.879	50.715	22.625	45.081	193.597
Design 2						
Full Ridge	9.098	8.708	7.715	16.668	15.390	12.514
Partial Ridge	17.404	98.570	299.451	46.902	291.052	928.765
Design 3 (p+1)						
Full Ridge	9.107	8.683	7.618	18.108	17.255	15.119
Partial Ridge	13.855	49.202	159.194	33.354	138.680	501.598

TABLE 3.5: Sum of individual MSEs for Full and Partial Ridge across each of the designs

	p=100			p=200		
Design 1	$\rho = 0.1$	$\rho = 0.4$	$\rho = 0.7$	$\rho = 0.1$	$\rho = 0.4$	$\rho = 0.7$
Full Ridge	10.207	14.761	26.701	19.099	26.814	45.003
Partial Ridge	10.941	16.308	32.216	21.976	32.804	64.873
Design 2						
Full Ridge	10.009	14.350	25.386	18.415	25.4845	41.381
Partial Ridge	10.067	10.749	12.415	20.112	21.854	25.645
Design 3 (p+1)						
Full Ridge	10.019	14.308	25.063	20.015	28.593	50.063
Partial Ridge	10.507	13.181	20.756	20.964	26.781	42.886

Tables 3.4 and 3.5 show show interesting results when evaluating expressions 3.34 and 3.35. While Table 3.5 shows that Partial Ridge dominates in the sum of individual MSEs for most settings under Designs 2 and 3 (when the correlation is greater than 0.1), the opposite is true when considering the excess risk prediction. The reason for dominance in the sum of MSEs was justified before but looking at the relativity of β_i to $\sum_{j\neq i}^{p} \beta_j$ for each coefficient as well as the fact that Partial Ridge has greater flexibility with the λ_i values, the results appear to not reflect this. This leads one to reconsider the importance of the second parts of 3.34 and 3.35, the bias cross products and covariances between the centred β_i estimates. Therefore 3.34 and 3.35 will be analysed further to seek answers with the first aim being to define the second part of these 2 equations as follows:

$$EPR_{FR}(2) = \sum_{k=1}^{p} \rho \sum_{i\neq k}^{p} Bias(\hat{\beta}_k(\lambda^{FR}))Bias(\hat{\beta}_i(\lambda^{FR})) + Cov(\bar{\beta}_k(\lambda^{FR}), \bar{\beta}_i(\lambda^{FR}))\,,$$

$$EPR_{PR}(2) = \sum_{k=1}^{p}\rho \sum_{i\neq k}^{p} Bias(\hat{\beta}_k(\lambda_k, S_k))Bias(\hat{\beta}_i(\lambda_i, S_i)) + Cov\big(\bar{\beta}_k(\lambda_k, S_k), \bar{\beta}_i(\lambda_i, S_i)\big)\,.$$

Using expressions 3.20-3.25 from the previous subsection, the above can be written as follows:

$$EPR_{FR}(2) = \sum_{k=1}^{p}\rho \sum_{i\neq k}^{p} \Big(\frac{(\rho\lambda^{FR}\sum_{j\neq i}^{p}\beta_j - \beta_i\lambda^{FR}(1+\lambda^{FR}+(p-2)\rho))(\rho\lambda^{FR}\sum_{j\neq k}^{p}\beta_j - \beta_k\lambda^{FR}(1+\lambda^{FR}+(p-2)\rho))}{(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+(p-1)\rho)^2} + \frac{\sigma_\epsilon^2\rho(\lambda_{FR}^2+(\rho-1)(1+(p-1)\rho))}{n(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+(p-1)\rho)^2}\Big)\,, \quad (3.36)$$

$$EPR_{PR}(2) = \sum_{k=1}^{p}\rho \sum_{i\neq k}^{p} \Big(\frac{(\rho\lambda_i\sum_{j\neq i}^{p}\beta_j)(\rho\lambda_k\sum_{j\neq k}^{p}\beta_j)}{(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)(1+\lambda_k+(p-2)\rho-(p-1)\rho^2)} + \frac{\sigma_\epsilon^2\rho(\lambda_i\lambda_k+(\rho-1)(1+(p-1)\rho))}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)(1+\lambda_k+(p-2)\rho-(p-1)\rho^2)}\Big)\,. \quad (3.37)$$

By looking at the difference between the corresponding entries in Tables 3.4 and 3.5 one can see that expression 3.36 for Full Ridge is always negative making the EPR lower than the sum of MSEs whereas for Partial Ridge expression 3.37 is always positive and becomes quite large in certain scenarios, especially when the correlation is high making 3.36 and 3.37 a more significant component of the EPR values.

Firstly, it is important to notice that, with Partial Ridge, the λ_is were obtained by minimising the individual MSEs but where $\sum_{j\neq i}^{p}\beta_j$ is small then so is the bias. This combined with how the variance in 3.25 decreases as λ_i increases leads to the optimal λ_i being very large. While this is good for the individual MSEs, expression 3.37 would suggest otherwise for the EPR. By looking closer at 3.36 and 3.37, expanding them and sorting by powers of λ^{FR}, λ_i and λ_k leads to the following expressions for them:

$$EPR_{FR}(2) = \sum_{k=1}^{p}\rho \sum_{i\neq k}^{p} \frac{(\beta_i\beta_k)\lambda_{FR}^4 + O(\lambda_{FR}^3)}{\lambda_{FR}^4 + O(\lambda_{FR}^3)}\,, \quad (3.38)$$

$$EPR_{PR}(2) = \sum_{k=1}^{p}\rho \sum_{i\neq k}^{p} \frac{\Big(\rho^2(\sum_{j\neq i}^{p}\beta_j)(\sum_{j\neq k}^{p}\beta_j) + \frac{\sigma_\epsilon^2\rho}{n}\Big)\lambda_i\lambda_k + O(1)}{\lambda_i\lambda_k + O(\lambda_i) + O(\lambda_k)}\,. \quad (3.39)$$

So one can see that choosing a single λ_i as being very large for a given β_i can have consequences for the EPR through the bias cross products and centred estimator covariances when the numerator coefficient of $\lambda_i\lambda_k$ is greater than 1. More importantly though, one can see that the sign and magnitude of the 2 expressions

above are influenced by different features of the true DGP. For Full Ridge the coefficient of the highest order of λ^{FR} in the numerator has its sign and magnitude determined by the cross product of individual β_i values. Whereas for Partial Ridge this is determined by the cross products of the $\sum_{j\neq i}^{p}\beta_j$ values. Such a feature is significant when explaining the results above as the aggregated nature of the $\sum_{j\neq i}\beta_j$ values means that not only with they likely be larger in magnitude than that of the β_i cross products but will also likely be the same sign for the majority of the predictors compared to β_is themselves which will have much more variability in sign. For example, in Design 3 the β_i and $\sum_{j\neq i}^{p}\beta_j$ cross products are given in the tables below, remembering that p is an odd number.

TABLE 3.6: Cross products of β_i values under Design 3

i	1	2	3	4	...	p
β_i	1	-1	1	-1	...	1
$\beta_i\beta_1$		-1	1	-1	...	1
$\beta_i\beta_2$	-1		-1	1	...	-1
$\beta_i\beta_3$	1	-1		-1	...	1
$\beta_i\beta_4$	-1	1	-1		...	-1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\beta_i\beta_p$	1	-1	1	-1	...	
$\sum_{k\neq i}\beta_i\beta_k$	0	-2	0	-2	...	0

TABLE 3.7: Cross products of $\sum_{j\neq i}^{p}\beta_j$ values under Design 3

i	1	2	3	4	...	p
β_i	1	-1	1	-1	...	1
$\sum_{j\neq i}^{p}\beta_j$	0	2	0	2	...	0
$(\sum_{j\neq i}^{p}\beta_j)(\sum_{j\neq 1}^{p}\beta_j)$		0	0	0	...	0
$(\sum_{j\neq i}^{p}\beta_j)(\sum_{j\neq 2}^{p}\beta_j)$	0		0	4	...	0
$(\sum_{j\neq i}^{p}\beta_j)(\sum_{j\neq 3}^{p}\beta_j)$	0	0		0	...	0
$(\sum_{j\neq i}^{p}\beta_j)(\sum_{j\neq 4}^{p}\beta_j)$	0	4	0		...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(\sum_{j\neq i}^{p}\beta_j)(\sum_{j\neq p}^{p}\beta_j)$	0	0	0	0	...	
$\sum_{k\neq i}\beta_i\beta_k$	0	$4\times\frac{p-3}{2}$	0	$4\times\frac{p-3}{2}$	...	0

So in this example, one can see that the sum of the β_i cross products is negative and small in value compared to the $\sum_{j\neq i}^{p}\beta_j$ cross products which are positive and larger in size. This explains why the results in Table 3.5 were promising for Partial Ridge but, when considering the EPR expression in its entirety, Table 3.4 showed Partial Ridge suffering, because the expression in 3.37 was positive and large in value compared to

3.36 being negative reducing the EPR of Full Ridge from its sum of MSEs.

Therefore, it is clear to see that while Partial Ridge has some nice properties when considering some individual parameters, its individualized nature means that when considering whole sets of predictors, it is highly likely to be inferior to Full Ridge when it comes to predictive accuracy. However, this does not mean that Partial Ridge should be discarded completely but, instead, raises the possibility of combining the 2 estimation methods to reap both of their advantages.

3.3.5 Hybrid Estimation Procedure (HEP)

As discussed, Partial Ridge does extremely well at estimating coefficients that have a true β_i value large relative to the sum of all other coefficients, $\sum_{j\neq i}^{p}\beta_j$. Therefore, intuitively, one might wish to estimate coefficients that are believed to have a high value of $\frac{\beta_i}{\sum_{j\neq i}^{p}\beta_j}$ with Partial Ridge and use Full Ridge for the rest. To formally define this hybrid approach, first let the coefficient matrix, X, be ordered such that the first s columns from the left represent the covariates whose coefficients are to be estimated via Partial Ridge with the remaining $p-s$ columns representing covariates having their coefficient estimated via Full Ridge. It is important to note that when estimating the β_i's for these $p-s$ covariates, Full Ridge is run on the entire predictor matrix and not just the $p-s$ variables described. This allows one to formalise the HEP as follows:

$$\hat{\beta}_{Hy} = \begin{bmatrix} \hat{\beta}_1(\lambda_1, S_1) \\ \vdots \\ \hat{\beta}_s(\lambda_s, S_s) \\ \hat{\beta}_{s+1}(\lambda^{FR}) \\ \vdots \\ \hat{\beta}_p(\lambda^{FR}) \end{bmatrix}. \tag{3.40}$$

So one can see that this is simply a mixture of the estimators defined in Proposition 3.2 meaning that much of the theoretical analysis before is relevant for this modified estimator. How one chooses which coefficients to estimate with Partial Ridge and Full Ridge is very important for the reasons discussed previously and there are multiple approaches that can be utilised. Firstly, one could run an initial Lasso regression on the full predictor set and using the model selection element to distinguish between predictors that are likely to have relatively large slopes and ones that will have small ones. Therefore, one possible rule could be to estimate all coefficients that are non-zero under Lasso by Partial Ridge and the rest with Full Ridge. Alternatively, one could use an initial stage where Full Ridge is applied to the data set and the estimated coefficients are sorted in order of magnitude. From here, one can choose the largest

w% of the coefficients and have these slopes estimated by Partial Ridge with the rest Full Ridge. Here, w can be flexible and allow the use of a priori knowledge, for example, if one believed that the original data set where sparse in nature then w would be relatively small. More on this topic is discussed when constructing the simulation experiment in Section 3.4 but a more formalised approach to this feature is left to future work.

To understand the properties of this estimation approach, one can repeat the procedure deriving the excess risk predictions shown in Proposition 3.4 to obtain a similar expression for the HEP. This is formally defined in Proposition 3.5 below followed with deeper analysis into the behaviour of the EPR under various DGP conditions.

Proposition 3.5: The excess risk prediction for the HEP estimator is given by the following expression:

$$EPR(\hat{\beta}_{Hy}) = \theta_1 + \theta_2\,, \tag{3.41}$$

where θ_1 and θ_2 are defined as follows.

$$\begin{aligned}\theta_1 = &\sum_{k=1}^{s} MSE(\hat{\beta}_k(\lambda_k, S_k)) \\ &+ \rho\Big[\sum_{i=1, i\neq k}^{s} Bias(\hat{\beta}_i(\lambda_i, S_i))Bias(\hat{\beta}_k(\lambda_k, S_k)) + Cov\big(\bar{\beta}_i(\lambda_i, S_i), \bar{\beta}_k(\lambda_k, S_k)\big)\big) \\ &\quad + \sum_{i=s+1}^{p} Bias(\hat{\beta}_i(\lambda^{FR}))Bias(\hat{\beta}_k(\lambda_k, S_k)) + Cov\big(\bar{\beta}_i(\lambda^{FR}), \bar{\beta}_k(\lambda_k, S_k)\big)\Big]\,,\end{aligned} \tag{3.42}$$

$$\begin{aligned}\theta_2 = &\sum_{k=s+1}^{p} MSE(\hat{\beta}_k(\lambda^{FR})) \\ &+ \rho\Big[\sum_{i=1}^{s} Bias(\hat{\beta}_i(\lambda_i, S_i))Bias(\hat{\beta}_k(\lambda^{FR})) + Cov\big(\bar{\beta}_i(\lambda_i, S_i), \bar{\beta}_k(\lambda^{FR})\big)\big) \\ &\quad + \sum_{i=s+1, i\neq k}^{p} Bias(\hat{\beta}_i(\lambda^{FR}))Bias(\hat{\beta}_k(\lambda^{FR})) + Cov\big(\bar{\beta}_i(\lambda^{FR}), \bar{\beta}_k(\lambda^{FR})\big)\Big]\,.\end{aligned} \tag{3.43}$$

Once again, the $\bar{\beta}_i$ values are the estimators themselves with their mean subtracted, for example, $\bar{\beta}_k(\lambda^{FR}) = \hat{\beta}_k(\lambda^{FR}) - E[\hat{\beta}_k(\lambda^{FR})]$. One can see that the MSEs of the individual coefficients themselves are present again but this time there are bias cross products and covariances of centred estimators combining Full and Partial Ridge. To understand this expression further, one must combine 3.41-3.43 with the toy model expressions from 3.20-3.25 as was done in section 3.3.3.

Since, the individual coefficient MSEs have already been considered, the analysis here focuses on the bias cross products and covariances with the square bracket components of 3.42 and 3.43 investigated. These square bracket terms from 3.42 and 3.43 are defined as $\theta_1(2)$ and $\theta_2(2)$ respectively and can be written as follows:

$$\theta_1(2) = \sum_{k=1}^{s}\rho\Big[\sum_{i=1,i\neq k}^{s}\frac{(\rho\lambda_i\sum_{j\neq i}^{p}\beta_j)(\rho\lambda_k\sum_{j\neq k}^{p}\beta_j)+\frac{\sigma_\epsilon^2}{n}\rho(\lambda_i\lambda_k+(\rho-1)(1+(p-1)\rho))}{(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)(1+\lambda_k+(p-2)\rho-(p-1)\rho^2)}$$
$$+\sum_{i=s+1}^{p}\frac{(\rho\lambda^{FR}\sum_{j\neq i}^{p}\beta_j-\beta_i\lambda^{FR}(1+\lambda^{FR}+(p-2)\rho))(\rho\lambda_k\sum_{j\neq k}^{p}\beta_j)}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)(1+\lambda_k+(p-2)\rho-(p-1)\rho^2)}$$
$$+\frac{\frac{\sigma_\epsilon^2}{n}\rho(\lambda_k\lambda^{FR}+(\rho-1)(1+(p-1)\rho))}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)(1+\lambda_k+(p-2)\rho-(p-1)\rho^2)}\Big], \quad (3.44)$$

$$\theta_2(2) = \sum_{k=s+1}^{p}\rho\Big[\sum_{i=1}^{s}\frac{(\rho\lambda_i\sum_{j\neq i}^{p}\beta_j)(\rho\lambda^{FR}\sum_{j\neq k}^{p}\beta_j-\beta_k\lambda^{FR}(1+\lambda^{FR}+(p-2)\rho))}{(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}$$
$$+\frac{\frac{\sigma_\epsilon^2}{n}\rho(\lambda_i\lambda^{FR}+(\rho-1)(1+(p-1)\rho))}{(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}+$$
$$\sum_{\substack{i=s+1\\ i\neq k}}^{p}\frac{(\rho\lambda^{FR}\sum_{j\neq i}^{p}\beta_j-\beta_i\lambda^{FR}(1+\lambda^{FR}+(p-2)\rho))(\rho\lambda^{FR}\sum_{j\neq k}^{p}\beta_j-\beta_k\lambda^{FR}(1+\lambda^{FR}+(p-2)\rho))}{(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+(p-1)\rho)^2}$$
$$+\frac{\frac{\sigma_\epsilon^2}{n}\rho(\lambda_{FR}^2+(\rho-1)(1+(p-1)\rho))}{(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+(p-1)\rho)^2}\Big]. \quad (3.45)$$

From here one can conduct similar analysis to that of the previous section whereby 3.42 and 3.43 are expanded and arranged into powers of λ^{FR} and λ_k. The first line of 3.42 and the last 2 lines of 3.43 were studied before in 3.38 and 3.39 with the conclusion that the magnitude of Full Ridge bias cross products is predominantly determined by the individual β_i values with Partial Ridge being more influenced by the $\sum_{j\neq i}\beta_j$ cross products. This results in almost identical expressions to that of 3.38 and 3.39 only with different summations, with the first expression below corresponding to the first line of 3.44 and the second expression representing the last 2 lines of 3.45,

$$\sum_{k=1}^{s}\rho\sum_{i=1,i\neq k}^{s}\frac{\Big(\rho^2(\sum_{j\neq i}^{p}\beta_j)(\sum_{j\neq k}^{p}\beta_j)+\frac{\sigma_\epsilon^2\rho}{n}\Big)\lambda_i\lambda_k+O(1)}{\lambda_i\lambda_k+O(\lambda_i)+O(\lambda_k)}, \quad (3.46)$$

$$\sum_{k=s+1}^{p}\rho\sum_{i=s+1,i\neq k}^{p}\frac{(\beta_i\beta_k)\lambda_{FR}^4+O(\lambda_{FR}^3)}{\lambda_{FR}^4+O(\lambda_{FR}^3)}. \quad (3.47)$$

This reveals a great degree of flexibility to the new HEP as 3.46 and 3.47 show that each coefficient estimate will influence the EPR through either the $\beta_i\beta_k$ or $(\sum_{j\neq i}^{p}\beta_j)(\sum_{j\neq k}^{p}\beta_j)$ cross products depending on whether they were estimated by Full or Partial Ridge. Based on what was discussed previously, it can be argued that gains

in the EPR for the HEP over Full Ridge alone can be found when carefully allocating each coefficient estimate to Full Ridge when the sum of the $\beta_i\beta_k$ cross products is small or even a negative value with large magnitude. Likewise, coefficients that have a small or large negative sum of the $\sum_{j\neq i}^{s}\beta_j$ cross products should be estimated with Partial Ridge. To make this clearer, consider the β profiles of the designs from section 3.3.3 with the β_is and their corresponding $\sum_{j\neq i}^{p}\beta_j$ values detailed below.

Design 1

β_i	$\sum_{j\neq i}^{p}\beta_j$
1	-3
-2	6
3	1
-4	8
5	-1
-6	10
7	-3
0	4
⋮	⋮
0	4
0	4

Design 2

β_i	$\sum_{j\neq i}^{p}\beta_j$
10	-10
-1	1
-1	1
-1	1
-1	1
⋮	⋮
-1	1
0	0
⋮	⋮
0	0
0	0

Design 3

β_i	$\sum_{j\neq i}^{p}\beta_j$
1	0
-1	2
1	0
-1	2
1	0
-1	2
1	0
-1	2
⋮	⋮
-1	2
1	0

Using what was mentioned before, one can look at the the β profile for each design and pick variables to be estimated by Full and Partial Ridge in an attempt to provide a lower EPR for the HEP than that of Full Ridge alone. Firstly, for Design 1, one can see that for the third, fifth and seventh coefficients it is such that $\beta_i > \sum_{j\neq i}^{p}\beta_j$ which will lead to the individual coefficients having lower MSEs under Partial Ridge that Full Ridge. However, this will also be more beneficial for 3.46 than for 3.47 since it can be shown that the sum of β_i cross products for these 3 coefficients will be greater than the sum of cross products for their corresponding $\sum_{j\neq i}^{p}\beta_j$ values. In addition, for these β_i slopes the $\sum_{j\neq i}^{p}\beta_j$ values are 1, -1 and -3 giving a mix of positive negative values contributing to a low sum of cross products (equating to -1). The other coefficients see the opposite where $\beta_i < \sum_{j\neq i}^{p}\beta_j$ meaning that not only will the sum of $MSE(\hat{\beta}_i(\lambda^{FR}))$ values be lower under Full Ridge but the expression in 3.46 will likely be greater than that of 3.47 making Full Ridge more appropriate. Therefore, using the HEP under Design 1, it would be most appropriate to use Partial Ridge for the third, fifth and seventh coefficients and Full Ridge for the rest of them.

For Design 2, the decision process is more unclear as it is such that $|\beta_i| = |\sum_{j\neq i}^{p}\beta_j|$ meaning there is very little difference in the MSEs of the Full and Partial Ridge estimates. In addition, when looking closer at what the sum of the β_i and $\sum_{j\neq i}^{p}\beta_j$ cross

products, it is tricky to allocate coefficients to Partial and Full Ridge that will dominate any other allocation through keeping the expressions in 3.46 and 3.47 as low as possible.

Finally, for Design 3 the concept is similar to that of Design 1; $\beta_i > \sum_{j\neq i}^{p} \beta_j$ when $\beta_i = 1$ and $\beta_i < \sum_{j\neq i}^{p} \beta_j$ when $\beta_i = -1$. Therefore, from and individual MSE perspective it would make sense to use Partial Ridge to estimate coefficients where $\beta_i = 1$ and Full Ridge when $\beta_i = -1$. This also works well for expression 3.39 where all $\sum_{j\neq i}^{p} \beta_j$ cross products will be 0 for the coefficients estimated by Partial Ridge.

To test this more decisively, the experiment run at the end of the previous subsection is repeated over Full Ridge and HEP. The DGP is exactly the same for each design with $\sigma_\epsilon^2 = 10$, $n = 100$ and p and ρ varied over the same set of values for each of the 3 designs. Once again, for Full Ridge, the optimal λ^{FR} is obtained by running the sum of individual coefficient MSEs over a grid a choosing the λ^{FR} that provides the lowest value. Also, for Partial Ridge, when estimating a single β_i value, the optimal penalty parameter is obtained by running the individual $MSE(\hat{\beta}_i(\lambda_i))$ over a grid of λ_i and choosing the one that provides the lowest MSE. Since the HEP estimator is simply a combination of Full and Partial Ridge estimates, this covers how all the penalty parameters are obtained for the new estimation procedure. Finally, as discussed previously, for Design 1, the third, fifth and seventh coefficients are estimated by Partial Ridge with rest by Full Ridge. For Design 2 all coefficients that have a non-zero true coefficient are estimated with Partial Ridge and the rest Full Ridge. In Design 3 all $\beta_i = 1$ are estimated with Partial Ridge and all $\beta_i = -1$ are estimated with Full Ridge. The tables below shows the EPR values as well as the sum of individual coefficient MSEs of Full Ridge and the HEP for each experiment.

TABLE 3.8: EPR for Full Ridge and the HEP across each of the designs

	p=100			p=200		
Design 1	$\rho = 0.1$	$\rho = 0.4$	$\rho = 0.7$	$\rho = 0.1$	$\rho = 0.4$	$\rho = 0.7$
Full Ridge	9.277	8.955	8.110	17.284	16.187	13.600
HEP	8.723	8.273	7.418	15.284	13.717	11.195
Design 2						
Full Ridge	9.098	8.708	7.715	16.668	15.390	12.514
HEP	8.570	8.614	10.523	14.353	13.111	12.776
Design 3 (p+1)						
Full Ridge	9.107	8.683	7.618	18.108	17.255	15.119
HEP	14.435	57.088	206.339	36.727	175.634	707.526

TABLE 3.9: Sum of individual MSEs for Full Ridge and the HEP across each of the designs, that is $\sum_{k=1}^{s} MSE(\hat{\beta}_k(\lambda_k)) + \sum_{k=s+1}^{p} MSE(\hat{\beta}_k(\lambda^{FR}))$

	p=100			p=200		
Design 1	$\rho = 0.1$	$\rho = 0.4$	$\rho = 0.7$	$\rho = 0.1$	$\rho = 0.4$	$\rho = 0.7$
Full Ridge	10.207	14.761	26.701	19.099	26.814	45.003
HEP	9.556	13.384	22.209	16.745	22.160	31.910
Design 2						
Full Ridge	10.009	14.350	25.386	18.415	25.4845	41.381
HEP	9.259	12.733	20.274	15.677	20.179	27.391
Design 3 (p+1)						
Full Ridge	10.019	14.308	25.063	20.015	28.593	50.063
HEP	10.073	12.332	17.930	20.094	25.057	37.137

Firstly, under Design 1 the Hybrid model dominates Full Ridge for all experiments as expected. This is due to the combination of running Partial Ridge on the coefficients where $\beta_i > \sum_{j\neq i}^{p} \beta_j$ to obtain the optimal individual coefficient MSEs. This along with the mixture of signs in the β_is and $\sum_{j\neq i}^{p} \beta_j$ values varying allowed the cross products signs to be negative in many cases. This contributed to expressions 3.44 and 3.45 being negative to reduce the EPR from the sum of individual coefficient MSEs as seen by looking at the entries in Table 3.9 corresponding to those of Table 3.8 for Design 1.

For Design 2, the best approach is more unclear, Table 3.8 shows that the HEP does well when the correlation is low and the the sparsity increases (p increasing since the additional variables have true coefficients of 0). While under higher correlation Full Ridge has a lower EPR, Table 3.9 shows that the sum of individual MSEs is lower for the HEP in all experiments. This can be explained by looking closer at expressions 3.46 and 3.47 along with the β profile for Design 2. Since it was decided that the 11 active predictors would be estimated by Partial Ridge, consider the main component of 3.46 below,

$$\sum_{k=1}^{11} \rho \sum_{i=1, i\neq k}^{11} \rho^2 (\sum_{j\neq i}^{p} \beta_j)(\sum_{j\neq k}^{p} \beta_j) \lambda_i^* \lambda_k^* = -55\rho^3 \bar{\lambda}\,, \tag{3.48}$$

where, here it is such that $\lambda_i \lambda_k$ is equal to some arbitrary constant ($\bar{\lambda}$) across all i and k for simplicity. Now consider the Full Ridge component of the HEP given in 3.47 with the main influential part given by the following:

$$\sum_{k=12}^{p} \rho \sum_{i=12, i\neq k}^{p} (\beta_i \beta_k) \lambda_{FR}^4\,. \tag{3.49}$$

Since all $\beta_i = 0$ for $i = 12, \ldots, p$ it is such that the above expression (as well as 3.47 will be equal to 0 under this design, therefore, for comparison purposes, the first

expression equating to $-55\rho^3\bar{\lambda}$ is most relevant here.

To compare this to the corresponding component of the Full Ridge EPR consider the main component of 3.38 detailed below,

$$\sum_{k=1}^{p}\rho \sum_{i=1,i\neq k}^{p}(\beta_i\beta_k)\lambda_{FR}^4 = -55\rho\lambda_{FR}^4 . \tag{3.50}$$

This is easily comparable to the HEP equivalent expression of $-55\rho^3\bar{\lambda}$ with the main difference lying in the power of ρ. This reveals how the correlation plays a key role in helping Full Ridge to dominate the HEP under higher correlation. Firstly, for $\rho > 0$ expressions 3.48 and 3.50 will always be negative explaining why the EPR values in Table 3.8 are smaller than the sum of MSEs in Table 3.9. When the correlation is small then expressions 3.48 and 3.50 are also small as a result making them less influential on the EPR as a whole which explains why the HEP dominates from having a lower sum of individual coefficient MSEs. However, as the correlation increases, the fact that $\rho^3 < \rho$ for $0 < \rho < 1$ means that the Full Ridge estimator EPR component in 3.50 will be a larger negative than that of 3.48. In more simple terms it is such that $-55\rho < -55\rho^3$ making 3.50 have greater relative power in keeping the EPR low in relation to the individual coefficient MSE sum. Therefore, while the HEP has desirable features one must be careful in certain situations where Full Ridge can use the high level of correlatedness to its advantage.

Finally, for Design 3 the results are surprising as despite the HEP dominating when concerned with the individual coefficient MSEs, the EPR suffers greatly compared to Full Ridge for all experiments with this worsened as ρ and p increase. As was the case with the previous 2 designs, one needs to look closer at the other components of the EPR starting with the Full Ridge estimator where 3.36 and 3.38 represent the secondary component of the EPR. Where one is running Full Ridge on all coefficients, the main component of 3.38 under Design 3 can be shown to be equal to the following since out of p coefficients, $\frac{p+1}{2}$ are equal to 1 and $\frac{p-1}{2}$ are equal to -1,

$$\sum_{k=1}^{p}\rho \sum_{i=1,i\neq k}^{p}(\beta_i\beta_k)\lambda_{FR}^4 = -\frac{p-1}{2}\rho\lambda_{FR}^4 . \tag{3.51}$$

For the HEP, one is concerned with the summation of expressions 3.38 and 3.39 given as follows:

$$\sum_{k=1}^{s}\rho\sum_{i=1,i\neq k}^{s}\frac{\left(\rho^2(\sum_{j\neq i}^{p}\beta_j)(\sum_{j\neq k}^{p}\beta_j)+\frac{\sigma_\epsilon^2\rho}{n}\right)\lambda_i\lambda_k+O(1)}{\lambda_i\lambda_k+O(\lambda_i)+O(\lambda_k)} + \sum_{k=s+1}^{p}\rho\sum_{i=s+1,i\neq k}^{p}\frac{(\beta_i\beta_k)\lambda_{FR}^4+O(\lambda_{FR}^3)}{\lambda_{FR}^4+O(\lambda_{FR}^3)}. \quad (3.52)$$

The main components of this above expression are given by the following:

$$\sum_{k=1}^{s}\rho\sum_{i=1,i\neq k}^{s}(\rho^2(\sum_{j\neq i}^{p}\beta_j)(\sum_{j\neq k}^{p}\beta_j))\lambda_i\lambda_k+\sum_{k=s+1}^{p}\rho\sum_{i=s+1,i\neq k}^{p}(\beta_i\beta_k)\lambda_{FR}^4. \quad (3.53)$$

Using the Table below 3.47 showing the β profile of Design 3 and the fact that Partial Ridge is run on all coefficients that have a true value of 1, one can see that the first part of 3.53 is equal to 0 since all $\sum_{j\neq i}^{p}\beta_j=0$ for $\beta_i=1$. Now using that $\beta_i=-1$ for all $i=s+1,\ldots,p$, where p being an odd number means that $s=\frac{p+1}{2}$, the second part of the expression can be evaluated to $\frac{(p-1)(p-3)}{8}\rho\lambda_{FR}^4$. This leads to expression 3.53 being evaluated as follows:

$$\sum_{k=1}^{s}\rho\sum_{i=1,i\neq k}^{s}\left(\rho^2(\sum_{j\neq i}^{p}\beta_j)(\sum_{j\neq k}^{p}\beta_j)\right)\lambda_i\lambda_k+\sum_{k=s+1}^{p}\rho\sum_{i=s+1,i\neq k}^{p}(\beta_i\beta_k)\lambda_{FR}^4=\frac{(p-1)(p-3)}{8}\rho\lambda_{FR}^4. \quad (3.54)$$

By comparing 3.54 with 3.51 it can be seen why, under Design 3, Full Ridge dominates the HEP despite the latter providing lower individual coefficient MSEs. Since $\rho>0$ the expression in 3.51 will always be negative and grow in magnitude as p and ρ increase. On the other hand, for the HEP, 3.54 will always be positive for $\rho>0$ and $p>3$ and will grow in size as both p and ρ increase. Going further, the magnitude of 3.54 will increase at a faster rate than that of 3.51 as p increases which explains how the relative performance of the HEP deteriorates as p and ρ increase. Therefore, it is clear that caution must be exercised when implementing the HEP and one must not only consider the relative magnitude of β_i to $\sum_{j\neq i}\beta_j$, but also the signs of the β_i and $\sum_{j\neq i}\beta_j$ cross products. Where there is a mixture of signs of the β_i values then their cross products will often be negative meaning that Full Ridge alone will do well but when the $\sum_{j\neq i}\beta_j$ values vary in sign then Partial Ridge has the potential to bring predictive gains when combined with Full Ridge in a HEP setting.

This section has provided a thorough investigation into the situations required for Partial Ridge to be used in a Hybrid approach setting as a means of outperforming Full Ridge alone when concerned with predictive risk. While one can likely think of more scenarios where the β_i cross products have more negative values compared to

that of the $\sum_{j\neq i}\beta_j$ counterparts, there are still many situations where there is a strong case for employing the Partial Ridge method. For example, when concerned with demand functions one might choose to include a mixture of complementary and substitute good variables as candidate predictors resulting in it being such that $\sum_{j=1}^{p}\beta_j \approx 0$. This would cause a mixture in sign of the $\sum_{j\neq i}^{p}\beta_j$ values and give opportunity for the Partial Ridge approach to provide gains as demonstrated in Design 1. In the next section, a Monte Carlo Simulation experiment is run which tests the HEP against Full Ridge in a more pragmatic setting since consideration is given to the fact that the analyst will not know the true β_i values. Therefore, one will need to use a realistic means of determining which variables to estimate with Partial Ridge and Full Ridge.

3.4 Monte Carlo Simulation Study

3.4.1 Design

In this section a simulation experiment is run replicating a scenario where an analyst if faced with a high dimensional data set and wishes to forecast 1 step ahead with the largest degree of accuracy possible. The data-generating-process is detailed as follows with the basic linear model below forming the base of the design,

$$y = X\beta + \epsilon\,, \tag{3.55}$$

where y is the $n \times 1$ vector of the dependent variable and X is the $n \times p$ predictor matrix defined as a multivariate normal random variable with 0 mean and unit variance as follows:

$$X \sim N(0, \Sigma_X)\,, \tag{3.56}$$

$$\Sigma_X = \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix}.$$

The β component of 3.55 is a $p \times 1$ vector of coefficient determined in a way similar to that of Design 1 from the previous section. The individual components of β are determined as follows for $i = 1, \dots, p$ and let s represent the degree of sparsity where the number of non-zero coefficients is equal to $s \times p$ for $s \in (0, 1)$,

$$\beta_i = \begin{cases} \frac{(-1)^{i+1}i}{50} & \text{for } i \leq sp \\ 0 & \text{otherwise} \end{cases}. \tag{3.57}$$

Therefore, one can see that there is a varying degree of both sign and magnitude in the coefficients which proved to be beneficial for Partial Ridge and the HEP in the previous section. However , unlike the experiment prior to this, the degree of sparsity, s, is now varied across experiments in order to add another perspective for comparison purposes. Finally, the ϵ component of 3.55 is simulated by an IID normal random variable with mean 0 and variance equal to σ_ϵ^2. The σ_ϵ^2 is determined in a way that maintains the Signal-to-noise ratio of the DGP at a value equal 10 using the following definition of SNR:

$$SNR = \frac{\beta' \Sigma_X \beta}{\sigma_\epsilon^2}. \tag{3.58}$$

Since this is an out-of-sample forecasting experiment, a total of 120 observations are simulated for y and each regressor with the coefficient estimation occurring using only the first 100 observations. Therefore, for each data set, each approach constructs 20 forecasts using the relevant row of the X matrix but with the coefficients estimated using only the first 100 observations to create 20 single forecasts, $\hat{y}_{100+k} = x_{100+k}\hat{\beta}$. where x_{100+k} is the (100+k)th row of of X for $k = 1, \ldots, 20$. For example, once the coefficients have been estimated the first forecast to be constructed is for the 101th observation of the y vector, $\hat{y}_{101} = x_{101}\hat{\beta}$. To assess the accuracy of the competing estimation procedures, the mean-squared-forecasting error (MSFE) is used for to evaluate all 20 forecasts for each data set and is defined as follows:

$$MSFE = \sum_{i=1}^{20} (y_{100+i} - \hat{y}_{100+i})^2. \tag{3.59}$$

This MSFE is computed for each approach in each simulated data set with the lowest value representing the estimation procedure that has has provided the most accurate forecasts across the out-of-sample period.

To formalise this study further, multiple experiments are run with a varied dimension (p), degree of sparsity (s) and predictor correlation (ρ). With the in-sample estimation period fixed at 100 observations, p is varied over (100,150,200) creating scenarios where OLS is not feasible and, hence, establishes a high dimensional nature. The level of predictor correlation ranges over $\rho \in (0.2, 0.5, 0.8)$ since it was seen in the previous section how this can influence the relative predictive performance of Full Ridge and the HEP. Finally, the level of sparsity in the true DGP takes 3 forms with $s \in (0.2, 0.5, 0.7)$ to replicate a sparse, middle-ground and a dense setting in order to add greater coverage to the conclusions of this experiment with regards to empirical applications that these motions might be faced with.

Finally, the models compared here include not only Full Ridge and HEP but also the Lasso of Tibshirani (1996) along with the OLS-post-Lasso approach, whereby one uses

the Lasso for model selection the run OLS on the variables that did not have their coefficient estimate set to 0 by Lasso. This adds wider perspective to the HEP as typically Ridge does well in dense settings with high correlation amongst the predictors while Lasso and other forms of model selection succeed in more sparse environments, therefore, comparing the HEP with these 3 models allows one to see the merits of this new approach in the bigger picture. When estimating the coefficients from the in-sample data, both Full Ridge and Lasso have their penalty parameter determined using 10-fold cross validation with these Lasso coefficients being used to decide which variables are included in a simplified OLS model as part of the OLS-post-Lasso approach. Since OLS is only feasible when $n > p$, in the case where Lasso results in more than n coefficients having a non-zero coefficient, the first $n - 1$ columns of the predictor matrix being used as predictors to run OLS on. In the case where Lasso allocates all coefficients a value of 0 then the resulting forecast is 0 as well.

For the HEP, one must first determine which variables are to be estimated using Partial Ridge and Full Ridge. In this study, it is such that the Full Ridge coefficients are arranged by their magnitude in descending order with the variables corresponding to the largest g coefficients being estimated with Partial Ridge with the rest being allocated their equivalent estimate from the Full Ridge model. Here, g is a grid of $(1, 2, 3, 4, 5, 10)$ meaning that there are 6 different HEP coefficient vectors. For each case of g, the coefficients estimated with Partial Ridge each require an individual penalty parameter in order to estimate the given β_i, however, for simplicity, here these penalty parameters are set as being equal for all g of the coefficients being estimated with Partial Ridge. To determine this universal λ^{PR}, a grid of the penalty parameters is created and then an in-sample MSE estimate for the HEP created using the resulting Partial Ridge coefficient estimates (along with the Full Ridge ones) defined as $MSE = \sum_{i=1}^{100}(y - x_i\hat{\beta}_{Hy})^2$. The λ^{PR} chosen is the one that minimises this in-sample MSE with the coefficients then estimated and an out-of-sample forecast being computed in the usual way. The grid of λ^{PR} varies depending on p; for $p = 100$ the grid ranges from 1 to 2000, 1 to 6000 for $p = 150$ and 1 to 10000 for $p = 200$.

3.4.2 Results

As mentioned previously, for each of the 100 data sets, and each method run, an MSFE is computed with the tables below reporting the average of the MSFEs across the 100 simulated data set scenarios. Once again, it is important to mention that the average of the Hybrid MSFEs is the average of the 100 lowest MSFEs taken from the 6 approaches used on each data set, with each approach corresponding to a given element of the g grid for every given data set.

TABLE 3.10: MSFE values ($\times 10^{-2}$) for experiments when p=100

	s=0.2			s=0.5			s=0.7		
ρ	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
FR	0.0779	0.0465	0.0186	1.0530	0.7059	0.2934	3.0395	1.8877	0.8640
Hy	0.0726	0.0428	0.0162	1.0011	0.6595	0.2795	2.9087	1.8214	0.8297
Las	0.0375	0.0229	0.0102	0.9983	0.6287	0.2600	3.4105	2.0894	0.9044
OPL	0.0443	0.0271	0.0124	1.1309	0.7250	0.3116	4.2262	2.4869	1.1016

TABLE 3.11: MSFE values ($\times 10^{-2}$) for experiments when p=150

	s=0.2			s=0.5			s=0.7		
ρ	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
FR	0.3591	0.2246	0.0880	5.4826	3.2991	1.4382	14.968	9.2192	4.2394
Hy	0.3249	0.2041	0.0778	5.3156	3.1884	1.3337	14.563	8.8675	4.1518
Las	0.1847	0.1210	0.0489	6.2732	3.8470	1.6345	20.167	12.112	5.1644
OPL	0.2201	0.1496	0.0573	7.9390	4.3448	1.9374	36.681	20.130	6.3099

TABLE 3.12: MSFE values ($\times 10^{-2}$) for experiments when p=200

	s=0.2			s=0.5			s=0.7		
ρ	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
FR	0.9835	0.6119	0.2537	14.905	9.3472	3.8176	42.485	27.033	11.248
Hy	0.9124	0.5882	0.2308	14.458	9.3359	3.6391	43.026	26.429	11.190
Las	0.7228	0.4447	0.1890	18.821	11.729	4.9654	57.425	35.546	14.155
OPL	0.9434	0.5292	0.2130	178.58	96.731	40.640	73.425	72.418	46.156

The general pattern that can be seen here is that the l_2 norm methods (Full Ridge and Hybrid) perform stronger in more dense environments ($s = 0.7$) While the l_1 norm methods (Lasso and OLS-post-Lasso) perform stronger in more sparse environments ($s = 0.2$) irrespective of p and ρ. Focusing on the more dense beta profile settings, it can be seen that the Hybrid approach nearly always dominates, although the relative margin is somewhat reduced once p reaches 200. Over $s = 0.5$ and $s = 0.7$ the percentage gains in MSFE for the Hybrid approach over Full Ridge vary from 3.5% to 7% for $p = 100$, 2.1% to 7.3% for $p = 150$ and -1.3% to 4.7% for $p = 200$. Therefore, the gains appear to diminish as p increases but are still significant given how similar the Hybrid approach is to Full Ridge in terms of the parameter estimates used to compute forecasts. One could even argue that bigger improvements are possible with the use of a more effective algorithmic procedure for determining the individual penalty parameters for the Partial Ridge coefficients given the simplicity of the method used in this study.

While these experiments have covered a large variety of DGP conditions which a forecaster might be faced with, there are still many common data set features, such as non-stationarity and mixed scaling amongst the predictors, which are not covered here. Therefore, the next section compares the methods in an empirical application to add further scrutiny to the Hybrid approach.

3.5 Empirical Application

3.5.1 Data

Building more on the analysis of the previous sections, an application to predicting with real-world data is carried out. Specifically, a data set concerning prices and construction costs of residential apartments in Tehran, Iran, completed between the year 73 Q2 and 90 Q3 of the Persian calendar are taken from the Machine Learning repository of Dua and Graff (2019). The data set includes a total of 103 predictors including 8 project physical and financial variables (although one concerning project locality is dropped) as well as 19 economic variables, each with 5 lags. Finally, there are 2 dependent variables including the price the apartment is sold for and the total construction cost, both in Iranian Rial. With 372 observations for each property, the sample is split into training and testing data with the training data used to estimate the parameters of the appropriate methods which are then used to generate predictions of the testing data dependent variables with the testing data predictor values. The observations are ordered by the time of completion with the property completed first acting as the first observation and the property completed last acting as the final observation. However, as this is only recorded to the nearest quarter of a given year, many properties are essentially completed at the same time meaning that one has to be cautious when splitting the training and testing data sample to avoid having 2 properties completed at the same time in 2 separate samples.

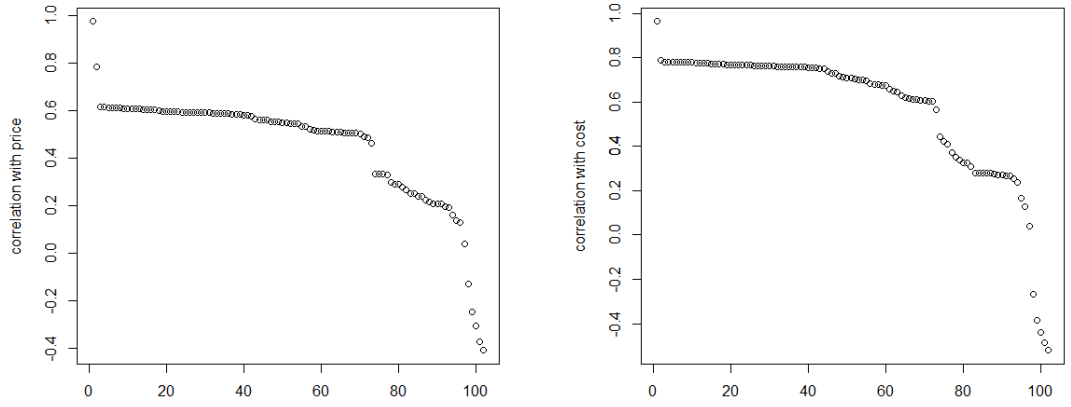
3.5.2 Prediction Experiment Design

To assess the Hybrid's predictive accuracy, the ordered data set is split into 2 parts with the first being used as the training data which allows the coefficients of the Hybrid approach and its competitors to be estimated. These coefficients are then used on the testing data predictor to obtain predictions, in the linear model form of $\hat{y}_{test} = X_{test}\hat{\beta}$, of either the selling price or construction cost of the property. The models compared to the Hybrid approach are, of course, Ridge Regression, along with the Lasso and OLS-post-Lasso procedure where, like in the previous section, the penalty parameter is determined using 10-fold cross validation. However, the parameters of the Hybrid approach estimated by Partial Ridge are determined in a

similar way to that which was used in the simulation study of the previous section whereby a grid of penalty parameters is created ranging from 0.1 to 2000 with varying increments. For each value of this grid, the in-sample prediction MSE is computed by estimating all Partial Ridge variables with the same given penalty parameter from the grid and the λ providing the lowest MSE is used to construct the coefficient vector for computing the predictions of the testing data.

For the Hybrid approach, a variety of approaches are used to rank covariates by their believed influence on y and, therefore, determine which coefficients are estimated by Partial Ridge and which are by Full Ridge. One is using the absolute value of the Lasso coefficients, for example, when choosing to estimate 10 coefficients by Partial Ridge, the variables that have their Lasso $\hat{\beta}$ training sample estimate in the top 10 when ranked by absolute value magnitude will be estimated by Partial Ridge. Secondly, the absolute value of the Full Ridge $\hat{\beta}$ values is also used in an identical fashion to that of the Lasso approach just discussed. Finally, variables are ranked by the magnitude of the absolute value of their training sample correlation with the dependent variable. Figure 3.1 below shows the correlation coefficient of each predictor with the given dependent variable ordered from the largest positive on the left through to the largest negative on the far right. Although this is for the entire data set, and not just the training data periods, it shows that many of the variables co-move with price and cost significantly and, therefore, should almost certainly be estimated by Partial Ridge in the Hybrid approach due to their likely large coefficient in the true DGP.

FIGURE 3.1: Ordered correlation of each predictor with price (left) and cost (right)



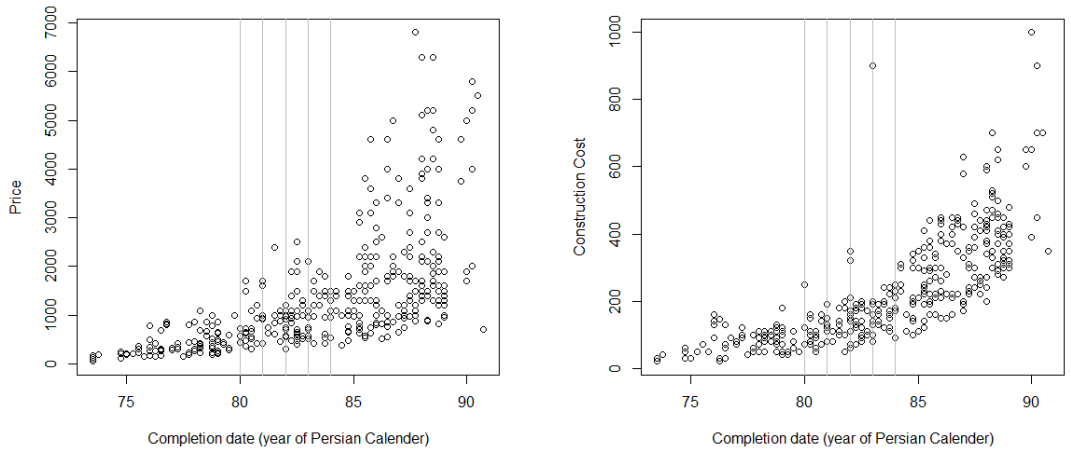
For robustness, multiple splits of the entire data set for the training and testing data sample are used with consideration given to how many groups of properties within the 372 observations have their completion date at the same time. Therefore, the

points of where the data set is split occur at the end of certain years of the Persian calendar and are detailed in the table below. It is seen from this that the splits result in larger means of the dependent variable for the testing data than the training data which leads on to another point about the standardization of variables. For all 4 methods used in this study, the estimation of the coefficients is carried out using the predictors and dependent variable after they have been standardized to have 0 mean and unit variance. However, when using the estimated slopes to construct predictions using the test data, one is required to compute values on the same scale as that of the non-standardized raw data set in the way detailed in section 25.12 of Stan User's guide (Stan 2011). Although it is assumed that the analyst would not know the mean and standard deviation of the testing data dependent variable, here, it is assumed that they know these quantities in order to compute non-standardized predictions more effectively as Table 3.13 and Figure 3.2 show a significant shift in these statistics from the training sample to the testing one. This is not a completely unrealistic assumption as forecasting average house prices as well as their volatility for a certain period is common so it is intuitive for one to compute predictions based off predicted mean prices and costs rather than that of the data available at a given time.

TABLE 3.13: Training and testing sample split summary data

Year of the Persian Calendar	79	80	81	82	83
Observation number of final property completed	84	107	129	162	184
$\frac{p}{n}$ split for training data	1.21	0.95	0.79	0.63	0.55
mean price training data	419.11	513.69	578.64	676.76	733.99
mean price testing data	1669.86	1740.23	1806.79	1935.67	2026.97
mean cost training data	79.23	85.65	96.32	109.78	117.53
mean cost testing data	277.40	292.00	305.02	327.43	345.32

FIGURE 3.2: Property selling price (left) and construction (right) across the entire data set period where the vertical grey lines represent that training and testing data split points



3.5.3 Results and Discussion

To compare predictive performance of these methods for each split of the data set, the predicted values of the testing data dependent variable values are compared to that of the true ones through the Mean-Squared-Prediction-Error (MSPE) as well as the out-of-sample R squared value defined respectively as follows where o is the number of the observation from the original data set used as the last observation of the training data,

$$MSPE = \frac{1}{372 - o} \sum_{i=o+1}^{372} (y_i - \hat{y}_i)^2 , \tag{3.60}$$

$$R^2 = 1 - \frac{\sum_{i=o+1}^{372} (y_i - \hat{y}_i)^2}{\sum_{i=o+1}^{372} (y_i - \hat{y}_i^{bcm})^2} , \tag{3.61}$$

where y_i represent the true values of the dependent variable, $\hat{y}_i$ are the predicted values by the given model and $\hat{y}_i^{bcm}$ represent the predictions from the benchmark model, which here is simply the mean of testing data dependent variable values (once again, this is assumed to be known). The MSPE is a standard measure for comparability amongst the models and the out-of-sample R squared adds context to the accuracy of each model as MSPE is subject to the scale of house prices and construction costs as raw variables. The tables below report the MSPE for each approach under each split of the original data set with the R squared values reported in the Appendix. The Hybrid approach is estimated 30 times under each of the 3 ranking criteria discussed above, for example, when using the ranked Ridge coefficient estimates to determine which coefficients to estimate with Partial Ridge, first one considers only estimating the variable with the largest Ridge $\hat{\beta}$ with Partial,

then the top 2, all the way up until the top 30. Finally, the numbers in brackets represent the number of predictors included for OLS-post-Lasso and the optimal number of predictors estimated with Partial Ridge for the Hybrid approach.

TABLE 3.14: MSPEs ($\times 10^{-2}$) for house price prediction

	Sample split (no. of observations in training data)				
	84	107	129	162	184
Full Ridge	1075	1088	965	1091	1373
Lasso	832	830	891	980	1096
OLS-post-Lasso	816 (11)	849 (11)	882 (15)	998 (21)	1075 (12)
Hybrid (using correlation)	724 (26)	796 (3)	762 (26)	1091 (1)	1132 (24)
Hybrid (using Lasso $\hat{\beta}s$)	971 (14)	992 (14)	777 (14)	903 (30)	1202 (26)
Hybrid (using Ridge $\hat{\beta}s$)	1218 (16)	890 (22)	778 (25)	943 (12)	1081 (17)

TABLE 3.15: MSPEs for construction cost prediction

	Sample split (no. of observations in training data)				
	84	107	129	162	184
Full Ridge	3383	3804	5844	7935	6955
Lasso	10141	5685	2261	2023	2199
OLS-post-Lasso	5659 (3)	4324 (4)	22110 (7)	2057 (14)	3466 (24)
Hybrid (using correlation)	6788 (1)	7796 (1)	6148 (1)	8155 (5)	5260 (18)
Hybrid (using Lasso $\hat{\beta}s$)	6047 (9)	7796 (1)	6120 (1)	7402 (5)	6337 (8)
Hybrid (using Ridge $\hat{\beta}s$)	6788 (1)	7796 (1)	4193 (20)	7111 (20)	5870 (21)

As can be seen above, the Hybrid approach is very successful in terms of relative prediction accuracy for the price of apartments with at least one of them dominating all other methods for each case, with the exception of the sample being split after the 184th observation where OLS-post-Lasso is better by a small margin. The appendix tables also show very high R squared values for all methods indicating that all models here make a respectable attempt at prediction and the Hybrid approach outperforming its competitors is even more significant. Although, it is interesting to see how there is no consistency in which method of selection for coefficients to be estimated with Partial Ridge is best, and sometimes this determines the Hybrid's success compared to its competitors, the main point is that the Hybrid approach has the means in a realistic way to improve upon well established methods of prediction.

For construction cost prediction, the results are less promising but there is likely reasonable justification for this. Firstly, it is seen that for for the 3 columns from the right the Hybrid approach can outperform Full Ridge but falls significantly short of the Lasso and OLS-post-Lasso. This is quite a common occurrence as much of the

literature describes Lasso as having greater potential in settings that are either sparse or inactive predictors have limited relation to active ones as discussed in Zhao and Yu (2006). One might suspect this in this setting where the majority of the predictors are economics variables and their lags which will likely heavily influence the property market through price. whereas for the cost of building properties, one would expect this to be more influenced by the project physical variables, which there are fewer of. For the first 2 sample splits (after the 84th and 107th observation), the results are somewhat more dubious as here Full Ridge dominates with the Hybrid approach, as well as the Lasso based methods, performing significantly inferior. While looking closer at the ranking methods (to determine which coefficients to estimate with Partial Ridge), it is difficult to spot and any issues with choosing the wrong variables to estimate with Partial Ridge compared to the cases where the training set is larger. Therefore, one can put this down to the very nature of the small training set compared to the testing set making performance unpredictable, especially how Figure 3.2 shows significantly greater variance on the right side of the graph. While the cost prediction appears to be disheartening for the Hybrid approach, it can be viewed as an important reminder that; firstly, l_2 penalisation struggles against the Lasso in certain situations and, secondly, that the nuisance parameters of the Hybrid approach become amplified when faced with a very small training sample relative to that of the testing one.

3.6 Discussion

This paper has compared the Partial Ridge and Hybrid approach to the ordinary Ridge Regression of Hoerl and Kennard (1970) through detailed theoretical analysis of the individual coefficient estimates as well as the overall prediction of the dependent variable, $X\hat{\beta}$. In addition, it has used a simulation experiment and an empirical application including other well-established models to compete with in order to justify its use as an alternative to Ridge when the bias-variance tradeoff is unfavourable. While, theoretically, there are situations where the Hybrid approach has the potential to outperform Ridge in terms of prediction accuracy, when faced with a data set alone, there are extra layers of contamination such as choosing which coefficients to estimate with Partial Ridge as well as the penalty parameter to use in each case. However, the simulations and application to Iranian residential property data show that the restraint of this does not prevent the Hybrid approach from outperforming its profound competitors. Therefore, there is potential for one wishing to forecast variables, in settings where they suspect the true coefficients vary greatly in sign and magnitude, to benefit from the adoption of the Hybrid approach as opposed to Ridge or Lasso alone.

Further research that would provide further support for this method include attempts to understand the sensitivity of parameter estimates to λ and ρ. Such an approach

could be based upon the work of Banerjee et al. (1999) where the GLS estimator sensitivity to covariance matrix misspecification was measured with the use of a novel test statistic. This is particularly relevant to high dimensional economic data sets where autoregressive processes are common and are key contributor to error covariance matrix misspecification. In addition, work analysing the asymptotic bias-variance behaviour with respect to the choice of variables estimated by PR in the HEP would support further establishment of this novel approach.

3.A Appendix

3.A.1 Derivation of MSE in Proposition 3.1

For Full Ridge the MSE can be computed by considering the bias and variance separately. As shown in Section 3.2, the Full Ridge estimator is given by the following expression:

$$\hat{\beta}_i(\lambda^{FR}) = \frac{x_i^{*'} y}{x_i^{*'} x_i + n\lambda^{FR}},$$

where $x_i^* = x_i - X_{-i}\left(\frac{X'_{-i}X_{-i}}{n} + \lambda^{FR} I_{p-1}\right)^{-1} \frac{X'_{-i}x_i}{n}$. Plugging 3.11 into the above gives the following:

$$\hat{\beta}_i(\lambda^{FR}) = \frac{x_i^{*'} x_i \beta_i + x_i^{*'} X_{-i}\beta_{-i} + x_i^{*'}\epsilon}{x_i^{*'} x_i + n\lambda^{FR}}.$$

Using the fact that $E[\epsilon] = 0$, the bias can be formulated as follows:

$$Bias(\hat{\beta}_i(\lambda^{FR})) = E[\hat{\beta}_i(\lambda^{FR})] - \beta_i = \frac{x_i^{*'} x_i \beta_i + x_i^{*'} X_{-i}\beta_{-i}}{x_i^{*'} x_i + n\lambda^{FR}} - \beta_i,$$

$$Bias(\hat{\beta}_i(\lambda^{FR})) = \frac{x_i^{*'} X_{-i}\beta_{-i} - n\lambda^{FR}\beta_i}{x_i^{*'} x_i + n\lambda^{FR}}.$$

Various components of the above can be evaluated using the SVD of $\frac{X_{-i}}{\sqrt{n}}$ described in Section 3.2. Consider the following:

$$\frac{x_i^*}{\sqrt{n}} = \frac{x_i}{\sqrt{n}} - \frac{X_{-i}}{\sqrt{n}}\left(\frac{X'_{-i}X_{-i}}{n} + \lambda^{FR} I_{p-1}\right)^{-1} \frac{X'_{-i}x_i}{n}.$$

Since $\frac{X'_{-i}X_{-i}}{n} = VSU'USV' = VS^2V'$, it is such that

$$\frac{x_i^*}{\sqrt{n}} = \frac{x_i}{\sqrt{n}} - USV'(VS^2V' + \lambda^{FR} VV')^{-1} VSU' \frac{x_i}{\sqrt{n}},$$

$$\frac{x_i^*}{\sqrt{n}} = \frac{x_i}{\sqrt{n}} - USV'(V(S^2 + \lambda^{FR} I_{p-1})V')^{-1} VSU' \frac{x_i}{\sqrt{n}},$$

$$\frac{x_i^*}{\sqrt{n}} = \frac{x_i}{\sqrt{n}} - USV'(V')^{-1}(S^2 + \lambda^{FR} I_{p-1})^{-1} V^{-1} VSU' \frac{x_i}{\sqrt{n}},$$

$$\frac{x_i^*}{\sqrt{n}} = \frac{x_i}{\sqrt{n}} - US(S^2 + \lambda^{FR} I_{p-1})^{-1} SU' \frac{x_i}{\sqrt{n}},$$

$$\frac{x_i^*}{\sqrt{n}} = \frac{x_i}{\sqrt{n}} - US^2(S^2 + \lambda^{FR} I_{p-1})^{-1} U' \frac{x_i}{\sqrt{n}}.$$

Now consider $\frac{x_i^{*'}x_i}{n}$,

$$\frac{x_i^{*'}x_i}{n} = \frac{x_i'x_i}{n} - \frac{x_i'}{\sqrt{n}}US^2(S^2+\lambda^{FR}I_{p-1})^{-1}U'\frac{x_i}{\sqrt{n}}\,.$$

Now let $\tilde{x}_i = \frac{U'x_i}{\sqrt{n}}$ and use the fact that $x_i'x_i = n$,

$$\frac{x_i^{*'}x_i}{n} = 1 - \tilde{x}_i'S^2(S^2+\lambda^{FR}I_{p-1})^{-1}\tilde{x}_i\,.$$

Now consider the following component of the Full Ridge bias expression,

$$\frac{x_i^{*'}X_{-i}\beta_{-i}}{n} = \Big(\frac{x_i'}{\sqrt{n}} - \frac{x_i'}{\sqrt{n}}US^2(S^2+\lambda^{FR}I_{p-1})^{-1}U'\Big)\frac{X_{-i}\beta_{-i}}{\sqrt{n}}\,,$$

$$\frac{x_i^{*'}X_{-i}\beta_{-i}}{n} = \frac{x_i'USV'\beta_{-i}}{\sqrt{n}} - \frac{x_i'}{\sqrt{n}}US^2(S^2+\lambda^{FR}I_{p-1})^{-1}U'USV'\beta_{-i}\,.$$

Now use the fact that $\tilde{x}_i = \frac{U'x_i}{\sqrt{n}}$ as well as let $\beta_{-i}^* = V'\beta_{-i}$,

$$\frac{x_i^{*'}X_{-i}\beta_{-i}}{n} = \tilde{x}_i'S\beta_{-i}^* - \tilde{x}_i'S^2(S^2+\lambda^{FR}I_{p-1})^{-1}S\beta_{-i}^*\,,$$

$$\frac{x_i^{*'}X_{-i}\beta_{-i}}{n} = \tilde{x}_i'S\big(I_{p-1} - S^2(S^2+\lambda^{FR}I_{p-1})^{-1}\big)\beta_{-i}^*\,,$$

$$\frac{x_i^{*'}X_{-i}\beta_{-i}}{n} = \lambda^{FR}\tilde{x}_i'S(S^2+\lambda^{FR}I_{p-1})^{-1}\beta_{-i}^*\,.$$

Therefore, the bias of the Full Ridge estimator can be given by the following expression:

$$Bias(\hat{\beta}_i(\lambda^{FR})) = \frac{\lambda^{FR}\tilde{x}_i'S(S^2+\lambda^{FR}I_{p-1})^{-1}\beta_{-i}^* - \lambda^{FR}\beta_i}{1-\tilde{x}_i'S^2(S^2+\lambda^{FR}I_{p-1})^{-1}\tilde{x}_i + \lambda^{FR}}\,.$$

For the variance it is such that

$$Var(\hat{\beta}_i(\lambda^{FR})) = E\Big[\big(\hat{\beta}_i(\lambda^{FR}) - E[\hat{\beta}_i(\lambda^{FR})]\big)^2\Big]\,,$$

$$Var(\hat{\beta}_i(\lambda^{FR})) = E\Big[\frac{x_i^{*'}\epsilon\epsilon'x_i^*}{(x_i^{*'}x_i+n\lambda^{FR})^2}\Big] = \sigma_\epsilon^2\frac{x_i^{*'}x_i^*}{(x_i^{*'}x_i+n\lambda^{FR})^2}\,.$$

Now $\frac{x_i^{*'}x_i^*}{n}$ is computed as follows,

$$\frac{x_i^{*'}x_i^*}{n} = \left(\frac{x_i}{\sqrt{n}} - US^2(S^2+\lambda^{FR}I_{p-1})^{-1}U'\frac{x_i}{\sqrt{n}}\right)'\left(\frac{x_i}{\sqrt{n}} - US^2(S^2+\lambda^{FR}I_{p-1})^{-1}U'\frac{x_i}{\sqrt{n}}\right)\,,$$

$$\frac{x_i^{*'}x_i^*}{n} = \frac{x_i'x_i}{n} - \frac{x_i'}{\sqrt{n}}US^2(S^2+\lambda^{FR})^{-1}U'\frac{x_i}{\sqrt{n}} - \frac{x_i'}{\sqrt{n}}US^2(S^2+\lambda^{FR})^{-1}U'\frac{x_i}{\sqrt{n}} + \frac{x_i'}{\sqrt{n}}US^2(S^2+\lambda^{FR})^{-1}U'US^2(S^2+\lambda^{FR}I_{p-1})^{-1}U'\frac{x_i}{\sqrt{n}}.$$

Now use the fact that $\tilde{x}_i = \frac{U'x_i}{\sqrt{n}}$ to obtain the following:

$$\frac{x_i^{*'}x_i^*}{n} = 1 - 2\tilde{x}_i'S^2(S^2+\lambda^{FR})^{-1}\tilde{x}_i + \tilde{x}_i'S^4(S^2+\lambda^{FR})^{-2}\tilde{x}_i,$$

$$\frac{x_i^{*'}x_i^*}{n} = 1 - \tilde{x}_i'S^2\Big(2(S^2+\lambda^{FR})^{-1} - S^2(S^2+\lambda^{FR})^{-2}\Big)\tilde{x}_i,$$

$$\frac{x_i^{*'}x_i^*}{n} = 1 - \tilde{x}_i'S^2(S^2+2\lambda^{FR}I_{p-1})(S^2+\lambda^{FR}I_{p-1})^{-2}\tilde{x}_i.$$

Now plug the relevant expressions into the following:

$$Var(\hat{\beta}_i(\lambda^{FR})) = \sigma_\epsilon^2\frac{x_i^{*'}x_i^*}{(x_i^{*'}x_i+n\lambda^{FR})^2} = \sigma_\epsilon^2\frac{\frac{x_i^{*'}x_i^*}{n}}{n(\frac{x_i^{*'}x_i}{n}+\lambda^{FR})^2},$$

$$Var(\hat{\beta}_i(\lambda^{FR})) = \frac{\sigma_\epsilon^2}{n}\frac{1-\tilde{x}_i'S^2(S^2+2\lambda^{FR}I_{p-1})(S^2+\lambda^{FR}I_{p-1})^{-2}\tilde{x}_i}{(1-\tilde{x}_i'S^2(S^2+\lambda^{FR}I_{p-1})^{-1}\tilde{x}_i+\lambda^{FR})^2}.$$

From here, one can use the relationship between the MSE, variance and bias to obtain the expression in Proposition 3.1.

For Partial Ridge, recall the following expression from Section 3.2,

$$\hat{\beta}_i(\lambda_i, S_i) = \frac{x_i^{*'}y}{x_i^{*'}x_i}.$$

Now plug 4.11 into the above to obtain the following:

$$\hat{\beta}_i(\lambda_i, S_i) = \frac{x_i^{*'}x_i\beta_i + x_i^{*'}X_{-i}\beta_{-i} + x_i^{*'}\epsilon}{x_i^{*'}x_i} = \beta_i + \frac{x_i^{*'}X_{-i}\beta_{-i} + x_i^{*'}\epsilon}{x_i^{*'}x_i}.$$

Using the fact that $E[\epsilon] = 0$, the bias can be written as follows,

$$Bias(\hat{\beta}_i(\lambda_i, S_i)) = E[\hat{\beta}_i(\lambda_i, S_i)] - \beta_i = \frac{x_i^{*'}X_{-i}\beta_{-i}}{x_i^{*'}x_i}.$$

Using the Full Ridge case, one can see that the numerator can be written as follows,

$$\frac{x_i^{*'}X_{-i}\beta_{-i}}{n} = \lambda_i\tilde{x}_i'S(S^2+\lambda_iI_{p-1})^{-1}\beta_{-i}^*.$$

Also from the proof for the Full Ridge case, the denominator can be expressed as follows,

$$\frac{x_i^{*'} x_i}{n} = 1 - \tilde{x}_i' S^2 (S^2 + \lambda_i I_{p-1})^{-1} \tilde{x}_i \,.$$

Therefore, the bias of the Partial Ridge estimator can be written as follows.

$$Bias(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\lambda_i \tilde{x}_i' S (S^2 + \lambda_i I_{p-1})^{-1} \beta^*_{-i}}{1 - \tilde{x}_i' S^2 (S^2 + \lambda^{FR} I_{p-1})^{-1} \tilde{x}_i} \,.$$

For the variance it is such that

$$Var(\hat{\beta}_i(\lambda_i, S_i)) = E\Big[\big(\hat{\beta}_i(\lambda_i, S_i) - E[\hat{\beta}_i(\lambda_i, S_i)]\big)^2 \Big] = E\Big[\frac{x_i^{*'} \epsilon \epsilon' x_i^*}{(x_i^{*'} x_i)^2} \Big] = \sigma_\epsilon^2 \frac{x_i^{*'} x_i^*}{(x_i^{*'} x_i)^2} \,.$$

Using the Full Ridge case, the main component of the numerator can be computed as follows,

$$\frac{x_i^{*'} x_i^*}{n} = 1 - \tilde{x}_i' S^2 (S^2 + 2\lambda_i I_{p-1})(S^2 + \lambda_i I_{p-1})^{-2} \tilde{x}_i \,.$$

Plugging these into the variance expression leads to the following,

$$Var(\hat{\beta}_i(\lambda_i, S_i)) = \sigma_\epsilon^2 \frac{\frac{x_i^{*'} x_i^*}{n}}{n(\frac{x_i^{*'} x_i}{n})^2} = \frac{\sigma_\epsilon^2}{n} \frac{\frac{x_i^{*'} x_i^*}{n}}{(\frac{x_i^{*'} x_i}{n})^2} \,,$$

$$Var(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\sigma_\epsilon^2}{n} \Big[\frac{1 - \tilde{x}_i' S^2 (S^2 + 2\lambda_i I_{p-1})(S^2 + \lambda_i I_{p-1})^{-2} \tilde{x}_i}{(1 - \tilde{x}_i' S^2 (S^2 + \lambda_i I_{p-1})^{-1} \tilde{x}_i)^2} \Big] \,.$$

From here, one can use the relationship between the MSE, variance and bias to obtain the expression in Proposition 3.1.

3.A.2 Derivation of expression in Proposition 3.2

For Full Ridge it is such that:

$$\begin{bmatrix} \hat{\beta}_1(\lambda^{FR}) \\ \vdots \\ \hat{\beta}_p(\lambda^{FR}) \end{bmatrix} = \frac{1}{n}\left(\frac{X'X}{n} + \lambda^{FR} I_p\right)^{-1} X'y\,.$$

Consider the matrix $\frac{1}{n}\left(\frac{X'X}{n} + \lambda^{FR} I_p\right)^{-1} X'y$ evaluated for various values of p below.

For $p = 3$

$$\frac{1}{n}\left(\frac{X'X}{n} + \lambda^{FR} I_p\right)^{-1} = \frac{1}{n}\begin{bmatrix} 1+\lambda^{FR} & \rho & \rho \\ \rho & 1+\lambda^{FR} & \rho \\ \rho & \rho & 1+\lambda^{FR} \end{bmatrix}^{-1}$$

$$= \frac{1}{n(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+2\rho)} \begin{bmatrix} (1+\lambda^{FR})^2-\rho^2 & -\rho(1+\lambda^{FR}-\rho) & -\rho(1+\lambda^{FR}-\rho) \\ -\rho(1+\lambda^{FR}-\rho) & (1+\lambda^{FR})^2-\rho^2 & -\rho(1+\lambda^{FR}-\rho) \\ -\rho(1+\lambda^{FR}-\rho) & -\rho(1+\lambda^{FR}-\rho) & (1+\lambda^{FR})^2-\rho^2 \end{bmatrix}$$

For $p = 4$

$$\frac{1}{n}\left(\frac{X'X}{n}+\lambda^{FR}I_p\right)^{-1} = \frac{1}{n}\begin{bmatrix} 1+\lambda^{FR} & \rho & \rho & \rho \\ \rho & 1+\lambda^{FR} & \rho & \rho \\ \rho & \rho & 1+\lambda^{FR} & \rho \\ \rho & \rho & \rho & 1+\lambda^{FR} \end{bmatrix}^{-1}$$

$$= \frac{1}{n(1+\lambda^{FR}-\rho)^3(1+\lambda^{FR}+3\rho)} \begin{bmatrix} (1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+2\rho) & -\rho(1+\lambda^{FR}-\rho)^2 & -\rho(1+\lambda^{FR}-\rho)^2 & -\rho(1+\lambda^{FR}-\rho)^2 \\ -\rho(1+\lambda^{FR}-\rho)^2 & (1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+2\rho) & -\rho(1+\lambda^{FR}-\rho)^2 & -\rho(1+\lambda^{FR}-\rho)^2 \\ -\rho(1+\lambda^{FR}-\rho)^2 & -\rho(1+\lambda^{FR}-\rho)^2 & (1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+2\rho) & -\rho(1+\lambda^{FR}-\rho)^2 \\ -\rho(1+\lambda^{FR}-\rho)^2 & -\rho(1+\lambda^{FR}-\rho)^2 & -\rho(1+\lambda^{FR}-\rho)^2 & (1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+2\rho) \end{bmatrix}$$

For $p = 5$

$$\frac{1}{n}\left(\frac{X'X}{n}+\lambda^{FR}I_p\right)^{-1} = \frac{1}{n}\begin{bmatrix} 1+\lambda^{FR} & \rho & \rho & \rho & \rho \\ \rho & 1+\lambda^{FR} & \rho & \rho & \rho \\ \rho & \rho & 1+\lambda^{FR} & \rho & \rho \\ \rho & \rho & \rho & 1+\lambda^{FR} & \rho \\ \rho & \rho & \rho & \rho & 1+\lambda^{FR} \end{bmatrix}^{-1} = \frac{1}{n(1+\lambda^{FR}-\rho)^4(1+\lambda^{FR}+4\rho)}$$

$$\times \begin{bmatrix} (1+\lambda^{FR}-\rho)^3(1+\lambda^{FR}+3\rho) & -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 \\ -\rho(1+\lambda^{FR}-\rho)^3 & (1+\lambda^{FR}-\rho)^3(1+\lambda^{FR}+3\rho) & -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 \\ -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 & (1+\lambda^{FR}-\rho)^3(1+\lambda^{FR}+3\rho) & -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 \\ -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 & (1+\lambda^{FR}-\rho)^3(1+\lambda^{FR}+3\rho) & -\rho(1+\lambda^{FR}-\rho)^3 \\ -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 & -\rho(1+\lambda^{FR}-\rho)^3 & (1+\lambda^{FR}-\rho)^3(1+\lambda^{FR}+3\rho) \end{bmatrix}$$

For estimation of $\hat{\beta}_i(\lambda^{FR})$, only the ith row of $\frac{1}{n}\left(\frac{X'X}{n}+\lambda^{FR}I_p\right)^{-1}$ must be considered and multiplied by $X'y$. From the above, recursively one can see that this leads the following expression for p in general,

$$\hat{\beta_1}(\lambda^{FR}) = \frac{1}{n(1+\lambda^{FR}-\rho)^{p-1}(1+\lambda^{FR}+(p-1)\rho)} \begin{bmatrix} (1+\lambda^{FR}-\rho)^{p-2}(1+\lambda^{FR}+(p-2)\rho) \\ -\rho(1+\lambda^{FR}-\rho)^{p-2} \\ \vdots \\ -\rho(1+\lambda^{FR}-\rho)^{p-2} \end{bmatrix}' \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_p' \end{bmatrix} y\,,$$

$$\hat{\beta_1}(\lambda^{FR}) = \frac{(1+\lambda^{FR}-\rho)^{p-2}(1+\lambda^{FR}+(p-2)\rho)x_1'y - \rho(1+\lambda^{FR}-\rho)^{p-2}\sum_{j\neq 1}^{p} x_j'y}{n(1+\lambda^{FR}-\rho)^{p-1}(1+\lambda^{FR}+(p-1)\rho)}\,,$$

$$\hat{\beta_1}(\lambda^{FR}) = \frac{(1+\lambda^{FR}+(p-2)\rho)x_1'y - \rho\sum_{j\neq 1}^{p} x_j'y}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}\,.$$

For Partial Ridge it is such that:

$$\begin{bmatrix} \hat{\beta}_1(\lambda_i, S_i) \\ \vdots \\ \hat{\beta}_p(\lambda_i, S_i) \end{bmatrix} = \frac{1}{n}\left(\frac{X'X}{n}+\lambda_i S_i\right)^{-1} X'y\,,$$

where S_i is the identity matrix but with the element (i,i) replaced with 0. For representational simplicity here, i=1. Consider the following evaluations of the matrix, $\frac{1}{n}\left(\frac{X'X}{n}+\lambda_i S_i\right)^{-1} X'y$, for various values of p.

For $p = 3$

$$\frac{1}{n}\left(\frac{X'X}{n}+\lambda_i S_i\right)^{-1} = \frac{1}{n}\begin{bmatrix}1 & \rho & \rho\\ \rho & 1+\lambda_i & \rho\\ \rho & \rho & 1+\lambda_i\end{bmatrix}^{-1}$$

$$= \frac{1}{n(1+\lambda_i-\rho)(1+\lambda_i+\rho-2\rho^2)}\begin{bmatrix}(1+\lambda_i-\rho)(1+\lambda_i+\rho) & -\rho(1+\lambda_i-\rho) & -\rho(1+\lambda_i-\rho)\\ -\rho(1+\lambda_i-\rho) & 1+\lambda_i-\rho^2 & \rho(\rho-1)\\ -\rho(1+\lambda_i-\rho) & \rho(\rho-1) & 1+\lambda_i-\rho^2\end{bmatrix}$$

For $p = 4$

$$\frac{1}{n}\left(\frac{X'X}{n}+\lambda_i S_i\right)^{-1} = \frac{1}{n}\begin{bmatrix}1 & \rho & \rho & \rho\\ \rho & 1+\lambda_i & \rho & \rho\\ \rho & \rho & 1+\lambda_i & \rho\\ \rho & \rho & \rho & 1+\lambda_i\end{bmatrix}^{-1} =$$

$$\frac{1}{n(1+\lambda_i-\rho)^2(1+\lambda_i+2\rho-3\rho^3)}\begin{bmatrix}(1+\lambda_i-\rho)^2(1+\lambda_i+2\rho) & -\rho(1+\lambda_i-\rho)^2 & -\rho(1+\lambda_i-\rho)^2 & -\rho(1+\lambda_i-\rho)^2\\ -\rho(1+\lambda_i-\rho)^2 & (1+\lambda_i-\rho)(1+\lambda_i+\rho-2\rho^2) & \rho(\rho-1)(1+\lambda_i-\rho) & \rho(\rho-1)(1+\lambda_i-\rho)\\ -\rho(1+\lambda_i-\rho)^2 & \rho(\rho-1)(1+\lambda_i-\rho) & (1+\lambda_i-\rho)(1+\lambda_i+\rho-2\rho^2) & \rho(\rho-1)(1+\lambda_i-\rho)\\ -\rho(1+\lambda_i-\rho)^2 & \rho(\rho-1)(1+\lambda_i-\rho) & \rho(\rho-1)(1+\lambda_i-\rho) & (1+\lambda_i-\rho)(1+\lambda_i+\rho-2\rho^2)\end{bmatrix}$$

For $p = 5$

$$\frac{1}{n}\left(\frac{X'X}{n}+\lambda_i S_i\right)^{-1} = \frac{1}{n}\begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1+\lambda_i & \rho & \rho & \rho \\ \rho & \rho & 1+\lambda_i & \rho & \rho \\ \rho & \rho & \rho & 1+\lambda_i & \rho \\ \rho & \rho & \rho & \rho & 1+\lambda_i \end{bmatrix}^{-1} = \frac{1}{n(1+\lambda_i-\rho)^3(1+\lambda_i+3\rho-4\rho^2)}$$

$$\times \begin{bmatrix} (1+\lambda_i-\rho)^3(1+\lambda_i+3\rho) & -\rho(1+\lambda_i-\rho)^3 & -\rho(1+\lambda_i-\rho)^3 & -\rho(1+\lambda_i-\rho)^3 & -\rho(1+\lambda_i-\rho)^3 \\ -\rho(1+\lambda_i-\rho)^3 & (1+\lambda_i-\rho)^2(1+\lambda_i+2\rho-3\rho^2) & \rho(\rho-1)(1+\lambda_i-\rho)^2 & \rho(\rho-1)(1+\lambda_i-\rho)^2 & \rho(\rho-1)(1+\lambda_i-\rho)^2 \\ -\rho(1+\lambda_i-\rho)^3 & \rho(\rho-1)(1+\lambda_i-\rho)^2 & (1+\lambda_i-\rho)^2(1+\lambda_i+2\rho-3\rho^2) & \rho(\rho-1)(1+\lambda_i-\rho)^2 & \rho(\rho-1)(1+\lambda_i-\rho)^2 \\ -\rho(1+\lambda_i-\rho)^3 & \rho(\rho-1)(1+\lambda_i-\rho)^2 & \rho(\rho-1)(1+\lambda_i-\rho)^2 & (1+\lambda_i-\rho)^2(1+\lambda_i+2\rho-3\rho^2) & \rho(\rho-1)(1+\lambda_i-\rho)^2 \\ -\rho(1+\lambda_i-\rho)^3 & \rho(\rho-1)(1+\lambda_i-\rho)^2 & \rho(\rho-1)(1+\lambda_i-\rho)^2 & \rho(\rho-1)(1+\lambda_i-\rho)^2 & (1+\lambda_i-\rho)^2(1+\lambda_i+2\rho-3\rho^2) \end{bmatrix}$$

For estimation of $\hat{\beta}_i(\lambda_i, S_i)$, only the ith row of $\frac{1}{n}\left(\frac{X'X}{n}+\lambda_i S_i\right)^{-1}$ must be considered and multiplied by $X'y$. From the above, recursively one can see that this leads the following expression for p in general,

$$\hat{\beta}_i(\lambda_i, S_i) = \frac{1}{n(1+\lambda_i-\rho)^{p-2}(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)} \begin{bmatrix} (1+\lambda_i-\rho)^{p-2}(1+\lambda_i+(p-2)\rho) \\ -\rho(1+\lambda_i-\rho)^{p-2} \\ \vdots \\ -\rho(1+\lambda_i-\rho)^{p-2} \end{bmatrix}' \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_p' \end{bmatrix} y\,,$$

$$\hat{\beta}_i(\lambda_i, S_i) = \frac{(1+\lambda_i-\rho)^{p-2}(1+\lambda_i+(p-2)\rho)x_i'y - \rho(1+\lambda_i-\rho)^{p-2}\sum_{j\neq i}^{p} x_j'y}{n(1+\lambda_i-\rho)^{p-2}(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)},$$

$$\hat{\beta}_i(\lambda_i, S_i) = \frac{(1+\lambda_i+(p-2)\rho)x_i'y - \rho\sum_{j\neq i}^{p} x_j'y}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)}.$$

3.A.3 Derivation of expressions in Proposition 3.3

3.A.3.1 Full Ridge Bias

Recall from 3.12 that

$$\hat{\beta}_i(\lambda^{FR}) = \frac{(1+\lambda^{FR}+(p-2)\rho)x_i'y - \rho\sum_{j\neq i}^{p} x_j'y}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}\,.$$

Replacing y with the true model from 3.1 leads to the following expression for the numerator,

$$\begin{aligned}(1+\lambda^{FR}+(p-2)\rho)\big(&x_i'x_1\beta_1 + x_i'x_2\beta_2 + \cdots + x_i'x_p\beta_p + x_i'\epsilon\big)\\ &-\rho\big(x_1'x_1\beta_1 + x_1'x_2\beta_2 + \cdots + x_1'x_p\beta_p + x_1'\epsilon\\ &+ x_2'x_1\beta_1 + x_2'x_2\beta_2 + x_2'x_3\beta_3 + \cdots + x_2'x_p\beta_p + x_2'\epsilon\\ &+\cdots+\\ &x_p'x_1\beta_1 + \cdots + x_p'x_p\beta_p + x_p'\epsilon\big)\,.\end{aligned}$$

Using A3 and collecting terms results in the following expression for the numerator,

$$(1+\lambda^{FR}+(p-2)\rho)\big(n\beta_i + n\rho\sum_{j\neq i}^{p}\beta_j + x_i'\epsilon\big) - \rho\sum_{j\neq i}^{p}\big(n\beta_j + n\rho\sum_{k=1,k\neq j}^{p}\beta_k + x_j'\epsilon\big)\,.$$

Using assumption A1 ($E(x_j'\epsilon) = 0$ for all $j = 1,\ldots,p$) and taking the expectation of $\hat{\beta}_i(\lambda^{FR})$ gives the following:

$$E[\hat{\beta}_i(\lambda^{FR})] = \frac{(1+\lambda^{FR}+(p-2)\rho)(n\beta_i + n\rho\sum_{j\neq i}^{p}\beta_j) - \rho\sum_{j\neq i}^{p}(n\beta_j + n\rho\sum_{k=1,k\neq j}^{p}\beta_k)}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}\,.$$

One can simplify the second expression of the numerator further,

$$\rho\sum_{j\neq i}^{p}(n\beta_j + n\rho\sum_{k=1,k\neq j}^{p}\beta_k) = \rho\big(n\rho(p-1)\beta_i + n\sum_{j\neq i}^{p}\beta_j + n\rho\sum_{j\neq i}^{p}\sum_{k=1,k\neq j}^{p}\beta_k\big)\,,$$

where one can show that $\sum_{j\neq i}^{p}\sum_{k=1,k\neq j}^{p}\beta_k = (p-2)\sum_{j\neq i}^{p}\beta_j$ resulting in the following:

$$\rho\sum_{j\neq i}^{p}(n\beta_j + n\rho\sum_{k=1,k\neq j}^{p}\beta_k) = \rho\big(n\rho(p-1)\beta_i + n(1+(p-2)\rho)\sum_{j\neq i}^{p}\beta_j\big)\,.$$

Therefore the estimator expectation can be written as

$$E[\hat{\beta}_i(\lambda^{FR})] = \frac{n(1+\lambda^{FR}+(p-2)\rho)(\beta_i+\rho\sum_{j\neq i}^{p}\beta_j)}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)} - \frac{n\rho\big(\rho(p-1)\beta_i+(1+(p-2)\rho)\sum_{j\neq i}^{p}\beta_j\big)}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)},$$

$$E[\hat{\beta}_i(\lambda^{FR})] = \frac{(1+\lambda^{FR}+(p-2)\rho)(\beta_i+\rho\sum_{j\neq i}^{p}\beta_j)}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)} - \frac{\rho\big(\rho(p-1)\beta_i+(1+(p-2)\rho)\sum_{j\neq i}^{p}\beta_j\big)}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)},$$

$$E[\hat{\beta}_i(\lambda^{FR})] = \frac{(1+\lambda^{FR}+(p-2)\rho-\rho^2(p-1))\beta_i+\rho\lambda^{FR}\sum_{j\neq i}^{p}\beta_j}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}.$$

To obtain the bias, the subtraction of β_i occurs as follows,

$$Bias(\hat{\beta}_i(\lambda^{FR})) = \frac{(1+\lambda^{FR}+(p-2)\rho-\rho^2(p-1))\beta_i+\rho\lambda^{FR}\sum_{j\neq i}^{p}\beta_j}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)} - \frac{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)\beta_i}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}.$$

Finally, since $(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho) =$
$(1+\lambda^{FR}+(p-2)\rho-\rho^2(p-1))+\lambda^{FR}(1+\lambda^{FR}+(p-2)\rho)$, the following expression is formed,

$$Bias(\hat{\beta}_i(\lambda^{FR})) = \frac{\rho\lambda^{FR}\sum_{j\neq i}^{p}\beta_j-\lambda^{FR}(1+\lambda^{FR}+(p-2)\rho)\beta_i}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}.$$

3.A.3.2 Partial Ridge Bias

Recall from 3.13 that:

$$\hat{\beta}_i(\lambda_i, S_i) = \frac{(1+\lambda_i+(p-2)\rho)x_i'y-\rho\sum_{j\neq i}^{p}x_j'y}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho)}.$$

Since the numerator of the Partial Ridge estimator is identical to that of Full Ridge (with λ^{FR} replaced with λ_i) the proof above can be used to short cut to the following

expression for the expectation of the Partial Ridge estimator,

$$E[\hat{\beta}_i(\lambda_i, S_i)] = \frac{n(1+\lambda_i+(p-2)\rho)(\beta_i+\rho\sum_{j\neq i}^{p}\beta_j)}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho)} - \frac{n\rho\big(\rho(p-1)\beta_i+(1+(p-2)\rho)\sum_{j\neq i}^{p}\beta_j\big)}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho)},$$

$$E[\hat{\beta}_i(\lambda_i, S_i)] = \frac{(1+\lambda_i+(p-2)\rho)(\beta_i+\rho\sum_{j\neq i}^{p}\beta_j)}{(1+\lambda_i+(p-2)\rho-(p-1)\rho)} - \frac{\rho\big(\rho(p-1)\beta_i+(1+(p-2)\rho)\sum_{j\neq i}^{p}\beta_j\big)}{(1+\lambda_i+(p-2)\rho-(p-1)\rho)},$$

$$E[\hat{\beta}_i(\lambda_i, S_i)] = \frac{(1+\lambda_i+(p-2)\rho-\rho^2(p-1))\beta_i+\rho\lambda_i\sum_{j\neq i}^{p}\beta_j}{(1+\lambda_i+(p-2)\rho-(p-1)\rho)}.$$

To obtain the bias, the subtraction of β_i occurs as follows,

$$Bias(\hat{\beta}_i(\lambda_i, S_i)) = \frac{(1+\lambda_i+(p-2)\rho-\rho^2(p-1))\beta_i+\rho\lambda_i\sum_{j\neq i}^{p}\beta_j}{(1+\lambda_i+(p-2)\rho-(p-1)\rho)} - \frac{(1+\lambda_i+(p-2)\rho-(p-1)\rho)\beta_i}{(1+\lambda_i+(p-2)\rho-(p-1)\rho)},$$

which simplifies to what is shown in 3.15,

$$Bias(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\rho\lambda_i\sum_{j\neq i}^{p}\beta_j}{(1+\lambda_i+(p-2)\rho-(p-1)\rho)}.$$

3.A.3.3 Full Ridge variance

Recall from the previous proofs that

$$\hat{\beta}_i(\lambda^{FR}) = \frac{(1+\lambda^{FR}+(p-2)\rho)\big(n\beta_i+n\rho\sum_{j\neq i}^{p}\beta_j+x_i'\epsilon\big)}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)} - \frac{\rho\sum_{j\neq i}^{p}\big(n\beta_j+n\rho\sum_{k=1,k\neq j}^{p}\beta_k+x_j'\epsilon\big)}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}.$$

Using the equivalence of $\sum_{j\neq i}^{p}\sum_{k=1,k\neq j}^{p}\beta_k = (p-2)\sum_{j\neq i}^{p}\beta_j$, the following expression can be obtained,

$$\hat{\beta}_i(\lambda^{FR}) = \frac{(1+\lambda^{FR}+(p-2)\rho)\left(n\beta_i + n\rho\sum_{j\neq i}^{p}\beta_j + x_i'\epsilon\right)}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)} - \frac{\rho\left(n\sum_{j\neq i}^{p}\beta_j + n\rho^2(p-2)\sum_{j\neq i}^{p}\beta_j + \sum_{j\neq i}^{p}x_j'\epsilon\right)}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}.$$

Taking the expectation and using A1 gives

$$E[\hat{\beta}_i(\lambda^{FR})] = \frac{(1+\lambda^{FR}+(p-2)\rho)\left(n\beta_i + n\rho\sum_{j\neq i}^{p}\beta_j\right)}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)} - \frac{\rho\left(n\sum_{j\neq i}^{p}\beta_j + n\rho^2(p-2)\sum_{j\neq i}^{p}\beta_j\right)}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}.$$

Now consider the following expression,

$$\hat{\beta}_i(\lambda^{FR}) - E[\hat{\beta}_i(\lambda^{FR})] = \frac{(1+\lambda^{FR}+(p-2)\rho)x_i'\epsilon - \rho\sum_{j\neq i}^{p}x_j'\epsilon}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)},$$

$$\hat{\beta}_i(\lambda^{FR}) - E[\hat{\beta}_i(\lambda^{FR})] = \frac{(1+\lambda^{FR}+(p-2)\rho)x_i'\epsilon - \rho x_1'\epsilon - \rho x_2'\epsilon - \cdots - \rho x_p'\epsilon}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}.$$

Since $Var(\hat{\beta}_i(\lambda^{FR})) = E[\left(\hat{\beta}_i(\lambda^{FR}) - E[\hat{\beta}_i(\lambda^{FR})]\right)\left(\hat{\beta}_i(\lambda^{FR}) - E[\hat{\beta}_i(\lambda^{FR})]\right)']$, consider the following expression,

$$\left(\hat{\beta}_i(\lambda^{FR}) - E[\hat{\beta}_i(\lambda^{FR})]\right)\left(\hat{\beta}_i(\lambda^{FR}) - E[\hat{\beta}_i(\lambda^{FR})]\right)' = \frac{(1+\lambda^{FR}+(p-2)\rho)x_i'\epsilon - \cdots - \rho x_p'\epsilon}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)} \times \frac{(1+\lambda^{FR}+(p-2)\rho)\epsilon' x_i - \cdots - \rho\epsilon' x_p}{n(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}.$$

The numerator of this expression can be expanded by multiplying terms together as follows,

$$\begin{aligned}
&(1+\lambda^{FR}+(p-2)\rho)^2 x_i'\epsilon\epsilon' x_i - \rho(1+\lambda^{FR}+(p-2)\rho)x_i'\epsilon\epsilon'\sum_{j\neq i}^{p}x_j \\
&\quad - \rho(1+\lambda^{FR}+(p-2)\rho)x_1'\epsilon\epsilon' x_i + \rho^2 x_1'\epsilon\epsilon' x_1 + \cdots + \rho^2 x_1'\epsilon\epsilon' x_p \\
&\quad - \rho(1+\lambda^{FR}+(p-2)\rho)x_2'\epsilon\epsilon' x_i + \rho^2 x_2'\epsilon\epsilon' x_1 + \cdots + x_2'\epsilon\epsilon' x_p \\
&\quad + \cdots + \\
&\quad - \rho(1+\lambda^{FR}+(p-2)\rho)x_p'\epsilon\epsilon' x_i + \rho^2 x_p'\epsilon\epsilon' x_1 + \cdots + \rho^2 x_p'\epsilon\epsilon' x_p.
\end{aligned}$$

Taking the expectation gives

$$\begin{aligned}(1+\lambda^{FR}+(p-2)\rho)^2 x_i' E(\epsilon\epsilon')x_i - \rho(1+\lambda^{FR}+(p-2)\rho)x_i' E(\epsilon\epsilon')\sum_{j\neq i}^{p} x_j \\ - \rho(1+\lambda^{FR}+(p-2)\rho)x_1' E(\epsilon\epsilon')x_i + \rho^2 x_1' E(\epsilon\epsilon')x_1 + \cdots + \rho^2 x_1' E(\epsilon\epsilon')x_p \\ - \rho(1+\lambda^{FR}+(p-2)\rho)x_2' E(\epsilon\epsilon')x_i + \rho^2 x_2' E(\epsilon\epsilon')x_1 + \cdots + x_2' E(\epsilon\epsilon')x_p \\ + \cdots + \\ - \rho(1+\lambda^{FR}+(p-2)\rho)x_p' E(\epsilon\epsilon')x_i + \rho^2 x_p'(\epsilon\epsilon')x_1 + \cdots + \rho^2 x_p' E(\epsilon\epsilon')x_p\,.\end{aligned}$$

Now using the fact that $E[\epsilon\epsilon'] = \sigma_\epsilon^2 I$ from A1 as well as $x_j' x_j = n$ and $x_j' x_h = n\rho$ from A3, the numerator can be simplified to the following:

$$\begin{aligned}\sigma_\epsilon^2 \Bigg(& n(1+\lambda^{FR}+(p-2)\rho)^2 - n\rho^2(p-1)(1+\lambda^{FR}+(p-2)\rho) \\ & - n\rho^2(p-1)(1+\lambda^{FR}+(p-2)\rho) + n(p-1)\rho^2 + n\rho^3(p-1)(p-2)\Bigg),\end{aligned}$$

where the fact that $x_i' \sum_{j\neq i}^{p} x_j = n(p-1)\rho$ was used in the first line. The variance term is as follows,

$$\begin{aligned}& Var(\hat{\beta}_i(\lambda^{FR})) = E[\left(\hat{\beta}_i(\lambda^{FR}) - E[\hat{\beta}_i(\lambda^{FR})]\right)\left(\hat{\beta}_i(\lambda^{FR}) - E[\hat{\beta}_i(\lambda^{FR})]\right)'] = \\ & \frac{\sigma_\epsilon^2\left((1+\lambda^{FR}+(p-2)\rho)^2 - 2\rho^2(p-1)(1+\lambda^{FR}+(p-2)\rho) + (p-1)\rho^2\left(1+\rho(p-2)\right)\right)}{n(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+(p-1)\rho)^2}.\end{aligned}$$

Attempting to factorise the numerator leads to the following:

$$\begin{aligned}& Var(\hat{\beta}_i(\lambda^{FR})) = \\ & \frac{\sigma_\epsilon^2\left(\left(1+\lambda^{FR}+(p-2)\rho-\rho^2(p-1)\right)^2 - \rho^4(p-1)^2 + (p-1)\rho^2 + \rho^3(p-1)(p-2)\right)}{n(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+(p-1)\rho)^2}.\end{aligned}$$

One further stage of simplification leads to the expression in Proposition 3.3 as follows,

$$\begin{aligned}& Var(\hat{\beta}_i(\lambda^{FR})) = \\ & \frac{\sigma_\epsilon^2\left(\left(\lambda^{FR}-(\rho-1)(1+(p-1)\rho)\right)^2 - (p-1)(\rho-1)\rho^2(1+(p-1)\rho)\right)}{n(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+(p-1)\rho)^2}.\end{aligned}$$

3.A.3.4 Partial Ridge variance

Recall from the previous proofs that

$$\hat{\beta}_i(\lambda_i, S_i) = \frac{(1+\lambda_i+(p-2)\rho)\left(n\beta_i + n\rho\sum_{j\neq i}^{p}\beta_j + x_1'\epsilon\right)}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)} - \frac{\rho\sum_{j\neq i}^{p}\left(n\beta_j + n\rho\sum_{k=1,k\neq j}^{p}\beta_k + x_j'\epsilon\right)}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)}.$$

Since the numerator is identical to that of Full Ridge with the exception of λ^{FR} being replaced with λ_i, one can skip to the following expression using the previous proof but using the Partial Ridge denominator,

$$\hat{\beta}_i(\lambda_i, S_i) - E[\hat{\beta}_i(\lambda_i, S_i)] = \frac{(1+\lambda_i+(p-2)\rho)x_i'\epsilon - \rho\sum_{j\neq i}^{p}x_j'\epsilon}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)},$$

$$\hat{\beta}_i(\lambda_i, S_i) - E[\hat{\beta}_i(\lambda_i, S_i)] = \frac{(1+\lambda_i+(p-2)\rho)x_i'\epsilon - \rho x_1'\epsilon - \rho x_2'\epsilon - \cdots - \rho x_p'\epsilon}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)}.$$

Since $Var(\hat{\beta}_i(\lambda_i, S_i)) = E[\left(\hat{\beta}_i(\lambda_i, S_i) - E[\hat{\beta}_i(\lambda_i, S_i)]\right)\left(\hat{\beta}_i(\lambda_i, S_i) - E[\hat{\beta}_i(\lambda_i, S_i)]\right)']$, consider the following expression,

$$\left(\hat{\beta}_i(\lambda_i, S_i) - E[\hat{\beta}_i(\lambda_i, S_i)]\right)\left(\hat{\beta}_i(\lambda_i, S_i) - E[\hat{\beta}_i(\lambda_i, S_i)]\right)' = \frac{(1+\lambda_i+(p-2)\rho)x_i'\epsilon - \cdots - \rho x_p'\epsilon}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)} \times \frac{(1+\lambda_i+(p-2)\rho)\epsilon' x_i - \cdots - \rho\epsilon' x_p}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)}.$$

From here, one can use the fact that the previous proof showed that the numerator can be written as follows,

$$\sigma_\epsilon^2\Big(n(1+\lambda_i+(p-2)\rho)^2 - n\rho^2(p-1)(1+\lambda_i+(p-2)\rho) - n\rho^2(p-1)(1+\lambda_i+(p-2)\rho) + n(p-1)\rho^2 + n\rho^3(p-1)(p-2)\Big).$$

Therefore, the variance expression can be written as

$$Var(\hat{\beta}_i(\lambda_i)) = E[\left(\hat{\beta}_i(\lambda_i, S_i) - E[\hat{\beta}_i(\lambda_i, S_i)]\right)\left(\hat{\beta}_i(\lambda_i, S_i) - E[\hat{\beta}_i(\lambda_i, S_i)]\right)'] = \frac{\sigma_\epsilon^2\Big((1+\lambda_i+(p-2)\rho)^2 - 2\rho^2(p-1)(1+\lambda_i+(p-2)\rho) + (p-1)\rho^2\left(1+\rho(p-2)\right)\Big)}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)^2}.$$

Simplifying to the following in a very similar fashion to that of the previous proof,

$$Var(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\sigma_\epsilon^2 \left(\left(\lambda_i - (\rho-1)(1+(p-1)\rho)\right)^2 - (p-1)(\rho-1)\rho^2(1+(p-1)\rho) \right)}{n(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)^2}.$$

3.A.4 Alternative Proof of Proposition 3.3

Using the results associated with Proposition 3.1, recall the Full Ridge bias expression from the proof of Proposition 3.1 as follows,

$$Bias(\hat{\beta}_i(\lambda^{FR})) = \frac{\lambda^{FR}\tilde{x}_i' S(S^2+\lambda^{FR}I_{p-1})^{-1}\beta^*_{-i} - \lambda^{FR}\beta_i}{1-\tilde{x}_i' S^2(S^2+\lambda^{FR}I_{p-1})^{-1}\tilde{x}_i + \lambda^{FR}}.$$

Using the fact that $\tilde{x}_i = \frac{U'x_i}{\sqrt{n}}$ and $\beta^*_{-i} = V'\beta_{-i}$, the numerator can be written as follows,

$$\lambda^{FR}\tilde{x}_i' S(S^2+\lambda^{FR}I_{p-1})^{-1}\beta^*_{-i} - \lambda^{FR}\beta_i = \\ -\lambda^{FR}[\beta_i - \frac{x_i'}{\sqrt{n}}USV'(V')^{-1}(S^2+\lambda^{FR}I_{p-1})^{-1}V'\beta_{-i}].$$

In order to simplify further, first consider the partitioned covariance matrix from previously,

$$\frac{X'X}{n} = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix} = \begin{bmatrix} \frac{x_i'x_i}{n} & \frac{x_i'X_{-i}}{n} \\ \frac{X_{-i}'x_i}{n} & \frac{X_{-i}'X_{-i}}{n} \end{bmatrix} = \begin{bmatrix} 1 & \frac{x_i'}{\sqrt{n}}USV' \\ VSU'\frac{x_i'}{\sqrt{n}} & VS^2V' \end{bmatrix}.$$

From this one can see that $\frac{x_i'}{\sqrt{n}}USV' = \rho e'$, were e is a $(p-1)\times 1$ vector of 1s.
Therefore, the numerator of the bias can be written as follows,

$$-\lambda^{FR}[\beta_i - \rho e'V(S^2+\lambda^{FR}I_{p-1})^{-1}V'\beta_{-i}].$$

Now, the denominator can be rewritten as follows,

$$1-\tilde{x}_i' S^2(S^2+\lambda^{FR}I_{p-1})^{-1}\tilde{x}_i + \lambda^{FR} = 1+\lambda^{FR} - \frac{x_i'}{\sqrt{n}}US^2(S^2+\lambda^{FR}I_{p-1})^{-1}U'\frac{x_i}{\sqrt{n}}.$$

Using the rearrangement of $\frac{X_{-i}}{\sqrt{n}} = USV'$ as $U = \frac{X_{-i}VS^{-1}}{\sqrt{n}}$, gives the following:

$$= 1 + \lambda^{FR} - \frac{x_i'}{\sqrt{n}}\frac{X_{-i}}{\sqrt{n}}VS^{-1}S^2(S^2 + \lambda^{FR}I_{p-1})^{-1}S^{-1}V'\frac{X_{-i}'}{\sqrt{n}}\frac{x_i}{\sqrt{n}},$$

$$1 + \lambda^{FR} - \rho e'VS(S^2 + \lambda^{FR}I_{p-1})^{-1}S^{-1}V'\rho e,$$

$$1 + \lambda^{FR} - \rho e'V(S^2 + \lambda^{FR}I_{p-1})^{-1}V'\rho e.$$

Therefore, the bias of Full Ridge can be expressed as follows,

$$Bias(\hat{\beta}_i(\lambda^{FR})) = \frac{-\lambda^{FR}[\beta_i - \rho e'V(S^2 + \lambda^{FR}I_{p-1})^{-1}V'\beta_{-i}]}{1 + \lambda^{FR} - \rho e'V(S^2 + \lambda^{FR}I_{p-1})^{-1}V'\rho e}.$$

For the Variance, the proof of Proposition 3.1 resulted in the following:

$$Var(\hat{\beta}_i(\lambda^{FR})) = \frac{\sigma_\epsilon^2}{n}\frac{1 - \tilde{x}_i'S^2(S^2 + 2\lambda^{FR}I_{p-1})(S^2 + \lambda^{FR}I_{p-1})^{-2}\tilde{x}_i}{(1 - \tilde{x}_i'S^2(S^2 + \lambda^{FR}I_{p-1})^{-1}\tilde{x}_i + \lambda^{FR})^2}.$$

For the numerator, using the $\tilde{x}_i$ expansion results in the following expression,

$$1 - \tilde{x}_i'S^2(S^2 + 2\lambda^{FR}I_{p-1})(S^2 + \lambda^{FR}I_{p-1})^{-2}\tilde{x}_i = 1 - \frac{x_i'}{\sqrt{n}}US^2(S^2 + 2\lambda^{FR}I_{p-1})(S^2 + \lambda^{FR}I_{p-1})^{-2}U'\frac{x_i}{\sqrt{n}},$$

$$= 1 - \frac{x_i'X_{-i}}{n}VS^{-1}S^2(S^2 + 2\lambda^{FR}I_{p-1})(S^2 + \lambda^{FR}I_{p-1})^{-2}S^{-1}V'\frac{X_{-i}'x_i}{n},$$

$$= 1 - \rho e'V(S^2 + 2\lambda^{FR}I_{p-1})(S^2 + \lambda^{FR}I_{p-1})^{-2}V'\rho e.$$

For the denominator component inside squared brackets, the following is considered,

$$1 - \tilde{x}_i'S^2(S^2 + \lambda^{FR}I_{p-1})^{-1}\tilde{x}_i + \lambda^{FR} = 1 + \lambda^{FR} - \frac{x_i'}{\sqrt{n}}US^2(S^2 + \lambda^{FR}I_{p-1})^{-1}U'\frac{x_i}{\sqrt{n}},$$

$$= 1 + \lambda^{FR} - \frac{x_i'X_{-i}}{n}VS^{-1}S^2(S^2 + \lambda^{FR}I_{p-1})^{-1}S^{-1}V'\frac{X_{-i}'x_i}{n},$$

$$= 1 + \lambda^{FR} - \rho e'V(S^2 + \lambda^{FR}I_{p-1})^{-1}V'\rho e.$$

Therefore, the variance term can be written as

$$Var(\hat{\beta}_i(\lambda^{FR})) = \frac{\sigma_\epsilon^2}{n}\frac{1 - \rho e'V(S^2 + 2\lambda^{FR}I_{p-1})(S^2 + \lambda^{FR}I_{p-1})^{-2}V'\rho e}{(1 + \lambda^{FR} - \rho e'V(S^2 + \lambda^{FR}I_{p-1})^{-1}V'\rho e)^2}.$$

For Partial Ridge the bias is given by

$$Bias(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\lambda_i \tilde{x}_i' S(S^2 + \lambda_i I_{p-1})^{-1} \beta^*_{-i}}{1 - \tilde{x}_i' S^2 (S^2 + \lambda_i I_{p-1})^{-1} \tilde{x}_i}.$$

The numerator can be rewritten as follows simply, as it is identical to that of the Full Ridge case only with λ^{FR} replaced with λ_i and the elimination of the $\lambda^{FR}\beta_i$ term,

$$\lambda_i \tilde{x}_i' S(S^2 + \lambda_i I_{p-1})^{-1} \beta^*_{-i} = \lambda_i \rho e' V (S^2 + \lambda_i I_{p-1})^{-1} V' \beta_{-i}.$$

For the denominator, it is a similar situation whereby the Partial Ridge case is identical to that of Full Ridge, only with λ^{FR} replaced with λ_i and the elimination of the λ^{FR} component,

$$1 - \tilde{x}_i' S^2 (S^2 + \lambda_i I_{p-1})^{-1} \tilde{x}_i = 1 - \rho e' V (S^2 + \lambda_i I_{p-1})^{-1} V' \rho e.$$

Therefore, the Partial Ridge bias term is given by the following expression,

$$Bias(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\lambda_i \rho e' V (S^2 + \lambda_i I_{p-1})^{-1} V' \beta_{-i}}{1 - \rho e' V (S^2 + \lambda_i I_{p-1})^{-1} V' \rho e}.$$

For the Variance, consider the expression arising in the proof of Proposition 3.1 as follows,

$$Var(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\sigma_\epsilon^2}{n} \Big[\frac{1 - \tilde{x}_i' S^2 (S^2 + 2\lambda_i I_{p-1})(S^2 + \lambda_i I_{p-1})^{-2} \tilde{x}_i}{(1 - \tilde{x}_i' S^2 (S^2 + \lambda_i I_{p-1})^{-1} \tilde{x}_i)^2} \Big].$$

Once again, the numerator,is identical to that of Full Ridge, only with λ^{FR} replaced with λ_i resulting in the following:

$$1 - \tilde{x}_i' S^2 (S^2 + 2\lambda_i I_{p-1})(S^2 + \lambda_i I_{p-1})^{-2} \tilde{x}_i = 1 - \rho e' V (S^2 + 2\lambda_i I_{p-1})(S^2 + \lambda_i I_{p-1})^{-2} V' \rho e.$$

For the denominator, the following expression can be obtained by using the usual similarities with the Full Ridge variance expression,

$$1 - \tilde{x}_i' S^2 (S^2 + \lambda_i I_{p-1})^{-1} \tilde{x}_i = 1 - \rho e' V (S^2 + \lambda_i I_{p-1})^{-1} V' \rho e.$$

Therefore the Partial Ridge variance is given by the following:

$$Var(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\sigma_\epsilon^2}{n} \frac{1 - \rho e' V (S^2 + 2\lambda_i I_{p-1})(S^2 + \lambda_i I_{p-1})^{-2} V' \rho e}{(1 - \rho e' V (S^2 + \lambda_i I_{p-1})^{-1} V' \rho e)^2}.$$

To formalise the above, the table shows the alternate forms for the bais and variance of a single Full and Partial Ridge estimate.

	Bias	Variance
Full Ridge	$\frac{-\lambda^{FR}[\beta_i - \rho e' V(S^2+\lambda^{FR} I_{p-1})^{-1} V' \beta_{-i}]}{1+\lambda^{FR} - \rho e' V(S^2+\lambda^{FR} I_{p-1})^{-1} V' \rho e}$	$\frac{\sigma_\epsilon^2}{n} \frac{1-\rho e' V(S^2+2\lambda^{FR} I_{p-1})(S^2+\lambda^{FR} I_{p-1})^{-2} V' \rho e}{(1+\lambda^{FR} - \rho e' V(S^2+\lambda^{FR} I_{p-1})^{-1} V' \rho e)^2}$
Partial Ridge	$\frac{\lambda_i \rho e' V(S^2+\lambda_i I_{p-1})^{-1} V' \beta_{-i}}{1-\rho e' V(S^2+\lambda_i I_{p-1})^{-1} V' \rho e}$	$\frac{\sigma_\epsilon^2}{n} \frac{1-\rho e' V(S^2+2\lambda_i I_{p-1})(S^2+\lambda_i I_{p-1})^{-2} V' \rho e}{(1-\rho e' V(S^2+\lambda_i I_{p-1})^{-1} V' \rho e)^2}$

For this equicorrelation environment, various values of p will be considered with each case involving a specific S and V matrix (as well as specialising the dimension of e) allowing the expressions in the table above to be evaluated.

Case 1: p=3

In this case, it is such that X_{-i} is a 2×2 matrix with the following SVD components (assuming $\rho > 0$),

$$S^2 = \begin{bmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{bmatrix},$$

$$V = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

First, consider the Full Ridge bias. The main component of the numerator can be written as follows,

$$\lambda^{FR} \rho e' V(S^2 + \lambda^{FR} I_{p-1})^{-1} V' \beta_{-i} = \begin{bmatrix} \rho & \rho \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{1+\rho+\lambda^{FR}} & 0 \\ 0 & \frac{1}{1-\rho+\lambda^{FR}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \beta_{-i}.$$

Multiplying out gives

$$\lambda^{FR} \rho e' V(S^2 + \lambda^{FR} I_{p-1})^{-1} V' \beta_{-i} = \frac{\lambda^{FR} \rho}{1 + \lambda^{FR} + \rho} \sum_{j \neq i} \beta_j.$$

Therefore, the bias numerator is given by the following expression,

$$\frac{\lambda^{FR} \rho}{1 + \lambda^{FR} + \rho} \sum_{j \neq i} \beta_j - \lambda^{FR} \beta_i.$$

For the denominator, recall the following,

$$1 + \lambda^{FR} - \rho e' V(S^2 + \lambda^{FR} I_{p-1})^{-1} V' \rho e.$$

The main component can be written as follows,

$$\rho e' V(S^2 + \lambda^{FR} I_{p-1})^{-1} V' \rho e = \begin{bmatrix} \rho & \rho \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{1+\rho+\lambda^{FR}} & 0 \\ 0 & \frac{1}{1-\rho+\lambda^{FR}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \rho \\ \rho \end{bmatrix}.$$

$$\rho e' V(S^2 + \lambda^{FR} I_{p-1})^{-1} V' \rho e = \frac{2\rho^2}{1 + \lambda^{FR} + \rho}$$

Therefore the denominator is given by the following,

$$1 + \lambda^{FR} - \frac{2\rho^2}{1 + \lambda^{FR} + \rho}.$$

Combining this with the numerator term allows one to compute the Full Ridge bias as follows,

$$Bias(\hat{\beta}_i(\lambda^{FR})) = \frac{\frac{\lambda^{FR}\rho}{1+\lambda^{FR}+\rho} \sum_{j \neq i} \beta_j - \lambda^{FR}\beta_i}{1 + \lambda^{FR} - \frac{2\rho^2}{1+\lambda^{FR}+\rho}} = \frac{\lambda^{FR}\rho \sum_{j \neq i} \beta_j - \lambda^{FR}\beta_i(1 + \lambda^{FR} + \rho)}{(1 + \lambda^{FR} - \rho)(1 + \lambda^{FR} + 2\rho)}.$$

For the variance, recall the following numerator expression,

$$1 - \rho e' V(S^2 + 2\lambda^{FR} I_{p-1})(S^2 + \lambda^{FR} I_{p-1})^{-2} V' \rho e.$$

This can be written as follows,

$$1 - \begin{bmatrix} \rho & \rho \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1+\rho+2\lambda^{FR}}{(1+\rho+\lambda^{FR})^2} & 0 \\ 0 & \frac{1-\rho+2\lambda^{FR}}{(1-\rho+\lambda^{FR})^2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \rho \\ \rho \end{bmatrix}.$$

Multiplying out gives the following expression for the numerator,

$$1 - \rho e' V(S^2 + 2\lambda^{FR} I_{p-1})(S^2 + \lambda^{FR} I_{p-1})^{-2} V' \rho e = 1 - \frac{2\rho^2(1 + 2\lambda^{FR} + \rho)}{(1 + \lambda^{FR} + \rho)^2}.$$

For the denominator, recall the following expression,

$$(1 + \lambda^{FR} - \rho e' V(S^2 + \lambda^{FR} I_{p-1})^{-1} V' \rho e)^2,$$

where this is the denominator of the bias but squared leading to the following:

$$(1 + \lambda^{FR} - \rho e' V(S^2 + \lambda^{FR} I_{p-1})^{-1} V' \rho e)^2 = \big(1 + \lambda^{FR} - \frac{2\rho^2}{1 + \lambda^{FR} + \rho}\big)^2.$$

Therefore, the variance can be written as follows,

$$Var(\hat{\beta}_i(\lambda^{FR})) = \frac{\sigma_\epsilon^2}{n} \frac{1 - \frac{2\rho^2(1+2\lambda^{FR}+\rho)}{(1+\lambda^{FR}+\rho)^2}}{\left(1+\lambda^{FR} - \frac{2\rho^2}{1+\lambda^{FR}+\rho}\right)^2} = \frac{\sigma_\epsilon^2}{n} \frac{(1+\lambda^{FR}+\rho)^2 - 2\rho^2(1+2\lambda^{FR}+\rho)}{\left((1+\lambda^{FR})(1+\lambda^{FR}+\rho) - 2\rho^2\right)^2},$$

which can be rearranged to obtain the following:

$$Var(\hat{\beta}_i(\lambda^{FR})) = \frac{\sigma_\epsilon^2}{n} \frac{\left(\lambda^{FR} - (\rho-1)(1+2\rho)\right)^2 - 2(\rho-1)\rho^2(1+2\rho)}{(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+2\rho)^2}.$$

For Partial Ridge, recall the bias numerator expression,

$$1 - \rho e' V (S^2 + \lambda_i I_{p-1})^{-1} V' \rho e,$$

where the numerator is almost identical to that of the Full Ridge case with the elimination of $\lambda^{FR}\beta_i$ and λ^{FR} replaced with λ_i resulting in the following:

$$\lambda_i \rho e' V (S^2 + \lambda_i I_{p-1})^{-1} V' \beta_{-i} = \frac{\lambda_i \rho \sum_{j \neq i} \beta_j}{1+\lambda_i+\rho}.$$

For the denominator, it is also very similar to that of Full Ridge leading to the following:

$$1 - \rho e' V (S^2 + \lambda_i I_{p-1})^{-1} V' \rho e = 1 - \frac{2\rho^2}{1+\lambda_i+\rho}.$$

Therefore, the Partial Ridge bias can be written as follows,

$$Bias(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\frac{\lambda_i \rho \sum_{j \neq i} \beta_j}{1+\lambda_i+\rho}}{1 - \frac{2\rho^2}{1+\lambda_i+\rho}} = \frac{\lambda_i \rho \sum_{j \neq i} \beta_j}{1+\lambda_i+\rho-2\rho^2}.$$

For the variance, recall the following numerator expression as being identical to that of Full Ridge but with different λ notation. This leads to the following:

$$1 - \rho e' V (S^2 + 2\lambda_i I_{p-1})(S^2 + \lambda_i I_{p-1})^{-2} V' \rho e = 1 - \frac{2\rho^2(1+2\lambda_i+\rho)}{(1+\lambda_i+\rho)^2}.$$

For the denominator, once again, the Partial Ridge expression is identical to that of Full Ridge with the elimination of the λ^{FR} term and λ_i instead of λ^{FR},

$$(1 - \rho e' V (S^2 + \lambda_i I_{p-1})^{-1} V' \rho e)^2 = \left(1 - \frac{2\rho^2}{1+\lambda_i+\rho}\right)^2.$$

Therefore, the Partial Ridge variance can be written as follows,

$$Var(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\sigma_\epsilon^2}{n} \frac{1 - \frac{2\rho^2(1+2\lambda_i+\rho)}{(1+\lambda_i+\rho)^2}}{\left(1 - \frac{2\rho^2}{1+\lambda_i+\rho}\right)^2} = \frac{\sigma_\epsilon^2}{n} \frac{(1+\lambda_i+\rho)^2 - 2\rho^2(1+2\lambda_i+\rho)}{(1+\lambda_i+\rho-2\rho^2)^2},$$

which can be simplified to the following:

$$Var(\hat{\beta}_i(\lambda_i, S_i)) = \frac{\sigma_\epsilon^2}{n} \frac{\left(\lambda_i - (\rho-1)(1+2\rho)\right)^2 - 2(\rho-1)\rho^2(1+2\rho)}{(1+\lambda_i+\rho-2\rho^2)^2}.$$

Case 2: p=4

In this case, it is such that X_{-i} is a 3×3 matrix with the following SVD components (assuming $\rho > 0$),

$$S^2 = \begin{bmatrix} 1+2\rho & 0 & 0 \\ 0 & 1-\rho & 0 \\ 0 & 0 & 1-\rho \end{bmatrix},$$

$$V = \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \end{bmatrix}.$$

First, consider the Full Ridge bias. The main component of the numerator can be written as follows,

$$\lambda^{FR}\rho e' V(S^2 + \lambda^{FR} I_{p-1})^{-1} V' \beta_{-i} =$$
$$\begin{bmatrix} \rho & \rho & \rho \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} \frac{1}{1+2\rho+\lambda^{FR}} & 0 & 0 \\ 0 & \frac{1}{1-\rho+\lambda^{FR}} & 0 \\ 0 & 0 & \frac{1}{1-\rho} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & \sqrt{\frac{2}{3}} & -\frac{1}{\sqrt{6}} \end{bmatrix} \beta_{-i}.$$

Multiplying out gives

$$\lambda^{FR}\rho e' V(S^2 + \lambda^{FR} I_{p-1})^{-1} V' \beta_{-i} = \frac{\lambda^{FR}\rho}{\rho 1 + \lambda^{FR} + 2\rho} \sum_{j \neq i} \beta_j.$$

Therefore, the bias numerator is given by the following:

$$\frac{\lambda^{FR}\rho}{1 + \lambda^{FR} + 2\rho} \sum_{j \neq i} \beta_j - \lambda^{FR}\beta_i.$$

For the denominator, recall the following:

$$1 + \lambda^{FR} - \rho e' V (S^2 + \lambda^{FR} I_{p-1})^{-1} V' \rho e \,.$$

The main component can be written as follows,

$$\rho e' V (S^2 + \lambda^{FR} I_{p-1})^{-1} V' \rho e =$$
$$\begin{bmatrix} \rho & \rho & \rho \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} \frac{1}{1+2\rho+\lambda^{FR}} & 0 & 0 \\ 0 & \frac{1}{1-\rho+\lambda^{FR}} & 0 \\ 0 & 0 & \frac{1}{1-\rho} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & \sqrt{\frac{2}{3}} & -\frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} \rho \\ \rho \\ \rho \end{bmatrix},$$

$$\rho e' V (S^2 + \lambda^{FR} I_{p-1})^{-1} V' \rho e = \frac{3\rho^2}{1 + \lambda^{FR} + 2\rho} \,.$$

Therefore, the denominator is given by the following:

$$1 + \lambda^{FR} - \frac{3\rho^2}{1 + \lambda^{FR} + 2\rho} \,.$$

Combining this with the numerator term allows one to compute the Full Ridge bias as follows,

$$Bias(\hat{\beta}_i(\lambda^{FR})) = \frac{\frac{\lambda^{FR}\rho}{1+\lambda^{FR}+2\rho} \sum_{j \neq i} \beta_j - \lambda^{FR} \beta_i}{1 + \lambda^{FR} - \frac{3\rho^2}{1+\lambda^{FR}+2\rho}} = \frac{\lambda^{FR}\rho \sum_{j \neq i} \beta_j - \lambda^{FR} \beta_i (1 + \lambda^{FR} + 2\rho)}{(1 + \lambda^{FR} - \rho)(1 + \lambda^{FR} + 3\rho)} \,.$$

For the variance, recall the following numerator expression,

$$1 - \rho e' V (S^2 + 2\lambda^{FR} I_{p-1})(S^2 + \lambda^{FR} I_{p-1})^{-2} V' \rho e \,.$$

This can be written as follows,

$$1 - \begin{bmatrix} \rho & \rho & \rho \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} \frac{1+2\rho+2\lambda^{FR}}{(1+2\rho+\lambda^{FR})^2} & 0 & 0 \\ 0 & \frac{1-\rho+2\lambda^{FR}}{(1-\rho+\lambda^{FR})^2} & 0 \\ 0 & 0 & \frac{1-\rho+2\lambda^{FR}}{(1-\rho+\lambda^{FR})^2} \end{bmatrix}$$
$$\times \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & \sqrt{\frac{2}{3}} & -\frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} \rho \\ \rho \\ \rho \end{bmatrix}.$$

Multiplying out gives the following expression for the numerator,

$$1 - \rho e' V (S^2 + 2\lambda^{FR} I_{p-1})(S^2 + \lambda^{FR} I_{p-1})^{-2} V' \rho e = 1 - \frac{3\rho^2 (1 + 2\lambda^{FR} + 2\rho)}{(1 + \lambda^{FR} + 2\rho)^2} \,.$$

For the denominator, recall the following expression,

$$(1+\lambda^{FR}-\rho e'V(S^2+\lambda^{FR}I_{p-1})^{-1}V'\rho e)^2\,,$$

where this is the denominator of the bias but squared leading to the following:

$$(1+\lambda^{FR}-\rho e'V(S^2+\lambda^{FR}I_{p-1})^{-1}V'\rho e)^2=\big(1+\lambda^{FR}-\frac{3\rho^2}{1+\lambda^{FR}+2\rho}\big)^2.$$

Therefore, the variance can be written as follows,

$$Var(\hat{\beta}_i(\lambda^{FR}))=\frac{\sigma_\epsilon^2}{n}\frac{1-\frac{3\rho^2(1+2\lambda^{FR}+2\rho)}{(1+\lambda^{FR}+2\rho)^2}}{\big(1+\lambda^{FR}-\frac{3\rho^2}{1+\lambda^{FR}+2\rho}\big)^2}=\frac{\sigma_\epsilon^2}{n}\frac{(1+\lambda^{FR}+2\rho)^2-3\rho^2(1+2\lambda^{FR}+2\rho)}{\big((1+\lambda^{FR})(1+\lambda^{FR}+2\rho)-3\rho^2\big)^2}\,,$$

which can be rearranged to obtain the following:

$$Var(\hat{\beta}_i(\lambda^{FR}))=\frac{\sigma_\epsilon^2}{n}\frac{\big(\lambda^{FR}-(\rho-1)(1+3\rho)\big)^2-3(\rho-1)\rho^2(1+3\rho)}{(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+3\rho)^2}\,.$$

For Partial Ridge, recall the bias numerator expression,

$$1-\rho e'V(S^2+\lambda_i I_{p-1})^{-1}V'\rho e\,,$$

where the numerator is almost identical to that of the Full Ridge case with the elimination of $\lambda^{FR}\beta_i$ and λ^{FR} replaced with λ_i resulting in the following:

$$\lambda_i\rho e'V(S^2+\lambda_i I_{p-1})^{-1}V'\beta_{-i}=\frac{\lambda_i\rho\sum_{j\neq i}\beta_j}{1+\lambda_i+2\rho}\,.$$

For the denominator, it is also very similar to that of Full Ridge leading to the following:

$$1-\rho e'V(S^2+\lambda_i I_{p-1})^{-1}V'\rho e=1-\frac{3\rho^2}{1+\lambda_i+2\rho}\,.$$

Therefore, the Partial Ridge bias can be written as follows,

$$Bias(\hat{\beta}_i(\lambda_i,S_i))=\frac{\frac{\lambda_i\rho\sum_{j\neq i}\beta_j}{1+\lambda_i+2\rho}}{1-\frac{3\rho^2}{1+\lambda_i+2\rho}}=\frac{\lambda_i\rho\sum_{j\neq i}\beta_j}{1+\lambda_i+2\rho-3\rho^2}\,.$$

For the variance, recall the following numerator expression as being identical to that of Full Ridge but with different λ notation. This leads to the following:

$$1-\rho e'V(S^2+2\lambda_i I_{p-1})(S^2+\lambda_i I_{p-1})^{-2}V'\rho e=1-\frac{3\rho^2(1+2\lambda_i+2\rho)}{(1+\lambda_i+2\rho)^2}\,.$$

For the denominator, once again, the Partial Ridge expression is identical to that of Full Ridge with the elimination of the λ^{FR} term and λ_i instead of λ^{FR}, seen as follows,

$$(1-\rho e' V(S^2+\lambda_i I_{p-1})^{-1}V'\rho e)^2 = \big(1-\frac{3\rho^2}{1+\lambda_i+2\rho}\big)^2.$$

Therefore, the Partial Ridge variance can be written as follows,

$$Var(\hat{\beta}_i(\lambda_i,S_i)) = \frac{\sigma_\epsilon^2}{n}\frac{1-\frac{3\rho^2(1+2\lambda_i+2\rho)}{(1+\lambda_i+2\rho)^2}}{\big(1-\frac{3\rho^2}{1+\lambda_i+2\rho}\big)^2} = \frac{\sigma_\epsilon^2}{n}\frac{(1+\lambda_i+2\rho)^2-3\rho^2(1+2\lambda_i+2\rho)}{(1+\lambda_i+2\rho-3\rho^2)^2},$$

which can be simplified to the following:

$$Var(\hat{\beta}_i(\lambda_i,S_i)) = \frac{\sigma_\epsilon^2}{n}\frac{\big(\lambda_i-(\rho-1)(1+3\rho)\big)^2-3(\rho-1)\rho^2(1+3\rho)}{(1+\lambda_i+2\rho-3\rho^2)^2}.$$

To formalise the above, the following table shows the bias and variance for each p under equicorrelation where one can see a general pattern that allows the expressions to be generalized with respect to p.

Full Ridge		
	Bias	Variance
$p=3$	$\frac{\lambda^{FR}\rho\sum_{j\neq i}\beta_j-\lambda^{FR}\beta_i(1+\lambda^{FR}+\rho)}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+2\rho)}$	$\frac{\sigma_\epsilon^2}{n}\frac{\left(\lambda^{FR}-(\rho-1)(1+2\rho)\right)^2-2(\rho-1)\rho^2(1+2\rho)}{(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+2\rho)^2}$
$p=4$	$\frac{\lambda^{FR}\rho\sum_{j\neq i}\beta_j-\lambda^{FR}\beta_i(1+\lambda^{FR}+2\rho)}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+3\rho)}$	$\frac{\sigma_\epsilon^2}{n}\frac{\left(\lambda^{FR}-(\rho-1)(1+3\rho)\right)^2-3(\rho-1)\rho^2(1+3\rho)}{(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+3\rho)^2}$
$\vdots$		
p	$\frac{\lambda^{FR}\rho\sum_{j\neq i}\beta_j-\lambda^{FR}\beta_i(1+\lambda^{FR}+(p-2)\rho)}{(1+\lambda^{FR}-\rho)(1+\lambda^{FR}+(p-1)\rho)}$	$\frac{\sigma_\epsilon^2}{n}\frac{\left(\lambda^{FR}-(\rho-1)(1+(p-1)\rho)\right)^2-(p-1)(\rho-1)\rho^2(1+(p-1)\rho)}{(1+\lambda^{FR}-\rho)^2(1+\lambda^{FR}+(p-1)\rho)^2}$
Partial Ridge		
$p=3$	$\frac{\lambda_i\rho\sum_{j\neq i}\beta_j}{1+\lambda_i+\rho-2\rho^2}$	$\frac{\sigma_\epsilon^2}{n}\frac{\left(\lambda_i-(\rho-1)(1+2\rho)\right)^2-2(\rho-1)\rho^2(1+2\rho)}{(1+\lambda_i+\rho-2\rho^2)^2}$
$p=4$	$\frac{\lambda_i\rho\sum_{j\neq i}\beta_j}{1+\lambda_i+2\rho-3\rho^2}$	$\frac{\sigma_\epsilon^2}{n}\frac{\left(\lambda_i-(\rho-1)(1+3\rho)\right)^2-3(\rho-1)\rho^2(1+3\rho)}{(1+\lambda_i+2\rho-3\rho^2)^2}$
$\vdots$		
p	$\frac{\lambda_i\rho\sum_{j\neq i}\beta_j}{1+\lambda_i+(p-2)\rho-(p-1)\rho^2}$	$\frac{\sigma_\epsilon^2}{n}\frac{\left(\lambda_i-(\rho-1)(1+(p-1)\rho)\right)^2-(p-1)(\rho-1)\rho^2(1+(p-1)\rho)}{(1+\lambda_i+(p-2)\rho-(p-1)\rho^2)^2}$

Therefore, from looking at the pattern over p when evaluating the expressions acting as the main components in Proposition 3.1, the results detailed in Proposition 3.3 are obtained.

3.A.5 Empirical Application R squared values

TABLE 3.A1: R squared values for house price prediction

	Sample split (no. of observations in training data)				
	84	107	129	162	184
Full Ridge	0.928	0.930	0.940	0.936	0.924
Lasso	0.945	0.947	0.945	0.943	0.939
OLS-post-Lasso	0.946 (11)	0.945 (11)	0.945 (15)	0.942 (21)	0.941 (12)
Hybrid (correlation)	0.952 (26)	0.949 (3)	0.953 (26)	0.936 (1)	0.937 (24)
Hybrid (Lasso $\hat{\beta}s$)	0.935 (15)	0.936 (15)	0.952 (18)	0.947 (30)	0.934 (26)
Hybrid (Ridge $\hat{\beta}s$)	0.919 (16)	0.943 (22)	0.952 (25)	0.945 (12)	0.940 (17)

TABLE 3.A2: R squared values for construction cost prediction

	Sample split (no. of observations in training data)				
	84	107	129	162	184
Full Ridge	0.863	0.842	0.754	0.621	0.654
Lasso	0.590	0.764	0.905	0.903	0.891
OLS-post-Lasso	0.771 (3)	0.821 (4)	0.070 (7)	0.902 (14)	0.828 (24)
Hybrid (correlation)	0.726 (1)	0.677 (1)	0.742 (1)	0.610 (5)	0.739 (18)
Hybrid (Lasso $\hat{\beta}s$)	0.756 (9)	0.677 (1)	0.743 (1)	0.646 (5)	0.685 (8)
Hybrid (Ridge $\hat{\beta}s$)	0.726 (1)	0.677 (1)	0.824 (20)	0.660 (20)	0.708 (21)

Chapter 4

Partial Random Projections, A Novel Approach to High-Dimensional Linear Regression in Economics

4.1 Introduction

The increased availability of data through automated collection as well as a wide host of other factors has meant that nearly all fields utilising quantitative analysis have faced larger data sets. For cross-sectional and longitudinal data sets this may involve a greater number of samples and time observations respectively, or it might be a case of there being a larger number of data attributes available. Both has presented new challenges for researchers ranging from data storage and manageability issues to the computational feasibility when using existing statistical models. Within the field of economics, the most significant area where this is noticed lies with linear regression analysis, where one is typically using a model similar to that detailed by the following:

$$y = X\beta + \epsilon\,, \tag{4.1}$$

where y is an $n \times 1$ vector representing the dependent variable and X is a $n \times p$ matrix with each of the p columns representing an independent variable. The ϵ term represents the disturbance component, typically assumed to be such that each element is IID with 0 mean and constant variance across all observations. Finally, β is a $p \times 1$ vector representing the coefficients and are what one has to estimate, whether this be for predictive purposes ($X\hat{\beta}$) or to analyse causal inference of certain covariates. Usually, this is done through methods based upon Ordinary Least Squares (OLS) resulting in the following unbiased and efficient estimator:

$$\hat{\beta} = (X'X)^{-1}X'y\,, \tag{4.2}$$

where one can see that the above leans upon the inversion of the covariance matrix, $X'X$. However, due to the increased availability of candidate predictors, it is often such that p is large relative to n resulting in 4.2 providing uncertain estimates and when $p > n$ then the above is no longer feasible, this is characterised as a high-dimensional setting. Throughout economics, there are 2 main branches of approaches that seek allow estimation of β for either predictive or causal inference purposes, the first being regularization techniques in the form of Penalized Least Squares such as as Ridge and Lasso (Hoerl and Kennard (1970) and Tibshirani (1996)). They work in a similar way to OLS, only they induce some bias in return for reduced variance to handle to overparameterisation aiming to make the trade off favourable in a mean-squared-error (MSE) sense.

The second branch of approaches, frequently used in the field of economics, is to directly reduce the size of the pool of predictors in order to make OLS feasible. The most widely used example of this is factor models, which make the assumption that all covariates are determined by a small number of underlying driving forces. While there have been many ways in which these factors can be estimated from a large set of predictors, Principal Components Analysis (PCA) is the most used method due to it not requiring any a priori knowledge as well as its ability to retain a significant amount of the information from the original data set as seen in Connor and Korajckzy (1988) and Stock and Watson (2002). Here, one computes the factors by transforming X using the eigenvectors of the predictor covariance matrix. This is followed by choosing a small number of factors to run OLS on with a significant amount of information from all p predictors maintained in the small number of factors used.

Other less common methods include model selection approaches that still allow OLS to be run on a subset of the original predictor by choosing what is believed to be the most relevant with respect to the dependent variable. Historically, this was carried out through various information criteria such as: AIC (Akaike (1969)), BIC (Schwarz (1978)) and MIC (Mallows (1973)), however, even these approaches suffer when p is too large as they require all possible submodels to be computed under the criteria in order to choose an optimal set of predictors. More modern approaches involve using the model selection element of the Lasso to choose a set of predictors to use OLS on (Belloni and Chernozhukov (2013)). Such estimation procedures are well established throughout the economics and statistics literature and are commonly adapted to fit the extent to which $p > n$ with often 2 or more Lasso models being applied in order to sufficiently reduce the submodel of predictors. The obvious criticism of such methods lies in the assumption of a certain degree of sparsity amongst the true predictors as one discards a significant amount of information (possibly useful) which can adversely effect parameter estimation. Finally, in the case where one is only interested

in predictions of the dependent variable, model averaging is nice alternative considered in work such as Timmermann (2006) and Elliott et al. (2013) with promising results.

Returning to the second branch of methods, one approach similar in nature to Principal Components, but originating from the machine learning literature is Random Projections (RP). Where PCA seeks to reduce the dimension of X by multiplying X by a matrix containing the k eigenvectors (where k is the size of the new dimension) corresponding to the k largest eigenvalues of $X'X$. RP does something similar in that X is multiplied by a matrix R in order to reduce the dimension of X, only here, the entries of R are independent but identically distributed random variables. To reduce the number of rows in X, a row-wise random projection was analysed in Dhillon et al. (2013) where the following transformation to a small subspace occurs:

$$R_{k\times n}X_{n\times p}\,. \tag{4.3}$$

Such a transformation is known as sketching and is very common throughout the machine learning literature where, very frequently, the issue faced by analysts lies in the number of data observations being too large to store and move the original data set around prior to applying the data to a model. However, from the perspective of linear regression analysis it is the number of covariates (p) being too large which is the issue requiring a column-wise transformation to a smaller subspace as follows:

$$X_{n\times p}R_{p\times k}\,, \tag{4.4}$$

where $k << p$ meaning that methods based upon OLS can now be carried out on the above transformed data set as is done in work such as Kaban (2014) and Thanei et al. (2017). Typically, one would compute R such that each entry is an IID standard normal distribution, however, other distributions have been considered in work such as Li et al. (2006) and Achlioptas (2003) in order to save computational time and is often known as Sparse Random Projections. However, in the majority work related to linear regression, the standard normal has been used and will be the main focus of this paper.

Intuitively, one might question how RP can be successful when the compressed dimension column are simply linear combinations of the original set with random weights compared to a data driven approach such as PCA. However, the Johnson-Lindenstrauss lemma of Johnson and Lindenstrauss (1984) shows how when this process is repeated over multiple draws of R, the expectation maintains the pairwise Euclidean distances between the points with other work on this including Matoušek (2008) and Dasgupta and Gupta (2003). Therefore, one can view this as a

way of side stepping the curse of dimensionality whilst maintaining all the information necessary for the desired form of analysis to take place.

This premise of Random Projections has been used for a wide variety of applications throughout the machine learning literature. A large area of work has focused on using RP as a means of reducing the dimension of image and text data in order to carry out Nearest Neighbour (NN) algorithms where one is might be interested in finding similar documents based on certain words appearing in similar papers or identifying images of identical objects or landmarks from a large pool of separate images based on the brightness of pixels. Work such as Fradkin and Madigan (2003), Yan et al. (2018) and Nabil (2017) have shown how RP can be useful in applying NN algorithms to large data sets with Deegalla and Bostrom (2006) comparing RP to PCA on medical image data sets. They find that, for the majority of data sets, PCA is more accurate but is more sensitive to the choice of subsample size compared to RP.

Another key area of the machine learning literature where RP has been used successfully is that of clustering and classification where huge data sets concerning text data, such as academic papers, social media posts and financial market announcements are often highly noisy causing standard algorithms to poorly group the observations concerned. RP has been applied to algorithms such as the Gaussian Mixture Model in work such as Anderlucci et al. (2010), Fern and Brodley (2003) and Dasgupta (2000). Kaski (1998) compared RP to PCA in a setting that organises text documents and and found that RP was equally as effective with regards to sorting accuracy but has the benefit of a faster run time.

However, in terms of how RP has been applied to linear regression settings, the literature is somewhat sparse and nearly all work focuses purely on prediction of y rather than individual β estimation. McWilliams et al. (2015) proposed an algorithm that combines RP with the Ridge Regression of Hoerl and Kennard (1970) in the so-called LOCO algorithm while Slawski (2018) compares the prediction error of RP to PCA in a Twitter data application highlighting the potential of RP. Bounds on excess predictive risk are constructed in Maillard and Munos (2009) and developed further in Kaban (2014) and Thanei et al. (2017) which reveal a similar bias-variance trade off that methods such and Ridge and Lasso experience. Where reducing the size of the new subspace, k, induces bias but reduces variance in a predictive sense. Finally, one of the only works to apply RP to a setting in economics is Boot and Nibbering (2019) who show RP outperforming many other commonly used method for high-dimensional linear regression in a setting concerned with forecasting with FRED-MD data. They also derive a predictive risk bound to support the case of RP

being useful for forecasting with macroeconomic indicators.

Therefore, it can be seen that such a method is very new to economics in a predictive sense. In addition, it is also noticeable that RP has not been adapted in any way that focuses primarily on individual parameter estimates that may be of interest in typical linear regression settings encountered by economists. Therefore, this paper seeks to propose an estimation procedure for individual parameters that may be of particular interest to an analyst. This procedure seeks to utilise the dimension reduction element of RP and, hence, is referred to as Partial Random Projections (PRP). In a similar fashion to that of many earlier methods, it aims to achieve a favourable bias-variance trade off when faced with data sets typically faced by economists and, therefore, are often characterised by significant contemporaneous correlation amongst predictors as well as sparsity in the true data-generating-process (DGP). More specifically, ordinary RP is known to is known to induce bias into the coefficients from reducing the dimension of the predictor matrix. PRP seeks to improve upon this from the perspective of a single parameter by using the RP mechanism to reduce the dimension of the X_{-j} without affecting x_j. The aim of this being to achieve more favourable properties in the estimate of β_j as a reflection of its true value.

This paper is organized as follows: Section 4.2 formally defines RP and the proposed estimation procedure as well as the toy model setting under which the properties of this approach will be analysed. Section 4.3 goes into depth on the behaviour of the bias and variance of PRP with comparisons drawn with Ridge. Section 4.4 provides simulation evidence to support the theoretical conclusions and Section 4.5 concludes.

4.2 Model Framework and Assumptions

4.2.1 Estimation Procedure

In order to define the proposed parameter estimation approach, the following true Data Generating Process (DGP) is defined as a reformulation of 4.1 with $X = (x_j \; X_{-j})$ and $\beta = (\beta_j \; \beta_{-j})'$ with β_j being the individual coefficient of interest at this given time,

$$y = x_j\beta_j + X_{-j}\beta_{-j} + \epsilon\,, \tag{4.5}$$

where x_j is an $n \times 1$ vector of the covariate whose coefficient is of interest while X_{-j} is a $n \times (p-1)$ matrix with each column representing another covariate which may or may not be active in determining the outcome, y. Therefore, β_j is a scalar and β_{-j} is a $(p-1) \times 1$ vector. The error term is denoted by ϵ with a more formal definition in the

assumptions below.

From here it is assumed that p is large relative to n resulting in OLS via the Frisch-Waugh-Lovell Theorem being infeasible since $p-1>n$. Therefore, in the nature of RP, X_{-j} is compressed to a $n \times k$ dimension where $k<n<p-1$ by multiply by the random matrix R^s for $s=1,\ldots,S$ where S is the total number of random draws of the R matrix used. Specifically, the Frisch-Waugh-Lovell theorem can can now be applied to the following linear model. So, essentially, one is estimating β_j by partialling out the covariates in X_{-j} with a RP mechanism as seen below,

$$y = x_j\beta_j + X_{-j}R^s\beta_{-j} + e\,, \tag{4.6}$$

where R^s is a $p-1 \times k$ matrix such that:

$$R^s_{ij} \sim_{IID} N(0,1)\,.$$

For $i=1,\ldots,p-1$ and $j=1,\ldots,k$. This makes $X_{-j}R^s$ an $n \times k$ matrix with $k << p$.

This process can be repeated for each variable in $j=1,\ldots,p$ so that a full profile of β can be estimated. This way one side steps the issue of OLS being infeasible while the potential higher bias arising from Random Projections alone due to how one compresses the original predictor then recompresses it after OLS has been carried out.

Proposition 4.1: The Partial Random Projections estimator for β_j for a single draw of R^s is given by the following expression:

$$\hat{\beta}^s_j = (x_j'x_j)^{-1}x_j'(y - X_{-j}R^s\hat{\beta}_{-j})\,, \tag{4.7}$$

where

$$\hat{\beta}^s_{-j} = (R^{s\prime}X'_{-j}M_xX_{-j}R^s)^{-1}R^{s\prime}X'_{-j}M_xy\,. \tag{4.8}$$

It is important to mention here that $\hat{\beta}^s_{-j}$ is a $k \times 1$ vector so once may not view this as a final estimator of β_{-j}. To obtain this estimate, one must multiply $\hat{\beta}^s_{-j}$ by R^s, however, this is irrelevant when one is only concerned with $\hat{\beta}^s_j$. Since this is repeated for S draws of R, a final estimate for $\hat{\beta}_j$ is computed by averaging the estimate for each draw as follows:

$$\hat{\beta}_j = E_R[\hat{\beta}^s_j] = \frac{1}{S}\sum_{s=1}^{S}\hat{\beta}^s_j\,.$$

A proof is provided in Appendix 4.A.1

It is worth noting that, from here onwards, the s superscript notation will be dropped as all expressions will be considered as an expectation (or variance) over all draws of R. Here, $M_x = I_n - x_j(x_j'x_j)^{-1}x_j'$. While these expressions mean little in isolation, the following subsections analyse the bias and variance of this method and draw comparison with other high-dimensional regression methods in a simplified setting. While the expression in Proposition 4.1 is for a single draw of R, the bias and variance analysis allows for one to construct numerous $\hat{\beta}_j$ estimates over multiple draws of R then average them to obtain a final estimate.

4.2.2 Toy Model

In the following section, expressions for the bias and variance of Partial Random Projections will be derived for a fixed design setting. However, on surface value these do not mean much in terms of understanding how they behave under various conditions of the data and true DGP. Therefore, some restrictions are imposed later on in order to more easily understand how the bias and variance will likely vary between settings in order to determine when this novel approach can outperform other well established methods.

The following assumptions are made:

A1 The error term components are independently and identically distributed with 0 mean and homoskedastic variance. $\epsilon \sim IID(0, \sigma_\epsilon^2 I_n)$.

A2 The predictor covariance matrix $X_{-j}'X_{-j}$ is positive semi-definite.

A3 All covariates are standardized to have mean 0 and unit variance ($x_i'x_i = n$ for all $i = 1, \ldots, p$).

A4 The correlation between x_j and all covariates in X_{-j} are equal. That is $\frac{X_{-j}'x_j}{n} = (\tau, \tau, \ldots, \tau)'$.

A5 There is equal correlation between all covariates in X_{-j} and this need not be the same as the correlation between x_j and each variable in X_{-j} ($\rho \neq \tau$). Therefore, with a slight abuse of notation, the covariance matrix of X_{-j} is given as follows:

$$\frac{X_{-j}'X_{-j}}{n} = \begin{bmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \rho & \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \ldots & 1 \end{bmatrix}.$$

Where one can see that in order for the above $p-1$ square matrix to be positive definite, the following condition must be satisfied for $p > 2$ (for $p = 2$ then only only $\rho^2 < 1$ is required),

$$\rho > -\frac{1}{p-2}\,.$$

Note that, in order for the covariance matrix for $X = (x_j\ X_{-j})$ to be positive definite, the following condition, derived from recursively computing the determinant of $X'X$, must hold concerning τ and ρ assuming n is fixed,

$$(p-2)\rho > (p-1)\tau^2 - 1\,.$$

As $p \to \infty$ then this is approximately

$$\rho > \tau^2\,.$$

Therefore, in this toy model setting, it is such that $\rho \geq 0$ which, although may seem restrictive, the results will be without loss of generality.

4.3 Estimator Properties

4.3.1 Bias

In this section, the bias of Partial Random Projections is derived under the toy model assumptions and compared to other commonly used approaches such as Ridge Regression. To begin with, the bias is derived from the expression in Proposition 4.1 using the formula of conditional expectations since both R and ϵ are stochastic. Therefore, as mentioned previously, from here onwards, the case where multiple draws of R are used and averaged over is considered with E_R denoting the mean across all draws of R.

4.3.1.1 Partial Random Projections

Since it is known that RP induces bias from projecting the predicting matrix onto a lower dimensional subspace, one would suspect that sparing x_j from projection would benefit the estimation of β_j. The following proposition and analysis investigate this claim further to understand how PRP may be able to hold an edge over RP in certain scenarios.

Proposition 4.2: Where R is the $p-1 \times k$ subspace transformation matrix, the bias expression for the Partial Random Projections estimator is given by the following general expression by taking an expectation over both R and ϵ,

$$Bias(\hat{\beta}_j) = (x_j' x_j)^{-1} x_j' X_{-j} \left(\beta_{-j} - E_R[R(R' X_{-j} M_x X_{-j} R)^{-1} R'] X_{-j}' M_x X_{-j} \beta_{-j} \right) . \quad (4.9)$$

A proof is provided in Appendix 4.A.2

From here, one can carry out an eigendecomposition of the matrix $X_{-j}' M_x X_{-j}$ as follows:

$$X_{-j}' M_x X_{-j} = UDU' , \quad (4.10)$$

$$D = diag(\lambda_1, \lambda_2, \ldots, \lambda_{p-1}) ,$$

where U is a $p-1 \times p-1$ matrix with each column representing an eigenvector of $X_{-j}' M_x X_{-j}$ and D is $p-1 \times p-1$ diagonal matrix with each diagonal element being the eigenvalues of $X_{-j}' M_x X_{-j}$ in descending order with the largest in the top left down to the smallest in the bottom right. Since $X_{-j} M_x X_{-j}$ is symmetric, it is such that U is orthogonal meaning that $U'U = UU' = I_{p-1}$ resulting in the following expression for the bias term:

$$Bias(\hat{\beta}_j) = (x_j' x_j)^{-1} x_j' X_{-j} U \left(I_{p-1} - E_R[R(R' D R)^{-1} R'] D \right) U' \beta_{-j} . \quad (4.11)$$

From here, the fundamental way to evaluate the expression above lies in the evaluation of $E_R[R(R'DR)^{-1}R']$. Due to the nature of how this transformation of the R components involves quadratic ratios, it is not possible to write this as a closed form expression as shown in Marzetta et al. (2011). However, it is well defined in the sense that it is a diagonal matrix with each diagonal element a function of $\frac{k}{p}$ (more specifically $\frac{k}{p-1}$) and the eigenvalues in D,

$$E_R[R(R'DR)^{-1}R'] = diag\left(\frac{1}{\eta_1}, \frac{1}{\eta_2}, \ldots, \frac{1}{\eta_{p-1}} \right) . \quad (4.12)$$

It is worth noting that a special case where all the eigenvalues of $X_{-j}' M_x X_{-j}$ are equal results in 4.11 being much more easily analyzed since it is such that $\lambda_1 = \lambda_2 = \cdots = \lambda = \eta_i$ for all i meaning that 4.12 can be written as $\frac{1}{\lambda} I_{p-1}$. However, the following theorem generalizes more under the toy model framework.

Theorem 4.1: Under the assumptions of the toy model detailed in Section 4.2.2, the bias of the PRP estimator is given by the following:

$$Bias(\hat{\beta}_j) = \tau \left(1 - \frac{1 + (p-2)\rho - (p-1)\tau^2}{\eta_1}\right) \sum_{i \neq j}^{p} \beta_i \,, \tag{4.13}$$

where the $\frac{1}{\eta_1}$ component comes from the assumption that $1 + (p-2)\rho - (p-1)\tau^2$ is the maximum eigenvalue which can be made without loss of generality.

A proof is provided in Appendix 4.A.3

From, here one can use Corollary 1 from Thanei et al. (2017) that allows the bias term to be bound from above.

Lemma 4.1: The bias term can be bounded above with the following expression:

$$Bias(\hat{\beta}_j) \leq \tau w \sum_{i \neq j}^{p} \beta_i \,, \tag{4.14}$$

where

$$w = \frac{\left(1 + \frac{1}{k}\right)\alpha^2 + \left(1 + \frac{2}{k}\right)\alpha + \frac{1}{k}}{\left(k + 2 + \frac{1}{k}\right)\alpha^2 + 2\left(1 + \frac{1}{k}\right)\alpha + \frac{1}{k}} \,,$$

$$\alpha = \frac{1 + (p-2)\rho - (p-1)\tau^2}{Trace(D)} \,.$$

Combining all the expressions in Lemma 4.1, the following inequality can be written.

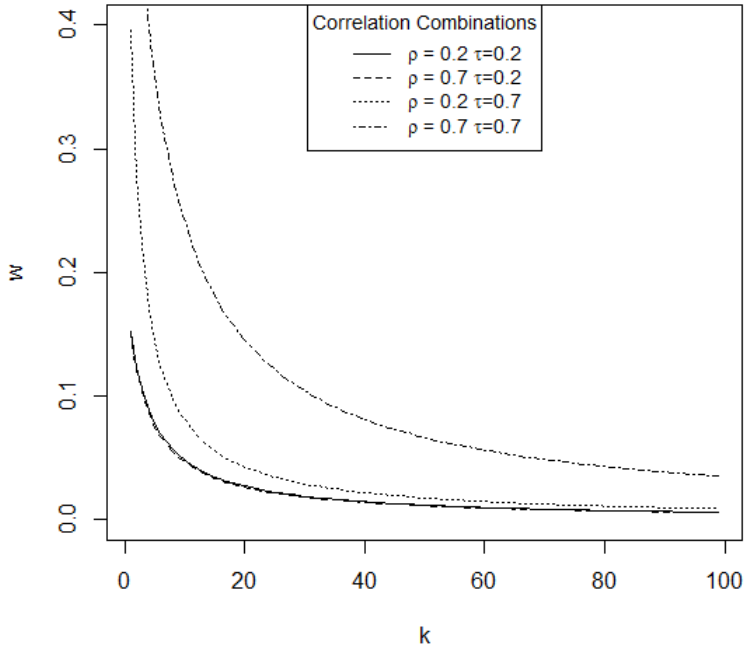
$$Bias(\hat{\beta}_j) \leq \frac{\tau(2 + 3(p-2)\rho - 2(p-1)\tau^2)}{2 + 3(p-2)\rho - 2(p-1)\tau^2 + k(1 + (p-2)\rho - (p-1)\tau^2)} \sum_{i \neq j}^{p} \beta_i \,. \tag{4.15}$$

One of the first features that stands out is how when $\tau = 0$ the bias (and squared bias) of PRP is 0 which hints that this methods looks promising for estimating coefficients of covariates that show little correlation with the other candidate predictors. This corresponds to the case where x_j is uncorrelated with all variables in X_{-j} so may be viewed as hypothetical but provides a good foundation for characterising when PRP will perform well. Another interesting property can be seen when considering the case where $\rho = 0$ allowing significant simplification leading to the following bound:

$$Bias(\hat{\beta}_j) \leq \frac{2\tau}{2 + k} \sum_{i \neq j}^{p} \beta_i \,. \tag{4.16}$$

This helps so show, more clearly, how the bias magnitude remains low when either τ, $\sum_{i \neq j}^{p} \beta_i$ or both are small and this can be assisted by choosing a larger subspace dimension (k). Such an expression can be easily compared to that of Ridge regression and will be discussed in the following subsection. Finally, one can see that the fraction term from 4.15 lies between 0 and 1 and will depend on ρ, τ and k. the figure below shows, under various magnitudes of ρ and τ, how the fraction component changes with k.

FIGURE 4.1: How the fraction component (τw) varies across k under various correlation conditions where p=100



As expected, the weight, and the magnitude of the bias, decreases as k approaches p. It is also interesting to see that the case where τ and ρ is 0.7 provides the highest weight compared to the cases where one or both ρ and τ being significantly smaller. It is also worth noting that cases where τ is smaller seem to result in the weight starting at a smaller value and staying lower than the rest for all values of k. So it seems that τ is more influential on the bound for the bias compared to ρ which can have implications for the optimal k as when τ is small, there is little improvement in the bound compared to when τ is larger. This can allow one to impose a lower k to keep the variance down unlike the case where $\tau = 0.7$ where the bound improvement slope does not start to flatten until $\frac{k}{p}$ reaches approximately 0.5.

4.3.1.2 Comparison to other methods

In this subsection, the bias of PRP will be compared with the marginal least squares estimator for β_j as well as that of Ridge regression with respect to features of the true DGP. Firstly, for marginal least squares, the estimator for β_j under 4.5 is given as follows:

$$\hat{\beta}_j^{mls} = \frac{x_j' y}{x_j' x_j}\,, \tag{4.17}$$

where one can easily show that the bias is given by the following:

$$Bias(\hat{\beta}_j^{mls}) = \frac{x_j' X_{-j}\beta_{-j}}{x_j' x_j} = \tau \sum_{i \neq j}^{p} \beta_i\,. \tag{4.18}$$

This can be shown to be always larger than the bound for the bias of PRP in 4.15 since the discussion below A5 showed that $1 + (p-2)\rho - (p-1)\tau^2 \geq 0$ then the following is true,

$$w = \frac{(2 + 3(p-2)\rho - 2(p-1)\tau^2)}{2 + 3(p-2)\rho - 2(p-1)\tau^2 + k(1 + (p-2)\rho - (p-1)\tau^2)} \leq 1\,. \tag{4.19}$$

Moreover, it can be shown that w is smaller when the degree to which $\rho > \tau^2$ is larger. Therefore, PRP has noticeable gains in bias over marginal least squares which is interesting given how the 2 approaches are similar in nature due to how they focus on the estimation of a single parameter at a time. This a crucial step for the PRP approach as marginal least squares should be viewed a benchmark and to take this further a comparison to the more sophisticated Ridge Regression will be made to consolidate this further.

Under the toy model assumptions, one can easily show that the bias expression of Ridge regression is given as follows:

$$Bias(\hat{\beta}_j(\lambda_R)) = \frac{\tau\lambda_R \sum_{i \neq j}^{p} \beta_i - \lambda_R \beta_j (1 + \lambda_R + (p-2)\rho)}{(\lambda_R + 1)^2 + \rho(p-2)(\lambda_R + 1) - (p-1)\tau^2}\,, \tag{4.20}$$

where, unlike for PRP, when $\tau = 0$ the bias in non-zero. More specifically, it is given by

$$Bias(\hat{\beta}_j(\lambda_R)) = -\frac{\lambda_R \beta_j}{\lambda_R + 1}\,. \tag{4.21}$$

Here, one can also see that the main component of this expression is the true value of β_j itself meaning that as β_j becomes larger in magnitude, PRP has bigger gains potential through the bias. It is also very similar for when $\sum_{i \neq j}^{p} \beta_i = 0$ where the PRP

is unbiased whereas the bias for Ridge is given by the following:

$$Bias(\hat{\beta}_j(\lambda_R)) = -\frac{\lambda_R\beta_j(1+\lambda_R+(p-2)\rho)}{(\lambda_R+1)(1+\lambda_R+(p-2)\rho)-(p-1)\tau^2}\,, \tag{4.22}$$

$$= -\frac{\lambda_R^2\beta_j+\lambda_R\beta_j+O(1)}{\lambda_R^2+((p-2)\rho+2)\lambda_R+O(1)}\,,$$

where, once again, this is typically non-zero and grows significantly in magnitude as the size of β_j does. Returning to the expression in 4.20, for the purpose of this comparison, consider the bias written in the following form:

$$Bias(\hat{\beta}_j(\lambda_R)) = \frac{\tau\sum_{i\neq j}^{p}\beta_j}{\frac{1}{\lambda_R}\left((\lambda_R+1)^2+\rho(p-2)(\lambda_R+1)-(p-1)\tau^2\right)} - \frac{\beta_j(1+\lambda_R+(p-2)\rho)}{\frac{1}{\lambda_R}\left((\lambda_R+1)^2+\rho(p-2)(\lambda_R+1)-(p-1)\tau^2\right)}\,. \tag{4.23}$$

Here the focus will be on the first fraction term where the numerator is similar to that seen in the bound for the PRP bias in 4.14 and 4.15 where $\tau\sum_{i\neq j}^{p}\beta_i$ is the key component. Looking closer, one can expand and rearrange the denominator term as follows where, henceforth, it is such that $h = 1+(p-2)\rho-(p-1)\tau^2$ $(h>0)$,

$$\lambda_R+2+(p-2)\rho+\frac{h}{\lambda_R}\,. \tag{4.24}$$

To see how this behaves, the first derivative is taken with respect to λ_R,

$$\frac{d}{d\lambda_R}\left(\lambda_R+2+(p-2)\rho+\frac{h}{\lambda_R}\right) = 1-\frac{h}{\lambda_R^2}\,. \tag{4.25}$$

Where it can be seen that this is decreasing in λ_R up until where $\lambda_R = h^{\frac{1}{2}}$ with a strictly positive second derivative. Even from here, the gradient is never greater than 1 meaning that there is the potential (when h is very large) for a very large λ_R required in order to sufficiently restrain this first component of 4.23. However, the second fraction of 4.23 has a λ_R term in the numerator and will always increase in magnitude with λ_R when $|\beta_j|>1$ since the numerator and denominator both have the same order of magnitude in λ_R.

Now, recall from 4.15 the bias bound that can be reformulated in a similar way to that of the first term from 4.23,

$$Bias(\hat{\beta}_j) \leq \frac{\tau\sum_{i\neq j}^{p}\beta_i}{1+\frac{kh}{2h+(p-2)\rho}}\,. \tag{4.26}$$

Therefore, one can assess these 2 bias terms by comparing 4.24 and the denominator of

4.26. This can be formalised with the following condition for when the denominator of 4.26 is greater than 4.24,

$$\frac{kh}{2h+(p-2)\rho} - \left(\lambda_R + \frac{h}{\lambda_R} + 1 + (p-2)\rho\right) > 0\,. \tag{4.27}$$

When this condition holds, one more step can be taken in establishing the bias of PRP being smaller in magnitude than that Ridge by revisiting 4.23. If the denominator of 4.26 is larger than that of 4.24, it can then be seen that if β_j has the opposite sign to that of $\tau \sum_{i\neq j}^{p} \beta_i$ then the second fraction in 4.23 will increase the magnitude of the squared bias for Ridge due to how the denominator is always positive for $\lambda_R > 0$.

Looking at 4.27 closer one can see that a large h makes this much more likely to hold or, at the very least, requires a very small λ_R in order to prevent it which would then leave little room for Ridge to demonstrate its variance gains as will be discussed in the following section. In addition, this may not still prevent the second fraction term in 4.23 from escalating the bias significantly when β_j and $\sum_{i\neq j}^{p} \beta_i$ are opposite signs. Therefore, one can argue that situations where ρ is much greater then τ^2 (leading to a large h) favour PRP over Ridge due to the bias. To summarise this more concretely; the larger h leads to a larger denominator for the bias bound from 4.26 which, combined with β_j being of opposite sign to that of $\sum_{i\neq j}^{p} \beta_i$ contributes to a bound of the PRP bias being much less than that of Ridge. So one can see that there are multiple channels to which PRP can improve over Ridge.

4.3.2 Variance

In this section, the behaviour of the variance of PRP under the toy model is investigated with respect to the subspace dimension as well as the relevant features of the true DGP. It is somewhat more complex than the bias due to how uncertainty comes from both the true DGP noise (ϵ) as well as the randomness from the dimension reduction matrix (R). Nonetheless, the Proposition below represents and expression for the variance in closed form.

4.3.2.1 Partial Random Projections

While it was established that PRP shows promising signs through its bias term, this may be undone should its variance suffer under certain conditions of the true DGP. The following proposition and analysis look closer into how its variance behaves

given that only X_{-j} has its dimension reduced.

Proposition 4.3: The variance expression of the PRP estimator is given by the following expression by considering the variance over the distribution of both R and ϵ,

$$Var(\hat{\beta}_j) = \sigma_\epsilon^2 \left((x_j'x_j)^{-1} + (x_j'x_j)^{-1} x_j' X_{-j} E_R[R(R'X_{-j}'M_x X_{-j} R)^{-1} R'] X_{-j}' x_j (x_j'x_j)^{-1} \right) + Var_R \left((x_j'x_j)^{-1} x_j' X_{-j} R(R'X_{-j}'M_x X_{-j} R)^{-1} R' X_{-j}' M_x X_{-j} \beta_{-j} \right) . \quad (4.28)$$

A proof is provided in Appendix 4.A.4

From first appearance of the expression above, one can see that there are 2 components which can be justified with intuition. The first line of 4.28 is similar in nature to the variance of a Frisch-Waugh OLS estimate for β_j being governed by the error variance and the covariance matrix of the transformed data set. While the second line is noticeably different and appears to be very similar in structure to that of the β_j estimate for a given draw of R seen in 4.7 or even the bias term in 4.9. Therefore, one can view this component as the variability in the estimate of β_j arising from the randomness of R.

Using a similar approach to that used for the bias, one can rewrite the expectation term in 4.28 as follows:

$$(x_j'x_j)^{-1} x_j' X_{-j} E_R[R(R'X_{-j}'M_x X_{-j} R)^{-1} R'] X_{-j}' x_j (x_j'x_j)^{-1} = \frac{(p-1)\tau^2}{\eta_1} . \quad (4.29)$$

Where η_1 corresponds to the maximum eigenvalue of $X_{-j}'M_x X_{-j}$. From here, one can use the following inequality from Thanei et al. (2017) to bound the above expression,

$$\left(1 - \frac{\lambda_i}{\eta_i}\right)^2 \leq w_i^2 \leq 1 , \quad (4.30)$$

$$\frac{1}{\eta_1} \leq \frac{2}{\lambda_1} .$$

Applying this to 4.29 with the maximum eigenvalue being $1 + (p-2)\rho - (p-1)\tau^2$, the following variance inequality can be stated. Due to what was discussed in 4.2.2 regarding the relationship between ρ and τ^2, when writing general expressions it will

be assumed that $\rho > \tau$. Therefore the variance is 4.28 can be bounded as follows:

$$Var(\hat{\beta}_j) \leq \frac{\sigma_\epsilon^2}{n}\left(1 + \frac{2(p-1)\tau^2}{1+(p-2)\rho-(p-1)\tau^2}\right) + Var_R\left((x_j'x_j)^{-1}x_j'X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'M_xX_{-j}\beta_{-j}\right). \quad (4.31)$$

While this bound does is not dependent on k, the first component in brackets on the right-hand side can be shown to be increasing in k and will always be greater than $\frac{\sigma_\epsilon^2}{n}$ apart from the case where $\tau = 0$ (since $1 + (p-2)\rho - (p-1)\tau^2 > 0$). Once again, it can be seen that this component is much lower when ρ is significantly greater than τ^2 along with τ alone being smaller in magnitude bringing the entire term in brackets closer to 1. This is an important benchmark as this is what the variance would be for OLS if it were feasible and will also be an important point of comparison to other methods as will be seen in the following subsection.

The second component of the above inequality is notably more challenging to evaluate for 2 reasons. Firstly, the complexity of it means that it is hard for one to simplify it into standard components, such as $R(R'X_{-j}'M_xX_{-j})^{-1}R'$, leaving a very complex transformation of Gaussian random variables. Secondly, the fact it is a polynomial function of Gaussian random variables with infinite support makes it challenging to bound. However, one can see that when $\tau = 0$ then the whole component goes to 0, stressing once more how PRP performs so well when the correlation between x_j and X_{-j} is relatively low.

To understand this more deeply, a small simulation study is carried out to investigate the behaviour of the variance component on the right-hand side of 4.31 with respect to ρ, τ and β_{-j}. This small study is divided into 4 separate classes, each considering a different profile of β_{-j}. For each design the number of of coefficients in β_{-j}, $p-1$, is fixed at 100 and the variance is taken over 100 draws of R. 3 separate combinations of ρ and τ are considered; one being where they are both small, one where ρ is large and τ is small and, finally, one where they are both large. It is worth noting that one can easily verify that, for each experiments, the conditions of A5 are satisfied with a varying magnitude of $(p-2)\rho - (p-1)\tau^2$. Regarding the profiles of β_{-j}, the following table summarises this.

Design	β_{-j} pattern	$\sum_{i=1}^{p-1}\beta_{-j,i}$
1	$\beta_i = 1$ for all i	$p-1$
2	50% of $\beta_i = 2$, 50% of $\beta_i = -2$	0
3	$\beta_i = i$	$\frac{p(-1)}{2}$
4	$\beta_i = i(-1)^{i+1}$	$-\frac{p-1}{2}$

So one can see that there is significant variation in profiles of β_{-j} that one might face in empirical settings. One would expect Design 3 to possibly produce the largest variances due to how 4.31 shows the variance component increasing with the elements of β_{-j} and this design has by far the largest sum of all the coefficients. The following tables report the value of this variance component using 100 draws of R for each design with each row representing a different experiments with varied ρ and τ combination while each column represents a different subspace dimension, k, for the matrix R.

TABLE 4.1: Design 1 Variance component of 4.31

	k								
	10	20	30	40	50	60	70	80	90
$\tau = 0.2, \rho = 0.2$	4.965	0.973	0.465	0.174	0.078	0.026	0.016	0.006	0.003
$\tau = 0.2, \rho = 0.7$	0.315	0.029	0.005	0.001	0.000	0.000	0.000	0.000	0.000
$\tau = 0.7, \rho = 0.7$	20.08	3.264	0.795	0.258	0.072	0.037	0.018	0.010	0.002

TABLE 4.2: Design 2 Variance component of 4.31

	k								
	10	20	30	40	50	60	70	80	90
$\tau = 0.2, \rho = 0.2$	0.163	0.133	0.068	0.056	0.037	0.017	0.015	0.010	0.004
$\tau = 0.2, \rho = 0.7$	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\tau = 0.7, \rho = 0.7$	0.284	0.154	0.094	0.048	0.039	0.032	0.020	0.014	0.005

TABLE 4.3: Design 3 Variance component of 4.31

	k								
	10	20	30	40	50	60	70	80	90
$\tau = 0.2, \rho = 0.2$	1129	2656	955.0	660.8	198.6	86.47	50.48	18.76	5.562
$\tau = 0.2, \rho = 0.7$	477.1	58.81	13.35	5.312	1.967	0.837	0.393	0.174	0.050
$\tau = 0.7, \rho = 0.7$	7075	7194	1001	421.8	177.5	86.98	43.80	20.12	7.121

TABLE 4.4: Design 4 Variance component of 4.31

	k								
	10	20	30	40	50	60	70	80	90
$\tau = 0.2, \rho = 0.2$	160.5	90.46	48.64	34.34	26.34	17.41	15.66	8.418	3.683
$\tau = 0.2, \rho = 0.7$	3.228	1.188	0.845	0.468	0.278	0.222	0.134	0.083	0.036
$\tau = 0.7, \rho = 0.7$	309.8	120.7	88.82	71.00	34.26	18.77	17.18	5.988	4.394

One of the first things that stands out from the results is how the variances improve as k increases and is likely due to how the expression inside the variance term is similar

in structure to that of the bias which improves as k approaches p. This is in contrast to the first component of 4.31 which can be seen to worsen as k increases so this feature makes bias-variance trade off of PRP somewhat unclear with regards to using k as a means of controlling this. Secondly, as predicted, it appears that when the sum of all the β_{-j} components are larger then so is the entire variance component. This can be seen how typically the values in Table 4.3 are greater than that of the counterparts in all other tables while Design 2 has the smallest values when previously it was shown that $\sum_{i=1}^{p-1} \beta_{-j,i} = 0$ for Design 2. Finally, it is clear from the results that the variance component is smallest when $\rho = 0.7$ and $\tau = 0.2$ for all designs. This is also consistent with the bias properties discussed in 4.3.1 and further supports the argument of there being situations that PRP can thrive in and act as a useful alternative to other methods.

4.3.2.2 Comparison to other methods

Similar to section 4.3.1.2 the variance of PRP will be compared to that of marginal least squares and Ridge. Firstly, one can see that from the estimator for marginal least squares in 4.17 the variance is given by the following:

$$Var(\hat{\beta}_j^{mls}) = \frac{\sigma_\epsilon^2}{n}. \tag{4.32}$$

Therefore, based on the discussion from Section 4.3.2.1, the variance of marginal least squares will always be smaller than that of PRP with the exception of when $\tau = 0$. This is interesting as in Section 4.3.1.2 it was shown that the PRP bias will always be smaller than that of marginal least squares, therefore, one can make the generalisation that PRP is more effective when $\sum_{i \neq j}^{p} \beta_i$ is large relative to σ_ϵ^2, or rather that the signal-to-noise ratio (SNR) is greater since the bias gains will outweigh the variance differences.

For Ridge, it is much more complex, under the toy model, the variance for the estimator of a single predictor using Ridge is given as follows:

$$Var(\hat{\beta}_j(\lambda_R)) = \\ \frac{\sigma_\epsilon^2}{n} \frac{(1+\lambda_R+(p-2)\rho)^2 - 2\tau\rho(p-1)(1+\lambda_R+(p-2)\rho) + \tau^2(p-1)(1+(p-2)\rho)}{((1+\lambda_R)^2+(p-2)\rho(1+\lambda_R)-(p-1)\tau^2)^2}. \tag{4.33}$$

Firstly, one can see how the above varies with respect to λ_R by considering separate derivatives of the numerator and denominator. The numerator first derivative is given

by the following expression:

$$2\big(\lambda_R + 1 + \rho\left((p-2) - \tau(p-1)\right)\big), \tag{4.34}$$

with the denominator equivalent given below,

$$2\big(2(1+\lambda_R) + \rho(p-2)\big)\big(1 + \lambda_R^2 + \rho(p-2) + \lambda_R\left(2 + (p-2)\rho\right) + \tau^2(p-1)\big). \tag{4.35}$$

This can be written as follows to facilitate comparisons between the 2 expressions,

$$\big(2(1+\lambda_R+\rho(p-2)) + 2(1+\lambda_R)\big)\big(1 + \lambda_R^2 + \rho(p-2) + \lambda_R\left(2 + (p-2)\rho\right) + \tau^2(p-1)\big),$$

which can be rewritten in the following form:

$$\begin{aligned} &2(1+\lambda_R+\rho(p-2))\big(1 + \lambda_R^2 + \rho(p-2) + \lambda_R\left(2 + (p-2)\rho\right) + \tau^2(p-1)\big) \\ &\quad + 2(1+\lambda_R)\big(1 + \lambda_R^2 + \rho(p-2) + \lambda_R\left(2 + (p-2)\rho\right) + \tau^2(p-1)\big). \end{aligned} \tag{4.36}$$

From here, one can use the fact from A5 that for $\lambda_R \geq 1$ then
$1 + \lambda_R^2 + \rho(p-2) + \lambda_R\left(2 + (p-2)\rho\right) + \tau^2(p-1) > 1$ (this will also very likely be the case for when $0 \leq \lambda_R \leq 1$). Therefore, it can be seen that for $\lambda_R \geq 1$ then the denominator derivative will be much larger than that of the numerator with both always being positive. As a result, it can be concluded that with λ_R increasing, the variance of Ridge shrinks rapidly allowing potential for large gains over PRP. To give this more perspective, a possible worst case scenario is considered whereby $\lambda_R = 0$ to see how the variance of Ridge behaves with respect to ρ and τ. When $\lambda_R = 0$, the variance expression given in 4.33 becomes the following:

$$\begin{aligned} &Var(\hat{\beta}_j(\lambda_R)) = \\ &\quad \frac{\sigma_\epsilon^2}{n}\frac{(1+(p-2)\rho)^2 - 2\tau\rho(p-1)(1+(p-2)\rho) + \tau^2(p-1)(1+(p-2)\rho)}{\left(1+(p-2)\rho-(p-1)\tau^2\right)^2}. \end{aligned} \tag{4.37}$$

Rearranging the above leads to the following expression making it more comparable to PRP,

$$Var(\hat{\beta}_j(\lambda_R)) = \frac{\sigma_\epsilon^2}{n}\Big(1 + \frac{\left(1+(p-2)\rho\right)(p-1)\left(3\tau^2 - 2\tau\rho\right) - (p-1)\tau^2}{\left(1+(p-2)\rho-(p-1)\tau^2\right)^2}\Big), \tag{4.38}$$

where one can see that for $0 \leq \tau \leq \frac{2}{3}\rho$ then the fraction term within the brackets becomes negative meaning that the Ridge variance term is less than the marginal least squares benchmark of $\frac{\sigma_\epsilon^2}{n}$ and, therefore, will always be lower than that of PRP. Based on the condition relating ρ and τ^2 in A5, this requirement will occur more of often than not making Ridge have a lower variance, even when $\lambda_R = 0$. This combined with the previous analysis revealing how the variance is always decreasing in λ_R for

$\lambda_R \geq 1$, it is clear that Ridge has a significant advantage over PRP in terms of variance.

This section has shown that Ridge (and marginal least squares) have an advantage over PRP when $\tau \neq 0$, therefore, for PRP to outperform them, its gains will come from its promising potential for a lower bias. This means that, already, one can characterise scenarios in which PRP provides a lower MSE by situations where the the SNR is greater since the bias magnitude, inflated by β_j and $\sum_{i \neq j}^{p} \beta_i$, is a larger component than the variance controlled by $\frac{\sigma_\epsilon^2}{n}$. The following section provides simulation evidence to see how these opposite approaches to the bias-variance trade off perform in an environment where one is focused on the estimation of a single parameter with a high dimensional set of controls mimicking a causal inference study.

4.4 Simulation Evidence

As mentioned previously, this section is devoted to replicating a causal inference study whereby one is interested in estimating the sign and magnitude of the treatment variable parameter as accurately as possible. A range of approaches will be used to compute this coefficient and will be compared based on how close their estimate of the treatment variable coefficient is to that of the value under the true DGP. Multiple experiments are carried out with the correlation amongst the predictors, signal-to-noise ratio and sparsity in the true coefficients varied throughout. The experiments are split into 2 separate classes concerned with the profile of true parameters. The first is concerned with all active coefficients being homogeneous while the second involves all non-zero coefficients varying significantly in sign and magnitude. From the analysis of the previous section, one would expect PRP to thrive more when the the coefficients have mixed sign and magnitude due to the bias gains over other methods. To facilitate this, a high SNR would likely also favour PRP as the relativity of the coefficients size to the error variance would mean that the bias is a larger proportion of the MSE where PRP has an advantage.

The DGP is defined by the linear expression identical to 4.1 as follows:

$$y = X\beta + \epsilon\,, \tag{4.39}$$

where, as mentioned previously, X is a $n \times p$ matrix of predictors and y is a $n \times 1$ vector of the dependent variable observations. The coefficient profile is given in the $p \times 1$ vector, β, and the stochastic error terms are given in the vector ϵ. To begin, the

predictor matrix X is simulated from a multivariate normal distribution as follows:

$$X_{n\times p} \sim N(0, \Sigma_X)\,, \tag{4.40}$$

where the covariance matrix is defined similar to that used in the toy model setting by the following:

$$\Sigma_X = \begin{bmatrix} 1 & \tau & \tau & \dots & \tau \\ \tau & 1 & \rho & \dots & \rho \\ \tau & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \tau & \rho & \rho & \dots & 1 \end{bmatrix}, \tag{4.41}$$

with τ and ρ varied across experiments. At this point, it is worth mentioning that the expression in 4.39 can also be reformulated in similar fashion to 4.5 as seen in 4.42. Here $X = (x_j\ X_{-j})$ and $\beta = (\beta_j\ \beta'_{-j})'$ with x_j and β_j being the jth column of X and jth element of β respectively while X_{-j} and β_{-j} represent all other covariates and coefficients respectively.

$$y = x_j\beta_j + X_{-j}\beta_{-j} + \epsilon\,. \tag{4.42}$$

This is particularly useful since the focus of this study is on the estimation of a single covariate of interest which is β_j in this case. It is also important to note that, by the design of 4.41, the covariance between x_j and all other predictors is τ while the correlation between all predictors besides x_j is given by ρ when $j = 1$. Therefore, an additional layer of fine tuning has been added to investigate how the relative correlations influence estimation accuracy as Section 4.3 showed how significant this is likely to be. Specifically, all experiments carried out consider 3 separate profiles of τ and ρ, all of which satisfy the conditions outline below A5 for covariance matrix feasibility. The first is where $\tau = 0.2$ and $\rho = 0.7$ as is this is believed to favour PRP over other methods as discussed previously and replicates a setting where the controls variable are closely correlated with each other but the treatment variable has little correlation with the controls. One might expect this to be the case where many of the control variables are similar measures, for example, in a macroeconomic setting one might wish to use inflation as a control but will likely have multiple price indices available which will be highly correlated with each other. Another setting is where both τ and ρ are equal to 0.2 replicating an environment with all-round little correlation. Although this is less realistic since, in high dimensional settings, there is usually significant correlation present purely by chance. Finally, a more realistic setting where both τ and ρ are equal to 0.7 is used to create an environment where all the variables are highly correlated, proving helpful for some methods but negatively impacting the estimation accuracy of others.

For the coefficients in β, 2 classes of profiles are considered with attention given to the

proportion of variables deemed as being active (variables that have a non-zero true coefficient) as well as the sign and magnitude of these active coefficients. Firstly, let s be defined as the proportion of coefficients in β that are non-zero for $s \in (0, 1)$. Multiple experiments are considered with s varied over $(0.1, 0.4, 0.8)$ to see how the degree of sparsity can influence the performance of PRP and its competitors. While there is no direct implication of this detailed in Section 4.3, one would expect that it has the potential to exaggerate the gap in computational accuracy's of PRP and the other various methods with implications on when PRP may be most suitable. For the sp active predictors, the 2 designs are as follows:

- **Design 1: Homogeneous Coefficients**
  Here, the first sp coefficients of β all take a value of $\frac{1}{\sqrt{n}}$ with the rest equal to 0. Under the formulation of 4.42, $\beta_1 = \frac{1}{\sqrt{n}}$ so one is always interested in the estimation of an active coefficient. As discussed earlier, this setting is less favourable to PRP due to less potential relative bias gains over methods such as Ridge, however, it is still interesting to see how its performance varies over other factors of the true DGP.

- **Design 2: Mixed Sign and Magnitude Coefficients**
  Under this design, the coefficients in β are computed as follows for $i = 1, \ldots, p$,

$$\beta_i = \begin{cases} \frac{(-1)^i(sp+1-i)}{\sqrt{n}} & \text{for } i \leq sp \\ 0 & \text{otherwise} \end{cases}. \tag{4.43}$$

  to illustrate this with a small example, when $sp = 5$ then the non-zero components of β would be $\beta^{active} = (-5, 4, -3, 2, -1)'$ (assuming $n = 1$ for simplicity here) and since β_1 is the first component of β then $\beta_1 = -5$. In this setting, one would expect PRP to estimate β_1 more accurately relative to other methods that suffer through the bias when the true coefficients vary significantly in sign and magnitude. Finally, in both of these designs, the inclusion of the reciprocated $\sqrt{n}$ term is to keep the coefficients local to 0 which is necessary to keep the bias from using a low dimensional subspace finite (Boot and Nibbering (2019)) for the purpose of RP and PRP.

Regarding the disturbance term in 4.39, it is such that each element of ϵ is simulated from an independent normal distribution such that $\epsilon \sim N(0, \sigma_\epsilon^2 I_n)$. Regarding the variance term, σ_ϵ^2, the following definition of signal-to-noise ratio (SNR) is first defined,

$$SNR = \frac{\beta' \Sigma_X \beta}{\sigma_\epsilon^2}. \tag{4.44}$$

Therefore, using the above equality, 3 separate experiments are carried out with the SNR fixed at $1, 5$ and 10 to replicate different DGP environments that might be face.

One would expect PRP to perform well in environment of a high SNR where the magnitude of β components is large relative to that of σ_ϵ^2 making the estimator biases account for a larger proportion of inaccuracy, where PRP is known to carry benefits over other approaches.

For all experiments, the analyst is faced with 200 candidate predictors ($p = 200$) and a sample size of $n = 100$ ensuring a high dimensional setting. As mentioned previously, one is interested solely in the estimation of the first coefficient in the coefficient matrix as seen in 4.42 with $j = 1$ since x_j is assumed to be the treatment variable and the methods used to estimate this include: Marginal Ordinary Least Squares (MOLS), Ridge, Lasso, Partial Ridge (from previous work), Random Projections (RP) and PRP. For MOLS, an estimate of β_1 is computed by regressing only x_1 on y while Ridge and Lasso estimate the whole profile of β followed by the estimate of β_1 taken out for comparison purposes. For both Ridge and Lasso, 10 fold cross validation is used to determine the penalty parameter while for Partial Ridge a grid with varying interval sizes from 0.1 to 2000 is used to compute multiple estimates of β_1. The penalty parameter that provides the lowest mean-square-error (MSE) of β_1 is used as the final estimate representing Partial Ridge with MSE defined shortly.

Finally, for RP and PRP, the subspace dimension, k, is also chosen by considering the grid $k \in (10, 20, 30, 50, 70, 90)$ meaning that for each simulation there will be 6 estimates for both RP and PRP with the one providing the lowest MSE reported. When computing an estimate for each subspace dimension size for either RP or PRP, 100 draws of the random matrix R are used leading to 100 preliminary estimates of β_1, these are then averaged to obtain a single estimate for that simulated data set. For each method in a given experiment, the best of the 6 estimates are reported so the value of k for each result is either $10, 20, 30, 50, 70$ or 90 and this need not be the same for both RP and PRP. It is also important to mention that, for all methods besides MOLS[1], both the dependent and independent variable are standardised before parameter estimation. Following this, the estimated value of β_1 is multiplied by the ratio of standard deviations, $\frac{sd(y)}{sd(x_1)}$, in order to obtain a parameter estimation on the same scale as the original data.

Regarding the metric of performance comparison, the main measure used is the mean-squared error (MSE) with the lowest MSE representing a higher accuracy. This is defined as the squared difference between the estimated coefficient from a given method ($\hat{\beta}_1$) and the true value of the coefficient (β_1) averaged over the total number

[1] Although the experiments were repeated when standardisation is used for MOLS but there are no obvious differences in the results.

of simulations, N. This is shown by the following expression:

$$MSE(\hat{\beta}_1) = \frac{1}{N}\sum_{l=1}^{N}(\hat{\beta}_1 - \beta_1)^2\,, \tag{4.45}$$

where, for the experiments in this study, there are a total of 100 simulations ($N = 100$). This can be viewed as the main way to measure estimation accuracy and will be the priority for making conclusions from the results. Other related measures used include the bias and variance of the estimators for each approach defined as follows:

$$Bias(\hat{\beta}_1) = E[\hat{\beta}_1] - \beta_1\,, \tag{4.46}$$

$$Var(\hat{\beta}_1) = \frac{1}{N}\sum_{l=1}^{N}(\hat{\beta}_1 - E[\hat{\beta}_1])^2\,, \tag{4.47}$$

where the mean of the chosen estimator across simulations is given by $E[\hat{\beta}_1] = \frac{1}{N}\sum_{l=1}^{N}\hat{\beta}_1$. One can use the well-known relationship to show that $MSE(\hat{\beta}_1) = Bias(\hat{\beta}_1)^2 + Var(\hat{\beta}_1)$ and this can be seen to be in the case in the results with only occasional minor discrepancies due to rounding. The measures in 4.46 and 4.47 will help provide insight into the composition of the MSE to facilitate the explanation of the MSE hierarchy for each experiment as it has already been shown that the main channel of improvement for PRP comes from the bias compared to other methods such as Ridge. The following 2 subsections report the MSEs for each experiment with relevant discussions with tables showing the bias and variance values relegated to the appendix.

4.4.1 Design 1: Homogeneous Coefficients

Tables 4.5-4.7 shows the MSEs for each method across each experiment for Design 1 of the β profile with each table representing a different environment with respect to the signal-to-noise ratio. As expected PRP generally struggles here for reasons associated with the relativity of β_1 to $\sum_{i\neq 1}^{p}\beta_i$ as discussed previously. Looking more closely, it can be seen from Table 4.5 that, for $s = 0.1$ and $\tau = 0.2$, PRP remains closely competitive with the other methods with the second results column showing it provides the second lowest MSE of 0.0045 (52% of the that of Lasso). However, as the density of the coefficient profile increases, the MSEs of methods such as Lasso and RP change little while the PRP increases significantly to 0.3381 in the second column from the right (649% of that of Lasso). This is understandable since s increasing means that the relativity of β_1 to $\sum_{i\neq 1}^{p}\beta_i$, $\frac{\beta_i}{\sum_{i\neq 1}^{p}\beta_i}$, becomes smaller removing its potential for bias gains. Such a feature is easy to see in this setting of homogeneous coefficients where the increase in s leads to a linear increase in the size of $\sum_{i\neq 1}^{p}\beta_i$ (with β_1 still fixed). Although, one would expect that, for a fixed SNR, the sparsity increasing would also reduce the error variance leading to more emphasis on the bias which favours PRP.

Therefore, this shows how the coefficient profile is more influential on performance than the SNR.

Regarding the correlation structure, it is visible that PRP performs relatively better when $\tau = 0.2$, particularly when $\tau = 0.2$ also. For example, consider Table 4.6 when $s = 0.1$ the MSE of RP is 73% of the PRP MSE when $\tau = 0.2$ and $\rho = 0.2$, 57% when $\tau = 0.2$ and $\rho = 0.7$, and 24% when τ and ρ are both 0.7 with and identical pattern seen in the majority of other settings. This supports what was discussed previously in how τ and the bias bound of PRP are positively linearly related. However, one may question this in light of the theoretical analysis where it was clearly shown that the bias and variance bounds of PRP improved upon the ratio of $\frac{\rho}{\tau}$ being larger. This is explained by how the design of the experiment being such that SNR determination of σ_ϵ^2 is influenced by Σ_X as seen in 4.44. Therefore, a higher ρ results in a larger error variance to keep the SNR fixed at a certain value since all true coefficients and ρ are positive under this design. This implies that when $\rho = 0.7$ the size of σ_ϵ^2 is larger relative the coefficient size than when $\rho = 0.2$ making the bias gains less apparent, hence, explaining why PRP performs stronger under $\rho = 0.2$ when $\tau = 0.2$. Despite this, in an empirical setting, where the SNR was not fixed, one would expect PRP to improve as $\frac{\rho}{\tau}$ increases with σ_ϵ^2 remaining unchanged.

Tables 4.A1-4.A3 show the bias and variance associated with each method for all the experiments and confirms the conclusions of the analysis before whereby PRP provides the lowest bias for almost all experiments when $s = 0.1$ reinforcing the justification for the best relative performance in MSEs. Across all SNR values, when $s = 0.1$ it is only when $\tau = \rho = 0.7$ and when the SNR is equal to 1 or 5 that it does not provide the lowest bias which is justified by the previous arguments. However, these tables also show how PRP very often has one of the largest variances, especially when $\tau = \rho = 0.7$ where it provides the highest the vast majority of the time. This reinforces what was discussed in Section 4.3.2 where the dual component of the PRP leaves it vulnerable to estimation uncertainty through even the value of true coefficients themselves, unlike most other approaches. One can see from the second component of 4.31 and the following study in Tables 4.1-4.4 that the total sum of true coefficients combined with an unfavourable correlation profile quickly leads to the variance term increasing rapidly unless k is very close to p.

TABLE 4.5: MSE for experiments under Design 1 with $SNR = 1$

	s=0.1			s=0.4			s=0.8		
τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	0.2068	0.1482	1.8182	3.2922	3.0081	31.245	13.533	10.494	126.93
Ridge	0.0216	0.0066	0.0260	0.3210	0.0696	0.3347	1.2834	0.2515	1.1026
Lasso	0.0133	0.0086	0.0172	0.0627	0.0196	0.0667	0.2135	0.0521	0.1994
PR	0.0206	0.0049	0.0560	0.2796	0.1124	0.8656	1.2994	0.3754	2.7868
RP	0.0035	0.0030	0.0034	0.0131	0.0027	0.0056	0.0321	0.0028	0.0110
PRP	0.0195	0.0045	0.0630	0.3148	0.1000	0.95551	1.3183	0.3381	2.8553

TABLE 4.6: MSE for experiments under Design 1 with $SNR = 5$

	s=0.1			s=0.4			s=0.8		
τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	0.1511	0.1595	1.7562	3.2927	2.6983	30.289	13.974	10.254	125.822
Ridge	0.0063	0.0030	0.0074	0.0683	0.0143	0.0922	0.2667	0.0429	0.3290
Lasso	0.0074	0.0052	0.0081	0.0177	0.0143	0.0325	0.0900	0.0218	0.0991
PR	0.0041	0.0037	0.0111	0.0547	0.0354	0.1630	0.2927	0.0931	0.8642
RP	0.0028	0.0024	0.0028	0.0045	0.0014	0.0035	0.0077	0.0010	0.0046
PRP	0.0049	0.0033	0.0115	0.0640	0.0306	0.1586	0.2947	0.0922	0.8660

TABLE 4.7: MSE for experiments under Design 1 with $SNR = 10$

	s=0.1			s=0.4			s=0.8		
τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	0.1542	0.1531	1.7869	2.8750	2.5730	30.305	11.790	10.238	126.31
Ridge	0.0038	0.0035	0.0059	0.0327	0.0085	0.0389	0.1177	0.0240	0.1873
Lasso	0.0055	0.0035	0.0087	0.0146	0.0101	0.0200	0.0401	0.0255	0.1112
PR	0.0016	0.0010	0.0061	0.0219	0.0234	0.1175	0.1249	0.0764	0.4758
RP	0.0020	0.0028	0.0032	0.0038	0.0014	0.0017	0.0040	0.0005	0.0026
PRP	0.0021	0.0011	0.0057	0.0270	0.0207	0.1020	0.1376	0.0721	0.3889

Overall, it is clear that under a coefficient profile of homogeneous coefficients, PRP rarely provides competitive estimation accuracy for the treatment variable parameter due to the ratio of β_1 to $\sum_{i\neq 1}^{p} \beta_i$ being unfavourable for the bias and very little scope to gain any edge through the variance term. While it is clear that environments where the predictor sparsity is low and the correlation between the treatment variable and

controls (x_1 and X_{-1} respectively) is also small then there is some hope, as this keeps the bias and variance of PRP down, ultimately, the coefficient profile is most significant in making PRP highly undesirable in this setting.

4.4.2 Design 2: Mixed Sign and Magnitude Coefficients

Tables 4.8-4.10 shows the MSEs for each method across each experiment for Design 2 of the β profile with each table representing a different environment with respect to the signal-to-noise ratio. Here, it can be seen that PRP is much more successful in providing a low MSE relative to the other methods and provides the lowest value in the majority of experiments. This makes sense based on how the coefficient profile has an even mixture of signs in β_{-j} contributing a smaller $\sum_{i\neq 1}^{p}\beta_i$ relative to β_1. Specifically, PRP is the most accurate method whenever $\tau = 0.2$ with the exception of when the $SNR = 1$ and $s = 0.8$ where RP dominates. This is likely due to the reasons mentioned previously where the low SNR makes the error variance high compared to the coefficients and the dense setting will make $\sum_{i\neq 1}^{p}\beta_i$ larger relative to β_i adversely impacting the bias of PRP making a worst case scenario for PRP in this class of experiments. Therefore, this does show that, while the coefficient profile is highly significant in the performance of PRP, the correlation can override this benefit in some cases.

The results are similar to before in the sense that PRP has an improving edge as the level of sparsity decreases. For example, in Table 4.9 for $\tau = 0.2$, $\rho = 0.7$ and $s = 0.8$ the MSE of PRP is 99% of the MSE of the second best approach (Partial Ridge) but for the corresponding correlation profile when $s = 0.1$ the MSE of PRP is 88% of the second lowest MSE (also Partial Ridge). At this stage it is also worth mentioning how PRP and Partial Ridge follow a similar performance pattern relative to the other approaches. This is likely due to how they are similar in that they estimate a single parameter while applying penalisation or dimension reduction to the remaining variables giving them similar statistical properties. This is particularly evident in how they both enjoy benefits in through their bias under the same coefficient conditions.

Tables 4.A4-4.A6 show that, with the exception of Partial Ridge, PRP provides a significantly lower bias than all other methods for all the experiments, even in the worst of scenarios for PRP such as the last column of Table 4.A4 PRP provides a bias of 2.3990 compared to 14.8678 and 10.7250 of Lasso and Ridge respectively. Unlike for Design 1, PRP is much more competitive on the variance with the other methods in this study, for example, in the first column of Table 4.A4, it provides the lowest variance of 0.0944. The only area where it still struggles is when τ and ρ are both equal

to 0.7 where very often it provides the highest variance regardless of the sparsity or SNR.

TABLE 4.8: MSE for experiments under Design 2 with $SNR = 1$

	s=0.1			s=0.4			s=0.8		
τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	0.1797	0.4889	0.6583	10.296	33.041	16.798	110.04	324.36	121.57
Ridge	0.4009	1.3521	1.3617	12.297	25.825	27.440	83.082	146.07	180.50
Lasso	0.3558	1.5856	1.6317	26.316	52.727	54.610	146.47	235.93	244.40
PR	0.0989	0.2798	0.2330	6.3112	20.631	8.8849	66.175	166.49	135.93
RP	0.4144	1.2095	1.2002	9.0488	19.303	21.132	51.618	89.707	110.38
PRP	0.0947	0.2798	0.2528	6.1357	19.754	10.378	65.925	162.596	145.53

TABLE 4.9: MSE for experiments under Design 2 with $SNR = 5$

	s=0.1			s=0.4			s=0.8		
τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	0.1253	0.2014	0.6006	6.6934	19.147	15.280	60.949	122.73	74.219
Ridge	0.3951	1.1095	1.2989	7.8088	20.678	20.89	42.478	93.477	111.63
Lasso	0.0766	0.2723	0.3868	9.4000	33.085	33.344	101.01	195.77	195.49
PR	0.0537	0.1022	0.0754	3.2308	8.7426	6.3569	30.407	41.761	94.453
RP	0.3821	1.1013	1.2640	7.0100	19.892	19.894	36.381	83.167	94.453
PRP	0.0473	0.0938	0.0924	2.9856	8.0680	7.5723	30.224	36.220	80.798

TABLE 4.10: MSE for experiments under Design 2 with $SNR = 10$

	s=0.1			s=0.4			s=0.8		
τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	0.1226	0.2801	0.5448	5.7558	13.081	12.707	37.336	106.33	64.050
Ridge	0.3685	1.2183	1.3302	6.2686	20.081	20.442	31.611	102.64	87.834
Lasso	0.0442	0.1916	0.2167	6.1641	29.654	28.464	74.503	180.76	169.92
PR	0.0274	0.1192	0.0926	1.9621	4.2853	5.4050	15.401	53.851	44.710
RP	0.3279	1.3080	1.2635	5.8501	20.678	19.638	28.676	82.047	78.920
PRP	0.0264	0.1078	0.0967	1.6219	4.0851	6.2565	13.305	53.755	48.205

Despite the first design of coefficients showing little hope for PRP in terms of estimation accuracy, the second design is much more promising with PRP being the superior method in the vast majority of situations. This can be brought down the bias

variance trade off in most cases where earlier it was discussed that the PRP has an advantage through lower bias when β_1 is not too small compared to $\sum_{i\neq 1}^{p} \beta_i$ compared to methods such as Ridge and Lasso. This can then be supported by a scenario where the correlation between x_1 and variables in X_{-1} is also low and a high SNR environment also helps the cause. Overall, this section has characterised situations where PRP can be the best approach to accurately estimating a single coefficient if one has a certain level of prior knowledge about the data set concerned.

4.5 Discussion

This paper has proposed a new estimation procedure for a single parameter in a linear regression framework that is characterised as being of a high-dimensional nature. Specifically, this has sought to adapt the widely used approach of Random Projections from the machine learning literature where previously the only work on Random Projections in a regression setting was focused around forecasting and prediction of the dependent variable. By incorporating this approach into a Frisch-Waugh style method for individual parameter estimation, this work has looked to define a method that not only overcomes the issue of high-dimensionality that hinders OLS based methods but also avoid the issue of inflated biases under certain DGP conditions that approaches such as Ridge face frequently. This has lead to the proposal of the Partial Random Projections estimation procedure which, through theoretical analysis of the bias and variance as well as finely tuned simulation experiments, has proved to be very effective at single parameter estimation compared to many well-established methods from the literature. More specifically, the conditions where this is most apparent include when: the true value of the coefficient of interest is not too small relative to the sum of all other true coefficients, the correlation between the variable of interest and all others is small and, finally, the ratio of signal-to-noise in the true DGP is higher then these gains are amplified. Situations where such conditions are present in economics settings can be easily identified, for example, if the predictor set includes price and quantity data on multiple complementary and substitute products then the sign of coefficients will likely vary significantly in conjunction with the setting created in the second design of the simulation study.

While this study acts as the initial stage for establishing an approach for computing individual coefficient estimates using Random Projections, one can easily realise the broad range of directions that could be explored through future work. One path could be devoted to investigating additional theoretical properties of the estimator itself. This might include hypothesis testing for statistical significance or determining the asymptotic distribution of the PRP estimator. In addition, comparisons could be made with the non-parametric coefficient distributions obtained from the bootstrap

approaches seen in Mammen (1993). Due to these approaches being similar in spirit to that of RP and PRP, a comparison between the asymptotic behaviour of these methods has the potential to uncover more about stochastic based shrinkage in a broader sense.

4.A Appendix

4.A.1 Proof of Proposition 4.1

The least squares regression of 4.2 results in the following estimator classification for a single draw of R,

$$\hat{\delta}^s_{PRP} = \begin{bmatrix} \hat{\beta}_j \\ \hat{\beta}_{-j} \end{bmatrix} = \begin{bmatrix} x_j'x_j & x_j'X_{-j}R^s \\ R^{s\prime}X_{-j}'x_{-j} & R^{s\prime}X_{-j}'X_{-j}R^s \end{bmatrix}^{-1} \begin{bmatrix} x_j'y \\ R^{s\prime}X_{-j}'y \end{bmatrix}.$$

Now let $\hat{\delta}^s_{PRP} = (W_R'W_R)^{-1}W_R'y$ where $W_R = [x_j \; X_{-j}R^s]$. One can use this with the Partial Ridge model approach to 4.1 as follows,

$$y - x\hat{\beta}_j - X_{-j}R^s\hat{\beta}_{-j} = y - W_R\hat{\delta}^s_{PRP} = y - W_R(W_R'W_R)^{-1}W_R'y$$

Letting $P_{W_R} = W_R(W_R'W_R)^{-1}W_R'$ the following expression can be seen,

$$y - x\hat{\beta}_j - X_{-j}R^s\hat{\beta}_{-j} = (I_n - P_{W_R})y\,.$$

Now define $P_x = x_j(x_j'x_j)^{-1}x_j'$, one can show that $P_{W_R}P_x = P_x$. Therefore, the above equation can be rearranged and rewritten as follows,

$$x_j\hat{\beta}_j = y - X_{-j}R^s\hat{\beta}_{-j} - (I_n - P_{W_R})y\,.$$

Now multiply both sides by P_x to obtain

$$P_xx_j\hat{\beta}_j = P_x(y - X_{-j}R^s\hat{\beta}_{-j}) - P_x(I_n - P_{W_R})y = P_x(y - X_{-j}R^s\hat{\beta}_{-j})\,.$$

Expanding the P_x term gives the following:

$$x_j(x_j'x_j)^{-1}x_j'x_j\hat{\beta}_j = x_j(x_j'x_j)^{-1}x_j'(y - X_{-j}R^s\hat{\beta}_{-j})\,,$$

$$x_{-j}\hat{\beta}_j = x_j(x_j'x_j)^{-1}x_j'(y - X_{-j}R^s\hat{\beta}_{-j})\,,$$

$$\hat{\beta}_j = (x_j'x_j)^{-1}x_j'(y - X_{-j}R^s\hat{\beta}_{-j})\,.$$

4.A.2 Proof of Proposition 4.2

Recall the expression for $\hat{\beta}_j$ as being given by the following when combining 4.3 and 4.4,

$$\hat{\beta}_j = (x_j'x_j)^{-1}x_j'\left(I_n - X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'M_x\right)y\,.$$

Replacing y with the true model from 4.1 and using the fact that $M_x x_j = 0$ gives

$$\hat{\beta}_j = \beta_j + (x_j' x_j)^{-1} x_j' \left(I_n - X_{-j} R (R' X_{-j}' M_x X_{-j} R)^{-1} R' X_{-j}' M_x \right) (X_{-j}\beta_{-j} + \epsilon) .$$

Taking the expectation with respect to ϵ allows the use of A1 leading to the following simplification,

$$E_\epsilon[\hat{\beta}_j | R] = \beta_j + (x_j' x_j)^{-1} x_j' \left(I_n - X_{-j} R (R' X_{-j}' M_x X_{-j} R)^{-1} R' X_{-j}' M_x \right) X_{-j}\beta_{-j} .$$

Now one can apply the law of conditional expectation to obtain a complete expression for the bias,

$$E_R \left[E_\epsilon[\hat{\beta}_j] \right] = \beta_j + (x_j' x_j)^{-1} x_j' \left(I_n - X_{-j} E_R[R (R' X_{-j}' M_x X_{-j} R)^{-1} R'] X_{-j}' M_x \right) X_{-j}\beta_{-j} .$$

Therefore, the bias can be expressed as follows,

$$Bias(\hat{\beta}_j) = E_R \left[E_\epsilon[\hat{\beta}_j] \right] - \beta_j = (x_j' x_j)^{-1} x_j' \left(I_n - X_{-j} E_R[R (R' X_{-j}' M_x X_{-j} R)^{-1} R'] X_{-j}' M_x \right) X_{-j}\beta_{-j} .$$

Which can be rearranged as the expression in Proposition 4.2,

$$Bias(\hat{\beta}_j) = (x_j' x_j)^{-1} x_j' X_{-j} \left(\beta_{-j} - E_R[R (R' X_{-j}' M_x X_{-j} R)^{-1} R'] X_{-j}' M_x X_{-j} \beta_{-j} \right) .$$

4.A.3 Proof of Theorem 4.1

Recall the bias expression for Partial Random Projections is given as follows from Proposition 4.2,

$$Bias(\hat{\beta}_j) = (x_j' x_j)^{-1} x_j' X_{-j} \left(\beta_{-j} - E_R[R (R' X_{-j}' M_x X_{-j} R)^{-1} R'] X_{-j}' M_x X_{-j} \beta_{-j} \right) .$$

From here one can decompose $X_{-j}' M_x X_{-j}$ as detailed below Proposition 4.2 and use the orthogonality of U to replace R with UR as shown in Marzetta et al. (2011),

$$Bias(\hat{\beta}_j) = (x_j' x_j)^{-1} x_j' X_{-j} U \left(I_{p-1} - E_R[R (R' D R)^{-1} R'] D \right) U' \beta_{-j} .$$

Using the result from Marzetta et al. (2011) and writing the above as matrices, the following is obtained,

$$Bias(\hat{\beta}_j) = \begin{bmatrix} \tau & \tau & \dots & \tau \end{bmatrix} U \begin{bmatrix} 1-\frac{\lambda_1}{\eta_1} & 0 & 0 & \dots & 0 \\ 0 & 1-\frac{\lambda_2}{\eta_2} & 0 & \dots & 0 \\ 0 & 0 & 1-\frac{\lambda_3}{\eta_3} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1-\frac{\lambda_p}{\eta_{p-1}} \end{bmatrix} U'\beta_{-j}.$$

By looking closer at $X'_{-j}M_xX_{-j}$, one can easily find the eigenvalues under the toy model assumptions,

$$X'_{-j}M_xX_{-j} = X'_{-j}X_{-j} - X'_{-j}x_j(x'_jx_j)^{-1}x'_jX_{-j} = \begin{bmatrix} 1-\tau^2 & \rho-\tau^2 & \rho-\tau^2 & \dots & \rho-\tau^2 \\ \rho-\tau^2 & 1-\tau^2 & \rho-\tau^2 & \dots & \rho-\tau^2 \\ \rho-\tau^2 & \rho-\tau^2 & 1-\tau^2 & \dots & \rho-\tau^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho-\tau^2 & \rho-\tau^2 & \rho-\tau^2 & \dots & 1-\tau^2 \end{bmatrix},$$

where one can show that $p-2$ of the eigenvalues are equal to $\rho-1$ and one is equal to $1+(p-2)\rho-(p-1)\tau^2$. Without loss of generality, it will be assumed that the largest eigenvalue is $\lambda_1 = 1+(p-2)\rho-(p-1)\tau^2$ meaning the D matrix is given as follows,

$$D = \begin{bmatrix} 1+(p-2)\rho-(p-1)\tau^2 & 0 & 0 & \dots & 0 \\ 0 & \rho-1 & 0 & \dots & 0 \\ 0 & 0 & \rho-1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \rho-1 \end{bmatrix}.$$

From here, the bias expression can be evaluated for small values of p to observe a pattern which then allows generalization of p.

Case 1: p-1=2 Here, the eigenvectors of $X'_{-j}M_xX_{-j}$ are given by the following as the U matrix,

$$U = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Therefore, the bias expression can be evaluated as follows,

$$Bias(\hat{\beta}_j) = \begin{bmatrix} \tau & \tau \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 - \frac{1+\rho-2\tau^2}{\eta_1} & 0 \\ 0 & 1 - \frac{1-\rho}{\eta_2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \beta_{-j}$$
$$= \tau \sum_{i \neq j}^{p} \beta_i \left(1 - \frac{1+\rho-2\tau^2}{\eta_1} \right).$$

Case 2: p-1=3 Here, the eigenvectors of $X'_{-j} M_x X_{-j}$ are given by the following as the U matrix,

$$U = \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \sqrt{\frac{2}{3}} & 0 \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Therefore, the bias expression can be evaluated as follows,

$$Bias(\hat{\beta}_j) = \begin{bmatrix} \tau & \tau & \tau \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \sqrt{\frac{2}{3}} & 0 \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 - \frac{1+2\rho-3\tau^2}{\eta_1} & 0 & 0 \\ 0 & 1 - \frac{1-\rho}{\eta_2} & 0 \\ 0 & 0 & 1 - \frac{1-\rho}{\eta_3} \end{bmatrix}$$
$$\times \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \sqrt{\frac{2}{3}} & -\frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \beta_{-j},$$

$$Bias(\hat{\beta}_j) = \tau \sum_{i \neq j}^{p} \beta_i \left(1 - \frac{1+2\rho-3\tau^2}{\eta_1} \right).$$

Case 3: p-1=4 Here, the eigenvectors of $X'_{-j} M_x X_{-j}$ are given by the following as the U matrix,

$$U = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2\sqrt{3}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} & 0 & 0 \\ \frac{1}{2} & -\frac{1}{2\sqrt{3}} & \sqrt{\frac{2}{3}} & 0 \\ \frac{1}{2} & -\frac{1}{s\sqrt{3}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Therefore, the bias expression can be evaluated as follows,

$$Bias(\hat{\beta}_j) = \begin{bmatrix}\tau\\ \tau\\ \tau\\ \tau\end{bmatrix}' \begin{bmatrix}\frac{1}{2} & -\frac{1}{2\sqrt{3}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}}\\ \frac{1}{2} & \frac{\sqrt{3}}{2} & 0 & 0\\ \frac{1}{2} & -\frac{1}{2\sqrt{3}} & \sqrt{\frac{2}{3}} & 0\\ \frac{1}{2} & -\frac{1}{s\sqrt{3}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}}\end{bmatrix} \begin{bmatrix}1-\frac{1+3\rho-4\tau^2}{\eta_1} & 0 & 0 & 0\\ 0 & 1-\frac{1-\rho}{\eta_2} & 0 & 0\\ 0 & 0 & 1-\frac{1-\rho}{\eta_3} & 0\\ 0 & 0 & 0 & 1-\frac{1-\rho}{\eta_4}\end{bmatrix}$$
$$\times \begin{bmatrix}\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2}\\ -\frac{1}{2\sqrt{3}} & \frac{\sqrt{3}}{2} & -\frac{1}{2\sqrt{3}} & -\frac{1}{2\sqrt{3}}\\ -\frac{1}{\sqrt{6}} & 0 & \sqrt{\frac{2}{3}} & -\frac{1}{\sqrt{6}}\\ -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}}\end{bmatrix} \beta_{-j},$$

$$Bias(\hat{\beta}_j) = \tau \sum_{i\neq j}^{p} \beta_i \left(1 - \frac{1+3\rho-4\tau^2}{\eta_1}\right).$$

Therefore, one can see from the pattern with p increasing, that the general case can be written as

$$Bias(\hat{\beta}_j) = \tau \sum_{i\neq j}^{p} \beta_i \left(1 - \frac{1+(p-2)\rho-(p-1)\tau^2}{\eta_1}\right).$$

4.A.4 Proof of Proposition 4.3

Recall the conditional variance formula which can be written as follows when applied to this framework,

$$Var(\hat{\beta}_j) = E_R\left[Var_\epsilon(\hat{\beta}_j|R)\right] + Var_R\left(E_\epsilon[\hat{\beta}_j|R]\right).$$

One can use the 2 expressions from Proposition 4.1 to write a closed form expression for $\hat{\beta}_j$ in terms of the true DGP as was done in the proof for Proposition 4.2. Recall the following expression from this proof,

$$\hat{\beta}_j = \beta_j + (x_j'x_j)^{-1}x_j'\left(I_n - X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'M_x\right)(X_{-j}\beta_{-j}+\epsilon).$$

Therefore, $\hat{\beta}_j - E_\epsilon[\hat{\beta}_j]$ is given by the following:

$$\hat{\beta}_j - E_\epsilon[\hat{\beta}_j|R] = (x_j'x_j)^{-1}x_j'\left(\epsilon - X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'M_x\epsilon\right).$$

From here, $Var_\epsilon(\hat{\beta}_j|R)$ can be computed as follows,

$$Var_\epsilon(\hat{\beta}_j|R) = E_\epsilon\left[(\hat{\beta}_j - E_\epsilon[\hat{\beta}_j|R])(\hat{\beta}_j - E_\epsilon[\hat{\beta}_j|R])'|R\right],$$

$$
\begin{aligned}
Var_\epsilon(\hat{\beta}_j|R) = E_\epsilon \Big[& (x_j'x_j)^{-1}x_j'\epsilon\epsilon' x_j(x_j'x_j)^{-1} \\
& - (x_j'x_j)^{-1}x_j'\epsilon\epsilon' M_x X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'x_j(x_j'x_j)^{-1} \\
& - (x_j'x_j)^{-1}x_jX_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'M_x\epsilon\epsilon' x_j(x_j'x_j)^{-1} + \\
& (x_j'x_j)^{-1}x_j'X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'M_x\epsilon\epsilon' M_xX_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'x_j(x_j'x_j)^{-1} \Big].
\end{aligned}
$$

Using assumption A1 and the fact that $M_x x_j = 0$ results in the following:

$$
Var_\epsilon(\hat{\beta}_j|R) = \sigma_\epsilon^2 \left((x_j'x_j)^{-1} + (x_j'x_j)^{-1}x_j'X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'x_j(x_j'x_j)^{-1} \right).
$$

Therefore, it is such that

$$
\begin{aligned}
& E_R[Var_\epsilon(\hat{\beta}_j|R)] = \\
& \qquad \sigma_\epsilon^2 \left((x_j'x_j)^{-1} + (x_j'x_j)^{-1}x_j'X_{-j}E_R\left[R(R'X_{-j}'M_xX_{-j}R)^{-1}R' \right] X_{-j}'x_j(x_j'x_j)^{-1} \right).
\end{aligned}
$$

For the second component from the conditional variance formula, recall that, from the proof of Proposition 4.2, the expectation of $\hat{\beta}_j$ conditional on R is given as follows,

$$
E_\epsilon[\hat{\beta}_j|R] = \beta_j + (x_j'x_j)^{-1}x_j' \left(I_n - X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'M_x \right) X_{-j}\beta_{-j}.
$$

Therefore, one can easily achieve the expression for the variance with respect to R seen in Proposition 4.3.

$$
\begin{aligned}
& Var_R(E_\epsilon[\hat{\beta}_j|R]) = \\
& \qquad Var_R\left(\beta_j + (x_j'x_j)^{-1}x_j' \left(I_n - X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'M_x \right) X_{-j}\beta_{-j} \right),
\end{aligned}
$$

$$
Var_R(E_\epsilon[\hat{\beta}_j|R]) = Var_R\left((x_j'x_j)^{-1}x_j'X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'M_xX_{-j}\beta_{-j} \right).
$$

Therefore, combining these 2 components achieves the expression stated in Proposition 4.3,

$$
\begin{aligned}
Var(\hat{\beta}_j) = {} & \sigma_\epsilon^2 \left((x_j'x_j)^{-1} + (x_j'x_j)^{-1}x_j'X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'x_j(x_j'x_j)^{-1} \right) \\
& + Var_R\left((x_j'x_j)^{-1}x_j'X_{-j}R(R'X_{-j}'M_xX_{-j}R)^{-1}R'X_{-j}'M_xX_{-j}\beta_{-j} \right).
\end{aligned}
$$

4.A.5 Simulation Biases and Variances

TABLE 4.A1: Bias and Variances for experiments under Design 1 with $SNR = 1$

		s=0.1			s=0.4			s=0.8		
	τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
	ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	Bias	0.03823	0.3670	1.3347	1.5688	1.6568	5.5410	3.2115	3.0994	11.1609
	Variance	0.0607	0.0135	0.0367	0.8312	0.2631	0.5433	3.2188	0.8878	2.3659
Ridge	Bias	-0.0258	-0.0533	-0.0484	-0.0346	-0.0238	0.0084	0.1819	-0.0288	0.0370
	Variance	0.0209	0.0038	0.0236	0.3198	0.0691	0.3346	1.2503	0.2507	1.1012
Lasso	Bias	-0.0684	-0.0801	-0.0771	-0.0633	-0.0556	-0.0719	-0.0331	-0.0371	-0.0169
	Variance	0.0086	0.0022	0.0113	0.0587	0.0165	0.0615	0.2124	0.0508	0.1991
PR	Bias	0.0219	0.0339	0.1177	0.0829	0.1788	0.4473	0.2905	0.3091	0.8301
	Variance	0.0202	0.0037	0.0421	0.2727	0.0804	0.6656	1.2150	0.2799	2.0977
RP	Bias	-0.0366	-0.0427	-0.0347	-0.0392	-0.0296	-0.0324	-0.0317	-0.0227	-0.0169
	Variance	0.0021	0.0012	0.0022	0.0116	0.0018	0.0046	0.0311	0.0023	0.0107
PRP	Bias	0.0133	0.0259	0.0730	0.0506	0.1508	0.3044	0.1778	0.2656	0.4685
	Variance	0.0194	0.0038	0.0577	0.3122	0.0773	0.8624	1.2866	0.2676	2.6358

TABLE 4.A2: Bias and Variances for experiments under Design 1 with $SNR = 5$

		s=0.1			s=0.4			s=0.8		
	τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
	ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	Bias	0.3481	0.3850	1.3203	1.6509	1.6018	5.4856	3.4006	3.1291	11.1712
	Variance	0.0299	0.0113	0.0131	0.5674	0.1325	0.1977	2.4100	0.4624	1.0272
Ridge	Bias	-0.0436	-0.0419	-0.0507	-0.0295	-0.0246	-0.0973	0.1155	-0.0262	0.0266
	Variance	0.0044	0.0012	0.0048	0.0675	0.0137	0.0827	0.2534	0.0422	0.3283
Lasso	Bias	-0.0638	-0.0458	-0.0713	-0.0698	-0.0302	-0.0552	-0.0226	-0.0425	-0.0155
	Variance	0.0033	0.0031	0.0031	0.0129	0.0134	0.0295	0.0895	0.0200	0.0988
PR	Bias	0.0043	0.0356	0.0589	0.0759	0.1187	0.2002	0.2272	0.1752	0.5122
	Variance	0.0041	0.0024	0.0076	0.0489	0.0213	0.1229	0.2411	0.0624	0.6018
RP	Bias	-0.0365	-0.0401	-0.0383	-0.0359	-0.0208	-0.0350	-0.0135	-0.0132	-0.0142
	Variance	0.0015	0.0008	0.0013	0.0032	0.0009	0.0022	0.0075	0.0009	0.0044
PRP	Bias	-0.0043	0.0316	0.0541	0.0403	0.1020	0.1450	0.1811	0.1563	0.4602
	Variance	0.0040	0.00233	0.0086	0.0624	0.0202	0.1376	0.2619	0.0678	0.6542

TABLE 4.A3: Bias and Variances for experiments under Design 1 with $SNR = 10$

		s=0.1			s=0.4			s=0.8		
	τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
	ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	Bias	0.3527	0.3795	1.3335	1.5511	1.5636	5.4863	3.1044	3.1296	11.2056
	Variance	0.0298	0.0091	0.0088	0.4690	0.1281	0.2064	2.1521	0.4436	0.7435
Ridge	Bias	-0.0374	-0.0538	-0.0583	-0.0429	-0.0339	-0.0271	0.0768	0.0045	-0.0298
	Variance	0.0024	0.0006	0.0025	0.0309	0.0073	0.0381	0.1118	0.0239	0.1864
Lasso	Bias	-0.0583	-0.0441	-0.0679	-0.0622	-0.0419	-0.0465	-0.0267	-0.0232	0.0404
	Variance	0.0021	0.0016	0.0041	0.0108	0.0084	0.0178	0.0394	0.0249	0.1096
PR	Bias	0.0083	0.0175	0.0426	0.0426	0.0900	0.2012	0.1689	0.1651	0.3924
	Variance	0.0016	0.0007	0.0043	0.0201	0.0153	0.0770	0.0963	0.0491	0.3218
RP	Bias	-0.0323	-0.0476	-0.0442	-0.0389	-0.0259	-0.0232	-0.0188	-0.0067	-0.0126
	Variance	0.0009	0.0005	0.0013	0.0023	0.0007	0.0012	0.0036	0.0005	0.0025
PRP	Bias	0.0013	0.0148	0.0371	0.0137	0.0784	0.1746	0.1168	0.1489	0.3346
	Variance	0.0021	0.0009	0.0044	0.0268	0.0145	0.0710	0.1240	0.0499	0.2770

TABLE 4.A4: Bias and Variances for experiments under Design 2 with $SNR = 1$

		s=0.1			s=0.4			s=0.8		
	τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
	ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	Bias	0.1441	0.1431	0.7209	0.4070	0.4174	2.9516	1.1487	2.1756	6.7179
	Variance	0.1589	0.4684	0.1385	10.1298	32.8666	8.0858	108.7230	319.6219	76.4391
Ridge	Bias	0.5412	1.0919	1.1034	2.1617	4.0632	4.6002	4.7326	8.8083	10.7250
	Variance	0.1080	0.1600	0.1442	7.6244	9.3148	6.2784	60.6841	68.4833	65.4742
Lasso	Bias	0.4270	1.0678	1.0830	3.5630	6.6230	7.0420	8.6673	13.4363	14.8678
	Variance	0.1734	0.4454	0.4588	13.6203	8.7748	5.0201	71.3513	55.3950	23.3543
PR	Bias	-0.0200	-0.0084	0.1144	-0.1804	-0.4838	0.5589	-0.1796	0.3305	3.4671
	Variance	0.0985	0.3119	0.2199	6.2786	20.3964	8.5726	66.1429	166.3761	123.9081
RP	Bias	0.5507	1.0374	1.0152	2.0889	3.5550	4.0013	5.1625	7.4507	8.8161
	Variance	0.1111	0.1332	0.1696	4.6853	6.6647	5.1219	24.9670	34.1945	32.6550
PRP	Bias	-0.0153	-0.0291	0.0571	-0.2498	-0.5610	0.3536	-0.4267	0.2585	2.3990
	Variance	0.0944	0.2790	0.2495	6.0733	19.4394	10.2534	65.7425	162.5291	139.7745

TABLE 4.A5: Bias and Variances for experiments under Design 2 with $SNR = 5$

		s=0.1			s=0.4			s=0.8		
	τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
	ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	Bias	0.1794	0.1651	0.7088	0.7711	1.0198	3.0064	1.7800	1.7117	5.7767
	Variance	0.0931	0.1741	0.0983	6.0988	18.1071	6.2412	57.7810	119.8009	40.8483
Ridge	Bias	0.5911	1.0268	1.1207	2.2237	4.1681	4.2811	4.2030	8.5476	9.0800
	Variance	0.0456	0.0552	0.0429	2.8638	3.3044	2.5666	24.8125	20.4160	29.1863
Lasso	Bias	0.1491	0.4060	0.4880	1.7914	4.6497	4.6519	6.3216	12.7750	11.9264
	Variance	0.0544	0.1075	0.1486	6.1908	11.4654	11.7031	61.0500	32.5727	53.2517
PR	Bias	-0.0025	-0.0121	0.0711	-0.1202	0.0121	0.2134	-0.4164	0.4653	1.3986
	Variance	0.0537	0.1020	0.0703	3.2164	8.7425	6.3114	30.2332	41.5446	64.2538
RP	Bias	0.5774	1.0153	1.0995	2.0685	4.0750	4.1247	4.4710	8.2065	8.4784
	Variance	0.0487	0.0705	0.0550	2.7313	3.2869	2.8802	16.3916	15.8197	22.5695
PRP	Bias	0.0011	-0.0390	0.0527	-0.0522	-0.0706	0.0123	-0.5121	0.3357	1.0967
	Variance	0.0473	0.0923	0.0896	2.9829	8.0630	7.5721	29.9618	36.1075	79.5951

TABLE 4.A6: Bias and Variances for experiments under Design 2 with $SNR = 10$

		s=0.1			s=0.4			s=0.8		
	τ	0.2	0.2	0.7	0.2	0.2	0.7	0.2	0.2	0.7
	ρ	0.7	0.2	0.7	0.7	0.2	0.7	0.7	0.2	0.7
MOLS	Bias	0.2137	0.1901	0.6911	0.3507	1.1284	2.6952	1.4197	1.7101	4.8844
	Variance	0.0769	0.2440	0.0672	5.6327	11.8075	5.4425	35.3208	103.4097	40.1926
Ridge	Bias	0.5824	1.0818	1.1337	2.1756	4.2352	4.2599	4.2820	8.8609	8.3133
	Variance	0.0293	0.0481	0.0448	1.5353	2.1443	2.2950	13.2754	24.1284	18.7222
Lasso	Bias	0.1293	0.3270	0.3676	1.3575	4.4432	4.0478	5.6837	12.1149	11.0866
	Variance	0.0275	0.0847	0.0816	4.3212	9.9120	12.0787	42.1986	33.9890	47.0100
PR	Bias	0.0185	0.0349	0.0756	-0.2017	0.2371	-0.0480	-0.2787	0.6638	-0.2510
	Variance	0.0270	0.1180	0.0869	1.9214	4.2291	5.4027	15.3229	53.4108	44.6474
RP	Bias	0.5305	1.1159	1.0969	2.0905	4.2640	4.1114	3.9800	7.9531	7.8386
	Variance	0.0464	0.0628	0.0603	1.4799	2.4961	2.7343	12.8358	18.7957	17.4772
PRP	Bias	0.0109	0.0174	0.0371	-0.1907	0.1357	-0.2322	-0.2197	0.3279	-0.6318
	Variance	0.0263	0.1075	0.0953	1.5855	4.0667	6.2026	13.2564	53.6470	47.8060

Chapter 5

Conclusion

This thesis has sought to contribute to the literature on high dimensional linear regression models in two different ways. One was by providing an overview of the methods available for economic forecasters, the other was by proposing modifications to existing statistical models for the improvement of practicality and reliability. Throughout this work, consideration has not only been given to the types of data that economists face, but also the purpose for which they require a statistical procedure. The first two essays were concerned with the case where one has to compute forecasts while the final one focused on where the estimation of individual slope parameters are of interest.

The first essay provided a base level of knowledge and first stage of direction in determining the most appropriate forecasting model when faced with a given high dimensional economic data set. This was executed through carefully designed simulation experiments that consider the influence of predictor correlation and temporal dependency separately. More specifically, the experiments mimic a setting where a forecaster is faced with a large set of predictors and wishing to compute one-step-ahead forecasts of the dependent variable value for the following period using a linear regression framework.

The first class of experiments focused on the correlation aspect of predictors in econometric models with Principal Components being the most successful across the experiments with a large p and ρ. With Ridge performing very similarly, this was expected given how the 2 models are shown to thrive when covariates co-move notably. While many attributes of this study can be argued as being realistic with respect to an empirical application, there are some aspects which can be questioned or, at the very least, provided a limited vision of how the methods truly perform. Firstly, the SNR is fixed throughout all experiments and it was seen in the later essays how

this influences the performance of models based on the estimator bias-variance trade off.

Another area that can be argued to be unrealistic is the assumption of sparsity. For this class of experiments it was assumed that only 5% of the covariates were active with the rest having no direct influence on y. Such an assumption has been debated throughout the field of economics with work such as Giannone et al. (2021) finding no theoretical evidence to support the assumption of sparsity. Future work could commit attention to covering a wider range of scenarios through diversifying the sparsity, SNR as well as allowing the true coefficients to vary in sign as magnitude, as in this essay they are either 0 or 1. In addition, it was assumed that the correlation between all predictors was equal and greater than 0. While this is relevant and facilitated easy interpretation of how correlation influences forecast reliability, it is not representative of a scenario likely to be faced with real data. As discussed in the second essay, methods such as Ridge can suffer when the sign of correlations are mixed and the concept of clusters of highly correlated predictors is another area worth attention in the context of high dimensional predictors sets in economic applications.

For the second design of experiments, the DGP is similar to that of the first, only with the degree of time series persistence varied instead of correlation. The results here are much more concrete with Random Projections dominating all other methods as soon as the level of persistence increases slightly. The limitations of this class of experiments are similar to that of the first with respect to the SNR, sign and magnitude of true coefficients and highly sparse DGP. While having an even split of persistent and purely stationary predictors was logical, future work could consider varying compositions of stationary and non-stationary covariates as well as uneven splits of the active predictors amongst these two sets (for example, 70% of active predictors stationary and the remaining 30% showing persistence).

The second essay concentrated, specifically, on the Ridge Regression approach of Hoerl and Kennard (1970) and proposed a modified approach to predictive regression with the aim of overcoming the bias issues that Ridge alone faces in certain situations. Using a Frisch-Waugh style approach, this essay proposed the Partial Ridge procedure for estimating a single coefficient applying Ridge penalisation but sparing the variable of interest. Through theoretical and simulation analysis it was demonstrated that Partial Ridge can achieve superior MSEs over Ridge for single covariates when β_i is large relative to $\sum_{j \neq i}^{p} \beta_j$. However, it is infeasible that this be the case for all of the true predictors, therefore, when considering the estimation of a full profile of slopes, Partial Ridge is inferior to Ridge.

As a result, a hybrid approach was proposed whereby one carries out an initial stage of screening to assess which true coefficients are likely to be largest. These select coefficients are then estimated with Partial Ridge and the rest with Full Ridge. Using a fixed design toy model setting with all predictors having equal correlation amongst them, the expected predictive risk of this method was compared to that of using Ridge alone. It was shown that, under specific DGPs and with a certain degree of accuracy when choosing how to estimate each parameter, the hybrid approach can provide a lower predictive risk than that from using Ridge alone.

These results were supported by an equicorrelation simulation setting that replicated a one-step-ahead forecasting scenario in similar fashion to that of the first essay. The study used Full Ridge as a way to screen and determine which coefficients to estimate with Partial Ridge and, therefore, can be deemed as realistic in comparison to the theoretical analysis which began with assumption that the analyst has some idea of which true coefficients are likely to be large. The results confirmed what was shown in the toy model in that when the true coefficients vary markedly in sign and magnitude then the hybrid approach outperforms Ridge alone. In addition, a higher SNR amplifies the gains in bias achieved by the use of Partial Ridge and, hence, provides strong justification for the use of this approach in economic applications with relatively little noise.

Finally, an empirical application concerning the prediction of residential apartment selling prices and construction costs was considered to complete the investigation of how appropriate this method can be for econometric modelling. For robustness, this was done by considering multiple splits of the original data set into the training and testing sample in addition to a careful but computationally reasonable strategy for determining penalty parameters. The newly proposed estimation procedure was highly competitive for the prediction of apartment selling prices providing further evidence to support its use. Although, one could argue that this was not a fair reflection of the suitability of HEP in practise due to how this study considered a range of forecasts using multiple choices of variable subsets to be estimated with Partial Ridge and reports the best in an MSFE sense. Therefore, this can be considered as a best case scenario but still reveals that the potential is there and inspires the need of further work concerning a formal procedure for choosing which parameters to estimate with Partial Ridge.

The final essay was motivated by the success of Random Projections in the first essay. However, the lack of work on RP as a way to estimate individual parameters meant there was a perfect opportunity to combine the Frisch-Waugh approach utilised in the second essay with the highly effective Johnson-Lindenstrauss Lemma to improve

estimation precision. This gave rise to the proposal of Partial Random Projections as an attempt to reduce individual estimation bias through avoiding distortion of the predictor of interest during the dimension reduction phase. Using a fixed design setting with predictor correlation restrictions imposed, the bias and variance features of PRP were studied and compared with that of Ridge and Marginal OLS. Similar to what was seen in the second essay, PRP has significant potential over its rivals when through its bias term. However, its variance can be seen to clearly suffer pointing towards a similar scenario seen in the previous essay whereby PRP will favour situations with a high SNR.

Generally, the lower the correlation of the variable of interest with all other covariates (τ), the more successful PRP was along with a low $\sum_{j\neq i}^{p} \beta_j$ amplifying its bias gains in similar fashion to that of Partial Ridge. This was demonstrated through a simulation study replicating causal inference studies but considering 2 separate profiles of the true coefficients. While PRP struggled to be competitive when the coefficients had little variation in sign and magnitude, it frequently provided the lowest MSE when the sign and size of the true coefficients was mixed and the correlation between the variable of interest τ was small relative to ρ.

Therefore, this essay provided the foundations of a new model for studying causal inference in high dimensional settings, similar to that of Galbraith and Zinde-Walsh (2020). While this essay showed that there are situations where PRP can be considered as the most appropriate method in terms of estimation accuracy, there are still a vast number of areas that require attention to support its reputation as a respectable model for causal inference. The main question one raises is concerned with the standard error of the PRP estimator. This was shown to depend on the value of true coefficients themselves and, therefore, poses significant challenges to one wishing to carry out hypothesis testing. Another area in need of further research is the choice of the subspace dimension, as while this is widely discussed for RP, this may be applied differently for PRP which suffers more through its variance and, therefore, possibly requires a smaller k than RP alone would.

This thesis has sought to contribute to the knowledge surrounding the use of high dimensional models with data from economics and finance. Following a broad overview of these methods in a generalised but detailed simulation study, the remaining essays attempt to adapt methods that showed promising signs in the first essay with the aim to improve their accuracy or provide an additional purpose for them with respect to applications in economics. The proposal of these new methodologies is ambitious but reflects the degree to which high dimensional linear regression models are still relatively unexplored in light of applications to economics.

Consequently, through the incorporation of existing literature in econometrics, statistics and machine learning, there is an extensive span of avenues worth investigating to assist econometricians with the analysis of high dimensional economic data.

Bibliography

Achlioptas, D. (2003) 'Database-friendly random projections: Johnson-Lindenstrauss with binary coins', *Journal of Computer and System Sciences*, 66(4), pp. 671-687.

Akaike, H. (1969) 'Fitting Autoregressions for Predictions', *Annals for the Institutes of Statistical Mathematics*, 21, pp. 243-247.

Anderlucci, L., Fortunato, F. and Montanari, A. (2010) 'High-Dimensional Clustering via Random Projections', *Journal of Classification*, 39(1), pp. 191-216.

Bach, F. (2022) '*Learning Theory from First Principles*', Available at: https://www.di.ens.fr/ fbach/ltfp_book.pdf (Accessed 4 October 2022).

Bacon, R. and Hausman, J. (1974) 'The Relationship Between Ridge Regression and the Minimum Mean Square Error Estimator of Chipman', *Oxford Bulletin of Economics and Statistics*, 36(2), pp. 115-124.

Banerjee, A. and Magnus, J. (1999) 'The Sensitivity of OLS when the Variance Matrix is (Partially) Unknown', *Journal of Econometrics*, 92(2), pp. 295-323.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012) 'Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain', *Econometrica*, 80(6), pp. 2369-2429.

Belloni, A. and Chernozhukov, V. (2013) 'Least Squares After Model Selection in High-Dimensional Sparse Models', *Bernoulli*, 19(2), pp. 521-547.

Bingham, E. and Mannila, H. (2001) 'Random Projection in Dimensionality Reduction: Applications to Image and Text Data', *Seventh ACM SIGKDD International Conference*

on Knowledge Discovery and Data Mining, San Francisco, California (USA), 26-29 August 2001, Association for Computing Machinery: New York (USA), pp. 245-250.

Böheim, R. and Stöllinger, P. (2020) 'Decomposition of the Gender Wage Gap Using the LASSO Estimator', *Applied Economics Letters*, 28(10), pp. 817-828.

Boivin, J. and Ng, S. (2005) 'Understanding and Comparing Factor-Based Forecasts', *International Journal of Central Banking*, 1(3), pp. 117-151.

Boot, T. and Nibbering, D. (2019) 'Forecasting Using Random Subspace Methods', *Journal of Econometrics*, 209(2), pp. 391-406.

Brown, W. and Beattie, B. (1975) 'Improving Estimates of Economic Parameters by Use of Ridge Regression with Production Function Applications', *American Journal of Agricultural Economics*, 57(1), pp. 21-32.

Brown, W. and Nawas, F. (1973) 'Impact of Aggregation on the Estimation of Outdoor Recreation Demand Functions', *American Journal of Agricultural Economics*, 55(2), pp. 246-249.

Campbell, J. and Mankiw, N. (1987) 'Are Output Fluctuations Transitory?', *Quarterly Journal of Economics*, 102, pp. 857-880.

Caporale, G. and Pittis, N. (2001) 'Persistence in Macroeconomic Time Series: is it a Model Invariant Property?', *Revista de Economia del Rosario*, 4(2), pp. 117-142.

Cheung, M.W-L., and Jak, S. (2016) 'Analyzing Big Data in Psychology: A Split/Analyze/Meta-Analyze Approach', *Frontiers in Psychology*, 7(738), DOI:https://doi.org/10.3389/fpsyg.2016.00738.

Chung, W. (2014) 'BizPro: Extracting and Categorizing business Intelligence Factors from Textual News Articles', *International Journal of Information Management*, 34(2), pp. 272-284.

Coniffe, D., Stone, J. and O'Neill, F. (1976) 'Application of Ridge Regression in Agricultural Economics', *Irish Journal of Agricultural Economics and Rural Sociology*, 6(1),

pp. 89-92.

Connor, G. and Korajczky, R. (1988) 'Risk and return in an equilibrium APT: Application of a new test methodology', *Journal of Financial Economics*, 21, pp. 255-289.

Connor, G. and Korajczky, R. (1993) 'A Test for the Number of Factors in an Approximate Factor Model', *The Journal of Finance*, 48(4), pp. 1263-1291.

Dasgupta, S. (2000) 'Experiments with Random Projection', *16th Conference on Uncertainty in Artificial Intelligence*, San Francisco (USA), 30 June - 3 July 2000, Morgan Kaufmann Publishers Inc., pp. 143-151.

Dasgupta, S. and Gupta, A. (2002) 'An elementary proof of a theorem of Johnson and Lindenstrauss', *Random Structures and Algorithms*, 22(1), pp.60-65.

Dash, S., Shakyawar, S.K., Sharma, M. and Kaushik, S. (2019) 'Big Data in Healthcare: Management, Analysis and Future Prospects', *Journal of Big Data*, 6(54), DOI; https://doi.org/10.1186/s40537-019-0217-0.

Deegalla, S. and Bostrom. H. (2006) 'Reducing High-Dimensional Data by Principal Component Analysis vs. Random Projection for Nearest Neighbor Classification', *5th International Conference on Machine Learning and Applications (ICMLA'06)*, Orlando (USA), 14-16 December 2006, IEEE, pp. 245-250.

Dhillon, P., Lu, Y., Foster, D. and Ungar, L. (2013) 'New Subsampling Algorithms for Fast Least Squares Regression', *26th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'13)*, Lake Tahoe, Nevada (USA), 5-10 December 2013, Curran Associates Inc: New York (USA), pp. 360-368.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) 'Least Angle Regression', *The Annals of Statistics*, 32(2), pp. 407-499.

Elliott, G., Gargano, A. and Timmermann, A. (2013) 'Complete Subset Regression', *Journal of Econometrics*, 177(2), pp. 357-373.

Fan, J. and Li, R. (2001) 'Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties', *Journal of American Statistical Association*, 96, pp. 1348-1360.

Fan, J., Lv, J. and Qi, L. (2011) 'Sparse High-Dimensional Models in Economics', *Annual Review of Economics*, 3, pp. 291-317.

Farrar, D. and Glauber, R. (1967) 'Multicollinearity in Regression Analysis: The Problem Revisited ', *The Review of Economics and Statistics*, 49(1), pp. 92-107.

Fern, X. and Brodley, C. (2003) 'Random projection for high dimensional data clustering: a cluster ensemble approach', *Twentieth International Conference on International Conference on Machine Learning*, Washington DC (USA), 21-24 August 2003, AAAI Press, pp. 186-193.

Ferson, W., Sarkissian, S. and Simin, T. (2003) 'Spurious Regressions in Financial Economics?', *The Journal of Finance*, 58(4), pp. 1393-1413.

Forni, M. and Lippi, M. (1997) *Aggregation and the Microfoundations of Dynamic Macroeconomics*, UK: Oxford University Press.

Forni, M. and Reichlin, L. (1998) 'Let's Get Real: A Factor Analytical Approach to Disaggregated Business Cycle Dynamics', *The Review of Economic Studies*, 65(3), pp. 453-473.

Foroni, C., Marcellino, M. and Schumacher, C. (2015) 'Unrestricted Mixed Data Sampling (MIDAS): MIDAS Regressions with Unrestricted Lag Polynomials', *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 178(1), pp. 57-82.

Fradkin, D. and Madigan, D. (2003) 'Experiments with random projections for machine learning', *Ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '03*, Washington (USA), 24-27 August 2003, Association for Computing Machinery: New York, United States, pp. 517-522.

Frank, I. and Friedman, J. (1993) 'A Statistical View of Some Chemometrics Regression Tools', *Technometrics*, 35(2), pp. 109-135.

Frisch, R. and Waugh, F. (1933) 'Partial Time Regressions as Compared with Individual Trends', *Econometricia*, 1(4), pp. 387-401.

Galbraith, J. and Zinde-Walsh, V. (2020) 'Simple and Reliable Estimators of Coefficients of Interest in a Model With High-Dimensional Confounding Effects', *Journal of Econometrics*, 218(2), pp. 609-632.

Geweke, J. (1977) 'The Dynamic Factor Analysis of Economic Time Series', in *Latent Variables in Socio-Economic Models*, eds. D. J. Aigner and A. S. Goldberger, Amsterdam: North-Holland.

Ghysels, E. and Marcellino, M. (2018) *Applied Economic Forecasting Using Time Series Methods*, New York: Oxford University Press.

Gonzalo, J. and Pitarakis, JY. (2020) 'Spurious Relationships in High Dimensional Time Series with Strong or Mild Persistence', *International Journal of Forecasting*, 37(4), pp. 1480-1497.

Ghysels, E., Santa-Clara, P. and Valkanov, R. (2004) 'The MIDAS Touch: Mixed Data Sampling Regression Models', CIRANO Working Papers 2004s-20, CIRANO.

Ghysels, E., Sinko, A. and Valkanov, R. (2007) 'MIDAS Regressions: Further Results and New Directions', *Econometric Reviews*, 26(1), pp. 53-90.

Giannone, D., Lenza, M. and Primiceri, G. (2021) 'Economic Predictions With Big Data: The Illusion of Sparsity', *Econometricia*, 89(5), pp. 2409-2437.

Granger, C. and Newbold, P. (1974) 'Spurious Regressions in Econometrics', *Journal of Econometrics*, 2(2), pp. 111-120.

Gunst, R., Webster, J. and Mason, R. (1976) 'A Comparison of Least Squares and Latent Root Regression Estimators', *Technometrics*, 18, pp. 75-83.

Hansen, P. and Lunde, L. (2014) 'Estimating the Persistence and the Autocorrelation Function of a Time Series that is Measured with Error', *Econometric Theory*, 30, pp. 60-93.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D., Scherf, U., Weinstein, J., Alizadeh, A., Staudt, L. and Botstein, D. (2000) 'Gene Shaving as a method for identifying distinct sets of genes with similar expression patterns', *Genome Biology*, 1(2), pp.1–21.

Hastie, T., Tibshirani, R. and Tibshirani, R.J. (2020) 'Best Subset, Forward Stepwise, or LASSO? Analysis and Recommendations Based on Extensive Comparisons', *Statistical Science*, 35(4), pp. 579-592.

Heinze, C., McWilliams, B., Meinshausen, N., and Krummenacher, G. (2014). 'LOCO: Distributing Ridge Regression with Random Projections', *arXiv: Machine Learning*.

Hoerl , A. and Kennard, R. (1970) 'Ridge Regression: Biased Estimation for Nonorthogonal Problems', *Technometrics*, 42(1), pp. 80-86.

Inoue, A. and Kilian, L. (2008) 'How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. Consumer Price Inflation', *Journal of the American Statistical Association*, 103(1), pp. 511-522.

Johnson, W. and Lindenstrauss, J. (1984) 'Extensions of Lipschitz mappings into a Hilbert space', *Contemporary mathematics*, 26(1), pp. 189-206.

Kaban, A. (2014) 'New Bounds on Compressive Linear Least Squares Regression', *Seventeenth International Conference on Artificial Intelligence and Statistics*, Reykjavik (Iceland), 22-25 April 2014, JMLR, pp. 448-456.

Kalina, J. (2018) 'Big Data, Biostatistics and Complexity Reduction', *European Journal for Biomedical Information*, 14(2), pp. 24-32.

Kaski, S. (1998) 'Dimensionality reduction by random mapping: fast similarity computation for clustering', *IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*,

Anchorage (USA), 4-9 May 1998, IEEE, pp. 413-418.

Kidwell, J. and Brown, L. (1982) 'Ridge Regression as a Technique for Analyzing Models with Multicollinearity', *Journal of Marriage and the Family*, 44(2), pp. 287-299.

Klikenberg, S., Straatemeier, M. and Van der Mass, H.L.J. (2011) 'Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation', *Computers and Education*, 57(2), pp. 1813-1824.

Knight, K. and Fu, W.J. (2000) 'Asymptotics for Lasso-type Estimators', *Annals of Statistics*, 28(5), pp. 1356-1378.

Kolleww, J. (2019) 'NHS data is worth billions – but who should have access to it?', *The Guardian UK Edition*, 10 June, Available at: https://www.theguardian.com/society/2019/jun/10/nhs-data-google-alphabet-tech-drug-firms#:~:text=The%20NHS%20database%20holds%20the,the%20attention%20of%20private%20businesses. (Accessed 15 November 2023).

Kruse, R., Ventosa-Santaulària, D. and Noriega, A. (2017) 'Changes in Persistence, Spurious Regressions and the Fisher Hypothesis', *Studies in Nonlinear Dynamics and Econometrics*, 21(3), pp. 1-28.

Laney, D. (2001) '3D Data Management: Controlling Data Volume, Velocity, and Variety', Technical Report, META Group.

Lehmann, B. and Modest, D. (1988) 'The Empirical Foundations of the Arbitrage Pricing Theory', *Journal of Financial Economics*, 21, pp. 213-254.

Li, P., Hastie, T. and Church, K. (2006) 'Very sparse random projections', *12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '06)*, Philadelphia (USA), 20-23 August 2006, Association for Computing Machinery, New York (USA), pp. 287–296.

Maillard, OA. and Munos, R. (2009) 'Compressed Least-Squares regression', *22nd International Conference on Neural Information Processing Systems (NIPS'09)*, Vancouver (Canada), 7-10 December 2009, New York: Curran Associates Inc., pp. 1213-1221.

Mallows, C. (1973) 'Some Commnets on Cp', *Technometrics*, 15(1), pp. 661-675.

Mammen, E. (1993) 'Bootstrap and Wild Bootstrap for High Dimensional Linear Models', *The Annals of Statistics*, 21(1), pp. 255-285.

Marzetta, T., Tucci, G. and Simon, S. (2011) 'A Random Matrix Theoretic Approach to Handling Singular Covariance Estimates', *IEEE Trans. Information Theory*.

Mason, R. and Brown, W. (1975) 'Multicollinearity Problems and Ridge Regression in Sociological Models', *Social Science Research*, 4(2), pp. 135-149.

Matoušek, J. (2008) 'On variants of the Johnson–Lindenstrauss lemma', *Random Structures and Algorithms*, 33(2), pp. 142-156.

Nabil, M. (2017) 'Random Projection and Its Applications', Number: arXiv:1710.03163.

Newhouse, J. and Oman, S. (1971) 'An Evaluation of Ridge Estimators'.

Ng, S. (2013) 'Chapter 14 – Variable Selection in Predictive Regressions', *Handbook of Economic Forecasting*, 2B, pp. 752-789.

Ng, S. (2016) 'Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data', Available at: DOI:http://dx.doi.org/10.2139/ssrn.2864302.

O'Neill, F. and Buttimer, D. (1972) 'The Domestic Demand for Irish Beef', *Irish Journal of Agricultural Economics and Rural Sociology* 4(1), pp. 35-54.

Peplow, M. (2016) 'The 100,000 Genomes Project', *BMJ*, 353, DOI: https://doi.org/10.1136/bmj.i1757

Rao, C. (1964) 'The use and interpretation of principal components in applied research', *Sankhya*, 26, pp. 329-358.

Ross, S. (1976) 'The Arbitrage Theory of Capital Asset Pricing', *Journal of Economic Theory*, 13, pp. 341-360.

Rosset, S., Zhu, J, and Hastie, T. (2004) 'Boosting as a regularized path to a maximum margin classifier', *Journal of Machine Learning Research*, 5, pp. 941-973.

Sargent, T. and Sims, C. (1977) 'Business Cycle Modelling Without Pretending to Have Too Much A-Priori Economic Theory', in *New Methods in Business Cycle Research*, ed. by C. Sims et al., Minneapolis: Federal Reserve Bank of Minneapolis.

Schwarz, G. (1978) 'Estimating the Dimension of a Model', *The Annals of Statistics*, 6(2), pp. 461-464.

Severn, K. (2023) 'Applied Multivariate Statistics', *MATH3030: Multivariate Statistics* Available at: https://rich-d-wilkinson.github.io/MATH3030/ (Accessed 7 August 2023).

Shao, J. and Deng, X. (2012) 'Estimation in High-Dimensional Linear Models with Deterministic Design Matrices', *The Annals of Statistics*, 40(2), pp. 812-831.

Shapiro, J.M. (2017) 'Is Big Data a Big Deal for Applied Microeconomics?', In *Advances in Economics and Econometrics: Eleventh World Congress of the Economic Society*, 2, pp. 35-52.

Sharpe, W. (1964) 'Capital asset prices: a theory of market equilibrium under conditions of risk', *Journal of Finance*, 19, pp. 425-442.

Shi, X., Liang, B. and Zhang, Q. (2020) 'Post-Selection Inference of Generalized Linear Models Based on the Lasso and the Elastic Net', *Communications in Statistics – Theory and Methods*, Ahead-of-print 1-18.
Doi: https://doi.org/10.1080/03610926.2020.1821892.

Simeon, A. and Olaiya, A. (2021) 'Multicollinearity Regularization Using Lasso and Ridge Regression on Economic Data', *Kasu Journal of Mathematical Sciences*, 2(2), pp. 43-54.

Sims, C. (1980) 'Macroeconomics and Reality', *Econometrica*, 48(1), pp. 1-48.

Slawski, M. (2018) 'On principal components regression, random projections, and column subsampling', *Electronic Journal of Statistics*, 12(2), pp. 3673-3712.

Smith, G. and Campbell, F. (1980) 'A Critique of Some Ridge Regression Methods', *Journal of the American Statistical Association*, 75(369), pp, 74-81.

Stock, J. and Watson, M. (2002a) 'Forecasting Using Principal Components from a Large Number of Predictors', *Journal of the American Statistical Association*, 97(468), pp. 1167-1179.

Stock, J. and Watson, M. (2002b) 'Macroeconomic Forecasting Using Diffusion Indexes', *Journal of Business and Economic Statistics*, 20(2), pp. 147-162.

Stock, J. and Watson, M. (2011) 'Dynamic Factor Models', in *Oxford Handbook of Economic Forecasting*, eds. Michael P. Clements and David F. Hendry,Oxford: Oxford University Press.

Schwarz, G. (1978) 'Estimating the Dimension of a Model', *The Annals of Statistics*, 6(2), pp. 461-464.

Taboga, M. (2017) *Ridge Regression*, Available at: https://www.statlect.com/fundamentals-of-statistics/ridge-regression (Accessed 23rd September 2021).

Taboga, M. (2021) *Multicollinearity*, Available at: https://www.statlect.com/fundamentals-of-statistics/multicollinearity (Accessed 5th January 2022).

Thanei, G.A., Heinze, C. and Meinshausen, N. (2017) 'Random Projections for Large Scale Regression', In *Big and Complex Data Analysis*, 51-68. Springer.

The Stan Development Team, (2011), 'Stan User's Guide', *25 Efficiency Tuning*. Available at: https://mc-stan.org/docs/stan-users-guide/

standardizing-predictors-and-outputs.html (Accessed: 1 December 2022).

Theobald, C. M. (1974) 'Generalizations of Mean Squared Error Applied to Ridge Regression', *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1), pp. 103-106.

Tibshirani, R. (1996) 'Regression Shrinkage and Selection via the Lasso', *Journal of Royal Statistical Society Series B*, 58(1), pp. 267-288.

Timmermann, A. (2006) 'Chapter 4 Forecast Combinations', *Handbook of Economic Forecasting*, 1, pp. 135-196.

Wold, H. and Lyttkens, E. (1969) 'Nonlinear Iterative Partial Least Squares (NIPALS) Estimation Procedures', *Bulletin of the International Statistical Institute*, 43, pp. 29-51.

Yan, D., Wang, J., Wang, H., Wang, H. and Li, Z. (2018) 'K-nearest Neighbor Search by Random Projection Forests', *IEEE International Conference on Big Data (Big Data)*, Seattle (USA), 10-13 December, IEEE, pp. 4775-4781.

Yousuf, K. and Ng, S. (2021) 'Boosting high dimensional predictive regressions with time varying parameters', *Journal of Econometrics*, 224(1), pp. 60-87.

Zhang, CH. (2010) 'Nearly Unbiased Variable Selection Under Minimax Concave Penalty', *Annals of Statistics*, 38(2), pp. 894-942.

Zhang, Y. and Politis, N. (2022) 'Debiased and Thresholded Ridge Regression for Linear Models with Heteroskedastic and Correlated Errors', *Journal of Royal Statistical Society Series B: Statistical Methodology*, pp. 1-29.

Zhao, P. and Yu, B. (2006) 'On Model Selection Consistency of Lasso', *Journal of Machine Learning Research*, 7, pp. 2541-2563.

Zou, H. (2006) 'The Adaptive Lasso and its Oracle Properties', *Journal of American Statistical Association*, 101(476), pp. 1418-1429.

Zou, H. and Hastie, T. (2005) 'Regularization and Variable Selection via the Elastic Net', *Journal of the Royal Statistical Society Series B*, 67(2), pp. 301-320.