Confidence intervals in high-dimensional regression based on regularized pseudoinverses

Tom Boot*  
University of Groningen

Didier Nibbering†  
Erasmus University Rotterdam
Tinbergen Institute

January 31, 2017

Abstract

In modern macroeconomic data sets, the number of available variables can greatly exceed the number of observations. In this paper we show how valid confidence intervals can be constructed for all coefficients by approximating the inverse covariance matrix by a scaled Moore-Penrose pseudoinverse, and then applying the lasso to perform a bias correction. In addition, we propose random least squares, a new regularization technique which yields narrower confidence intervals with the same theoretical validity. Random least squares estimates the inverse covariance matrix using multiple low-dimensional projections of the data. This is shown to be equivalent to a generalized form of ridge regularization. The methods are illustrated in Monte Carlo experiments and an empirical example using the quarterly data from the FRED-QD database, where gross domestic product is explained by a large number of macroeconomic and financial indicators.

Keywords: high-dimensional regression, confidence intervals, random projection, Moore-Penrose pseudoinverse

JEL codes: C12, C13, O40

*Department of Economics, Econometrics and Finance, University of Groningen, Nettelbosje 2, 9747 AE Groningen, The Netherlands, e-mail: t.boot@rug.nl
†Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands, e-mail: nibbering@ese.eur.nl
1 Introduction

The increase in available macroeconomic indicators has led to a statistically challenging situation where the number of explanatory variables often exceeds the number of available observations. This is commonly observed in cross-sectional datasets on economic growth such as Barro and Lee (1993); Sala-i Martin (1997); Fernandez et al. (2001), but also in macroeconomic time series data with a low measurement frequency as in Stock and Watson (2002) and McCracken and Ng (2016). Standard inference techniques are hampered by the resulting rank deficiency of the empirical covariance matrix. Hence, new methods are needed to handle this type of ‘wide’ data.

This paper deals with the estimation of coefficients and the construction of the corresponding confidence intervals in high-dimensional linear regression models, where the number of unknown coefficients are allowed to increase almost exponentially with the number of observations. There are two key ingredients to our approach: (1) We approximate the inverse covariance matrix by a diagonally scaled Moore-Penrose pseudo-inverse or a regularized variant, and (2) a bias correction step is implemented based on the lasso.

After establishing the validity for the method when using the Moore-Penrose inverse, we show how regularization of this inverse can reduce the size of the confidence intervals. We use a novel approach to regularization, where the regressor matrix is projected multiple times onto a low-dimensional subspace by post-multiplying with a random matrix with independent standard normal entries. The estimates of the inverse covariance matrix in the low-dimensional subspaces are aggregated to yield an estimate of the inverse covariance matrix. We show that this approach is equivalent to a specific type of generalized ridge regularization and yields valid confidence intervals when the projection dimension is chosen sufficiently close to the sample size.

An alternative regularization strategy is offered the commonly used ridge estimator, which adds a constant diagonal matrix to the covariance matrix before inverting. With an appropriately chosen penalty parameter, this again leads to valid confidence intervals.

There are several estimators available which can deal with high-dimensional wide data sets. Among the most well-known methods are the lasso estimator (Tibshirani, 1996), the adaptive lasso estimator (Zou, 2006), the Dantzig selector (Candes and Tao, 2007), and penalized likelihood methods (Fan et al., 2004). A comprehensive overview of theoretical results is provided by Bühlmann and Van De Geer (2011). The construction of standard errors around the resulting estimates remains challenging.

There are broadly two different approaches in the literature to constructing confidence intervals for a high-dimensional linear regression model. The
first one is based on inference regarding a low-dimensional set of parameters following model selection or regularization among the high-dimensional set of all available regressors (Belloni and Chernozhukov, 2009; Belloni et al., 2011; Chernozhukov et al., 2015). The two-stage procedure consists of, for instance, a lasso type estimator as first stage, and ordinary least squares estimation on the maintained variables in the second stage. These post-model selection estimators only provide confidence intervals for coefficients corresponding to the regressors which are included in the low-dimensional model, but do not yield an uncertainty measure for the majority of the coefficients.

The second approach, related to the approach we take in this paper, provides confidence intervals for all estimated coefficients. These estimators are based on various ways to construct an approximate inverse for the empirical covariance matrix and an ex-post bias correction. Javanmard and Montanari (2014) propose an optimization procedure that yields a sufficiently accurate approximate covariance inverse such that the resulting estimator is approximately normal and unbiased. Under slightly stricter assumptions Van de Geer et al. (2014) propose an estimator with the same theoretical validity based on the group-wise lasso, which parallels Zhang and Zhang (2014). Both estimators are computationally quite intensive, especially in the large-scale applications that they are intended for.

We contribute to the literature in a number of ways. First, we show that approximating the inverse of the empirical covariance matrix with a diagonally scaled Moore-Penrose inverse results in an estimator that is approximately normally distributed. The corresponding confidence intervals of this estimator are valid under standard conditions on the regressor matrix and a sparsity assumption, which constrains the number of non-zero coefficients. We explicitly allow for non-gaussian regression errors.

The Moore-Penrose estimator was advocated in Wang and Leng (2015) to set up a variable screening technique. We extend their results by introducing a diagonal scaling matrix, which is essential to show that the bias from using an approximate inverse covariance matrix vanishes. Approximately unbiased estimates are directly obtained when the non-zero coefficients are all local-to-zero. When some coefficients are large, the same results are obtained when the lasso is used to perform a bias-correction, as proposed in Zhang and Zhang (2014); Van de Geer et al. (2014); Javanmard and Montanari (2014). The noise level of the estimator is shown to be of the familiar $O(n^{-1/2})$.

Second, we show that random least squares can improve upon the scaled Moore-Penrose estimator in terms of statistical power. We extend the results of Marzetta et al. (2011) and show that for a subspace dimension close to the number of observations, random least squares approximates the scaled Moore-Penrose estimator and yields valid confidence intervals. Under a suit-
able choice of the random subspace dimension, the width of the confidence intervals of random least squares is at most as large as the width of the confidence intervals under the Moore-Penrose inverse.

Third, we show that the results remain valid for a ridge adjusted inverse instead of the Moore-Penrose inverse. A ridge based inverse covariance matrix was previously considered by Bühlmann et al. (2013) to construct conservative p-values, and by Wang and Leng (2015) to enhance variable screening efficiency.

The theoretical results are confirmed in a set of Monte Carlo experiments, in which we vary the specification of the covariance matrix, the amount of sparsity, and the signal strength. In line with the theoretical results, we find that even though the number of regressors is twice the number of observations, the coverage rates are close to the nominal rate of 95% for all settings under consideration. In general, the estimated values for the coefficients that are zero are very close to zero. The estimated values for the nonzero coefficients are slightly downward biased for random least squares and ridge regression, but more efficient in terms of standard errors and power, relative to the generalized inverse. The experiments confirm that for small values of the coefficients, a lasso bias correction is not needed.

We apply the methods to the FRED-QD, a quarterly dataset consisting of 254 macroeconomic and financial series of the United States economy, available from the second quarter of 1987. We analyze the relation between the real gross domestic product and the other variables provided in this dataset in a linear regression framework. Although the number of regressors largely exceeds the number of observations, our methods have enough power to distinguish significant effects from which the largest relate to the productivity and the number of hours worked in the business sector.

The outline of this paper is as follows. Section 2 sets up the general estimation approach and introduces our methods. The theoretical properties of the generalized inverse, ridge regression, and random least squares are derived and presented in Section 3. Section 4 illustrates these results through Monte Carlo simulations and Section 5 applies the methods on the FRED-QD dataset. Section 6 concludes.

2 Methods

This section outlines how to construct confidence intervals in a high-dimensional linear regression model, where the number of variable greatly exceeds the number of observations. For ease of exposition, all theoretical results are deferred to Section 3. We discuss three different methods for constructing
confidence intervals. All methods are based on an approximate inverse for the singular covariance matrix in combination with a bias correction based on a suitable initial estimator. For the approximate inverse we consider (a) the Moore-Penrose pseudo inverse, (b) a novel method called random least squares, and (c) a ridge adjusted inverse. We take the lasso estimator as the initial estimator used for bias correction.

2.1 General Framework

Consider the data generating process

\[ y = X\beta + \epsilon, \quad \epsilon \sim \text{i.i.d.}(0, \sigma^2 I), \quad (1) \]

where \( y \) is an \( n \times 1 \) response vector, \( X \) an \( n \times p \) regressor matrix, and \( \beta = (\beta_1, \ldots, \beta_p)' \) a \( p \times 1 \) vector of unknown regressor coefficients. The empirical covariance matrix is denoted by \( \hat{\Sigma} = \frac{1}{n}X'X \).

Define \( M \) as a \( p \times n \) matrix and consider estimators for \( \beta \) of the form

\[
\hat{\beta} = \frac{1}{n}My = \frac{1}{n}MX\beta + \frac{1}{n}M\epsilon \\
= \beta + \left( \frac{1}{n}MX - I \right)\beta + \frac{1}{n}M\epsilon.
\]

The second term represents a bias, which arises if \( \frac{1}{n}M \) is not an exact inverse for \( X \).

When \( p < n \), ordinary least squares yields unbiased estimates by choosing \( M = \hat{\Sigma}^{-1}X' \). When \( p > n \) it is no longer possible to choose \( M \) such that the bias is identically zero. However, suppose we have an initial estimator \( \hat{\beta}^{\text{init}} \), then we can adjust the bias in (2) as

\[
\hat{\beta}^c = \frac{1}{n}My - \left( \frac{1}{n}MX - I \right)\hat{\beta}^{\text{init}} \\
= \beta + \left( \frac{1}{n}MX - I \right)\left( \beta - \hat{\beta}^{\text{init}} \right) + \frac{1}{n}M\epsilon.
\]

The initial estimator is taken to be the lasso estimator proposed by Tibshirani (1996), with a cross-validated penalty term. The lasso estimator is very accurate, but its distribution is not tractable and hence there is no simple procedure to construct confidence intervals.

The purpose of this paper is to introduce several choices for \( M \) that, in combination with an accurate initial estimator \( \hat{\beta}^{\text{init}} \), yield a bias that
vanishes compared to the noise. In addition, we show that for a broad class of distributions of the error term, \( \frac{1}{\sqrt{n}} M \varepsilon \sim N(0, \sigma^2 MM') \). Then, if a consistent estimator of the noise level \( \sigma^2 \) is available, \( (1 - \alpha) \cdot 100\% \) confidence intervals can be readily constructed as

\[
\left[ \hat{\beta}_i^c - z_{\alpha/2} \sqrt{\hat{\sigma}^2 m'_i m_i / n}, \hat{\beta}_i^c + z_{\alpha/2} \sqrt{\hat{\sigma}^2 m'_i m_i / n} \right]
\]

(4)

where \( z_{\alpha/2} \) is the \( \alpha/2 \) critical value for the standard normal distribution, \( \hat{\sigma}^2 \) is a consistent estimator of \( \sigma^2 \), and \( m_i \) the \( i \)-th row of \( M \).

A key element of the proposed methods is that the diagonal elements of \( \frac{1}{n} MX - I \) can be set identical to zero by introducing a diagonal matrix \( D \) and taking

\[
\frac{1}{n} M = \frac{1}{n} D \hat{M}, \quad D_{ii} = n[\hat{M}X]^{-1}_{ii}
\]

(5)

We now discuss three ways of constructing \( M \) that are all of the form defined in (5), with \( \hat{M} \) chosen as: (a) the Moore-Penrose pseudo inverse, (b) random least squares: a novel technique to construct an approximate inverse based on low-dimensional projections of the regressor matrix \( X \), or (c) the Moore-Penrose pseudo inverse adjusted by a ridge penalty term. The first method has the benefit of being completely tuning free. The last two methods can both be seen as a penalized form of the Moore-Penrose inverse and require a choice of the penalization strength.

### 2.1.1 Moore-Penrose pseudo inverse

The most simple choice for \( \hat{M} \) is the Moore-Penrose pseudo inverse. When \( p < n \), and the columns of \( X \) are independent, then \( \hat{M} = (X'X)^{-1}X' \). In the high-dimensional setting where \( p > n \), the matrix \( X \) has linearly dependent columns by default. By assumption, the rows of \( X \) are linearly independent, in which case the pseudo inverse equals \( X'(XX')^{-1} \). The estimator we consider is therefore

\[
\frac{1}{n} M = \frac{1}{n} D^{MP} X'(XX')^{-1}
\]

(6)

where the elements of the diagonal scaling matrix \( D^{MP} \) equal

\[
D_{ii}^{MP} = \left( \frac{1}{n} X'(XX')^{-1} X \right)_{ii}^{-1}
\]

(7)
2.1.2 Random Least Squares

The Moore-Penrose pseudoinverse is a simple, yet noisy estimator. Therefore we consider more efficient ways to construct an approximate inverse covariance matrix. As a novel alternative to the pseudo inverse estimator, we propose the random least squares (RLS) estimator. This method is based on projecting the high-dimensional regressor matrix $X$ onto a $k < p$ dimensional subspace by post-multiplying with a $p \times k$ matrix $R$ with independently standard normally distributed elements,

$$[R_i] \sim \mathcal{N}(0, 1), \quad 1 \leq m \leq p, \quad 1 \leq n \leq k. \quad (8)$$

The idea behind the low-dimensional random projection is the following. Instead of considering the high-dimensional model in which all $p$ regressors are included, we consider the low-dimensional model

$$y = XR\gamma + u \quad (9)$$

Least squares estimation of $\gamma_R$ yields

$$\hat{\gamma}_R = (R'X'XR)^{-1}R'X'y, \quad (10)$$

which is related to the estimator of $\beta$ based on a single realization of $R$ by $\beta_R = R\hat{\gamma}_R$. Since $R$ is random, relying on a single realization is suboptimal. Therefore we average over different realizations of $R$ to arrive at an estimator of $\beta$,

$$\hat{\beta} = \mathbb{E}_R[R\hat{\gamma}_R] = \mathbb{E}_R[R(R'X'XR)^{-1}R']X'y = \frac{1}{n}My. \quad (11)$$

In the context of prediction, where one is interested in approximating $X\beta$, the accuracy of this estimator was investigated in Maillard and Munos (2009) and Kabán (2014).

From (11) it can be seen that random least squares uses a random construction method to obtain a regularized inverse covariance matrix $M$ as

$$\frac{1}{n}M = \frac{1}{n}D^{RLS}\mathbb{E}_R[R(R'X'XR)^{-1}R']X', \quad (12)$$

with

$$D^{RLS}_{ii} = \left(\frac{1}{n}\mathbb{E}_R[R(R'X'XR)^{-1}R']X'X\right)^{-1} \quad (13)$$

We will show in Section 3.3.2 that when the projection dimension $k = n$, then the expression in (12) reduces to the Moore-Penrose pseudoinverse.
2.1.3 Ridge regression

If the Moore-Penrose pseudoinverse is noisy, a well-known regularization method is to implement a ridge adjustment in which case

\[
\frac{1}{n} M = \frac{1}{n} D_{RI} \left( X'X + \gamma I \right)^{-1} X',
\]

(14)

where \( \gamma \) denotes the ridge penalty and the elements of the diagonal scaling matrix \( D_{RI} \) equal

\[
D_{RI}^{ii} = \left( \frac{1}{n} (X'X + \gamma I)^{-1} X'X \right)_{ii}^{-1}
\]

(15)

As with random least squares, the regularization in (14) can be related to the Moore-Penrose pseudo inverse, since we can write

\[
X'(XX')^{-1} = \lim_{\gamma \to 0} X' (X'X + \gamma I) X.
\]

(16)

2.2 Consistent estimation of the noise level

A consistent estimator of the noise level \( \sigma^2 \) is obviously crucial to construct a confidence interval. Since the lasso estimator \( \hat{\beta}_{\text{lasso}} \) is a consistent estimator, it can be used to construct such a consistent estimator of \( \sigma^2 \),

\[
\hat{\sigma}^2 = \frac{1}{n - \hat{s}} \hat{\varepsilon}' \hat{\varepsilon},
\]

(17)

where \( \hat{s} \) is the number of nonzero elements in \( \hat{\beta}_{\text{lasso}} \) and \( \hat{\varepsilon} = y - X\hat{\beta}_{\text{lasso}} \). This estimator is shown in Reid et al. (2013) to be more reliable than the frequently used scaled lasso by Sun and Zhang (2012).

3 Theoretical results

This section derives the main results of the paper. Proofs to the theorems can be found in Appendix C. First, Section 3.1 provides the necessary assumptions on the regressor matrix and the coefficient vector. These assumptions guarantee that the bias correction by the initial estimator is effective when we use the lasso estimator. In addition, they are used to show that the proposed approximate inverses are sufficiently accurate. Section 3.2 uses these assumptions to present the main result of the paper.

The main results is Theorem 1, which states that approximating the high dimensional inverse covariance matrix with a scaled Moore-Penrose inverse,
or regularizing with random least squares or ridge regression, results in estimators which are approximately normally distributed. From this theorem it follows that confidence intervals of coefficient estimates of these estimators can be constructed by standard procedures. Section 3.3 shows how we arrive at this result. Theorem 2 and 3 guarantee that the bias of the Moore Penrose estimator is small and vanishes compared to the variance. Similar results for random least squares are derived in Theorem 4. Another key result of the paper is presented in Section 3.4 by Theorem 5, which argues that for a particular choice of the diagonal scaling matrix, the width of the confidence intervals for random least squares is at most as large as the width of the confidence intervals under the Moore-Penrose inverse.

### 3.1 Assumptions

This section provides the assumptions from which we derive the theoretical results in the paper. First, we provide the necessary assumptions on the regressor matrix \( X \), which parallel those in Fan and Lv (2008) and Wang and Leng (2015)

**Assumption 1** The regressor matrix \( X \) can be written as

\[
X = Z \Sigma^{1/2} = V S U \Sigma^{1/2},
\]

where the rows of the \( n \times p \) matrix \( Z \) are generated from a spherically symmetric distribution, \( V \) is an \( n \times n \) orthogonal matrix, \( S \) is an \( n \times p \) matrix of singular values, and \( U \) a \( p \times p \) orthogonal matrix. Standard properties of these matrices can be found in Appendix A. The condition number of the population covariance matrix \( \Sigma \) is bounded by a constant, i.e.

\[
\kappa(\Sigma) = \frac{\lambda_{\text{max}}(\Sigma)}{\lambda_{\text{min}}(\Sigma)} \leq c_{\kappa}
\]

Second, we impose an assumption on the number of non-zero coefficients \( ||\beta||_0 = s_0 \). For lasso consistency, one usually takes \( s_0 = o(\sqrt{n \log p}) \). However, as noted in Van de Geer et al. (2014) and Javanmard and Montanari (2014), the following, slightly stronger, assumption is needed when constructing confidence intervals

**Assumption 2** The sparsity satisfies \( s_0 = o\left(\sqrt{n \log p}\right) \)

Third, Bühlmann et al. (2013) shows that Assumption 1 implies the compatibility condition, under which strong results on the \( l_1 \) and \( l_2 \) norm of the error of the lasso estimator compared to the true \( \beta \) are available.
Assumption 3 When $s_0$ is the true set of non-zero coefficients, then the compatibility condition is satisfied for this set if

$$||\beta_{S_0}||_1 \leq \frac{\sqrt{s_0}||X\beta||_2}{\sqrt{n}\phi_0}$$

for all $\beta$ for which $||\beta_{S_0}||_1 \leq 3||\beta_{S_0}||_1$ and $\phi_0 > 0$.

Under Assumption 3, the lasso estimator satisfies

$$||\beta - \hat{\beta}_{\text{lasso}}||_1 = O\left(s_0\sqrt{\frac{\log p}{n}}\right).$$

Using this bound, we deduce in Section 3.3 the required accuracy of the approximate inverse in order for the bias to vanish.

3.2 Distribution of the estimators

This section presents the main result of the paper; the approximate distributions of the estimators discussed in Section 2. Define the following diagonal matrices

$$D_{ii}^{MP} = \left(\frac{1}{n}X'(XX')^{-1}X\right)_{ii}^{-1}$$

$$D_{ii}^{RLS} = \left(\frac{1}{n}E_R[R(R'X'R)^{-1}R']X'X\right)_{ii}^{-1}$$

$$D_{ii}^{RI} = \left(\frac{1}{n}(X'X + \gamma I_p)^{-1}X'X\right)_{ii}^{-1}$$

where $R$ is a $p \times k$ matrix of independent standard normal entries.

Furthermore, set the penalty parameters for respectively random least squares and the ridge adjusted inverse such that

$$k = k^* = \left(1 - c\sqrt{\frac{\log p}{n}}\right)(n - 1)$$

$$\gamma = \gamma^* = O\left(\sqrt{\frac{\log p}{n}}\right)$$

where $c$ is a positive constant. Now the following theorem holds.
Theorem 1 Consider the linear model in (1), and suppose Assumption 1 and Assumption 2 hold. When \( \hat{\beta}^c \) is defined as \( \hat{\beta}^c = \frac{1}{n} My - C \), we have

\[
\sqrt{n}(\hat{\beta}^c - \beta) \sim N\left(0, \sigma^2 M M'\right),
\]

with \( C = \left(\frac{1}{n} MX - I_p\right)\beta_{\text{lasso}} \), and for the following choices of \( M \) and the relevant penalty parameter:

\[
M = D^M P X'(X X')^{-1}
\]

\[
M = D^{RLS} E_R [R(R' X' X R)^{-1} R'] X'
\]

\[
M = D^{RI} (X' X + \gamma^* I_p)^{-1} X'
\]

where \( R \) is a \( p \times k^* \) matrix with independent standard normal entries.

From Theorem 1 it follows that we can use the classical procedure to construct confidence intervals, given in (4).

3.3 Approximate inverse accuracy

This section provides details on how we arrive at Theorem 1. First we show which properties of the estimators need to be proven for Theorem 1 to hold. We show that it is sufficient for the approximate inverse to satisfy certain bounds. Second, we prove that these properties hold for each of the methods used to construct an approximate inverse. Proofs of the theorems can be found in Appendix C.

For notational purposes, we rewrite the estimator in (3) as

\[
\sqrt{n} \left(\hat{\beta}^c - \beta\right) = Z + \Delta
\]

where

\[
\Delta = \sqrt{n} \left(\frac{1}{n} MX - I\right) (\beta - \hat{\beta}_{\text{lasso}})
\]

\[
Z = \frac{1}{\sqrt{n}} M \varepsilon
\]

When \( ||\Delta||_\infty = o(Z) \) for a specific construction method for \( M \), the corrected estimator is approximately unbiased. Later in this section we will show that under each proposed specification for the approximate inverse \( M \), we have that \( Z = O(1) \) under a wide class of distributions of \( \varepsilon \). This implies that the bias term should satisfy \( ||\Delta||_\infty = o(1) \). Since the bias \( ||\Delta||_\infty \) can be decomposed as

\[
||\Delta||_\infty \leq \sqrt{n} \left||n^{-1} MX - I\right||_{\text{max}} ||\beta - \hat{\beta}_{\text{lasso}}||_1,
\]

11
it follows from Assumption 2 and (21) that the approximate inverse $M$ is sufficiently accurate when

$$\left\| \frac{1}{n} MX - I \right\|_{\text{max}} = O\left( \sqrt{\frac{\log p}{n}} \right).$$

(27)

Note that if we are willing to assume that all coefficients are local-to-zero, i.e. $\max_{i=1,\ldots,p} \beta_i = O\left( \sqrt{n^{-1} \log p} \right)$, then no lasso correction term is needed. This might not be an unrealistic assumption when working with economic data, where effects are generally small.

The following paragraphs show for each proposed method that the bias term $\Delta$ is small, and vanishes compared to the variance $Z$. In other words, we prove that (27) holds and $Z = O(1)$.

### 3.3.1 Diagonally scaled Moore-Penrose inverse

The next theorem ensures that the bias of the scaled Moore-Penrose inverse estimator is small with high probability.

**Theorem 2** Define $D_{ii}^{MP} = \left( \left[ \frac{1}{n} X'(XX')^{-1} X \right]_{ii} \right)^{-1}$ for $i = 1, \ldots, n$, then we have

$$P \left( \left| \frac{1}{n} D_{ii}^{MP} X'(XX')^{-1} X - I \right|_{ij} > a \sqrt{\frac{\log p}{n}} \right) = O(p^{-\hat{c}})$$

(28)

with $\hat{c} = a^2 \frac{cp}{c_p - \kappa} - 2$ where $c = 1 + \epsilon$, $\kappa = \frac{\lambda_{\text{max}}(\Sigma)}{\lambda_{\text{min}}(\Sigma)}$ and $c_p < \frac{p}{n}$

In addition to a small bias, we now need to show that the bias vanishes compared to the variance of the estimator. The following theorem guarantees that this is indeed the case under a wide class of distributions of the error term.

**Theorem 3** The variance of each estimated coefficient satisfies

$$\left[ \frac{1}{\sqrt{n}} D^{MP} X'(XX') \varepsilon \right]_i = O(1)$$

(29)

Furthermore, if the errors $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d$(0, \sigma^2)$ distributed with a subexponential norm of $K < \infty$, then as $n \rightarrow \infty$

$$\left[ \frac{1}{\sqrt{n}} D X'(XX') \varepsilon \right]_i \sim N \left( 0, \sigma^2 \frac{d^2}{n} e_i X'(XX')^{-2} X e_i \right)$$

(30)
3.3.2 Diagonally scaled random least squares

The key to the behavior of the regularized covariance matrix in random least squares, is the projection dimension $k$. The following theorem suggests a choice of $k$ for which the bias remains small relative to the variance.

**Theorem 4** Define $\hat{\kappa} = \frac{\lambda_{\text{max}}(\hat{\Lambda})}{\lambda_{\text{min}}(\hat{\Lambda})}$. If we choose

$$k = \left(1 - \frac{a}{\hat{\kappa}} \sqrt{\frac{\log p}{n}}\right)(n - 1)$$

(31)

then

$$P\left(\left|\frac{1}{n}D^{RLS}E[R(R'X'XR)^{-1}R']X'X - I\right|_{ij} > a\sqrt{\frac{\log p}{n}}\right) \leq p^{-\tilde{c}}$$

(32)

with $\tilde{c}$ as in Theorem 1. Furthermore

$$\left|\frac{1}{\sqrt{n}}D^{RLS}E[R(R'X'XR)^{-1}R']X'\epsilon\right|_{i} = O(1)$$

(33)

In the proof of Theorem 4, we show that when $k$ is sufficiently close to $n$ the regularized inverse approximates the Moore-Penrose inverse. The results from Section 3.3.1 can then be invoked to show that the bias of the random least squares estimator remains small. Note that to reduce the variance, we want to choose $k$ as small as possible.

By drawing the entries of $R$ from a standard normal distribution, the columns and rows of $R$ are only orthogonal with high probability. One might wonder if the lack of exact orthogonality in the columns of $R$ poses a problem for the proof of Theorem 4. Because of the following result this is not the case. Consider the QR decomposition $R = QT$ where the columns of $Q$ is a $p \times k$ matrix with orthogonal columns, and $T$ is an invertible $k \times k$ matrix. Since $T$ is invertible, we have the following equality

$$E_{R}[R(R'X'XR)^{-1}R'X'] = E_{Q}[Q'X'XQ]^{-1}Q'X'X$$

(34)

Because of this result, we can implicitly assume the columns of $R$ to be orthogonal.

3.3.3 Diagonally scaled ridge adjustment

Because of the well known relation between the generalized inverse and ridge adjusted covariance matrices displayed in (16), intuition suggests that for a sufficiently small penalty parameter, the results under a Moore-Penrose inverse carry over to a ridge adjusted estimator. This is indeed the case under the penalty parameter as in Theorem 1. A proof is provided in the Appendix.
3.4 Power increase by regularization

This section presents the second main result of the paper, stating that the width of the confidence intervals for random least squares is at most as large as the width of the confidence intervals under the Moore-Penrose inverse.

**Theorem 5** For the choice of $k$ as in Theorem 3, we have for $d_i = D_{ii}^{MP}$ or $d_i = D_{ii}^{RLS}$

$$\frac{d_i}{\sqrt{n}} \left\| \epsilon_i E \left[ R(R'X'XR)^{-1}R' \right] X' \right\|_2 - \frac{d_i}{\sqrt{n}} \left\| \epsilon_i X'(X'X)^{-1} \right\|_2 \leq 0 \quad (35)$$

Note that using $D_{ii}^{MP}$ for the Moore-Penrose inverse and $D_{ii}^{RLS}$ for the random least squares estimator yields valid confidence intervals, but no ordering in terms of power can be analytically obtained. However, in most cases we have encountered, the inequality in Theorem 5 is satisfied when using the diagonal matrix specific to the estimator under consideration.

4 Monte Carlo Experiments

We now examine the finite sample behaviour of the Moore-Penrose, random least squares, and ridge estimator in a Monte Carlo experiment.

4.1 Monte Carlo set-up

The data generating process takes the form

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \quad (36)$$

where $y$ is a $n \times 1$ response vector, $X$ a $n \times p$ regressor matrix, and $\beta$ a $p \times 1$ vector of unknown regressor coefficients. The rows of $X$ are fixed i.i.d. realizations from $\mathcal{N}_p(0, \Sigma)$. We specify two different covariance matrices:

- Equicorrelated: $\Sigma_{jk} = 0.8, \quad \forall j \neq k, \quad \Sigma_{jj} = 1 \quad \forall j,$
- Toeplitz: $\Sigma_{jk} = 0.9^{|j-k|}, \quad \forall j, k.$

The strength of the individual predictors is considered local-to-zero by setting $\beta = \sqrt{\frac{\sigma^2}{n}} \cdot b \cdot \iota_s$ for a fixed constant $b$. The vector $\iota_s$ contains $s$ randomly chosen non-zero elements that are equal to one. We refer to $s$ as the sparsity of the coefficient vector. We vary the signal strength $b$ and the sparsity $s$ across different Monte Carlo experiments.
Results are based on 1,000 replications of the data generating process (36). In each replication the predictors in $X$ and the coefficients in $\beta$ are generated. The number of predictors is set equal to $p = 200$, which is much larger than the sample size $n = 100$. The different experiments vary over signal strength $b$, sparsity $s$, and covariance matrix $\Sigma$. We report average results for nonzero coefficients and zero coefficients.

We estimate the coefficients by the Moore-Penrose pseudoinverse, random least squares, and ridge estimator, which are corrected by a lasso estimator. In the random least squares estimator we average over $N = 1000$ realizations of the regularized covariance matrix with subspace dimension $k = 90$. We set the penalty parameter in the lasso estimator as the penalty parameter corresponding to the lowest mean squared error over a grid of one hundred values. We compare the performance of the random least squares estimator against ridge regression with a penalty parameter equal to $1/n$ as in Bühlmann and Van De Geer (2011).

4.2 Simulation Results

Table 1 shows the Monte Carlo simulation results for the set of experiments with an equicorrelated covariance matrix and Table 2 with a Toeplitz covariance matrix. The tables report the estimated coefficients, standard errors, coverage rates, and power of the Pseudo Inverse, Random Least Squares, and Ridge Regression. Settings vary over the number ($s = 3, 15$) and signal strength ($b = 2, 5$, which corresponds to coefficients of size 0.5 and 0.2 respectively) of nonzero coefficients.

We find a downward bias for the nonzero coefficients for all methods in all settings under consideration. This bias is attenuated when we remove the bias correction with the lasso estimator. The bias decreases in sparsity, which means that nonzero coefficients are more precisely estimated when there are relatively few of them. Compared to the benchmark models, random least squares seems to estimate slightly more biased coefficients. For all methods, the coefficients which are set to zero in the data generating process are estimated very close to zero.

Random least squares produces the most efficient estimates relative to ridge regression and pseudo-inverse regression. Standard errors of the random least squares estimates are lower than the benchmark models in all experiments, where ridge is again a more efficient estimator relative to the pseudo-inverse. The same ordering holds for the power, for which random least squares also performs best and the pseudo-inverse worst. The highest power is achieved in a sparse setting with a strong signal strength.

The considered methods obtain a coverage rate close to the nominal rate
### Table 1: Monte Carlo simulation: Equicorrelated Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>s=3</th>
<th>s=15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coef.</td>
<td>SE</td>
</tr>
<tr>
<td>Pinv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b=2</td>
<td>0.191</td>
<td>0.303</td>
</tr>
<tr>
<td>b=0</td>
<td>0.000</td>
<td>0.303</td>
</tr>
<tr>
<td>RLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b=2</td>
<td>0.190</td>
<td>0.283</td>
</tr>
<tr>
<td>b=0</td>
<td>0.000</td>
<td>0.282</td>
</tr>
<tr>
<td>Ridge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b=2</td>
<td>0.191</td>
<td>0.291</td>
</tr>
<tr>
<td>b=0</td>
<td>0.000</td>
<td>0.291</td>
</tr>
</tbody>
</table>

Note: this table reports the estimated coefficients (coef.), standard errors (SE), coverage rates (CR), and power of the Pseudo Inverse (Pinv), Random Least Squares (RLS), and Ridge Regression (Ridge) for different Monte Carlo experiments where the regressors have an equicorrelated covariance matrix as specified in (37). Settings vary over the number ($s = 3, 15$) and signal strength ($b = 2, 5$) of nonzero coefficients. The number of observations is $n = 100$ and the number of regressors $p = 200$. The subspace dimension in RLS is set equal to $k = 0.9n$ and we average over $N = 1000$ realizations of projection matrices, and the penalty parameter in Ridge $\lambda = 1/n$. Results are based on 1000 Monte Carlo replications.

of 95%. In case of the equicorrelated covariance matrix, the coverage rate increases with sparsity for the nonzero coefficients and decreases with sparsity for the zero coefficients. These relations seem only to hold for a strong signal in case of the Toeplitz covariance matrix. In general, the quality of the results seem to deteriorate when the data is generated using a Toeplitz instead of an equicorrelated covariance matrix. The bias and standard errors increase, while the power declines.

In Section 2 we note that when the effects of regressors on the variable of interest are small, no lasso correction term is needed. To examine the effect of the correction term on the estimation results, we compare the bias corrected simulation results in Tables 1 and 2 with uncorrected results in Appendix D. In general, we find again that bias decreases in sparsity and random least squares seem to be more biased than the benchmark models. However, except for high sparsity and a weak signal, the uncorrected estimates have
Table 2: Monte Carlo simulation: Toeplitz Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>s=3</th>
<th></th>
<th></th>
<th></th>
<th>s=15</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coef.</td>
<td>SE</td>
<td>CR</td>
<td>power</td>
<td>coef.</td>
<td>SE</td>
<td>CR</td>
<td>power</td>
</tr>
<tr>
<td>Pinv</td>
<td>b=2</td>
<td>0.182</td>
<td>0.352</td>
<td>0.952</td>
<td>0.080</td>
<td>0.169</td>
<td>0.339</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.002</td>
<td>0.351</td>
<td>0.949</td>
<td>0.005</td>
<td>0.167</td>
<td>0.339</td>
<td>0.938</td>
</tr>
<tr>
<td>RLS</td>
<td>b=2</td>
<td>0.181</td>
<td>0.309</td>
<td>0.951</td>
<td>0.088</td>
<td>0.167</td>
<td>0.298</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.002</td>
<td>0.309</td>
<td>0.950</td>
<td>0.006</td>
<td>0.167</td>
<td>0.298</td>
<td>0.949</td>
</tr>
<tr>
<td>Ridge</td>
<td>b=2</td>
<td>0.182</td>
<td>0.323</td>
<td>0.952</td>
<td>0.086</td>
<td>0.168</td>
<td>0.312</td>
<td>0.937</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.002</td>
<td>0.323</td>
<td>0.950</td>
<td>0.005</td>
<td>0.168</td>
<td>0.311</td>
<td>0.949</td>
</tr>
<tr>
<td>Pinv</td>
<td>b=5</td>
<td>0.455</td>
<td>0.350</td>
<td>0.934</td>
<td>0.274</td>
<td>0.418</td>
<td>0.340</td>
<td>0.907</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.002</td>
<td>0.349</td>
<td>0.950</td>
<td>0.008</td>
<td>0.339</td>
<td>0.953</td>
<td>0.316</td>
</tr>
<tr>
<td>RLS</td>
<td>b=5</td>
<td>0.450</td>
<td>0.308</td>
<td>0.929</td>
<td>0.331</td>
<td>0.411</td>
<td>0.299</td>
<td>0.892</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.002</td>
<td>0.307</td>
<td>0.950</td>
<td>0.009</td>
<td>0.298</td>
<td>0.952</td>
<td>0.298</td>
</tr>
<tr>
<td>Ridge</td>
<td>b=5</td>
<td>0.452</td>
<td>0.321</td>
<td>0.933</td>
<td>0.312</td>
<td>0.414</td>
<td>0.312</td>
<td>0.898</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.002</td>
<td>0.321</td>
<td>0.950</td>
<td>0.009</td>
<td>0.312</td>
<td>0.953</td>
<td>0.298</td>
</tr>
</tbody>
</table>

Note: this table reports the results for different Monte Carlo experiments where the regressors have a toeplitz covariance as specified in (38). For additional information, see the note following Table 1.

an upward bias for the nonzero coefficients. This upward bias seems smaller than the downward bias found for the lasso corrected estimates, but the estimated coefficients for the zero coefficients are less precise. Due to the smaller bias, power increases, where the coverage rates are less close to the nominal value. In contrast to corrected results, the coverage rates increase in sparsity and decrease with signal strength.

In sum, we find that even in small samples with high dimensional regressor matrix, the three proposed estimators in this paper provide valid confidence intervals with coverage rates very close to the nominal value of 95%. Moreover, the simulation results seem to confirm the theoretical result that random least squares improves in efficiency relative to the Moore-Penrose estimator. We find a small downward bias in the coefficient estimates of the estimators. However, the estimation results with no lasso correction term are less biased in the considered experiments and achieve higher power. This is even the case for $b = 5$ which corresponds with $\beta_i = 0.5 \gg \sigma \sqrt{\frac{\log p}{n}} = 0.231$. However, the increase in power comes at the cost of smaller coverage rates and a potential less severe but upward bias.
5 Empirical Application

This section applies the proposed estimators to a macroeconomic dataset. We examine the relation between a large number of macroeconomic and financial indicators and the real gross domestic product of the U.S. economy.

5.1 Data

We use the FRED-QD database consisting of 254 quarterly macroeconomic and financial series running from the second quarter of 1987 through the third quarter of 2015. Less variables are available before this time period and because records of the variables with FRED mnemonic SPCS20RSA, ACOGNOx, and EXUSEU have a later starting point, we exclude these variables from our analysis. The data can be grouped in fourteen different categories: national income and product accounts (1), industrial production (2), employment and unemployment (3), housing (4), inventories, orders, and sales (5), prices (6), earnings and productivity (7), interest rates (8), money and credit (9), household balance sheets (10), exchange rates (11), other (12), stock markets (13) and non-household balance sheets (14). The data is available from the website of the Federal Reserve Bank of St. Louis, together with code for transforming the series to render them stationary and to remove severe outliers. The data and transformations are described in detail by McCracken and Ng (2016). After transformation, we find a small number of missing values, which are recursively replaced by the value in the previous time period of that variable. Finally, we subtract the mean of each variable and divide the variables by their standard deviation.

5.2 Regression

The coefficients $\beta$ are estimated in the regression equation

$$y = Z\delta + X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

where $y$ equals the real gross domestic product of the U.S. economy (FRED mnemonic GDPC96), $Z$ includes an intercept along with four lags of the quarterly dependent variable $y$, and $X$ consists of the remaining variables in the database which are not in the same group as $y$. Since we are only interested in the macroeconomic relations in $\beta$, we partial out the variables in $Z$ before estimating $\beta$. After initialization and the loss in degrees of freedom by partialling out $Z$, we are left with $n = 105$ observations.

When estimating by random least squares, we choose the subspace dimension $k = 95$ and $N = 1000$ realizations of the regularized covariance
Table 3: Significant effects on Real Gross Domestic Product

<table>
<thead>
<tr>
<th>group</th>
<th>FRED mnemonic</th>
<th>Pinv</th>
<th>RLS</th>
<th>Ridge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Coef.</td>
<td>SE</td>
<td>Coef.</td>
</tr>
<tr>
<td>2</td>
<td>IPFINAL</td>
<td>0.072</td>
<td>0.037</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>USWTRADE</td>
<td>-0.061</td>
<td>0.029</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>HOABS</td>
<td>0.746</td>
<td>0.262</td>
<td>0.759</td>
</tr>
<tr>
<td>3</td>
<td>HOANBS</td>
<td>0.162</td>
<td>0.034</td>
<td>0.162</td>
</tr>
<tr>
<td>5</td>
<td>NAPMII</td>
<td>0.051</td>
<td>0.023</td>
<td>0.051</td>
</tr>
<tr>
<td>7</td>
<td>OPHNFB</td>
<td>0.067</td>
<td>0.019</td>
<td>0.067</td>
</tr>
<tr>
<td>7</td>
<td>OPHPBS</td>
<td>0.767</td>
<td>0.060</td>
<td>0.752</td>
</tr>
<tr>
<td>8</td>
<td>CPF3MTB3Mx</td>
<td>0.060</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>MED3TB3Mx</td>
<td>0.065</td>
<td>0.027</td>
<td>0.065</td>
</tr>
<tr>
<td>9</td>
<td>AMBSLREALx</td>
<td>0.027</td>
<td>0.013</td>
<td>0.027</td>
</tr>
<tr>
<td>12</td>
<td>UMCSNTx</td>
<td>0.064</td>
<td>0.029</td>
<td>0.064</td>
</tr>
<tr>
<td>13</td>
<td>VXOCLX</td>
<td>0.051</td>
<td>0.023</td>
<td>0.051</td>
</tr>
<tr>
<td>14</td>
<td>GFDEGDQ188S</td>
<td>-0.041</td>
<td>0.020</td>
<td>-0.041</td>
</tr>
</tbody>
</table>

Note: this table reports the estimated coefficients (Coef.) and standard errors (SE) which are significantly different from zero on a five percent significance level, estimated by Pseudo-Inverse (Pinv), Random Least Squares (RLS), and Ridge Regression (Ridge). The group numbers correspond to the FRED variable categories, the variable names correspond to the fred mnemonics, and variable descriptions are given in Appendix E.

matrix. The penalty parameter in the lasso estimator for the lasso correction corresponds to the lowest mean squared error over a grid of one hundred values, and the penalty parameter in ridge regression is equal to $1/n$.

### 5.3 Empirical Results

Table 3 shows the estimated coefficients and standard errors which are significantly different from zero on a five percent significance level in the regression of the economic indicators on the real gross domestic product. In general, random least squares estimates lower standard errors compared to the benchmark methods. Ridge regression finds 11 out of the 231 coefficients to be significant, which is slightly higher for random least squares with 13 coefficients. Ridge regression differs only for three variables with random least squares. The Pseudo-inverse regression estimates only 2 coefficients to be significant, which corresponds to the theoretical finding that the random least squares and ridge estimators are more efficient in terms of statistical power compared to the Moore-Penrose estimator.
We find that employment and productivity have the largest effect on real gross domestic product. Hours of all persons worked in the business sector (HOABS), real output per hour of all persons in the business sector (OPHBS), and hours of all persons worked in the nonfarm business sector (HOANBS) have large positive coefficients of respectively 0.759, 0.162, and 0.752 for random least squares. The descriptions of the other variables can be found in Appendix E. Figure 1 shows that except for these three variables concerning employment and earnings, all coefficients are close to zero. We do not find any significant effect of variables in the categories housing (4), prices (6), household balance sheets (10), and exchange rates (11). Random least squares finds two significant negative effects on the real gross domestic product; all employees in wholesale trade (USWTRADE) and the total public debt as percent of GDP (GFDEGDQ188S). Ridge regression only finds a significant negative effect for the public debt. The negative effect assigned to the number of employees in wholesale trade is remarkable, but note that employment also effects GDP positively via HOABS and HOANBS, which
makes the net effect of employment on real GDP positive.

6 Conclusion

This paper proposes three novel methods for constructing confidence intervals in high-dimensional linear regression models, where the number of unknown coefficients are allowed to increase almost exponentially with the number of observations. We approximate the inverse covariance matrix by a Moore-Penrose pseudo-inverse, random least squares, or a ridge regularization. We prove that these estimators, after a diagonal scaling and a lasso bias correction, are approximately normally distributed. From this result follows that confidence intervals of coefficient estimates can be constructed using standard procedures. Another key result of the paper is that the regularization by random least squares and ridge result in confidence intervals that are smaller than those obtained under the Moore-Penrose inverse.

Monte Carlo experiments show that, even in small samples with high dimensional regressor matrix, the proposed estimators indeed provide valid confidence intervals with coverage rates very close to the nominal value of 95%. Moreover, the simulation results seem to confirm that random least squares and ridge improve in efficiency relative to the Moore-Penrose estimator. The empirical application provides similar findings. In a high-dimensional regression of macroeconomic and financial indicators on the real gross domestic product of the United States economy, the Pseudo-inverse regression estimates only 2 coefficients to be significant. However, the more efficient estimators random least squares and ridge regression find respectively 13 and 11 out of the 231 coefficients to be significant.

References


A Assumptions on the regressor matrix

Denote by $\mathcal{O}(p)$ the group of $p \times p$ orthogonal matrices. We introduce a matrix $Z$ with rows that are generated from a spherically symmetric distribution, that is

$$Z \overset{(d)}{=} ZT, \quad T \in \mathcal{O}(p)$$  \hspace{1cm} (40)

Then $Z$ can be decomposed by a singular value decomposition as

$$Z = VSU'$$  \hspace{1cm} (41)

where $V \in \mathcal{O}(n)$, $S$ is a $n \times p$ matrix of singular values and $U \in \mathcal{O}(p)$. Since $Z$ is invariant under right multiplication with an orthogonal matrix, $U$ is uniformly distributed on $\mathcal{O}(p)$.

The matrix of singular values $S$ contains at most $n$ non-zero singular values located on the diagonal of the left $n \times n$ block of $S$. Therefore, an equivalent expression to (41) is

$$Z = VS_nU'_n$$  \hspace{1cm} (42)

where $S_n$ is an $n \times n$ matrix with the non-zero singular values on its diagonal, and $U_n$ is a $p \times n$ matrix that satisfies

$$U_n' = [I_n, O_{n,p-n}]U'$$  \hspace{1cm} (43)

Since $U$ is uniformly distributed over $\mathcal{O}_p$, it follows (Fan and Lv, 2008) that $U_n$ is uniformly distributed over the Stiefel manifold $V_{n,p}$ defined as

$$V_{n,p} = \{ X \in \mathbb{R}^{p \times n} : X'X = I_n \}$$  \hspace{1cm} (44)

B Auxiliary definitions and lemma’s

Definition 1 (Matrix Angular Gaussian Distribution, Chikuse (1990))
Suppose the entries of a $p \times n$ matrix $Z$ are standard normal and independently distributed. Define $H_{\Sigma^{1/2}Z} = \Sigma^{1/2}Z(Z'\Sigma Z)^{-1/2}$. Then $H_Z$ has the density function

$$f_{H_Z} = |\Sigma|^{-n/2}|H_Z'\Sigma^{-1}H_Z|^{-p/2}$$  \hspace{1cm} (45)

This distribution is called the matrix angular Gaussian distribution with parameter $\Sigma$ defined on the Stiefel manifold $V_{n,p}$ and denoted as MACG($\Sigma$).

Lemma 1 (Reduction to standard normal random variables) Define $Z$ as a $p \times n$ matrix with independent standard normal entries. For any matrix $U_n$ that is distributed uniformly over $V_{n,p}$, we have that (Chikuse (2012), page 29)

$$U_n = Z(Z'Z)^{-1/2}$$  \hspace{1cm} (46)
Lemma 2 (Manifold decomposition, Chikuse (2012)) Let $H$ be a $p \times n$ random matrix on the Stiefel manifold $V_{n,p}$, which is decomposed as

$$H = (h_1, H_2)$$

where $h_1$ is a $p \times 1$ vector and $H_2$ is a $p \times n - 1$ matrix. Then we can write

$$h_1 = G(H_2)T$$

where $G(H_2)$ is any $p \times p - n + 1$ matrix chosen so that $[H_2, G(H_2)] \in O(p)$. As $H_2$ takes values in $V_{n-1,p}$, $T$ takes values in $V_{1,p-n+1}$ and the relationship is one-to-one.

Lemma 3 (Wang and Leng (2015)) Let $H$ be a $p \times n$ random matrix on the Stiefel manifold $V_{n,p}$, which follows the Matrix Angular Central Gaussian (MACG) distribution with covariance matrix $\Sigma$. After decomposing the Stiefel manifold $H = (G(H_2)T, H_2)$, with $T$ a $(p - n + 1) \times 1$ and $H_2$ a $p \times (n - 1)$ matrix, we have

$$T|H_2 \sim ACG(G(H_2)'\Sigma G(H_2))$$

As $G(H_2)T$ is simply a linear transformation of $T$, if we define

$$\tilde{\Sigma} = G(H_2)G(H_2)'\Sigma G(H_2)G(H_2)'$$

then we also have

$$h_1|H_2 \sim ACG(\tilde{\Sigma})$$

This implies that if $h_1 = (h_{11}, \ldots, h_{p1})'$, then

$$h_{11}|H_2 \overset{(d)}{=} \frac{z_i}{\sqrt{z_1^2 + \ldots + z_p^2}}$$

where $z \sim N(0, \tilde{\Sigma})$

Lemma 4 (Fan and Lv (2008); Wang and Leng (2015)) Define $e_i$ as a standard basis vector with its $i$-th entry equal to 1 and all others equal to zero. Note that

$$e_1HH'e_2 = e_1HQ'H'e_2, \quad Q \in O(n)$$

Now define $Q \in O(n - 1)$ and $\tilde{Q} = \begin{pmatrix} 1 & 0_{1 \times n-1} \\ 0_{n-1 \times 1} & Q \end{pmatrix}$ and choose $Q$ such that it rotates $H \rightarrow \tilde{H} = HQ$ into a frame where $e_1\tilde{H} = [\tilde{h}_{11}, 0_{1 \times n-1}]$. In terms of the rotated frame, we have

$$e_1HH'e_2 = e_1'\tilde{H}\tilde{H}e_2 = \tilde{h}_{11}\tilde{h}_{12}$$
implying that
\[ e_1HH'e_2^{(d)} = h_{11}h_{12} | e_1H = h_{11} \] (55)

Denote the first row of \( H \) by \([H]_{1,1:n} = [h_{11}, h'_{2}]\). Then \( e_1HH'e_1 = h_{11}^2 + h'_2h_2 \) and thus \( e_1H = [h_{11}, 0_{1 \times n-1}] \leftrightarrow e_1HH'e_1 = h_{11}^2 \). We can then rewrite (55) into
\[ e_1HH'e_2^{(d)} = h_{11}h_{12} | e_1HH'e_1 = h_{11}^2 \] (56)

**Lemma 5 (Berstein’s inequality)** We will use the following identity on the sum of independent \( \chi^2(1) \) variables
\[
P\left( \frac{1}{n} \sum_i \chi_i^2(1) > 1 + \epsilon \right) \leq \exp\left( -\frac{\epsilon^2 n}{8} \right)
\] (57)
\[
P\left( \frac{1}{n} \sum_i \chi_i^2(1) < 1 - \epsilon \right) \leq \exp\left( -\frac{\epsilon^2 n}{8} \right)
\]

**C Proofs**

**C.1 Proof of Theorem 1**

Using (41) we have that
\[ X'(XX')^{-1}X = \Sigma^{1/2}U_n(U_n'\Sigma U_n)^{-1}U_n'\Sigma \] (58)

By Lemma 1, we can write
\[ U_n = Z(Z'Z)^{-1/2} \] (59)

with the elements of \( Z \) standard normal and independently distributed. This implies
\[ X'(XX')^{-1}X = \Sigma^{1/2}Z(Z'\Sigma Z)^{-1}Z'\Sigma = HH' \] (60)

where we defined \( H = \Sigma^{1/2}Z(Z'\Sigma Z)^{-1/2} \).

We will separately bound the diagonal and off-diagonal elements of \( HH' \). The proof follows the approach by Wang and Leng (2015), with the exception of the diagonal scaling matrix that we apply.

**Diagonal terms of \( HH' \)** The diagonal elements of \( HH' \) are themselves not of particular interest, as we choose the diagonal matrix \( D \) such that the diagonal elements of \( \frac{1}{n}MX \) are all equal to one. However, to bound the off-diagonal elements, we nevertheless also require a bound on the diagonal
elements of $HH'$. We first construct bounds under the assumption that $\Sigma = I_p$. In a second step, we connect the more general case $\Sigma \neq I_p$ to these bounds.

If $\Sigma = I_p$, then each element of $H$ is distributed as

$$H_{ij} \overset{(d)}{=} \frac{z_{ij}}{\sqrt{\sum_{j=1}^{p} z_{ij}^2}} \quad (61)$$

Define $e_i$ the unit vector with its $i$-th element equal to 1 and all others equal to zero, then

$$e_i'HH'e_i = \frac{z_{i1}^2 + \ldots + z_{ip}^2}{z_{i1}^2 + \ldots + z_{ip}^2} \quad (62)$$

Using now the Bernstein inequality for $\chi^2$ random variables, (57), we have the following upper bound with high probability

$$P\left(\frac{z_{i1}^2 + \ldots + z_{ip}^2}{z_{i1}^2 + \ldots + z_{ip}^2} > \frac{n}{p} \frac{1 + \epsilon}{1 - \epsilon}\right) \leq \exp\left(-\frac{\epsilon^2 n}{8}\right) + \exp\left(-\frac{\epsilon^2 p}{8}\right) \quad (63)$$

where the last line holds because $p > n$. In a similar fashion we establish a lower bound

$$P\left(\frac{z_{i1}^2 + \ldots + z_{ip}^2}{z_{i1}^2 + \ldots + z_{ip}^2} < \frac{n}{p} \frac{1 - \epsilon}{1 + \epsilon}\right) \leq 2 \exp\left(-\frac{\epsilon^2 n}{8}\right) \quad (64)$$

Using Bonferonni's inequality this yields

$$P\left(e_1 U_n U_n' e_1 > \frac{c}{p} n \quad \cup \quad e_1 U_n U_n' e_1 < \frac{1}{c} \frac{n}{p}\right) \leq 4 \exp\left(-\frac{\epsilon^2 n}{8}\right) \quad (65)$$

with $c = \frac{1 + \epsilon}{1 - \epsilon} > 1$.

We will now use these results to establish a bound when $\Sigma \neq I$. The diagonal terms are easily bounded by the $l_2$ norm of $H'$, which satisfies

$$||H'v||_2^2 = vHH'v$$

$$= v\Sigma^{1/2} U(U'\Sigma U)^{-1} U'\Sigma^{1/2}$$

$$\leq \kappa v' U_n U_n' v \quad (66)$$

where the condition number $\kappa = \frac{\lambda_{\text{max}}(\Sigma)}{\lambda_{\text{min}}(\Sigma)} < \infty$. Similarly

$$vHH'v \geq \frac{1}{\kappa} v' U_n U_n' v \quad (67)$$
Since $U_n \overset{d}{=} QU_n$ with $Q \in \mathcal{O}(p)$, upon choosing $Q$ such that $Qv = e_1$, we immediately obtain

$$P \left( e_1 H H' e_1 > \frac{c \kappa n}{p} \right) \cup e_1 H H' e_1 < \frac{1}{c \kappa p} \right) \leq 4 \exp \left( -\frac{c^2 n}{8} \right)$$

with $c = \frac{1+\epsilon}{1-\epsilon}$ and $\kappa = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$.

**Off-diagonal elements** The proof for the off-diagonal elements is somewhat more involved. Because we multiply with the diagonal matrix $D$, we are interested in bounding $|e_1 H H' e_2|$. We separate three cases: (a) $e_1 H H' e_1 \leq \frac{c \kappa n}{p}$, (b) $\frac{c \kappa n}{p} < e_1 H H' e_1 \leq \frac{1}{c \kappa p}$, and (c) $e_1 H H' e_1 > \frac{c \kappa n}{p}$. Conditioning on these three cases and using the trivial fact that for any probability $P(\cdot) \leq 1$, it follows that

$$P \left( \left| \frac{e_1 H H' e_2}{e_1 H H' e_1} \right| > t \right) \leq P \left( e_1 H H' e_1 \leq \frac{c \kappa n}{p} \right) + P \left( e_1 H H' e_1 \geq \frac{1}{c \kappa p} \right)$$

$$+ \int_{\frac{1}{c \kappa p}}^{c \kappa \frac{n}{p}} P \left( \left| \frac{e_1 H H' e_2}{e_1 H H' e_1} \right| > t \bigg| e_1 H H' e_1 = t_1^2 \right) P \left( e_1 H H' e_1 = t_1^2 \right) dt_1$$

As before

$$e_1 H H' e_1 \overset{d}{=} h_{11} h_{12} \big| h_{11}^2 = e_1 H H' e_1$$

In which case

$$e_1 H H' e_2 \mid e_1 H H' e_1 = t_1^2 \overset{d}{=} h_{11} h_{12} \big| h_{11}^2 = t_1^2$$

Then after rewriting we have

$$h_{11} h_{11}^2 = t_1^2 \overset{d}{=} \frac{\sqrt{1 - t_1^2} \lambda_i}{\sqrt{\lambda_1^2 + \ldots + \lambda_p^2}} \overset{72}{\text{where}} (y_2, \ldots, y_p) \sim N(0, \Sigma).$$

Now we establish the following upper bound

$$P \left( \left| \frac{e_1 H H' e_2}{e_1 H H' e_1} \right| > t \bigg| h_{11}^2 = t_1^2 \right) = P \left( \frac{|h_{11} h_{12}|}{h_{11}^2} > t \bigg| h_{11}^2 = t_1^2 \right)$$

$$= P \left( \frac{\sqrt{1 - t_1^2} |y_1|}{\sqrt{y_1^2 + \ldots + y_p^2}} > |t_1| t \right) \overset{73}{=}$
Standard bounds on normal and $\chi^2$ distributed variables can be applied, and furthermore using the fact that $\tilde{y}$ has a rank $p - n$ degenerate covariance matrix, we have

$$P \left( |y_2| > \eta \sqrt{\lambda_{\max}(\Sigma)} \right) \leq 2e^{-\frac{\eta^2}{2}}$$

$$P \left( \sqrt{y_2^2 + \ldots + y_p^2} \leq \sqrt{\lambda_{\min}(\Sigma)} \sqrt{p - n (1 - \tilde{\eta})} \right) \leq e^{-\frac{1}{2} (p - n) \tilde{\eta}^2}$$

(74)

Then we know that

$$P \left( \frac{|y_2|}{\sqrt{y_2^2 + \ldots + y_p^2}} > \frac{\eta}{1 - \tilde{\eta}} \sqrt{\frac{\kappa}{p - n}} \right) \leq 2e^{-\frac{\eta^2}{2}} + e^{-\frac{1}{2} (p - n) \tilde{\eta}^2}$$

(75)

So we can choose

$$t = \sqrt{\frac{1 - t_1^2}{t_1^2}} \frac{\eta}{1 - \tilde{\eta}} \sqrt{\frac{\kappa}{p - n}}$$

(76)

In fact any $t_- < t$ is sufficient, so that we take $t_1^2 = c\kappa p$.

$$t > \sqrt{\frac{1 - \frac{p}{c p - n}}{\frac{n}{p - n}} \frac{\kappa}{\sqrt{n}}}$$

(77)

Now assume $\frac{p}{n} > c_p$, then

$$\sqrt{\frac{1}{c p / n - 1}} - \frac{1}{p / n - 1} \frac{\eta}{\sqrt{n}} > \sqrt{\frac{c p - \kappa}{c_p - 1}} \frac{\eta}{\sqrt{n}}$$

(78)

Choosing

$$\eta = a \sqrt{\frac{c_p - 1}{c_p - \kappa} \log p}$$

(79)

results in

$$P \left( \frac{|e_i H H^t e_2|}{e_i H H^t e_1} > a \sqrt{\frac{\log p}{n}} \right) \leq 2e^{-a^2 \frac{c_p - 1}{c_p - \kappa} \log p} + e^{-\frac{1}{2} (c_0 - 1) a^2 \tilde{\eta}^2}$$

(80)

And finally, taking the union bound over all pairs $e_i, e_j$ we have that for all $i, j$

$$P \left( \frac{|e_i H H^t e_j|}{e_i H H^t e_i} > a \sqrt{\frac{\log p}{n}} \right) \leq O \left( p^{-c} \right)$$

(81)

with $c = a^2 \frac{c_p - 1}{c_p - \kappa} - 2$. ■
C.2 Proof of Theorem 2

For all the estimators we should have
\[
\frac{1}{\sqrt{n}}M\epsilon = O(1) \tag{82}
\]
Following the proof by Wang and Leng (2015), with in addition the diagonal scaling matrix \(D\), we denote the individual error terms of the estimator of \(\beta_i\) as
\[
\eta_i = \frac{d_i}{\sqrt{n}} e_i'X(XX')^{-1}\epsilon
\]
\[
= \frac{d_i}{\sqrt{n}}||e_i'X(XX')^{-1}||_2 \sigma e_iX'(XX')^{-1}u \tag{83}
\]
where we define \(u \sim i.i.d. (0, 1)\).

We first bound the norm term
\[
\frac{d_i}{\sqrt{n}}||e_i'X(XX')^{-1}||_2 = \sqrt{n}||e_i'X(XX')^{-1}||_2 \tag{84}
\]
And
\[
||e_i'X(XX')^{-1}||_2^2 = e_iX'(XX')^{-2}Xe_i
\]
\[
\leq \frac{1}{\lambda_{\min}(XX')} e_iX'(XX')^{-1}Xe_i \tag{85}
\]
\[
||e_i'X(XX')^{-1}||_2^2 \geq \frac{1}{\lambda_{\max}(XX')} e_iX'(XX')^{-1}Xe_i
\]
The eigenvalues of \(XX' = ZZ'\) can be bounded
\[
\lambda_{\max}(ZZ') \leq \lambda_{\max}(\Sigma)\lambda_{\max}(ZZ')
\]
\[
\lambda_{\min}(ZZ') \geq \lambda_{\min}(\Sigma)\lambda_{\min}(ZZ') \tag{86}
\]
For spherically symmetric \(Z\), the results in Vershynin (2010) give
\[
P \left( \lambda_{\max} \left( \frac{1}{p}ZZ' \right) > (1 + \epsilon) \right) \leq e^{-\frac{\epsilon^2 p}{8}}
\]
\[
P \left( \lambda_{\min} \left( \frac{1}{p}ZZ' \right) < (1 - \epsilon) \right) \leq e^{-\frac{\epsilon^2 p}{8}} \tag{87}
\]
Finally, using the previously established bounds
\[
P \left( e_iX'(XX')^{-1}Xe_1 > \frac{c\kappa}{p} \frac{n}{p} \right) \leq 2e^{-\frac{c^2 n}{8}}
\]
\[
P \left( e_iX'(XX')^{-1}Xe_1 < \frac{1}{c\kappa} \frac{n}{p} \right) \leq 2e^{-\frac{c^2 n}{8}} \tag{88}
\]
we have that with high probability
\begin{equation}
\left( \frac{1}{\lambda_{\max}(\Sigma)} \frac{n}{p} \frac{1}{cn^2} \right)^{1/2} \leq \frac{d_i}{\sqrt{n}} \|e_iX'(XX')^{-1}\|_2 \leq \left( \frac{1}{\lambda_{\min}(\Sigma)} \frac{n}{p} \frac{1}{cn^2} \right)^{1/2}
\end{equation}
which shows that
\begin{equation}
\frac{d_i}{\sqrt{n}} \|e_iX'(XX')^{-1}\|_2 = O(1)
\end{equation}
We now turn to the second term of (83)
\begin{equation}
\frac{e_iX'(XX')^{-1}u}{||e_iX'(XX')^{-1}||_2}
\end{equation}
where \( u \sim i.i.d(0, 1) \).
Define \( a_i = e_1X'(XX')^{-1} = (a_{i1}, \ldots, a_{in}) \). Then we can write
\begin{equation}
\frac{e_iX'(XX')^{-1}u}{||e_iX'(XX')^{-1}||_2} = \frac{\sum_{j=1}^{n} a_{ij}u_j}{\sqrt{\sum_{j=1}^{n} a_{ij}^2}}
\end{equation}
Define
\begin{equation}
w_{ij} = \frac{a_{ij}}{\sqrt{\sum_{j=1}^{n} a_{ij}^2}}
\end{equation}
Then if \( u_1, \ldots, u_n \) are independent and identically distributed centered subexponential random variable with \( \sup_{p} p^{-1}(E|u_j|)^{p} = K \), then
\begin{equation}
P \left( |w_{ij}u_j| > t \right) \leq 2 \exp \left( -c \min \left( \frac{t^2}{K||w_{ij}||_2^2}, \frac{t}{K||w_{ij}||_\infty} \right) \right)
\end{equation}
\begin{equation}
= 2 \exp \left( -\frac{\tilde{c}}{K} t^2 n \right)
\end{equation}
This implies Lindebergs’s condition
\begin{equation}
\lim_{n \to \infty} \frac{\sum_{j=1}^{n} E \left[ a_{ij}^2 u_j^2 I \left[ \frac{a_{ij}u_j}{\sqrt{\sum_{j=1}^{n} a_{ij}^2}} > \delta \right] \right]}{\sum_{j=1}^{n} a_{ij}^2} \to 0
\end{equation}
So that as \( n \to \infty \)
\begin{equation}
\eta_i \overset{(d)}{=} N \left( 0, \frac{d^2}{n} e_iX'(XX')^{-2} X e_i \right)
\end{equation}
From which we conclude two things. First, \( \eta_i = O(1) \), so the error term is indeed larger than the bias. And second, the construction of the confidence intervals is valid under a wide class of distributions of the error \( e_i \).
C.3 Proof of Theorem 3

A convenient property of the inverse covariance matrix in (12) is that the eigenvectors of $\frac{1}{n} X'X = \hat{U} \hat{\Lambda} U'$ are maintained (Marzetta et al. (2011))

$$E[R(R'\hat{\Sigma}R)^{-1}R'] = \hat{U}E[\Phi'\hat{\Lambda}\Phi]^{-1}\Phi'\hat{U}'$$

(97)

where $\Phi = \hat{U}'R$. In addition the matrix $E[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']$ is a diagonal matrix. Now rewrite

$$\frac{1}{n}MX = U E[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi'] \hat{\Lambda} U'$$

(98)

with $L$ a diagonal matrix with the following elements on the diagonal

$$L_{ii} = \phi_i' (\Phi'\hat{\Lambda}\Phi)^{-1} \phi_i$$

$$= \phi_i' \left( \sum_{j \neq i} \hat{\lambda}_j \phi_j' \phi_j' + \hat{\lambda}_i \phi_i' \phi_i' \right)^{-1} \phi_i$$

$$= \phi_i' (A_{-i} + \hat{\lambda}_i \phi_i' \phi_i')^{-1} \phi_i, \quad A_{-i} \equiv \sum_{j \neq i} \hat{\lambda}_j \phi_j' \phi_j'$$

$$= \phi_i' \left( A_{-i}^{-1} - \frac{\hat{\lambda}_i A_{-i}^{-1} \phi_i' \phi_i' A_{-i}^{-1}}{1 + \hat{\lambda}_i \phi_i' \phi_i' A_{-i}^{-1} \phi_i} \right) \phi_i$$

$$= \frac{\phi_i' A_{-i}^{-1} \phi_i}{1 + \hat{\lambda}_i \phi_i' A_{-i}^{-1} \phi_i}$$

$$= \frac{\mu_i}{1 + \hat{\lambda}_i \mu_i}$$

(99)

Since $\mu_i$ is a quadratic form, we have

$$L_{ii} \hat{\lambda}_i = \frac{\mu_i \hat{\lambda}_i}{1 + \mu_i \hat{\lambda}_i} \leq 1$$

(100)

Random Least Squares can therefore been seen as a generalized form of ridge regression which shrinks each singular value of $X$ with a specific factor.

In light of the results for the Moore-Penrose inverse, one expects that the random least squares estimator can be used to construct valid confidence intervals if $\mu_i$ is sufficiently large. We therefore need to establish if there exists a choice of $k$ for which this is the case. In order to do so, we use the following identity (Marzetta et al. (2011))

$$E[\Phi'\hat{\Lambda}\Phi]^{-1}\Phi'\hat{\Lambda} = I - W$$

(101)

32
where
\[ W = E[\Psi(\Psi'\hat{\Lambda}^{-1}\Psi)^{-1}\Psi']\hat{\Lambda}^{-1} \]  
with \( \Psi \) is a \( p \times n - k \) matrix with independent standard normal entries. The matrix in (102) is diagonal, as can be seen by verifying that for a unitary diagonal matrix \( \Omega \), we have
\[ \Omega\Psi(\Psi'\hat{\Lambda}^{-1}\Psi)^{-1}\Psi'\hat{\Lambda}^{-1}\Omega^* \overset{(d)}{=} \Psi(\Psi'\hat{\Lambda}^{-1}\Psi)^{-1}\Psi'\hat{\Lambda}^{-1} \]  
using that \( \Omega\Psi(\Psi'\hat{\Lambda}^{-1}\Psi)^{-1}\Psi'\hat{\Lambda}^{-1}\Omega = \Psi \) for any unitary matrix \( \Omega \).

Using again the Sherman-Morrison inverse as in (99) to evaluate the diagonal elements (102), we obtain
\[ [W]_{ii} = \left[ \frac{\hat{\lambda}_i \nu_i}{1 + \frac{1}{\hat{\lambda}_i} \nu_i} \right] \]  
where
\[ \nu_i = \psi_i \left( \Psi_i\hat{\Lambda}^{-1}_i\Psi_i \right)^{-1} \psi_i \]  
Now we can use Jensen’s inequality and the fact that \( x/(1 + x) \) with \( x > 0 \) is a concave function to show that
\[ [W]_{ii} \leq \frac{\frac{1}{\hat{\lambda}_i} E[\nu_i]}{1 + \frac{1}{\hat{\lambda}_i} E[\nu_i]} \leq \frac{\hat{\kappa}_n k^{-1}}{1 + \hat{\kappa} n^{-k-1}} \]  
where \( \hat{\kappa} = \frac{\hat{\lambda}_{\max}}{\hat{\lambda}_{\min}} = \frac{p+n}{p+n} \).

If we require \( [W]_{ii} = a \sqrt{\log p/n} \), then we have
\[ k = \frac{\hat{\kappa}}{\hat{\kappa} + \frac{a \sqrt{\log p/n}}{1 - a \sqrt{\log p/n}}} (n - 1) \]
\[ \approx \left( 1 - \frac{a}{\hat{\kappa}} \sqrt{\frac{\log p}{n}} \right) (n - 1) \]  
showing that we can use \( k \) slightly lower than \( n \).

Under the suggested choice for \( k \), we have
\[ D_{ii}^{RLS} = D_{ii}^{MP} \left( 1 - O \left( \sqrt{\frac{\log p}{n}} \right) \right) \]  
Equation (108) implies that the established bounds on \( D(XX')^{-1}X - I \) remain valid up to a vanishing term of order \( O \left( \sqrt{\frac{\log p}{n}} \right) \).
C.4 Proof of Theorem 4

This follows after some rewriting. Define the diagonal matrix \( A = E[R(R'\hat{\Lambda}R)^{-1}R']\hat{\Lambda}, \) then
\[
\|e_i\hat{U}E[R(R'\hat{\Lambda}R)^{-1}R']X\|_2^2 = e_i\hat{U}A\hat{\Lambda}^{-1}A\hat{U}'e_i \\
= e_i\hat{U}\hat{\Lambda}^{-1/2}A^2\hat{\Lambda}^{-1/2}\hat{U}'e_i
\] (109)

For the generalized inverse we have
\[
\|e_iX'(XX')^{-1}X\|_2^2 = e_iX'(XX')^{-2}Xe_i \\
= e_i\hat{U}\hat{\Lambda}^{-1}\hat{U}'e_i
\] (110)

And since \( A^2 \) is a diagonal matrix with the diagonal elements satisfying \( 0 \leq A^2_{ii} \leq 1, \) the claim in Theorem 4 follows.

C.5 Validity of the ridge adjusted inverse

The proof relies on the results in Wang and Leng (2015), with some necessary adjustments.

First note that \((XX + \gamma I_p)^{-1}X'X = X'(XX + \gamma I_n)^{-1}X.\) Then we analyze first the bias term
\[
\Delta = \frac{1}{\sqrt{n}}D^{RI}(X'X + \gamma I_p)^{-1}X'X - I \\
= \frac{1}{\sqrt{n}}D^{RI}X'(XX + \gamma I_n)^{-1}X - I
\] (111)

Substituting \(X = VSU'\Sigma^{1/2} \) and using a von Neumann expansion of the inverse, Wang and Leng (2015) obtain
\[
X'(XX' + \gamma I_n)X = HH' + E
\] (112)

where \(HH' = X'(XX')^{-1}X\) and \(E\) is a matrix that satisfies
\[
P \left( \max(E) > \frac{C \gamma}{p - C \gamma} \right) \leq e^{-\tilde{C}n}
\] (113)

with \(C, \tilde{C} > 0.\) Then, choosing \(\gamma\) small enough pushes \(E\) towards zero. If
\[
\frac{C \gamma}{p - C \gamma} = O \left( \lambda_{\max}(HH') \sqrt{\frac{\log p}{n}} \right)
\] (114)
then by the same arguments as for random least squares, the results obtained using the Moore-Penrose pseudo-inverse carry over. Now $\lambda_{\text{max}}(HH') = O\left(\frac{n^2}{p}\right)$, so we should choose $\gamma$ such that

$$\frac{C\gamma}{p - C\gamma} = O\left(\frac{n}{p} \sqrt{\frac{\log p}{n}}\right)$$

(115)

Noting that $p \gg C\gamma$, gives

$$\gamma = O\left(n \sqrt{\frac{\log p}{n}}\right)$$

(116)

Under this choice the variance of the estimator also remains finite.

## D Simulation Results Without Correction

Table 4: Monte Carlo simulation: Equicorrelated Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>s=3</th>
<th></th>
<th></th>
<th>s=15</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coef.</td>
<td>SE</td>
<td>CR</td>
<td>power</td>
<td>coef.</td>
<td>SE</td>
</tr>
<tr>
<td>Pinv</td>
<td>b=2</td>
<td>0.203</td>
<td>0.303</td>
<td>0.935</td>
<td>0.115</td>
<td>0.215</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.003</td>
<td>0.303</td>
<td>0.939</td>
<td>0.016</td>
<td>0.215</td>
</tr>
<tr>
<td>RLS</td>
<td>b=2</td>
<td>0.204</td>
<td>0.282</td>
<td>0.937</td>
<td>0.123</td>
<td>0.218</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.004</td>
<td>0.282</td>
<td>0.938</td>
<td>0.019</td>
<td>0.218</td>
</tr>
<tr>
<td>Ridge</td>
<td>b=2</td>
<td>0.203</td>
<td>0.291</td>
<td>0.937</td>
<td>0.120</td>
<td>0.216</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.003</td>
<td>0.290</td>
<td>0.938</td>
<td>0.017</td>
<td>0.216</td>
</tr>
<tr>
<td>Pinv</td>
<td>b=5</td>
<td>0.508</td>
<td>0.305</td>
<td>0.934</td>
<td>0.403</td>
<td>0.537</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.008</td>
<td>0.304</td>
<td>0.935</td>
<td>0.039</td>
<td>0.306</td>
</tr>
<tr>
<td>RLS</td>
<td>b=5</td>
<td>0.509</td>
<td>0.284</td>
<td>0.936</td>
<td>0.453</td>
<td>0.544</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.009</td>
<td>0.284</td>
<td>0.934</td>
<td>0.047</td>
<td>0.285</td>
</tr>
<tr>
<td>Ridge</td>
<td>b=5</td>
<td>0.509</td>
<td>0.292</td>
<td>0.933</td>
<td>0.434</td>
<td>0.540</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.009</td>
<td>0.292</td>
<td>0.935</td>
<td>0.043</td>
<td>0.294</td>
</tr>
</tbody>
</table>

Note: this table reports the results for different Monte Carlo experiments without bias correction where the regressors have a equicorrelated covariance as specified in (37). For additional information, see the note following Table 1.
Table 5: Monte Carlo simulation: Toeplitz Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>s=3</th>
<th>s=15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coef.</td>
<td>SE</td>
</tr>
<tr>
<td>Pinv</td>
<td>b=2</td>
<td>0.196</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.003</td>
</tr>
<tr>
<td>RLS</td>
<td>b=2</td>
<td>0.197</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.004</td>
</tr>
<tr>
<td>Ridge</td>
<td>b=2</td>
<td>0.197</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>b=5</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.008</td>
</tr>
<tr>
<td>RLS</td>
<td>b=5</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.009</td>
</tr>
<tr>
<td>Ridge</td>
<td>b=5</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>b=0</td>
<td>0.009</td>
</tr>
</tbody>
</table>

Note: this table reports the results for different Monte Carlo experiments without bias correction where the regressors have a toeplitz covariance as specified in (38). For additional information, see the note following Table 1.

E Variable Descriptions

Table 6: Variable Descriptions Table 3

<table>
<thead>
<tr>
<th>FRED mnemonic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPFINAL</td>
<td>Industrial Production: Final Products (Market Group) (Index 2012=100)</td>
</tr>
<tr>
<td>USWTRADE</td>
<td>All Employees: Wholesale Trade (Thousands of Persons)</td>
</tr>
<tr>
<td>HOABS</td>
<td>Business Sector: Hours of All Persons (Index 2009=100)</td>
</tr>
<tr>
<td>HOANBS</td>
<td>Nonfarm Business Sector: Hours of All Persons (Index 2009=100)</td>
</tr>
<tr>
<td>NAPMII</td>
<td>ISM Manufacturing: Inventories Index</td>
</tr>
<tr>
<td>OPHNFB</td>
<td>Nonfarm Business Sector: Real Output Per Hour of All Persons (Index 2009=100)</td>
</tr>
<tr>
<td>OPHPFB</td>
<td>Business Sector: Real Output Per Hour of All Persons (Index 2009=100)</td>
</tr>
<tr>
<td>CPF3MTB3MnX</td>
<td>3-Month Commercial Paper Minus 3-Month Treasury Bill, secondary market (Percent)</td>
</tr>
<tr>
<td>MED3TB3MnX</td>
<td>3-Month Eurodollar Deposit Minus 3-Month Treasury Bill, secondary market (Percent)</td>
</tr>
<tr>
<td>AMBSLREALx</td>
<td>St. Louis Adjusted Monetary Base (Billions of 1982-84 Dollars), dated by CPI</td>
</tr>
<tr>
<td>UMCSNENTx</td>
<td>University of Michigan: Consumer Sentiment (Index 1st Quarter 1966=100)</td>
</tr>
<tr>
<td>VXOCLSX</td>
<td>CBOE S&amp;P 100 Volatility Index: VXO</td>
</tr>
<tr>
<td>GFDEGDQ1888</td>
<td>Federal Debt: Total Public Debt as Percent of GDP (Percent)</td>
</tr>
</tbody>
</table>

Note: this table reports the variable descriptions corresponding to the fred mnemonics in Table 3.