TI 2017-032/III
Tinbergen Institute Discussion Paper

Confidence Intervals in High-Dimensional Regression Based on Regularized Pseudoinverses

Tom Boot[1]
Didier Nibbering[2]

[1] *University of Groningen, The Netherlands*
[2] *Erasmus University Rotterdam, and Tinbergen Institute, The Netherlands*

Confidence intervals in high-dimensional regression based on regularized pseudoinverses

Tom Boot*
University of Groningen

Didier Nibbering†
Erasmus University Rotterdam
Tinbergen Institute

March 13, 2017

Abstract

In modern 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 by approximating the inverse covariance matrix by a scaled Moore-Penrose pseudoinverse, and using 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 random 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 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 economic indicators has led to a challenging situation where the number of explanatory variables approaches, and often even 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). Even more extreme examples can be found in recent studies on the relation between the human genome and life outcomes such as educational attainment by Rietveld et al. (2013). Standard inference techniques are hampered by the resulting rank deficiency of the empirical covariance matrix. New methods are therefore needed to handle this type of high-dimensional data.

This paper deals with the estimation of coefficients and the construction of corresponding confidence intervals in high-dimensional linear regression models, where the number of unknown coefficients increases almost exponentially in 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 pseudoinverse or a regularized variant, and (2) a bias correction step is implemented based on the lasso.

We first show how valid confidence intervals can be constructed when using the Moore-Penrose inverse, which has the benefit of being tuning parameter free. We then 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 repeatedly projected 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, which we refer to as random least squares, is equivalent to a type of generalized ridge regularization, and yields valid confidence intervals when the projection dimension is chosen sufficiently close to the sample size. In addition, we show that the standard ridge regularized inverse covariance matrix with an appropriately chosen penalty parameter, also provides valid confidence intervals.

There are several existing estimators that allow for estimation in high-dimensional 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 has proven to be challenging.

Currently, there are broadly three different approaches to construct confidence intervals in a high-dimensional linear regression model. The first 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 et al., 2013, 2010; 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 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 remaining 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, in combination with an accurate 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). The bias correction is most commonly based on the lasso estimator, but can also be replaced with a suitable alternative as proposed by Caner and Kock (2014). Both estimators are computationally intensive, requiring a numerical optimization procedure for each of the available regressors. Especially in the large-scale applications that they are intended for this might be prohibitively slow.

A third approach, recently developed by Lan et al. (2016), estimates confidence intervals for each coefficient separately. Confouding effects are reduced by identifying a set of regressors which is correlated to the regressor of interest using partial correlations. Subsequently, the regressor of interest is orthogonalized with respect to this set. Under assumptions required to accurately identify the set of correlated regressors, Lan et al. (2016) show that this approach yields correct inference. These assumptions limit the approach to high-dimensional settings where only a small set of regressors is correlated to the regressor of interest.

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 yields an estimator that is approximately normally distributed. The corresponding confidence intervals of this estimator are valid under conditions on the regressor matrix which are typical in high-dimensional analysis (Fan and Lv, 2008), and a sparsity assumption, which bounds 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 is suitably small. In combination with an accurate initial estimator, such as the lasso, the bias vanishes compared to the noise. This approach substantially simplifies previous 'desparsification' methods developed by Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014), is tuning parameter free, and alleviates strong computational constraints associated with previous methods.

Second, we show that regularization by random least squares can improve upon the scaled Moore-Penrose estimator in terms of statistical power. We show that for a projection dimension close to the number of observations, random least squares approximates the scaled Moore-Penrose estimator and yields valid confidence intervals. The width of the confidence intervals of random least squares can then be shown to be equal to or smaller than 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 adjusted 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. We extend these results to allow for the construction of valid confidence intervals.

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, coefficients that are zero in the data generating process, are estimated very close to zero. Random least squares and ridge regression incur a slight downward bias compared to using a Moore-Penrose pseudoinverse, but yield narrower confidence intervals. For all proposed methods, coverage rates are much closer to the nominal coverage rate compared with existing alternatives by van de Geer et al. (2014), Javanmard and Montanari (2014) and Lan et al. (2016).

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 greatly 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 Section 3 introduces our methods. The theoretical properties of the Moore-Penrose pseudoinverse, random least squares, and ridge regression are derived and presented in Section 4. Section 5 illustrates these results through Monte Carlo simulations and Section 6 applies the methods on the FRED-QD dataset. Section 7 concludes.

Notation We use the following notation throughout the paper: For any $n \times 1$ vector $a = (a_1, \ldots, a_n)'$, the l_q-norm is defined as $||a||_q := (\sum_{i=1}^n |a_i|^q)^{1/q}$ for $q > 0$ and $||a||_0$ denotes the number of nonzero elements of a. The maximum norm is written as $||a||_\infty = \max(|a_1|, \ldots, |a_n|)$. For a $p \times n$ matrix A, the l_q-norm is defined as $||A||_q := \sup_x \{||Ax||_q, ||x||_q = 1\}$ and the maximum norm is written as $||A||_{\max} = \max_{i=1,\ldots,n, j=1,\ldots,p} |A_{ij}|$. The $n \times n$ identity matrix is denoted by I_n. For the regressor matrix X, we index the rows with the subscript $i = 1, \ldots, n$ and the columns with the subscript $j = 1, \ldots, p$. If U is a $p \times p$ orthogonal matrix, we write $U \in \mathcal{O}(p)$. When two random variables X and Y have the same distribution, we denote $X \overset{(d)}{=} Y$.

2 High-dimensional regression

Consider the data generating process

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \tag{1}$$

where y is an $n \times 1$ response vector, X an $n \times p$ regressor matrix, $\beta = (\beta_1, \ldots, \beta_p)'$ a $p \times 1$ vector of unknown regressor coefficients, and ε an $n \times 1$ vector of errors

which are independent and identically normally distributed with variance σ^2. The empirical covariance matrix is denoted by $\hat{\Sigma} = \frac{1}{n}X'X$. The normality of the errors can be relaxed to independent and identically distributed in the limit where $n \to \infty$.

Approximate inverse and bias correction Define $\frac{1}{n}M$ as a $p \times n$ matrix for which $\frac{1}{n}MX$ is close to the $p \times p$ identity matrix, in a sense that will be made precise below. We refer to $\frac{1}{n}M$ as an approximate inverse for X. Consider estimators for β of the form

$$\begin{aligned}\hat{\beta} &= \frac{1}{n}My \\ &= \frac{1}{n}MX\beta + \frac{1}{n}M\varepsilon \\ &= \beta + \left(\frac{1}{n}MX - I_p\right)\beta + \frac{1}{n}M\varepsilon.\end{aligned} \tag{2}$$

The second term of (2) represents a bias. When $p < n$, ordinary least squares yields unbiased estimates by choosing $M = \hat{\Sigma}^{-1}X'$. When $p > n$, $\hat{\Sigma}$ is singular. This forces one to consider approximate inverses to construct M. In this case, the bias will not be identically zero, but will depend on the accuracy of the approximate inverse.

Suppose we have an accurate initial estimator $\hat{\beta}^{\text{init}}$, then we can reduce the bias in (2) by applying the correction

$$\begin{aligned}\hat{\beta}^c &= \frac{1}{n}My - \left(\frac{1}{n}MX - I_p\right)\hat{\beta}^{\text{init}} \\ &= \beta + \left(\frac{1}{n}MX - I_p\right)\left(\beta - \hat{\beta}^{\text{init}}\right) + \frac{1}{n}M\varepsilon.\end{aligned} \tag{3}$$

The goal of this paper is to introduce choices of M and $\hat{\beta}_{\text{init}}$ such that the remaining bias is of lower order than the variance.

Required properties We can rescale and rewrite the estimator in (3) as

$$\sqrt{n}\left(\hat{\beta}^c - \beta\right) = Z + \Delta, \tag{4}$$

where

$$\begin{aligned}\Delta &= \sqrt{n}\left(\frac{1}{n}MX - I_p\right)\left(\beta - \hat{\beta}^{\text{init}}\right) \\ Z &= \frac{1}{\sqrt{n}}M\varepsilon.\end{aligned} \tag{5}$$

We propose specifications for the approximate inverse M, for which we show that $Z = O_p(1)$. This implies that in order to obtain an approximately unbiased estimator, the bias term should satisfy $||\Delta||_\infty = o_p(1)$.

The infinity norm $||\Delta||_\infty$ can be bounded by

$$||\Delta||_\infty \leq \sqrt{n}\left|\left|n^{-1}MX - I_p\right|\right|_{\max} ||\beta - \hat{\beta}^{\text{init}}||_1. \tag{6}$$

Using an accurate estimator such as the lasso, Tibshirani (1996), one can achieve a bound on the l_1 norm of $||\beta - \hat{\beta}^{\text{init}}||_1 = O_p\left(s_0\sqrt{n^{-1}\log p}\right)$. Here $s_0 = ||\beta||_0$, the number of non-zero coefficients. As van de Geer et al. (2014) and Javanmard and Montanari (2014), we take $s_0 = o(\sqrt{n}/\log p)$. Then, the approximate inverse M is sufficiently accurate when

$$\left|\left|\frac{1}{n}MX - I_p\right|\right|_{\max} = O\left(\sqrt{\frac{\log p}{n}}\right). \tag{7}$$

Confidence intervals In addition to choosing M such that (7) is satisfied, we show that under mild conditions, $\frac{1}{\sqrt{n}}M\varepsilon \sim N(0, \sigma^2 MM')$. Then, if a consistent estimator of the noise level σ^2 is available, $(1-\alpha)\cdot 100\%$ confidence intervals can be constructed as

$$\left[\hat{\beta}_j^c - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2 m_j' m_j}{n}}, \hat{\beta}_j^c + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2 m_j' m_j}{n}}\right], \tag{8}$$

where $z_{\alpha/2}$ is the $\alpha/2$ critical value for the standard normal distribution, $\hat{\sigma}^2$ is a consistent estimator of σ^2, and m_j the j-th row of M with $j = 1, \ldots, p$.

3 Methods

This section outlines how to construct confidence intervals in a high-dimensional linear regression model, where the number of variables greatly exceeds the number of observations. Results on the theoretical validity of these methods are deferred to Section 4.

All methods considered in this section are based on an approximate inverse for the singular empirical covariance matrix in combination with a bias correction based on a suitable initial estimator. Section 3.1 proposes for the approximate inverse the Moore-Penrose pseudoinverse (MPI). This approach has the benefit of being completely tuning parameter-free. To reduce potential noise in the Moore-Penrose pseudoinverse, we propose a novel regularization method called random least squares (RLS), and a ridge regularized inverse (RID). Throughout, we take the lasso estimator as the initial estimator used for bias correction. Section 3.2 discusses how the noise level in the confidence intervals can be estimated.

3.1 Approximate inverse construction

Equation (7) requires that each element of $\frac{1}{n}MX - I_p$ has to be sufficiently small with high probability. As a first step, we ensure that the diagonal terms are exactly equal to zero by introducing a $p \times p$ diagonal matrix D, with diagonal elements d_j, and taking

$$\frac{1}{n}M = \frac{1}{n}D\tilde{M}, \qquad d_j = n(\tilde{m}_j' x_j)^{-1}, \tag{9}$$

with $\tilde{m}_j'$ the j-th row of $\tilde{M}$. It then remains to be shown that the off-diagonal elements are of order $O\left(\sqrt{\frac{\log p}{n}}\right)$ with high probability. Section 4 shows that this is indeed the case.

We consider three ways of constructing M that are all of the form defined in (9), with $\tilde{M}$ chosen as: (a) the Moore-Penrose pseudoinverse of X, (b) the random least squares inverse covariance matrix estimator, or (c) the inverse of the ridge adjusted empirical covariance matrix.

3.1.1 Moore-Penrose pseudoinverse

An attractive, tuning parameter free choice for $\tilde{M}$ is the Moore-Penrose pseudoinverse. When $p < n$, and the columns of X are independent, then $\tilde{M} = (X'X)^{-1}X'$. 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 pseudoinverse equals $X'(XX')^{-1}$. The estimator we consider is therefore

$$\frac{1}{n}M^{\text{MPI}} = \frac{1}{n}D^{\text{MPI}}X'(XX')^{-1}, \tag{10}$$

where the diagonal elements d_j^{MPI} of the diagonal scaling matrix D^{MPI} equal

$$d_j^{\text{MPI}} = n\left[x_j'(XX')^{-1}x_j\right]^{-1}. \tag{11}$$

The Moore-Penrose pseudoinverse relies on the inverse of XX'. Almost ironically, the accuracy of this estimator deteriorates when p approaches n from above.

3.1.2 Random Least Squares

As a novel alternative to the possibly noisy pseudoinverse estimator, we propose the random least squares (RLS) estimator. This method is based on projecting the high-dimensional regressor matrix X onto a $k < n$ dimensional subspace by post-multiplying with a $p \times k$ matrix R with independently standard normally distributed elements,

$$R_{jl} \sim N(0,1), \quad j = 1,\dots p, \quad l = 1,\dots,k \tag{12}$$

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

$$y = XR\gamma_R + u. \tag{13}$$

Least squares estimation of γ_R yields

$$\hat{\gamma}_R = (R'X'XR)^{-1}R'X'y, \tag{14}$$

which is related to the estimator of β 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 β,

$$\hat{\beta} = \mathrm{E}_R[R\hat{\gamma}_R] = \mathrm{E}_R[R(R'X'XR)^{-1}R']X'y. \tag{15}$$

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 (15) follows that random least squares uses low-dimensional projections to obtain an approximate inverse covariance matrix. We therefore take $\tilde{M} = \mathrm{E}_R[R(R'X'XR)^{-1}R']X'$ in (10). After scaling we arrive at

$$\frac{1}{n}M^{\mathrm{RLS}} = \frac{1}{n}D^{\mathrm{RLS}}\mathrm{E}_R\left[R(R'X'XR)^{-1}R'\right]X', \tag{16}$$

with

$$d_j^{\mathrm{RLS}} = n\left\{\mathrm{E}_R[r_j'(R'X'XR)^{-1}R']X'x_j\right\}^{-1}. \tag{17}$$

3.1.3 Ridge regression

An alternative regularization method is to implement a ridge adjustment which overcomes the rank deficiency of the empirical covariance matrix by adding a scalar multiple of the identity matrix

$$\frac{1}{n}M^{\mathrm{RID}} = \frac{1}{n}D^{\mathrm{RID}}(X'X + \gamma I_p)^{-1}X', \tag{18}$$

where γ denotes the ridge penalty and the elements of the diagonal scaling matrix D^{RID} equal

$$d_j^{\mathrm{RID}} = n\left(v_j'X'x_j\right)^{-1}, \tag{19}$$

with v_j the j-th row of $(X'X + \gamma I_p)^{-1}$.

The regularization in (18) can be related to the Moore-Penrose pseudoinverse, since we can write

$$X'(XX')^{-1} = \lim_{\gamma \to 0}\left(X'X + \gamma I_p\right)^{-1}X'. \tag{20}$$

3.2 Estimation of the noise level

A consistent estimator of the noise level σ^2 is crucial to construct valid confidence intervals. Existing methods, such as van de Geer et al. (2014) and Javanmard and Montanari (2014) rely on the scaled lasso developed by Sun and Zhang (2012). For this estimator, under a sparsity constraint $s_0\frac{\log p}{n} = o(1)$ and a compatibility condition, it holds that $\left|\frac{\hat{\sigma}}{\sigma} - 1\right| = o_P(1)$. By Assumption 3 and Assumption 2 discussed in Section 4.1, these conditions are indeed satisfied.

In the Monte Carlo simulations in Section 5, and in line with findings by Reid et al. (2016), we find the scaled lasso to be unreliable in many settings. An alternative is to use

$$\hat{\sigma}^2_{\mathrm{lasso}} = \frac{1}{n - \hat{s}}\hat{\varepsilon}'\hat{\varepsilon}, \tag{21}$$

with $\hat{s}$ the number of non-zero coefficients retained by the lasso, and $\hat{\varepsilon}$ the $n \times 1$ vector of lasso regression errors. Corresponding to the results in Reid et al. (2016), we find that this leads to more robust estimation of the noise level.

4 Theoretical results

This section provides the main results of the paper. Proofs for the theorems in this section are given in Appendix B.

First, Section 4.1 provides the necessary assumptions on the regressor matrix and the coefficient vector. Section 4.2 uses these assumptions to present the main result of the paper. Theorem 1 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 unbiased and normally distributed. It follows that confidence intervals of coefficient estimates of these estimators can be constructed by standard procedures.

Section 4.3 shows how we arrive at this result and Section 4.4 shows 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. Section 4.5 generalizes the results for high-dimensional regression models with normally distributed errors to models with independent and identically distributed errors, and Section 4.6 concludes with a discussion of the computational complexity of the proposed methods.

4.1 Assumptions

By allowing for the number of variables to exceed the number of observations, it is necessary to make some assumptions in addition to standard assumptions on the linear regression model. These assumptions are standard in high-dimensional regression problems, see for example Fan and Lv (2008) and Wang and Leng (2015).

The first assumptions provides an upper bound on the number of variables relative to the number of observations.

Assumption 1 *The number of variables grows near exponentially with the number of observations, i.e.*

$$\frac{\log p}{n} = o(1). \tag{22}$$

Next, we restrict the number of non-zero coefficients $||\beta||_0 = s_0$. For lasso consistency, one usually takes $s_0^2 = o\left(n/\log p\right)$. 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* $s_0 = o\left(\frac{\sqrt{n}}{\log p}\right)$.

The following assumption on the regressors are imposed

Assumption 3 *The regressor matrix* X *can be written as*

$$X = Z\Sigma^{1/2} = VSU'\Sigma^{1/2}, \tag{23}$$

where the rows of the $n \times p$ *matrix* Z *are generated independently from a spherically symmetric distribution,* V *is an* $n \times n$ *orthogonal matrix, i.e.* $V \in \mathcal{O}(n)$,

S is an $n \times p$ matrix of singular values, and $U \in \mathcal{O}(p)$. The eigenvalues of ZZ' can be bounded with high probability

$$P\left(\frac{1}{p}ZZ' > c_Z \quad \cup \quad \frac{1}{p}ZZ' < \frac{1}{c_Z}\right) \leq e^{-C_Z n}. \tag{24}$$

with c_Z, C_Z positive constants.

Spherical symmetry for example holds when the rows of X follow a multivariate normal distribution, but also allows for heavier tailed distributions such as the power exponential distribution and multivariate t distributions (Dasgupta et al., 2012; Serfling, 2006).

Finally, the eigenvalues of the population covariance matrix are assumed to be finite and independent of the dimensions n and p.

Assumption 4 *The condition number of the population covariance matrix Σ is bounded by a constant, i.e.*

$$\kappa = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \leq c_\kappa < \infty. \tag{25}$$

This assumption can be relaxed at the expense of a slower convergence rate.

4.2 Approximate unbiasedness and normality

This section presents the main result of the paper; the unbiasedness and normality of the estimators discussed in Section 3. For each method, define the diagonal scaling matrices by the following diagonal elements

$$\begin{aligned} d_j^{\text{MPI}} &= n\left[x_j'(XX')^{-1}x_j\right]^{-1}, \\ d_j^{\text{RLS}} &= n\left\{\mathrm{E}_R\left[r_j'(R'X'XR)^{-1}R'\right]X'x_j\right\}^{-1}, \\ d_j^{\text{RID}} &= n\left(v_j'X'x_j\right)^{-1}, \end{aligned} \tag{26}$$

where $j = \{1, \ldots, p\}$, R is a $p \times k$ matrix of independent standard normal entries, and v_j the j-th row of $(X'X + \gamma I_p)^{-1}$.

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

$$\begin{aligned} k = k^* &\geq \left(1 - c_k\sqrt{\frac{\log p}{n}}\right)(n-1), \\ \gamma = \gamma^* &\leq c_\gamma n\sqrt{\frac{\log p}{n}}, \end{aligned} \tag{27}$$

where c_k and c_γ are positive constants defined in Theorem 4 and 5, respectively. Now the following theorem holds.

Theorem 1 *Suppose Assumption 2 and Assumption 3 hold, and an initial estimator $\hat{\beta}^{init}$ is available which satisfies the l_1 bound $||\hat{\beta}^{init} - \beta||_1^2 = O_p\left(s_0^2 \log(p)/n\right)$. Let $\hat{\beta}^c = \frac{1}{n}My - \left(\frac{1}{n}MX - I_p\right)\hat{\beta}^{init}$, and take M as one of the following*

$$\begin{aligned} M^{MPI} &= D^{MPI}X'(XX')^{-1}, \\ M^{RLS} &= D^{RLS}E_R\left[R(R'X'XR)^{-1}R'\right]X', \\ M^{RID} &= D^{RID}(X'X + \gamma^* I_p)^{-1}X', \end{aligned} \tag{28}$$

where the elements of the diagonal matrices D are defined in (26)*, R is a $p \times k^*$ matrix with independent standard normal entries, and k^* and γ^* as in* (27)*. Then,*

$$\begin{aligned}\sqrt{n}(\hat{\beta}^c - \beta) &= Z + \Delta \\ Z = \frac{1}{\sqrt{n}} M\varepsilon &\sim N\left(0, \sigma^2 MM'\right) \\ \Delta &= o_p(1).\end{aligned}$$

From Theorem 1 it follows that we can use the classical procedure to construct confidence intervals, given in (8). The following sections provide details on how we arrive at Theorem 1 and give explicit probability bounds for Δ under different choices of M and D.

4.3 Accuracy of the initial estimator and approximate inverses

Initial estimator For the initial estimator we take the standard lasso estimator. In Appendix B.1, we prove that under Assumption 3 the following compatibility condition (Bühlmann and Van De Geer, 2011) holds: when S_0 is the true set of $s_0 = ||S_0||_0$ non-zero coefficients, then the compatibility condition is satisfied for this set if

$$||\beta_{S_0}||_1 \leq \frac{\sqrt{s_0}||X\beta||_2}{\sqrt{n}\phi_0}, \tag{29}$$

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

The lasso estimator subsequently satisfies the following l_1 accuracy bound (Bühlmann and Van De Geer (2011), corollary 6.2):

Lemma 1 *Suppose the compatibility condition holds for S_0, and the penalty parameter is chosen as $\lambda \geq 4\sigma\sqrt{\frac{t^2+2\log p}{n}}$, then with probability exceeding $1-\alpha$, where*

$$\alpha = 2\exp(-t^2/2) + P(\hat{\sigma} < \sigma), \tag{30}$$

we have

$$||\beta - \hat{\beta}^{lasso}||_1 = O_p\left(s_0\sqrt{\frac{\log p}{n}}\right). \tag{31}$$

The lasso estimator therefore satisfies the l_1 bound required in Theorem 1.

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

Theorem 2 *Define d_j^{MPI} as in* (26)*, then we have*

$$P\left(\left|\frac{1}{n} d_j^{MPI} x_j'(XX')^{-1}x_l - \delta_{jl}\right| > a\sqrt{\frac{\log p}{n}}\right) = O(p^{-\tilde{c}}), \tag{32}$$

with $\delta_{jl} = 1$ if $j = l$ and 0 otherwise, $\tilde{c} = a^2\frac{c_p - 1}{c_p/c_\epsilon - \kappa} - 2$ where $a > 0$, $c_\epsilon = \frac{1+\epsilon}{1-\epsilon}$ with $\epsilon > 0$, $\kappa = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$, and $c_p < \frac{p}{n}$.

A proof is provided in Appendix B.2. The theorem shows that the Moore-Penrose pseudoinverse yields a matrix with lower order off-diagonal terms, which makes it a suitable approximate inverse.

In addition, we need to show that the bias vanishes compared to the variance of the estimator. The following theorem guarantees that this is indeed the case.

Theorem 3 *For $i = 1, \ldots, p$ we have*

$$\frac{1}{\sqrt{n}} d_j^{MPI} x_j'(XX')^{-1}\varepsilon = O_p(1). \tag{33}$$

A proof is presented in Appendix B.3.

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 $\hat{\kappa} = \frac{\lambda_{\max}(\hat{\Lambda})}{\lambda_{\min\neq 0}(\hat{\Lambda})}$. Let*

$$k = \left(1 - c_\kappa \sqrt{\frac{\log p}{n}}\right)(n-1), \tag{34}$$

where $c_\kappa = \frac{c_V}{c_Z^2 c_\epsilon^2 \kappa}$ with $0 \leq c_V \leq \frac{a}{c_\zeta + c_\epsilon^2 \kappa^2}$ and $c_\zeta > a\sqrt{\frac{\log p}{n}}$, c_Z from Assumption 3, and $c_\epsilon = \frac{1+\epsilon}{1-\epsilon}$ for $\epsilon > 0$. then

$$P\left(\left|\frac{1}{n} d_j^{RLS} E[r_j'(R'X'XR)^{-1}R']X'x_l - \delta_{jl}\right| > \tilde{a}\sqrt{\frac{\log p}{n}}\right) = O\left(p^{-\tilde{c}}\right), \tag{35}$$

where $\tilde{a} = a - c_V(c_\zeta + c_\epsilon^2\kappa^2)$, with c_V, c_ζ and c_ϵ constants independent of n and p, and $\tilde{c}$ as in Theorem 2 with a replaced by $\tilde{a}$. Furthermore,

$$\frac{1}{\sqrt{n}} d_j^{RLS} E[r_j(R'X'XR)^{-1}R']X'\varepsilon = O_p(1). \tag{36}$$

In the proof of Theorem 4 given in Appendix B.4, we show that when k is sufficiently close to n the regularized inverse approximates the Moore-Penrose inverse. The results from Section 4.3 can then be used to show that the bias of the random least squares estimator remains small. Theorem 6 shows that regularization can be used to increase the power over using the Moore-Penrose pseudoinverse. Moreover, the proof of Theorem 4 also shows that random least squares is equivalent to a generalized form of ridge regression, where the penalty parameter is allowed to be different for each coefficient.

Ridge regularization Because of the relation between the generalized inverse and ridge regularized covariance matrices displayed in (20), 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 when setting the penalty parameter as defined in Theorem 1.

Theorem 5 *Define* $\gamma \leq c_\gamma n\sqrt{\frac{\log p}{n}}$, *where* $c_\gamma = c_V c_Z c_\epsilon \lambda_{\min}(\Sigma)$ *a positive constant, then*

$$P\left(\left|\frac{1}{n}d_j^{RID}(X'X+\gamma I_p)^{-1}X'X-\delta_{jl}\right| > \tilde{a}\sqrt{\frac{\log p}{n}}\right) = O\left(p^{-\tilde{c}}\right), \tag{37}$$

with $\tilde{a}$ *and* $\tilde{c}$ *as in Theorem 4. Furthermore,*

$$\frac{1}{\sqrt{n}}d_j^{RID}(X'X+\gamma I_p)^{-1}X'\varepsilon = O_p(1). \tag{38}$$

A proof is provided in Appendix B.5.

4.4 Power increase by regularization

One would expect that regularization by random least squares or ridge, yields smaller confidence intervals compared to using the Moore-Penrose pseudoinverse. This is indeed the case if the estimators all use the same diagonal scaling matrix

Theorem 6 *For the choice of* k *as in Theorem 4, or* γ *as in Theorem 5, we have for* $d_i = d_i^{MPI}$

$$\frac{d_j}{\sqrt{n}}\left|\left|E\left[r_j'(R'X'XR)^{-1}R'\right]X'\right|\right|_2 - \frac{d_j}{\sqrt{n}}\left|\left|x_j'(XX')^{-1}\right|\right|_2 \leq 0, \tag{39}$$

and

$$\frac{d_j}{\sqrt{n}}\left|\left|v_jX'\right|\right|_2 - \frac{d_j}{\sqrt{n}}\left|\left|x_j'(XX')^{-1}\right|\right|_2 \leq 0, \tag{40}$$

where v_j *is the* j*-th row of* $(X'X+\gamma I_p)^{-1}$

The proof is given in Appendix B.6.

Note that using D^{MPI} for the Moore-Penrose inverse, D^{RLS} for the random least squares estimator, and D^{RID} for the ridge regularized inverse, yields valid confidence intervals, but no ordering in terms of power. However, in most cases we have encountered, the inequality in Theorem 6 is satisfied when using the diagonal matrix specific to the estimator under consideration. This is also evident from the Monte Carlo results in Section 5.

4.5 Non-gaussian errors

The result of the previous paragraphs can be generalized to the case where the errors are independent and identically distributed.

Lemma 2 *Let the errors* $\varepsilon_i \sim i.i.d(0,\sigma^2)$, *then as* $n \to \infty$, *Theorem 1 holds.*

A proof is provided in Appendix B.7 and follows from the application of a central limit theorem.

Table 1: Computational complexity comparison

Method	Complexity	Method	Complexity
MPI	$n \cdot p^2$	van de Geer et al. (2014)	p^4
RLS	$n \cdot p^2$	Javanmard and Montanari (2014)	p^4
RID	$n \cdot p^2$	Lan et al. (2016)	$n \cdot p^2$

4.6 Computational complexity

This section shows that the computational complexity of the proposed methods is several orders of magnitude in the number of variables lower compared to existing methods. This greatly extends the applicability of the procedure.

All methods, except Lan et al. (2016), use an initial estimator, which is provided by the lasso. With $r = \min\{n, p\}$, this costs $O(r^3)$ at a fixed value of the penalization parameter. For the initial estimator we apply K-fold cross-validation, and $r = n$, such that the computational complexity of all methods is at least $O(K \cdot n^3)$. However, for none of the methods this is the leading order term.

(Ridge regularized) Moore-Penrose pseudoinverse For the Moore-Penrose pseudoinverse, the leading order term in the computation complexity is the matrix multiplication $X(XX')^{-1}X'$, which has a computational complexity of $O(n \cdot p^2)$. The same holds for the ridge regularized version.

Random least squares estimator For the random least squares estimator, evaluation of the expectation operator in (15) is made efficient by using the singular value decomposition of $X = \hat{V}\hat{S}_n\hat{U}_n'$. Fast SVD algorithms require $O(n^2 \cdot p)$. Define $\hat{\Lambda} = \hat{S}_n^2$ and the $n \times k$ matrix $W_n = \hat{\Lambda}^{\frac{1}{2}}R_n$, where R_n consists of the first n rows of the $p \times k$ matrix R. Then Appendix B.4 shows that

$$\mathrm{E}[R(R'X'XR)^{-1}R']X'X = \hat{U}_n\mathrm{E}[W_n(W_n'W_n)^{-1}W_n']\hat{U}_n'. \tag{41}$$

The Ahlswede and Winter (2002) theorem can be used to show that $n\log(n)$ draws of $W_n(W_n'W_n)^{-1}W_n'$ accurately approximate the expectation by a finite sum. The computational complexity required for the product is $O(n^3)$, so that we end up with $O(n^4 \cdot \log(n))$ to calculate the expectation. The resulting multiplication $\hat{U}_n\mathrm{E}[W_n(W_n'W_n)^{-1}W_n']\hat{U}_n$ is however of order $O(n \cdot p^2)$, which is the leading term when $p \gg n$.

Existing alternatives van de Geer et al. (2014) solve a penalized regression problem with $p - 1$ explanatory variables for every column of the empirical covariance matrix. A fast solver is provided by the lars algorithm of Efron et al. (2004), which has a complexity of $O(p^3)$ for each column. Therefore, the complexity for obtaining the approximate inverse covariance matrix equals $O(p^4)$ in total. Javanmard and Montanari (2014) state that this is equivalent to the complexity of their method.

The method by Lan et al. (2016) requires for each regressor a matrix multiplication with the remaining regressors. The complexity of this multiplication is $O(n \cdot p)$. Potentially this multiplication has to be repeated until a set size is

found for which the stopping rule defined by Lan et al. (2016) is satisfied. This potentially adds another factor n to the complexity, but we will omit it here. The above complexity is required to calculate a single coefficient, such that the total complexity is $O(n \cdot p^2)$.

The complexity of the different methods is summarized in Table 1. We conclude that the methods proposed in this paper are two orders of magnitude faster in terms of the number of variables compared to the methods of van de Geer et al. (2014) and Javanmard and Montanari (2014). In addition, they are equally efficient as the method by Lan et al. (2016).

5 Monte Carlo Experiments

This section examines the finite sample behaviour of the proposed estimators in a Monte Carlo experiment.

5.1 Monte Carlo set-up

Data generating process The data generating process takes the form

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \tag{42}$$

where y is a $n \times 1$ response vector, X a $n \times p$ regressor matrix, and β 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:

$$\text{Equicorrelated: } \Sigma_{jk} = 0.8, \quad \forall j \neq k, \quad \Sigma_{jj} = 1 \quad \forall j, \tag{43}$$

$$\text{Toeplitz: } \Sigma_{jk} = 0.9^{|j-k|}, \quad \forall j, k. \tag{44}$$

The strength of the individual predictors is considered local-to-zero by setting $\beta = \sqrt{\sigma_\varepsilon^2 / n} \cdot b\iota_s$ for a fixed constant b. The vector ι_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 (42). In each replication the predictors in X and the coefficients in β 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 Σ. We report average results for nonzero coefficients and zero coefficients.

Estimation We estimate the coefficients by the Moore-Penrose pseudoinverse, random least squares, and ridge estimator. The lasso estimator uses a penalty term that minimizes the mean squared error under tenfold cross-validation. In the random least squares estimator we average over $N = 1000$ realizations of the regularized covariance matrix with subspace dimension $k = 90$. We compare the performance of the random least squares estimator against ridge regression with a penalty parameter $\gamma = 1$ as in Bühlmann et al. (2013).

Moreover, we compare the proposed estimators to other methods for constructing confidence intervals in high-dimensional regression for all coefficients. The method of van de Geer et al. (2014) (GBRD) serves as the first benchmark,

in which M is constructed by performing lasso for each column in X on the remaining columns in X. For each lasso estimation the penalty parameter is selected by tenfold cross-validation. The method of Zhang and Zhang (2014) boils down to this approach for linear regression problems as in (42). Second, Javanmard and Montanari (2014) (JM) construct M by solving a convex program. We set the tuning parameter $\mu = 2\sqrt{n^{-1}\log p}$, which is equal to the value used in their simulation studies. Both benchmark methods also make use of a bias correction by an initial estimator, for which we use the lasso estimator. Finally, we compare the performance against the recently developed Correlated Predictors Screening (CPS) method by Lan et al. (2016). In this method, for each regressor x_j we find highly correlated regressors from the set of remaining columns in the regressor matrix. We then orthogonalize both y and x_j with respect to this set. Stopping rules for the size of the correlated set and estimation of the noise level can be found in Lan et al. (2016).

Both for our proposed methods and for JM and GBRD we estimate the noise level σ^2 using an estimator based on the lasso as defined in (21).

Evaluation The coverage rate is calculated as the percentage of cases in which the value of the coefficient in the data generating process falls inside the 95% confidence interval. The statistical power is calculated as the percentage of Monte Carlo replications in which zero is not included in the confidence interval of nonzero coefficients.

5.2 Simulation Results

5.2.1 Sparsity and signal strength

Table 2 shows the Monte Carlo simulation results for the set of experiments with an equicorrelated covariance matrix and Table 3 with a Toeplitz covariance matrix. The tables report the estimated coefficients, standard errors, coverage rates, and power of the Moore-Penrose pseudoinverse, random least squares, and ridge regression. Settings vary over the number ($s = 3, 15$) and signal strength ($b = 2, 5$; corresponding to coefficients of size 0.2 and 0.5) of nonzero coefficients.

The proposed methods obtain a coverage rate close to the nominal rate of 95%. The coverage rates are most precise in case of an equicorrelated covariance matrix in a sparse setting with a weak signal. We observe the largest deviations from the nominal rate for a Toeplitz covariance matrix in a non-sparse setting with a strong signal. In general, the quality of the results seem to be higher when an equicorrelated covariance matrix is used. Both the bias and the standard errors are smaller, and the coverage rate is very close to the nominal rate.

We find that the ridge regularization results in an increase in power relative to the Moore-Penrose pseudoinverse estimator, but both estimators are outperformed by random least squares in all considered settings. Even in the high-dimensional settings of Table 2 and 3, where the number of variables is twice as large as the number of observations, the proposed methods achieve a reasonable amount of power, varying from 0.10 to 0.40. The highest power is achieved in a sparse setting with a strong signal strength. In almost all cases, power is larger in settings with equicorrelated covariance matrix instead of Toeplitz.

We find some downward bias for the nonzero coefficients for the proposed methods in this paper. The bias decreases in sparsity, which means that nonzero

Table 2: Monte Carlo simulation: Equicorrelated Covariance Matrix

		$s = 3$				$s = 15$			
method	b	coef.	SE	CR	power	coef.	SE	CR	power
MPI	2	0.19	0.30	0.95	0.10	0.17	0.29	0.94	0.10
	0	0.00	0.30	0.95		0.00	0.29	0.95	
RLS	2	0.19	0.28	0.95	0.10	0.17	0.27	0.94	0.11
	0	0.00	0.28	0.95		0.00	0.27	0.95	
RID	2	0.19	0.29	0.95	0.10	0.17	0.28	0.94	0.11
	0	0.00	0.29	0.95		0.00	0.28	0.95	
GBRD	2	0.17	0.20	0.94	0.13	0.16	0.20	0.93	0.14
	0	0.00	0.20	0.95		0.01	0.20	0.96	
JM	2	0.06	0.05	0.15	0.14	0.09	0.05	0.21	0.27
	0	0.02	0.05	0.96		0.03	0.05	0.91	
CPS	2	0.26	0.23	0.94	0.21	0.68	0.28	0.58	0.70
	0	0.10	0.23	0.92		0.52	0.28	0.55	
MPI	5	0.47	0.30	0.94	0.35	0.44	0.34	0.94	0.27
	0	0.00	0.30	0.95		0.01	0.34	0.96	
RLS	5	0.46	0.28	0.93	0.40	0.43	0.31	0.93	0.30
	0	0.00	0.28	0.95		0.01	0.31	0.96	
RID	5	0.46	0.29	0.93	0.38	0.44	0.33	0.94	0.28
	0	0.00	0.29	0.95		0.01	0.33	0.96	
GBRD	5	0.43	0.20	0.89	0.53	0.42	0.23	0.87	0.44
	0	0.01	0.20	0.96		0.03	0.23	0.96	
JM	5	0.22	0.05	0.14	0.64	0.34	0.06	0.25	0.77
	0	0.02	0.05	0.95		0.11	0.94	0.70	
CPS	5	0.67	0.24	0.89	0.79	1.70	0.47	0.27	0.94
	0	0.26	0.25	0.82		1.30	0.50	0.26	

Note: this table reports the average over the estimated coefficients (coef.), standard errors (SE), coverage rates (CR) and statistical power of the Moore-Penrose pseudoinverse (MPI), random least squares (RLS), ridge regression (RID), and the methods of van de Geer et al. (2014) (GBRD), Javanmard and Montanari (2014) (JM) and Lan et al. (2016) (CPS) for different Monte Carlo experiments where the regressors have an equicorrelated covariance matrix as specified in (43). The coverage rate is calculated as the percentage of cases in which the value of the coefficient in the data generating process falls inside the 95% confidence interval. The statistical power is calculated as the percentage of Monte Carlo replications in which zero is not included in the confidence interval of nonzero coefficients. Estimates are averaged over Monte Carlo replications and coefficients with the same signal strength. 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 $k = 0.9n$, we average over $N = 1000$ realizations of projection matrices, and the penalty parameter in Ridge equals $\gamma = 1$. Results are based on 1000 Monte Carlo replications.

coefficients are more precisely estimated when there are relatively few of them. Compared to the other estimators, random least squares seems to estimate slightly more biased coefficients. For all methods, the coefficients which are set

Table 3: Monte Carlo simulation: Toeplitz Covariance Matrix

		$s = 3$				$s = 15$			
method	b	coef.	SE	CR	power	coef.	SE	CR	power
MPI	2	0.19	0.35	0.95	0.08	0.17	0.34	0.94	0.09
	0	0.00	0.35	0.95		0.00	0.34	0.95	
RLS	2	0.19	0.30	0.95	0.09	0.17	0.29	0.94	0.10
	0	0.00	0.30	0.95		0.01	0.29	0.95	
RID	2	0.19	0.32	0.95	0.09	0.17	0.31	0.94	0.10
	0	0.00	0.32	0.95		0.01	0.31	0.95	
GBRD	2	0.18	0.21	0.94	0.15	0.15	0.20	0.94	0.13
	0	0.01	0.20	0.95		0.02	0.20	0.96	
JM	2	0.10	0.05	0.41	0.31	0.10	0.05	0.28	0.32
	0	0.01	0.05	0.95		0.03	0.95	0.92	
CPS	2	0.19	0.31	0.95	0.10	0.19	0.44	0.95	0.08
	0	0.00	0.32	0.95		0.00	0.45	0.95	
MPI	5	0.46	0.35	0.94	0.28	0.42	0.34	0.91	0.26
	0	0.00	0.35	0.95		0.01	0.66	0.95	
RLS	5	0.45	0.30	0.93	0.35	0.42	0.30	0.89	0.33
	0	0.00	0.30	0.95		0.01	0.70	0.95	
RID	5	0.46	0.32	0.93	0.32	0.42	0.31	0.90	0.30
	0	0.00	0.32	0.95		0.01	0.69	0.95	
GBRD	5	0.42	0.20	0.86	0.55	0.37	0.20	0.77	0.47
	0	0.01	0.20	0.96		0.02	0.80	0.96	
JM	5	0.29	0.05	0.25	0.82	0.29	0.05	0.22	0.73
	0	0.01	0.05	0.95		0.03	0.95	0.89	
CPS	5	0.50	0.37	0.95	0.28	0.48	0.84	0.95	0.09
	0	0.00	0.41	0.95		-0.01	0.88	0.95	

Note: this table reports the results for different Monte Carlo experiments where the regressors have a Toeplitz covariance as specified in (44). For additional information, see the note following Table 2.

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 Moore-Penrose pseudoinverse regression. Standard errors of the random least squares estimates are lower than these estimators in all experiments, where ridge is again a more efficient estimator relative to the pseudoinverse. Except for the non-sparse setting with a strong signal, standard errors are larger for a Toeplitz than an equicorrelated covariance matrix.

Compared to the benchmark models, the proposed models are less (downward) biased and obtain coverage rates much closer to the nominal rate. In all settings under consideration, the methods proposed in this paper produce coverage rates that are closer to the nominal rates than the method of van de Geer et al. (2014). This can be explained by the large bias of the GBRD estimator and the remarkably small standard errors. The JM method produces coefficient

estimates and standard errors that are both close to zero, which results in low coverage rates for the nonzero coefficients. Javanmard and Montanari (2014) present better results under the same choice for the tuning parameter. However, they only run simulation studies for low-dimensional settings, where the number of variables does not exceed the number of observations. The method developed by Lan et al. (2016) performs well for Toeplitz designs. We see only a minor bias in the coefficient estimates, but substantially larger standard errors compared to the methods proposed in this paper when the signal strength and/or the number of nonzero coefficients increase. For the equicorrelated design the coverage rates deteriorate and bias increases severely. Clearly this design does not satisfy the condition that each regressor is only correlated with a small set of the remaining regressors. This condition turns out to be essential for the approach to work well.

5.2.2 Varying signal strength

Since many economic processes can be characterized by a small number of large effects and a large number of small effects on the variable of interest, we now consider a setting in which the signal strength varies over the nonzero coefficients in the data generating process. Table 4 shows the Monte Carlo simulation results for this set of experiments for an equicorrelated and Toeplitz covariance matrix. The sparsity s equals 15 and we randomly assign $b = 10$ to three nonzero coefficients and $b = 2$ to the 12 remaining nonzero coefficients.

In general, the findings for the proposed methods are similar to the settings discussed in the previous paragraph. The nonzero coefficients are estimated with some downward bias, which is larger in the Toeplitz setting relative to the equicorrelated covariance matrix. Estimates of coefficients that are zero in the data generating process are again estimated very close to zero. Although there is a large variation in signal strength, the standard errors are almost the same for coefficients of different strength and we find the same ranking in efficiency; random least squares produces the smallest standard errors, followed by the ridge regularized estimator.

The coverage rates for the zero coefficients are close to the nominal rate. The coverage rates for coefficients with a weak and moderately strong signal are slightly too low. The decrease in coverage rates holds especially for the Toeplitz setting, where standard errors are relatively larger, but also the bias increases relative to data generated from an equicorrelated covariance matrix.

We find that the power for coefficients with intermediate signal strength ($b = 2$) is comparable to settings with a constant signal strength in Table 2 and 3. As expected, the power for the strong signals is much larger, varying between 0.75 and 0.86. In general, power increases for data generated from an equicorrelated covariance matrix relative to a Toeplitz.

Compared to the benchmark estimators, the proposed estimators show also superior performance in the settings with varying signal strength. The distance between the nominal coverage rate and the coverage rate attained by the methods GBRD and JM is in any case larger than for MPI, RLS, and RID. For the Toeplitz design, the coverage rate of CPS is excellent, but the standard errors are almost two times as large as for the competing methods.

Table 4: Monte Carlo simulation: Varying signal strength

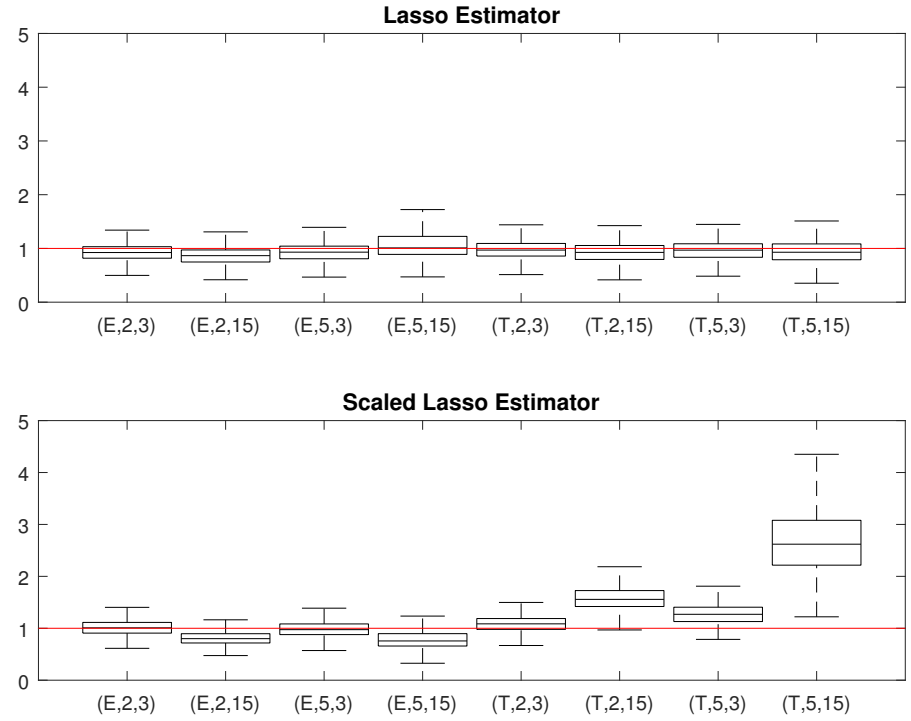
		Equicorrelated				Toeplitz			
method	b	coef.	SE	CR	power	coef.	SE	CR	power
MPI	10	0.94	0.31	0.93	0.84	0.92	0.34	0.91	0.75
	2	0.18	0.31	0.95	0.09	0.17	0.34	0.94	0.09
	0	0.00	0.31	0.95		0.01	0.34	0.95	
RLS	10	0.93	0.28	0.93	0.89	0.91	0.29	0.89	0.85
	2	0.18	0.28	0.95	0.10	0.17	0.29	0.94	0.10
	0	0.00	0.28	0.96		0.01	0.29	0.95	
RID	10	0.94	0.29	0.93	0.86	0.91	0.31	0.90	0.81
	2	0.18	0.29	0.95	0.09	0.17	0.31	0.94	0.09
	0	0.00	0.29	0.96		0.01	0.31	0.95	
GBRD	10	0.90	0.21	0.85	0.97	0.87	0.20	0.81	0.95
	2	0.16	0.21	0.94	0.13	0.15	0.20	0.93	0.13
	0	0.02	0.21	0.96		0.02	0.20	0.96	
JM	10	0.75	0.05	0.20	0.99	0.77	0.05	0.25	0.99
	2	0.11	0.05	0.24	0.34	0.10	0.05	0.26	0.31
	0	0.05	0.05	0.84		0.03	0.05	0.92	
CPS	10	1.76	0.36	0.43	1.00	0.98	0.65	0.95	0.34
	2	1.09	0.40	0.39	0.78	0.19	0.74	0.95	0.06
	0	0.93	0.41	0.37		-0.01	0.75	0.95	0.00

Note: this table reports the results for Monte Carlo experiments with an equicorrelated and Toeplitz covariance matrix, where the nonzero coefficients of the regressors have different signal strengths. Three randomly chosen coefficients out of the 15 nonzero coefficients have signal strength $b = 10$ and the remaining 12 coefficients $b = 2$. For additional information, see the note following Table 2.

Estimation of the noise level The validity of confidence intervals depends on a consistent estimator of the noise level σ^2. Figure 1 shows for each setting of the Monte Carlo experiments a box plot of the estimated σ^2 in each replication. We find that the noise level estimated by scaled lasso can be strongly biased, especially in settings where the data is generated from a Toeplitz covariance matrix, where the lasso estimator results in estimates that are always within one standard deviation from the true value. Therefore, the results in Table 2 and 3 are based on the estimator for the noise level σ^2 as defined in (21).

Computational complexity In addition to accuracy, the proposed estimators also reduce the computational costs. Figure 2 shows on a logarithmic scale for each Monte Carlo experiment the total time spend at constructing the different estimators. The Moore-Penrose pseudoinverse estimator and the ridge regularized estimator demand very little computation time. Since the random least squares estimator is constructed from averaging over $N = 1000$ realizations of the regularized covariance matrix, this estimator takes considerably more time. However, even the computational costs of this demanding estimator is negligible compared to the computation time needed to perform p times a lasso in GBRD and solve p optimization problems in JM.

Figure 1: Estimates noise level Monte Carlo experiments



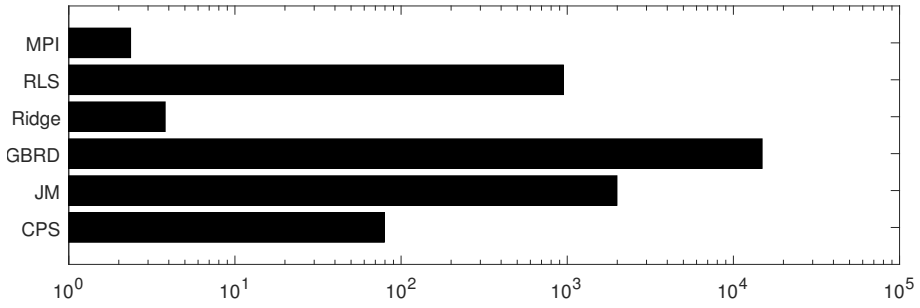
Note: this figure shows for each Monte Carlo experiment a box plot for the estimates of the noise level σ^2 in each replication. The first panel shows these plots for the estimator based on lasso, as defined in (21), and the second panel for the estimator based on scaled lasso as in Sun and Zhang (2012). The red horizontal line indicates the value of $\sigma^2 = 1$ in the data generating process. Settings are indicated by (covmat,b,s), where the covariance matrix covmat varies between equicorrelated (E) and Toeplitz (T), the signal strength ($b = 2, 5$) and sparsity ($s = 3, 15$). For additional information, see the note following Table 2.

In sum, we find that even in small samples with a 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 pseudoinverse estimator. The proposed estimators improve upon existing methods for constructing high-dimensional confidence intervals in terms of coverage rate, and are more robust to different choices of the data generating process. In addition, they significantly reduce the computational complexity that is required, thereby making them more suitable for high-dimensional data.

6 Empirical Application

This section applies the proposed estimators to a macroeconomic dataset. We examine the relation between the real gross domestic product of the U.S. econ-

Figure 2: Computation time estimators Monte Carlo experiments



Note: this figure shows on a logarithmic scale the total time spend at constructing the estimators for each method, averaged over all Monte Carlo experiments in Table 2 and 3.

omy and a large number of macroeconomic and financial indicators.

6.1 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.

6.2 Estimation

The coefficients β are estimated in the regression equation

$$y = Z\delta + X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \tag{45}$$

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

Table 5: Significant effects on Real Gross Domestic Product

		MPI		RLS		RID	
gr.	variable	coef.	SE	coef.	SE	coef.	SE
2	industrial production	0.087	0.042	0.055	0.028	0.074	0.034
3	employees wholesale trade			-0.047	0.022		
3	hours worked business	0.752	0.036	0.750	0.025	0.750	0.030
3	hours worked nonfarm	0.152	0.037	0.134	0.025	0.144	0.031
4	housing starts			0.029	0.014		
5	retail sales	0.042	0.019	0.037	0.016	0.040	0.018
5	manufacturing inventories			0.038	0.018	0.044	0.022
6	GDP deflator	-0.038	0.018	-0.028	0.013	-0.033	0.016
7	productivity nonfarm	0.050	0.023	0.051	0.016	0.051	0.019
7	productivity business	0.776	0.023	0.763	0.016	0.771	0.019
7	labour costs			-0.042	0.020	-0.045	0.023
8	rate commercial paper	0.050	0.024	0.051	0.020	0.048	0.021
8	rate Eurodollar deposit	0.069	0.030	0.055	0.021	0.062	0.025
9	real money stock	-0.044	0.022	-0.039	0.017	-0.042	0.019
12	consumer sentiment	0.070	0.032	0.055	0.023	0.065	0.028
13	stock price volatility	0.058	0.027	0.043	0.018	0.049	0.022
14	federal debt	-0.046	0.022	-0.038	0.017	-0.043	0.020

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 the Moore-Penrose pseudoinverse estimator (MPI), random least squares (RLS), and ridge regularization (RID). The group numbers (gr.) correspond to the FRED-QD variable categories. The fred mnemonics and variable descriptions corresponding to the variable names are given in Appendix D.

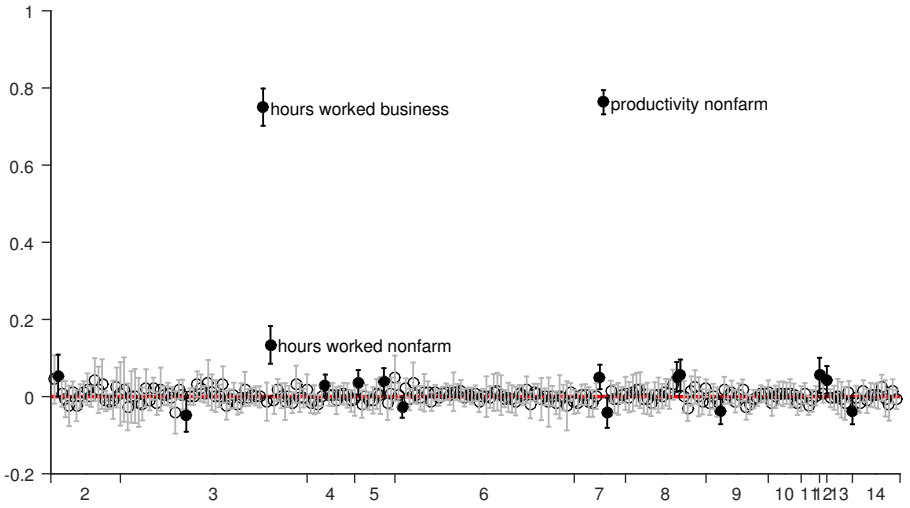
the macroeconomic relations in β, we partial out the variables in Z using the Frisch-Waugh theorem before estimating β. We note that Assumption 3 and Assumption 4 are now imposed on $M_Z X$ with M_Z the projection matrix orthogonal to the columns of Z. The proof for Theorem 1 carries through with n replaced by $n - n_z$ with n_z the number of columns of Z. After initialization and the loss in degrees of freedom by partialling out Z, we are left with a 110×231 matrix $X^* = M_Z X$ which has rank $n - n_z = 105$.

When estimating by random least squares, we choose the subspace dimension $k = 95$ and $N = 1000$ realizations of the regularized covariance 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 set to $\gamma = 1$.

6.3 Empirical Results

Table 5 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 yields lower standard errors compared to the benchmark methods. Ridge regression finds 15 out of the 231 coefficients to be significant, which is slightly higher for random least squares with 17 coefficients. The Moore-Penrose pseudoinverse regression estimates 13 coefficients to be significant, which cor-

Figure 3: Confidence Intervals Coefficients Regression GDP



Note: this figure shows the 95% confidence intervals together with the estimated coefficients in the regression of the FRED-QD variables on real GDP. Boldfaced coefficients are significantly different from zero on a five percent significance level. The numbers on the x-axis indicate the FRED categories associated with the effects.

responds to the theoretical finding that the random least squares and ridge estimators yield higher 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 (hours worked business), real output per hour of all persons in the business sector (productivity business), and hours of all persons worked in the nonfarm business sector (hours worked nonfarm) have large positive coefficients of respectively 0.750, 0.134, and 0.763 for random least squares. More elaborate descriptions of the remaining variables can be found in Appendix D. Figure 3 shows that the remaining coefficients are close to zero. We do not find any significant effect of variables in the categories household balance sheets (10), and exchange rates (11). Random least squares finds five significant negative effects on the real gross domestic product; all employees in wholesale trade (employees wholesale trade), gross domestic product: chain-type price index (GDP deflator), unit labor cost in the business sector (labour costs), real MZM (money-zero-maturity) money stock (real money stock), and the total public debt as percentage of GDP (Federal debt). Ridge regression does not find a significantly negative effect of all employees in wholesade trade. The negative effect assigned to the number of employees in wholesale trade found by random least squares is remarkable, but note that employment also effects GDP positively via hours worked in the business and nonfarm sector, which makes the net effect of employment on real GDP positive.

7 Conclusion

This paper proposes novel methods for constructing confidence intervals in high-dimensional linear regression models, where the number of unknown coefficients increases almost exponentially with the number of observations. We approximate the inverse of the singular empirical covariance matrix by a diagonally scaled Moore-Penrose pseudoinverse. After a bias correction with the lasso this yields an unbiased, normally distributed estimator. Confidence intervals can then be constructed using standard procedures. In order to increase efficiency, we consider two regularized variants: random least squares, which relies on low-dimensional random projections of the data, and ridge regularization. These estimators are shown to have the same theoretical validity under suitable choices of the regularization parameters.

Monte Carlo experiments show that, even in small samples with high dimensional regressor matrix, the proposed estimators indeed provide valid confidence intervals with correct coverage rates. 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, we find a large positive effect from variables in the employment and productivity categories. A small negative effect is found from federal debt.

References

Ahlswede, R. and Winter, A. (2002). Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579.

Barro, R. J. and Lee, J.-W. (1993). International comparisons of educational attainment. *Journal of Monetary Economics*, 32(3):363–394.

Belloni, A., Chernozhukov, V., et al. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., and Hansen, C. (2010). Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics*, 3.

Bühlmann, P. et al. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, pages 2313–2351.

Caner, M. and Kock, A. B. (2014). Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *arXiv preprint arXiv:1410.4208*.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1):649–688.

Chikuse, Y. (1990). The matrix angular central Gaussian distribution. *Journal of Multivariate Analysis*, 33(2):265–274.

Chikuse, Y. (2012). *Statistics on special manifolds*, volume 174. Springer Science & Business Media.

Dasgupta, S., Hsu, D., and Verma, N. (2012). A concentration theorem for projections. *arXiv preprint arXiv:1206.6813*.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32(3):928–961.

Fernandez, C., Ley, E., and Steel, M. F. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5):563–576.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909.

Kabán, A. (2014). New bounds on compressive linear least squares regression. In *AISTATS*, pages 448–456.

Lan, W., Zhong, P.-S., Li, R., Wang, H., and Tsai, C.-L. (2016). Testing a single regression coefficient in high dimensional linear models. *Journal of Econometrics*, 195(1):154–168.

Maillard, O. and Munos, R. (2009). Compressed least-squares regression. In *Advances in Neural Information Processing Systems*, pages 1213–1221.

Marzetta, T. L., Tucci, G. H., and Simon, S. H. (2011). A random matrix-theoretic approach to handling singular covariance estimates. *IEEE Transactions on Information Theory*, 57(9):6256–6271.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, 26:35–67.

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, 340(6139):1467–1471.

Sala-i-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review*, pages 178–183.

Serfling, R. J. (2006). Multivariate symmetry and asymmetry. *Encyclopedia of Statistical Sciences*.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202.

Wang, X. and Leng, C. (2015). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

A Properties of spherically symmetric matrices

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

$$Z \overset{(d)}{=} ZT, \quad T \in \mathcal{O}(p). \tag{46}$$

Then Z can be decomposed by a singular value decomposition as

$$Z = VSU', \tag{47}$$

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 (47) is

$$Z = VS_nU_n', \tag{48}$$

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

$$U_n' = [I_n, O_{n,p-n}]U'. \tag{49}$$

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

$$V_{n,p} = \left\{A \in R^{p \times n} : A'A = I_n\right\}. \tag{50}$$

B Proofs

This appendix provides the proofs of the theorems in Section 4. Auxiliary definitions and lemmas used in these proofs are in Appendix C.

B.1 Proof of the compatibility condition

Note that $||\beta_{S_0}||_1 \leq \sqrt{s_0}||\beta_{S_0}||_2$, so it suffices to show

$$||\beta||_2^2 \leq \frac{\beta' \frac{1}{n} X'X\beta}{\phi_0}. \tag{51}$$

Using (23), we have

$$\begin{aligned} \beta' \frac{1}{n} X'X\beta &= \beta'\Sigma^{1/2} \frac{1}{n} US'SU'\Sigma^{1/2}\beta \\ &\geq \frac{1}{c_Z} \frac{p}{n} v'U_nU_n'v, \end{aligned} \tag{52}$$

where the last line holds since the non-zero eigenvalues $S'S$ are the same as the eigenvalues of ZZ' which are bounded by Assumption 3. Again by Assumption 3, U_nU_n' is a projection matrix that is uniformly distributed over

the Grassmannian manifold $G(n,p)$, see Chikuse (2012). For the l_2 norm of a projected vector, strong concentration bounds exists that guarantee

$$\begin{aligned} P\left(\sqrt{\frac{p}{n}}||U_n'v||_2 \geq (1+\epsilon)||v||_2\right) &\leq 2e^{-\frac{1}{4}\epsilon^2 n}, \\ P\left(\sqrt{\frac{p}{n}}||U_n'v||_2 \leq (1-\epsilon)||v||_2\right) &\leq 2e^{-\frac{1}{4}\epsilon^2 n}. \end{aligned} \tag{53}$$

Therefore, with high probability

$$\begin{aligned} \beta'\frac{1}{n}X'X\beta &\geq \frac{1}{c_Z}\frac{1-\epsilon}{1+\epsilon}\beta'\Sigma\beta \\ &\geq \frac{1}{c_Z}\frac{1-\epsilon}{1+\epsilon}\lambda_{\min}(\Sigma)||\beta||_2^2. \end{aligned} \tag{54}$$

Choosing $\phi_0 \leq \frac{1}{c_Z}\frac{1-\epsilon}{1+\epsilon}\lambda_{\min}(\Sigma)$ yields the desired result. ■

B.2 Proof of Theorem 2

Using Assumption 3 and (48) we have that

$$X'(XX')^{-1}X = \Sigma^{1/2}U_n(U_n'\Sigma U_n)^{-1}U_n'\Sigma^{1/2}, \tag{55}$$

By Lemma 3 in Appendix C, we can write

$$U_n = W(W'W)^{-1/2}, \tag{56}$$

with the elements of W standard normal and independently distributed. This implies

$$X'(XX')^{-1}X = \Sigma^{1/2}W(W'\Sigma W)^{-1}W'\Sigma^{1/2} = HH', \tag{57}$$

where we define $H = \Sigma^{1/2}W(W'\Sigma W)^{-1/2}$.

We will separately bound the diagonal and off-diagonal elements of HH'. The proof extends the approach by Wang and Leng (2015).

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

$$H_{ij} \overset{(d)}{=} \frac{w_{ij}}{\sqrt{\sum_{j=1}^p w_{ij}^2}}, \tag{58}$$

where w_{ij} denotes the elements of W.

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

$$e_i'HH'e_i \overset{(d)}{=} \frac{w_{i1}^2 + \ldots + w_{in}^2}{w_{i1}^2 + \ldots w_{ip}^2}, \tag{59}$$

Using now the Bernstein inequality for χ^2 random variables in (137) in Appendix C, we have the following high probability bound

$$\begin{aligned} P\left(\frac{w_{i1}^2+\ldots+w_{in}^2}{w_{i1}^2+\ldots w_{ip}^2} > \frac{n}{p}\frac{1+\epsilon}{1-\epsilon}\right) &\leq \exp\left(-\frac{\epsilon^2 n}{8}\right) + \exp\left(-\frac{\epsilon^2 p}{8}\right) \\ &\leq 2\exp\left(-\frac{\epsilon^2 n}{8}\right), \end{aligned} \tag{60}$$

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

$$P\left(\frac{w_{i1}^2+\ldots+w_{in}^2}{w_{i1}^2+\ldots w_{ip}^2} < \frac{n}{p}\frac{1-\epsilon}{1+\epsilon}\right) \leq 2\exp\left(-\frac{\epsilon^2 n}{8}\right), \tag{61}$$

Using Bonferonni's inequality this yields

$$P\left(e_1' U_n U_n' e_1 > c_\epsilon \frac{n}{p} \quad \cup \quad e_1' U_n U_n' e_1 < \frac{1}{c_\epsilon}\frac{n}{p}\right) \leq 4\exp\left(-\frac{\epsilon^2 n}{8}\right), \tag{62}$$

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

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

$$\begin{aligned} ||H'v||_2^2 &= v'HH'v \\ &= v'\Sigma^{\frac{1}{2}} U_n (U_n'\Sigma U_n)^{-1} U_n' \Sigma^{\frac{1}{2}} v \\ &\leq \kappa v' U_n U_n' v, \end{aligned} \tag{63}$$

where the condition number $\kappa = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} < \infty$ by Assumption 3. Similarly

$$v'HH'v \geq \frac{1}{\kappa} v' U_n U_n' v \tag{64}$$

Since $U_n \overset{(d)}{=} QU_n$ with $Q \in \mathcal{O}(p)$, upon choosing Q such that $Qv = e_1$, we immediately obtain

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

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 with high probability $\frac{|e_i' HH' e_j|}{e_i' HH' e_i}$. We proof this for $i = 1$ and $j = 2$, and then apply a union bound over all i and j. We separate three cases: (a) $e_1' HH e_1 > c_\epsilon \kappa \frac{n}{p}$, (b) $c_\epsilon \kappa \frac{n}{p} > e_1' HH' e_1 > \frac{1}{c_\epsilon \kappa}\frac{n}{p}$, and (c) $e_1' HH e_1 < c_\epsilon \kappa \frac{n}{p}$. Conditioning on these three cases and using the trivial fact that for any probability $P(\cdot) \leq 1$, it follows that

$$\begin{aligned} P\left(\frac{|e_1' HH' e_2|}{e_1' HH' e_1} > t\right) &\leq P\left(e_1' HH' e_1 > c_\epsilon \kappa \frac{n}{p}\right) + P\left(e_1' HH' e_1 < \frac{1}{c_\epsilon \kappa}\frac{n}{p}\right) \\ &+ \int_{\frac{1}{c_\epsilon \kappa}\frac{n}{p}}^{c_\epsilon \kappa \frac{n}{p}} P\left(\frac{|e_1' HH' e_2|}{e_1' HH' e_1} > t \middle| e_1' HH' e_1 = t_1^2\right) P\left(e_1' HH' e_1 = t_1^2\right) dt_1^2. \end{aligned} \tag{66}$$

Lemma 6 in Appendix C states that

$$e_1' HH' e_2 \stackrel{(d)}{=} h_{11}h_{12} \mid \left\{ h_{11}^2 = e_1' HH' e_1 \right\}. \tag{67}$$

In which case

$$e_1' HH' e_2 | \left\{ e_1' HH' e_1 = t_1^2 \right\} \stackrel{(d)}{=} h_{11}h_{12} | \left\{ h_{11}^2 = t_1^2 \right\}. \tag{68}$$

After rewriting this expression using Lemma 5 in Appendix C we have

$$h_{1i} | \left\{ h_{11}^2 = t_1^2 \right\} \stackrel{(d)}{=} \frac{\sqrt{1-t_1^2} y_i}{\sqrt{y_2^2 + \ldots + y_p^2}} \text{ for } i = 2, \ldots, p, \tag{69}$$

where $\tilde{y} = (y_2, \ldots, y_p) \sim N(0, \tilde{\Sigma})$. Now we establish the following upper bound

$$\begin{aligned} P\left(\frac{|e_1' HH' e_2|}{e_1' HH' e_1} > t \middle| h_{11}^2 = t_1^2 \right) &= P\left(\frac{|h_{11}h_{12}|}{h_{11}^2} > t \middle| h_{11}^2 = t_1^2 \right) \\ &= P\left(\frac{\sqrt{1-t_1^2}|y_2|}{\sqrt{y_2 + \ldots + y_p}} > |t_1| t \right) \\ &= P\left(\frac{|y_2|}{\sqrt{y_2 + \ldots + y_p}} > \sqrt{\frac{t_1^2}{1-t_1^2}} t \right). \end{aligned} \tag{70}$$

Standard bounds on normal and χ^2 distributed variables can be applied, and furthermore using the fact that $\tilde{y}$ has a rank $p-n$ degenerate covariance matrix, we have

$$\begin{aligned} P\left(|y_2| > \eta \sqrt{\lambda_{\max}(\Sigma)} \right) &\leq 2e^{-\frac{\eta^2}{2}} \\ P\left(\sqrt{y_2^2 + \ldots + y_p^2} \leq \sqrt{\lambda_{\min}(\Sigma)} \sqrt{p-n}(1-\tilde{\eta}) \right) &\leq e^{-\frac{1}{2}(p-n)\tilde{\eta}^2}. \end{aligned} \tag{71}$$

Then we know that

$$P\left(\frac{|y_2|}{\sqrt{y_2^2 + \ldots + y_p^2}} > \frac{\eta}{1-\tilde{\eta}} \sqrt{\frac{\kappa}{p-n}} \right) \leq 2e^{-\frac{\eta^2}{2}} + e^{-\frac{1}{2}(p-n)\tilde{\eta}^2}. \tag{72}$$

So we can choose

$$t = \sqrt{\frac{1-t_1^2}{t_1^2}} \frac{\eta}{1-\tilde{\eta}} \sqrt{\frac{\kappa}{p-n}}. \tag{73}$$

In fact any $t_- < t$ is sufficient, so that we take $t_1^2 = c_\epsilon \kappa \frac{n}{p}$,

$$t > \sqrt{\frac{1}{c_\epsilon} \frac{p}{p-n} - \frac{n}{p-n}\kappa} \frac{\eta}{\sqrt{n}}. \tag{74}$$

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

$$\sqrt{\frac{1}{c_\epsilon} \frac{p/n}{p/n-1} - \frac{1}{p/n-1}\kappa} \frac{\eta}{\sqrt{n}} > \sqrt{\frac{\frac{c_p}{c_\epsilon} - \kappa}{c_p - 1}} \frac{\eta}{\sqrt{n}}. \tag{75}$$

Choosing

$$\eta = a\sqrt{\frac{c_p - 1}{c_p/c_\epsilon - \kappa}}\sqrt{\log p}, \tag{76}$$

results in

$$P\left(\frac{|e_1' HH' e_2|}{e_1' HH' e_1} > a\sqrt{\frac{\log p}{n}}\right) \leq 2e^{-\frac{a^2}{2}\frac{c_p-1}{c_p/c_\epsilon-\kappa}\log p} + e^{-\frac{1}{2}(p-n)\tilde{\eta}^2}. \tag{77}$$

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

$$P\left(\frac{|e_i' HH' e_j|}{e_i' HH' e_i} > a\sqrt{\frac{\log p}{n}}\right) \leq O\left(p^{-\tilde{c}}\right), \tag{78}$$

with $\tilde{c} = \frac{a^2}{2}\frac{c_p-1}{c_p/c_\epsilon-\kappa} - 2$. ■

B.3 Proof of Theorem 3

For all the estimators we should have

$$\frac{1}{\sqrt{n}} M\epsilon = O_p(1). \tag{79}$$

Building on the proof by Wang and Leng (2015), we denote the individual error terms of the estimator of β_i as

$$\begin{aligned} \eta_i &= \frac{d_i}{\sqrt{n}} e_i' X'(XX')^{-1}\varepsilon \\ &\overset{(d)}{=} \frac{d_i}{\sqrt{n}} ||e_i' X'(XX')^{-1}||_2 \frac{\sigma e_i' X'(XX')^{-1} u}{||e_i' X'(XX')^{-1}||_2}, \end{aligned} \tag{80}$$

where $u \sim i.i.d.(0, 1)$.

We first bound the norm term

$$\frac{d_i}{\sqrt{n}} ||e_i' X'(XX')^{-1}||_2 = \sqrt{n}\frac{||e_i' X'(XX')^{-1}||_2}{e_i' X'(XX')^{-1} X e_i}, \tag{81}$$

Using standard norm inequalities, we have

$$\begin{aligned} ||e_i' X'(XX')^{-1}||_2^2 &= e_i' X'(XX')^{-2} X e_i \\ &\leq \frac{1}{\lambda_{\min}(XX')} e_i' X'(XX')^{-1} X e_i, \\ ||e_i' X'(XX')^{-1}||_2^2 &\geq \frac{1}{\lambda_{\max}(XX')} e_i' X'(XX')^{-1} X e_i. \end{aligned} \tag{82}$$

The eigenvalues of $XX' = Z\Sigma Z'$ satisfy

$$\begin{aligned} \lambda_{\max}(Z\Sigma Z') &\leq \lambda_{\max}(\Sigma)\lambda_{\max}(ZZ'), \\ \lambda_{\min}(Z\Sigma Z') &\geq \lambda_{\min}(\Sigma)\lambda_{\min}(ZZ'), \end{aligned} \tag{83}$$

and are bounded by Assumption 3

$$\begin{aligned} P\left(\lambda_{\max}\left(\frac{1}{p}ZZ'\right) > c_Z\right) &\leq e^{-C_Z n}, \\ P\left(\lambda_{\min}\left(\frac{1}{p}ZZ'\right) < \frac{1}{c_Z}\right) &\leq e^{-C_Z n}. \end{aligned} \tag{84}$$

Finally, using the previously established bounds in (65)

$$\begin{aligned} P\left(e_i'X'(XX')^{-1}Xe_i > c_\epsilon\kappa\frac{n}{p}\right) &\leq 2e^{-\frac{\epsilon^2 n}{8}}, \\ P\left(e_i'X'(XX')^{-1}Xe_i < \frac{1}{c_\epsilon\kappa}\frac{n}{p}\right) &\leq 2e^{-\frac{\epsilon^2 n}{8}}, \end{aligned} \tag{85}$$

it follows that with probability exceeding $1 - 4\exp(-\epsilon^2 n/8) - 2\exp(-C_Z n)$ we have that

$$\left(\frac{1}{\lambda_{\max}(\Sigma)}\frac{n}{p}\frac{1}{c_\epsilon\kappa\frac{n}{p}}\right)^{1/2} \leq \frac{d_i}{\sqrt{n}}||e_i'X'(XX')^{-1}||_2 \leq \left(\frac{1}{\lambda_{\min}(\Sigma)}\frac{n}{p}\frac{1}{\frac{1}{c_\epsilon\kappa}\frac{n}{p}}\right)^{1/2}, \tag{86}$$

which shows that

$$\frac{d_i}{\sqrt{n}}||e_i'X'(XX')^{-1}||_2 = O(1). \tag{87}$$

We now turn to the second term of (80)

$$\frac{\sigma e_i'X'(XX')^{-1}u}{||e_i'X'(XX')^{-1}||_2} = \frac{\sigma e_i'\frac{1}{\sqrt{n}}X'\left(\frac{1}{p}XX'\right)^{-1}u}{\left|\left|e_i'\frac{1}{\sqrt{n}}X'\left(\frac{1}{p}XX'\right)^{-1}\right|\right|_2}. \tag{88}$$

When $u \sim N.i.d(0,1)$, it is immediately clear that

$$\frac{1}{\sqrt{n}}X'\left(\frac{1}{p}XX'\right)^{-1}u \overset{(d)}{\to} N\left[0, \lim_{n\to\infty}\frac{1}{n}X'\left(\frac{1}{p}XX'\right)^{-2}X\right], \tag{89}$$

and hence, as $n \to \infty$,

$$\frac{\sigma e_i'\frac{1}{\sqrt{n}}X'\left(\frac{1}{p}XX'\right)^{-1}u}{\left|\left|e_i'\frac{1}{\sqrt{n}}X'\left(\frac{1}{p}XX'\right)^{-1}\right|\right|_2} \overset{(d)}{\to} N(0,\sigma^2). \tag{90}$$

B.4 Proof of Theorem 4

We show that the results for the Moore-Penrose inverse in Theorem 2 and Theorem 3 also hold for random least squares.

Size of the bias First, we will show that

$$P\left(\left|\frac{1}{n}d_i^{\text{RLS}}\text{E}[r_i(R'X'XR)^{-1}R']X'x_j - \delta_{ij}\right| > \tilde{a}\sqrt{\frac{\log p}{n}}\right) = O\left(p^{-c}\right) \tag{91}$$

We first list several properties of the expectation $\mathrm{E}[R(R'X'XR)^{-1}R']X'X$.

Consider the eigenvalue decomposition

$$\frac{1}{n}X'X = \hat{U}\hat{\Lambda}\hat{U}' \tag{92}$$

Then we have

$$\begin{aligned}\mathrm{E}[R(R'X'XR)^{-1}R']X'X &= \mathrm{E}[R(R'\hat{U}\hat{\Lambda}\hat{U}'R)^{-1}R']\hat{U}\hat{\Lambda}\hat{U}' \\ &\overset{(d)}{=} \hat{U}\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\hat{\Lambda}\hat{U}'\end{aligned} \tag{93}$$

using that for any fixed unitary matrix $\hat{U}'R \overset{(d)}{=} \Phi$, with Φ again a matrix of independent standard normal random variables.

Furthermore, Marzetta et al. (2011) show that $\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\hat{\Lambda}$ is a diagonal matrix. This can proven by noting that a matrix A is diagonal if and only if for all diagonal unitary matrices Ω, we have that $\Omega A\Omega^* = A$ with Ω^* the complex conjugate of Ω. Indeed

$$\begin{aligned}\Omega\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\hat{\Lambda}\Omega^* &= \Omega\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\Omega^*\hat{\Lambda} \\ &= \Omega\mathrm{E}[\Phi(\Phi'\Omega^*\Omega\hat{\Lambda}\Omega^*\Omega\Phi)^{-1}\Phi']\Omega^*\hat{\Lambda} \\ &\overset{(d)}{=} \mathrm{E}[\Psi(\Psi'\hat{\Lambda}\Psi)^{-1}\Psi']\hat{\Lambda}\end{aligned} \tag{94}$$

where Ψ is again a matrix of standard normals, and using as above that $\Omega\Phi \overset{(d)}{=} \Psi$ for any unitary matrix Ω.

Finally, Marzetta et al. (2011) establish the following relationship

$$\mathrm{E}[\Psi(\Psi'\hat{\Lambda}\Psi)^{-1}\Psi']\hat{\Lambda} = I - V, \tag{95}$$

where

$$V = \mathrm{E}[(\Xi'\hat{\Lambda}^{-1}\Xi)^{-1}\Xi']\hat{\Lambda}^{-1} \tag{96}$$

a diagonal matrix with Ξ is a $p \times (n-k)$ matrix with independent standard normal entries.

Using (95), it follows that

$$\mathrm{E}_R[R(R'X'XR)^{-1}R']X'X = \hat{U}(I-V)\hat{U}' \tag{97}$$

Now, $\hat{U}\hat{U}'$ is the Moore-Penrose pseudoinverse post-multiplied by X, which is identical to (57), so that we have

$$\hat{U}\hat{U}' = X'(XX')^{-1}X = HH' \tag{98}$$

Therefore, one expects that if the entries of $\hat{U}V\hat{U}'$ are sufficiently small compared to $\hat{U}\hat{U}'$, then the results obtained under the Moore-Penrose inverse will maintain to hold.

We can use the following string of inequalities

$$
\begin{aligned}
P\left(\frac{|\hat{u}_i'(I-V)\hat{u}_j|}{\hat{u}_i'(I-V)\hat{u}_i} > \zeta\right) &\leq P\left(\frac{|\hat{u}_i'(I-V)\hat{u}_j|}{\hat{u}_i'\hat{u}_i(1-||V||_2)} > \zeta\right) \\
&\leq P\left(\frac{|\hat{u}_i'\hat{u}_j|}{\hat{u}_i'\hat{u}_i} + \frac{|\hat{u}_i'V\hat{u}_j|}{\hat{u}_i'\hat{u}_i} > \zeta(1-||V||_2)\right) \\
&\leq P\left(\frac{|\hat{u}_i'\hat{u}_j|}{\hat{u}_i'\hat{u}_i} + ||V||_2\frac{||\hat{U}'||_2^2}{\hat{u}_i'\hat{u}_i} > \zeta(1-||V||_2)\right) \\
&\leq P\left(\frac{|\hat{u}_i'\hat{u}_j|}{\hat{u}_i'\hat{u}_i} + ||V||_2\frac{\lambda_{\max}(\hat{U}\hat{U}')}{\lambda_{\min}(\hat{U}\hat{U}')} > \zeta(1-||V||_2)\right)
\end{aligned}
\tag{99}
$$

We can bound this by using the high probability eigenvalue bounds established earlier. Denote by $\mathcal{E}$ the event that $\lambda_{\max}(\hat{U}\hat{U}') < c_\epsilon\kappa\frac{n}{p}$, $\lambda_{\min}(\hat{U}\hat{U}') > (c_\epsilon\kappa)^{-1}\frac{n}{p}$, then the string of inequalities proceeds as

$$
\begin{aligned}
&\leq P\left(\frac{|\hat{u}_i'\hat{u}_j|}{\hat{u}_i'\hat{u}_i} + ||V||_2\left(c_\epsilon\kappa\right)^2 > \zeta(1-||V||_2)\middle|\mathcal{E}\right)\left[1-4\exp\left(-\frac{\epsilon^2 n}{8}\right)\right] + \\
&\quad + 4\exp\left(-\frac{\epsilon^2 n}{8}\right) \\
&= P\left(\frac{|\hat{u}_i'\hat{u}_j|}{\hat{u}_i'\hat{u}_i} > \zeta - ||V||_2\left[\zeta + \left(c_\epsilon\kappa\right)^2\right]\right)\left[1-4\exp\left(-\frac{\epsilon^2 n}{8}\right)\right] + \\
&\quad + 4\exp\left(-\frac{\epsilon^2 n}{8}\right)
\end{aligned}
\tag{100}
$$

Now choose $\zeta = a\sqrt{\frac{\log p}{n}}$ and define scale free constants $c_\zeta > \zeta$ and c_V such that $a > c_V(c_\zeta + c_\epsilon^2\kappa^2) > 0$, then if we have

$$
||V||_2 = c_V\sqrt{\frac{\log p}{n}} \tag{101}
$$

Then

$$
\begin{aligned}
P\left(\frac{|\hat{u}_i'(I-V)\hat{u}_j|}{\hat{u}_i'(I-V)\hat{u}_i} > \zeta\right) &\leq P\left(\frac{|\hat{u}_i'\hat{u}_j|}{\hat{u}_i'\hat{u}_i} > \tilde{a}\sqrt{\frac{\log p}{n}}\right)\left[1-4\exp\left(-\frac{\epsilon^2 n}{8}\right)\right] \\
&\quad + 4\exp\left(-\frac{\epsilon^2 n}{8}\right) \\
&= O\left(p^{-\tilde{c}}\right)
\end{aligned}
\tag{102}
$$

with $\tilde{a} = a - c_V(c_\zeta + c_\epsilon^2\kappa^2)$, and $\tilde{c}$ as in Theorem 2 with a replaced by $\tilde{a}$. The last line of (102) follows directly from Theorem 2.

What remains to be shown is that for a specific choice of k, the elements of the diagonal matrix V are $v_i \leq c_V\sqrt{\frac{\log p}{n}}$, and therefore $||V||_2 \leq c_V\sqrt{\frac{\log p}{n}}$.

Denote by $\hat{\lambda}_i$ the i-th diagonal element of the diagonal matrix of empirical

eigenvalues $\hat{\Lambda}$, ξ_i the i-th column of Ξ, and $A_{-i} \equiv \sum_{j \neq i} \hat{\lambda}_j^{-1} \xi_j \xi_j'$. It holds that

$$\begin{aligned} v_i &= \hat{\lambda}_i^{-1} \xi_i' (\Xi' \hat{\Lambda}^{-1} \Xi)^{-1} \xi_i \\ &= \hat{\lambda}_i^{-1} \xi_i' \left(\sum_{j \neq i} \hat{\lambda}_i^{-1} \xi_j \xi_j' + \hat{\lambda}_i^{-1} \xi_i \xi_i' \right)^{-1} \xi_i \\ &= \hat{\lambda}_i^{-1} \xi_i' \left(A_{-i} + \hat{\lambda}_i^{-1} \xi_i \xi_i' \right)^{-1} \xi_i \\ &= \hat{\lambda}_i^{-1} \xi_i' \left(A_{-i}^{-1} - \frac{\hat{\lambda}_i^{-1} A_{-i}^{-1} \xi_i \xi_i' A_{-i}^{-1}}{1 + \hat{\lambda}_i^{-1} \xi_i' A_{-i}^{-1} \xi_i} \right) \xi_i \\ &= \frac{\hat{\lambda}_i^{-1} \nu_i}{1 + \hat{\lambda}_i^{-1} \nu_i}, \end{aligned} \tag{103}$$

where

$$\nu_i = \xi_i' A_{-i}^{-1} \xi_i = \xi_i' \left(\Xi_{-i} \hat{\Lambda}_{-i}^{-1} \Xi_{-i} \right)^{-1} \xi_i > 0. \tag{104}$$

This shows that random least squares performs a generalized type of ridge regression, where the penalty is different for each eigenvalue. Now we can use Jensen's inequality and the fact that $x/(1+x)$ with $x > 0$ is a concave function to show that

$$v_i \leq \frac{\hat{\lambda}_i^{-1} \mathrm{E}[\nu_i]}{1 + \hat{\lambda}_i^{-1} \mathrm{E}[\nu_i]} \leq \frac{\hat{\kappa} \frac{n-k-1}{k}}{1 + \hat{\kappa} \frac{n-k-1}{k}}, \tag{105}$$

where $\hat{\kappa} = \frac{\lambda_{\max}(\hat{\Lambda})}{\lambda_{\min}(\hat{\Lambda})} \leq c_Z^2 c_\epsilon^2 \kappa$.

If we require $v_i \leq c_V \sqrt{\frac{\log p}{n}}$, then we need to choose

$$\begin{aligned} k &= \frac{\hat{\kappa}}{\hat{\kappa} + \frac{c_V \sqrt{\frac{\log p}{n}}}{1 - c_V \sqrt{\frac{\log p}{n}}}} (n-1) \\ &\approx \left(1 - c_k \sqrt{\frac{\log p}{n}} \right) (n-1), \end{aligned} \tag{106}$$

where $c_k = \frac{c_V}{c_Z^2 c_\epsilon^2 \kappa}$.

Order of the variance term What remains to be shown is that the variance term satisfies

$$\frac{d_i^{\mathrm{RLS}}}{\sqrt{n}} e_i \mathrm{E} \left[R (R' X' X R)^{-1} R' \right] X' = O_p(1). \tag{107}$$

Again we split this as follows

$$\frac{d_i^{\mathrm{RLS}}}{\sqrt{n}} || e_i \mathrm{E} \left[R (R' X' X R)^{-1} R' \right] X' ||_2 \frac{e_i \mathrm{E} \left[R (R' X' X R)^{-1} R' \right] X'}{|| e_i \mathrm{E} \left[R (R' X' X R)^{-1} R' \right] X' ||_2}. \tag{108}$$

Using the results of the previous section allows us to rewrite the (squared) norm as

$$\left(\frac{d_i^{\mathrm{RLS}}}{\sqrt{n}} || e_i \mathrm{E} \left[R (R' X' X R)^{-1} R' \right] X' ||_2 \right)^2 = n \frac{e_i' \hat{U}_n (I - V) \hat{\Lambda}^{-1} (I - V) \hat{U}_n' e_i}{(e_i' \hat{U}_n (I - V) \hat{U}_n' e_i)^2}, \tag{109}$$

which can be lower and upper bounded as

$$\begin{aligned}\frac{1}{\lambda_{\max}(\hat{\Lambda})} n \frac{e_i'\hat{U}_n(I-V)^2\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2} &\leq n\frac{e_i'\hat{U}_n(I-V)\hat{\Lambda}^{-1}(I-V)\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2} \\ &\leq \frac{1}{\lambda_{\min}(\hat{\Lambda})} n \frac{e_i'\hat{U}_n(I-V)^2\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2}.\end{aligned} \tag{110}$$

Under stated assumptions, all eigenvalues are $O(p)$. Also, from (101) we know that the elements of V satisfy $0 \leq v_i \leq \frac{c_V}{c_\varsigma + c_\epsilon^2\kappa^2}\sqrt{\frac{\log p}{n}}$. Then

$$\begin{aligned}O\left(\frac{n}{p}(1-v_{\max})^2\frac{1}{e_i'\hat{U}_n\hat{U}_n'e_i}\right) &\leq n\frac{e_i'\hat{U}_n(I-V)\hat{\Lambda}^{-1}(I-V)\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2} \\ &\leq O\left(\frac{n}{p}\frac{1}{(1-v_{\max})^2}\frac{1}{e_i'\hat{U}_n\hat{U}_n'e_i}\right).\end{aligned} \tag{111}$$

Finally, we use (62), such that with high probability we have

$$\underline{c} \leq n\frac{e_i'\hat{U}_n(I-V)\hat{\Lambda}^{-1}(I-V)\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2} \leq \overline{c}. \tag{112}$$

Since this norm determines the variance of the estimator, this yields the required result. ■

B.5 Proof of Theorem 5

Order of bias term First, we need to show that

$$P\left(\left|\frac{1}{n}d_i^{\text{RI}}e_i'(X'X+\gamma I_p)^{-1}X'Xe_i - \delta_{ij}\right| > \tilde{a}\sqrt{\frac{\log p}{n}}\right) = O(p^{-\tilde{c}}), \tag{113}$$

with $\tilde{c}$ as in Theorem 4. The proof largely follows the strategy under random least squares. We first show that $(X'X+\gamma I_p)^{-1}X'X$ also satisfies the right-hand side of (97).

By substituting $X = \hat{V}\hat{S}\hat{U}$ and defining $\hat{\Lambda} = \hat{S}'\hat{S}$, we have

$$\begin{aligned}(X'X+\gamma I_p)^{-1}X'X &= (\hat{U}\hat{\Lambda}\hat{U}' + \gamma I_p)^{-1}\hat{U}\hat{\Lambda}\hat{U}' \\ &= \left(\hat{U}\left(\hat{\Lambda}-\gamma I_p\right)\hat{U}'\right)^{-1}\hat{U}\hat{\Lambda}\hat{U}' \\ &= \hat{U}_n\left(I_n - (\hat{\Lambda}_n - \gamma I_n)^{-1}\gamma I_n\right)\hat{U}_n' \\ &= \hat{U}_n(I-V)\hat{U}_n',\end{aligned} \tag{114}$$

where $\hat{\Lambda}_n$ is a diagonal matrix with on the diagonal the nonzero eigenvalues of $X'X$ and U_n consists of the first n rows of $\hat{U}$. It is clear that V is diagonal matrix with diagonal elements v_i

$$v_i = \frac{\gamma}{\hat{\lambda}_i + \gamma}. \tag{115}$$

Now as before, V should satisfy

$$||V||_2 = c_V \sqrt{\frac{\log p}{n}}. \tag{116}$$

Since V is diagonal

$$\begin{aligned} ||V||_2 &= \max_{i=1,\ldots,n} \frac{\gamma}{\hat{\lambda}_i + \gamma} \\ &= \frac{\gamma}{\min_i \hat{\lambda}_i + \gamma} \\ &= \frac{\gamma}{\frac{n}{c_Z c_\epsilon}\lambda_{\min}(\Sigma) + \gamma}. \end{aligned} \tag{117}$$

Equating (116) with (117) yields

$$\gamma \leq c_\gamma n \sqrt{\frac{\log p}{n}}, \tag{118}$$

where $c_\gamma = c_V c_Z c_\epsilon \lambda_{\min}(\Sigma)$, which completes the proof. ■

Order of the variance What remains to be shown is

$$\frac{d_i^{\mathrm{RI}}}{\sqrt{n}} e_i'(X'X + \gamma I_p)^{-1} X'\varepsilon = O(1). \tag{119}$$

This follows from the same argument as made for random least squares.

B.6 Proof of Theorem 6

This follows after some rewriting. Define the diagonal matrix $A = \mathrm{E}[R(R'\hat{\Lambda}R)^{-1}R']\hat{\Lambda}$, then

$$\begin{aligned} ||e_i\hat{U}\mathrm{E}[R(R'\hat{\Lambda}R)^{-1}R'X'||_2^2 &= e_i\hat{U}A\hat{\Lambda}^{-1}A\hat{U}'e_i \\ &= e_i\hat{U}\hat{\Lambda}^{-1/2}A^2_{\mathrm{RLS}}\hat{\Lambda}^{-1/2}\hat{U}'e_i, \end{aligned} \tag{120}$$

where A^2_{RLS} is a diagonal matrix with diagonal elements $0 \leq A^2_{ii} \leq 1$.

Similarly, for the ridge regularized inverse, we have

$$\begin{aligned} ||e_i(X'X + \gamma I_p)^{-1}X'||_2^2 &= e_i(X'X + \gamma I_p)^{-1}X'X(X'X + \gamma I_p)^{-1}e_i \\ &= e_i\hat{U}_n(\hat{\Lambda} + \gamma I_p)^{-2}\hat{\Lambda}\hat{U}_n'e_i \\ &= e_i\hat{U}_n\hat{\Lambda}^{-1/2}A^2_{\mathrm{RID}}\hat{\Lambda}^{-1/2}\hat{U}_n'e_i, \end{aligned} \tag{121}$$

with A^2_{RID} is a diagonal matrix with the diagonal elements satisfying $0 \leq A^2_{ii} \leq 1$. For the Moore-Penrose pseudoinverse we have

$$\begin{aligned} ||e_iX'(XX')^{-1}||_2^2 &= e_iX'(XX')^{-2}Xe_i \\ &= e_i\hat{U}\hat{\Lambda}^{-1}\hat{U}'e_i. \end{aligned} \tag{122}$$

And since for both RLS and RID A^2 is a diagonal matrix with the diagonal elements satisfying $0 \leq A^2_{ii} \leq 1$, the claim in Theorem 6 follows. ■

B.7 Proof of Lemma 2

When $u \sim i.i.d(0,1)$ and $\mathrm{E}\left\{u_t \left[\left(\frac{1}{p}XX'\right)^{-1} X\right]_t\right\} = 0$, a central limit theorem applies ensuring that as $n \to \infty$,

$$\frac{1}{\sqrt{n}} X' \left(\frac{1}{p}XX'\right)^{-1} u \xrightarrow{(d)} N\left[0, \lim_{n\to\infty} \frac{1}{n} X' \left(\frac{1}{p}XX'\right)^{-2} X\right], \tag{123}$$

and hence, as $n \to \infty$,

$$\frac{\sigma e_i' \frac{1}{\sqrt{n}} X' \left(\frac{1}{p}XX'\right)^{-1} u}{\left\| e_i' \frac{1}{\sqrt{n}} X' \left(\frac{1}{p}XX'\right)^{-1} \right\|_2} \xrightarrow{(d)} N(0, \sigma^2), \tag{124}$$

which completes the proof. ■

C Auxiliary definitions and lemma's

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

$$f_{H_W} = |\Sigma|^{-n/2} |H_W' \Sigma^{-1} H_W|^{-p/2}. \tag{125}$$

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

Lemma 3 (Reduction to standard normal random variables) *Define W 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), p. 29)*

$$U_n = W(W'W)^{-1/2}. \tag{126}$$

Lemma 4 (Manifold decomposition, Chikuse (2012)) *Let H be a $p \times n$ random matrix on the Stiefel manifold $V_{n,p}$, which is decomposed as*

$$H = (h_1, H_2), \tag{127}$$

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

$$h_1 = G(H_2)T, \tag{128}$$

where $G(H_2)$ is any $p \times p-n+1$ matrix chosen so that $[H_2, G(H_2)] \in \mathcal{O}(p)$, and T a $(p-n+1) \times 1$ vector. 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 5 (Wang and Leng (2015)) *Let H be a $p \times n$ random matrix on the Stiefel manifold $V_{n,p}$, which follows the $MACG(\Sigma)$. 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*

$$T|H_2 \sim ACG(G(H_2)'\Sigma G(H_2)). \tag{129}$$

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

$$\tilde{\Sigma} = G(H_2)G(H_2)'\Sigma G(H_2)G(H_2)', \tag{130}$$

then we also have

$$h_1|H_2 \sim ACG(\tilde{\Sigma}). \tag{131}$$

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

$$h_{i1}|H_2 \overset{(d)}{=} \frac{z_i}{\sqrt{z_1^2 + \ldots + z_p^2}}, \tag{132}$$

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

Lemma 6 (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*

$$e_1'HH'e_2 = e_1'HQQ'H'e_2, \quad Q \in \mathcal{O}(n). \tag{133}$$

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

$$e_1'HH'e_2 = e_1'\tilde{H}\tilde{H}e_2 = \tilde{h}_{11}\tilde{h}_{12}, \tag{134}$$

implying that

$$e_1'HH'e_2 \overset{(d)}{=} h_{11}h_{12} \Big| \{e_1'H = h_{11}\}. \tag{135}$$

Denote the first row of H by $h_1' = [h_{11}, h_{12}']$. Then $e_1'HH'e_1 = h_{11}^2 + h_2'h_2$ and thus $e_1'H = [h_{11}, 0_{1\times n-1}] \leftrightarrow e_1'HH'e_1 = h_{11}^2$. We can then rewrite (135) into

$$e_1HH'e_2 \overset{(d)}{=} h_{11}h_{12} \Big| \left\{e_1HHe_1 = h_{11}^2\right\}. \tag{136}$$

Lemma 7 (Berstein's inequality) *We will use the following identity on the sum of independent $\chi^2(1)$ variables*

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^{n}\chi_i^2(1) > 1+\epsilon\right) &\leq \exp\left(-\frac{\epsilon^2 n}{8}\right), \\ P\left(\frac{1}{n}\sum_{i=1}^{n}\chi_i^2(1) < 1-\epsilon\right) &\leq \exp\left(-\frac{\epsilon^2 n}{8}\right). \end{aligned} \tag{137}$$

D Variable Descriptions

Table 6: Variable Descriptions Table 5

variable	FRED mnemonic	Description
industrial production	IPFINAL	Industrial Production: Final Products (Market Group) (Index 2012=100)
employees wholesale trade	USWTRADE	All Employees: Wholesale Trade (Thousands of Persons)
hours worked business	HOABS	Business Sector: Hours of All Persons (Index 2009=100)
hours worked nonfarm	HOANBS	Nonfarm Business Sector: Hours of All Persons (Index 2009=100)
housing starts	HOUSTS	Housing Starts in South Census Region (Thousand of Units)
retail sales	RSAFSx	Real Retail and Food Services Sales (Millions of Chained 2009 Dollars), deflated by Core PCE
manufacturing inventories	NAPMII	ISM Manufacturing: Inventories Index
GDP deflator	GDPCTPI	Gross Domestic Product: Chain-type Price Index (Index 2009=100)
productivity nonfarm	OPHNFB	Nonfarm Business Sector: Real Output Per Hour of All Persons (Index 2009=100)
productivity business	OPHPBS	Business Sector: Real Output Per Hour of All Persons (Index 2009=100)
labour costs	ULCBS	Business Sector: Unit Labor Cost (index 2009=100)
rate commercial paper	CPF3MTB3Mx	3-Month Commercial Paper Minus 3-Month Treasury Bill, secondary market (Percent)
rate Eurodollar deposit	MED3TB3Mx	3-Month Eurodollar Deposit Minus 3-Month Treasury Bill, secondary market (Percent)
real money stock	MZMREALx	Real MZM (money of zero maturity) Money Stock (Billions of 1982-84 Dollars), deflated by CPI
consumer sentiment	UMCSENTx	University of Michigan: Consumer Sentiment (Index 1st Quarter 1966=100)
stock price volatility	VXOCLSX	CBOE S&P 100 Volatility Index: VXO
Federal debt	GFDEGDQ188S	Federal Debt: Total Public Debt as Percent of GDP (Percent)

Note: this table reports the variable descriptions and FRED mnemonics corresponding to the variables in Table 5.