TI 2017-032/III
Tinbergen Institute Discussion Paper

Inference in high-dimensional linear regression models

Revision: 05-07-2017

Tom Boot[1]
Didier Nibbering[2]

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

Inference in high-dimensional linear regression models

Tom Boot[*]
University of Groningen

Didier Nibbering[†]
Erasmus University Rotterdam
Tinbergen Institute

June 5, 2017

Abstract

We introduce an asymptotically unbiased estimator for the full high-dimensional parameter vector in linear regression models where the number of variables exceeds the number of available observations. The estimator is accompanied by a closed-form expression for the covariance matrix of the estimates that is free of tuning parameters. This enables the construction of confidence intervals that are valid uniformly over the parameter vector. Estimates are obtained by using a scaled Moore-Penrose pseudoinverse as an approximate inverse of the singular empirical covariance matrix of the regressors. The approximation induces a bias, which is then corrected for using the lasso. Regularization of the pseudoinverse is shown to yield narrower confidence intervals under a suitable choice of the regularization parameter. The methods are illustrated in Monte Carlo experiments and in an empirical example where gross domestic product is explained by a large number of macroeconomic and financial indicators.

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

[*] 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

We would like to thank Paul Bekker, Patrick Groenen, Trevor Hastie, Christiaan Heij, Richard Paap, Andreas Pick, and participants of the workshop on statistical learning and econometrics at the Erasmus University Rotterdam for helpful comments.

1 Introduction

The increasing number of economic indicators confronts researchers with data where the number of explanatory variables approaches, and often even exceeds, the number of available observations. This is commonly observed in cross-sectional data 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). The ratio of observations to parameters is even smaller in studies on the relation between the human genome and later in life outcomes such as educational attainment by Rietveld et al. (2013).

Estimation of high-dimensional models has been intensively studied in recent years. Well-known estimators include ridge regression (Hoerl and Kennard, 1970), lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006), Dantzig selector (Candes and Tao, 2007), and penalized likelihood methods by (Fan et al., 2004). An overview of theoretical results is provided by Bühlmann and Van De Geer (2011). The adequacy of these estimators is argued through accuracy bounds measuring the difference between the true and estimated parameter vector. The distribution of these estimators is however intractable, and the construction of standard errors and valid confidence intervals remains a challenging problem.

In this paper, we develop an asymptotically unbiased estimator for the full high-dimensional parameter vector in a linear regression model where the number of variables p greatly exceeds the number of observations n. The estimator is accompanied by a closed form expression for the covariance matrix of the estimated parameters which is free of tuning parameters. This enables the construction of uniformly valid confidence intervals, hypothesis testing, estimation of forecast uncertainty, and efficient adjustments for multiple testing which fully take the dependence between the estimates into account. Standard errors are shown to decrease at the familiar $n^{-1/2}$ rate.

The estimator uses a diagonally scaled Moore-Penrose pseudoinverse to obtain parameter estimates, and implements a bias correction based on the lasso. The scaled Moore-Penrose pseudoinverse approximates the inverse of the singular high-dimensional covariance matrix of the regressors, and the lasso corrects for the bias resulting from this approximation. The remaining bias can be factorized into a term which reflects the accuracy of the pseudoinverse, and a term measuring the lasso estimation error. The product of these two components is of lower order compared to the variance of the estimator, yielding an asymptotically unbiased estimator. The proof relies on several extensions of the results of Fan and Lv (2008) and Wang and Leng (2015), who use the Moore-Penrose pseudoinverse to set up a variable screening technique.

Using the Moore-Penrose pseudoinverse is especially effective when the number of variables is much larger than the number of observations. If p is relatively close to n, regularization of the inverse can reduce the standard errors while the bias remains negligible. This motivates an extension to two regularized variants of the Moore-Penrose pseudoinverse; random least squares and ridge regularization. For a suitable choice of the regularization parameters, these estimators yield smaller standard errors while maintaining the same theoretical validity.

Random least squares projects the columns of the regressor matrix onto a low-dimensional subspace by post-multiplying with a matrix with independently standard normally distributed elements. Repeatedly applying this procedure yields an estimate of the full parameter vector. Mean squared error properties of this estimator are studied by Maillard and Munos (2009) based on the lemma by Johnson and Lindenstrauss (1984), and refined by Kabán (2014). We show that random least squares results in a form of generalized ridge regularization on the empirical covariance matrix. The regularization strength is inversely related to the projection dimension, which for inferential procedures should be chosen close to the sample size.

The second regularization method we consider is ridge regularization. In order to show that the bias of the estimator remains sufficiently small, we exploit the relation of the ridge regularized estimator to the Moore-Penrose inverse when the regularization strength is small. This extends the results of Bühlmann et al. (2013), who uses the ridge estimator to construct conservative p-values, and Wang and Leng (2015) who focus on variable screening.

The results depend on a sparsity assumption with regard to the high-dimensional parameter vector, and a mild restriction on the distributional class of the regressor matrix. We assume the sparsity of the parameter vector to be of the same order as in recent studies on high-dimensional inference by Zhang and Zhang (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014). Furthermore, we require the rows of the regressor matrix to be generated from the class of elliptical distributions. This class includes the multivariate normal, power exponential and Student's t-distribution. To facilitate applications to time series data or data with cross-sectional dependence, we allow for correlation between and within the regressors. Results are provided for both gaussian and non-gaussian regression errors.

Our approach to high-dimensional inference builds upon Zhang and Zhang (2014), van de Geer et al. (2014), and Javanmard and Montanari (2014). Under an additional sparsity assumption on the elements of the inverse covariance matrix, Zhang and Zhang (2014) and van de Geer et al. (2014) use the lasso for each column of the regressor matrix to estimate the inverse covariance matrix. As an alternative, Javanmard and Montanari (2014) rely on direct numerical optimization to find an accurate approximate inverse. Both methods lead to standard errors which depend on one or more additional regularization parameters that potentially influence the results.

We consider situations where interest lies in performing inference on the full high-dimensional parameter vector. Alternatively, one can focus on a low-dimensional subvector of the high-dimensional parameter vector. A sequence of papers (Belloni et al., 2013, 2010; Chernozhukov et al., 2015) introduces a multistage procedure that uses the lasso to select control variables in such a way that variable selection errors do not affect the distribution of the estimates of interest. This approach is effective when both the number of control variables related to the dependent variable, as well as the number of control variables related to the variables of interest, are limited. Strengthening this assumption such that every variable is correlated with only a small number of the remaining variables, Lan et al. (2016) provide a method to construct confidence intervals for the full parameter vector.

In this paper, we do not limit inference to a low-dimensional subvector of the parameter vector. The proposed method relaxes the assumption that only a small number of control variables is related to the variables of interest. This relaxation might come at the cost of a potential power loss, although this is not reflected in the rate at which the standard errors decrease as the sample size tends to infinity.

We confirm our theoretical results with a set of Monte Carlo experiments. We vary the specification of the covariance matrix, the amount of sparsity of the parameter vector, and the signal strength. In line with the theoretical results, we find that even in small samples where the number of regressors is twice the number of observations, coverage rates are close to the nominal rate of 95%. Random least squares and ridge regression yield narrower confidence intervals compared to using a Moore-Penrose pseudoinverse, but this comes at the expense of a slight downward bias. Coverage rates are substantially closer to the nominal rate compared to existing methods.

We apply the methods to the FRED-QD, a quarterly data set consisting of 254 macroeconomic and financial series of the United States economy, available from the second quarter of 1987 onwards. We analyze the relation between the real gross domestic product and the other variables provided in this data set in a linear regression framework. Although the number of regressors 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 introduces the estimation approach and the proposed estimators. The theoretical properties of the Moore-Penrose pseudoinverse, random least squares, and ridge regression are presented in Section 3. Section 4 illustrates these results through Monte Carlo simulations and Section 5 applies the methods on the FRED-QD dataset. Section 6 concludes.

Notation We use the following notation throughout the paper: For any $n \times 1$ vector $a = (a_1, \dots, a_n)'$, 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 $||a||_\infty = \max(|a_1|, \dots, |a_n|)$. For a $p \times n$ matrix A, the l_q-norm is defined as $||A||_q := \sup_{x, ||x||_q = 1} \{||Ax||_q\}$ and the maximum norm is written as $||A||_{\max} = \max_{i=1,\dots,n, j=1,\dots,p} |A_{ij}|$. The $n \times n$ identity matrix is denoted by I_n. The vector e_i has its i-th entry equal to 1 and zeros everywhere else. For the regressor matrix X, we index the rows with the subscript $i = 1, \dots, n$ and the columns with the subscript $j = 1, \dots, p$. If U is a $p \times p$ orthogonal matrix, we write $U \in \mathcal{O}(p)$. When two random variables X and Y follow the same distribution, this is denoted as $X \overset{(d)}{=} Y$.

2 High-dimensional linear 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, $\beta = (\beta_1, \dots, \beta_p)'$ a $p \times 1$ vector of unknown regressor coefficients, and ε an $n \times 1$ vector of errors which are independent and normally distributed with variance σ^2. The empirical covariance matrix of X is denoted by $\hat{\Sigma} = \frac{1}{n} X'X$. We will show how the normality assumption on the errors can be relaxed.

2.1 Approximate inverse and bias correction

Define M as a $p \times n$ matrix for which MX is close to the $p \times p$ identity matrix I_p, in a sense that will be made precise below. We refer to M as an approximate inverse for X.

We start by considering estimators for β of the form

$$\begin{aligned}\hat{\beta} &= My \\ &= MX\beta + M\varepsilon \\ &= \beta + (MX - I_p)\,\beta + M\varepsilon.\end{aligned} \tag{2}$$

The second term of (2) represents a bias which depends on the accuracy of the approximate inverse M. When $p \leq n$, ordinary least squares yields unbiased estimates by choosing $M = (X'X)^{-1}X'$. When $p > n$, the matrix $X'X$ is singular, and we have to resort to an expression for M for which the bias is not equal to zero.

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

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

For the initial estimator $\hat{\beta}^{\text{init}}$ we take the lasso estimator of Tibshirani (1996). Alternative initial estimators can be used, as long as they satisfy a sufficiently tight accuracy bound on the l_1 norm of $\beta - \hat{\beta}^{\text{init}}$.

The goal of this paper is to introduce choices of M such that the bias of the estimator $\hat{\beta}^{c}$ is of lower order than the variance. Anticipating the usual $\sqrt{n}$ rate of convergence, we rescale the estimator in (3) as

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}^{c} - \beta\right) &= \Delta + Z \\ \Delta &= \sqrt{n}\,(MX - I_p)\left(\beta - \hat{\beta}^{\text{init}}\right) \\ Z &= \sqrt{n}M\varepsilon\end{aligned} \tag{4}$$

The term Δ reflects the bias of the corrected estimator. To ensure asymptotic unbiasedness, Δ should be of lower order than the noise term Z. We propose specifications for the approximate inverse M for which $Z|X \sim N(0, \sigma^2\Omega)$ with $\Omega = nMM'$ and the variance $\Omega_{jj} = O_p(1)$. This shows that the standard errors of the estimator $\hat{\beta}^{c}$ decrease at the familiar $n^{-1/2}$ rate.

In order for the bias to vanish compared to the variance term, given that $\Omega_{jj} = O_p(1)$, we now need $||\Delta||_\infty = o_p(1)$. Under a sparsity assumption on β, we show that this is indeed the case, which implies that $\hat{\beta}^c$ is an asymptotically unbiased estimator. Combined with a closed-form expression for the covariance matrix Ω, confidence intervals can be constructed for the j-th parameter as

$$\left[\hat{\beta}^c_j - z_{\alpha/2}\sqrt{\sigma^2\Omega_{jj}}, \quad \hat{\beta}^c_j + z_{\alpha/2}\sqrt{\sigma^2\Omega_{jj}}\right], \tag{5}$$

where $z_{\alpha/2}$ is the $\alpha/2$ critical value for the standard normal distribution. We discuss estimation of σ in Section 2.3.

The estimator defined in (3) occurs in a different form in Zhang and Zhang (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014), who consider $\hat{\beta}^c = \hat{\beta}^{\text{lasso}} + \frac{1}{n}\bar{M}X'(y - X\hat{\beta}^{\text{lasso}})$. This leads to an interpretation of $\hat{\beta}^c$ as a 'desparsified' version of the lasso estimator. An alternative to the standard lasso estimator is put forward by Caner and Kock (2014). The matrix $\bar{M}$ serves as an approximate inverse to the empirical covariance matrix $\frac{1}{n}X'X$, which is found by a series of lasso regressions in Zhang and Zhang (2014) and van de Geer et al. (2014), or direct numerical optimization in Javanmard and Montanari (2014). As a consequence of the complex estimation procedures, standard errors are not available in closed form, and their validity depends on the appropriate selection of one or more tuning parameters.

2.2 Choosing the approximate inverse M

This section proposes specifications of M for which the bias $||\Delta||_\infty$ in (4) is small. We ensure that the diagonal terms of $MX - I_p$ are identically equal to zero by introducing a $p \times p$ diagonal matrix D, with diagonal elements d_j, and taking

$$M = D\tilde{M}, \qquad d_j = (\tilde{m}_j' x_j)^{-1}, \tag{6}$$

with $\tilde{m}_j'$ the j-th row of $\tilde{M}$. We first choose M in the form defined in (6), with $\tilde{M}$ specified as the Moore-Penrose pseudoinverse of X. Subsequently, we consider regularized alternatives obtained by random least squares and ridge regression.

2.2.1 The Moore-Penrose pseudoinverse

A tuning parameter free choice for $\tilde{M}$ in (6) is the Moore-Penrose pseudoinverse. When $p \leq n$, and the columns of X are linearly independent, $\tilde{M} = (X'X)^{-1}X'$. In the high-dimensional setting where $p > n$, the matrix X has linearly dependent columns by default. In this case the pseudoinverse equals $X'(XX')^{-1}$, and

$$M^{\text{MPI}} = D^{\text{MPI}} X'(XX')^{-1}. \tag{7}$$

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

$$d_j^{\text{MPI}} = \left[x_j'(XX')^{-1}x_j\right]^{-1}. \tag{8}$$

This provides a closed-form expression for the approximate inverse. In addition, since the bias term of the estimator is of lower order compared to the variance, the covariance of $\hat{\beta}^c$ is available in closed form as well,

$$V(\hat{\beta}^c) = D^{\text{MPI}} X'(XX')^{-2} X D^{\text{MPI}}. \tag{9}$$

2.2.2 Regularizing the Moore-Penrose pseudoinverse

The accuracy of the Moore-Penrose pseudoinverse depends on the concentration of the eigenvalues of the matrix XX', which can be weak when p is close to n. Regularizing the approximate inverse can improve in accuracy, with smaller standard errors as a result. This section introduces two regularization techniques, for which Section 3 shows the appropriate choice for the regularization strength.

Random Least Squares 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,

$$R_{jl} \sim N(0,1), \quad j = 1, \dots p, \quad l = 1, \dots, k. \tag{10}$$

The multiplication yields a low-dimensional analogue to (1),

$$y = XR\gamma_R + u. \tag{11}$$

Least squares estimation of γ_R is straightforward as

$$\hat{\gamma}_R = (R'X'XR)^{-1}R'X'y, \tag{12}$$

from which an estimator for β can be constructed by $\hat{\beta}_R = R\hat{\gamma}_R$. Since R is random, Jensen's inequality can be used to show that the accuracy of this estimator can be improved by averaging over different realizations of R. We then arrive at the following estimator of β,

$$\hat{\beta}_{\bar{R}} = \mathrm{E}_R[R\hat{\gamma}_R] = \mathrm{E}_R[R(R'X'XR)^{-1}R']X'y. \tag{13}$$

From equation (13), we recognize that random least squares yields an approximate inverse covariance matrix of X. Defining $\tilde{M} = \mathrm{E}_R[R(R'X'XR)^{-1}R']X'$ in (6) yields

$$M^{\mathrm{RLS}} = D^{\mathrm{RLS}}\mathrm{E}_R\left[R(R'X'XR)^{-1}R'\right]X', \tag{14}$$

with

$$d_j^{\mathrm{RLS}} = \left\{\mathrm{E}_R[r_j'(R'X'XR)^{-1}R']X'x_j\right\}^{-1}. \tag{15}$$

Ridge regression An alternative regularization strategy is to use a ridge adjustment,

$$M^{\mathrm{RID}} = D^{\mathrm{RID}}(X'X + \gamma I_p)^{-1}X', \tag{16}$$

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

$$d_j^{\mathrm{RID}} = \left(v_j'X'x_j\right)^{-1}, \tag{17}$$

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

The regularization in (16) can be related to the Moore-Penrose pseudoinverse, since the latter is defined as

$$\begin{aligned} X'(XX')^{-1} &= \lim_{\gamma\to 0}\left(X'X + \gamma I_p\right)^{-1}X' \\ &= \lim_{\gamma\to 0} X'\left(XX' + \gamma I_n\right)^{-1}. \end{aligned} \tag{18}$$

which can be shown using the singular value decomposition of X as in Albert (1972).

2.3 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 which holds that $\left|\frac{\hat{\sigma}}{\sigma} - 1\right| = o_p(1)$ under Assumption A1 and Assumption A2 discussed in Section 3.1.

However, in the Monte Carlo simulations in Section 4, 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

$$\hat{\sigma}^2_{\text{lasso}} = \frac{1}{n - \hat{s}} \hat{\varepsilon}' \hat{\varepsilon}, \tag{19}$$

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.

3 Theoretical results

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

3.1 Assumptions

Performing inference in a linear regression model with more variables than observations requires additional assumptions over its low-dimensional counterpart. Our assumptions parallel Fan and Lv (2008) and Wang and Leng (2015). We will provide a discussion below.

A1. *The sparsity $s_0 = ||\beta||_0$ satisfies $s_0 = o\left(\frac{\sqrt{n}}{\log p}\right)$.*

A2. *The regressor matrix X is generated from an elliptical distribution, i.e.*

$$X = \Sigma_1^{1/2} Z \Sigma_2^{1/2} = \Sigma_1^{1/2} V S U' \Sigma_2^{1/2}, \tag{20}$$

where the $n \times n$ population covariance matrix Σ_1 and the $p \times p$ population covariance matrix Σ_2 determine the dependence between the rows and columns of X, respectively. The elements of the $n \times p$ matrix Z are generated independently from a spherically symmetric distribution, $V \in \mathcal{O}(n)$, S is an $n \times p$ matrix of singular values, and $U \in \mathcal{O}(p)$.

Furthermore,

$$P\left(\lambda_{\max}(p^{-1}ZZ') \geq c_Z, \quad \lambda_{\min}(p^{-1}ZZ') \leq c_Z^{-1}\right) \leq e^{-C_Z n}, \tag{21}$$

where $\lambda_{\max}(.)$ and $\lambda_{\min}(.)$ are the largest and smallest eigenvalues of a matrix respectively, and c_Z, C_Z are positive constants.

A3. *For both the population covariance matrices Σ_1 and Σ_2, the eigenvalues are bounded by a constant, i.e. for $i = 1, 2$,*

$$0 < c_{i,1} \leq \lambda_{\min}(\Sigma_i) \leq \lambda_{\max}(\Sigma_i) \leq c_{i,2} < \infty. \tag{22}$$

Assumption A1 imposes a sparsity constraint which restricts the number of non-zero coefficients in β by $s_0 = ||\beta||_0$. For lasso consistency, it is required that $s_0^2 = o(n/\log p)$. As noted in van de Geer et al. (2014) and Javanmard and Montanari (2014), a slightly stronger assumption is needed when constructing confidence intervals.

In recent work, for example by Chernozhukov et al. (2015), assumption A1 is relaxed to allow for approximate sparsity, arguably a more realistic assumption in practical applications. This restricts only the number of large non-zero coefficients, and allows the remaining coefficients to be sufficiently small. Since our results only depend on the l_1 norm of the lasso estimation error, which does not change under approximate sparsity, they remain valid under approximate sparsity.

Assumption A2 requires that the regressors are generated from an elliptical distribution. The class of elliptical distributions includes the multivariate normal distribution, but also allows for heavier tailed distributions such as the power exponential distribution and the multivariate t distribution (Serfling, 2006; Dasgupta et al., 2012). This class precludes X to consist of binomial variables. However, our results rely on the distribution of the elements of $X'(XX')^{-1}X$, which consist of sums of binomial variables. It is possible that one can use the convergence of these sums towards a normal distribution to extend the results towards binomial regressors.

The matrices Σ_1 and Σ_2 in Assumption A2 allows for dependence between the rows and the columns of X, respectively. Assumption A3 states that the eigenvalues of these population covariance matrices are finite and independent of the dimensions n and p. This assumption can be relaxed by replacing $c_{i,2}$ with $c_{i,2}n^{\alpha}$. The standard errors then decrease at the rate of $1/\sqrt{n^{1-\alpha}}$ instead of $1/\sqrt{n}$.

3.2 Asymptotic unbiasedness and normality using the Moore-Penrose pseudoinverse

To prove that $\hat{\beta}^c$ in (3) based on the Moore-Penrose pseudoinverse is an asymptotically unbiased estimator, we show that with high probability the bias term in (4) is small and of lower order than the noise. Moreover, the construction of confidence intervals as in (5) requires $Z|X$ to follow a normal distribution. Efficiency of the estimator is ensured by showing that the standard errors decrease at the usual $n^{-1/2}$ rate.

The first requirement follows from bounding the bias term of the estimator in

(4) by a norm inequality,

$$||\Delta||_\infty \leq \sqrt{n}\, ||MX - I_p||_{\max}\, ||\beta - \hat{\beta}^{\text{init}}||_1, \tag{23}$$

which is an element-wise bound on $MX - I_p$ together with an l_1 accuracy bound on $\beta - \hat{\beta}^{\text{init}}$.

The following lemma bounds on the first term in probability.

Lemma 1. *Suppose Assumption A2 and A3 hold. Define* $M^{MPI} = D^{MPI} X'(XX')^{-1}$ *with* D^{MPI} *a diagonal matrix with elements* $d_j^{MPI} = (x_j'(XX')^{-1}x_j)^{-1}$*, then we have*

$$P\left(||M^{MPI}X - I_p||_{\max} \geq a\sqrt{\frac{\log p}{n}}\right) = O(p^{-\tilde{c}}), \tag{24}$$

with $\tilde{c} = \frac{c}{2c_s}a^2 - 2$ *where* $a, c, c_s > 0$*.*

A proof is presented in Appendix B.1. Note that the diagonal elements of $M^{\text{MPI}}X - I_p$ are identically zero, due to the diagonal scaling with D^{MPI}. Lemma 1 is therefore a statement on the off-diagonal elements of $M^{\text{MPI}}X - I_p$.

Next we show that the l_1 norm of the initial estimation error, in the second term in the bound for $||\Delta||_\infty$ in (23), is bounded with high probability. As the initial estimator we use the lasso estimator by Tibshirani (1996), which is defined as

$$\hat{\beta}^{\text{lasso}} = \arg\min_b \left[\frac{1}{n}(y - Xb)'(y - Xb) + \lambda||b||_1\right]. \tag{25}$$

The following bound applies to the l_1-error of the lasso estimator.

Lemma 2. *Suppose Assumption A1 and Assumption A2 hold. Consider the lasso estimator* (25) *with* $\lambda \geq 8\sigma\sqrt{\frac{\log p}{n}}$*, then with probability exceeding* $1 - 2p^{-1}$ *we have*

$$\left|\left|\beta - \hat{\beta}^{lasso}\right|\right|_1 = O_p\left(s_0\sqrt{\frac{\log p}{n}}\right). \tag{26}$$

A proof is presented in Appendix B.2. As shown in Bühlmann and Van De Geer (2011), this bound applies under a so-called compatibility condition on X. The proof amounts to showing that the compatibility condition is indeed satisfied under Assumption A1 and Assumption A2.

Combining Assumption A1, Lemma 1, and Lemma 2, we see that the bias can be bounded by

$$||\Delta||_\infty = O_p\left(s_0\frac{\log p}{\sqrt{n}}\right) = o_p(1). \tag{27}$$

In order for the estimator to be asymptotically unbiased, it is necessary that the bias in (27) is of lower order than the noise term of the estimator, given by Z in (4). The following lemma states that this is indeed the case.

Lemma 3. *Suppose Assumption A2 and A3 hold. For $j = 1, \dots, p$ we have*

$$\begin{aligned} Z_j &= \sqrt{n} d_j^{MPI} x_j'(XX')^{-1}\varepsilon, \\ Z_j|X &\sim N(0, \sigma^2 \Omega_j), \\ ||\Omega_{jj}||_2 &= O_p(1), \end{aligned} \tag{28}$$

where $\Omega_{jj} = n m_j' m_j$ with m_j' the j-th row of $M^{MPI} = D^{MPI} X'(XX')^{-1}$ and D^{MPI} a diagonal matrix with $d_j^{MPI} = [x_j'(XX')^{-1}x_j]^{-1}$.

A proof is presented in Appendix B.3. Appendix B.4 shows that under additional assumptions this result also holds for independent and identically distributed errors ε_i.

Combining Lemma 3 with (27) yields the central theorem of this paper.

Theorem 1. *Suppose A1-A3 hold. Let $\hat{\beta}^c = My - (MX - I_p)\,\hat{\beta}^{init}$, with $\hat{\beta}^{init}$ such that $||\hat{\beta}^{init} - \beta||_1 = O_p\left(s_0\sqrt{\log(p)/n}\right)$, and take M as*

$$M^{MPI} = D^{MPI} X'(XX')^{-1},$$

where D^{MPI} is a diagonal matrix with elements $d_j^{MPI} = n\left[x_j'(XX')^{-1}x_j\right]^{-1}$. Then,

$$\begin{aligned} \sqrt{n}(\hat{\beta}^c - \beta) &= Z + o_p(1), \\ Z|X &\sim N\left(0, \sigma^2\Omega\right), \end{aligned}$$

where $\Omega = n M^{MPI} M^{MPI'}$ and $\Omega_{jj} = O_p(1)$.

This theorem shows that the estimator $\hat{\beta}^c$ in (3) is asymptotically unbiased with covariance matrix Ω, and standard errors that decrease at the usual $n^{-1/2}$ rate. Theorem 1 allows for the construction of confidence intervals that are uniformly valid over j. Uniformity is guaranteed since the bound on the lasso estimator given in Lemma 2 holds uniformly over all sets S_0 of size $s_0 = o(\sqrt{n}/\log p)$, see van de Geer et al. (2014) for a discussion.

Since the resulting covariance matrix of the estimator is available in closed form, efficient multiple testing procedures as in Bühlmann et al. (2013) can be employed, together with joint tests on estimated coefficients, as well as confidence intervals around predictions for future values of the dependent variable.

3.3 Regularized approximate inverse

When the number of variables is of the same order as the number of observations, the concentration of the eigenvalues in Assumption A2 might not be very tight. In this case, regularization of the pseudoinverse can increase the accuracy. We therefore analyze two regularization approaches.

Random least squares The key to the behavior of the regularized covariance matrix in repeated least squares, is the projection dimension k. The following lemma parallels Lemma 1 and Lemma 3 for an appropriate choice of the projection dimension.

Lemma 4. *Define* $M^{RLS} = D^{RLS} E_R \left[R(R'X'XR)^{-1}R'\right] X'$ *where* D^{RLS} *is a diagonal matrix with diagonal elements* $d_j^{RLS} = \left\{E_R[r_j'(R'X'XR)^{-1}R']X'x_j\right\}^{-1}$*, and* R *a* $p \times k$ *matrix with normally and independently distributed entries. Choose the projection dimension* k *as*

$$k = \left(1 - c_\kappa \sqrt{(\log p)/n}\right)(n-1), \tag{29}$$

where c_k *is a positive constant.*

Then we have

$$P\left(||M^{RLS}X - I_p||_{\max} \geq a\sqrt{\frac{\log p}{n}}\right) = O\left(p^{-\tilde{c}}\right), \tag{30}$$

with $\tilde{c}$ *as in Lemma 1 with* a *replaced by* $\tilde{a} < a$*. Furthermore, for* $Z = \sqrt{n} d_j^{RLS} E[r_j(R'X'XR)^{-1}R']X'\varepsilon$*, we have*

$$\begin{aligned} Z|X &\sim N(0, \sigma^2 \Omega^{RLS}), \\ \Omega^{RLS} &= nM^{RLS}M^{RLS'}, \\ \Omega_{jj}^{RLS} &= O_p(1). \end{aligned} \tag{31}$$

The proof of Lemma 4 given in Appendix B.5 relies on showing that when k is sufficiently close to n, the regularized inverse approximates the Moore-Penrose inverse. The results from Section 3.2 can then be used to show that regularizing using random least squares does not adversely affect the bias. The proof of Lemma 4 also elicits that random least squares is equivalent to a generalized form of ridge regression, where the regularization strength is dependent on the eigenvalues of the regressor matrix X. Details on the constant c_k are provided in the proof.

Ridge regularization Because of the relation between the Moore-Penrose pseudoinverse and ridge regularized covariance matrices displayed in (18), intuition suggests that for a sufficiently small penalty parameter λ, the results under a Moore-Penrose inverse carry over to a ridge adjusted estimator. The following lemma formalizes this intuition.

Lemma 5. *Define $M^{RID} = D^{RID}(X'X + \gamma I_p)^{-1}X'$, with the elements of the diagonal scaling matrix D^{RID} equal to $d_j^{RID} = \left(e_j'(X'X + \gamma I_p)^{-1}X'x_j\right)^{-1}$. If the ridge penalty parameter satisfies $\gamma \leq c_\gamma p\sqrt{\frac{\log p}{n}}$, where c_γ is a positive constant, then we have*

$$P\left(||M^{RID}X - I_p||_\infty \geq a\sqrt{\frac{\log p}{n}}\right) = O\left(p^{-\tilde{c}}\right), \tag{32}$$

with $\tilde{a}$ and $\tilde{c}$ as in Lemma 4.

Furthermore, for $Z = \sqrt{n}d_j^{RID}(X'X + \gamma I_p)^{-1}X'\varepsilon$, we have

$$\begin{aligned} Z|X &\sim N(0, \sigma^2\Omega^{RID}), \\ \Omega^{RID} &= nM^{RID}M^{RID'}, \\ \Omega_{jj}^{RID} &= O_p(1). \end{aligned} \tag{33}$$

A proof is provided in Appendix B.6, which also gives a more detailed description of the constant c_γ.

Inference using a regularized approximate inverse Using Lemma 4 and Lemma 5, we arrive at the following theorem for the regularized estimators.

Theorem 2. *Suppose A1-A3 hold. Let $\hat{\beta}^c = My - (MX - I_p)\,\hat{\beta}^{init}$, with $\hat{\beta}^{init}$ such that $||\hat{\beta}^{init} - \beta||_1 = O_p\left(s_0\sqrt{\log(p)/n}\right)$, and take M as either $M^{RLS} = D^{RLS}E_R\left[R(R'X'XR)^{-1}R'\right]X'$ or $M^{RID} = D^{RID}(X'X + \gamma^* I_p)^{-1}X'$, where the elements of the diagonal matrices D are defined in Lemma 4 and Lemma 5, R is a $p \times k^*$ matrix with independent standard normal entries, $k^* = k$ as in Lemma 4, and $\gamma^* = \gamma$ as in Lemma 5. Then,*

$$\begin{aligned} \sqrt{n}(\hat{\beta}^c - \beta) &= Z + o_p(1), \\ Z|X &\sim N\left(0, \sigma^2\Omega\right), \\ \Omega &= nMM', \\ \Omega_{jj} &= O_p(1). \end{aligned}$$

This theorem follows directly from Lemma 4 and Lemma 5. It confirms that when k is close to n and γ is sufficiently small, the estimator in (3) is asymptotically unbiased with covariance matrix Ω, and standard errors that decrease at the usual $n^{-1/2}$ rate.

The reason one would opt for the regularized variants despite the additional tuning parameters is provided by the following theorem. Here we compare the variance of Z in equation (4) for the different estimators.

Theorem 3. *Denote the variance of the estimator $\hat{\beta}_j^c$ under a diagonal scaling matrix D by $\Omega_{jj}(D)$. For the choice of k as in Lemma 4, or γ as in Lemma 5, we have*

$$\Omega_{jj}(D)^{RLS} - \Omega_{jj}(D)^{MPI} \leq 0, \qquad \Omega_{jj}(D)^{RID} - \Omega_{jj}(D)^{MPI} \leq 0. \tag{34}$$

The proof is given in Appendix B.7.

Note that Theorem 3 requires the regularized estimator and the estimator based on the Moore-Penrose pseudoinverse to use the same diagonal scaling matrix. Using D^{MPI} for the Moore-Penrose inverse, D^{RLS} for the repeated least squares estimator, and D^{RID} for the ridge regularized inverse, does not yield an ordering in terms of power. However, in all cases we have encountered, the inequality in Theorem 3 is satisfied when using the diagonal matrix specific to the estimator under consideration. This is also evident from the Monte Carlo results in Section 4.

3.4 Consistency

Although our focus in this paper is on the construction of confidence intervals, the estimator $\hat{\beta}^c$ can be shown to be consistent when we restrict the growth rate of the number of variables relative to the number of observations.

A4. *The number of variables grows near exponentially with the number of observations, i.e.*

$$\frac{\log p}{n} = o(1). \tag{35}$$

Since Z_i is (asymptotically) normal, we have that $\max_{i=1,\dots,j} |Z_i| = O_p(\sqrt{\log p})$. Since $\hat{\beta}^c = \beta + \frac{1}{\sqrt{n}}(\Delta + Z)$, Assumption A4 then guarantees that $\lim_{n\to\infty} \hat{\beta}^c = \beta$.

If one is only interested in consistency, then Assumption A2 can potentially be relaxed. In that case the bias is not required to be of lower order compared to the variance.

4 Monte Carlo Experiments

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

4.1 Monte Carlo set-up

Data generating process The data generating process takes the form

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

where y is an $n \times 1$ vector, X an $n \times p$ 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 Σ:

$$\text{Equicorrelated: } \Sigma_{jk} = 0.8, \quad \forall j \neq k, \quad \Sigma_{jj} = 1 \quad \forall j, \tag{37}$$

$$\text{Toeplitz: } \Sigma_{jk} = 0.9^{|j-k|}, \quad \forall j, k. \tag{38}$$

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 vary signal strength b, sparsity s, and covariance matrix Σ across different Monte Carlo experiments.

To align the simulation experiment with the setting in the economic application of Section 5, we set the number of predictors $p = 200$ and the sample size $n = 100$. In each replication the predictors in X and the coefficients in β are generated. We report average results for nonzero coefficients and zero coefficients, based on 1000 replications of the data generating process in (36).

Estimation We use (3) to 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. The random least squares estimator averages over $N = 1000$ realizations of the regularized covariance matrix and projects onto a subspace dimension with $k = 90$. The ridge regression based estimator sets its penalty parameter as $\gamma = 1$, following Bühlmann et al. (2013).

The proposed estimators are compared to three existing 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) is equivalent to this method for linear regression problems considered here. 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 again 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 (19).

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.

4.2 Simulation Results

4.2.1 Sparsity and signal strength

Table 1 shows the Monte Carlo simulation results for the set of experiments with an equicorrelated covariance matrix and Table 2 with a Toeplitz covariance matrix. The tables report the estimated coefficients, standard errors, coverage rates, and power of the 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

Table 1: 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). Results are based on 1000 replications of the linear model (36), with equicorrelated regressors as in (37). Results are provided separately for non-zero ($b \neq 0$) and zero ($b = 0$) coefficients. The number of observations is $n = 100$ and the number of regressors $p = 200$. The subspace dimension in RLS is $k = 0.9n$, we average over $N = 1000$ low-dimensional projections, and the penalty parameter for ridge regression is $\gamma = 1$. We vary the number ($s = 3, 15$) and signal strength ($b = 2, 5$) of nonzero coefficients.

are smaller, and the coverage rate is very close to the nominal rate.

We find that 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 though the number of

Table 2: 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 (38). For additional information, see the note following Table 1.

variables is twice as large as the number of observations, the proposed methods achieve nontrivial 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 coefficients are more precisely estimated when there are relatively few of them. For all methods, the coefficients which are set to zero in the data generating process are estimated very close to zero.

Random least squares produces the most efficient estimates relative to ridge regression and Moore-Penrose pseudoinverse regression. Standard errors of the random least squares estimates are lower than these estimators in all experiments. Ridge is again a more efficient estimator relative to the pseudo-inverse, in line with Theorem 3. 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 substantially 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 in combination with 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, their simulation study considers a low-dimensional setting, 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 necessary conditions underlying the validity of CPS.

4.2.2 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 3 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 gen-

Table 3: 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 1.

erating 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 1 and 2. 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.

Estimation of the noise level The validity of confidence intervals depends on a consistent estimator of the noise level σ^2. Appendix C 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 1 and 2 are based on the estimator for the noise level σ^2 as defined in (19).

5 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. economy and a large number of macroeconomic and financial indicators.

5.1 Data

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

5.2 Estimation

The coefficients β are estimated in the regression equation

$$y = Z\delta + X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \tag{39}$$

where y equals the real gross domestic product of the U.S. economy (FRED mnemonic GDPC96), Z includes an intercept along with four lags of the quarterly dependent variable y, and X consists of the remaining variables in the database which are not in the same group as y. Since we are only interested in the macroeconomic relations in β, we partial out the variables in Z using the Frisch-Waugh theorem before estimating β. We note that Assumption A2 and Assumption A3 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 $n - n_z = 105$.

The high-dimensional regression theory in Section 3 allows both the rows and the columns of X^* to be correlated. The assumption that the rows of X^* are generated from the class of elliptical distributions only excludes very erratic behavior, for which we account by the data transformations described in McCracken and Ng (2016). Finally, a sparsity assumption on β imposes that not all variables in X influence y.

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 $\gamma = 1$ as in Bühlmann et al. (2013).

Table 4: 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

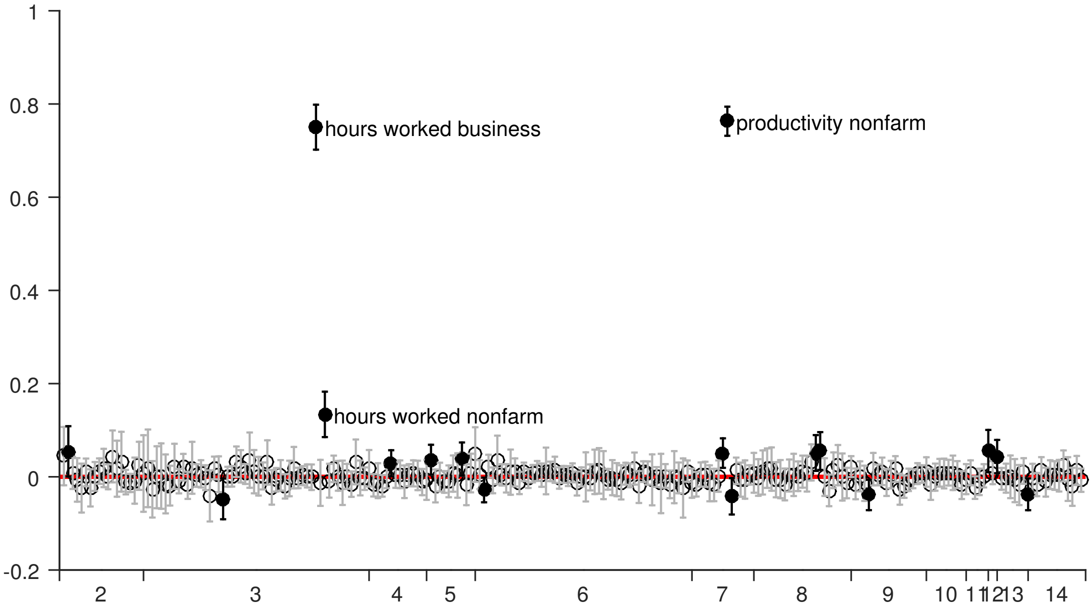
Note: this table reports the estimated coefficients (coef.) and standard errors (SE) which are significantly different from zero on a five percent significance level, estimated by the Moore-Penrose pseudoinverse estimator (MPI), random least squares (RLS), and ridge regularization (RID). The group numbers (gr.) correspond to the FRED-QD variable categories. The fred mnemonics and variable descriptions corresponding to the variable names are given in Appendix D.

5.3 Empirical Results

Table 4 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 corresponds 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

Figure 1: 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.

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 1 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.

6 Conclusion

This paper proposes 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 of the regressors by a diagonally scaled Moore-Penrose pseudoinverse. After a bias correction with the lasso this yields an asymptotically unbiased and normally distributed estimator. The covariance matrix of the estimates is available in closed form and free of tuning parameters. Confidence intervals can then be constructed using standard procedures.

We also consider two regularized estimators; 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 a high dimensional regressor matrix, the proposed estimators provide valid confidence intervals with correct coverage rates. 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.

References

Albert, A. (1972). *Regression and the Moore-Penrose pseudoinverse*. Elsevier.

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*.

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

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

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

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1.

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*, 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.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

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

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*, 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 Preliminary lemmas

A.1 Concentration bounds

Lemma 6. *Let $z_1^2, \ldots, z_p^2$ be independent subexponential variables with $E[z_i^2] = 1$. Define by $c_s > 0$ a constant such that $\sup_{l\geq 1} l^{-1/2} \left(E[|z_i|^l]\right)^{1/l} \leq c_s$. Then for every $\epsilon \geq 0$,*

$$P\left(\left|\frac{1}{p}\sum_{i=1}^{p} z_i^2 - 1\right| \geq \epsilon\right) \leq 2\exp\left[-cp\min\left(\frac{\epsilon^2}{4c_s^2}, \frac{\epsilon}{2c_s}\right)\right] \tag{40}$$

with $c > 0$ an absolute constant.

Proof: see Vershynin (2010), Proposition 5.16.

Lemma 7 (Variant Johnson and Lindenstrauss (1984) lemma)**.** *Let v be a fixed $p \times 1$ vector, and U_n a $p \times n$ matrix that is distributed uniformly over the Stiefel manifold $V_{n,p}$. Then for c_s as in Lemma 6 and $0 \leq \epsilon \leq 2c_s$,*

$$P\left(\frac{v'U_nU_n'v}{v'v} \geq (1+\epsilon)\frac{n}{p}, \quad \frac{v'U_nU_n'v}{v'v} \leq (1-\epsilon)\frac{n}{p}\right) \leq 2\exp\left(-\frac{c}{4c_s^2}\epsilon^2 n\right) \tag{41}$$

for $c, c_s > 0$.

Proof: Since $U_n \in V_{n,p}$, we have that $U_n'U_n = I_n$. Then $v'U_nU_n'v = ||P_{U_n}v||_2^2$, with the orthogonal projection matrix $P_{U_n} = U_n(U_n'U_n)^{-1}U_n'$. As U_n is uniformly distributed on $V_{n,p}$, P_{U_n} is uniformly distributed on the Grassmannian manifold $G_{n,p}$ (Chikuse (2012), theorem 2.2.2).

Instead of taking P_{U_n} random and v fixed, we can take the projection fixed and consider a random v. This holds as for any fixed $n \times n$ matrix $P \in G_{n,p}$ and Q uniformly distributed in $\mathcal{O}(p)$, the product QPQ' is uniformly distributed on the Grassmannian $G_{n,p}$ (Chikuse (2012), theorem 2.2.2). Then, for uniformly random P_{U_n}, $v'P_{U_n}'v \overset{(d)}{=} v'QPQv$ where P is fixed.

Since Q is uniformly distributed on $\mathcal{O}(p)$, $Qv \overset{(d)}{=} z$ with z uniformly on the unit sphere S^{p-1}. Without loss of generality, assume that the fixed projection matrix P projects z on its first n coordinates. Then

$$\mathrm{E}\left[||Pz||_2^2\right] = \mathrm{E}\left[\sum_{i=1}^{n} z_i^2\right] = \sum_{i=1}^{n} \mathrm{E}\left[z_i^2\right] \tag{42}$$

Since z is uniformly distributed S^{p-1}, $E[z'z] = \mathrm{E}\left[\sum_{i=1}^{p} z_i^2\right] = pE[z_1^2] = 1$. Then it follows from (42) that

$$\mathrm{E}[||Pz||_2^2] = \frac{n}{p} \tag{43}$$

To prove Lemma 7, we need a concentration result around this expectation. Since z is uniformly distributed on the unit sphere, z is subgaussian. The subvector consisting of the first m coordinates is also subgaussian, as this is simply a linear transformation of z. The product of two subgaussian random variables is subexponential (Vershynin, 2010), and hence, we can invoke Lemma 6. We have $E[z_i^2] = \frac{1}{p}$, such that

$$P\left(\left|\frac{p}{n}\sum_{i=1}^{n} z_i^2 - 1\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{c}{4c_s^2}\epsilon^2 n\right), \tag{44}$$

for $\epsilon, c, c_s > 0$. Note that we assume that $\epsilon^2/(4c_s^2) \leq \epsilon/(2c_s)$, which is satisfied for sufficiently small ϵ. ■

A.2 Properties of elliptical distributions

Under Assumption A2, the concentration results from Appendix A.1 bound the elements of the diagonally scaled Moore-Penrose pseudoinverse. To show this, we first introduce properties of matrices generated from elliptical and spherically symmetric distributions.

For Z an $n \times p$ matrix with rows generated from a spherically symmetric distribution, $Z \overset{(d)}{=} ZT$ for $T \in \mathcal{O}(p)$. The matrix Z can be decomposed by a singular value decomposition as

$$Z = VSU', \tag{45}$$

where $V \in \mathcal{O}(n)$, S the $n \times p$ 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 $\mathcal{O}(p)$. When $n < p$,

$$Z = VS_nU_n', \tag{46}$$

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 $U_n' = [I_n, O_{n,p-n}]U'$. Since U is uniformly distributed over $\mathcal{O}(p)$, U_n is uniformly distributed over the Stiefel manifold $V_{n,p}$ defined as $V_{n,p} = \{A \in R^{p\times n} : A'A = I_n\}$.

Definition 1 (Matrix Angular Central Gaussian distribution, Chikuse (1990))**.** *Suppose the entries of a $p \times n$ matrix W are independent standard normally distributed, and Σ an invertible $p \times p$ matrix. Define $H = \Sigma^{1/2}W(W'\Sigma W)^{-1/2}$. Then H has the density function*

$$f_H = |\Sigma|^{-n/2}|H'\Sigma^{-1}H|^{-p/2}, \tag{47}$$

and is generated from the Matrix Angular Central Gaussian distribution with parameter Σ, denoted as $MACG(\Sigma)$, and defined on the Stiefel manifold $V_{n,p}$. For $n = 1$, this reduces to the Angular Central Gaussian distribution $ACG(\Sigma)$ on the unit sphere S^{p-1}.

Lemma 8 (Chikuse (2012))**.** *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*

$$U_n = W(W'W)^{-1/2}. \tag{48}$$

Lemma 9 (Chikuse (2012))**.** *Let H be a $p \times n$ random matrix on the Stiefel manifold $V_{n,p}$, which is decomposed as*

$$H = [h_1, H_2], \tag{49}$$

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

$$h_1 = G(H_2)T, \tag{50}$$

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 10 (Wang and Leng (2015))**.** *Let H be a $p \times n$ random matrix on the Stiefel manifold $V_{n,p}$. Suppose $H \sim MACG(\Sigma)$. Decompose the Stiefel manifold $H = (G(H_2)T, H_2)$ as in Lemma 9, with T a $(p-n+1)\times 1$ and H_2 a $p \times (n-1)$ matrix. Then,*

$$T|H_2 \sim ACG(G(H_2)'\Sigma G(H_2)). \tag{51}$$

Since $h_1 = G(H_2)T$, which is a linear transformation of T,

$$h_1|H_2 \sim ACG(\tilde{\Sigma}). \tag{52}$$

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

Lemma 11 (Fan and Lv (2008); Wang and Leng (2015))**.** *Denote the first row of H by $h_1' = [h_{11}, h']$. We then have*

$$e_1HH'e_2 \overset{(d)}{=} h_{11}h_{21}\Big| \left\{e_1HHe_1 = h_{11}^2\right\}. \tag{53}$$

Proof: For $Q \in \mathcal{O}(n)$

$$e_1'HH'e_2 = e_1'HQQ'H'e_2. \tag{54}$$

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

$$e_1'HH'e_2 = e_1'\tilde{H}\tilde{H}e_2 = \tilde{h}_{11}\tilde{h}_{21}, \tag{55}$$

implying that

$$e_1'HH'e_2 \overset{(d)}{=} h_{11}h_{21}\Big| \left\{e_1'H = h_{11}\right\}. \tag{56}$$

Denote the first row of H by $h_1' = [h_{11}, h']$. Then $e_1'HH'e_1 = h_{11}^2 + h'h$ and thus $e_1'H = [h_{11}, 0_{1\times n-1}]$ if and only if $e_1'HH'e_1 = h_{11}^2$. Substituting this into (56) completes the proof. ■

B Proofs

B.1 Proof of Lemma 1

Under Assumption A2 and the decomposition (46) in Appendix A.2,

$$X'(XX')^{-1}X = \Sigma_2^{1/2}U_n(U_n'\Sigma_2U_n)^{-1}U_n'\Sigma_2^{1/2}. \tag{57}$$

By Lemma 8 in Appendix A, we can write $U_n = W(W'W)^{-1/2}$ with the elements of W standard normal and independently distributed. Substituting into (57) gives

$$X'(XX')^{-1}X = \Sigma_2^{1/2}W(W'\Sigma_2W)^{-1}W'\Sigma_2^{1/2} = HH', \tag{58}$$

where $H = \Sigma_2^{1/2}W(W'\Sigma_2W)^{-1/2}$.

We 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 MX are all equal to one. However, to bound the off-diagonal elements, we require a bound on the diagonal elements of HH'. We first construct bounds under the assumption that $\Sigma = I_p$, and then connect these to the case where $\Sigma_2 \neq I_p$.

When $\Sigma_2 = I_p$, we can invoke Lemma 7 in Appendix A to show that

$$P\left(e_1'U_nU_n'e_1 > c_\epsilon\frac{n}{p}, \quad e_1'U_nU_n'e_1 < \frac{1}{c_\epsilon}\frac{n}{p}\right) \leq 2\exp\left(-\frac{c}{4c_s^2}\epsilon^2 n\right), \tag{59}$$

with $c, c_s > 0$, and $c_\epsilon = \frac{1+\epsilon}{1-\epsilon} > 1$ is introduced to reduce notation.

We will now use these results to establish a bound when $\Sigma_2 \neq I$. The diagonal terms can be bounded by noting that for any vector v,

$$\begin{aligned} v'HH'v &= v'\Sigma_2^{\frac{1}{2}}U_n(U_n'\Sigma_2U_n)^{-1}U_n'\Sigma_2^{\frac{1}{2}}v \\ &\leq \kappa v'U_nU_n'v, \end{aligned} \tag{60}$$

where the condition number $\kappa = \frac{\lambda_{\max}(\Sigma_2)}{\lambda_{\min}(\Sigma_2)} < \infty$ by Assumption A3. Similarly

$$v'HH'v \geq \frac{1}{\kappa}v'U_nU_n'v. \tag{61}$$

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

$$P\left(e_1'HH'e_1 > c_\epsilon\kappa\frac{n}{p}, \quad e_1'HH'e_1 < \frac{1}{c_\epsilon\kappa}\frac{n}{p}\right) \leq 2\exp\left(-\frac{c}{4c_s^2}\epsilon^2 n\right). \tag{62}$$

Off-diagonal elements The proof for the off-diagonal elements is more involved. For $i = 1$ and $j = 2$, we bound with high probability the ratio $\frac{|e_i'HH'e_j|}{e_i'HH'e_i}$. A union bound is used to extend the results to arbitrary i and j.

We separate three cases: (a) $e_1'HHe_1 \geq c_\epsilon\kappa\frac{n}{p}$, (b) $c_\epsilon\kappa\frac{n}{p} > e_1'HH'e_1 > \frac{1}{c_\epsilon\kappa}\frac{n}{p}$, and (c) $e_1'HHe_1 \leq c_\epsilon\kappa\frac{n}{p}$. Conditioning on these three cases and using the trivial fact that for any probability $P(\cdot) \leq 1$, it follows that

$$\begin{aligned} P\left(\frac{|e_1'HH'e_2|}{e_1'HH'e_1} \geq t\right) &\leq P\left(e_1'HH'e_1 \geq c_\epsilon\kappa\frac{n}{p}\right) + P\left(e_1'HH'e_1 \leq \frac{1}{c_\epsilon\kappa}\frac{n}{p}\right) \\ &+ \int_{\frac{1}{c_\epsilon\kappa}\frac{n}{p}}^{c_\epsilon\kappa\frac{n}{p}} P\left(\frac{|e_1'HH'e_2|}{e_1'HH'e_1} \geq t \,\middle|\, e_1'HH'e_1 = t_1^2\right) P\left(e_1'HH'e_1 = t_1^2\right) dt_1^2 \\ &\leq P\left(e_1'HH'e_1 \geq c_\epsilon\kappa\frac{n}{p}\right) + P\left(e_1'HH'e_1 \leq \frac{1}{c_\epsilon\kappa}\frac{n}{p}\right) \\ &+ P\left(\frac{|e_1'HH'e_2|}{e_1'HH'e_1} \geq t \,\middle|\, e_1'HH'e_1 = t_*^2\right). \end{aligned} \tag{63}$$

where t_* is the value of t_1 that maximizes $P\left(\frac{|e_1'HH'e_2|}{e_1'HH'e_1} \geq t \,\middle|\, e_1'HH'e_1 = t_1^2\right)$.

The first two terms of (63) are bounded by (62), so we focus on the final term of (63). Denote the i,j-th element of H by h_{ij}. Lemma 11 in Appendix A states that

$$e_1'HH'e_2 \overset{(d)}{=} h_{11}h_{21} \mid \left\{h_{11}^2 = e_1'HH'e_1\right\}, \tag{64}$$

from which it follows that

$$e_1'HH'e_2 | \left\{e_1'HH'e_1 = t_1^2\right\} \overset{(d)}{=} h_{11}h_{21} | \left\{h_{11}^2 = t_1^2\right\}. \tag{65}$$

We decompose $H = [h_1, H_2]$, with h_1 a $p \times 1$ vector, and H_2 a $p \times n-1$ matrix. As in Lemma 10, $h_1 = G(H_2)T$ with $G(H_2)$ such that $[H_2, G(H_2)] \in \mathcal{O}(p)$. Then by Lemma 10 in Appendix A, $h_1|H_2 \overset{(d)}{=} \frac{y}{\sqrt{y_1^2+\ldots+y_p^2}}$, where $y = (y_1, \ldots, y_p) \sim N(0, \tilde{\Sigma})$ with $\tilde{\Sigma} = G(H_2)G(H_2)'\Sigma G(H_2)G(H_2)'$.

Using the above results, $h_{11}h_{21}| \left\{h_{11}^2 = t_1^2\right\} \overset{(d)}{=} \frac{y_1y_2}{y_1^2+\ldots+y_p^2}$. Since $\frac{y_1^2}{y_1^2+\ldots+y_p^2} = t_1^2$, we have $y_1^2 = \frac{t_1^2}{1-t_1^2}\left(y_2^2 + \ldots + y_p^2\right)$. Then

$$\frac{|y_1y_2|}{y_1^2 + \ldots + y_p^2} = \frac{(1-t_1^2)|y_1y_2|}{y_2^2 + \ldots + y_p^2} \leq \frac{\sqrt{1-t_1^2}|t_1||y_2|}{\sqrt{y_2^2 + \ldots + y_p^2}}. \tag{66}$$

Now we establish the following upper bound

$$
\begin{aligned}
P\left(\frac{|e_1'HH'e_2|}{e_1'HH'e_1}\geq t\middle| h_{11}^2=t_1^2\right) &= P\left(\frac{|h_{11}h_{21}|}{h_{11}^2}\geq t\middle| h_{11}^2=t_1^2\right)\\
&\leq P\left(\frac{\sqrt{1-t_1^2}|y_2|}{\sqrt{y_2^2+\ldots+y_p^2}}\geq |t_1|t\right)\\
&= P\left(\frac{|y_2|}{\sqrt{y_2^2+\ldots+y_p^2}}\geq t\sqrt{\frac{t_1^2}{1-t_1^2}}\right) \qquad (67)\\
&\leq P\left(\frac{|y_2|}{\sqrt{y_2^2+\ldots+y_p^2}}\geq t\sqrt{\frac{1}{c_\epsilon\kappa}\frac{n}{p}}\right).
\end{aligned}
$$

where we use that $t_1^2/(1-t_1^2)$ is a monotonically increasing function in t_1^2, and the minimum value of t_1^2 that we need to consider equals $\frac{1}{c_\epsilon\kappa}\frac{n}{p}$. This is then our choice for t_* in (63).

Since by definition, $G(H_2)'G(H_2)=I_{p-n+1}$, $\lambda_{\max}(\tilde{\Sigma})\leq\lambda_{\max}(\Sigma)$. Similarly, we have $\lambda_{\min}(\tilde{\Sigma})\geq\lambda_{\min}(\Sigma)$. Then by Lemma 6 in Appendix A,

$$
\begin{aligned}
P\left(|y_2|\geq\sqrt{\lambda_{\max}(\Sigma)}\sqrt{1+\epsilon_1}\right) &\leq 2e^{-\frac{c}{2c_s}\epsilon_1}\\
P\left(\sqrt{y_2^2+\ldots+y_p^2}\leq\sqrt{\lambda_{\min}(\Sigma)(p-n)(1+\epsilon_2)}\right) &\leq 2e^{-\frac{c}{4c_s^2}\epsilon_2^2(p-n)},
\end{aligned}
\qquad (68)
$$

where we assumed that ϵ_1 is such that $\epsilon_1/(2c_s)<\epsilon_1^2/(4c_s^2)$, which will be justified below, and ϵ_2 such that $\epsilon_2^2/(4c_s^2)\leq\epsilon_2/(2c_s)$.

Using Bonferonni's inequality, (68) implies

$$
P\left(\frac{|y_2|}{\sqrt{y_2^2+\ldots+y_p^2}}\geq\sqrt{\kappa\frac{1+\epsilon_1}{1+\epsilon_2}\frac{1}{p-n}}\right)\leq 2e^{-\frac{c}{2c_s}\epsilon_1}+2e^{-\frac{c}{4c_s^2}\epsilon_2^2(p-n)}. \qquad (69)
$$

Take c_p a constant such that $p/n\geq c_p>1$, then also

$$
P\left(\frac{|y_2|}{\sqrt{y_2^2+\ldots+y_p^2}}\geq\sqrt{\frac{\kappa}{(1-c_p^{-1})}\frac{1+\epsilon_1}{1+\epsilon_2}\frac{1}{p}}\right)\leq 2e^{-\frac{c}{2c_s}\epsilon_1}+2e^{-\frac{c}{4c_s^2}\epsilon_2^2(p-n)}. \qquad (70)
$$

We are interested in the case where $\sqrt{\frac{\kappa}{(1-c_p^{-1})}\frac{1+\epsilon_1}{1+\epsilon_2}\frac{1}{p}}=t\sqrt{\frac{1}{c_\epsilon\kappa}\frac{n}{p}}$, which holds for

$$
t=\kappa\sqrt{\frac{c_pc_\epsilon}{c_p-1}\frac{1+\epsilon_1}{1+\epsilon_2}\frac{1}{n}}. \qquad (71)
$$

Since $\kappa^2 c_\epsilon c_p/(c_p-1) > 1$, we can take $\epsilon_2 = \kappa^2 c_\epsilon c_p/(c_p-1) - 1$. Then choosing $\epsilon_1 = a^2 \log p - 1$, we have

$$P\left(\frac{|e_1' HH' e_2|}{e_1' HH' e_1} > a\sqrt{\frac{\log p}{n}}\right) \leq 2e^{-\frac{c}{2c_s}a^2 \log p} + 2e^{-\frac{c}{4c_s^2}[\kappa^2 c_\epsilon c_p/(c_p-1)-1]^2(p-n)}. \tag{72}$$

Note that for this choice of ϵ_1, for p sufficiently large $\epsilon_1/(2c_s) < \epsilon_1^2/(4c_s^2)$, which was used in (68).

Finally, taking the union bound over all pairs e_i, e_j we have that

$$P\left(\frac{|e_i' HH' e_j|}{e_i' HH' e_i} > a\sqrt{\frac{\log p}{n}}\right) = O\left(p^{-\tilde{c}}\right) \qquad \forall i, j \in \{1, \ldots, p\} \tag{73}$$

with $\tilde{c} = \frac{c}{2c_s}a^2 - 2$. ■

B.2 Proof of Lemma 2

The bound in Lemma 2 is shown by Bühlmann and Van De Geer (2011) to hold under the following compatibility condition

Definition 2 (Compatibility condition). *Denote by S_0 the true set of $s_0 = ||S_0||_0$ non-zero coefficients, then the compatibility condition is satisfied for this set if*

$$||\beta_{S_0}||_1 \leq \frac{\sqrt{s_0}||X\beta||_2}{\sqrt{n}\phi_0}, \tag{74}$$

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

This condition is satisfied under Assumption A2. Note that $||\beta_{S_0}||_1 \leq \sqrt{s_0}||\beta_{S_0}||_2$, so it is sufficient if

$$||\beta||_2^2 \leq \frac{\beta' \frac{1}{n} X' X \beta}{\phi_0}. \tag{75}$$

Using Assumption A2, we have

$$\begin{aligned} \beta' \frac{1}{n} X' X \beta &= \beta' \Sigma^{1/2} \frac{1}{n} U S' S U' \Sigma^{1/2} \beta \\ &\geq \frac{1}{c_Z} \frac{p}{n} v' U_n U_n' v, \end{aligned} \tag{76}$$

where $v = \Sigma^{1/2}\beta$, and the last line holds since the non-zero eigenvalues $S'S$ are the same as the eigenvalues of ZZ' which are bounded by Assumption A2. Now we can invoke Lemma 7, such that with probability at least $1 - 2\exp(-\frac{c}{4c_s^2}\epsilon^2 n)$, we have that

$$\begin{aligned}\beta'\frac{1}{n}X'X\beta &\geq \frac{1}{c_Z c_\epsilon}\beta'\Sigma\beta \\ &\geq \frac{1}{c_Z c_\epsilon}\lambda_{\min}(\Sigma)||\beta||_2^2.\end{aligned} \tag{77}$$

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

B.3 Proof of Lemma 3

Building on Wang and Leng (2015), we rewrite the noise term of $\hat{\beta}_i^c$ as

$$Z_i = \sqrt{n}d_i x_i'(XX')^{-1}\varepsilon \overset{(d)}{=} \sqrt{n}d_i||x_i'(XX')^{-1}||_2 \frac{\sigma x_i'(XX')^{-1}u}{||x_i'(XX')^{-1}||_2}, \tag{78}$$

where $u \sim N(0, I_n)$.

We first bound the norm term

$$\sqrt{n}d_i||x_i'(XX')^{-1}||_2 = \sqrt{n}\frac{||x_i'(XX')^{-1}||_2}{x_i'(XX')^{-1}x_i}. \tag{79}$$

Using standard norm inequalities, we have

$$\frac{1}{\lambda_{\max}(XX')}x_i'(XX')^{-1}x_i \leq ||x_i'(XX')^{-1}||_2^2 \leq \frac{1}{\lambda_{\min}(XX')}x_i'(XX')^{-1}x_i. \tag{80}$$

The eigenvalues of $XX' = \Sigma_1^{1/2}Z\Sigma_2 Z'\Sigma_1^{1/2}$ satisfy

$$\begin{aligned}\lambda_{\max}(\Sigma_1^{1/2}Z\Sigma_2 Z'\Sigma_1^{1/2}) &\leq \lambda_{\max}(\Sigma_1)\lambda_{\max}(\Sigma_2)\lambda_{\max}(ZZ'), \\ \lambda_{\min}(\Sigma_1^{1/2}Z\Sigma_2 Z'\Sigma_1^{1/2}) &\geq \lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_2)\lambda_{\min}(ZZ').\end{aligned} \tag{81}$$

The eigenvalues of ZZ' are bounded by Assumption A2, and using (62), it follows that with probability exceeding $1 - 2e^{-c\epsilon^2 n} - 2e^{-C_Z n}$ we have that

$$\begin{aligned}\left(\frac{1}{\lambda_{\max}(\Sigma_1)\lambda_{\max}(\Sigma_2)}\frac{n}{p}\frac{1}{c_\epsilon \kappa \frac{n}{p}}\right)^{1/2} &\leq \sqrt{n}d_i||x_i'(XX')^{-1}||_2 \\ &\leq \left(\frac{1}{\lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_2)}\frac{n}{p}\frac{1}{\frac{1}{c_\epsilon \kappa}\frac{n}{p}}\right)^{1/2},\end{aligned} \tag{82}$$

By Assumption A3, the eigenvalues of Σ_1 and Σ_2 are finite. Then

$$\sqrt{n}d_i||x_i'(XX')^{-1}||_2 = O_p(1). \tag{83}$$

We now turn to the second term of (78)

$$\frac{\sigma x_i'(XX')^{-1}u}{||x_i'(XX')^{-1}||_2} = \frac{\sigma \frac{1}{\sqrt{n}} x_i' \left(\frac{1}{p} XX'\right)^{-1} u}{\left|\left|\frac{1}{\sqrt{n}} x_i' \left(\frac{1}{p} XX'\right)^{-1}\right|\right|_2}. \tag{84}$$

When $u \sim N(0, I_n)$, it is clear that

$$\frac{1}{\sqrt{n}} X' \left(\frac{1}{p} XX'\right)^{-1} u \sim N\left[0, \frac{1}{n} X' \left(\frac{1}{p} XX'\right)^{-2} X\right]. \tag{85}$$

and hence $Z_i \sim N(0, \sigma^2 \Omega_{ii})$ with $\Omega_{ii} = n \frac{x_i'(XX')^{-2} x_i}{(x_i'(XX')^{-1} x_i)^2} = O_p(1)$. ■

B.4 Proof of Lemma 3 for non-gaussian errors

Lemma 12. *Suppose assumptions A2 and A3 hold. The errors ε_i are independent and identically distributed with variance σ^2, and satisfy*

$$E\left[|\varepsilon_i|^{2+\delta}\right] \le c < \infty \tag{86}$$

for $i = 1, \dots, n$. Then as $n \to \infty$,

$$\frac{1}{\sqrt{n}} e_i' M \varepsilon \overset{(d)}{\to} N(0, \sigma^2 e_i' M M' e_i / n). \tag{87}$$

Proof: When $u_i \sim i.i.d(0, 1)$, we will show that Lyupanov's condition is satisfied, and therefore a central limit theorem applies ensuring that, as $n \to \infty$,

$$\frac{\sigma x_i' \left(XX'\right)^{-1} u}{\left|\left|x_i' \left(XX'\right)^{-1}\right|\right|_2} \overset{(d)}{\to} N(0, \sigma^2). \tag{88}$$

Define

$$r_{ik} = \frac{\left[\left(XX'\right)^{-1} x_i\right]_k}{||\left(XX'\right)^{-1} x_i||_2} \tag{89}$$

where the numerator denotes the k-th component of the n-dimensional vector $\left(XX'\right)^{-1} x_i$. Furthermore, we have $\mathrm{E}[r_{ik} u_k] = 0$, $\mathrm{Var}[r_{ik} u_k] = \left(||\left(XX'\right)^{-1} x_i||_2^{-1} \left[\left(XX'\right)^{-1} x_i\right]_k\right)^2$, $s_n^2 = \sum_{k=1}^n \mathrm{Var}[r_{ik} u_k] = 1$. To prove that a central limit theorem applies to $\sum_{k=1}^n r_{ik} u_k$ we prove that Lyapunov's condition,

$$LC = \lim_{n \to \infty} \sum_{k=1}^n |r_{ik} u_k|^{2+\delta} = 0, \tag{90}$$

holds. By assumption we have

$$LC \leq c \lim_{n\to\infty} \sum_{k=1}^{n} |r_{ik}|^{2+\delta}. \tag{91}$$

By Assumption 2 the summand satisfies with probability exceeding $1-\exp(-C_Z n)$

$$|r_{ik}| \leq c_Z^2 \frac{||x_i||_\infty}{||x_i||_2}. \tag{92}$$

By the results in Appendix B.2, we have that, again with high probability, $||x_i||_2 \geq \frac{\lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_2)}{c_Z} c_\epsilon^{-1} n$. We can then continue our string of inequalities as

$$|r_{ik}| \leq c_Z^3 c_{\kappa,1} c_{\kappa_2} c_\epsilon \frac{||z_i||_\infty}{n}, \tag{93}$$

where z_i denotes the i-th row of the matrix Z defined in Assumption 2.

Since by assumption each element of Z is independent and identically distributed with variance 1, following Chebyshev's inequality

$$P(|z_{ik}| \geq a) \leq a^{-2}. \tag{94}$$

Then applying a union bound over $k \in \{1, \ldots, n\}$ gives

$$P(||z_i||_\infty \geq a) \leq na^{-2}. \tag{95}$$

Choosing $a = c_a n^{1/2(1+\alpha)}$, the right-hand side tends to zero, and uniformly over k,

$$|z_{ik}| \leq c_Z^3 c_{\kappa,1} c_{\kappa_2} c_\epsilon n^{-1/2(1-\alpha)}. \tag{96}$$

In this case

$$LC \leq c_Z^3 c_{\kappa,1} c_{\kappa_2} c_\epsilon n^{\alpha-\delta/2+\alpha\delta/2}, \tag{97}$$

which tends to zero as $n \to \infty$ if

$$\alpha - \delta/2 + \alpha\delta/2 < 0 \Rightarrow \alpha \leq \frac{\delta}{2+\delta}. \tag{98}$$

This shows that for an individual parameter β_i,

$$\sum_{k=1}^{n} r_{ik} u_k \xrightarrow{d} N(0,1) \tag{99}$$

which completes the proof. The extension to a fixed subset of β follows from a union bound of the size of the subset. ■

We can extend the results of Lemma 12 to hold uniformly over $i \in \{1, \ldots, p\}$, by making the additional assumption that the rows of Z are subgaussian and the number of variables does not increase too fast with the number of observations.

Lemma 13. *Suppose assumptions A2 and A3 hold, but strenghten Assumption A2 such that the rows of Z are also subgaussian. As in Lemma 12, suppose that the errors ε_i are independent and identically distributed with variance σ^2, and satisfy $E\left[|\varepsilon_i|^{2+\delta}\right] \leq c < \infty$ for $i = 1, \ldots, n$. In addition, the number of regressors grows at a rate*

$$\log p = o\left(n^{1-\frac{1}{2+\delta}}\right). \tag{100}$$

Then, as $n \to \infty$

$$\frac{1}{\sqrt{n}} e_i' M \varepsilon \overset{(d)}{\to} N(0, \sigma^2 e_i' M M' e_i / n). \tag{101}$$

and this result holds uniformly over $i \in \{1, \ldots, p\}$.

Proof: In this case, instead of Chebyshev's inequality (94) we use

$$P(|z_{ik}| \geq a) \leq 2 \exp\left(-a^2/2\right). \tag{102}$$

Applying again a union bound over all $k \in \{1, \ldots, n\}$ and $i \in \{1, \ldots, p\}$ gives uniformly over i

$$P(||Z||_{\max} \geq a) \leq 2 \exp\left(-a^2/2 + \log p + \log n\right) \tag{103}$$

The right-hand side now goes to zero if $a > \sqrt{2(\log p + \log n)}$. In this case, we have

$$LC \leq c \lim_{n \to \infty} \left(\frac{\log p + \log n}{n^{1-\frac{1}{2+\delta}}}\right)^{2+\delta} \tag{104}$$

Ignoring the lower order term $\log n$, we see that $LC \to 0$ uniformly over $i \in \{1, \ldots, p\}$, when $n \to \infty$ and $\log p = o\left(n^{1-\frac{1}{2+\delta}}\right)$. This completes the proof. ■

B.5 Proof of Lemma 4: random least squares

Size of the bias Consider the eigenvalue decomposition

$$\frac{1}{n} X'X = \hat{U}_n \hat{\Lambda} \hat{U}_n'. \tag{105}$$

where $\hat{U}_n$ is a $p \times n$ matrix, and $\hat{\Lambda}$ an $n \times n$ diagonal matrix of eigenvalues. We list three properties of the expectation $\mathrm{E}[R(R'X'XR)^{-1}R']X'X$ established in Marzetta et al. (2011).

First, using the eigenvalue decomposition (105) and the fact that only n eigenvalues are non-zero,

$$\mathrm{E}[R(R'X'XR)^{-1}R']X'X \overset{(d)}{=} \hat{U}_n\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\hat{\Lambda}\hat{U}_n', \tag{106}$$

with Φ an $n \times k$ matrix of independent standard normal random variables. The proof relies on the fact that for any orthogonal matrix $\hat{U}$ independent of R, we have that $\hat{U}'R \overset{(d)}{=} \Phi$.

Second, $\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\hat{\Lambda}$ is a diagonal matrix. This follows since 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,

$$\begin{aligned}\Omega\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\hat{\Lambda}\Omega^* &= \Omega\mathrm{E}[\Phi(\Phi'\hat{\Lambda}\Phi)^{-1}\Phi']\Omega^*\hat{\Lambda} \\ &= \Omega\mathrm{E}[\Phi(\Phi'\Omega^*\Omega\hat{\Lambda}\Omega^*\Omega\Phi)^{-1}\Phi']\Omega^*\hat{\Lambda} \\ &\overset{(d)}{=} \mathrm{E}[\Psi(\Psi'\hat{\Lambda}\Psi)^{-1}\Psi']\hat{\Lambda},\end{aligned} \tag{107}$$

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

The final property is that we can rewrite

$$\mathrm{E}[\Psi(\Psi'\hat{\Lambda}\Psi)^{-1}\Psi']\hat{\Lambda} = I - V, \tag{108}$$

where

$$V = \mathrm{E}[\Xi(\Xi'\hat{\Lambda}^{-1}\Xi)^{-1}\Xi']\hat{\Lambda}^{-1} \tag{109}$$

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

Using (108), it follows that

$$\mathrm{E}_R[R(R'X'XR)^{-1}R']X'X = \hat{U}(I-V)\hat{U}'. \tag{110}$$

Now, $\hat{U}\hat{U}'$ is the Moore-Penrose pseudoinverse post-multiplied by X, which is identical to (58) in Appendix B.1, so that we have

$$\hat{U}\hat{U}' = X'(XX')^{-1}X = HH'. \tag{111}$$

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 continue to hold.

Denote by $\hat{u}_i = \hat{U}'e_i$. We can use the following string of inequalities

$$\begin{aligned} P\left(\frac{|\hat{u}_i'(I-V)\hat{u}_j|}{\hat{u}_i'(I-V)\hat{u}_i} \geq t\right) &\leq P\left(\frac{|\hat{u}_i'(I-V)\hat{u}_j|}{\hat{u}_i'\hat{u}_i(1-||V||_2)} \geq t\right) \\ &\leq P\left(\frac{|\hat{u}_i'\hat{u}_j|}{\hat{u}_i'\hat{u}_i} + \frac{|\hat{u}_i'V\hat{u}_j|}{\hat{u}_i'\hat{u}_i} \geq t(1-||V||_2)\right) \\ &\leq P\left(\frac{|\hat{u}_i'\hat{u}_j|}{\hat{u}_i'\hat{u}_i} + ||V||_2\sqrt{\frac{\hat{u}_j'\hat{u}_j}{\hat{u}_i'\hat{u}_i}} \geq t(1-||V||_2)\right). \end{aligned} \tag{112}$$

For $\hat{u}_i'\hat{u}_i = e_i'HH'e_i$ and $\hat{u}_j'\hat{u}_j = e_j'HH'e_j$, we can apply the bounds established in (62) in Appendix B.1. Denote by $\mathcal{E}$ the event that $e_j'HH'e_j \leq c_\epsilon\kappa\frac{n}{p}$, $e_j'HH'e_j \geq (c_\epsilon\kappa)^{-1}\frac{n}{p}$, then the string of inequalities (112) proceeds as

$$\begin{aligned} &\leq P\left(\frac{|e_i'HH'e_j|}{e_i'HH'e_i} + ||V||_2c_\epsilon\kappa \geq t(1-||V||_2)\middle|\mathcal{E}\right)\left(1-2e^{-\frac{c}{4c_s^2}\epsilon^2 n}\right) + 2e^{-\frac{c}{4c_s^2}\epsilon^2 n} \\ &= P\left(\frac{|e_i'HH'e_j|}{e_i'HH'e_i} \geq t - ||V||_2\left(t + c_\epsilon\kappa\right)\right)\left(1-2e^{-\frac{c}{4c_s^2}\epsilon^2 n}\right) + 2e^{-\frac{c}{4c_s^2}\epsilon^2 n}. \end{aligned} \tag{113}$$

We now need to find a choice of the projection dimension k such that $t(1-||V||_2) - ||V||_2c_\epsilon\kappa = \tilde{a}\sqrt{\log p/n}$. This will then allow us to apply the previously derived bounds on $|e_i'HH'e_j|/e_i'HH'e_i$.

We first analyze the l_2 norm $||V||_2$ is more detail. Denote by $\hat{\lambda}_i$ the i-th diagonal element of the diagonal matrix of empirical eigenvalues $\hat{\Lambda}$, ξ_i the i-th row of Ξ defined in (109), and $A_{-i} \equiv \sum_{j\neq i}\hat{\lambda}_j^{-1}\xi_j\xi_j'$. It holds that

$$\begin{aligned} [V]_{ii} &= \hat{\lambda}_i^{-1}\xi_i'(\Xi'\hat{\Lambda}^{-1}\Xi)^{-1}\xi_i \\ &= \hat{\lambda}_i^{-1}\xi_i'\left(A_{-i} + \hat{\lambda}_i^{-1}\xi_i\xi_i'\right)^{-1}\xi_i \\ &= \frac{\hat{\lambda}_i^{-1}\nu_i}{1+\hat{\lambda}_i^{-1}\nu_i}, \end{aligned} \tag{114}$$

where

$$\nu_i = \xi_i'A_{-i}^{-1}\xi_i = \xi_i'\left(\Xi_{-i}\hat{\Lambda}_{-i}^{-1}\Xi_{-i}\right)^{-1}\xi_i > 0, \tag{115}$$

and the Sherman-Morrison formula is used to obtain the last line of (114). This shows that random least squares performs a type of generalized ridge regression, where the penalty is different for each eigenvalue. By Jensen's inequality and the fact that $x/(1+x)$ with $x > 0$ is a concave function,

$$[V]_{ii} \leq \frac{\hat{\lambda}_i^{-1}\mathrm{E}[\nu_i]}{1+\hat{\lambda}_i^{-1}\mathrm{E}[\nu_i]} \leq \frac{\hat{\kappa}\frac{n-k-1}{k}}{1+\hat{\kappa}\frac{n-k-1}{k}}, \tag{116}$$

where $\hat{\kappa} = \max_i \hat{\lambda}_i / \min_i \hat{\lambda}_i \leq c_\kappa$ by (81). We can now solve for which k we have that $t(1 - ||V||_2) - ||V||_2 c_\epsilon \kappa = \tilde{a}\sqrt{\log p/n}$. In order for the bias of the estimator to vanish compared to the noise, we require $t = a\sqrt{\log p/n}$. After some rewriting, we then find that k should satisfy

$$k = \left(1 + \frac{a - \tilde{a}}{\tilde{a}c_\kappa} \frac{\tilde{a}(c_\epsilon \kappa)^{-1}\sqrt{\log p/n}}{1 + \tilde{a}(c_\epsilon \kappa)^{-1}\sqrt{\log p/n}}\right)^{-1} (n-1). \tag{117}$$

Assuming $\tilde{a}(c_\epsilon \kappa)^{-1}\sqrt{\log p/n}$ to be sufficiently small, we have

$$k = \left(1 - c_k \sqrt{\frac{\log p}{n}}\right)(n-1), \tag{118}$$

with $c_k = (a - \tilde{a})/(\kappa c_\epsilon c_\kappa)$ a positive constant. Under this choice of k, the approximate inverse obtained by random least squares satisfies

$$\begin{aligned} P\left(\frac{|\hat{u}_i'(I-V)\hat{u}_j|}{\hat{u}_i'(I-V)\hat{u}_i} \geq a\sqrt{\frac{\log p}{n}}\right) &\leq P\left(\frac{|e_i'HH'e_j|}{e_i'HH'e_i} \geq \tilde{a}\sqrt{\frac{\log p}{n}}\right) \\ &= O\left(p^{-\tilde{c}}\right) \end{aligned} \tag{119}$$

with $\tilde{c}$ as in Lemma 1 with a replaced by $\tilde{a} < a$.

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

$$||\sqrt{n} d_i^{\mathrm{RLS}} e_i \mathrm{E}\left[R(R'X'XR)^{-1}R'\right] X'||_2 = O_p(1). \tag{120}$$

We rewrite this as

$$\left(\sqrt{n} d_i^{\mathrm{RLS}} ||e_i \mathrm{E}\left[R(R'X'XR)^{-1}R'\right] X'||_2\right)^2 = n\frac{e_i'\hat{U}_n(I-V)\hat{\Lambda}^{-1}(I-V)\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2}, \tag{121}$$

which can be lower and upper bounded as

$$\begin{aligned} \frac{1}{\lambda_{\max}(\hat{\Lambda})} n\frac{e_i'\hat{U}_n(I-V)^2\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2} &\leq n\frac{e_i'\hat{U}_n(I-V)\hat{\Lambda}^{-1}(I-V)\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2} \\ &\leq \frac{1}{\lambda_{\min}(\hat{\Lambda})} n\frac{e_i'\hat{U}_n(I-V)^2\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2}. \end{aligned} \tag{122}$$

Under stated assumptions, the eigenvalues satisfy $c_1 p \leq \lambda_{\min}(\hat{\Lambda}) \leq \lambda_{\max}(\hat{\Lambda}) \leq c_2 p$ for $0 \leq c_1 \leq c_2$. Also, from the previous paragraph we know that the elements of

V satisfy $0 \leq [V]_{ii} \leq c_V\sqrt{\frac{\log p}{n}} \leq 1$ for some $c_V > 0$. Then,

$$\frac{1}{c_2}\left[\frac{n}{p}\left(1 - c_V\sqrt{\frac{\log p}{n}}\right)^2 \frac{1}{e_i'\hat{U}_n\hat{U}_n'e_i}\right] \leq n\frac{e_i'\hat{U}_n(I-V)\hat{\Lambda}^{-1}(I-V)\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2}$$
$$\leq \frac{1}{c_1}\left[\frac{n}{p}\left(1 - c_V\sqrt{\frac{\log p}{n}}\right)^{-2} \frac{1}{e_i'\hat{U}_n\hat{U}_n'e_i}\right]. \tag{123}$$

Finally, (59) in Appendix A and the fact that $e_i'\hat{U}_n\hat{U}_n'e_i = e_iHH'e_i$ shows that

$$n\frac{e_i'\hat{U}_n(I-V)\hat{\Lambda}^{-1}(I-V)\hat{U}_n'e_i}{(e_i'\hat{U}_n(I-V)\hat{U}_n'e_i)^2} = O_p(1). \tag{124}$$

This completes the proof. ■

B.6 Proof of Lemma 5: ridge regression

Order of bias term 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 (110) in Appendix B.5.

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

$$\begin{aligned}(X'X + \gamma I_p)^{-1}X'X &= (\hat{U}\hat{\Lambda}\hat{U}' + \gamma I_p)^{-1}\hat{U}\hat{\Lambda}\hat{U}' \\ &= \hat{U}_n(I_n - V)\hat{U}_n',\end{aligned} \tag{125}$$

where $\hat{\Lambda}_n$ is a diagonal matrix with on the diagonal the nonzero eigenvalues of $X'X$, U_n consists of the first n rows of $\hat{U}$ and $V = (\hat{\Lambda}_n + \gamma I_n)^{-1}\gamma I_n$.

Now following (113), V should be such that $t(1-||V||_2)-||V||_2 c_\epsilon\kappa = \tilde{a}\sqrt{\log p/n}$ for $t = a\sqrt{\log p/n}$. This implies $||V||_2 = \frac{(a-\tilde{a})\sqrt{\log p/n}}{a\sqrt{\log p/n + c_\epsilon\kappa}}$. Since V is diagonal, and the non-zero eigenvalues satisfy $c_1p \leq \lambda_{\min}(\hat{\Lambda}) \leq \lambda_{\max}(\hat{\Lambda}) \leq c_2p$ for $0 \leq c_1 \leq c_2$,

$$||V||_2 = \max_{i=1,\ldots,n} \frac{\gamma}{\hat{\lambda}_i + \gamma} \leq \frac{\gamma}{c_1p + \gamma}. \tag{126}$$

It follows that we need to set

$$\gamma \leq c_1p\frac{||V||_2}{1 - ||V||_2}, \tag{127}$$

Using the expression for $||V||_2$ and assuming $\tilde{a}\sqrt{\log p/n}/(c_\epsilon\kappa)$ sufficiently small, we have

$$\gamma = c_\gamma \sqrt{\frac{\log p}{n}} p, \tag{128}$$

with $c_\gamma = c_1(a - \tilde{a})/(c_\epsilon \kappa)$. ■

Order of the variance What remains to be shown is

$$||\sqrt{n} d_i^{\mathrm{RI}} e_i'(X'X + \gamma I_p)^{-1} X'||_2 = O_p(1). \tag{129}$$

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

B.7 Proof of Theorem 3

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

$$\begin{aligned} ||e_i\hat{U}\mathrm{E}[R(R'\hat{\Lambda}R)^{-1}R'X'||_2^2 &= e_i\hat{U}A\hat{\Lambda}^{-1}A\hat{U}'e_i \\ &= e_i\hat{U}\hat{\Lambda}^{-1/2}A_{\mathrm{RLS}}^2\hat{\Lambda}^{-1/2}\hat{U}'e_i, \end{aligned} \tag{130}$$

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

Similarly, for the ridge regularized inverse, we have

$$\begin{aligned} ||e_i(X'X + \gamma I_p)^{-1}X'||_2^2 &= e_i(X'X + \gamma I_p)^{-1}X'X(X'X + \gamma I_p)^{-1}e_i \\ &= e_i\hat{U}_n(\hat{\Lambda} + \gamma I_p)^{-2}\hat{\Lambda}\hat{U}_n'e_i \\ &= e_i\hat{U}_n\hat{\Lambda}^{-1/2}A_{\mathrm{RID}}^2\hat{\Lambda}^{-1/2}\hat{U}_n'e_i, \end{aligned} \tag{131}$$

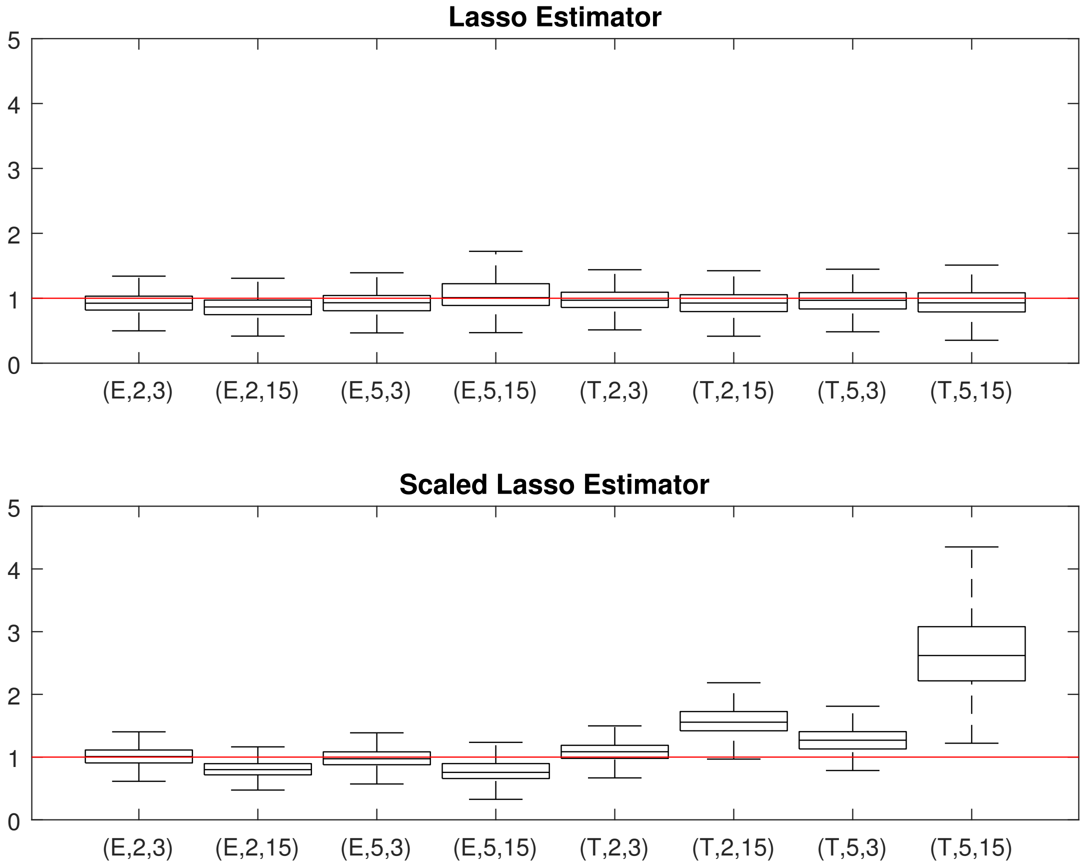
with A_{RID}^2 is a diagonal matrix with the diagonal elements satisfying $0 \leq A_{ii}^2 \leq 1$. For the Moore-Penrose pseudoinverse we have

$$\begin{aligned} ||e_iX'(XX')^{-1}||_2^2 &= e_iX'(XX')^{-2}Xe_i \\ &= e_i\hat{U}\hat{\Lambda}^{-1}\hat{U}'e_i. \end{aligned} \tag{132}$$

Since for both RLS and RID A^2 is a diagonal matrix with diagonal elements satisfying $0 \leq A_{ii}^2 \leq 1$, the claim in Theorem 3 follows. ■

C Estimation of the noise level

Figure 2: 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 (19), 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 1.

D Variable descriptions

Table 5: Variable Descriptions Table 4

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 4.