TI 2016-073/III
Tinbergen Institute Discussion Paper

Forecasting Using Random Subspace Methods

Revision: 11-08-2017

Tom Boot[1]
Didier Nibbering[2]

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

Forecasting Using Random Subspace Methods

Tom Boot*
University of Groningen

Didier Nibbering†
Erasmus University Rotterdam
Tinbergen Institute

August 11, 2017

Abstract

Random subspace methods are a novel approach to obtain accurate forecasts in high-dimensional regression settings. Forecasts are constructed from random subsets of predictors or randomly weighted predictors. We provide a theoretical justification for these strategies by deriving bounds on their asymptotic mean squared forecast error, which are highly informative on the scenarios where the methods work well. Monte Carlo simulations confirm the theoretical findings and show improvements in predictive accuracy relative to widely used benchmarks. The predictive accuracy on monthly macroeconomic FRED-MD data increases substantially, with random subspace methods outperforming all competing methods for at least 66% of the series.

Keywords: dimension reduction, forecasting, random subspace
JEL codes: C32, C38, C53, C55

*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 Andreas Pick, Richard Paap, Michel van der Wel, and conference participants at the 27th $(EC)^2$ Conference on Big Data (2016) for helpful discussions. We thank SURFsara for access to the Lisa Compute Cluster.

1 Introduction

Due to the increase in available macroeconomic data, dimension reduction methods have become an indispensable tool for accurate forecasting. One well-known approach to reduce the dimension of the predictor set is to identify a small set of factors that drive most of the variation in the high-dimensional predictor set, as in Stock and Watson (2002, 2006) and Bai and Ng (2006, 2008). Whether one uses the original predictor set or the extracted factors, selection of the relevant predictors is commonly subject to substantial uncertainty. Consequently, employing model selection and shrinkage methods that estimate inclusion weights for the predictors increases the forecast variance (Ng, 2013).

A seemingly naive strategy is to forgo data-based shrinkage or selection, and assign random weights to the predictors. Although a priori there seems to be little reason to expect this approach to lead to accurate forecasts, empirical evidence suggests otherwise. For example, Elliott et al. (2013, 2015) find that averaging over forecasts constructed from many randomly selected subsets of fixed size substantially lowers the mean squared forecast error compared with data-driven alternatives. The theoretical justification of such randomized approaches is not completely understood. We provide both theoretical and extensive empirical evidence for the intriguingly strong performance of random subspace methods.

We distinguish two different approaches to constructing a random subspace. The first method we consider is random subset regression, where a randomly chosen subset of predictors is used to estimate a low-dimensional approximation to the original model and construct a forecast. The forecasts from many such submodels are then combined in order to lower the mean squared forecast error (MSFE).

Instead of selecting a subset of available predictors, random projection regression forms a low-dimensional subspace by averaging over predictors using random weights drawn from a standard normal distribution. Although not required in the setup here, the justification for this method is usually derived from the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984), which has very recently inspired several applications in the econometric literature on discrete choice models by Chiong and Shum (2016), forecasting product sales by Schneider and Gupta (2016), and forecasting using large vector autoregressive models by Koop et al. (2016) based on the framework of Guhaniyogi and Dunson (2015).

There are many random sampling methods which are widely used in the statistical and machine learning literature but rather new to economics (Ng, 2015).

Bagging or bootstrap aggregation also selects a subset of available predictors, but differs from random subset regression in that each submodel is subject to some form of model selection. Averaging over the submodels serves to smooth selection errors (Breiman, 1996; Bühlmann and Yu, 2002; Inoue and Kilian, 2008). Similar to random projections, Frieze et al. (2004) and Mahoney and Drineas (2009) construct a new set of predictors by using predictor weights. However, these weights are drawn from distributions that depend on the original set of predictors. Ma et al. (2015) discuss related sampling methods focusing on a large number of observations instead of a large set of predictors. Furthermore, the random subspace methods we consider in this paper differ from alternatives by using random weights that are independent of the data, involve a single tuning parameter, are less time consuming, and are extremely simple to implement.

We derive expressions for the upper bound on the asymptotic MSFE for random subset and random projection regression and use these bounds to determine in which settings the methods are most effective. A direct comparison between the two random subspace methods can be made when the predictors are uncorrelated. This setting nevertheless brings out the main features we observe in general settings studied in Monte Carlo experiments. The bounds elicit that random projection regression shares certain properties with ridge regression. It achieves a low MSFE when highly variable predictors are the ones that are most strongly related to the dependent variable. On the other hand, the bound for random subset regression only depends on the aggregate signal and not on the variance of the individual predictors. When the relevant predictors have a lower than average variance, the bound for random subset regression is lower compared to random projection regression.

For random subset regression, the construction of an upper bound on the asymptotic MSFE appears new. For random projection regression, bounds are only available for the in-sample mean squared error under fixed regressors by Maillard and Munos (2009), Kabán (2014) and Thanei et al. (2017). Our out-of-sample bound improves upon the existing results for the in-sample mean squared error.

The bounds are derived for forecasts that take the expected value over the random subspaces. In practice, we have to settle for a finite number of draws. We show that this has a negligible effect on the asymptotic MSFE when the number of draws scales linearly with the number of predictors, up to a logarithmic factor. This explains why Elliott et al. (2013) find no deterioration in performance when not all subsets are used, which would require a number of draws exponential in

the number of predictors.

The theoretical findings are confirmed in a set of Monte Carlo experiments, which also compare the performance of the randomized methods to several well-known alternatives: principal component regression, based on Pearson (1901), partial least squares by Wold (1982), ridge regression by Hoerl and Kennard (1970) and the lasso by Tibshirani (1996). Both randomized methods offer superior forecast accuracy over principal component regression, even in some cases when the data generating process is specifically tailored to suit this method. The random subspace methods outperform the lasso unless there is a small number of very large non-zero coefficients. Ridge regression is outperformed for a majority of the settings where the coefficients are not very weak. When the data exhibits a factor structure, and factors associated with intermediate eigenvalues drive the dependent variable, random subset regression is the only method that outperforms the historical mean of the data.

We empirically test the theoretical and Monte Carlo findings using monthly macroeconomic series in the FRED-MD dataset, introduced by McCracken and Ng (2016). Random subset regression provides the lowest MSFE relative to the benchmark models for at least 66% of the 130 series, followed by random projection regression. For both random subspace methods, the accuracy is shown to be substantially less dependent on the dimension of the reduced subspace than it is in case of principal component regression. Moreover, the dimension of the subspace should be chosen relatively large (≥ 20). This stands in stark contrast to what is common for principal component regression, where one often uses a small number of factors, see for example Stock and Watson (2012). We show how the average weights of the predictors in the random subspaces provide insight in the main drivers of the forecasts of the random subspace methods.

The article is structured as follows. Section 2 introduces the random subspace methods. The theoretical results on the forecast performance of these methods are derived in Section 3. A Monte Carlo study in Section 4 highlights the performance of the techniques under different model specifications. Section 5 considers an extensive empirical application using monthly macroeconomic data. Section 6 concludes.

2 Methods

Consider the model

$$y_{t+1} = w_t'\beta_w + x_t'\beta_x + \varepsilon_{t+1}, \tag{1}$$

where w_t is a $p_w \times 1$ vector of variables that are always included in the model, x_t is a $p_x \times 1$ vector of variables which potentially contain information on y_{t+1}, and the forecast error is denoted by ε_{t+1}. The time index t runs from $t = 0, \ldots, T$.

We assume that $E[\varepsilon_{t+1}|w_t, x_t] = 0$ and $E[\varepsilon_{t+1}^2|w_t, x_t] = \sigma^2$. Further assumptions on the sequence $\{w_t, x_t, \varepsilon_{t+1}\}$ will be given in Section 3. Under these assumptions, both w_t and x_t can contain lags of y_{t+1} or they can consist of factors derived from an additional set of observed variables.

We study the properties of point forecasts $\hat{y}_{T+1}$ for y_{T+1} when the number of available predictors p is large and fixed, the predictors in x_t are weakly related to y_{t+1}, and $T \to \infty$. The predictors $z_t = (w_t', x_t')'$, with $t = 0, \ldots, T-1$, are used in the estimation of the $p \times 1$ parameter vector $\beta = (\beta_w', \beta_x')'$, and $z_T = (w_T', x_T')'$ is only used for the construction of the forecast for y_{T+1}.

Estimating β by ordinary least squares (OLS) yields the following forecast,

$$\hat{y}_{T+1}^{\text{OLS}} = z_T'\hat{\beta} = z_T'(Z'Z)^{-1}Z'y, \tag{2}$$

where $y = (y_1, \ldots, y_T)'$, $Z = (z_0, \ldots, z_{T-1})'$, and $\hat{\beta}$ is the OLS estimator. The asymptotic mean squared forecast error equals, see for example Elliott et al. (2015),

$$\mathrm{E}_\varepsilon \left[\lim_{T\to\infty} T\mathrm{E}_{z_T} \left[\left(y_{T+1} - z_T'\hat{\beta} \right)^2 \right] \right] = \sigma^2 + \sigma^2 p. \tag{3}$$

The first term on the right-hand side arises from the noise term ε_{T+1}, which is incurred by any forecasting method. To save on notation, we set $\varepsilon_{T+1} = 0$ in the remainder of this paper.

2.1 Random subspace methods

Since the MSFE under OLS estimates increases with the number of estimated coefficients, the forecast in (2) gets inaccurate when x_t contains a large number of predictors. To prevent this, we reduce the dimensionality of the predictor set by multiplying x_t with a $p_x \times k$ matrix R, where $k < p_x$, to obtain the approximating model

$$y_{t+1} = w_t'\beta_w + x_t'R\beta_{x,R} + u_{t+1}. \tag{4}$$

The construction of the matrix R is often data-driven. Model selection methods based on information criteria effectively estimate R as a selection matrix based on the available data. Principal component regression takes R as the matrix of principal component loadings corresponding to the k largest eigenvalues from

the sample covariance matrix of the regressors x_t. The key to random subspace methods is to generate the elements of R from a probability distribution that is independent of the data. We consider the following two choices for R, which yield random subset regression and random projection regression.

2.1.1 Random subset regression

In random subset regression (RS), the matrix R is a random selection matrix that selects a random set of k predictors out of the original p_x available predictors. For example, if $p_x = 5$ and $k = 3$, a possible realization of R is

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5}$$

More in general, define an index $l = 1, \dots k$ with k the dimension of the subspace, and a scalar $c(l)$ such that $1 \leq c(l) \leq p_x$. Denote by $e_{c(l)}$ a p_x-dimensional vector with all zeros except for the $c(l)$-th entry that equals one, then random subset regression is based on random matrices of the form

$$\left[e_{c(1)}, \dots, e_{c(k)}\right], \quad e_{c(m)} \neq e_{c(n)} \text{ if } m \neq n. \tag{6}$$

2.1.2 Random projection regression

Instead of selecting a subset of predictors, we can also take weighted averages to construct a new set of predictors. Random projection regression (RP) chooses the weights at random from a normal distribution. In this case, each entry of R is independent and identically distributed as

$$[R]_{ij} \sim N(0, 1), \quad 1 \leq i \leq p_x, \ 1 \leq j \leq k. \tag{7}$$

2.2 Forecasts from low-dimensional models

We rewrite the approximating model (4) as

$$y_{t+1} = z_t' S_R \beta_R + u_{t+1}, \text{ with } S_R = \begin{pmatrix} I_{p_w} & O \\ O & R \end{pmatrix}. \tag{8}$$

The least squares estimator of β_R is denoted by $\hat{\beta}_R$ and given by

$$\hat{\beta}_R = (S_R' Z' Z S_R)^{-1} S_R' Z' y. \tag{9}$$

Using this estimate, we construct a forecast as

$$\hat{y}_{T+1,R} = z_T' S_R \hat{\beta}_R. \tag{10}$$

If R is a random matrix, then intuitively, relying on a single realization is suboptimal and we can improve upon (10). By Jensen's inequality, we find that averaging over forecasts based on different realizations of R will lower the expected asymptotic MSFE compared to an individual forecast,

$$\mathrm{E}_\varepsilon \left[\lim_{T\to\infty} T \mathrm{E}_{z_T} \left[\left(y_{T+1} - \mathrm{E}_R \left[\hat{y}_{T+1,R} \right] \right)^2 \right] \right] \leq \mathrm{E}_R \left[\mathrm{E}_\varepsilon \left[\lim_{T\to\infty} T \mathrm{E}_{z_T} \left[\left(y_{T+1} - \hat{y}_{T+1,R} \right)^2 \right] \right] \right],$$

where E_R denotes the expectation with respect to the random matrix R. Therefore, we forecast y_{T+1} as

$$\hat{y}_{T+1} = \mathrm{E}_R \left[\hat{y}_{T+1,R} \right]. \tag{11}$$

In practice, we need to replace the expectation with a finite sum. In Section 3.2, we show that this does not affect the mean squared forecast error as long as the number of draws of R is of $O\left(\frac{p_x \log p_x}{k}\right)$. This also implies that for a sufficient number of draws, forecasters that use a different sequence of random matrices will obtain the same forecast accuracy.

3 Theoretical results

The results in this section are based on the linear regression model defined in (1) and the following additional assumptions on the regressors z_t and error terms ε_{t+1}. Consider the time index $t = 0, \ldots, T$, and the parameter index $i = 1, \ldots, p$. Denote by Δ a finite constant independent of the dimensions p and T.

A1 $\{z_t', \varepsilon_{t+1}\}$ *is a strong mixing sequence of size* $a = -r/(r-2)$, $r > 2$.

A2 $E[\varepsilon_{t+1} | z_{ti}] = 0$.

A3 $E[\varepsilon_{t+1}^2 | z_{ti}] = \sigma^2$.

A4 $E|z_{ti}\varepsilon_{t+1}|^r \leq \Delta < \infty$.

A5 $E[z_t z_t'] = \Sigma_z = \begin{pmatrix} \Sigma_w & \Sigma_{wx} \\ \Sigma_{xw} & \Sigma_x \end{pmatrix}$ *is positive definite.*

A6 $V_n = Var(T^{-1/2}Z'\varepsilon)$ *is uniformly positive definite.*

A7 $E|z_{ti}^2|^{r/2+\delta} \leq \Delta < \infty$.

Under these assumptions we derive theoretical results that apply to weakly dependent time series models. In particular, they allow both w_t and x_t to contain lagged values of the dependent variable.

The mixing size a in Assumption 1 is defined as in White (1984), Definition 3.42. In addition to standard results on asymptotic normality, the strong mixing assumption allows us to establish independence between z_T and the estimation error $\sqrt{T}(\hat{\beta} - \beta)$, as we show in Appendix A.1. This independence is essential to the proof of our main theorem. The necessity for this independence has been noted in Hansen (2008), and appears to be implied in Equation (2.2) of Hirano and Wright (2017).

Together, Assumptions A1-A7, guarantee that

$$\frac{1}{\sqrt{T}}Z'\varepsilon \xrightarrow{(d)} N(0, \Sigma_z), \quad \operatorname*{plim}_{T\to\infty} \frac{1}{T}Z'Z = \Sigma_z, \tag{12}$$

see for example White (1984).

We make one additional assumption with regard to the strength of the predictors, which rules out the possibility to consistently estimate β as $T \to \infty$.

A8 *The parameter vector* β *is local-to-zero, i.e.*

$$\beta_x = \frac{1}{\sqrt{T}}\beta_{x,0}, \tag{13}$$

where $\beta_{x,0} = O(1)$.

Under local-to-zero coefficients, the bias induced by using a low-dimensional subspace is finite, see Claeskens and Hjort (2008). When coefficients are stronger than in Assumption A8, the forecast based on OLS estimation in (2) using p variables is asymptotically the optimal forecast.

The theoretical results also suit forecasting models that assume a factor structure in x_t, such as the diffusion index model (Stock and Watson, 2002). In this case, if the factors are only weakly related to the dependent variable as in Assumption A8, the diffusion index model can be treated along the same lines as (1) upon replacing x_t with p_f common factors in f_t. It is common to treat p_f as fixed and let p_x grow with T. The forecast error distribution for this model is derived by Bai and Ng (2006). Their results show that if $p_x/T \to \infty$, estimation

of the factors does not affect the forecast distribution. If $p_x/T = O(1)$, an additive term enters due to the estimation error in the factors. This term is not affected by the methods in this paper, so that the MSFE only incurs an additional term independent of R.

3.1 MSFE for forecasts from low-dimensional models

Denote the asymptotic mean squared forecast error of (11) as

$$\rho(k) = E_\varepsilon \left[\lim_{T\to\infty} T \mathrm{E}_{z_T} \left[\left(z_T'\beta - z_T' \mathrm{E}_R \left[S_R \hat{\beta}_R \right] \right)^2 \right] \right]. \tag{14}$$

The following theorem provides a bound on the asymptotic mean squared forecast error for matrices R which can be deterministic or random.

Theorem 1 *Let $R \in \mathbb{R}^{p_x \times k}$ be a matrix such that $E_R[RR'] = \frac{k}{p_x} I_{p_x}$. The asymptotic mean squared forecast error $\rho(k)$ in (14) under (1) satisfying Assumption A1-A8, is upper bounded by*

$$\rho(k) \leq \sigma^2(p_w + k) + \beta_{x,0}'\Sigma_x\beta_{x,0} - \beta_{x,0}\Sigma_x \left(\frac{p_x}{k} E_R \left[RR'\Sigma_x RR' \right] \frac{p_x}{k} \right)^{-1} \Sigma_x\beta_{x,0}. \tag{15}$$

A proof is presented in Appendix A.2. Theorem 1 holds for general matrices R after suitable scaling.

The first term of (15) represents the variance of the estimates. This can be compared to the variance that is achieved by forecasting using OLS estimates for β, which is equal to $\sigma^2 p = \sigma^2(p_w + p_x)$. In empirical applications, we expect p_w to be small, as w_t usually only contains a constant and a small number of lags. The number of additional variables p_x can however be large, and hence, the reduction in variance to k can be substantial.

The remaining terms in (15) reflect the bias that arises by projecting x_t to a low-dimensional subspace. If any signal is present, this bias is strictly smaller then the bias of the naive estimator that does not use any of the predictors, which equals $\beta_{x,0}'\Sigma_x\beta_{x,0}$.

Loosely speaking, the product $\frac{p_x}{k} RR'\Sigma_x RR' \frac{p_x}{k}$ first projects Σ_x to a k-dimensional subspace by multiplying with R from the left and the right, and then re-inflates by another multiplication with R. If little information is lost in this procedure, then the expectation will be close to Σ_x, in which case the bias is small.

For both random subset regression and random projection regression, the bound in (15) can be evaluated explicitly. We start with random subset regression.

3.1.1 MSFE bound for random subset regression

For the random selection matrices in (6) we have the following result.

Lemma 1 *Let $R \in \mathbb{R}^{p_x \times k}$ be a random selection matrix and Σ_x a positive definite matrix. Then, $E_R[RR'] = \frac{k}{p_x} I_{p_x}$, and*

$$E_R\left[RR'\Sigma_x RR'\right] = \frac{k}{p_x}\left(\frac{k-1}{p_x-1}\Sigma_x + \frac{p_x-k}{p_x-1}D_{\Sigma_x}\right), \tag{16}$$

where $[D_{\Sigma_x}]_{ii} = [\Sigma_x]_{ii}$, and $[D_{\Sigma_x}]_{ij} = 0$ if $i \neq j$.

A proof is provided in Appendix A.3.

Using Lemma 1 in the bound from Theorem 1, we obtain the following bound on the MSFE for random subset regression.

Corollary 1 *For random subset regression, the asymptotic mean squared forecast error $\rho(k)$ in (14) under (1) satisfying Assumption A1-A8, is upper bounded by*

$$\rho(k) \leq \sigma^2(p_w + k) + \beta'_{x,0}\Sigma_x\beta_{x,0} - \frac{k}{p_x}\beta'_{x,0}\Sigma_x\left(\frac{k-1}{p_x-1}\Sigma_x + \frac{p_x-k}{p_x-1}D_{\Sigma_x}\right)^{-1}\Sigma_x\beta_{x,0}.$$

The bound for random subset regression depends on a convex combination $u\Sigma_x + (1-u)D_{\Sigma_x}$, for $0 \leq u \leq 1$. All weight is put on D_{Σ_x} when $k = 1$, which implies that all information on cross-correlations is lost in the low-dimensional subspace. When $k = p_x$, the bound reduces to the exact expression for OLS using p predictors as in (3).

3.1.2 MSFE bound for random projection regression

When R is constructed as in (7), the columns are not exactly orthogonal. Potentially, the lack of orthogonality of R results in an unnecessary loss of information compared to the use of a $p_x \times k$ matrix Q with orthogonal columns. However, the following lemma states that no such loss occurs.

Lemma 2 *Suppose R is a $p_x \times k$ matrix of independent standard normal random variables, $Q = R(R'R)^{-1/2}$ a $p_x \times k$ matrix with orthogonal columns, and $P = (R'R)^{1/2}$ an invertible $k \times k$ matrix, then*

$$\rho(k) = E_\varepsilon\left[\lim_{T\to\infty} T E_{z_T}\left[\left(z'_T\beta - z'_T E_Q\left[S_Q\hat{\beta}_Q\right]\right)^2\right]\right]. \tag{17}$$

A proof is provided in Appendix A.4.

By Lemma 2 we can replace R in Theorem 1 by Q, even though we are using R in the construction of the estimator. To complete the bound from Theorem 1, we then need the following.

Lemma 3 *Let $R \in \mathbb{R}^{p_x \times k}$ be a matrix of independent standard normal entries, and define $Q = R(R'R)^{-1/2} \in \mathbb{R}^{p_x \times k}$. Furthermore, let Σ_x be a positive definite matrix. Then, $E_Q[QQ'] = \frac{k}{p_x} I_{p_x}$, and*

$$E_Q[QQ'\Sigma_x QQ'] = \frac{k}{p_x}\left(\frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\Sigma_x + \frac{(p_x-k)p_x}{(p_x+2)(p_x-1)}\frac{trace(\Sigma_x)}{p_x} I_{p_x}\right).$$

A proof is provided in Appendix A.5, which relies on somewhat tedious calculations of the fourth order moments of the elements of the matrix Q.

Using Lemma 3 in the bound from Theorem 1, we obtain a bound on the asymptotic mean squared forecast error for random projection regression.

Corollary 2 *For random projection regression, the asymptotic mean squared forecast error $\rho(k)$ in* (14) *under* (1) *satisfying Assumption A1-A8, is upper bounded by*

$$\begin{aligned}\rho(k) \leq\ & \sigma^2(p_w + k) + \beta'_{x,0}\Sigma_x\beta_{x,0} \\ & - \frac{k}{p_x}\beta'_{x,0}\Sigma_x\left(\frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\Sigma_x + \frac{(p_x-k)p_x}{(p_x+2)(p_x-1)}\frac{trace(\Sigma_x)}{p_x} I_{p_x}\right)^{-1}\Sigma_x\beta_{x,0}.\end{aligned}$$

The bound for random projection regression depends on a convex combination $u\Sigma_x + (1-u)\frac{\text{trace}(\Sigma_x)}{p_x}$. When $k = 1$, nearly all weight is put on $\text{trace}(\Sigma_x)$, while when $k = p_x$, all weight is put on Σ_x and the bound reduces to (3).

Maillard and Munos (2009) provide a bound on the in-sample mean squared error under fixed regressors for random projection regression, which was subsequently improved by Kabán (2014). Thanei et al. (2017) arrive at a similar expression as in (15), and use the expressions in Kabán (2014) to evaluate the expectation. However, their bound is suboptimal. For example, it has the unattractive feature of not reducing to (3) when k is set equal to p_x. The bound in Corollary 2 solves this problem, by noting that we can rely on the matrix Q, which has orthogonal columns, instead of R in the calculations. Appendix A.6 shows that the resulting bound is uniformly tighter than the currently available bounds.

3.1.3 Comparison between the MSFE bounds of RS and RP

Based on the difference between the expressions for the MSFE bounds for random subset and random projection regression in Corollary 1 and 2, we show that there exists no covariance matrix Σ_x for which one of the methods offers a superior bound uniformly over all possible parameter vectors $\beta_{x,0}$.

The difference in the bounds is given by

$$\Delta = \beta_{x,0}\Sigma_x \left(M_{RP}^{-1} - M_{RS}^{-1}\right) \Sigma_x \beta_{x,0}, \tag{18}$$

where

$$\begin{aligned} M_{RP} &= \frac{k}{p_x}\left(\frac{k-1}{p_x-1}\Sigma_x + \frac{p_x-k}{p_x-1}D_{\Sigma_x}\right)^{-1}, \\ M_{RS} &= \frac{k}{p_x}\left(\frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\Sigma_x + \frac{(p_x-k)p_x}{(p_x+2)(p_x-1)}\frac{\text{trace}(\Sigma_x)}{p_x}I_{p_x}\right)^{-1}. \end{aligned}$$

If $\Delta > 0$, then the bound for random projection regression lies above the bound for random subset regression. Denote $A - B \succ 0$ if $A - B$ is a positive definite matrix. If $M_{RP}^{-1} - M_{RS}^{-1} \succ 0$, then $\Delta > 0$ uniformly over the choice of $\beta_{x,0}$. This occurs if and only if $M_{RP} - M_{RS} \prec 0$, where

$$M_{RP} - M_{RS} = \frac{p_x-k}{p_x-1}\left[\frac{2}{p_x+2}\left(\Sigma_x - D_{\Sigma_x}\right) + \frac{p_x}{p_x+2}\left(\frac{\text{trace}(\Sigma_x)}{p_x}I_{p_x} - D_{\Sigma_x}\right)\right].$$

Unless Σ_x is a multiple of the identity matrix, subtracting D_{Σ_x} yields an indefinite matrix. This is easily seen as the sum of the eigenvalues of $\Sigma_x - D_{\Sigma_x}$ equals the trace, which is identically equal to zero. Similarly, unless D_{Σ_x} is a multiple of the identity matrix, the second term yields an indefinite matrix. Hence, there does not exist a covariance matrix Σ_x for which $M_{RP} - M_{RS} \prec 0$, and hence where one of the methods outperforms the other uniformly over the choice of $\beta_{x,0}$.

However, we can distinguish cases in which the random subspace methods are expected to perform equally well or outperform each other when we take the relation between the covariance matrix of the regressors and the coefficients of the regressors into account. We consider a simplified setting based on (1) with Σ_x a diagonal $p_x \times p_x$ matrix, for which the bounds in Corollary 1 and 2 simplify to

$$\begin{aligned} \rho(k)^{RS} &\leq \sigma^2(p_w+k) + \beta_{x,0}'\Sigma_x\beta_{x,0}\left(1 - \frac{k}{p_x}\right), \\ \rho(k)^{RP} &\leq \sigma^2(p_w+k) + \beta_{x,0}'\Sigma_x\left[I_{p_x} - \frac{k}{p_x}D(\Sigma_x)\right]\beta_{x,0}, \end{aligned} \tag{19}$$

respectively, where

$$[D(\Sigma_x)]_{ii} = \frac{\lambda_i}{u\lambda_i + (1-u)\bar{\lambda}}, \quad u = \frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}, \quad \bar{\lambda} = \frac{1}{p_x}\sum_{i=1}^{p_x}\lambda_i, \tag{20}$$

where $\lambda_1, \ldots, \lambda_{p_x}$ are the eigenvalues of Σ_x in decreasing order.

For a well-conditioned covariance matrix, i.e. $\lambda_i \approx \bar{\lambda}$ which means that the eigenvalues of the covariance matrix are of the same size, we have $D(\Sigma_x) \approx I$. From (19) we infer that in this scenario, the methods are expected to perform equally well.

When the eigenvalues of the covariance matrix of the regressors are not of the same size, two things can happen. First, consider a typical principal component regression setting where the nonzero values of $\beta_{x,0}$ are associated with eigenvalues that are larger than the average eigenvalue. For random projection regression, $[D(\Sigma_x)]_{ii} > 1$ when $\lambda_i > \bar{\lambda}$. Therefore random projection will offer a superior bound compared to random subset regression in this case. In this sense, the behavior of random projection regression appears similar to that of ridge regression, in that it performs most shrinkage on small eigenvalues.

In contrast, it is also possible that the factor associated with the largest eigenvalue of the covariance matrix is not associated with the dependent variable. Random subset regression does not assume that large eigenvalues in Σ_x are informative on the relative importance with respect to y. Since in the bound for random projection regression it holds that $[D(\Sigma_x)]_{ii} < 1$ if $\lambda_i < \bar{\lambda}$, random subset regression now offers a superior bound.

3.1.4 Comparison between the MSFE of RS, RP, and OLS

Here we study the performance of the random subspace methods relative to OLS for different signal strength, in the same setting as the previous section.

Based on the MSFE bound, we find that for small signal strength, random subset regression outperforms OLS. Equating the exact MSFE of OLS in (2) to the bound for the MSFE of random subset regression results in the following condition,

$$\frac{\beta'_{x,0}\Sigma_x\beta_{x,0}}{\sigma^2 p_x} = 1, \tag{21}$$

which implies that random subset regression outperforms OLS when the average signal strength falls below 1.

The relative performance of random projection regression to OLS depends not

only on the signal strength, but also on which coefficients in $\beta_{x,0}$ are non-zero. Therefore, condition (21) does not apply to random projection regression. If non-zero coefficients are related to larger than average eigenvalues, the bound is lower than the MSFE under OLS as long as $\frac{\beta'_{x,0}\Sigma_x\beta_{x,0}}{\sigma^2 p_x} < 1+u$, for some $u > 0$. When non-zero coefficients are related to smaller than average eigenvalues, we obtain $\frac{\beta'_{x,0}\Sigma_x\beta_{x,0}}{\sigma^2 p_x} < 1-u$, for $u > 0$.

3.1.5 Quality MSFE bounds

To provide insight in the quality of the bounds obtained in Corollary 1 and 2, we consider a setting in which we obtain an expression for the exact MSFE. For random subset regression this is achieved when the regressors are independent. If in addition we assume the variances of the regressors to be equal, we also obtain an exact expression under random projection regression.

When $\Sigma_z = D_{\Sigma_z}$, we have that $\mathrm{E}_R[S_R\hat{\beta}_R]$ in (14) boils down to

$$\mathrm{E}_R[S_R\hat{\beta}_R] = \begin{pmatrix} (W'W)^{-1}W' \\ \mathrm{E}_R[R(R'X'XR)^{-1}R']X' \end{pmatrix} y, \tag{22}$$

where $W = (w_0, \ldots, w_{T-1})'$ and $X = (x_0, \ldots, x_{T-1})'$. When $\frac{1}{T}X'X$ converges to a diagonal matrix, we can explicitly evaluate the expectation for random subset regression,

$$\mathrm{E}_R[R(R'D_{\Sigma_x}R)^{-1}R'] = \frac{k}{p_x}D_{\Sigma_x}^{-1}, \tag{23}$$

where R is a random permutation matrix. This follows from the fact that each diagonal element of Σ_x is selected with probability k/p_x in random subset regression, see Appendix A.3.

We obtain the same result for random projection regression under independent predictors with equal variance, $\Sigma_x = cI_{p_x}$,

$$c^{-1}\mathrm{E}_R[R(R'R)^{-1}R'] = c^{-1}\mathrm{E}_Q[QQ'] = \frac{k}{p_x}c^{-1}I_{p_x}, \tag{24}$$

where the expression for the second moment follows from Lemma 3.

Subsequently, the exact MSFE for both random subspace methods is given by

$$\rho(k) = \sigma^2\left(p_w + k\frac{k}{p_x}\right) + \beta'_{x,0}\Sigma_x\beta_{x,0}\left(1-\frac{k}{p_x}\right)^2. \tag{25}$$

In case of independent regressors, the bounds in Corollary 1 and 2 simplify to

(19). Since we assume $\Sigma_x = c \cdot I_{p_x}$ for the bound of random projection regression, $[D(\Lambda)]_{ii} = 1$, and the bounds of the random subspace methods are identical.

Comparing the exact MSFE from (25) to the bounds in (19), we see that the bounds overestimate the variance by a factor p_x/k, and the bias by a factor $(1 - k/p_x)^{-1}$. The difference is maximized for $\frac{k}{p_x} = \frac{1}{2}$ in which case the bounds are conservative by at most a factor $\frac{1}{2}$.

As an alternative to the upper bound on the MSFE in Theorem 1, the MSFE can be bounded by bounding the eigenvalues of the expectation over the random matrix R. Using the eigenvalue inequalities in Appendix A.7, we derive both a conservative upper and lower bound on the MSFE in Appendix A.8. Since these bounds ignore the eigenvalue structure of the covariance matrix of the predictors, these bounds are in almost all cases uninformative. Furthermore, the bounds are identical for random subset regression and random projection regression. They therefore do not elicit the difference between the two methods.

3.2 Feasibility of the MSFE bounds

The bounds from the previous section are based on forecasts that depend on the expectation over the random matrix R. In practice, we need to approximate this expectation by using a finite number of draws of the matrix R. For the feasibility of the method in practice, it is important that the required number of draws is not too large. If one would have to draw all possible subsets of size k from p_x predictors, the number of required draws is exponential in p_x, limiting the practical use of the methods. The following theorem guarantees that in order to get close to the expectation, we only require a number of draws that is linear in p_x, up to logarithmic factors.

Theorem 2 *Let $\hat{y}_{T+1,S} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_{T+1,R_i}$, with $\hat{y}_{T+1,R_i}$ as in (10) where R_i is a realization of the random matrix R, and $\hat{y}_{T+1}$ as in (11). Denote by $\rho_S(k)$ the asymptotic mean squared forecast error based on $\hat{y}_{T+1,S}$, and denote by $\rho(k)$ the asymptotic mean squared forecast error based on $\hat{y}_{T+1}$ as in (14). Furthermore, let $N = O\left(\frac{p_x \log p_x}{k}\right)$. Then for an arbitrarily small constant ϵ,*

$$\rho_S(k) = (1 + \epsilon)\rho(k). \tag{26}$$

A proof is provided in Appendix A.9.

This result shows the feasibility of random subset regression in practice. It also provides a theoretical justification of the results obtained in Elliott et al. (2013)

and Elliott et al. (2015), where it was found that little prediction accuracy is lost by using a relatively small number of random subsets instead of all available subsets. Instead of drawing a number of subsets exponential in p_x, $N = \binom{p_x}{k} = O\left(\left[\frac{p_x}{k}\right]^k\right)$, which is the case for complete subset regression, we only require a number of draws linear in p_x.

4 Monte Carlo experiments

We examine the practical implications of the theoretical results in a Monte Carlo experiment. In a first set of experiments we show the effect of sparsity and signal strength on the MSFE, and a second set of experiments shows in which settings one of the random subspace methods is preferred over the other. The prediction accuracy of the random subspace methods is evaluated relative to several widely used alternative regularization techniques.

4.1 Monte Carlo set-up

The set-up we employ parallels Elliott et al. (2015). The data generating process takes the form

$$y_{t+1} = x_t'\beta_x + \varepsilon_{t+1}, \tag{27}$$

where x_t is a $p_x \times 1$ vector with predictors, β_x a $p_x \times 1$ coefficient vector, ε_{t+1} an error term with $\varepsilon_{t+1} \sim N(0, \sigma_\varepsilon^2)$, and $t = 0, \ldots, T$. In each replication of the Monte Carlo simulations, predictors are generated by drawing $x_t \sim N(0, \Sigma_x)$, after which we standardize the predictor matrix. The covariance matrix of the predictors equals $\Sigma_x = \frac{1}{p_x} P'P$, where P is a $p_x \times p_x$ matrix whose elements are independently and randomly drawn from a standard normal distribution. As argued by Elliott et al. (2015), this ensures that the eigenvalues of the covariance matrix are reasonably spaced.

The strength of the individual predictors is considered local-to-zero by setting $\beta_x = \sqrt{\sigma_\varepsilon^2/T} \cdot b\iota_s$ for a fixed constant b. The vector ι_s contains s non-zero elements that are equal to one. We refer to s as the sparsity of the coefficient vector. We vary the signal strength b and the sparsity s across different Monte Carlo experiments. In all experiments, the error term of the forecast period ε_{T+1} is set to zero, as this only yields an additional noise term σ^2 which is incurred by all forecasting methods.

We employ two sets of experimental designs, which mimic the high-dimensional setting in the empirical application by choosing the number of predictors $p_x = 100$ and the sample size $T = 200$. Results are based on $M = 10{,}000$ replications of the data generating process (27).

In the first set of experiments, we vary the signal to noise ratio b and the sparsity s over the grids $b \in \{0.5, 1.0, 2.0\}$ and $s \in \{10, 50, 100\}$. This allows us to study the effect of sparsity and signal strength on the MSFE and the optimal subspace dimension.

The second set of experiments reflects scenarios where random subset and random projection regression are expected to differ based on the discussion in Section 3.1.3. In this case we replace x_t in (27) by factors extracted from x_t, $t = 0, \dots, T$, using principal component analysis. Denote by f_i for $i = 1, \dots, p_x$ the extracted factors sorted by the explained variation in the predictors. In the first three experiments, we associate nonzero coefficients with the 10 factors that explain most of the variation in the predictors. We refer to this setting as the top factor setting. This setting is expected to suit random projection over random subset regression. In the remaining experiments, we associate the nonzero coefficients with factors $\{f_{46}, \dots, f_{55}\}$, which are associated with intermediately sized eigenvalues. This setting is referred to as the intermediate factor setting and expected to suit random subset regression particularly well. In both the top and intermediate factor setting, the coefficient strength b is again varied as $b \in \{0.5, 1.0, 2.0\}$.

We generate one-step-ahead forecasts by means of random projection and random subset regression using equation (4) in which we vary the subspace dimension over $k = \{1, \dots, p_x\}$. The subspace methods, as well as the benchmark models discussed below, estimate (27) with the inclusion of an intercept that is not subject to the dimension reduction or shrinkage procedure. We average over $N = 1{,}000$ predictions of the random subspace methods to arrive at a one-step-ahead forecast. This is in line with the findings in Section 3.2 which suggest to use $O(p_x \log p_x) = O(100 \cdot \log 100) = O(460)$ draws.

Benchmark models We compare the performance of the random methods with principal component (PC) regression and partial least squares (PL) regression introduced by Wold (1982). Both methods approximate the data generating process (27) as

$$y_{t+1} = w_t' \beta_w + \sum_{i=1}^{k} f_{ti} \beta_{f,i} + \eta_t, \tag{28}$$

where $k \in \{1, \ldots, p_x\}$ and w_t includes an intercept. The methods differ in their construction of the factors f_{ti}. Principal component regression is implemented by extracting the factors from the standardized predictors x_t with $t = 0, \ldots, T$ using principal component analysis. This is a diffusion index model along the lines of Stock and Watson (2002). Partial least squares uses a two-step procedure to construct the factors, as described for example by Groen and Kapetanios (2016). We use the static approach as discussed by Fuentes et al. (2015), who find good forecast performance for a similar macroeconomic forecasting exercise as in Section 5, in which the factors are extracted by applying partial least squares between the target variable y_{t+1} and the predictors x_t. We then estimate for both methods (28) and generate a forecast as $\hat{y}_{T+1} = w_T'\hat{\beta}_w + \sum_{i=1}^{k} f_{Ti}\hat{\beta}_{f,i}$. Note that the principal component regression model is correctly specified for the top factor setting in the second set of experiments.

In addition to comparing the random subspace methods to principal component regression and partial least squares, we include two widely used alternatives: ridge (RI) regression (Hoerl and Kennard, 1970) and the lasso (LA) (Tibshirani, 1996). We generate one-step-ahead forecasts using these methods by $\hat{y}_{T+1} = w_T'\hat{\beta}_w + x_T'\hat{\beta}_x$, with

$$(\hat{\beta}_w, \hat{\beta}_x) = \arg\min_{\beta_w, \beta_x} \left(\frac{1}{n} \sum_{t=0}^{T-1} (y_{t+1} - w_t'\beta_w - x_t'\beta_x)^2 + kP(\beta_x) \right). \tag{29}$$

The penalty term $P(\beta_x) = \sum_{i=1}^{p_x} \frac{1}{2}\beta_{x,i}^2$ in case of ridge regression and $P(\beta_x) = \sum_{i=1}^{p_x} |\beta_{x,i}|$ for the lasso. The penalty parameter k controls the amount of shrinkage. In contrast to the previous subspace methods, the values of k are not bounded to integers nor is there a natural grid. We consider forecasts based on equally spaced grids for $\ln k$ of 100 values; $\ln k \in \{-30, \ldots, 0\}$ for lasso and $\ln k \in \{-15, \ldots, 15\}$ for ridge regression. In general, we expect lasso to do well when the model contains a small number of large coefficients. Ridge regression, on the other hand, is expected to do well when we have many weak predictors.

Evaluation criterion We evaluate forecasts by reporting their MSFE relative to that of the prevailing mean model that takes $\bar{y}_{T+1} = \frac{1}{T}\sum_{t=0}^{T-1} y_{t+1}$. The mean squared forecast error is computed as

$$MSFE = \frac{1}{M} \sum_{j=1}^{M} \left(y_{T+1}^{(j)} - \hat{y}_{T+1}^{(j)} \right)^2, \tag{30}$$

Table 1: Simulation results: MSFE optimal subspace dimension

b	RP	RS	PC	PL	RI	LA
			$s = 10$			
0.5	0.967 (2)	0.966 (2)	1.253 (1)	9.592 (1)	0.966 (-3.3)	1.000 (-29.7)
1.0	0.864 (8)	0.865 (8)	1.056 (1)	3.099 (1)	0.864 (-2.1)	0.959 (-27.9)
2.0	0.638 (21)	0.637 (21)	0.929 (7)	0.961 (1)	0.640 (-0.6)	0.669 (-27.3)
			$s = 50$			
0.5	0.815 (10)	0.815 (10)	1.034 (1)	2.377 (1)	0.814 (-1.8)	0.961 (-27.9)
1.0	0.568 (25)	0.569 (25)	0.885 (12)	0.805 (1)	0.570 (-0.6)	0.706 (-27.3)
2.0	0.300 (46)	0.301 (46)	0.453 (43)	0.374 (2)	0.301 (0.6)	0.366 (-26.4)
			$s = 100$			
0.5	0.710 (16)	0.709 (16)	0.980 (2)	1.372 (1)	0.710 (-1.2)	0.877 (-27.6)
1.0	0.422 (36)	0.423 (35)	0.663 (29)	0.535 (1)	0.423 (0.0)	0.539 (-26.7)
2.0	0.188 (56)	0.189 (56)	0.268 (59)	0.227 (3)	0.189 (1.2)	0.242 (-26.1)

Note: this table reports the MSFE relative to the benchmark of the prevailing mean, for the subspace dimension corresponding to the minimum MSFE which is given in parentheses.

where $y_{T+1}^{(j)}$ is the realized value and $\hat{y}_{T+1}^{(j)}$ the predicted value in the jth replication of the Monte Carlo simulation. The number of replications M is set equal to $M = 10{,}000$.

4.2 Simulation results

4.2.1 Sparsity and signal strength

Table 1 shows the Monte Carlo simulation results for the first set of experiments for the value of k that yields the lowest MSFE. Results for different values of k are provided in Table 7 in Appendix B. The predictive performance of each forecasting method is reported relative to the prevailing mean. Values below one indicate that the benchmark model is outperformed.

We find that in general, a lower degree of sparsity results in a lower relative MSFE. Since the predictability increases in s, it is not surprising that a less sparse setting results in better forecast performance relative to the prevailing mean, which ignores all information in the predictors. Similarly, the prediction accuracy also clearly increases with increasing signal strength. The results for different values of k, reported in Table 7 in Appendix B, show that increasing the subspace dimension in case of a weak signal worsens the performance, due to the increasing effect of the parameter estimation error when the predictive signal is small. This dependency

on k tends to decrease for large values of s and b, where we observe smaller differences between the predictive performance over the different values of k.

Comparing the random subspace methods, we find that in these experiments, as expected, the predictive performance of random projection regression and random subset regression is almost the same. Table 1 shows that when choosing the optimal subspace dimension, these methods outperform both the prevailing mean as principal component regression and partial least squares for each setting. Lasso is not found to perform well. Only in the extremely sparse settings where $s = 10$ and b increases, its performance tends towards the random subspace methods. Ridge regression yields similar prediction accuracy as the random subspace methods. For strong signals, the random subspace methods perform better, whereas for very weak signals ridge regression appears to have a slight edge.

Table 1 shows that the optimal subspace dimension increases with both the sparsity s and the signal strength governed by b. Interestingly, random subset regression and random projection regression select, apart from one setting, exactly the same subspace dimension. The number of factors selected in principal component regression is lower for almost all settings. The results for partial least squares reflect that in settings with a small number of weak predictors, the factors cannot be constructed with sufficient accuracy. In these settings, more accurate forecasts are therefore obtained by ignoring the factors altogether. Note that where the parameter k has an intuitive appeal in the dimension reduction methods, the values in the grid of k for lasso and ridge regression methods lack interpretation.

4.2.2 Experiments using a factor design

The small differences between random subset and random projection regression in the previous experiments stand in stark contrast with the findings on the factor structured experiments. The relative MSFE for the choice of k that yields the lowest MSFE compared to the prevailing mean is reported in Table 2. Table 8 in in Appendix B shows results for different values of k. We observe precisely what was anticipated based on the discussion in Section 3.1.3. In the top factor setting, where the nonzero coefficients are associated with the factors corresponding to the largest 10 eigenvalues, random projection regression outperforms random subset regression by a wide margin. For a weak signal, when $b = 0.5$, it even outperforms principal component regression, which is correctly specified in this set-up. When $b = 2$, we are in a setting where we have a small number of large coefficients. As expected, this favors lasso, although not to the extend that it outperforms

Table 2: Simulation results: MSFE optimal subspace dimension - factor design

b	RP	RS	PC	PL	RI	LA
	Top factor setting					
0.5	0.722 (10)	0.955 (9)	0.992 (2)	2.466 (1)	0.721 (-1.8)	0.887 (-27.9)
1.0	0.428 (21)	0.842 (28)	0.300 (10)	0.495 (1)	0.429 (-0.9)	0.485 (-27.6)
2.0	0.205 (33)	0.580 (60)	0.078 (10)	0.139 (1)	0.206 (0.0)	0.150 (-27.3)
	Intermediate factor setting					
0.5	1.013 (1)	0.998 (1)	1.501 (1)	16.347 (1)	1.000 (-14.7)	1.000 (-29.7)
1.0	1.003 (1)	0.981 (4)	1.176 (1)	7.140 (1)	1.000 (-7.5)	1.000 (-29.1)
2.0	1.001 (1)	0.923 (16)	1.060 (1)	2.969 (1)	1.000 (-14.7)	1.000 (-29.7)

Note: this table reports the MSFE relative to the benchmark of the prevailing mean, for the subspace dimension corresponding to the minimum MSFE which is given in parentheses.

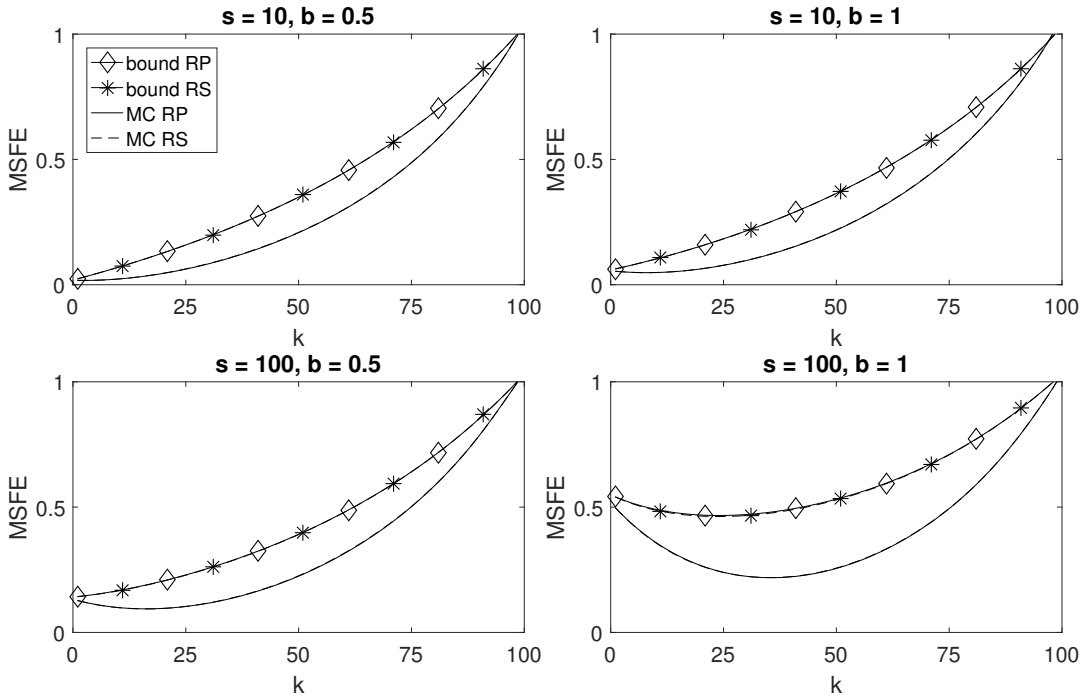
principal component regression. The findings are almost completely reversed in the intermediate factor setting, when the nonzero coefficients are associated with factors $f_{46}, \ldots, f_{55}$. Here we observe that random subset regression outperforms random projection. In fact, random subset regression is the only method that is able to extract an informative signal from the predictors and outperform the prevailing mean benchmark.

The difference in predictive performance is reflected in the optimal subspace dimension reported in parentheses in Table 2. For the top factor setting, when $b = \{1, 2\}$, we observe that the MSFE for random subset regression is minimized at substantially larger values than for random projection regression. This evidently increases the forecast error variance, and the added predictive content is apparently too small to outweigh this. Principal component regression, in turn, selects the correct number of factors when $b = \{1, 2\}$. In the intermediate factor setting, the dimension of random subset is again larger than for random projection, with an impressive difference when $b = 2$. Here, random projection is apparently not capable to pick up any signal and selects $k = 1$, while random subset regression uses a subspace dimension of $k = 16$. Lasso and ridge both choose such a strong penalization that they reduce to the prevailing mean benchmark for all choices of b.

4.3 Simulation results versus theoretical bounds

The qualitative correspondence between the simulation results and the theoretical results show that the bounds are useful to determine settings where the random

Figure 1: Simulation results: comparison with theoretical bounds



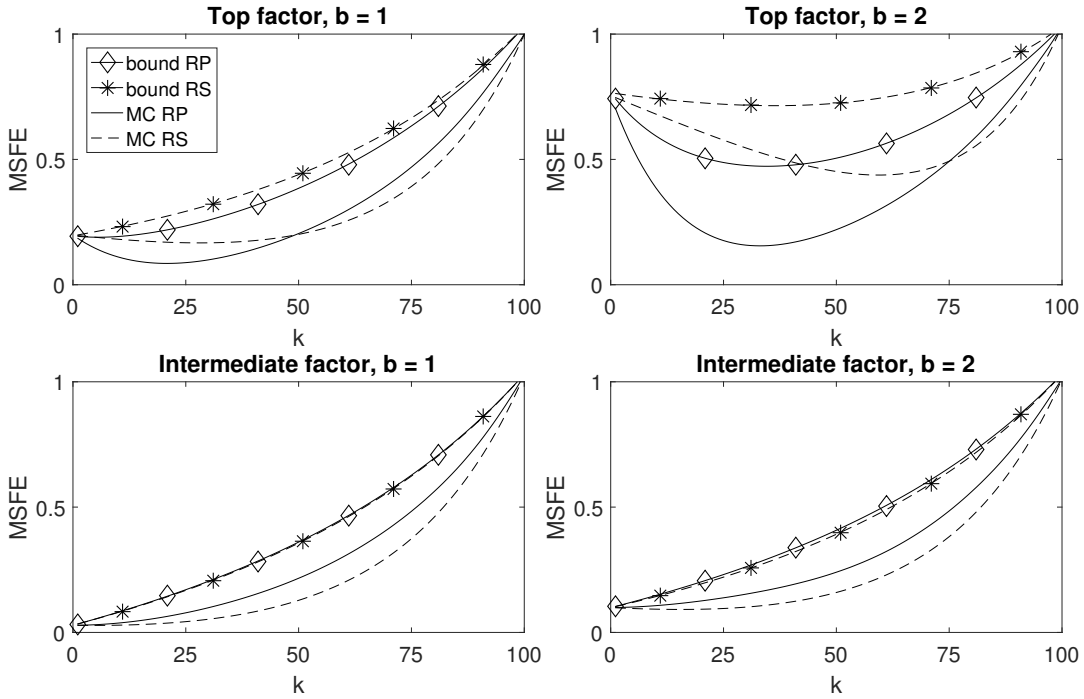
Note: this figure shows the MSFE for different values of the subspace dimension k, along with the theoretical upper bounds on the MSFE derived in Section 3.1 after a small sample size correction. The different lines correspond to the upper bound for random projections (bound RP, diamond marker), upper bound for random subsets (bound RS, asterisk marker), and the evaluation criteria for the dimension reduction methods random projections (MC RP, solid) and random subsets (MC RS, dashed). The four panels correspond to settings in which the sparsity s alternates between 10 and 100, and the signal to noise ratio parameter b between 0.5 and 1.

subspace methods are expected to do well. In this section, we investigate how close the bounds are to the exact MSFE obtained in the Monte Carlo experiments.

Figure 1 shows the MSFE over different subspace dimensions of random projection and random subset regression, along with the theoretical upper bounds on the MSFE derived in Section 3.1, for the first set of experiments described above. As we found in Table 7 in Appendix B, the values of the MSFE of the random subspace methods are almost identical to each other over the whole range of k. This also holds for the bounds. The bounds differ most from the exact MSFE from the Monte Carlo experiments for intermediate values of k when there is a strong signal and no sparsity.

In Figure 2 we show the bounds for the factor settings. Here we see that the bounds correctly indicate which method is expected to yield better results in the settings under consideration. The upper panels, corresponding to the top factor structure, show the bound for random projection to be lower. The lower panels

Figure 2: Simulation results: comparison with theoretical bounds - factor design



Note: this figure shows the MSFE for different values of the subspace dimension k, along with the theoretical upper bounds on the MSFE derived in Section 3.1 for the top and intermediate factor settings. For additional information, see the note following Figure 1.

display the MSFE in the intermediate factor setting. We observe that both the bounds and the exact simulation results indicate that random subset regression is best suited in this case.

5 Empirical application

This section evaluates the forecast performance of the random subspace methods in a macroeconomic application.

5.1 Data

We use the FRED-MD database consisting of 130 monthly macroeconomic and financial series running from January 1960 through December 2014. The data can be grouped in eight different categories: output and income (1), labor market (2), consumption and orders (3), orders and inventories (4), money and credit (5), interest rate and exchange rates (6), prices (7), and stock market (8). 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. The FRED-MD can be seen as an updated version of the Stock and Watson (2005) dataset. For completeness, Section 5.5 also applies the methods to the original Stock and Watson (2005) data.

5.2 Forecasting framework

We generate forecasts for each of the 130 macroeconomic time series using the following equation

$$y_{t+1} = w_t'\beta_w + x_t'\beta_x + \varepsilon_{t+1},$$

where w_t is a $p_w \times 1$ vector with predictors which are always included in the model and not subject to the dimension reduction methods, and x_t a $p_x \times 1$ vector with possible predictors.

We follow Bai and Ng (2008) in considering up to six lags of the dependent variable and evaluating the forecast performance relative to an AR(4) model. The dependent variable y_{t+1} is one of the macroeconomic time series, w_t includes an intercept and the first four lags of the dependent variable y_{t+1}, and x_t consists of the fifth and sixth lag of y_{t+1}, and all 129 remaining variables in the database. In Section 5.4, x_t also includes the second up to the sixth lag of the 129 remaining variables in the database.

We apply dimension reduction to the predictors in x_t using four different methods: random projection regression (RP), random subset regression (RS), principal component regression (PC), and partial least squares (PL). In addition, we compare the performance to lasso (LA) and ridge regression (RI) as described in Section 4.1. Predictive accuracy is measured by the MSFE defined in (30).

We standardize the predictors in each estimation window. In case of RP and RS we average over $N = 1{,}000$ forecasts to obtain one prediction. In some cases, random subset regression encounters substantial multicollinearity between the original predictors. Insofar this leads to estimation issues due to imprecise matrix inversion, these are discarded from the average. The models generate forecasts with subspace dimension k running from 0 to 100 and, as in Elliott et al. (2013), we recursively select the optimal k based on past predictive performance, using a burn-in period of 60 observations. Note that when $k = 0$, no additional predictors

Table 3: FRED-MD: percentage best forecast performance

		percentage loss							
		RP	RS	PC	PL	RI	LA	AR	All
percentage wins	RP		40.77	86.15	80.77	57.69	65.38	85.38	17.69
	RS	56.92		89.23	81.54	66.92	70.77	83.85	40.00
	PC	11.54	8.46		47.69	12.31	29.23	69.23	3.85
	PL	17.69	16.92	50.77		21.54	30.00	63.85	7.69
	RI	42.31	33.08	87.69	78.46		60.77	84.62	4.62
	LA	34.62	29.23	70.77	70.00	39.23		80.77	18.46
	AR	14.62	16.15	30.77	32.31	15.38	19.23		7.69

Note: this table shows the percentage wins in terms of lowest MSFE of the method listed in the rows over the method listed in the columns, and with respect to all other methods (last column). The percentages are calculated over forecasts for all 130 series in FRED-MD. Ties occur if only $k = 0$ is selected by both methods throughout the evaluation period, which is why losses and wins do not necessarily add up to 100.

are included and we estimate an AR(4) model.

We use an expanding window to produce 420 forecasts, from January 1980 to December 2014. Due to the burn-in period, the initial estimation sample runs from January 1960 to December 1975 and contains 180 observations, from which we discard the first six observations to estimate the lags. This is larger than the initial estimation sample in, for instance, Bai and Ng (2008), since the theory requires the number of variables $p_w + p_x = 136$ to be smaller than the sample size T.

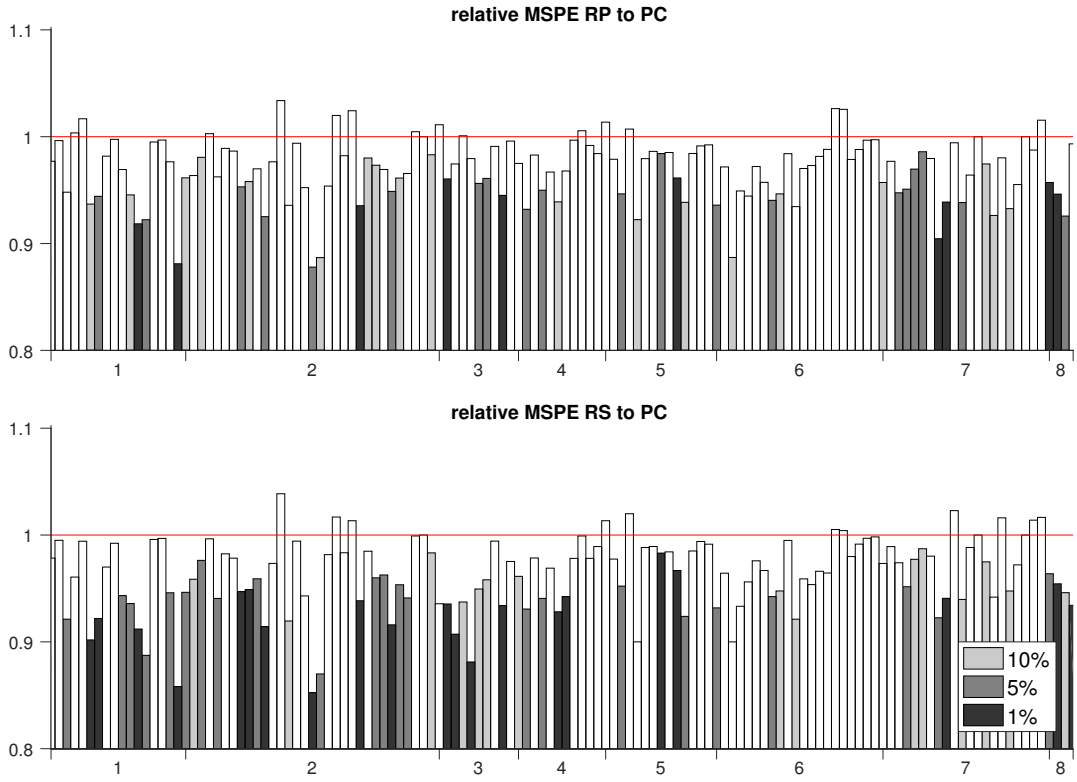
We report aggregate statistics over all 130 series, as well as detailed results for 4 major macroeconomic indicators out of the 130 series; industrial production index (INDP), unemployment rate (UNR), inflation (CPI), and the three-month Treasury Bill rate (3mTB). These series correspond to the FRED mnemonics INDPRO, UNRATE, CPIAUCSL, and TB3MS, respectively.

5.3 Empirical results

5.3.1 Aggregate statistics

We obtain series of forecasts for 130 macroeconomic variables generated by seven different methods. Table 3 shows the percentage wins of a method in terms of lowest MSFE compared to each of the other methods. The last column reports the percentage of the series for which a method outperforms all other methods. We find that random subset regression is more accurate than the other methods

Figure 3: FRED-MD: forecast accuracy relative to principal component regression



Note: this figure shows the MSFE of the forecasts for all series in the FRED-MD dataset produced by random projection regression (upper panel) and random subset regression (lower panel), scaled by the MSFE of principal component regression. Series are grouped in different macroeconomic indicators as described in McCracken and Ng (2016). Values below one prefer the method over principal components. Colors of the bars different from white indicate that the difference from one is significant at the 10% level (grey), 5% level (dark-grey), or 1% level (black), based on a two-sided Diebold-Mariano test.

for 40% of the series. This is a substantial difference with random projections and lasso that win in approximately 18% of the cases. Principal component regression, partial least squares, ridge regression, and the AR(4) model score at most 8%.

If a model is the second most accurate on all series, this cannot be observed in the overall comparison. For this reason, we analyze the relative performance of the methods in a bivariate comparison. Table 3 shows again that random subset regression achieves the best results, outperforming the benchmark models for at least 66% of the series. Interestingly, a close competitor is random projection, which itself is also more accurate than all five benchmarks for a majority of the series. Out of the benchmark models, ridge regression appears closest to random subset regression, which is nevertheless outperformed for more than 66% of the series.

In addition to the ranking of the methods, we are also interested in the relative MSFE of the methods. To get an overview of the forecast performance of the random subspace methods sorted by category, Figure 3 shows relative forecast performance compared with principal component regression, for all series available in the FRED-MD dataset. The MSFE is calculated for the subspace dimension as determined by past predictive performance. The upper panel shows the relative MSFE of random projection regression to principal component regression and the lower panel compares random subset to principal component regression. Values below one, indicate that the random method is preferred over the benchmark. As found in Table 3, the random subspace methods outperform the principal components in most of the cases. For random subset regression this happens in 89% of the cases, which is slightly lower for random projections with 86%. Figure 3 also shows the significance of the differences between the methods. The color of the bar indicates significance as determined by a Diebold and Mariano (1995) test. We see that for series where principal component regression is more accurate, the difference with the random methods is never significant, even at a 10% level. Random projection regression shows the largest improvements in forecast performance in category 7, including price indicators, and random subset regression in category 1 and 2, which contain output, income, and labor market.

Principal component regression is known for its good forecast performance in the presence of instabilities in the data (Rossi, 2013). However, the principal components are outperformed for almost all macroeconomic variables, indicating that random subspace methods are not disproportionately affected by these instabilities.

5.3.2 A case study of four key macroeconomic indicators

We look more closely into the forecast accuracy of the different methods for four key macroeconomic indicators: industrial production index (INDP), unemployment rate (UNR), inflation (CPI), and the three-month Treasury Bill rate (3TB). In Table 4 we show the MSFE relative to the AR(4) model for different values of the subspace dimension or penalty parameter k. The first row of each panel shows the relative MSFE corresponding to the recursively selected optimal value of k, denoted by k_R. The last column of each panel shows the average relative MSFE over all series.

Consistent with our previous findings, random subset regression performs best over all series when the optimal subspace dimension is selected. However, some

Table 4: FRED-MD: forecast accuracy relative to the AR(4)-model

	INDP	UNR	CPI	3TB	Avg.		INDP	UNR	CPI	3TB	Avg.
k	Random projection regression					k	Random subset regression				
k_R	0.843	0.842	0.870	0.892	0.929	k_R	0.820	0.823	0.888	0.906	0.923
1	0.982	0.975	0.992	0.979	0.987	1	0.978	0.968	0.991	0.977	0.984
5	0.930	0.910	0.968	0.930	0.955	5	0.916	0.894	0.968	0.931	0.948
10	0.891	0.871	0.945	0.900	0.937	10	0.870	0.852	0.947	0.910	0.931
15	0.868	0.849	0.928	0.887	0.930	15	0.846	0.832	0.931	0.898	0.925
30	0.841	0.827	0.886	0.892	0.937	30	0.818	0.811	0.898	0.893	0.929
50	0.859	0.846	0.875	0.951	0.983	50	0.822	0.828	0.890	0.918	0.966
100	1.195	1.145	1.080	1.242	1.309	100	1.110	1.097	1.030	1.087	1.245
k	Principal component regression					k	Partial least squares				
k_R	0.890	0.875	0.962	1.006	0.959	k_R	0.898	0.891	0.872	0.945	0.965
1	0.926	0.886	1.002	0.956	0.972	1	0.907	0.856	0.987	0.938	0.973
5	0.880	0.872	0.963	1.008	0.957	5	1.009	0.925	0.928	1.152	1.108
10	0.898	0.858	0.938	0.954	0.968	10	1.173	1.111	0.993	1.253	1.273
15	0.902	0.832	0.933	1.015	0.977	15	1.272	1.209	1.074	1.354	1.378
30	0.943	0.847	0.956	1.127	1.030	30	1.429	1.344	1.168	1.432	1.511
50	0.977	0.898	0.928	1.121	1.107	50	1.465	1.357	1.180	1.423	1.546
100	1.390	1.258	1.191	1.387	1.469	100	1.521	1.369	1.185	1.414	1.560
$\ln k$	Ridge regression					$\ln k$	Lasso				
k_R	0.844	0.842	0.901	0.900	0.930	k_R	0.826	0.848	0.897	0.894	0.935
-6	0.993	0.990	0.997	0.991	0.995	-28	0.864	0.846	0.920	0.894	0.947
-4	0.966	0.952	0.984	0.959	0.975	-27	0.831	0.830	0.880	0.927	0.949
-2	0.880	0.859	0.935	0.896	0.933	-26	0.887	0.898	0.902	1.022	1.022
0	0.847	0.832	0.869	0.930	0.961	-25	1.005	1.014	0.975	1.156	1.148
4	0.946	0.946	0.931	1.080	1.099	-22	1.273	1.229	1.113	1.254	1.358
8	1.216	1.173	1.102	1.261	1.340	-15	1.666	1.520	1.277	1.389	1.644
12	1.463	1.361	1.226	1.334	1.532	-5	1.841	1.651	1.370	1.484	1.788

Note: this table shows the relative MSFE, which equals values below one when the particular method outperforms the benchmark AR(4) model, for different values of subspace dimension k and the recursively selected optimal value of k denoted by k_R. For ridge regression and lasso, the penalty parameter runs over a grid of values k. The relative MSFE is reported for the dependent variables industrial production (INDP), unemployment rate (UNR), inflation (CPI), three month treasury bill rate (3TB), and the average over all series.

differences are observed when analyzing the four individual series. For predicting inflation and the treasury bill rate, random projection yields a lower MSFE compared to random subset regression. Principal component regression is worse than the random methods in predicting all four series and substantially worse on average over all series. The same holds for partial least squares, with the excep-

tion of inflation, where it outperforms random subset, but not random projection regression.

With regard to the lasso and ridge regression benchmarks, the results show that on average, these methods are outperformed by both random subset and random projection regression. Random projection regression has a slight edge on ridge regression, which is in line with our findings in Section 4. For the individual series reported here, the evidence is mixed. Random projection regression outperforms both ridge and lasso on these series, except for industrial production. Random subset regression is only outperformed by ridge or lasso when predicting the treasury bill rate.

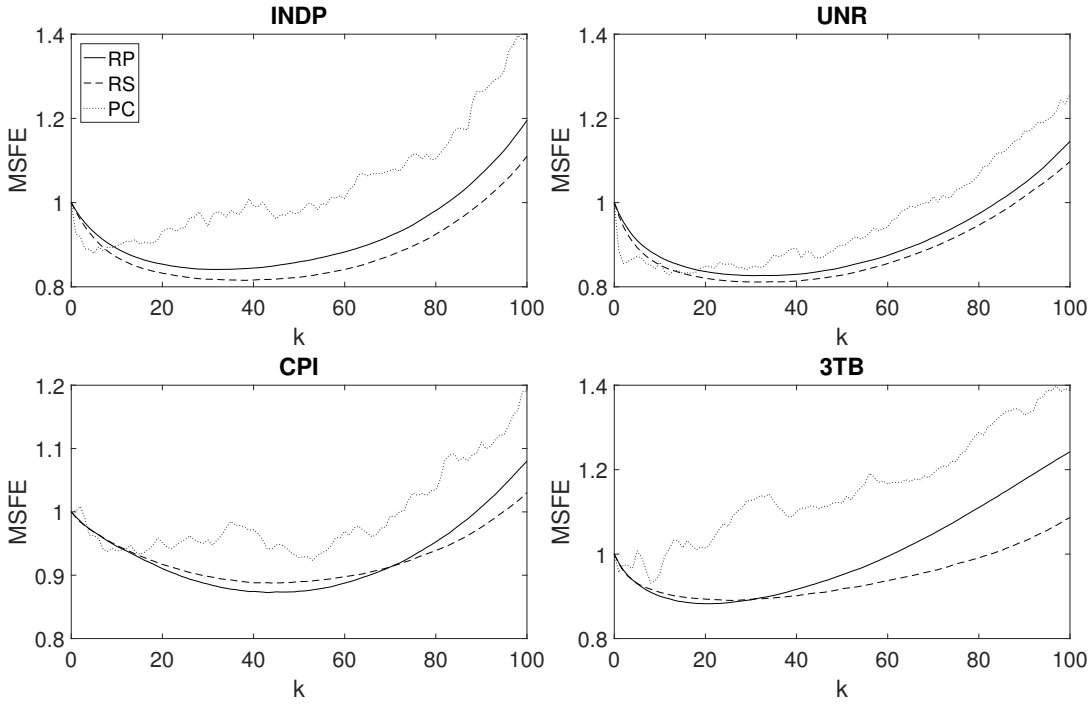
Table 4 also shows the dependence of the MSFE on the value of k if we were to pick the same k throughout the forecasting period. Apart from the treasury bill rate, the random subspace methods outperform the AR(4) benchmark model for almost all subspace dimensions, even for very large values of k. Compared to principal component regression and partial least squares, we again see that the random methods select much larger values of k.

To visualize the dependence on k for the different dimension reduction methods, Figure 4 shows the results for all subspace dimensions ranging from 0 to 100. The first thing to notice is the distinct development of the MSFE of forecasts generated by principal components compared to the random subspace methods. The MSFE evolves smoothly over subspace dimensions for random projections and random subsets, where the MSFE of the principal components changes rather erratically.

Figure 4 shows that the random subspace methods reach their minimum for relatively large values of k. The selected value is substantially larger than the selected dimension when using principal component regression. The difference is especially clear for industrial production in the upper left panel, where principal components suggests to use six factors, while the random methods reach their minimum when using a subspace of dimension larger than 30. Apparently, the information in the additional random factors outweigh the increase in parameter uncertainty and contain more predictive content than higher order principal components. In general, the MSFE of the random subspace methods seems to be lower for most values of k.

In practice, we do not know the optimal subspace dimension. Therefore, real-time forecasts are based on recursively selected values for k based on past performance. Figure 5 shows the selection of the subspace dimension over time. In line with the ex-post optimal subspace dimension, the selected value of k based on

Figure 4: FRED-MD: forecast accuracy for different subspace dimensions

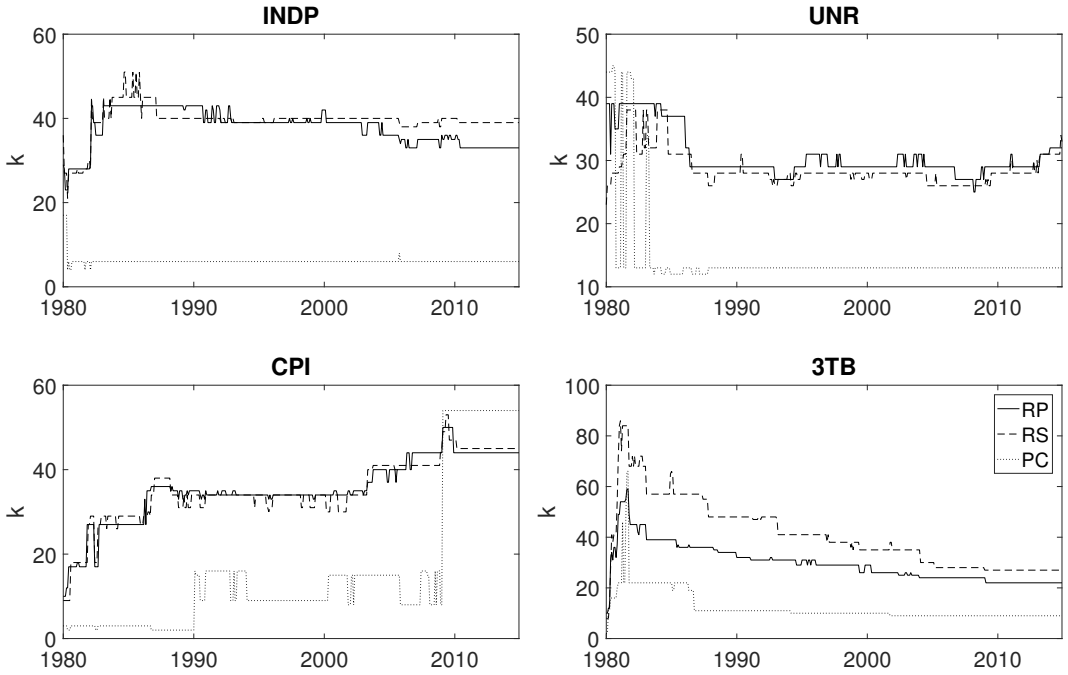


Note: this figure shows the relative MSFE for different values of the subspace dimension k. The different lines correspond to the evaluation criterium for the dimension reduction methods random projection (RP, solid), random subset (RS, dashed), and principal component regression (PC, dotted). The models at $k = 0$ corresponds to an autoregressive model of order four. The four panels correspond to four dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and three month treasury bill rate (3TB).

past predictive performance is smallest for principal component regression. The selected subspace dimension for random subset regression and random projection regression is very similar, but we do find quite some variation over time.

The left upper panel shows that for industrial production, the subspace dimension increases from approximately 30 to 40, where it is quite constant since the mid eighties. The dimension of random projection regression gradually declines back to 33 since the early 2000s. For the unemployment rate in the right upper panel, we observe that more factors seem to be selected since 2008 for both randomized methods, although this has not risen above historically observed values. This is in contrast with the inflation series in the lower left panel. Since the early 2000s both random methods choose gradually larger subspaces, while principal components shows a single sharp increase in 2009. The right lower panel shows that for the treasury bill rate, as one might expect, the subspace dimension decreases over time, reaching its minimum after the onset of the global financial crisis. The

Figure 5: FRED-MD: recursive selection of subspace dimensions

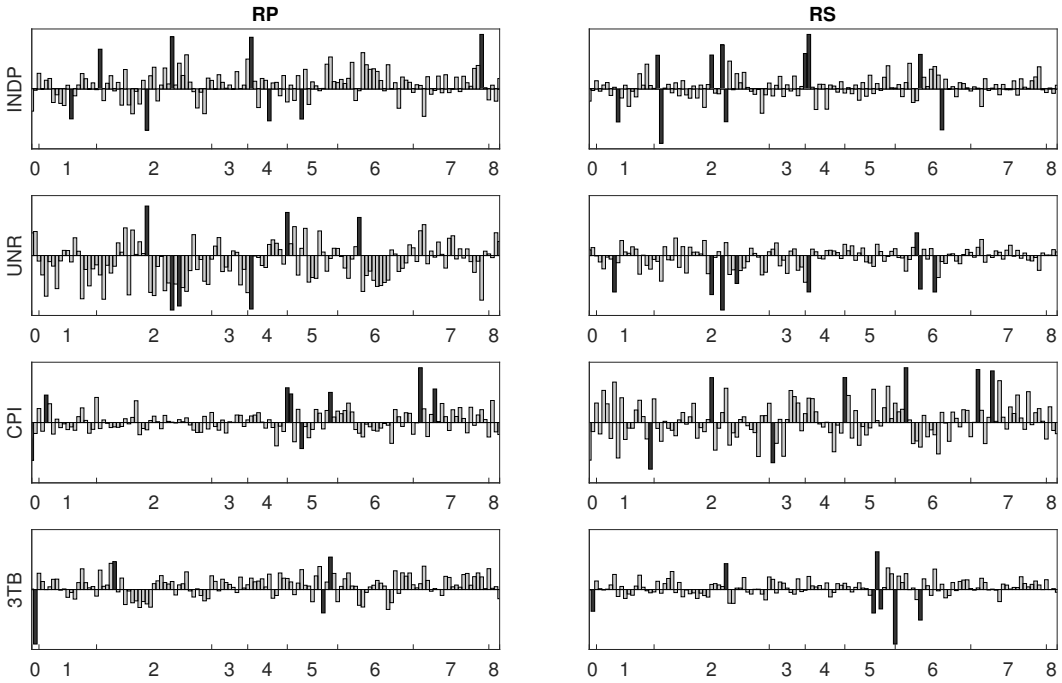


Note: this figure shows the selection of subset dimension k. The different lines correspond to the dimension reduction methods random projection (RP, solid), random subset (RS, dashed), and principal component regression (PC, dotted). At each point in time the subspace dimension is selected based on its past predictive performance up to that point in time. The four panels correspond to four dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and the three month treasury bill rate (3TB).

historical low can be explained by the lack of predictive content in the data since the zero lower bound of the interest rate impedes most variation in the dependent variable.

Figure 6 provides insights in the relation between the predictors and the macroeconomic indicator of interest. We find that random projections and random subset regression estimate different values for the average coefficients. For instance, random projections assigns most weight to lagged values of the three month treasury bill rate to predict this variable, where random subsets mostly explains the one-step-ahead forecast by indicators for money and credit (5) and interest rate and exchange rates (6). The average coefficients also differ over the different series. Where industrial production and unemployment rate are related to variables from all indicator groups, inflation rate seems best explained by indicators for money and credit (5) and prices (7), especially for random projection regression.

Figure 6: FRED-MD: relative weight predictors in random subspace methods



Note: this figure shows the average coefficients of the predictors in x_t in random projection regression (RP) in the left column and random subset regression (RS) in the right column, estimated by $\mathrm{E}_R\left[R\hat{\beta}_{x,R}\right]$ for the optimal subspace dimension in the last estimation sample. Series are grouped in different macroeconomic indicators as described in McCracken and Ng (2016) and the 'zero' group represents the lagged values of the dependent variable. The rows correspond to four dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and the three month treasury bill rate (3TB). Dark coloured bins indicate coefficients which differ two standard deviations from the average over all coefficients.

5.4 Lagged predictors

Although the theoretical results in Section 3 assume $T > p$, we empirically find that the random subspace methods also outperform benchmark methods for $p > T$. Following Bai and Ng (2008) among others, we include lags of the predictors in the forecasting model. We extend x_t with five lags of the variables in the database, such that we have six time periods for each macroeconomic indicator in the database in x_t. The first estimation sample contains 174 observations, while we have 781 regressors. We average over $N = 6{,}000$ forecasts to obtain one prediction in the random subspace methods.

The random subspace methods without including the extra lags of predictors show the best performance. Comparing the numbers in Table 5 to the relative MSFE for the optimal subspace dimension in Table 4, we find that random subset regression shows the overall best performance for industrial production and un-

Table 5: FRED-MD: forecast accuracy with lagged predictors

	RP	RS	PC	PL	RI	LA
INDP	0.894	0.878	0.849	0.914	0.890	0.884
UNR	0.872	0.848	0.872	0.871	0.873	0.868
CPI	0.905	0.895	0.943	0.973	0.904	0.957
3TB	0.958	0.978	1.158	1.047	0.976	0.971

Note: this table shows the relative MSFE generated by the optimal subspace dimension k of different methods using six lags of the predictors in x_t, for the dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and the three month treasury bill rate (3TB).

employment rate, and random projection regression for inflation and the treasury bill rate. Only principal component regression and partial least squares improve in some cases in forecast accuracy by including lagged predictors.

Table 5 shows no conclusive outcome for the relative forecast accuracy of the methods for the different macroeconomic indicators. Principal component regression is most accurate for industrial production, random subset regression for unemployment rate and inflation, and random projection for the treasury bill rate. Using random subspace methods in this high-dimensional setting increases the forecast performance for three out of the four macroeconomic indicators we consider.

5.5 Benchmark dataset

We perform the same analysis as discussed in 5.2 to the Stock and Watson (2005) data, which is used by many researchers to examine macroeconomic forecast accuracy of their methods (Stock and Watson, 2006; Bai and Ng, 2008; Fuentes et al., 2015). The 132 monthly time series run from January 1960 to December 2003. Because we consider six lags of y_{t+1}, the first estimation sample of ten years starts in June 1960. After the burn-in period, we generate forecasts from November 1973 to December 2003. Apart from the starting date, the design mimicks the empirical application in Bai and Ng (2008), where the first estimation sample starts in March 1960. Note that for the first 38 forecasts, the parameters are estimated in a setting where $p > T$.

Just as we found for the FRED-MD data, random subset regression performs best in terms of MSFE. Table 6 shows that random subset regression outperforms the other methods for industrial production and unemployment rate, and ranks

Table 6: Stock and Watson (2005) data: forecast accuracy

	RP	RS	PC	PL	RI	LA
INDP	0.837	0.804	0.852	0.892	0.837	0.813
UNR	0.824	0.809	0.816	0.810	0.824	0.815
CPI	0.986	0.988	0.992	1.027	0.988	1.018
3TB	0.903	0.900	0.936	0.893	0.906	0.935

Note: this table shows the relative MSFE generated by the optimal subspace dimension k of different methods for the dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and the three month treasury bill rate (3TB).

second in terms of lowest MSFE for inflation and treasury bill rate. Random projection regression is more accurate in predicting inflation, and partial least squares in predicting the three month treasury bill rate.

6 Conclusion

In this paper we study two random subspace methods that offer a promising way of dimension reduction to construct accurate forecasts. The first method randomly selects many different subsets of the original variables to construct a forecast. The second method constructs predictors by randomly weighting the original predictors. Although counterintuitive at first, we provide a theoretical justification for these strategies by deriving bounds on their asymptotic mean squared forecast error. These bounds are highly informative on the scenarios where one can expect the two methods to work well and where one is to be preferred over the other.

The theoretical findings are confirmed in a Monte Carlo simulation, where in addition we show that the predictive accuracy increases for nearly all settings under consideration relative to several widely used benchmarks: principal component regression, partial least squares, lasso regularization and ridge regression. In the empirical application, random subset regression generates more accurate forecasts than the benchmarks for no less than 66% of the 130 macroeconomic indicators, and random projection regression outperforms the benchmarks in at least 57% of the series.

References

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

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, 30(4):927–961.

Chiong, K. X. and Shum, M. (2016). Random projection estimation of discrete-choice models with large choice sets. *USC-INET Research Paper*, 2016(16-14).

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Elliott, G., Gargano, A., and Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, 177(2):357–373.

Elliott, G., Gargano, A., and Timmermann, A. (2015). Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control*, 54:86–110.

Frieze, A., Kannan, R., and Vempala, S. (2004). Fast Monte-Carlo algorithms for finding low-rank approximations. *Journal of the Association for Computing Machinery*, 51(6):1025–1041.

Fuentes, J., Poncela, P., and Rodríguez, J. (2015). Sparse partial least squares in time series for macroeconomic forecasting. *Journal of Applied Econometrics*, 30(4):576–595.

Golden, S. (1965). Lower bounds for the Helmholtz function. *Physical Review*, 137(4B):B1127.

Groen, J. J. and Kapetanios, G. (2016). Revisiting useful approaches to data-rich macroeconomic forecasting. *Computational Statistics & Data Analysis*, 100:221–239.

Guhaniyogi, R. and Dunson, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512):1500–1514.

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics*, 146(2):342–350.

Hirano, K. and Wright, J. H. (2017). Forecasting with model uncertainty: Representations and risk reduction. *Econometrica*, 85(2):617–643.

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

Inoue, A. and Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of US consumer price inflation. *Journal of the American Statistical Association*, 103(482):511–522.

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. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, PMLR*, 33:448–456.

Koop, G., Korobilis, D., and Pettenuzzo, D. (2016). Bayesian compressed vector autoregressions. *Available at SSRN 2754241*.

Ma, P., Mahoney, M. W., and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16(1):861–911.

Mahoney, M. W. and Drineas, P. (2009). CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702.

Maillard, O. and Munos, R. (2009). Compressed least-squares regression. In *Advances in Neural Information Processing Systems*, volume 22, pages 1213–1221.

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

Ng, S. (2013). Variable selection in predictive regressions. *Handbook of Economic Forecasting*, 2(Part B):752–789.

Ng, S. (2015). Opportunities and challenges: Lessons from analyzing terabytes of scanner data. Technical report, Columbia University.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Rossi, B. (2013). Advances in forecasting under instability. *Handbook of Economic Forecasting*, 2(Part B):1203–1324.

Schneider, M. J. and Gupta, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, 32(2):243–256.

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.

Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for VAR analysis. Technical report, National Bureau of Economic Research.

Stock, J. H. and Watson, M. W. (2006). Forecasting with many predictors. *Handbook of Economic Forecasting*, 1:515–554.

Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493.

Thanei, G.-A., Heinze, C., and Meinshausen, N. (2017). Random projections for large-scale regression. In *Big and Complex Data Analysis*, pages 51–68. Springer.

Thompson, C. J. (1965). Inequality with applications in statistical mechanics. *Journal of Mathematical Physics*, 6(11):1812–1813.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288.

White, H. (1984). *Asymptotic Theory for Econometricians*. Academic Press.

Wigderson, A. and Xiao, D. (2008). Derandomizing the Ahlswede-Winter matrix-valued Chernoff bound using pessimistic estimators, and applications. *Theory of Computing*, 4(1):53–76.

Wold, H. (1982). Soft modelling: the basic design and some extensions. *Systems under indirect observation, Part II*, pages 36–37.

A Proofs

A.1 Independence between predictor and estimation error

We need the following independence result to derive properties on the forecast accuracy of the random subspace methods.

Lemma 4 *For the regression model in* (1) *under Assumption A1-A7,* z_T *is independent of* $\sqrt{T}(\hat{\beta} - \beta)$ *as* $T \to \infty$.

Proof: We have T observations available for estimation of the parameter vector β. For some $\alpha > 0$, take $T_1 = (1 - T^{-\alpha})T$, such that $T_1/T = O(1)$, $(T - T_1)/T = o(1)$. We require $T - T_1 \to \infty$, such that $\alpha < 1$. The estimation error is given by

$$\sqrt{T}(\hat{\beta} - \beta) = \left(\frac{1}{T} \sum_{t=0}^{T-1} z_t z_t' \right)^{-1} \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} z_t \varepsilon_{t+1}. \tag{31}$$

We split $\frac{1}{\sqrt{T}} \sum_t z_t \varepsilon_{t+1}$ into a part that is independent of z_T and one that is dependent of z_T, but negligible as $T \to \infty$.

$$\frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} z_t \varepsilon_{t+1} = \sqrt{\frac{T_1}{T}} \frac{1}{\sqrt{T_1}} \sum_{t=0}^{T_1} z_t \varepsilon_{t+1} + \sqrt{\frac{T - T_1}{T}} \frac{1}{\sqrt{T - T_1}} \sum_{t=T_1+1}^{T-1} z_t \varepsilon_{t+1}. \tag{32}$$

By Assumption A4, $\text{var}(z_{ti}\varepsilon_{t+1}) = E[(z_{ti}\varepsilon_{t+1})^2] < \Delta < \infty$. By Chebyshev's inequality $P(|z_{it}\varepsilon_{t+1}| \geq T^{\frac{1}{4}}) \leq T^{-\frac{1}{2}}\Delta$. Using Bonferroni's inequality, we then have $P(\max_{t=T_1+1,\ldots,T-1} |z_{it}\varepsilon_{t+1}| \geq T^{\frac{1}{4}}) \leq T^{\frac{1}{2}-\alpha}\Delta$. For this to hold almost surely when $T \to \infty$, we require $\alpha > \frac{1}{2}$. Then,

$$\begin{aligned} \sqrt{\frac{T - T_1}{T}} \frac{1}{\sqrt{T - T_1}} \sum_{t=T_1+1}^{T-1} z_{it}\varepsilon_{t+1} &\leq \sqrt{\frac{T - T_1}{T}} \frac{1}{\sqrt{T - T_1}} \sum_{t=T_1+1}^{T-1} |z_{it}\varepsilon_{t+1}| \\ &\leq T^{-\frac{1}{2}+\frac{1}{4}+1-\alpha}. \end{aligned} \tag{33}$$

Choosing $\alpha > \frac{3}{4}$, we have that

$$\frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} z_{it}\varepsilon_{t+1} = \sqrt{\frac{T_1}{T}} \frac{1}{\sqrt{T_1}} \sum_{t=0}^{T_1} z_{it}\varepsilon_{t+1} + o_p(1). \tag{34}$$

Since under Assumptions A1-A7 a central limit theorem yields $\frac{1}{\sqrt{T}} \sum_{T=1}^{T} z_{it}\varepsilon_{t+1} \sim N(0, \Sigma_z)$, the left-hand size is $O_p(1)$. This implies that the first term on the right-hand side is $O_p(1)$. Since $\{(z_t', \varepsilon_{t+1})\}$ is strong mixing by Assumption A1, and $T - T_1 \to \infty$ for $\alpha < 1$, we have that z_T is independent of the first term of the right-hand side in the limit where $T \to \infty$. Then z_T is also independent of the left-hand side when $T \to \infty$.

The same argument can be used to show that z_T is asymptotically independent of $\frac{1}{T} \sum_{t=0}^{T-1} z_t z_t'$. This shows that as $T \to \infty$, z_T is independent of $\sqrt{T}(\hat{\beta} - \beta)$. ■

A.2 Proof of Theorem 1

By Jensen's inequality, the asymptotic MSFE can be bounded as

$$\begin{aligned} \rho(k) &= \mathrm{E}_\varepsilon \left[\lim_{T\to\infty} T \mathrm{E}_{z_T} \left[\left(z_T'\beta - z_T' \mathrm{E}_R \left[S_R \hat{\beta}_R \right] \right)^2 \right] \right] \\ &\leq \mathrm{E}_R \left[\mathrm{E}_\varepsilon \left[\lim_{T\to\infty} T \mathrm{E}_{z_T} \left[\left(z_T'\beta - z_T' S_R \hat{\beta}_R \right)^2 \right] \right] \right]. \end{aligned} \tag{35}$$

We define the expectation operator $E_{R,\varepsilon} = E_R[E_\varepsilon[.]]$ and rewrite the bound as

$$\begin{aligned} \rho(k) &\leq \mathrm{E}_{R,\varepsilon} \left[\lim_{T\to\infty} T \mathrm{E}_{z_T} \left[\mathrm{trace} \left\{ z_T z_T' (\beta - S_R\hat{\beta}_R)(\beta - S_R\hat{\beta}_R)' \right\} \right] \right] \\ &= \mathrm{E}_{R,\varepsilon} \left[\lim_{T\to\infty} T \mathrm{trace} \left\{ \mathrm{E}_{z_T} \left[z_T z_T' (\beta - A_R\hat{\beta})(\beta - A_R\hat{\beta})' \right] \right\} \right], \end{aligned} \tag{36}$$

where we use the linearity of the trace and define $A_R \equiv S_R(S_R' Z'Z S_R)^{-1} S_R' Z'Z$. We now invoke the asymptotic independence of z_T and $\hat{\beta}$ established in Lemma 4 in Appendix A.1 to evaluate the expectation with respect to z_T. Using that $\mathrm{E}[z_T z_T'] = \Sigma_z$, we then continue as

$$\begin{aligned} \rho(k) &\leq \mathrm{E}_{R,\varepsilon} \left[\lim_{T\to\infty} T(\beta - A_R\hat{\beta})' \Sigma_z (\beta - A_R\hat{\beta}) \right] \\ &= \mathrm{E}_{R,\varepsilon} \left[\lim_{T\to\infty} (\beta - A_R\hat{\beta})' Z'Z (\beta - A_R\hat{\beta}) R \right], \end{aligned} \tag{37}$$

where the second line follows from $\text{plim}_{T\to\infty}\frac{1}{T}Z'Z = \Sigma_z$ in (12), and Slutsky's theorem. Since $A_R\hat{\beta} = S_R\hat{\beta}_R$, the bound can be rewritten to

$$\begin{aligned}\rho(k) &\leq \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty}(\beta - S_R\hat{\beta}_R)'Z'Z(\beta - S_R\hat{\beta}_R)\right] \\ &= \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty}(y - \varepsilon - ZS_R\hat{\beta}_R)'(y - \varepsilon - ZS_R\hat{\beta}_R)\right] = \\ &\mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty}\left(\varepsilon'\varepsilon + (y - ZS_R\hat{\beta}_R)'(y - ZS_R\hat{\beta}_R) - 2\varepsilon'(y - ZS_R\hat{\beta}_R)\right)\right].\end{aligned} \tag{38}$$

To proceed, note that $\hat{\beta}_R = \arg\min_u(y - ZS_Ru)'(y - ZS_Ru)$. Therefore, it holds for an arbitrary $p \times 1$ vector v that

$$\begin{aligned}\rho(k) &\leq \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty}\left(\varepsilon'\varepsilon + (y - ZS_Rv)'(y - ZS_Rv) - 2\varepsilon'(y - ZS_R\hat{\beta}_R)\right)\right] \\ &= \mathrm{E}_{R,\varepsilon}\left[\lim_{T\to\infty}\left((\beta - S_Rv)'Z'Z(\beta - S_Rv) + 2\varepsilon'(ZS_R\hat{\beta}_R - ZS_Rv)\right)\right].\end{aligned} \tag{39}$$

Since we are free to choose v, we choose

$$v = \begin{pmatrix} \beta_w \\ \frac{1}{\sqrt{T}}R'u \end{pmatrix} + (S_R'Z'ZS_R)^{-1}S_R'Z'\varepsilon, \tag{40}$$

with u a fixed $p_x \times 1$ vector. Using (12), $\frac{1}{\sigma^2}\varepsilon'ZS_R(S_R'Z'ZS_R)^{-1}S_R'Z'\varepsilon \overset{(d)}{\to} \chi^2(p_w+k)$. Substituting (40) into (39) and taking the expectation with respect to ε conditional on R gives

$$\rho(k) \leq \sigma^2(p_w + k) + \mathrm{E}_R\left[(\beta_{x,0} - RR'u)'\Sigma_x(\beta_{x,0} - RR'u)\right]. \tag{41}$$

The bound in (41) is valid for any choice of u. After taking the expectation with respect to R, we can therefore minimize the bound with respect to u. Together with the fact that $\mathrm{E}_R[RR'] = \frac{k}{p_x}I_{p_x}$, this yields

$$\rho(k) \leq \sigma^2(p_w + k) + \beta_{x,0}'\Sigma_x\beta_{x,0} - \beta_{x,0}'\Sigma_x\left(\frac{p_x}{k}\mathrm{E}_R\left[RR'\Sigma_xRR'\right]\frac{p_x}{k}\right)^{-1}\Sigma_x\beta_{x,0}. \tag{42}$$

■

A.3 Proof of Lemma 1

Note that RR' is a $p_x \times p_x$ diagonal matrix with k diagonal elements equal to 1, and the remaining elements equal to zero. This implies that

$$[RR'\Sigma_x RR']_{ij} = \begin{cases} [\Sigma_x]_{ij} & \text{if } [RR']_{ii}[RR']_{jj} = 1, \\ 0 & \text{if } [RR']_{ii}[RR']_{jj} = 0. \end{cases} \tag{43}$$

Because the non-zero entries are selected uniformly at random, $P\left([RR']_{ii} = 1\right) = \frac{k}{p_x}$ and $P\left([RR']_{ii}[RR']_{jj} = 1\right) = \frac{k}{p_x}\frac{k-1}{p_x-1}$ for $i \neq j$. This yields $\mathrm{E}_R[RR'] = \frac{k}{p_x} I_{p_x}$ and

$$\mathrm{E}\left[[RR'\Sigma_x RR']_{ii}\right] = \frac{k}{p_x}[\Sigma_x]_{ii}, \qquad \mathrm{E}\left[[RR'\Sigma_x RR']_{ij}\right] = \frac{k}{p_x}\frac{k-1}{p_x-1}[\Sigma_x]_{ij}. \tag{44}$$

We summarize this as

$$\begin{aligned} \mathrm{E}\left[RR'\Sigma_x RR'\right] &= \frac{k}{p_x}\frac{k-1}{p_x-1}\Sigma_x + \frac{k}{p_x}\left(1 - \frac{k-1}{p_x-1}\right) D_{\Sigma_x} \\ &= \frac{k}{p_x}\left(\frac{k-1}{p_x-1}\Sigma_x + \frac{p_x-k}{p_x-1} D_{\Sigma_x}\right), \end{aligned} \tag{45}$$

where $[D_{\Sigma_x}]_{ii} = [\Sigma_x]_{ii}$, and $[D_{\Sigma_x}]_{ij} = 0$ if $i \neq j$. ■

A.4 Proof of Lemma 2

Define $Q = R(R'R)^{-1/2}$ and $P = (R'R)^{1/2}$. Furthermore, define the matrices $W = (w_0, \ldots, w_{T-1})'$ and $X = (x_0, \ldots, x_{T-1})'$. We have

$$\begin{aligned} S_R\hat{\beta}_R &= \begin{pmatrix} I_{p_w} & O \\ O & R \end{pmatrix} \begin{pmatrix} W'W & W'XR \\ R'X'W & R'X'XR \end{pmatrix}^{-1} \begin{pmatrix} W' \\ R'X' \end{pmatrix} y \\ &= \begin{pmatrix} (W'W)^{-1}W' - (W'W)^{-1}W'XV_RX'M_W \\ V_RX'M_W \end{pmatrix} y, \end{aligned} \tag{46}$$

where $M_W = I - W(W'W)^{-1}W'$ and $V_R = R(R'X'M_WXR)^{-1}R'$. Using now that $R = QP$ with P an $k \times k$ invertible matrix, we immediately see that $V_R = Q(Q'X'M_WXQ)^{-1}Q'$. Hence, $S_R\hat{\beta}_R = S_Q\hat{\beta}_Q$, which completes the proof. ■

A.5 Proof of Lemma 3

Consider a matrix R with independent standard normal entries, and a matrix $Q = R(R'R)^{-1/2}$ with the following property.

Lemma 5 *Let R be a $p_x \times k$ matrix with independent standard normal entries. Consider the decomposition $R = QP$, where $Q(R) = R(R'R)^{-1/2}$ and $P(R) = (R'R)^{1/2}$. When we write $U \in \mathcal{O}(p)$ if U is a $p \times p$ orthogonal matrix, we have*

1. *$Q(R) \overset{(d)}{=} H_{p_x} Q(R)$ for $H_{p_x} \in \mathcal{O}(p_x)$.*
2. *$Q(R) \overset{(d)}{=} Q(R) H_k$ for $H_k \in \mathcal{O}(k)$.*

Proof: (Part 1) We have

$$Q(H_{p_x} R) = H_{p_x} R (R' H_{p_x}' H_{p_x} R)^{-1/2} = H_{p_x} Q(R). \tag{47}$$

Also, $H_{p_x} R \overset{(d)}{=} R$. This can be seen from the fact that the matrix variate normal distribution only depends on R through the trace of $R'R$. Then $Q(H_{p_x} R) \overset{(d)}{=} Q(R)$. Combining this with (47), we see that $H_{p_x} Q(R) \overset{(d)}{=} Q(R)$. (Part 2) Decompose $R'R = U \Lambda U'$, where $U \in \mathcal{O}(k)$. Note that $(H_k' U \Lambda U' H_k)^{1/2} = H_k' U \Lambda^{1/2} U' H_k$, and $(H_k' U \Lambda U' H_k)^{-1/2} = H_k' U \Lambda^{-1/2} U' H_k$. Now we have

$$Q(RH_k) = RH_k (H_k' R' R H_k)^{-1/2} = RH_k H_k' (R'R)^{-1/2} H_k = Q(R) H_k. \tag{48}$$

Also $RH_k \overset{(d)}{=} R$, by the same arguments as before. Then $Q(RH_k) \overset{(d)}{=} Q(R) \overset{(d)}{=} Q(R) H_k$. ■

We use Lemma 5 and the eigenvalue decomposition of $\Sigma_x = H \Lambda H'$, where $H \in \mathcal{O}(p_x)$, to rewrite

$$\begin{aligned} \mathrm{E}_Q[QQ' \Sigma_x QQ'] &= \mathrm{E}_Q[QQ' H \Lambda H' QQ'] \\ &= \mathrm{E}_Q[HH' QQ' H \Lambda H' QQ' HH'] = H \mathrm{E}_Q[QQ' \Lambda QQ'] H'. \end{aligned} \tag{49}$$

The elements of the matrix $M = QQ' \Lambda QQ'$, $m_{ii'}$, are a function of the eigenvalues of Σ_x, λ_i, and the elements of Q, q_{ij}, for $i, i' = 1, \ldots, p_x$ and $j = 1, \ldots, k$:

$$\begin{aligned} m_{ii} &= \lambda_i \left(\sum_{j=1}^{k} q_{ij}^4 + \sum_{j \neq j'} q_{ij}^2 q_{ij'}^2 \right) + \sum_{l \neq i} \lambda_l \left(\sum_{j=1}^{k} q_{ij}^2 q_{lj}^2 + \sum_{j \neq j'} q_{ij} q_{lj} q_{ij'} q_{lj'} \right), \\ m_{ii'} &= \lambda_i \left(\sum_{j=1}^{k} q_{ij}^3 q_{i'j} + \sum_{j \neq j'} q_{ij}^2 q_{ij'} q_{i'j'} \right) + \lambda_{i'} \left(\sum_{j=1}^{k} q_{i'j}^3 q_{ij} + \sum_{j \neq j'} q_{i'j}^2 q_{i'j'} q_{ij'} \right) \\ &\quad + \sum_{l \neq \{i,i'\}} \lambda_l \left(\sum_{j=1}^{k} q_{ij} q_{i'j} q_{lj}^2 + \sum_{j \neq j'} q_{ij} q_{i'j'} q_{lj} q_{lj'} \right). \end{aligned} \tag{50}$$

From (50) it follows that we need the (mixed) moments of q_{ij} up to fourth order

to evaluate $\mathrm{E}_Q[QQ'\Lambda QQ']$. These are provided in the following lemma.

Lemma 6 *Suppose we have a $p_x \times k$ matrix Q for which $Q'Q = I_k$ and the i,j-th entry of Q is denoted by q_{ij}, where $i = 1,\ldots,p_x$, $j = 1,\ldots,k$, and $i \neq i', j \neq j'$. For any fixed $p_x \times p_x$ orthogonal matrix H_{p_x} and $k \times k$ orthogonal matrix H_k the matrix Q satisfies the invariance property $H_{p_x} Q H_k \overset{(d)}{=} Q$. Then the non-zero (mixed) moments up to fourth-order are*

$$\begin{aligned} E\left[q_{ij}^2\right] &= \frac{1}{p_x}, \\ E\left[q_{ij}^4\right] &= \frac{3}{p_x(p_x+2)}, \\ E\left[q_{ij}^2 q_{ij'}^2\right] = E\left[q_{ij}^2 q_{i'j}^2\right] &= \frac{1}{p_x(p_x+2)}, \\ E\left[q_{ij}^2 q_{i'j'}^2\right] &= \frac{p_x+1}{p_x(p_x-1)(p_x+2)}, \\ E[q_{ij}q_{ij'}q_{i'j}q_{i'j'}] &= -\frac{1}{p_x(p_x-1)(p_x+2)}. \end{aligned} \tag{51}$$

Note that none of the non-zero (mixed) moments appear in the expression for $m_{ii'}$, such that $\mathrm{E}[m_{ii'}] = 0$.

Proof: We consider the orthogonal matrix H with fixed indices r and $r' \neq r$, and define the elements of H as

$$h_{ij} = \begin{cases} 1 & \text{if } i = j, i \neq r, i \neq r', \\ \sin(\theta) & \text{if } i = j = r, \text{ or } i = j = r', \\ \cos(\theta) & \text{if } i = r, j = r', \\ -\cos\theta & \text{if } i = r', j = r, \\ 0 & \text{otherwise,} \end{cases} \tag{52}$$

where for H_{p_x}, $i,j = 1,\ldots,p_x$ and for H_k, $i,j = 1,\ldots,k$. H_{p_x} sets $\theta = \theta_1$ and H_k sets $\theta = \theta_2$. Throughout this proof, we use the notation that for any index $i' \neq i$. From the invariance property $H_{p_x} Q H_k \overset{(d)}{=} Q$ follows that the elements of Q satisfy

$$\begin{aligned} q_{ij} \overset{(d)}{=} \sin(\theta_1)\sin(\theta_2)q_{ij} + \cos(\theta_1)\sin(\theta_2)q_{i'j} \\ - \sin(\theta_1)\cos(\theta_2)q_{ij'} - \cos(\theta_1)\cos(\theta_2)q_{i'j'}. \end{aligned} \tag{53}$$

First moment Choosing $\theta_1 = \theta_2 = \pi$, we get $q_{ij} \overset{(d)}{=} q_{i'j'}$. Similary, choosing $\theta_1 = 0$ and $\theta_2 = \frac{\pi}{2}$, we get $q_{ij} \overset{(d)}{=} q_{i'j}$. Proceeding in this manner, we conclude that the elements q_{ij} are identically distributed. Furthermore, choosing $\theta_1 = \theta_2 = 0$,

we see that $q_{ij} \overset{(d)}{=} -q_{i'j'}$. Since $\mathrm{E}[q_{ij}] = \mathrm{E}[q_{i'j'}] = -\mathrm{E}[q_{ij}]$, we have $\mathrm{E}[q_{ij}] = 0$.

Second moment We have $Q'Q = I_k$, which implies that $\sum_{i=1}^{p_x} q_{ij}^2 = 1$ for every j. Taking the expectations on both sides and noting that the elements of Q are identically distributed, we have $\mathrm{E}[q_{ij}^2] = \frac{1}{p_x}$. We now proceed to the mixed moments. Take $\theta_2 = \pi/2$ and $\theta_1 = \theta$ in (53), such that $q_{ij} \overset{(d)}{=} \sin(\theta)q_{ij} + \cos(\theta)q_{i'j}$. Then $q_{ij}^2 \overset{(d)}{=} \sin^2(\theta)q_{ij}^2 + \cos^2(\theta)q_{i'j}^2 + 2\sin(\theta)\cos(\theta)q_{ij}q_{i'j}$. Since $\mathrm{E}[q_{ij}^2] = \mathrm{E}[q_{i'j}^2]$, $\mathrm{E}[q_{ij}q_{i'j}] = 0$. Similary, taking $\theta_1 = \pi/2$ and $\theta_2 = \theta$ yields $\mathrm{E}[q_{ij}q_{ij'}] = 0$. Considering then the case for general θ_1 and θ_2 and using the previously derived results, we find $\mathrm{E}[q_{ij}q_{i'j'}] = 0$. Summarizing,

$$\mathrm{E}[q_{ij}^2] = \frac{1}{p_x}, \quad \mathrm{E}[q_{ij}q_{i'j}] = 0, \quad \mathrm{E}[q_{ij}q_{ij'}] = 0, \quad \mathrm{E}[q_{ij}q_{i'j'}] = 0. \tag{54}$$

Fourth moment Setting $\theta_2 = \pi/2$ and $\theta_1 = \theta$ in (53) yields

$$\begin{aligned} q_{ij}^4 \overset{(d)}{=} \sin^4(\theta)q_{ij}^4 + \cos^4(\theta)q_{i'j}^4 + 6\sin^2(\theta)\cos^2(\theta)q_{ij}^2q_{i'j}^2 \\ + 4\sin^3(\theta)\cos(\theta)q_{ij}^3q_{i'j} + 4\sin(\theta)\cos^3(\theta)q_{ij}q_{i'j}^3. \end{aligned} \tag{55}$$

Since all the elements of Q are identically distributed, $\mathrm{E}[q_{ij}^4] = \mathrm{E}[q_{i'j}^4]$, and we have

$$\begin{aligned} \mathrm{E}[q_{ij}^4] &= [\sin^4(\theta) + \cos^4(\theta)]\mathrm{E}[q_{ij}^4] + 6\sin^2(\theta)\cos^2(\theta)\mathrm{E}[q_{ij}^2q_{i'j}^2] \\ &\quad + 4\sin^3\theta\cos(\theta)\mathrm{E}[q_{ij}^3q_{i'j}] + 4\sin(\theta)\cos^3\theta\mathrm{E}[q_{ij}q_{i'j}^3] \\ &= \mathrm{E}[q_{ij}^4] + 2\sin^2(\theta)\cos^2(\theta)\left(3\mathrm{E}[q_{ij}^2q_{i'j}^2] - \mathrm{E}[q_{ij}^4]\right) \\ &\quad + 4\sin^3(\theta)\cos(\theta)\mathrm{E}[q_{ij}^3q_{i'j}] + 4\sin(\theta)\cos^3(\theta)\mathrm{E}[q_{ij}q_{i'j}^3] = \\ &\mathrm{E}[q_{ij}^4] + 2\sin^2(\theta)\cos^2(\theta)\left(3\mathrm{E}[q_{ij}^2q_{i'j}^2] - \mathrm{E}[q_{ij}^4]\right) + 4\sin(\theta)\cos(\theta)\mathrm{E}[q_{ij}^3q_{i'j}], \end{aligned} \tag{56}$$

where we use that $\mathrm{E}[q_{ij}^3q_{i'j}] = \mathrm{E}[q_{ij}q_{i'j}^3]$. For the equality in (56) to hold, we require

$$\mathrm{E}[q_{ij}^4] = 3\mathrm{E}[q_{ij}^2q_{i'j}^2], \qquad \mathrm{E}[q_{ij}^3q_{i'j}] = 0. \tag{57}$$

We use that $Q'Q = I_k$. For any j,

$$1 = \sum_{i=1}^{p_x} q_{ij}^2 = \left(\sum_{i=1}^{p_x} q_{ij}^2\right)^2 = \sum_{i=1}^{p_x} q_{ij}^4 + \sum_{i\neq i'} q_{ij}^2q_{i'j}^2. \tag{58}$$

Taking the expectation and using (57), we have that $1 = p_x\mathrm{E}[q_{ij}^4] + \frac{p_x(p_x-1)}{3}\mathrm{E}[q_{ij}^4]$, which yields $\mathrm{E}[q_{ij}^4] = \frac{3}{p_x(p_x+2)}$, and $\mathrm{E}[q_{ij}^2q_{i'j}^2] = \frac{1}{p_x(p_x+2)}$. For $\theta_1 = \pi/2$ and $\theta_2 = \theta$,

analogous calculations yield

$$\mathrm{E}[q_{ij}^2 q_{ij'}^2] = \frac{1}{p_x(p_x+2)}, \quad \mathrm{E}[q_{ij}^3 q_{ij'}] = 0. \tag{59}$$

To obtain the remaining fourth order moments, we consider general θ_1 and θ_2 in (53). Using previously derived expressions, we arrive after tedious calculations at

$$\begin{aligned}\mathrm{E}[q_{ij}^4] = \mathrm{E}[q_{ij}^4] - a(\theta_1,\theta_2)\mathrm{E}[q_{ij}^3 q_{i'j'}] + b(\theta_1,\theta_2)\mathrm{E}[3q_{ij}^2 q_{i'j'}^2 + 6q_{ij}q_{ij'}q_{i'j}q_{i'j'} - q_{ij}^4] \\ + c(\theta_1,\theta_2)\left\{\mathrm{E}[q_{ij}^2 q_{ij'}q_{i'j}] + 2\mathrm{E}[q_{ij}^2 q_{i'j'}q_{i'j}]d(\theta_1) - 2\mathrm{E}[q_{ij}^2 q_{ij'}q_{i'j'}]d(\theta_2)\right\},\end{aligned} \tag{60}$$

where

$$\begin{aligned} a(\theta_1,\theta_2) &= 4\cos(\theta_1)\cos(\theta_2)\sin(\theta_1)\sin(\theta_2)(2\cos(\theta_1)^2\cos(\theta_2)^2 - 1), \\ b(\theta_1,\theta_2) &= 4\cos(\theta_1)^2\cos(\theta_2)^2\sin(\theta_1)^2\sin(\theta_2)^2, \\ c(\theta_1,\theta_2) &= -12\cos(\theta_1)\cos(\theta_2)\sin(\theta_1)\sin(\theta_2), \\ d(\theta) &= \sin(\theta)\cos(\theta). \end{aligned} \tag{61}$$

Again, since the expectations should be independent of θ_1 and θ_2, this implies that $\mathrm{E}[q_{ij}^3 q_{i'j'}] = \mathrm{E}[q_{ij}^2 q_{ij'}q_{i'j}] = \mathrm{E}[q_{ij}^2 q_{i'j}q_{i'j'}] = \mathrm{E}[q_{ij}^2 q_{ij'}q_{i'j'}] = 0$, and that

$$\mathrm{E}[3q_{ij}^2 q_{i'j'}^2 + 6q_{ij}q_{ij'}q_{i'j}q_{i'j'} - q_{ij}^4] = 0. \tag{62}$$

Since the off-diagonal elements of $Q'Q$ are equal to zero, we have for any $j' \neq j$,

$$0 = \sum_{i=1}^{p_x} q_{ij}q_{ij'} = \left(\sum_{i=1}^{p_x} q_{ij}q_{ij'}\right)^2 = \sum_{i=1}^{p_x} q_{ij}^2 q_{ij'}^2 + \sum_{i\neq i'} q_{ij}q_{ij'}q_{i'j}q_{i'j'}. \tag{63}$$

Taking the expectation and using (59), we get $\frac{1}{p_x+2} = -\sum_{i\neq i'}\mathrm{E}[q_{ij}q_{ij'}q_{i'j}q_{i'j'}]$. Since the expectation should not depend on our choice of i, j, i', j' as long as $i \neq i'$ and $j \neq j'$, we have that $\mathrm{E}[q_{ij}q_{ij'}q_{i'j}q_{i'j'}] = -\frac{1}{p_x(p_x-1)(p_x+2)}$. Then from (62) we obtain $\mathrm{E}[q_{ij}^2 q_{i'j'}^2] = \frac{p_x+1}{p_x(p_x-1)(p_x+2)}$. There is one final identity that we need. We found that $\mathrm{E}\left[q_{ij}^2\right] = \frac{1}{p_x}$ from which follows that $\mathrm{E}_Q[QQ'] = \frac{k}{p_x}I_{p_x}$. Then also $\mathrm{E}_Q[QQ'QQ'] = \frac{k}{p_x}I_{p_x}$. For the off-diagonal elements

$$\begin{aligned}[QQ'QQ']_{mm'} = \sum_{i=1}^{k} q_{mi}^3 q_{m'i} + \sum_{i\neq i'} q_{mi}^2 q_{mi'}q_{m'i'} + \sum_{i=1}^{k} q_{m'i}^3 q_{mi} \\ + \sum_{i\neq i'} q_{m'i}^2 q_{m'i'}q_{mi'} + \sum_{l\neq\{m,m'\}}^{p_x}\left(\sum_{i=1}^{k} q_{mi}q_{m'i}q_{li}^2 + \sum_{i\neq i'} q_{mi}q_{li}q_{m'i'}q_{li'}\right).\end{aligned} \tag{64}$$

We know that $E_Q[QQ'QQ']_{mm'} = 0$, and the only term on the right-hand side for which we have no expression is the final one. This implies that $\mathrm{E}[q_{mi}q_{m'i'}q_{li}q_{li'}] = 0$, which completes the calculation of the moments of q_{ij} up to fourth order. ■

Since $Q'Q = I_k$ for $Q = R(R'R)^{-1/2}$, and Lemma 5 shows that this choice for Q satisfies the invariance property, we can apply Lemma 6 to Q. Lemma 6 states that $\mathrm{E}\left[q_{ij}^2\right] = \frac{1}{p_x}$ from which follows that $\mathrm{E}_Q[QQ'] = \frac{k}{p_x} I_{p_x}$.

Substituting the moments in Lemma 6 in the expectation of (50), we have

$$\begin{aligned} m_{ii} &= \frac{k}{p_x}\left(\frac{2+k}{p_x+2}\lambda_i + \frac{p_x-k}{(p_x+2)(p_x-1)}\sum_{l\neq i}\lambda_l\right) \\ &= \frac{k}{p_x}\left(\frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\lambda_i + \frac{p_x-k}{(p_x+2)(p_x-1)}\sum_{l=1}^{p_x}\lambda_l\right). \end{aligned} \tag{65}$$

Substituting this expression in (49), we arrive at

$$E_Q[QQ'\Sigma_x QQ'] = \frac{k}{p_x}\left(\frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\Sigma_x + \frac{(p_x-k)p_x}{(p_x+2)(p_x-1)}\frac{\mathrm{tr}(\Sigma_x)}{p_x} I_{p_x}\right). \tag{66}$$

■

A.6 Uniform improvement MSFE bound RP

Define R to be a $p_x \times k$ matrix with independent normal entries. We set the variance equal to $1/p_x$ to ensure that $\mathrm{E}[RR'] = \frac{k}{p_x} I_p$. Take $Q = R(R'R)^{-1/2}$ a random orthogonal matrix. To show that the use of Q in Theorem 1 yields a uniform improvement over using R, we need to show that $\Delta = \mathrm{E}[RR'\Sigma_x RR'] - \mathrm{E}[QQ'\Sigma_x QQ'] \succ 0$. From Kabán (2014), Lemma 2, we have that

$$\mathrm{E}[RR'\Sigma_x RR'] = \frac{k}{p_x}\left[\frac{k+1}{p_x}\Sigma_x + \frac{\mathrm{tr}(\Sigma_x)}{p_x} I_{p_x}\right]. \tag{67}$$

Then

$$\begin{aligned} \Delta &= \frac{k}{p_x}\left[\left(\frac{k+1}{p_x} - \frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\right)\Sigma_x + \left(1 - \frac{(p_x-k)p_x}{(p_x+2)(p_x-1)}\right)\frac{\mathrm{tr}(\Sigma_x)}{p_x} I_{p_x}\right] \\ &= \frac{k}{p_x}\left[\frac{p_x(k+1)-2+2(p_x-k)}{p_x(p_x+2)(p_x-1)}\Sigma_x + \frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}\frac{\mathrm{tr}(\Sigma)}{p_x} I_{p_x}\right]. \end{aligned}$$

For the first term, $p_x(k+1) \geq 2$, with equality only when $p_x = k = 1$. Also, $p_x - k \geq 0$. For the second term, again $p_x(k+1) \geq 2$. We see that when $p_x > 1$, $\Delta = a\Sigma_x + bI_{p_x}$ with $a, b > 0$. Since Σ_x is positive definite, this implies Δ is

positive definite. ■

A.7 Eigenvalue bounds

Lemma 7 *Let R be a $p_x \times k$ random selection or random projection matrix, Σ a $p_x \times p_x$ positive definite matrix and $V_R = \Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}$. Then*

$$\frac{\lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma)}\frac{k}{p_x}\frac{1}{\eta} \leq \lambda_{\min}(E_R[V_R]) \leq \lambda_{\max}(E_R[V_R]) \leq \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{k}{p_x}\eta, \tag{68}$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote respectively the minimum and maximum eigenvalue of A, $\eta = 1$ for R a random projection matrix, and $\eta = 2$ for R a random selection matrix.

We provide separate proofs for random projections and random subsets.

Random projections Since both Σ and $\mathrm{E}_R[R(R'\Sigma R)^{-1}R']$ are positive definite,

$$\begin{aligned}\lambda_{\min}(\Sigma)\lambda_{\min}(\mathrm{E}_R[R(R'\Sigma R)^{-1}R']) \leq& \lambda_{\min}(\mathrm{E}_R[V_R]) \\ \leq& \lambda_{\max}(\Sigma)\lambda_{\max}(\mathrm{E}_R[R(R'\Sigma R)^{-1}R']).\end{aligned} \tag{69}$$

As discussed in Section 3.1.2, we can replace R by $Q = R(R'R)^{-1/2}$ and $\mathrm{E}_R[V_R] = \mathrm{E}_Q[V_Q]$. Furthermore, we use the singular value decomposition of $\Sigma = U\Lambda U'$, with $U \in \mathcal{O}(p_x)$, and apply Lemma 5 in Appendix A.5 which says that $UQ \overset{(d)}{=} Q$. Then

$$\begin{aligned}\lambda_{\max}(\mathrm{E}_Q\left[Q(Q'\Sigma Q)^{-1}Q'\right]) &= \lambda_{\max}(U\mathrm{E}_Q\left[Q(Q'\Lambda Q)^{-1}Q'\right]U') \\ &= \lambda_{\max}(\mathrm{E}_Q\left[Q(Q'\Lambda Q)^{-1}Q'\right]).\end{aligned} \tag{70}$$

Now we apply the following lemma to $\mathrm{E}_Q\left[Q(Q'\Lambda Q)^{-1}Q'\right]$:

Lemma 8 *Suppose we have a $p \times p$ matrix A. If $\Omega A \Omega = A$ for any $p \times p$ diagonal matrix Ω with elements randomly drawn from $\{-1, 1\}$, the matrix A is diagonal.*

Proof: Since $\Omega A\Omega = A$, the elements of A satisfy $a_{ij} = \omega_{ii}\omega_{jj}a_{ij}$. Since this holds for any Ω, there always is an Ω such that $\omega_{ii} = -\omega_{jj}$, in which case $a_{ij} = 0$. ■

Pick Ω as in Lemma 8, then,

$$\begin{aligned}\Omega\mathrm{E}_Q\left[Q(Q'\Lambda Q)^{-1}Q'\right]\Omega &= \mathrm{E}_Q\left[\Omega Q(Q'\Omega\Omega\Lambda\Omega\Omega Q)^{-1}Q'\Omega\right] \\ &= \mathrm{E}_Q\left[\Omega Q(Q'\Omega\Lambda\Omega Q)^{-1}Q'\Omega\right] \\ &= \mathrm{E}_Q\left[Q(Q'\Lambda Q)^{-1}Q'\right],\end{aligned} \tag{71}$$

where we use that Ω is an orthogonal matrix, and hence $\Omega Q \overset{(d)}{=} Q$. This proves the diagonality of $\mathrm{E}_Q[Q(Q'\Lambda Q)^{-1}Q']$. We upper bound the eigenvalues of this matrix as

$$\begin{aligned}
\mathrm{E}_Q[q_i'(Q'\Lambda Q)^{-1}q_i] &= \mathrm{E}_Q[q_i'(Q'Q)^{-1/2}((Q'Q)^{-1/2}Q'\Lambda Q(Q'Q)^{-1/2})^{-1}(Q'Q)^{-1/2}q_i] \\
&\leq \mathrm{E}_Q[\lambda_{\max}([(Q'Q)^{-1/2}Q'\Lambda Q(Q'Q)^{-1/2}]^{-1})q_i'(Q'Q)^{-1}q_i] \\
&= \mathrm{E}_Q[(\lambda_{\min}[(Q'Q)^{-1/2}Q'\Lambda Q(Q'Q)^{-1/2}])^{-1}q_i'(Q'Q)^{-1}q_i] \\
&\leq \frac{1}{\lambda_{\min}(\Lambda)}\mathrm{E}_Q[q_i'(Q'Q)^{-1}q_i] = \frac{1}{\lambda_{\min}(\Sigma)}\frac{k}{p_x},
\end{aligned}$$

where the introduction of $(Q'Q)^{-1/2} = I_{p_x}$ emphasizes that we can use the Poincaré separation lemma to obtain the fourth line. Using (69), this gives the bound

$$\lambda_{\max}(E_Q[\Sigma^{1/2}Q(Q'\Sigma Q)^{-1}Q'\Sigma^{1/2}]) \leq \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{k}{p_x}. \tag{72}$$

The proof for the lower bound on the minimum eigenvalue follows analogously. ■

Random subsets We first establish a lower bound on $\lambda_{\min}(\mathrm{E}_R\,[R(R'\Sigma R)^{-1}R'])$. Define a $p_x \times p_x$ random permutation matrix $P_1 = [R_1, R_2, \ldots, R_m]$, with $m = \frac{p_x}{k}$. Take the $p_x \times (p_x + r)$ matrix $P = [P_1, P_2]$, where P_2 is a $p_x \times r$ random selection matrix such that $\tilde{m} = (p_x + r)/k$ is an integer and $r < p_x$. Furthermore, define a $p_x \times p_x$ random matrix $S = D_{\tilde{m}} \otimes (\iota_k \iota_k')$, where $D_{\tilde{m}}$ is a random $\tilde{m} \times \tilde{m}$ matrix where each diagonal element is equal to 1 with probability $1/\tilde{m}$ and a draw of D has only one nonzero element on its diagonal. The $\otimes$ denotes the Kronecker product and ι_k is a $k \times 1$ vector of ones. Note that $\mathrm{E}[S] = \frac{1}{\tilde{m}}B$, where $B = I_{\tilde{m}} \otimes (\iota_k \iota_k')$ is a $p_x \times p_x$ matrix. Then,

$$R(R'\Sigma R)^{-1}R' \overset{(d)}{=} P[S \circ (B \circ P'\Sigma P)^{-1}]P', \tag{73}$$

where $\circ$ denotes the Hadamard product, and hence

$$\begin{aligned}
\mathrm{E}_R[R(R'\Sigma R)^{-1}R'] =& \mathrm{E}_{P,S}[P[S \circ (B \circ P'\Sigma P)^{-1}]P'] \\
=& \mathrm{E}_P[P\mathrm{E}_S[S \circ (B \circ P'\Sigma P)^{-1}|P]P'].
\end{aligned} \tag{74}$$

For the minimum eigenvalue of $\mathrm{E}_R\left[R(R'\Sigma R)^{-1}R'\right]$ now follows

$$\begin{aligned}\lambda_{\min}(\mathrm{E}_R[R(R'\Sigma R)^{-1}R']) &\geq \mathrm{E}_P[\lambda_{\min}(P\mathrm{E}_S[S\circ(B\circ P'\Sigma P)^{-1}|P]P')] \\ &\geq \mathrm{E}_P[\lambda_{\min}(\mathrm{E}_S[S\circ(B\circ P'\Sigma P)^{-1}|P])] \\ &\geq \frac{1}{2}\frac{k}{p_x}\mathrm{E}[\lambda_{\min}((B\circ P'\Sigma P)^{-1})] \\ &\geq \frac{1}{2}\frac{k}{p_x}\lambda_{\min}(\Sigma^{-1}) = \frac{1}{2}\frac{k}{p_x}\frac{1}{\lambda_{\max}(\Sigma)},\end{aligned} \tag{75}$$

where in the first line we use that the minimal eigenvalue is a concave function. For the second inequality we use that for any matrix PAP', we have $\lambda_{\min}(PAP') = \min_v \frac{v'PAP'v}{v'v}$, with $\frac{v'PAP'v}{v'v} = \frac{\tilde{v}A'\tilde{v}}{\tilde{v}'\tilde{v}-v'P_2P_2'v} \geq \frac{\tilde{v}A'\tilde{v}}{\tilde{v}'\tilde{v}}$. Then $\lambda_{\min}(PAP') \geq \lambda_{\min}(A)$. Next, we use that $E[S] = \frac{k}{p_x+r}B \geq \frac{1}{2}\frac{k}{p_x}B$. Finally, on the fourth line, we use that $B\circ P'\Sigma P$ is block diagonal, so that its eigenvalues are bounded by the eigenvalues of the blocks. The blocks itself are inverses of $k\times k$ principal submatrices of Σ, with their eigenvalues bounded by the eigenvalues of Σ^{-1}.

We derive an upper bound on $\lambda_{\max}(\mathrm{E}_R\left[R(R'\Sigma R)^{-1}R'\right])$ in a similar way. Define a $p_x\times p_x$ random permutation matrix $P_1 = [P, P_2]$. We take P to be a $p_x\times(p_x-r)$ random selection matrix such that $(p_x-r)/k$ is an integer. Note that $r < \frac{1}{2}p_x$. We now repeat the argument above. For any matrix PAP' we have $\lambda_{\max}(PAP') = \max_v \frac{v'PAP'v}{v'v}$, with $\frac{v'PAP'v}{v'v} = \frac{\tilde{v}A'\tilde{v}}{\tilde{v}'\tilde{v}+v'P_2P_2'v} \leq \frac{\tilde{v}A'\tilde{v}}{\tilde{v}'\tilde{v}}$. Then $\lambda_{\max}(PAP') \leq \lambda_{\max}(A)$. Moreover, $\mathrm{E}[S] = \frac{k}{p_x-r}B \leq \frac{2k}{p_x}B$. This results in

$$\lambda_{\max}(\mathrm{E}[R(R'\Sigma R)^{-1}R'] \leq 2\frac{1}{\lambda_{\max}(\Sigma)}\frac{k}{p_x}. \tag{76}$$

Combining the bounds on the eigenvalues of $\mathrm{E}_R\left[R(R'\Sigma R)^{-1}R'\right]$ with (69) completes the proof. ■

A.8 Lower bound on MSFE

We rewrite $\rho(k)$ in(14) using a bias-variance decomposition and Lemma 4 in Appendix A.1,

$$\rho(k) = \mathrm{E}_\varepsilon[Y]'\Sigma_z\mathrm{E}_\varepsilon[Y] + \mathrm{E}_\varepsilon[(Y-\mathrm{E}_\varepsilon[Y])'\Sigma_z(Y-\mathrm{E}_\varepsilon[Y])], \tag{77}$$

where we introduce $\sqrt{T}(\beta - \mathrm{E}_R[S_R\hat{\beta}_R]) \xrightarrow{(d)} Y$ to shorten notation. We separately bound the bias and variance term in (77).

Using (46) from Appendix A.4, we rewrite the bias term to

$$\begin{aligned}\mathrm{E}_\varepsilon[Y]'\Sigma_z\mathrm{E}_\varepsilon[Y] = \lim_{T\to\infty} T^{-1}\beta_0'Z'V'Z'ZVZ\beta_0 = \beta_{w,0}'\Sigma_w\beta_{w,0} + \beta_{x,0}'\Sigma_{xw}\beta_{w,0} \\ +\beta_{w,0}'\Sigma_{wx}\beta_{x,0} + \beta_{x,0}'\Sigma_{xw}\Sigma_w^{-1}\Sigma_{wx}\beta_{x,0} + \beta_{x,0}'\Sigma V_R\Sigma V_R\Sigma\beta_{x,0},\end{aligned} \tag{78}$$

where $\Sigma = \mathrm{plim}_{T\to\infty}\frac{1}{T}X'M_wX$, and

$$\begin{aligned}V &= \begin{pmatrix} (W'W)^{-1}W' - (W'W)^{-1}W'XV_RX'M_w \\ V_RX'M_w \end{pmatrix} \\ V_R &= \mathrm{E}_R[R(R'\Sigma R)^{-1}R'], \quad M_w = I_T - P_w, \quad P_w = W(W'W)^{-1}W'.\end{aligned} \tag{79}$$

The last term in (78) can be lower bounded by $\beta_{x,0}\Sigma'\beta_{x,0}\lambda_{\min}(\Sigma^{1/2}V_R\Sigma^{1/2})^2$, and upper bounded by the same expression with the minimum eigenvalues replaced by maximum eigenvalues.

For the variance, we have

$$\begin{aligned}\mathrm{E}_\varepsilon[(Y-\mathrm{E}_\varepsilon[Y])'\Sigma_z(Y-\mathrm{E}_\varepsilon[Y])] =& \mathrm{E}[\lim_{T\to\infty}\varepsilon'V'Z'ZV\varepsilon] \\ =& \mathrm{E}[\lim_{T\to\infty}\varepsilon'(P_w + T^{-1}M_wXV_R\Sigma V_RX'M_w)\varepsilon] \\ =& \sigma^2p_w + \mathrm{E}[\lim_{T\to\infty}T^{-1}\varepsilon'M_wXV_R\Sigma V_RX'M_w\varepsilon],\end{aligned} \tag{80}$$

where we use that $\varepsilon'P_w\varepsilon \xrightarrow{(d)} \sigma^2\chi^2(p_w)$. Since $T^{-1}\varepsilon'M_wX\Sigma^{-1}X'M_w\varepsilon \xrightarrow{(d)} \sigma^2\chi^2(p_x)$, the last term in (80) can be lower bounded by $\sigma^2p_x\lambda_{\min}(\Sigma^{1/2}V_R\Sigma^{1/2})^2$.

Using the bounds on $\lambda_{\min}(\Sigma^{1/2}V_R\Sigma^{1/2})$ in Lemma 7 in Appendix A.7 together with the expressions for the bias and variance terms in (78) and (80), we have the following lower bound on the MSFE

$$\begin{aligned}\rho(k) \geq& \beta_{w,0}'\Sigma_w\beta_{w,0} + \beta_{x,0}'\Sigma_{xw}\beta_{w,0} + \beta_{w,0}'\Sigma_{wx}\beta_{x,0} + \beta_{x,0}'\Sigma_{xw}\Sigma_w^{-1}\Sigma_{wx}\beta_{x,0} + \\ & (\beta_{x,0}'\Sigma\beta_{x,0})\frac{\lambda_{\min}(\Sigma)^2}{\lambda_{\max}(\Sigma)^2}\frac{k^2}{p_x^2}\frac{1}{\eta^2} + \sigma^2\left(p_w + \frac{\lambda_{\min}(\Sigma)^2}{\lambda_{\max}(\Sigma)^2}\frac{1}{\eta^2}\frac{k^2}{p_x}\right),\end{aligned} \tag{81}$$

which completes the proof. ■

Although in many settings weaker than the bound in Theorem 1, we also directly obtain an upper bound on the MSFE:

$$\begin{aligned}\rho(k) \leq& \beta_{w,0}'\Sigma_w\beta_{w,0} + \beta_{x,0}'\Sigma_{xw}\beta_{w,0} + \beta_{w,0}'\Sigma_{wx}\beta_{x,0} + \beta_{x,0}'\Sigma_{xw}\Sigma_w^{-1}\Sigma_{wx}\beta_{x,0} + \\ & (\beta_{x,0}'\Sigma\beta_{x,0})\frac{\lambda_{\max}(\Sigma)^2}{\lambda_{\min}(\Sigma)^2}\frac{k^2}{p_x^2}\eta^2 + \sigma^2\left(p_w + \frac{\lambda_{\max}(\Sigma)^2}{\lambda_{\min}(\Sigma)^2}\frac{k^2}{p_x}\eta^2\right).\end{aligned} \tag{82}$$

A.9 Proof of Theorem 2

First, we use Lemma 4 in Appendix A.1 to write $\rho_S(k)$ as

$$\begin{aligned}\rho_S(k) &= \mathrm{E}_\varepsilon\left[\lim_{T\to\infty} T\mathrm{E}_{z_T}\left[\left(z_T'\beta - z_T'\frac{1}{N}\sum_{i=1}^{N} S_{R_i}\hat{\beta}_{R_i}\right)^2\right]\right] \\ &= \mathrm{E}_\varepsilon\left[\lim_{T\to\infty} T\left(\beta - \frac{1}{N}\sum_{i=1}^{N} S_{R_i}\hat{\beta}_{R_i}\right)'\Sigma_z\left(\beta - \frac{1}{N}\sum_{i=1}^{N} S_{R_i}\hat{\beta}_{R_i}\right)\right].\end{aligned} \tag{83}$$

Define the $p\times 1$ vector d such that,

$$\frac{1}{N}\sum_{i=1}^{N} S_{R_i}\hat{\beta}_{R_i} = \mathrm{E}[S_R\hat{\beta}_R] + \frac{1}{\sqrt{T}}\Sigma_z^{-1/2}\tilde{\epsilon}d. \tag{84}$$

Substituting (84) into (83) yields

$$\rho_S(k) = \rho(k) + \tilde{\epsilon}^2\mathrm{E}_\varepsilon[\lim_{T\to\infty} d'd] - 2\tilde{\epsilon}\mathrm{E}_\varepsilon[\lim_{T\to\infty}\sqrt{T}d'\Sigma_z^{1/2}(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])], \tag{85}$$

where $\rho(k) = \mathrm{E}_\varepsilon\left[\lim_{T\to\infty} T(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])'\Sigma_z(\beta - \mathrm{E}_R[S_R\hat{\beta}_R]\right]$ follows again from Lemma 4 in Appendix A.1. We upper bound the last term in (85) as

$$\begin{aligned}&|2\mathrm{E}_\varepsilon[\lim_{T\to\infty}\sqrt{T}d'\Sigma_z^{1/2}(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])]| \\ &\quad\le 2\mathrm{E}_\varepsilon\left[\lim_{T\to\infty}\sqrt{Td'd(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])'\Sigma_z(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])}\right] \\ &\quad\le \mathrm{E}_\varepsilon\left[\lim_{T\to\infty} d'd\right] + \rho(k),\end{aligned} \tag{86}$$

where we use the Cauchy-Schwarz inequality in the first line, and $a^2+b^2 > 2ab$ in the second line. Combining (85) and (86) results in a bound on $\rho_S(k)$;

$$\rho_S(k) \le (1+\tilde{\epsilon})\rho(k) + (\tilde{\epsilon}+\tilde{\epsilon}^2)\mathrm{E}_\varepsilon[\lim_{T\to\infty} d'd] = (1+\tilde{\epsilon})\left(1 + \frac{\tilde{\epsilon}\mathrm{E}_\varepsilon[\lim_{T\to\infty} d'd]}{\rho(k)}\right)\rho(k).$$

For $\rho_S(k) = (1+\epsilon)\rho(k)$ to hold, we need $\tilde{\epsilon}\mathrm{E}_\varepsilon[\lim_{T\to\infty} d'd]$ to be smaller than the lower bound on $\rho(k)$ which we derive in Appendix A.8.

It suffices to show that

$$\mathrm{E}[\lim_{T\to\infty} d'd] \le \sigma^2\left(\frac{\lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma)}\right)^2\frac{k^2}{p_x}. \tag{87}$$

We construct an upper bound on $\mathrm{E}[\lim_{T\to\infty} d'd]$ and show that this bound satisfies

the bound in (87). By definition

$$\begin{aligned} d =& \sqrt{T}\Sigma_z^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N} S_{R_i}\hat{\beta}_{R_i} - \mathrm{E}_R[S_R\hat{\beta}_R]\right) \\ =& \sqrt{T}\Sigma_z^{1/2}\begin{bmatrix} -(W'W)^{-1}W'X\Delta X'M_W \\ \Delta X'M_W \end{bmatrix} y = \sqrt{T}\Sigma_z^{1/2}V_\Delta y, \end{aligned} \tag{88}$$

where $\Delta = \frac{1}{N}\sum_{i=1}^{N} R_i(R_i'\Sigma R_i)^{-1}R_i' - \mathrm{E}_R[R(R'\Sigma R)^{-1}R']$ with $\Sigma = X'M_WX$. Then

$$\begin{aligned} \mathrm{E}_\varepsilon[\lim_{T\to\infty} d'd] &= \mathrm{E}_\varepsilon[\lim_{T\to\infty} y'V_\Delta'Z'ZV_\Delta y] = \mathrm{E}_\varepsilon[\lim_{T\to\infty} y'M_wX\Delta\Sigma\Delta X'M_wy] \\ &\leq \lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2})^2\mathrm{E}_\varepsilon[\lim_{T\to\infty}(Z\beta+\varepsilon)'M_wX\Sigma^{-1}X'M_w(Z\beta+\varepsilon)] \\ &= \lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2})^2(\beta_0'\Sigma_z\beta_0 + \mathrm{E}_\varepsilon[\lim_{T\to\infty}\varepsilon'M_wX\Sigma^{-1}X'M_w\varepsilon]) \\ &\leq \lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2})^2(\beta_0'\beta_0\lambda_{\max}(\Sigma_z) + \sigma^2p_x) \\ &\leq c\lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2})^2\sigma^2p_x, \end{aligned}$$

since $\varepsilon'M_WX(\Sigma)^{-1}X'M_W\varepsilon \xrightarrow{(d)} \sigma^2\chi^2(p_x)$, and $c > 0$ is a constant independent of p_x. To satisfy (87), we require $\lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2}) \leq c\frac{k}{p_x}$. We apply the following lemma:

Lemma 9 (Ahlswede and Winter (2002), Theorem 19) *Let X_i be a $p_x \times p_x$ independent symmetric positive definite matrix with $\lambda_{\max}(X_i) \leq 1$ almost surely and $i = 1, \ldots, N$. Let $S_N = \sum_{i=1}^{N} X_i$ and $\Omega = \sum_{i=1}^{N}\lambda_{\max}(E[X_i])$, then for all $\epsilon \in (0,1)$*

$$P\left(\lambda_{\max}(S_N - E[S_N]) \geq \epsilon\Omega\right) \leq 2p\exp(-\epsilon^2\Omega/4). \tag{89}$$

This lemma is a non-trivial generalization of a Chernoff bound for sums of independent random variables. For an expository proof, see Section 2 of Wigderson and Xiao (2008). The main technical obstacle is that the proof for scalar random variables relies on the fact that scalars are commutative. To circumvent this, the Golden-Thompson inequality (Golden, 1965; Thompson, 1965) is used.

We define $X_i = \Sigma^{1/2}R_i(R_i'\Sigma R_i)^{-1}R_i'\Sigma^{1/2}$. Since X_i is a projection matrix we have $\lambda_{\max}(X_i) = 1$. We apply Lemma 9 and set

$$\Omega = N\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right]), \tag{90}$$

$$e = \epsilon\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right]). \tag{91}$$

Then plugging in (90) and (91) into Lemma 9, we obtain

$$\begin{aligned} &\mathrm{P}\left(\lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2}) \geq \epsilon\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right])\right) \\ &\leq 2p_x \exp\left(-\frac{\epsilon^2}{4}N\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right])\right). \end{aligned} \tag{92}$$

For $\lambda_{\max}(\Sigma^{1/2}\Delta\Sigma^{1/2}) \leq c\frac{k}{p_x}$ to hold, we need $\epsilon\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right]) \leq \frac{k}{p_x}$ which is guaranteed by Lemma 7 in Appendix A.7. Moreover, the right-hand side of (92) needs to be close to zero, which requires for some $\delta \in (0,1)$ that

$$2p_x \exp\left(-\frac{\epsilon^2}{4}N\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right])\right) \leq \delta. \tag{93}$$

This implies that we need to choose the number of samples

$$N \geq \frac{4}{\epsilon^2\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right])}\log\left(\frac{2p_x}{\delta}\right). \tag{94}$$

We lower bound $\lambda_{\max}(\mathrm{E}_R\left[\Sigma^{1/2}R(R'\Sigma R)^{-1}R'\Sigma^{1/2}\right])$ by using the lower bound on the minimum eigenvalue in Lemma 7 in Appendix A.7, for both random projection and random permutation matrices . We substitute the bound into (94). The result is that for both random permutation matrices and random projection matrices, we need

$$N = O\left(\frac{p_x \log p_x}{k}\right), \tag{95}$$

draws. ■

B Monte Carlo experiments

Table 7: Monte Carlo simulation: relative MSFE

		Random projections - k				Random subsets - k			
s	b	1	10	25	50	1	10	25	50
10	0.5	0.978	1.278	3.504	11.684	0.976	1.286	3.543	11.740
	0.1	0.967	0.872	1.389	3.909	0.967	0.874	1.399	3.927
	2.0	0.964	0.732	0.646	1.127	0.963	0.729	0.646	1.130
50	0.5	0.965	0.815	1.133	3.045	0.964	0.815	1.140	3.065
	0.1	0.962	0.712	0.568	0.885	0.962	0.710	0.569	0.890
	2	0.962	0.684	0.415	0.304	0.961	0.681	0.413	0.306
100	0.5	0.963	0.750	0.781	1.694	0.962	0.748	0.783	1.705
	0.1	0.962	0.693	0.463	0.493	0.962	0.690	0.462	0.496
	2.0	0.961	0.675	0.379	0.194	0.961	0.670	0.376	0.194
		Principal components - k				Partial least squares - k			
s	b	1	10	25	50	1	10	25	50
10	0.5	1.253	3.736	8.780	19.402	9.592	40.512	48.769	51.515
	0.1	1.056	1.665	3.073	6.297	3.099	13.265	15.882	16.731
	2.0	0.972	0.950	1.152	1.828	0.961	3.472	4.186	4.425
50	0.5	1.034	1.424	2.422	4.979	2.377	10.253	12.409	13.107
	0.1	0.972	0.900	0.962	1.428	0.805	2.693	3.248	3.415
	2.0	0.966	0.739	0.537	0.457	0.432	0.685	0.856	0.907
100	0.5	0.983	1.095	1.529	2.742	1.372	5.506	6.647	7.002
	0.1	0.968	0.778	0.677	0.775	0.535	1.364	1.683	1.765
	2.0	0.958	0.685	0.440	0.276	0.356	0.329	0.409	0.431
		Ridge regression - $\ln k$				Lasso - $\ln k$			
s	b	-6	-4	-2	0	-28	-27	-26	-25
10	0.5	0.995	0.971	1.108	4.821	1.073	3.684	11.159	25.064
	0.1	0.993	0.953	0.864	1.787	0.959	1.553	3.793	8.342
	2.0	0.992	0.943	0.768	0.705	0.796	0.708	1.187	2.280
50	0.5	0.993	0.949	0.826	1.428	0.961	1.378	3.073	6.522
	0.1	0.992	0.940	0.750	0.594	0.844	0.713	1.004	1.789
	2.0	0.990	0.925	0.686	0.326	0.617	0.393	0.377	0.519
100	0.5	0.993	0.944	0.781	0.915	0.921	1.011	1.827	3.600
	0.1	0.991	0.934	0.721	0.423	0.767	0.547	0.610	0.950
	2.0	0.988	0.907	0.633	0.241	0.506	0.299	0.242	0.280

Note: this table shows the MSFE relative to the prevailing mean, for random projection regression, random subset regression, principal component regression, partial least squares, ridge regression, and lasso under the data generating process (27) based on 10,000 replications, for increasing values of the subspace dimension k. The coefficient size varies over $b = \{0.5, 1.0, 2.0\}$, and $s = \{10, 50, 100\}$ out of $p = 100$ coefficients are non-zero.

Table 8: Monte Carlo simulation: relative MSFE under a factor design

		Random projections - k				Random subsets - k			
s	b	1	10	25	50	1	10	25	50
Top	0.5	0.944	0.722	1.243	3.931	0.991	0.955	1.136	2.591
	0.1	0.937	0.558	0.444	1.029	0.990	0.915	0.844	1.013
	2.0	0.935	0.513	0.233	0.291	0.990	0.902	0.764	0.598
Int.	0.5	1.013	1.841	5.724	19.064	0.998	1.199	2.735	10.897
	0.1	1.003	1.305	2.739	7.565	0.992	1.013	1.507	4.481
	2.0	1.001	1.075	1.390	2.418	0.991	0.934	0.961	1.604
		Principal components - k				Partial least squares - k			
s	b	1	10	25	50	1	10	25	50
Top	0.5	0.996	1.097	2.905	6.486	2.466	13.526	16.322	17.152
	0.1	0.917	0.300	0.749	1.685	0.495	3.461	4.260	4.470
	2.0	0.886	0.078	0.202	0.448	0.139	0.947	1.156	1.206
Int.	0.5	1.501	6.065	14.467	31.146	16.347	65.901	77.865	82.446
	0.1	1.176	2.948	6.438	12.808	7.140	24.905	29.846	31.725
	2.0	1.060	1.639	2.770	4.172	2.969	7.333	8.545	9.048
		Ridge regression - $\ln k$				Lasso - $\ln k$			
s	b	-6	-4	-2	0	-28	-27	-26	-25
Top	0.5	0.989	0.918	0.734	1.675	0.887	1.729	4.143	9.125
	0.1	0.987	0.903	0.614	0.527	0.539	0.577	1.142	2.399
	2.0	0.984	0.880	0.531	0.206	0.194	0.166	0.312	0.661
Int.	0.5	1.001	1.023	1.486	7.887	1.796	7.268	19.791	43.692
	0.1	1.000	1.007	1.178	3.556	1.335	3.400	7.835	16.719
	2.0	1.000	1.003	1.049	1.577	1.114	1.512	2.543	4.954

Note: this table shows the MSFE relative to the prevailing mean, for random projection regression, random subset regression, principal component regression, partial least squares, ridge regression, and lasso in the Monte Carlo simulations when the underlying model has a factor structure. In the experiments referred to with 'High', we associate nonzero coefficients with the 10 factors that explain most of the variation in the predictors. In the remaining experiments referred to with 'Int.' we associate the nonzero coefficients with intermediate factors $\{f_{46}, \ldots, f_{55}\}$. For additional information, see the note following Table 7.