Boot, Tom; Nibbering, Didier

Working Paper

Forecasting using Random Subspace Methods

Tinbergen Institute Discussion Paper, No. 16-073/III

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Boot, Tom; Nibbering, Didier (2016) : Forecasting using Random Subspace Methods, Tinbergen Institute Discussion Paper, No. 16-073/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/149477

TI 2016-073/III
Tinbergen Institute Discussion Paper

Forecasting using Random Subspace Methods

Tom Boot
Didier Nibbering

Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute, the Netherlands.

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Forecasting Using Random Subspace Methods

Tom Boot* Didier Nibbering†

September 1, 2016

Abstract

Random subspace methods are a novel approach to obtain accurate forecasts in high-dimensional regression settings. We provide a theoretical justification of the use of random subspace methods and show their usefulness when forecasting monthly macroeconomic variables. We focus on two approaches. The first is random subset regression, where random subsets of predictors are used to construct a forecast. Second, we discuss random projection regression, where artificial predictors are formed by randomly weighting the original predictors. Using recent results from random matrix theory, we obtain a tight bound on the mean squared forecast error for both randomized methods. We identify settings in which one randomized method results in more precise forecasts than the other and than alternative regularization strategies, such as principal component regression, partial least squares, lasso, and ridge regression. The predictive accuracy on the high-dimensional macroeconomic FRED-MD data set increases substantially when using the randomized methods, with random subset regression outperforming any one of the above mentioned competing methods for at least 66% of the series.

Keywords: dimension reduction, random projections, random subset regression, principal components analysis, forecasting
JEL codes: C32, C38, C53, C55

1 Introduction

Due to the increase in available macroeconomic data, dimension reduction methods have become an indispensable tool for accurate forecasting. Following Stock and Watson (2002), principal component analysis is widely used to construct a small number of factors from a high-dimensional set of

*Erasmus University Rotterdam, Tinbergen Institute, boot@ese.eur.nl
†Erasmus University Rotterdam, Tinbergen Institute, nibbering@ese.eur.nl
We would like to thank Andreas Pick and Richard Paap for helpful discussions. We thank SURFsara for access to the Lisa Compute Cluster.

predictors. For a recent overview of theoretical results and empirical applications, see Stock and Watson (2006). Instead of combining predictors based on principal component loadings, different combination strategies can be followed. If the underlying factor model is relatively weak, estimation of the factors by principal component analysis is inconsistent as shown by Kapetanios and Marcellino (2010) and one can consider partial least squares as argued by Groen and Kapetanios (2016).

Both principal component regression and partial least squares construct factors by combining the original predictors using data-dependent weights. An intriguing alternative is offered by fully randomized combination strategies. Here, the projection matrix to the low-dimensional subspace is independent of the data and sampled at random from a prespecified probability distribution. In this paper, we establish theoretical properties of two randomized methods and study their behavior in Monte Carlo simulations and in an extensive application to forecasting monthly macroeconomic data.

The first method we consider is random subset regression, which uses an arbitrary subset of predictors to estimate the model and construct a forecast. The forecasts from many such low-dimensional submodels are then combined in order to lower the mean squared forecast error (MSFE). Previous research by Elliott et al. (2013) focused on the setting where one estimates all possible submodels of fixed dimension. However, when the number of predictors increases, estimating all possible subsets rapidly becomes infeasible. As a practical solution, Elliott et al. (2013) and Elliott et al. (2015) propose to draw subsets at random and average over the obtained forecasts. We show that there are in fact strong theoretical arguments for this approach, and establish tight bounds on the resulting MSFE. Using a concentration inequality by Ahlswede and Winter (2002), we also show that it is possible to get arbitrarily close to this bound using a finite and relatively small number of random subsets, explaining why Elliott et al. (2013) find a similar performance when not all subsets are used.

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 normal distribution. Interest in this method sparked by the lemma by Johnson and Lindenstrauss (1984), which states that the geometry of the predictor space is largely preserved under a range of random weighting schemes. This lemma 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). Despite the strong relation to the Johnson-Lindenstrauss lemma, Kabán (2014) shows that in a linear regression model, the underlying assumptions of the lemma are overly restrictive to derive bounds on the in-sample MSFE and that improved bounds can be obtained which eliminate

a factor logarithmic in the number of predictors from earlier work by Maillard and Munos (2009). We show that such improved bounds apply to the out-of-sample MSFE as well.

The derived bounds for the two randomized methods can be used to determine in which settings the methods are expected to work well. For random subset regression, the leading bias term depends on the complete eigenvalue structure of the covariance matrix of the data in relation to the non-zero coefficients, while for random projection it depends only on the average of the eigenvalues multiplied by the average coefficient size. This is shown to imply that in settings where the eigenvalues of the population covariance matrix are roughly equal, the difference between both methods will be small. On the other hand, when the model exhibits a factor structure, the methods deviate. If the regression coefficients associated with the most important factors are non-zero, a typical setting for principal component regression, random projection is preferred as the average of the eigenvalues will be small, driving down the MSFE. If on the other hand the relation between the factor structure and the non-zero coefficients is reversed, random subset regression yields more accurate forecasts.

Of practical importance is our finding, both in theory and practice, that the dimension of the subspace should be chosen relatively large. This 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). Instead, in an illustrative example, we find the optimal subspace dimension k^* to be of order $O(\sqrt{ps})$ with p the number of predictors and s the number of non-zero coefficients. In our empirical setting where $p = 130$, even if $s = 10$, the optimal subspace dimension equals $k^* = 36$.

The theoretical findings are confirmed in a Monte Carlo simulation, which also compares 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). We consider a setup where the non-zero coefficients are not related to the eigenvalues of the covariance matrix to study the effect of sparsity and signal strength. In addition, we consider two settings where a small number of non-zero coefficients is either associated with the principal components corresponding to large eigenvalues, or to moderately sized eigenvalues.

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, but factors associated with intermediate eigenvalues drive the dependent variable, random subset regression is the only method that

outperforms the historical mean of the data.

The theoretical and Monte Carlo findings are empirically tested using the FRED-MD dataset introduced by McCracken and Ng (2015). As the derived theoretical bounds suggest, random subset regression and random projection regression provide similarly accurate forecasts with a clear benefit for random subset regression. This accuracy is shown to be substantially less dependent on the dimension of the reduced subspace than it is in case of principal component regression. In a one-by-one comparison, random subset regression outperforms principal component regression in 88% of the series, partial least squares in 70%, Lasso in 82% and Ridge in 67%. Random projection regression likewise outperforms the benchmarks for a majority of the series and is more accurate than principal component regression in 85% of the series, partial least squares in 56%, Lasso in 82% and Ridge in 57%. Random subset regression is more accurate than random projection regression in 65% of the series, indicating that the factor scenario in the Monte Carlo study where non-zero coefficients are associated with intermediate eigenvalues, is empirically more relevant.

The article is structured as follows. Using results from random matrix theory, Section 2 provides tight bounds on the MSFE under random subset regression and random projection regression. A Monte Carlo study is carried out in Section 3, which highlights the performance of the techniques under different model specifications. Section 4 considers an extensive empirical application using monthly macroeconomic data obtained from the FRED-MD database. Section 5 concludes.

2 Theoretical results

In this section, we start by setting up a general dimension reduction framework, that naturally fits both deterministic and random methods. We subsequently introduce two different randomized reduction methods: random subset regression and random projection regression. We derive bounds on the MSFE under general projection matrices, after which we specialize to the case where these matrices are random. The resulting bounds turn out to be highly informative on scenarios where the methods can be expected to work well.

Consider the data generating process (DGP)

$$y_{t+1} = x_t'\beta + \varepsilon_{t+1} \tag{1}$$

for $t = 1, \dots, T$, and where x_t' is a vector of predictors in $\mathbb{R}^p$. We assume that the errors satisfy $\varepsilon_t \sim i.i.d.(0, \sigma^2)$. We regard the predictors x_t as weakly exogenous, which is not overly restrictive as one typically does not average over lagged terms of the dependent variable. The DGP in (1) can

be straightforwardly adjusted to the situation where some predictors always need to be included.

Since the variance of ordinary least squares (OLS) estimates increases with the number of estimated coefficients, forecasts can get inaccurate when large numbers of predictors are available. As a solution, we project the p-dimensional vector of predictors x_t on a k-dimensional subspace using a matrix $R_i \in \mathbb{R}_i^{p \times k}$

$$\tilde{x}_t' = x_t' R_i \tag{2}$$

A frequently used choice for R_i in order to reduce the number of predictors, is to take the matrix of principal component loadings corresponding to the k largest eigenvalues from the sample covariance matrix $\frac{1}{T-1}\sum_{t=1}^{T-1} x_t x_t'$. Instead of using a single deterministic matrix, randomized methods sample a large number of different realizations of R_i from a prespecified probability distribution. As mentioned above, we consider two different methods to generate R_i: random subset regression and random projection regression.

Random subset regression In random subset regression, the matrix R_i is a random permutation matrix that selects a random set of k predictors out of the original p available predictors. For example, if $p = 5$ and $k = 3$, a possible realization of R_i is

$$R_i = \sqrt{\frac{5}{3}} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \tag{3}$$

For a single realization of R_i, the probability that a diagonal element is non-zero equals $k/p = 3/5$. The scaling factor thus ensures that $\mathrm{E}[R_i R_i'] = I$, which is required in the following sections. More formally, define an index $l = 1, \dots k$ with k the dimension of the subspace, and a scalar $c(l)$ such that $1 \leq c(l) \leq p$. Denote by $e_{c(l)}$ the p-dimensional unit vector with the $c(l)$-th entry equal to one, then random subset regression is based on random projection matrices of the form

$$R_i = \sqrt{\frac{p}{k}} \left[e^i_{c(1)}, \dots, e^i_{c(k)} \right] \quad e^i_{c(m)} \neq e^i_{c(n)} \text{ if } m \neq n \tag{4}$$

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 chooses the weights at random from a normal distribution. In this case, each entry of R_i is independent and identically distributed as

$$[R_i]_{mn} \sim N\left(0, \frac{1}{\sqrt{k}}\right) \quad 1 \leq m \leq p,\ 1 \leq n \leq k \tag{5}$$

where the scaling is again introduced to ensure $\mathrm{E}[R_i R_i'] = I_p$. In fact, a broader class of sampling distributions is allowed. For the results below, it is only required that the entries have zero mean and finite fourth moment.

2.1 Mean squared forecast error bound

We now derive a bound on the mean squared forecast error for general projection matrices R_i, which can be deterministic or random. Following the ideas set out by Kabán (2014), we rewrite the data generating process (1) as

$$y_{t+1} = x_t' R_i R_i' \beta + x_t'(I - R_i R_i')\beta + \varepsilon_{t+1} \tag{6}$$

Instead of (6) we estimate the low-dimensional model

$$y_{t+1} = x_t' R_i \gamma_i + \tilde{\varepsilon}_{t+1} \tag{7}$$

where $\gamma_i \in \mathbb{R}^k$ denotes the optimal parameter vector in the k-dimensional subproblem, that is

$$\gamma_i = \arg\min_u \mathrm{E}\left[\sum_{t=1}^{T-1} \left(y_{t+1} - x_t' R_i u\right)^2 \middle| R_i\right] \tag{8}$$

The least squares estimator of γ_i is denoted by $\hat{\gamma}_i$ and given by

$$\hat{\gamma}_i = \left(\sum_{t=1}^{T-1} R_i' x_t x_t' R_i\right)^{-1} \left(\sum_{t=1}^{T-1} R_i' x_t y_{t+1}\right) \tag{9}$$

Using this estimate, we construct a forecast as

$$\hat{y}_{T+1}^i = x_T' R_i \hat{\gamma}_i \tag{10}$$

If R_i is random, then intuitively, relying on a single realization of the random matrix R_i is suboptimal. By Jensen's inequality, we indeed find that averaging over different realizations of R_i will improve the accuracy

$$\begin{aligned}
&\mathrm{E}\left[(\mathrm{E}_{R_i}\left[\hat{y}_{T+1}^i\right] - x_T'\beta)^2\right] = \\
&\quad = \mathrm{E}\left[\mathrm{E}_{R_i}[\hat{y}_{T+1}^i]^2\right] - 2\mathrm{E}\left[\mathrm{E}_{R_i}[\hat{y}_{T+1}^i]x_T'\beta\right] + \mathrm{E}\left[(x_T'\beta)^2\right] \\
&\quad \leq \mathrm{E}_{R_i}\left[\mathrm{E}\left[\hat{y}_{T+1}^i\right]^2\right] - 2\mathrm{E}_{R_i}\left[\mathrm{E}[\hat{y}_{T+1}^i]x_T'\beta\right] + \mathrm{E}\left[(x_T'\beta)^2\right] \\
&\quad \leq \mathrm{E}_{R_i}\left[\mathrm{E}\left[(\hat{y}_{T+1}^i - x_T'\beta)^2\right]\right]
\end{aligned} \tag{11}$$

where E_{R_i} denotes the expectation with respect to the random variable R_i. For ease of exposition we ignore the variance term ε_{T+1}.

Following (11), we consider the MSFE after averaging over different realizations of the projection matrix R_i. For a single, deterministic projection matrix, this expectation is obviously superfluous. The following bound can

be established on the mean squared forecast error

Theorem 1 Let x_t a vector of predictors for which $\frac{1}{T}\sum_{t=1}^{T-1} x_t x_t' \xrightarrow{p} \Sigma_X$ and $\mathrm{E}[x_t x_t'] = \Sigma_X$ for all t, then

$$\begin{aligned} &E\left[\left(x_T'\beta - x_T'\mathrm{E}_{R_i}\left[R_i\hat{\gamma}_i\right]\right)^2\right] = \\ &\quad \leq \sigma^2\frac{k}{T} + \mathrm{E}_{R_i}\left[\beta'(I - R_iR_i')\Sigma_X(I - R_iR_i')\beta\right] + o_p(T^{-1}) \end{aligned} \tag{12}$$

A proof is presented in Appendix A.

The first term of (12) 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 $\sigma^2\frac{p}{T}$.

The second term reflects the bias that arises by estimating β in a low-dimensional subspace. Loosely speaking, if in (12) the product R_iR_i' concentrates tightly around I under a particular choice of sampling distribution, then the bias term will be small. It is exactly this concentration that underlies the power of randomized methods.

The effect of the choice of k on the bias, can be anticipated from (12). The elements of the matrix R_iR_i' are averages of k products of random entries. Intuitively, as k increases, the concentration of R_iR_i' around its expected value I will tighten. Indeed, we show below that the bias is a decreasing function of k, emphasizing the bias-variance trade-off governed by the choice of the subspace dimension k.

We now specialize to the two different randomized methods, in which case analytic expression are available for the expectation in the bias term.

2.1.1 MSFE bound for random subset regression

For random subset regression, the dimension of the original data space is reduced using a random permutation matrix R_i defined in (4). For this type of matrices we have the following result by Tucci and Wang (2011)

Theorem 2: Let $R_i \in \mathbb{R}^{p\times k}$ be a random permutation matrix, scaled such that $\mathrm{E}[R_iR_i'] = I$. Then

$$\begin{aligned} &\mathrm{E}_{R_i}^{RS}\left[(I - R_iR_i')\Sigma_X(I - R_iR_i')\right] = \\ &\quad = \frac{p}{k}\left(\left[\frac{k-1}{p-1} - \frac{k}{p}\right]\Sigma_X + \frac{p-k}{p-1}D_{\Sigma_X}\right) \end{aligned} \tag{13}$$

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

Substituting this expression into (12), we obtain that for random subset

regression

$$
\begin{aligned}
&E\left[\left(x_T'\beta - x_T'\mathrm{E}_{R_i}^{RS}\left[R_i\hat{\gamma}_i\right]\right)^2\right] = \\
&\quad \leq \frac{\sigma^2 k}{T} + \frac{p-k}{k}\frac{p}{p-1}\left[\beta' D_{\Sigma_X}\beta - \frac{1}{p}\beta'\Sigma_X\beta\right] + o_p(T^{-1})
\end{aligned}
\tag{14}
$$

We observe that as $k \to p$, the bias decreases and we obtain the variance formula for the OLS estimates of β when $k = p$. In many high-dimensional settings, we expect $p \gg k$ and $p, k \gg 1$, such that the leading bias term is $\frac{p}{k}\beta' D_{\Sigma_X}\beta$. We will discuss this term in more depth in an illustrating example below.

2.1.2 MSFE bound for random projection regression

For random projection defined in (5), the following theorem is derived by Kabán (2014)

Theorem 3 For $R_i \in \mathbb{R}^{p\times k}$ and $[R_i]_{mn} = N\left(0, \frac{1}{\sqrt{k}}\right)$ and Σ_X a positive semi-definite matrix

$$
\begin{aligned}
&\mathrm{E}_{R_i}^{RP}\left[(I - R_i R_i')\Sigma_X(I - R_i R_i')\right] = \\
&\quad = \frac{p}{k}\left[\left(\frac{k+1}{p} - \frac{k}{p}\right)\Sigma_X + \frac{1}{p}\mathrm{trace}(\Sigma_X)I\right]
\end{aligned}
\tag{15}
$$

This result holds when the assumption on the entries of the random matrix is weakened, requiring only that they are drawn from a symmetric distribution with zero mean and finite fourth moments.

Substituting (15) into (12), the mean squared forecast error that follows from random projection regression satisfies the following bound

$$
\begin{aligned}
&E\left[\left(x_T'\beta - x_T'\mathrm{E}_{R_i}^{RP}\left[R_i\hat{\gamma}_i\right]\right)^2\right] = \\
&\quad \leq \frac{\sigma^2 k}{T} + \frac{1}{k}\left[\beta'\Sigma_X\beta + \mathrm{trace}(\Sigma_X)\beta'\beta\right] + o_p(T^{-1})
\end{aligned}
\tag{16}
$$

A notable difference with random subset regression is that the bias term remains non-zero even when $p = k$. The reason is that the columns of the projections matrix are not exactly orthogonal, and therefore might span a smaller space than the original predictor matrix. Indeed, when the columns are orthogonalized, the following theorem by Marzetta et al. (2011) guarantees that the bias is identically zero when $k = p$.

Theorem 4 Let R_i a random matrix with i.i.d. normal entries such that $R_i' R_i = \frac{p}{k} I_k$ and Σ_X a positive semi-definite matrix, then

$$
\begin{aligned}
&\mathrm{E}_{R_i}^{ORP}\left[(I - R_i R_i')\Sigma_X(I - R_i R_i')\right] = \\
&\quad = \frac{p}{k}\left[\left(\frac{pk-1}{p^2-1} - \frac{k}{p}\right)\Sigma_X + \frac{p-k}{p^2-1}\mathrm{trace}(\Sigma_X)I\right]
\end{aligned}
\tag{17}
$$

Hence, the MSFE after orthogonalization is bounded by

$$E\left[\left(x_T'\beta - x_T'\mathrm{E}_{R_i}^{ORP}\left[R_i\hat{\gamma}_i\right]\right)^2\right] = \\ \leq \frac{\sigma^2 k}{T} + \frac{p-k}{k}\frac{p^2}{p^2-1}\left[\frac{\mathrm{trace}(\Sigma_X)}{p}\beta'\beta - \frac{1}{p}\beta'\Sigma_X\beta\right] + o_p(T^{-1}) \tag{18}$$

where the second term equals zero when $p = k$. Orthogonalization leads to an improved bound compared to (16), since the difference in MSFE between random projection and its orthogonalized form satisfies

$$E\left[\left(x_T'\beta - x_T'\mathrm{E}_{R_i}^{RP}\left[R_i\hat{\gamma}_i\right]\right)^2\right] - E\left[\left(x_T'\beta - x_T'\mathrm{E}_{R_i}^{ORP}\left[R_i\hat{\gamma}_i\right]\right)^2\right] \geq 0 \tag{19}$$

which is derived in Appendix B. However, orthogonalization is computationally costly and in many examples the dimensions of the problem are such that the gain in predictive accuracy will be negligible.

A second important difference with the results for random subset regression, is that when $p \gg k$ and $p, k \gg 1$, the leading bias term equals $\frac{\mathrm{trace}(\Sigma_X)}{k}\beta'\beta$. For random subset regression the leading term was found to be $\frac{p}{k}\beta' D_{\Sigma_X}\beta$. This points out a conceptual difference between the two methods that is further analyzed in the next section.

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

To gain intuition for the performance of the randomized methods compared with unrestricted estimation by ordinary least squares (OLS), and to show when one of the randomized methods is preferred over the other, we consider a simplified setting. This setting nevertheless brings out the main features we observe in the more sophisticated set-up studied in the Monte Carlo simulations described in Section 3.

Suppose $p \gg k$ and $p, k \gg 1$, then from (14) we have that the leading bias term for random subset regression is $\frac{p}{k}\beta' D_{\Sigma_X}\beta$. For random projection, we have from (16) that the leading bias term equals $\frac{\mathrm{trace}(\Sigma_X)}{k}\beta'\beta$. Suppose that the population covariance matrix is given by

$$\Sigma_X = \begin{pmatrix} 1+\alpha & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{20}$$

For notational convenience, assume that $\frac{\sigma^2}{T} = 1$. In this setting, the MSFE for random subset regression is given by

$$E\left[\left(x_T'\beta - x_T'\mathrm{E}_{R_i}^{RS}\left[R_i\hat{\gamma}_i\right]\right)^2\right] \leq k + \frac{p}{k}(\alpha\beta_1^2 + \beta'\beta) \tag{21}$$

This expression depends explicitly on the size of the coefficient β_1. This in contrast with random projection regression, for which the MSFE is given by

$$E\left[\left(x_T'\beta - x_T'\mathrm{E}_{R_i}^{RP}\left[R_i\hat{\gamma}_i\right]\right)^2\right] \leq k + \frac{p+\alpha}{k}\beta'\beta \tag{22}$$

RS and RP versus OLS The simplest scenario is when $\alpha = 0$ in (20), and $\beta_i = c$ for $i = 1, \ldots, s$, with $s \leq p$, and zero otherwise. We refer to s as the sparsity of the coefficient vector β. Both for RS and RP the bound on the MSFE reduces to

$$E\left[\left(x_T'\beta - x_T'\mathrm{E}_{R_i}\left[R_i\hat{\gamma}_i\right]\right)^2\right] \leq \frac{\sigma^2}{T}\left[k + \frac{ps}{k}c^2\right] \tag{23}$$

When using the optimal value of k derived in Appendix C, $k^* = c\sqrt{ps}$, this reduces to

$$E\left[\left(x_T'\beta - x_T'\mathrm{E}_{R_i}\left[R_i\hat{\gamma}_i\right]\right)^2\right] \leq 2c\sqrt{ps} \tag{24}$$

Note that the optimal size is of order $O(\sqrt{ps})$, which can be much larger than what one might expect based on findings when forecasting using factor models where typically around 5 factors are selected, as for example in Stock and Watson (2012). In the empirical setting of Section 4, we have $p = 130$ such that even at a sparsity level of 10%, the optimal model size is $k^* = 36$.

Under the optimal value of k, the relative performance compared to OLS is given by $2c\sqrt{\frac{s}{p}}$. As one might expect, the increase in accuracy of the randomized methods compared to OLS is larger when the coefficient size and the number of non-zero coefficients are small.

RS versus RP To examine the relative performance of RS and RP, we analyze the difference in MSFE obtained from (21) and (22)

$$\Delta = \frac{p}{k}\alpha\left[\beta_1^2 - \frac{\beta'\beta}{p}\right] \tag{25}$$

If all coefficients are of the same size, then $\beta_1^2 \approx \frac{\beta'\beta}{p}$ and the methods are expected to perform equally well. The same happens if the covariance matrix is well-conditioned, i.e. $\alpha \to 0$ and all eigenvalues of the covariance matrix are of the same size.

For non-zero α, two things can happen. First, consider a typical principal component regression setting where β_1 is large while all other coefficients are close or equal to zero. Here, the MSFE for random projection is only affected by the large coefficient β_1 through the inner product $\frac{\beta'\beta}{p}$. Random subset regression on the other hand suffers, as the MSFE depends explicitly on the product $\alpha\beta_1^2$. This setting therefore favors random projection. The difference between the two methods increases as β_1 and/or α grow larger.

In contrast with the previous setting, it is also possible that the factor associated with the largest eigenvalue of Σ_X is not associated with the dependent variable. This is the case when α is large, while $\beta_1 = 0$. If any signal is present in the remaining factors, random subset regression will outperform random projection.

In addition to the contrast in MSFE, there is also a difference in the optimal subspace dimension. We have

$$\begin{aligned} k^*_{RS} &= \sqrt{p(\alpha\beta_1^2 + \beta'\beta)} \\ k^*_{RP} &= \sqrt{(p+\alpha)\beta'\beta} \end{aligned} \tag{26}$$

In the factor setting where both α and β_1 are large, the optimal dimension for random subset regression can be much larger. If on the other hand β_1 is close to or equal to zero, random projection chooses a larger subspace dimension when $\alpha > 0$.

2.2 Feasibility of the MSFE bounds

The bounds from the previous section are calculated using expectations over the random matrix R_i. In reality we have to settle for a finite number of draws. We therefore need the average over these draws to concentrate around the expectation, i.e. with high probability it should hold that

$$\Delta = \left|\left| \frac{1}{N} \sum_{i=1}^{N} R_i R_i' \Sigma_X R_i R_i' - \mathrm{E}\left[R_i R_i' \Sigma_X R_i R_i' \right] \right|\right| < e \tag{27}$$

where $||\cdot||$ denotes the Euclidean norm and e is some small, positive number. Such a concentration can be proven both for random projections and for random subset regression using the following theorem by Ahlswede and Winter (2002)

Theorem 5 Let X_i, $i = 1, \ldots, N$ be a $p \times p$ independent random positive semi-definite matrix with $||X_i|| \leq 1$ almost surely. Let $S_N = \sum_{i=1}^{N} X_i$ and $\Omega = \sum_{i=1}^{N} ||\mathrm{E}[X_i]||$, then for all $\epsilon \in (0, 1)$

$$\mathrm{P}\left(||S_N - \mathrm{E}[S_N]|| \geq \epsilon\Omega \right) \leq 2p \exp(-\epsilon^2 \Omega/4) \tag{28}$$

Since this holds for all $\epsilon \in (0, 1)$, we can make $\epsilon\Omega$ arbitrarily small, which we use to show that (27) holds with high probability for small e. Using the same approach, it is then straightforward to show that

$$\tilde{\Delta} = \left|\left| \frac{1}{N} \sum_{i=1}^{N} R_i R_i' - \mathrm{E}\left[R_i R_i' \right] \right|\right| < e \tag{29}$$

for some finite number N.

Random subset regression Consider random permutation matrices $R_i \in \mathbb{R}^{p\times k}$ suitably scaled by a factor $\sqrt{\frac{p}{k}}$ to ensure that $\mathrm{E}[R_i R_i'] = I$. Let $Q_i = R_i R_i' \Sigma_X R_i R_i'$, then

$$||Q_i|| \leq ||\Sigma_X|| \cdot ||R_i R_i'||^2 = \left(\frac{p}{k}\right)^2 ||\Sigma_X|| \tag{30}$$

using that for any draw of R_i, the Euclidean norm of the outer product satisfies $||R_i R_i'|| = \frac{p}{k}$. Define now $X_i = Q_i/||Q_i||$. Then

$$\Omega = N \frac{||\mathrm{E}\left[R_i R_i' \Sigma_X R_i R_i'\right]||}{\left(\frac{p}{k}\right)^2 ||\Sigma_X||} \tag{31}$$

where we use that $||\mathrm{E}\left[R_i R_i' \Sigma_X R_i R_i'\right]||$ is independent of i which can be observed from (13). We can simply plug this expression into (28) to obtain

$$\begin{aligned} \mathrm{P}\left(||\Delta|| \geq \epsilon \frac{||\mathrm{E}\left[R_i R_i' \Sigma_X R_i R_i'\right]||}{\left(\frac{p}{k}\right)^2 ||\Sigma_X||}\right) = \\ \leq 2p \exp\left(-\epsilon^2 N \frac{||\mathrm{E}\left[R_i R_i' \Sigma_X R_i R_i'\right]||}{4\left(\frac{p}{k}\right)^2 ||\Sigma_X||}\right) \end{aligned} \tag{32}$$

Now, to satisfy (27) with high probability, we need the right hand side to be close to zero. If we require for some $\delta \in (0,1)$ that

$$2p \exp\left(-\epsilon^2 N \frac{||\mathrm{E}\left[R_i R_i' \Sigma_X R_i R_i'\right]||}{4\left(\frac{p}{k}\right)^2 ||\Sigma_X||}\right) \leq \delta \tag{33}$$

then we should choose the number of samples

$$N \geq \frac{4||\Sigma_X||}{\epsilon^2 ||\mathrm{E}\left[R_i R_i' \Sigma_X R_i R_i'\right]||} \left(\frac{p}{k}\right)^2 \log\left(\frac{2p}{\delta}\right) \tag{34}$$

For the term in the denominator we know by Theorem 2 that

$$||\mathrm{E}\left[R_i R_i' \Sigma_X R_i R_i'\right]|| = O\left(\frac{p}{k}\right) \tag{35}$$

Hence, we need

$$N = O(p \log p) \tag{36}$$

draws of the random matrix to obtain results that are close to the bounds of the previous paragraph. 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 finite number of random draws of the subsets.

Random projection regression For random projection regression, similar bounds to the ones we found for random subset regression have been established when R_i is a random projection matrix. The proof in this case is somewhat more involved as one needs additional concentration inequalities to bound the Euclidean norm $||R_i R_i'||$ with high probability. A complete proof of the following theorem can be found in Kabán et al. (2015)

Theorem 6: Let Σ_X be a positive semi-definite matrix of size $p \times p$ and rank r. Furthermore, let R_i, $i = 1, \dots N$ be independent random projections with $[R_i]_{jk} \sim \frac{1}{\sqrt{k}} N(0,1)$. Define Δ as in (27), then for all $\epsilon \in (0,1)$

$$\begin{aligned} &P\left(\Delta \geq \epsilon \frac{||\mathrm{E}\left[R_i R_i' \Sigma_X R_i R_i'\right]||}{K}\right) \\ &\quad \leq 2p \exp\left(-\epsilon^2 N \frac{||\mathrm{E}\left[R_i R_i' \Sigma_X R_i R_i'\right]||}{4K}\right) + 4N \exp\left(-\frac{N^{1/3}}{2}\right) \end{aligned} \tag{37}$$

where

$$K = ||\Sigma_X|| \left[\left(1 + \sqrt{\frac{p}{k}}\right) + \frac{1}{\sqrt{k}}\right]^2 \left[\left(\sqrt{\frac{r}{k}} + \sqrt{\frac{p}{k}}\right) + \frac{1}{\sqrt{k}}\right]^2 \tag{38}$$

If we neglect the last term of (37), then by the same arguments as above it can be shown that the required order of draws is the same as for random subset regression, i.e. $N = O(p \log p)$. The additional term on the right-hand side of (37) implies that we need a slightly larger number of draws for random projection regression. In practice however, we found no difference in the behavior for a finite number of draws between the two methods.

3 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 mean squared forecast error, 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.

3.1 Monte Carlo set-up

The set-up we employ is similar to the one by Elliott et al. (2015). The data generating process takes the form

$$y_{t+1} = x_t' \beta + \varepsilon_{t+1}, \tag{39}$$

where x_t is a $p \times 1$ vector with predictors, β a $p \times 1$ coefficient vector, and ε_{t+1} an error term with $\varepsilon_{t+1} \sim N(0, \sigma_\varepsilon^2)$.

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 $\Sigma_X = \frac{1}{p} P'P$, where P is a $p \times p$ 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 $\beta = \sqrt{\sigma_\varepsilon^2 / T} \cdot b\iota_s$ for a fixed constant b The vector ι_s contains s nonzero 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 mimick the high-dimensional setting in the empirical application by choosing the number of predictors $p = 100$ and the sample size $T = 200$. Results are based on $M = 10,000$ replications of the data generating process (39).

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 2.1.3. In this case we replace x_t in the DGP (39) by a subset of the factors extracted from the sample covariance matrix $\frac{1}{T} \sum_{t=1}^{T} x_t x_t'$ using principal component analysis. Denote by f_i for $i = 1, \ldots, p$ 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 $\{f_{46}, \ldots, f_{55}\}$, 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 (7) in which we vary the subspace dimension over $k = \{1, \ldots, p\}$. The subspace methods, as well as the benchmark models discussed below, estimate (39) 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 meth-

ods to arrive at a one-step-ahead forecast. This is in line with the findings in Section 2.2 which suggest to use $O(p \log p) = O(100 \cdot \log 100) = O(460)$ draws.

Benchmark models We compare the performance of the random methods with principal component regression, and partial least squares regression introduced by Wold (1982). Both methods approximate the data generating process (39) as

$$y_{t+1} = z_t'\delta^f + \sum_{i=1}^{k} f_{ti}\beta_i^f + \eta_t \tag{40}$$

where $k \in \{1, \dots, p\}$. 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 $t = 1, \dots, T$ using principal component analysis. We then estimate (40) and generate a forecast as $\hat{y}_{T+1} = z_t'\hat{\delta}^f + \sum_{i=1}^{k} f_{Ti}\hat{\beta}_i^f$. Note that for the top factor setting in the second set of experiments, the principal component regression model is thus correctly specified.

Partial least squares uses a two-step procedure to construct the factors, as described by Groen and Kapetanios (2016). We orthogonalize both the standardized predictors x_t and the dependent variable y_{t+1} with respect to z_t for $t = 1, \dots, T-1$. We then calculate the covariance of each predictor x_{it} with y_{t+1} which yields weights $w = \{w_1, \dots, w_p\}$. The first factor is readily constructed as $f_{t1} = x_t'w$. We then orthogonalize x_{it} and y_{t+1} with respect to this factor and repeat the procedure with the corresponding residuals until the required number of factors $f_{t1}, \dots, f_{tk}$ is obtained. To construct a forecast we require f_T for which the above procedure is repeated now taking $t = 1, \dots, T$. Calculating the covariance with y_{T+1} naturally is infeasible, such that the same weights w_i are used as obtained before.

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

$$(\hat{\delta}_k, \hat{\beta}_k) = \arg\min_{\delta,\beta} \left(\frac{1}{T-1} \sum_{t=1}^{T-1} (y_{t+1} - z_t'\delta - x_t'\beta)^2 + kP(\beta) \right), \tag{41}$$

where z_t includes an intercept. The penalty term $P(\beta) = \sum_{j=1}^{p} \frac{1}{2}\beta_j^2$ in case of ridge regression and $P(\beta) = \sum_{j=1}^{p} |\beta_j|$ 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; $\ln k \in \{-30, \dots, 0\}$ for lasso and $\ln k \in \{-15, \dots, 15\}$ 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 mean squared forecast error relative to that of the prevailing mean model that takes $\bar{y}_{T+1} = \frac{1}{T-1}\sum_{t=1}^{T-1} y_{t+1}$. The mean squared forecast error is computed as

$$MSFE = \frac{1}{M}\sum_{j=1}^{M}(y_{T+1}^{(j)} - \hat{y}_{T+1}^{(j)})^2, \tag{42}$$

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 $M = 10,000$.

3.2 Simulation results

3.2.1 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 5 in the appendix. 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 5 in the appendix, show that in case of a weak signal, increasing the subspace dimension 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 decreases 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 projections and random subsets 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, when $b = 2$ the

Table 1: Monte Carlo simulation: MSFE under optimal subspace dimension

b	RP	RS	PC	PL	RI	LA
			$s = 10$			
0.5	0.966 (2)	0.966 (2)	1.259 (1)	9.698 (1)	0.969 (-3.8)	1.000 (-30.0)
1.0	0.866 (8)	0.867 (8)	1.052 (1)	3.087 (1)	0.860 (-2.3)	0.960 (-28.2)
2.0	0.630 (22)	0.629 (22)	0.953 (7)	0.962 (1)	0.632 (-1.1)	0.648 (-27.6)
			$s = 50$			
0.5	0.831 (10)	0.829 (10)	1.049 (1)	2.492 (1)	0.829 (-2.0)	0.974 (-28.2)
1.0	0.574 (25)	0.574 (25)	0.869 (14)	0.796 (1)	0.579 (-0.8)	0.724 (-27.6)
2.0	0.289 (46)	0.290 (46)	0.428 (43)	0.372 (2)	0.304 (0.5)	0.369 (-26.7)
			$s = 100$			
0.5	0.715 (16)	0.714 (16)	0.998 (1)	1.383 (1)	0.712 (-1.4)	0.872 (-27.9)
1.0	0.436 (35)	0.436 (35)	0.667 (25)	0.535 (1)	0.438 (-0.2)	0.569 (-27.3)
2.0	0.195 (56)	0.195 (56)	0.277 (61)	0.236 (3)	0.200 (0.8)	0.259 (-26.4)

Note: this table reports the MSFE relative to the benchmark of the prevailing mean, for the optimal value of k corresponding to the minimum MSFE which is given in brackets. For additional information, see the note following Figure 5

random subspace methods perform better, whereas for very weak signals with $b = 0.5$ 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 exactly the same subspace dimension. Principal components is observed to select less factors 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 all together. Note that where the parameter k has a intuitive appeal in the dimension reduction methods, the values in the grid of k for lasso and ridge regression methods lack interpretation.

3.2.2 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 6 in the appendix shows results for different values of k. We observe precisely what was anticipated based on the discussion in Section 2.1.3. In the top factor setting, where the nonzero coefficients are associated

Table 2: Monte Carlo Simulation: optimal subspace dimension under a factor design

b	RP	RS	PC	PL	RI	LA
			Top factor setting			
0.5	0.713 (10)	0.959 (9)	0.952 (3)	2.466 (1)	0.712 (-2.0)	0.861 (-28.2)
1.0	0.421 (21)	0.853 (27)	0.297 (10)	0.501 (1)	0.419 (-1.1)	0.474 (-27.9)
2.0	0.202 (33)	0.573 (60)	0.075 (10)	0.133 (1)	0.202 (-0.5)	0.147 (-27.6)
			Intermediate factor setting			
0.5	1.010 (1)	0.998 (1)	1.489 (1)	16.766 (1)	1.000 (-15.0)	1.000 (-29.7)
1.0	1.002 (1)	0.982 (4)	1.181 (1)	7.034 (1)	1.000 (-6.5)	1.000 (-29.4)
2.0	1.001 (1)	0.916 (16)	1.063 (1)	2.894 (1)	1.000 (-15.0)	1.000 (-30.0)

Note: this table shows the out-of-sample performance of random projection (RP), random subset (RS), principal component (PC), partial least squares (PL), ridge (RI), and lasso (LA) in the Monte Carlo simulations using a factor design and selecting the value of k that yields the minimum MSFE compared to forecasting using the prevalent mean. For additional information, see the note following Table 6.

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 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 brackets 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.

3.3 Relation between theoretical bounds and Monte Carlo experiments

The qualitative correspondence between the results from the Monte Carlo experiments and the theoretical results show that the bounds are useful to determine settings where the random 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 2.1, for the first set of experiments described above. As we found in Table 5, the values of the MSFE of the random subspace methods are almost identical to each other over the whole range of k. The bounds are closest to the exact MSFE from the Monte Carlo experiments when the signal is not too strong and for large values of k. The bound for random subset regression is tighter than the bound for random projection regression due to the lack of exact orthogonality of the projection matrix. From the Monte Carlo results, it appears that this lack of orthogonality is not a driving force behind the difference between both methods.

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 panel, corresponding to the top factor structure, shows the bound for random projection to be lower. In line with our theoretical results, the optimal subspace dimension for random projection regression is found to be lower. In the lower panel displays the MSFE in the intermediate factor setting. We observe that both the bounds and the exact Monte Carlo results indicate that random subset regression is best suited in this case.

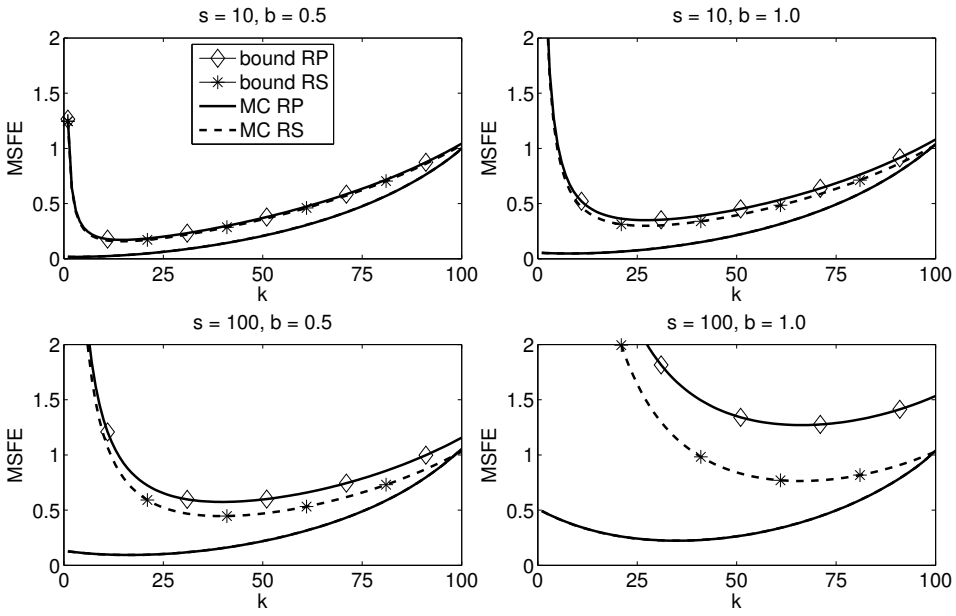
4 Empirical application

This section evaluates the predictive performance of the discussed methods in a macroeconomic application.

4.1 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),

Figure 1: Monte Carlo simulation: 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 2.1 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.

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 (2015). 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.

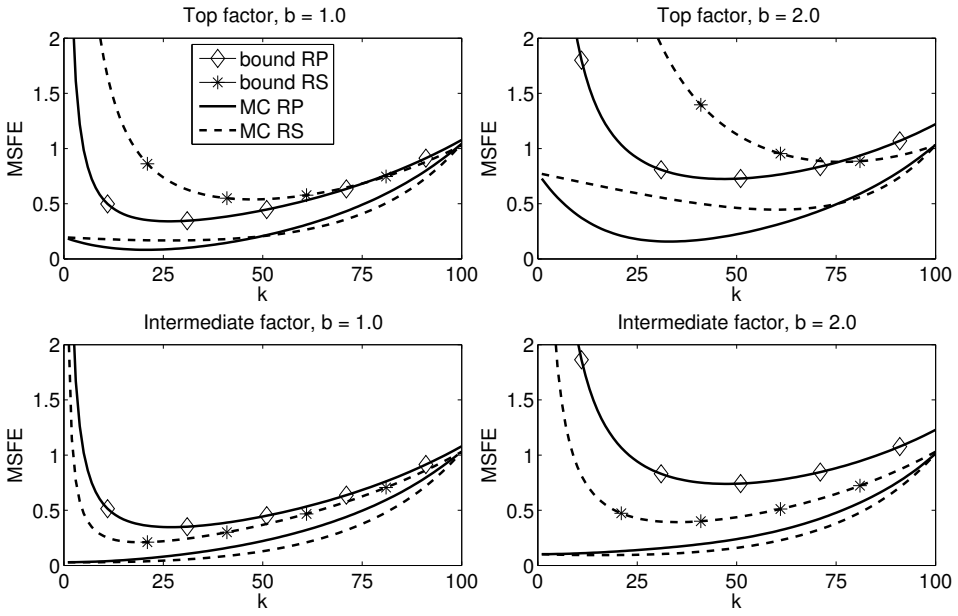
4.2 Forecasting framework

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

$$y_{t+1} = z_t'\delta + x_t' R_i \gamma_i + u_{t+1},$$

where z_t is a $q \times 1$ vector with predictors which are always included in the model and not subject to the dimension reduction methods, x_t a $p \times 1$ vector

Figure 2: Monte Carlo simulation: 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 2.1 for the top and intermediate factor settings. For additional information, see the note following 1.

with possible predictors, and R_i a $p \times k$ projection matrix. In this application y_{t+1} is one of the macroeconomic time series, z_t includes an intercept along with twelve lags of the dependent variable y_{t+1}, and x_t consists of all 129 remaining variables in the database. The predictors in x_t are projected on a low-dimensional subspace using four different projection methods whose projection matrices are discussed in Section 2: random projection regression (RP), random subset regression (RS), principal component regression (PC) and partial least squares (PL). In addition, we again compare the performance to lasso (LA) and ridge regression (RI) as described in Section 3.1, as well as to the baseline AR(12) model (AR). Predictive accuracy is measured by the MSFE defined in (42).

We use an expanding window to produce 348 forecasts, from January 1985 to December 2014. The initial estimation sample contains 312 observations and runs from January 1960 to December 1984. 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

Table 3: FRED-MD: percentage best predictive performance

		percentage loss							
		RP	RS	PC	PL	RI	LA	AR	All
percentage wins	RP		34.62	84.62	82.31	56.92	56.15	72.31	5.38
	RS	65.38		87.69	81.54	66.92	70.00	73.08	42.31
	PC	15.38	12.31		46.92	16.15	22.31	50.77	5.38
	PL	17.69	17.69	53.08		16.92	20.00	39.23	4.62
	RI	43.08	33.08	83.85	83.08		58.46	72.31	3.85
	LA	43.85	30.00	77.69	80.00	41.54		69.23	20.00
	AR	27.69	26.15	49.23	50.00	27.69	30.77		18.46

Note: this table shows the percentage wins of a method in terms of lowest MSFE compared to other methods separately, and with respect to all other methods (last column). 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. The percentages are calculated over forecasts for all 130 series in FRED-MD generated by random projections (RP), random subsets (RS), principal components (PC), partial least squares (PL), lasso (LA), ridge regression (RI), and an AR(12) model (AR). The numbers represent the percentage wins of the method listed in the rows over the method listed in the columns.

inversion, these are discarded from the average. The models generate forecasts with subspace dimension k running from 0 to 100, and 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 are included and we estimate an AR(12) model.

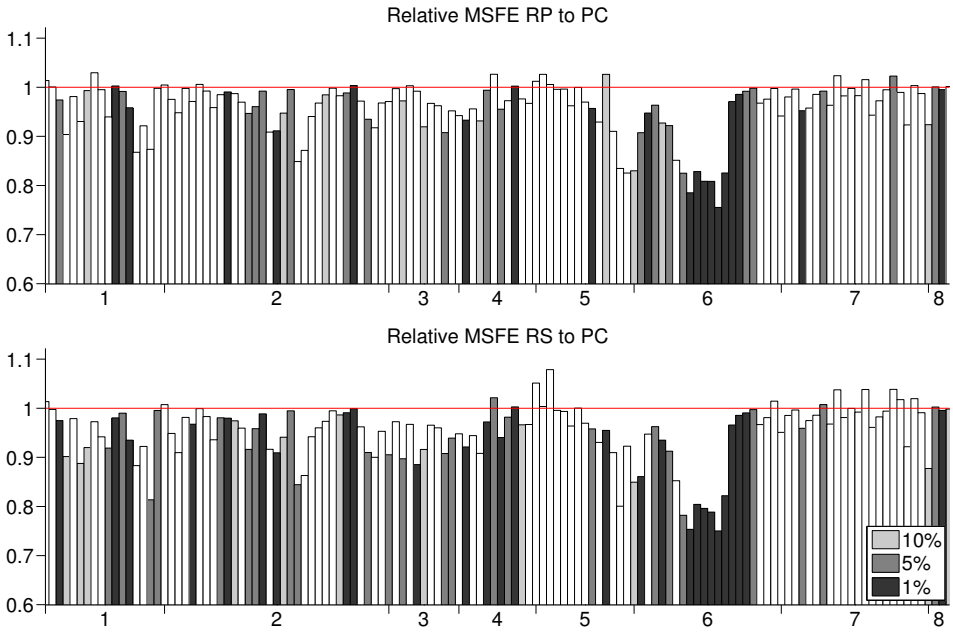
We report aggregate statistics over all 130 series, as well as detailed results for 4 major macroeconomic indicators out of the 130 series; industrial production index (INDP), unemployment rate (UNR), inflation (CPI), and the three-month Treasury Bill rate (3mTB). These series correspond to the FRED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

4.3 Empirical results

4.3.1 Aggregate statistics

We obtain series of forecasts for 130 macroeconomic variables generated by six 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 for 42% of the series. This is a substantial difference with lasso and the AR(12) model that win in approximately 20% of the cases. Random projection, principal component regression, ridge regression

Figure 3: FRED-MD: predictive accuracy of random subspace methods compared with PCR



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 (2015). 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.

and partial least squares score approximately equally well at 5%.

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 alternatives for at least 65% of the series. Interestingly, its closest 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 predictive performance of the random methods sorted by category, Figure 3 shows relative

predictive performance compared with principal component regression, for all series available in the FRED-MD dataset over the period from January 1985 through December 2014. The MSFE is calculated for the subspace dimension as determined by past predictive performance. The upper panel shows the relative MSFE of random subset regression to principal component regression and the lower panel compares random projection to principal component regression. Values below one, indicate that the random method is preferred over the benchmark. As found in Table 3, the random methods outperform the deterministic principal components in most of the cases. For random subset regression this happens in 88% of the cases, which is slightly lower for random projections with 85%. 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 almost never significant, even at a 10% level. The random methods show the largest improvements in forecast performance in category 6, which contains the interest rate and exchange rate series.

4.3.2 A case study of four key macroeconomic indicators

We look more closely into the predictive performance of the different methods on four key macroeconomic indicators: industrial production index (INDP), unemployment rate (UNR), inflation (CPI), and the three-month Treasury Bill rate (3mTB). In Table 4 we show the MSFE relative to the AR(12) model for different values of the subset 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 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 exception of the three month Treasury bill rate, 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. For the individual series reported here, the evidence is mixed. Random subset regression outperforms both lasso and ridge on industrial production and the unemployment rate series,

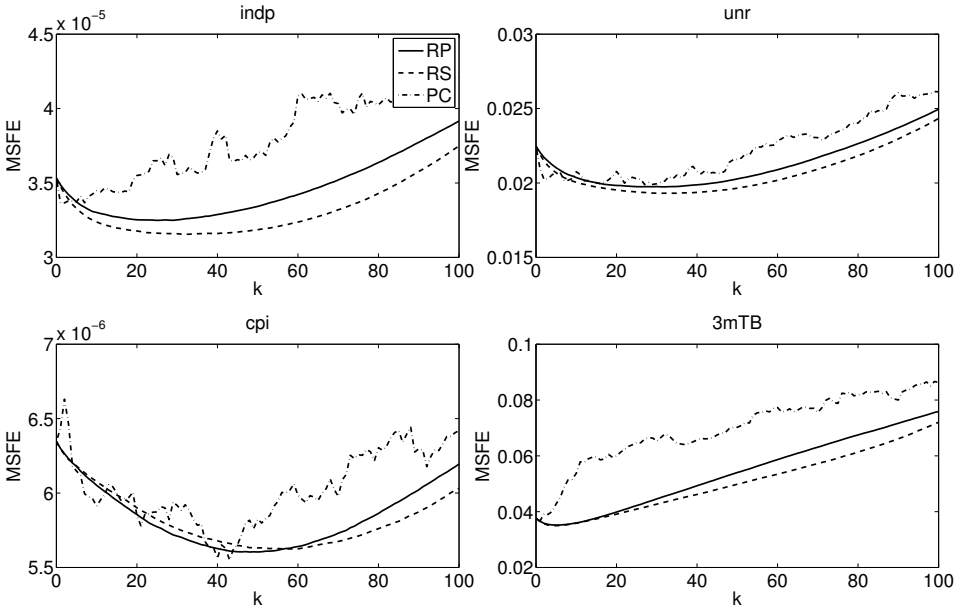
Table 4: FRED-MD: predictive accuracy relative to the AR(12)-model

	INDP	UNR	CPI	3TB	Avg.		INDP	UNR	CPI	3TB	Avg.
k	Random projection regression					k	Random subset regression				
k_R	0.955	0.884	0.899	1.123	0.969	k_R	0.912	0.863	0.915	1.255	0.962
1	0.987	0.982	0.993	0.969	0.990	1	0.984	0.976	0.992	0.966	0.987
5	0.955	0.936	0.974	0.934	0.969	5	0.942	0.921	0.974	0.929	0.964
10	0.935	0.906	0.954	0.954	0.962	10	0.917	0.892	0.958	0.952	0.957
15	0.926	0.891	0.938	1.001	0.963	15	0.905	0.878	0.943	0.993	0.957
30	0.921	0.879	0.900	1.184	0.987	30	0.894	0.860	0.908	1.133	0.972
50	0.946	0.902	0.883	1.434	1.049	50	0.902	0.875	0.887	1.323	1.017
100	1.109	1.111	0.976	2.016	1.324	100	1.061	1.083	0.950	1.913	1.278
k	Principal component regression					k	Partial least squares				
k_R	1.027	0.922	0.938	1.360	1.017	k_R	1.027	0.917	0.949	1.224	1.011
1	0.953	0.933	1.014	0.974	1.003	1	0.964	0.917	0.998	0.997	1.011
5	0.955	0.921	0.969	1.136	1.007	5	1.110	1.013	0.943	2.066	1.254
10	0.976	0.924	0.932	1.426	1.019	10	1.162	1.143	0.988	2.285	1.357
15	0.973	0.891	0.946	1.585	1.040	15	1.190	1.181	1.002	2.328	1.415
30	1.007	0.888	0.932	1.732	1.102	30	1.209	1.257	1.030	2.359	1.507
50	1.049	0.961	0.918	1.864	1.178	50	1.243	1.287	1.033	2.447	1.541
100	1.192	1.163	1.012	2.290	1.417	100	1.248	1.305	1.045	2.462	1.541
$\ln k$	Ridge regression					$\ln k$	Lasso				
k_R	0.953	0.881	0.898	1.140	0.974	k_R	0.963	0.888	0.905	1.100	0.979
-6	0.997	0.995	0.998	0.990	0.997	-28	0.956	0.934	0.962	0.953	0.979
-4	0.983	0.973	0.989	0.957	0.985	-27	0.917	0.883	0.891	1.127	0.971
-2	0.936	0.907	0.954	0.956	0.962	-26	0.927	0.901	0.901	1.435	1.024
0	0.927	0.881	0.887	1.287	1.008	-25	1.004	0.979	0.924	1.694	1.126
4	1.118	1.118	0.983	2.056	1.341	-22	1.227	1.280	1.038	2.369	1.514
8	1.261	1.324	1.058	2.464	1.592	-15	1.305	1.390	1.079	2.612	1.639
12	1.305	1.392	1.079	2.606	1.641	-5	1.305	1.392	1.080	2.613	1.641

Note: this table shows the out-of-sample performance of random projections, random subsets, principal components, lasso, and Ridge regression relative to the benchmark of an autoregressive model of order twelve, for different values of subspace dimension k and the recursively selected optimal value of k denoted by k_R. For lasso and ridge regression, the penalty parameter runs over a grid of values k. The predictive accuracy is reported for the dependent variables industrial production (INDP), unemployment rate (UNR), inflation (CPI), three month treasury bill rate (3TB), and the average over the mean squared forecast errors for all series. The predictive accuracy is measured by relative MSFE, which equals values below one when the particular method outperforms the benchmark model.

while the situation is reversed on the inflation and treasury bill rate. Random projection has a slight edge when predicting the treasury bill rate, but is close to ridge regression, which is in line with our findings in Section 3,

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



Note: this figure shows the 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 $k = 0$ correspond to the benchmark of an autoregressive model of order twelve. The four panels correspond to four dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and three month treasury bill rate (3mTB).

and lasso on all four series.

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(12) benchmark model for almost all subspace dimensions, even for very large values of k. Compared to PC and PL, we again see that the random methods select much larger values of k.

To visualize the dependence on k for the different projection 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 confirms that the random methods reach their minimum for relatively large values of k as discussed in Section 2. The selected value is

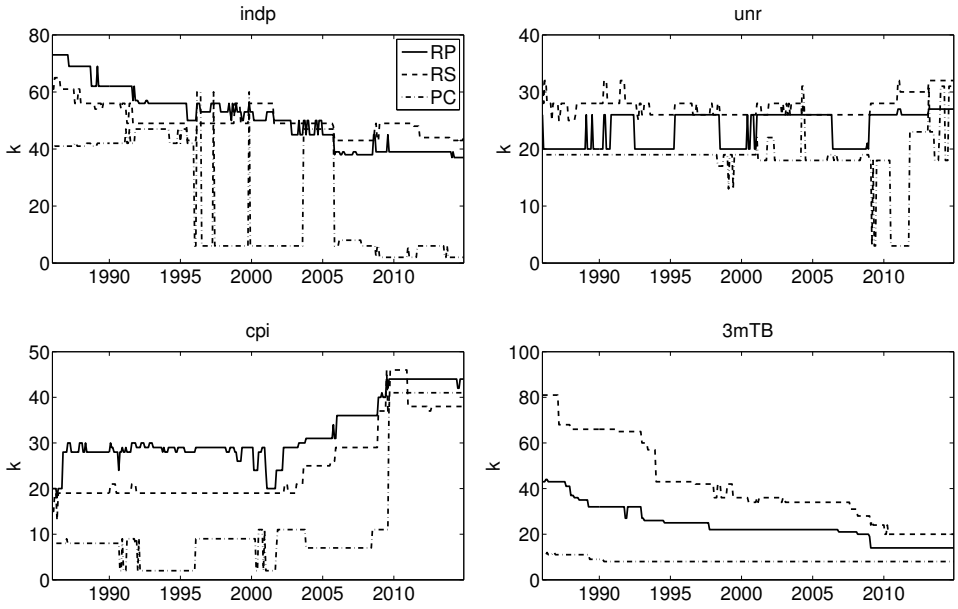
substantially larger than the selected dimension when using principal component regression. The difference is especially clear for industrial production in the upper left panel, where principal components suggests to use a single factor, while the random methods reach their minimum when using a subspace of dimension 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 methods seems to be lower for most values of k, except for inflation where a large principal component model yields more accurate results.

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. We found in Figure 4 that the minimum MSFE is lower for random subset than for random projection regression for all four series but inflation. However, the MSFE of the treasury bill rate corresponding to the recursively selected optimal value of k is lower for random projections while for all fixed k random subsets perform better. This shows that the selection of k plays an important role in the practical predictive performance of the methods.

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 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 has been gradually decreasing over time. While starting at a very large dimension around 70 in 1985, this has since dropped to values around 40. A minor effect of the global financial crisis is observed on random subset regression. 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 large 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 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.

The dimension reduction methods are expected to trade of bias and variance when the subspace dimension k varies. One would typically expect the forecast variance to be decreasing with k, while the bias is increasing with k. Figure 6 plots the bias-variance trade-off of the dimension reduction

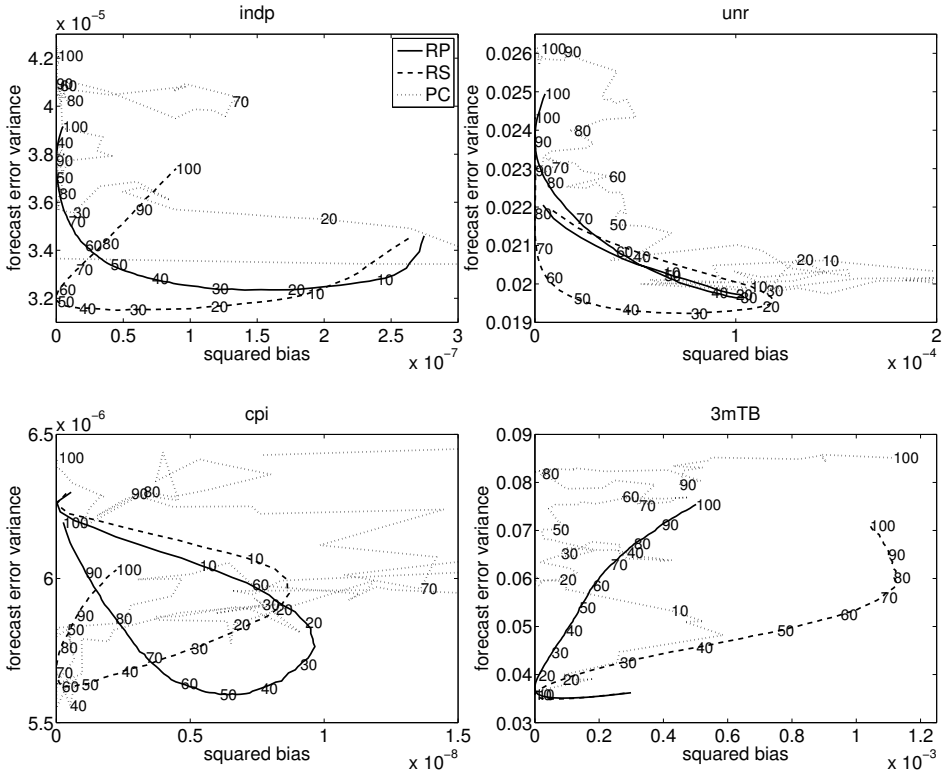
Figure 5: FRED-MD: recursive selection of subspace dimensions



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 subset 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 (3mTB).

methods. It is immediately clear that for PC, the behavior is very erratic. Although in general a large number of factors translates into a larger forecast variance, this increase is by no means uniform. For random subset regression and random projection regression, we find values for k where both the variance and the bias are smaller relative to principal components, explaining the better performance of the random method. The relationship between forecast error variance and squared bias follows a much smoother pattern over k for the random methods. Nevertheless, it is striking that also for both random methods the forecast error variance does not monotonically increase in k, and the bias not automatically declines with increasing subspace dimension. This observation is explained by the fact that the forecasts are constructed as averages over draws of projection matrices. The reported forecast error variance only includes the 'explained' part of the variance, the variance over the averaged predictions. However, there is also an unexplained part, due to the variance over the predictions within the averages.

Figure 6: FRED-MD: bias-variance trade-off



Note: this figure plots the forecast error variance against the squared bias for different values of the subspace dimension k. The different lines correspond to the dimension reduction methods random projection (RP, solid), random subset (RS, dashed), and principal component (PC, dotted) regression. The four panels correspond to four dependent variables, industrial production (INDP), unemployment rate (UNR), inflation (CPI), and the three month treasury bill rate (3mTB).

Appendix D shows that the sum of the explained and unexplained part, the total forecast error variance, increases in the subspace dimension, but due to the variance from the draws of the projection matrix, the observed forecast error variance can be decreasing in k.

5 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 tight bounds on their 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 compare the predictive accuracy to several widely used benchmarks: principal component regression, partial least squares, lasso regularization and ridge regression. The performance increases for nearly all settings under consideration compared to principal component regression and lasso regularization. Compared to ridge regression, we find large differences when we impose a factor structure on the model. When nonzero coefficients are associated with factors that explain most of the variance, random projection regression gives results very similar to ridge regression, but random subset regression is clearly outperformed. On the other hand, when the nonzero coefficients are associated with intermediate factors, random subset regression is the only method that is capable of beating the historical mean.

In the application, it seems this last scenario is prevalent, with random subset regression providing more accurate forecasts in 45% of the series. In method-by-method comparison, it outperforms the benchmarks in no less than 67% of the series. It also outperforms random projection regression in 65% of the cases. Random projection regression itself is more accurate than the benchmarks in at least 56% 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.

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

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.

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.

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

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

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

Kabán, A., Bootkrajang, J., and Durrant, R. J. (2015). Toward large-scale continuous EDA: A random matrix theory perspective. *Evolutionary computation*.

Kapetanios, G. and Marcellino, M. (2010). Factor-gmm estimation with large sets of possibly weak instruments. *Computational Statistics & Data Analysis*, 54(11):2655–2675.

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

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

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

McCracken, M. W. and Ng, S. (2015). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, (just-accepted).

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.

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. (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.

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

Tucci, G. H. and Wang, K. (2011). New methods for handling singular sample covariance matrices. *arXiv preprint arXiv:1111.0235*.

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

A Proof of Theorem 1

We start by noting that by Jensen's inequality

$$\mathrm{E}\left[\left(x_T'\beta - x_T'\mathrm{E}_{R_i}\left[R_i\hat{\gamma}_i\right]\right)^2\right] \leq \mathrm{E}_{R_i}\mathrm{E}\left[\left(x_T'\beta - x_T'R_i\hat{\gamma}_i\right)^2\middle| R_i\right] \tag{43}$$

Furthermore, since by assumption $\mathrm{E}[x_T x_T'] = \Sigma_X$ and $\hat{\gamma}_i$ is independent of x_T we have that

$$\begin{aligned}&\mathrm{E}\left[\left(x_T'\beta - x_T'R_i\hat{\gamma}_i\right)^2\right]\\ &\quad= \mathrm{E}\left[\left(\beta - R_i\hat{\gamma}_i\right)'\Sigma_X\left(\beta - R_i\hat{\gamma}_i\right)\right] + o_p(T^{-1})\end{aligned} \tag{44}$$

For the MSFE, we now have

$$\begin{aligned}&\mathrm{E}\left[\left(x_T'\left(\beta - \mathrm{E}_{R_i}\left[R_i\hat{\gamma}_i\right]\right)\right)^2\right] =\\ &\quad\leq \mathrm{E}_{R_i}\mathrm{E}\left[\left|\left|\Sigma_X^{1/2}\left(\beta - R_i\hat{\gamma}_i\right)\right|\right|^2\middle| R_i\right] + o_p(T^{-1})\\ &\quad= \mathrm{E}_{R_i}\mathrm{E}\left[\left|\left|\Sigma_X^{1/2}\left(\beta - R_i\gamma_i - R_i(\hat{\gamma}_i - \gamma_i)\right)\right|\right|^2\middle| R_i\right] + o_p(T^{-1})\\ &\quad= \mathrm{E}_{R_i}||\Sigma_X^{1/2}(\beta - R_i\gamma_i)||^2 + \mathrm{E}_{R_i}\mathrm{E}\left[\left|\left|\Sigma_X^{1/2}R_i(\hat{\gamma} - \gamma_i)\right|\right|^2\middle| R_i\right]\\ &\qquad- 2\mathrm{E}_{R_i}\mathrm{E}\left[(\beta - R_i\gamma_i)'\Sigma_X R_i(\hat{\gamma}_i - \gamma_i)\middle| R_i\right] + o_p(T^{-1})\end{aligned} \tag{45}$$

The parameter γ_i is estimated by OLS and we have

$$XR_i(\hat{\gamma}_i - \gamma_i) = P_{XR_i}X(\beta - R_i\gamma_i) + P_{XR_i}\varepsilon \tag{46}$$

where P_{XR_i} denotes the projection matrix on the subspace spanned by the columns of XR_i. The crucial step, observed in Kabán (2014), is that γ_i is the optimal parameter vector in the low-dimensional subproblem, defined as

$$\gamma_i = \arg\min_u \mathrm{E}\left[\sum_{t=1}^{T-1}\left(y_{t+1} - x_t'R_iu\right)^2\middle| R_i\right] \tag{47}$$

This implies the following inequality

$$||X\beta - XR_i\gamma_i||^2 \leq ||X\beta - XR_iR_i'\beta||^2 \tag{48}$$

Substituting (48) and (46) into (45) and using that $\frac{1}{T}X'X = \Sigma_X + o_p(T^{-1})$ we obtain

$$\begin{aligned}
&E\left[(x_T'\beta - x_T'\mathrm{E}_{R_i}\left[R_i\hat{\gamma}_i\right])^2\right] \\
&\quad\leq \sigma^2\frac{k}{T} + \mathrm{E}_{R_i}\left[(\beta - R_i\gamma_i)'\Sigma_X(\beta - R_i\gamma_i)\right] \\
&\qquad - \mathrm{E}_{R_i}\left[\left|\left|P_{\Sigma_X^{1/2}R_i}\Sigma_X^{1/2}(\beta - R\gamma_i)\right|\right|^2\right] + o_p(T^{-1}) \\
&\quad\leq \sigma^2\frac{k}{T} + \mathrm{E}_{R_i}\left[\beta'(I - R_iR_i')\Sigma_X(I - R_iR_i')\beta\right] \\
&\qquad - \mathrm{E}_{R_i}\left[\left|\left|P_{\Sigma_X^{1/2}R_i}\Sigma_X^{1/2}(\beta - R_i\gamma_i)\right|\right|^2\right] + o_p(T^{-1})
\end{aligned} \tag{49}$$

Finally, (47) has a simple solution

$$\gamma_i = \left(\frac{1}{T-1}\sum_{t=1}^{T-1}R_i'x_tx_t'R_i\right)^{-1}\left(\frac{1}{T-1}\sum_{t=1}^{T-1}R_i'x_tx_t'\beta\right) \tag{50}$$

Hence

$$\Sigma_X^{1/2}(\beta - R_i\gamma_i) = \left(I - P_{\Sigma_X^{1/2}R_i}\right)\Sigma_X^{1/2}\beta \tag{51}$$

which shows that the last term of (49) is identically zero.

B Derivation of equation (19)

For the difference between the MSFE under random projection and orthogonalized random projection we have that

$$\begin{aligned}
\Delta &= \frac{1}{k}\left[\beta'\Sigma_X\beta + \mathrm{trace}(\Sigma_X)\beta'\beta\right] \\
&\quad - \frac{p-k}{k}\frac{1}{p^2-1}\left[p\mathrm{trace}(\Sigma_X)\beta'\beta - \beta'\Sigma_X\beta\right] \\
&= \frac{1}{k}\left[\beta'\Sigma_X\beta\left(1 + \frac{p-k}{p^2-1}\right) - \mathrm{trace}(\Sigma_X)||\beta||^2\frac{kp-1}{p^2-1}\right] \\
&\geq \frac{\beta'\Sigma_X\beta}{k}\frac{p^2-1+p-k-kp+1}{p^2-1} \\
&= \frac{\beta'\Sigma_X\beta}{k}\frac{p-k}{p-1} \\
&\geq 0
\end{aligned} \tag{52}$$

In the third line we use the fact that $\Sigma_X = U\Lambda U'$ with U an orthogonal matrix and Λ a diagonal matrix consisting of the non-negative eigenvalues of Σ_X. Then

$$\begin{aligned}\beta'\Sigma_X\beta - \text{trace}(\Sigma_X)||\beta||^2 &= \beta'(\Sigma_X - \text{trace}(\Sigma_X))\beta \\ &= \beta' U\left[\Lambda - \left(\sum_{i=1}^{p}\lambda_i\right) I\right] U'\beta \\ &\leq 0\end{aligned} \tag{53}$$

where the last inequality holds since each term on the diagonal satisfies $\lambda_i - \sum_{j=1}^{p}\lambda_j = -\sum_{j\neq i}\lambda_j < 0$.

C Optimal bounds

The optimal, but infeasible, choice of k that minimizes the bounds is given by

$$\begin{aligned}k^*_{RSR} &= \left[\frac{T}{\sigma^2} p \frac{p}{p-1}\left(\beta' D_{\Sigma_X}\beta - \frac{1}{p}\beta\Sigma_X\beta\right)\right]^{1/2} \\ k^*_{RP} &= \left[\frac{T}{\sigma^2}\left(\beta'\Sigma_X\beta + \text{trace}(\Sigma_X)\beta'\beta\right)\right]^{1/2}\end{aligned} \tag{54}$$

The optimal choice of k leads to the following bound for random subset regression

$$\begin{aligned}&E\left[\left(x_T'\beta - x_T'\mathrm{E}^{RS}_{R_i}\left[R_i\hat{\gamma}_i\right]\right)^2\right] = \\ &\leq 2\left[\frac{\sigma^2}{T} p \frac{p}{p-1}\left(\beta' D_{\Sigma_X}\beta - \frac{1}{p}\beta\Sigma_X\beta\right)\right]^{1/2} \\ &\quad - \frac{p}{p-1}\sum_{j=1}^{p}\left(\beta' D_{\Sigma_X}\beta - \frac{1}{p}\beta\Sigma_X\beta\right) + o_p(T^{-1})\end{aligned} \tag{55}$$

For random projection regression under $k = k^*_{RP}$ we have

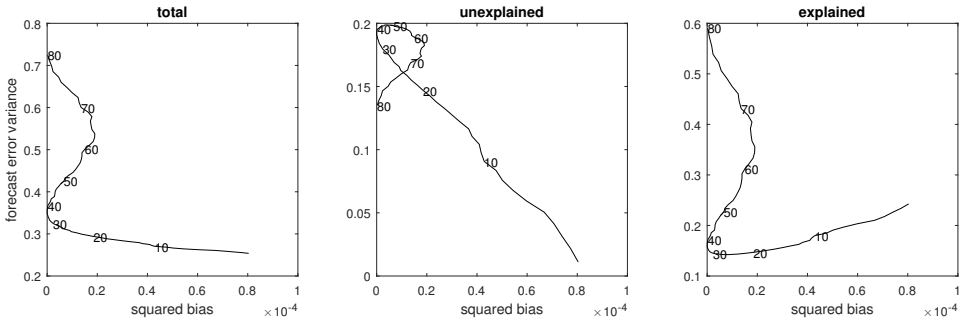
$$\begin{aligned}&E\left[\left(x_T'\beta - x_T'\mathrm{E}^{RP}_{R_i}\left[R_i\hat{\gamma}_i\right]\right)^2\right] = \\ &\leq 2\left[\frac{\sigma^2}{T}\left(\beta'\Sigma_X\beta + \text{trace}(\Sigma_X)\beta'\beta\right)\right]^{1/2} + o_p(T^{-1})\end{aligned} \tag{56}$$

D Application: bias-variance tradeoff

The mean squared forecast error can be decomposed in a bias and a variance component:

$$\mathrm{E}\left[(y_{t+1} - \mathrm{E}_{R_i}[\hat{y}^i_{T+1}])^2\right] = \mathrm{E}\left[y_{t+1} - \mathrm{E}_{R_i}[\hat{y}^i_{T+1}]\right]^2 + \mathrm{Var}\left[y_{t+1} - \mathrm{E}_{R_i}[\hat{y}^i_{T+1}]\right]$$

Figure 7: Bias-variance Trade-off



Note: this figure plots the forecast error variance against the squared bias for different values of the subspace dimension k. The three panels show the total forecasts error variance, the unexplained and the explained part. The forecasts are generated by random projections in a simulation design as discussed in Section 3.1, where we us $M = 1000$ replications with $b = 1$ and $s = 50$.

The first term equals the squared bias of the forecasts and the second term the forecast error variance. However, since we average over realizations of R_i, the second term only includes the explained component of the forecast error variance. This can be illustrated by applying the law of total variance on the forecast error of all generated predictions:

$$\text{Var}\left[y_{t+1} - \hat{y}^i_{T+1}\right] = \text{E}\left[y_{t+1} - \text{Var}_{R_i}[\hat{y}^i_{T+1}]\right] + \text{Var}\left[y_{t+1} - \text{E}_{R_i}[\hat{y}^i_{T+1}]\right]$$

where the left term equals the unexplained and the right term the explained component of the forecasts error variance.

Because of computational constraints, we do not store predictions for all different projection matrices in the empirical application. Hence, we setup a Monte Carlo experiment to investigate the behaviour of the unexplained and explained components of the forecast error variance. The simulation design is a small scale version of the experiments explained in Section 3.1, where we use $M = 1000$ replications with $b = 1$ and $s = 50$ to generate forecasts with random projection regressions. Figure 7 shows the bias-variance trade-off for the total variance and the unexplained and explained variance components. The total variance behaves as expected; the forecast error variance increase with the subspace dimension k. The unexplained component shows unpredictable behaviour, which causes that the explained variance is not always increasing in k. The third panel of Figure 7 shows similar patterns as we find in Figure 6, which shows the bias-variance trade-off in the empirical example. The empirical findings can be explained by the fact that the reported forecast error variance leaves out the unexplained part, leading to a forecast error variance that can decrease for larger subspace dimensions.

Table 5: Monte Carlo simulation: MSFE relative to prevailing mean

		Random projections - k				Random subsets - k			
s	b	1	10	25	50	1	10	25	50
10	0.5	0.977	1.291	3.584	11.861	0.977	1.301	3.626	11.938
	1.0	0.968	0.875	1.382	3.873	0.967	0.876	1.396	3.889
	2.0	0.964	0.732	0.635	1.091	0.964	0.729	0.635	1.096
50	0.5	0.965	0.831	1.188	3.160	0.965	0.829	1.196	3.174
	1.0	0.963	0.716	0.574	0.885	0.962	0.714	0.574	0.889
	2.0	0.962	0.682	0.408	0.293	0.961	0.679	0.406	0.293
100	0.5	0.964	0.756	0.781	1.668	0.963	0.753	0.782	1.673
	1.0	0.962	0.697	0.473	0.512	0.962	0.693	0.472	0.513
	2.0	0.961	0.678	0.386	0.202	0.961	0.674	0.384	0.202
		Principal components - k				Partial least squares - k			
s	b	1	10	25	50	1	10	25	50
10	0.5	1.259	3.883	8.929	19.732	9.698	41.613	50.135	52.279
	1.0	1.052	1.696	3.143	6.385	3.087	13.005	15.610	16.278
	2.0	0.990	0.961	1.085	1.733	0.962	3.455	4.192	4.408
50	0.5	1.049	1.477	2.584	5.231	2.492	10.157	12.189	12.732
	1.0	0.979	0.886	0.941	1.416	0.796	2.781	3.371	3.525
	2.0	0.960	0.733	0.518	0.438	0.438	0.679	0.821	0.864
100	0.5	0.998	1.097	1.493	2.761	1.383	5.241	6.326	6.642
	1.0	0.971	0.783	0.667	0.790	0.535	1.345	1.621	1.703
	2.0	0.959	0.690	0.451	0.287	0.371	0.335	0.424	0.444
		Ridge regression - $\ln k$				Lasso - $\ln k$			
s	b	-6	-4	-2	0	-28	-27	-26	-25
10	0.5	0.993	0.972	1.370	7.359	1.526	6.449	16.136	27.703
	1.0	0.990	0.948	0.873	2.370	0.998	2.239	4.950	8.326
	2.0	0.989	0.937	0.707	0.818	0.677	0.818	1.475	2.378
50	0.5	0.990	0.945	0.829	1.981	1.006	1.953	4.111	6.877
	1.0	0.988	0.934	0.685	0.689	0.760	0.803	1.257	1.911
	2.0	0.985	0.917	0.599	0.306	0.516	0.374	0.404	0.521
100	0.5	0.989	0.940	0.741	1.115	0.872	1.197	2.225	3.585
	1.0	0.988	0.929	0.648	0.449	0.671	0.569	0.720	1.007
	2.0	0.982	0.900	0.546	0.218	0.425	0.281	0.262	0.300

Note: this table shows the MSFE divided by that of the prevailing mean forecast, for random projection regression, random subset regression, principal component regression, partial least squares, lasso, and ridge regression under the data generating process (39) 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 6: 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
	0.5	0.942	0.713	1.217	3.872	0.992	0.959	1.145	2.599
Top	1.0	0.936	0.552	0.438	1.062	0.991	0.917	0.854	1.053
	2.0	0.935	0.510	0.230	0.287	0.990	0.903	0.764	0.595
	0.5	1.010	1.834	5.749	19.192	0.998	1.213	2.797	11.190
Int.	1.0	1.002	1.299	2.735	7.629	0.993	1.015	1.497	4.435
	2.0	1.001	1.068	1.363	2.336	0.990	0.929	0.947	1.558
		Principal components - k				Partial least squares - k			
s	b	1	10	25	50	1	10	25	50
	0.5	0.976	1.082	2.774	6.390	2.466	13.681	16.341	17.293
Top	1.0	0.901	0.297	0.745	1.719	0.501	3.704	4.393	4.602
	2.0	0.883	0.075	0.192	0.449	0.133	0.936	1.125	1.181
	0.5	1.489	5.917	14.398	32.184	16.766	66.078	78.234	82.060
Int.	1.0	1.181	2.943	6.388	12.876	7.034	24.611	29.615	31.166
	2.0	1.063	1.637	2.722	4.077	2.894	7.410	8.587	8.970
		Ridge regression - $\ln k$				Lasso - $\ln k$			
s	b	-6	-4	-2	0	-28	-27	-26	-25
	0.5	0.983	0.908	0.712	2.296	2.367	5.358	9.181	13.919
Top	1.0	0.981	0.891	0.517	0.680	0.737	1.516	2.582	3.879
	2.0	0.976	0.867	0.417	0.226	0.201	0.391	0.648	0.968
	0.5	1.001	1.025	1.931	11.236	10.792	25.880	45.308	68.811
Int.	1.0	1.000	1.007	1.340	4.761	4.749	10.264	17.290	25.895
	2.0	1.000	1.002	1.083	1.774	1.772	3.053	4.856	7.162

Note: this table shows the out-of-sample performance of random projection regression (RP), random subset regression (RS), principal component regression (PC), partial least squares (PL), ridge regression (RI), and lasso (LA) 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 5.