Journal of Econometrics 209 (2019) 391–406

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Forecasting using random subspace methods

Tom Boot [a,*], Didier Nibbering [b]

[a] Department of Economics, Econometrics and Finance, University of Groningen, P.O. box 800, 9700 AV Groningen, The Netherlands
[b] Department of Econometrics & Business Statistics, Monash University, Clayton VIC 3800, Australia

ARTICLE INFO

Article history:
Received 23 December 2016
Received in revised form 17 July 2018
Accepted 25 January 2019
Available online 8 February 2019

JEL classification:
C32
C38
C53
C55

Keywords:
Dimension reduction
Forecasting
Random subspace

ABSTRACT

Random subspace methods are a new approach to obtain accurate forecasts in high-dimensional regression settings. Forecasts are constructed by averaging over forecasts from many submodels generated by random selection or random Gaussian weighting of predictors. This paper derives upper bounds on the asymptotic mean squared forecast error of these strategies, which show that the methods are particularly suitable for macroeconomic forecasting. An empirical application to the FRED-MD data confirms the theoretical findings, and shows random subspace methods to outperform competing methods on key macroeconomic indicators.

1. Introduction

Due to the increase in available macroeconomic data, dimension reduction methods have become an important tool to obtain accurate forecasts. A well-known approach is to reduce the dimension of the predictor set by identifying a small set of factors that drive most of the variation in the high-dimensional predictor set, as in Stock and Watson (2002a, 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. A priori there seems no reason to expect this to lead to accurate forecasts, but empirical evidence suggests otherwise. For example, Elliott et al. (2013, 2015) find that averaging over forecasts constructed from randomly selected subsets of predictors substantially lowers the mean squared forecast error compared with data-driven alternatives. The theoretical justification of these randomized approaches is not completely understood. We provide new theoretical results which suggest that random subspace methods are particularly useful for macroeconomic forecasting, and provide empirical evidence that supports these conclusions.

We distinguish two different approaches to construct a random subspace. Random subset regression uses randomly selected subsets of predictors to estimate many low-dimensional approximations to the original model. The forecasts from these submodels are then combined in order to lower the mean squared forecast error (MSFE). Instead of selecting subsets of available predictors, Gaussian random projection regression forms a low-dimensional subspace by averaging over predictors using random weights drawn from a standard normal distribution.

* Corresponding author.
E-mail addresses: t.boot@rug.nl (T. Boot), didier.nibbering@monash.edu (D. Nibbering).

https://doi.org/10.1016/j.jeconom.2019.01.009
0304-4076/

We derive upper bounds on the asymptotic MSFE for random subset regression and Gaussian random projection regression, and use these bounds to determine in which settings the methods are most effective. We find that when factors that explain most variation in the predictors drive the dependent variable, the bias in the forecast is relatively small. This suggests that random subspace methods are particularly suited for macroeconomic forecasting. The bounds furthermore indicate that in this set-up random subset regression is preferred over Gaussian random projection regression. A numerical study shows that the exact mean squared forecast error reflects the intuition gained from the bounds.

The bounds are derived for forecasts that take the expected value over the random matrix used to generate the subspace. 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.

An application to 130 macroeconomic time series in the FRED-MD data set of McCracken and Ng (2016) confirms the theoretical findings. The random subspace methods improve forecast accuracy compared to well-known alternatives. Compared to principal component regression, the random subspace methods appear more robust to the selected dimension of the subspace.

Random sampling methods are widely used in the statistics and machine learning literature but rather new to economics (Ng, 2015). Similar to random subset regression, bootstrap aggregation methods fit models on subsets of the data, and then aggregate the forecasts of the submodels to a final prediction. However, bootstrap aggregation methods originally select among observations instead of predictors. When drawing subsets of observations with replacement, this is known as 'bagging' (Breiman, 1996), and when drawing without replacement as 'pasting' (Breiman, 1999). Alternative methods to reduce the number of observations are provided by Ma et al. (2015). Inoue and Kilian (2008) show that bagging does not convincingly outperform econometric benchmark models in forecasting inflation.

Random subset regression is a form of bootstrap aggregation where one draws subsets of predictors instead of observations. This is also referred to as 'feature bagging'. In contrast to a computational motivation (Lu et al., 2013), we focus on statistical gains achieved by reducing the dimension of the set of predictors. We also do not apply further model selection methods to the low-dimensional models, as is often the case for bootstrap aggregation.

The justification for random projection regression is usually derived from the Johnson–Lindenstrauss lemma (Johnson and Lindenstrauss, 1984), which has recently inspired economic applications on discrete choice models by Chiong and Shum (2016), forecasting product sales by Schneider and Gupta (2016), and forecasting using large vector autoregressions by Koop et al. (2016) based on the framework of Guhaniyogi and Dunson (2015). In contrast to alternative methods to construct a new set of predictors, as in Frieze et al. (2004) and Mahoney and Drineas (2009), random projection regression uses predictor weights independent of the data, involves only a single tuning parameter, and is simple to implement. Upper bounds are derived on 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 these results, and applies to a range of time-series models.

This paper 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. The intuition that follows from the theoretical results is verified in a numerical experiment in Section 4. Section 5 discusses an empirical application to 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 predictors thought to be essential to the model, such as a constant and lags of y_{t+1}, x_t is a $p_x \times 1$ vector of possibly relevant predictors, and the forecast error is denoted by ε_{t+1}. The time index t runs from $t = 0, \ldots, T$. We assume that $\mathrm{E}[\varepsilon_{t+1}|w_t, x_t] = 0$ and $\mathrm{E}[\varepsilon_{t+1}^2|w_t, x_t] = \sigma^2$. 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 asymptotic mean squared forecast error 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 $\hat{y}_{T+1}$.

For any estimator $\tilde{\beta}$ for β, we define the asymptotic mean squared forecast error as

$$\rho = \lim_{T\to\infty} T\mathrm{E}\left[\left(y_{T+1} - z_T'\tilde{\beta}\right)^2 - \sigma^2\right] = \lim_{T\to\infty} T\mathrm{E}\left[(z_T'\beta - z_T'\tilde{\beta})^2\right], \tag{2}$$

following for example Hansen (2010) and Hirano and Wright (2017). The error variance σ^2 is subtracted as it arises from the error ε_{T+1}, which is unpredictable for any method, and the operator $\mathrm{E}[.]$ denotes the expectation with respect to $\{(z_t', \varepsilon_{t+1})\}_{t=0}^T$. The methods in this paper reduce the dimension of the predictors in x_t to $k < p_x$, while maintaining the dimension of the predictors in w_t. We indicate this by writing $\rho(p_w, k)$ for the asymptotic mean squared forecast error.

Denote the ordinary least squares (OLS) estimator without dimension reduction as $\hat{\beta}$. This estimator yields the following forecast,

$$\hat{y}_{T+1}^{\text{OLS}} = z_T'\hat{\beta} = z_T'(Z'Z)^{-1}Z'y, \tag{3}$$

where $y = (y_1, \ldots, y_T)'$, $Z = (z_0, \ldots, z_{T-1})'$. Then,

$$\begin{aligned}\rho(p_w, p_x) &= \lim_{T\to\infty} TE\left[(\hat{\beta}-\beta)'z_T z_T'(\hat{\beta}-\beta)\right] \\ &= \lim_{T\to\infty} TE\left[(\hat{\beta}-\beta)'\Sigma_z(\hat{\beta}-\beta)\right] \\ &= \sigma^2(p_w + p_x),\end{aligned} \tag{4}$$

where $\Sigma_z = \text{E}[z_t z_t']$ for $t = 0, \ldots, T$. The second line holds since z_T is asymptotically independent of $\sqrt{T}(\hat{\beta}-\beta)$, which we show in Appendix A.1, and the final line is implied by the assumptions in Section 3, under which $\sqrt{T}(\hat{\beta}-\beta) \Rightarrow N(0, \sigma^2\Sigma_z^{-1})$.

Since the asymptotic MSFE in (4) increases with the number of estimated coefficients, the forecast accuracy might benefit from dimension reduction techniques when x_t contains a large number of predictors. Although dimension reduction generally lowers the forecast variance, it may induce omitted variable bias that increases the MSFE.

2.1. Random subspace methods

A generic approach to dimension reduction is to multiply x_t with a $p_x \times k$ matrix R, where $k < p_x$, to obtain the approximating model

$$y_{t+1} = w_t'\beta_w + x_t'R\beta_{x,R} + u_{t+1}. \tag{5}$$

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 predictors x_t. Alternatively, random subspace methods generate the elements of R from a probability distribution that is independent of the data. By averaging over forecasts based on different draws of R, random subspace methods reduce the forecast variance, while maintaining most of the signal. We consider 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. Define an index $l = 1, \ldots, k$ 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 its $c(l)$th entry equal to one, then random subset regression is based on random matrices

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

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. This paper considers Gaussian random projections, which draws the weights from a normal distribution. In this case, each entry of R is independent and identically distributed as

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

Alternative choices not considered in this paper are for example the database-friendly random projections proposed by Achlioptas (2003), or the fast Johnson–Lindenstrauss transform by Ailon and Chazelle (2009).

2.2. Forecasts from low-dimensional models

We rewrite the approximating model (5) 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 given by

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

Using this estimate, we construct a forecast for y_{T+1} as

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

If R is a random matrix, relying on a single realization is suboptimal. Jensen's inequality shows that averaging over forecasts based on different realizations of R will lower the MSFE,

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

where E_R denotes the expectation only with respect to the random matrix R. We therefore 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. Theorem 2 in Section 3.2 shows that this does not affect the mean squared forecast error as long as the number of draws of R is of $O(p_x \log p_x)$. 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. $\left\{(z_t',\varepsilon_{t+1})\right\}$ 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. $E[z_tz_t']=\Sigma_z=\begin{pmatrix}\Sigma_w & \Sigma_{wx}\\ \Sigma_{xw} & \Sigma_x\end{pmatrix}$ is positive definite.

A6. $Var[T^{-1/2}Z'\varepsilon]=\sigma^2\Sigma_z$.

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

The mixing size a in Assumption A1 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. The necessity for this independence in deriving the asymptotic MSFE has been noted in Hansen (2008), and appears implied in equation (2.2) of Hirano and Wright (2017).

Under stated assumptions, our theoretical results 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. Assumptions A1–A7, guarantee that

$$\frac{1}{\sqrt{T}}Z'\varepsilon\xrightarrow{(d)}N(0,\sigma^2\Sigma_z),\quad \operatorname*{plim}_{T\to\infty}\frac{1}{T}Z'Z=\Sigma_z, \tag{12}$$

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 $\beta_{x,0}$ is a fixed $p_x\times 1$ vector.

Under local-to-zero coefficients, the bias induced by using a low-dimensional subspace is finite, see Claeskens and Hjort (2008). When coefficients tend to zero at a slower rate compared to Assumption A8, the forecast based on OLS estimation in (3) using all predictors is asymptotically the optimal forecast.

The theoretical results allow forecasting models that assume a factor structure in x_t, such as the diffusion index model (Stock and Watson, 2002a). 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. Bai and Ng (2006) 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.

3.1. MSFE for forecasts from low-dimensional models

The asymptotic mean squared forecast error of the forecast (11) is

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

The following theorem provides an upper bound on (14).

Theorem 1. *Let $R \in \mathbb{R}^{p_x\times k}$ be a matrix such that $E_R[RR'] = (k/p_x)I_{p_x}$. Define $\Sigma = \Sigma_x - \Sigma_{xw}\Sigma_w^{-1}\Sigma_{wx}$. The asymptotic mean squared forecast error $\rho(p_w, k)$ in (14) under (1) satisfying Assumptions A1–A8, is upper bounded by*

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

A proof is presented in Appendix A.2.

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 $\sigma^2(p_w + p_x)$ as shown in (4). 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, Theorem 1 shows that this bias is strictly smaller than the bias of the naive estimator that does not use any of the predictors in x_t, which equals $\beta_{x,0}'\Sigma\beta_{x,0}$. We show below that for random subset regression and random projection regression the bound in (15) reduces to (4) when $k = p_x$.

Loosely speaking, the product $(p_x/k)RR'\Sigma RR'(p_x/k)$ first projects Σ to a k-dimensional subspace by multiplying with R from the left and the right, and then re-inflates by multiplication with $(p_x/k)R$. If little information is lost in this procedure, the final term in (15) will be close to $\beta_{x,0}'\Sigma\beta_{x,0}$, and the bias is small.

3.1.1. MSFE bound for random subset regression

For both random subset regression and random projection regression, the bound in (15) can be evaluated explicitly. For the random selection matrices in (6) we have the following result.

Lemma 1. *For random subset regression, the asymptotic MSFE $\rho(p_w, k)$ given in (14) under (1) satisfying Assumptions A1–A8, is upper bounded by*

$$\rho(p_w, k) \leq \sigma^2(p_w + k) + \beta_{x,0}'\Sigma\beta_{x,0} - \frac{k}{p_x}\beta_{x,0}'\Sigma\left[w_s\Sigma + (1 - w_s)D_\Sigma\right]^{-1}\Sigma\beta_{x,0},$$

where $w_s = (p_x - 1)^{-1}(k - 1)$, and D_Σ is a diagonal matrix with $[D_\Sigma]_{ii} = \Sigma_{ii}$.

A proof is given in Appendix A.3.

The bound for random subset regression depends on a convex combination of the full covariance matrix Σ and its diagonal elements D_Σ. When $k = 1$, all weight is put on D_Σ. In this case, 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 (4).

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*

$$\lim_{T\to\infty} TE\left[\left(z_T'\beta - z_T'E_R\left[S_R\hat{\beta}_R\right]\right)^2\right] = \lim_{T\to\infty} TE\left[\left(z_T'\beta - z_T'E_Q\left[S_Q\hat{\beta}_Q\right]\right)^2\right]. \tag{16}$$

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. The bound on the asymptotic MSFE for random projection regression follows from the fourth order moments of the elements of Q. This approach improves upon in-sample mean squared error bounds of Maillard and Munos (2009), Kabán (2014), and Thanei et al. (2017).

Lemma 3. *For random projection regression, the asymptotic MSFE $\rho(p_w, k)$ given in (14) under (1) satisfying Assumptions A1–A8, is upper bounded by*

$$\rho(p_w, k) \leq \sigma^2(p_w + k) + \beta_{x,0}'\Sigma\beta_{x,0} - \frac{k}{p_x}\beta_{x,0}'\Sigma\left[w_p\Sigma + (1 - w_p)\frac{tr(\Sigma)}{p_x}I_{p_x}\right]^{-1}\Sigma\beta_{x,0},$$

where $w_p = [(p_x + 2)(p_x - 1)]^{-1}[p_x(k + 1) - 2]$, and $tr(\Sigma)$ denotes the trace of Σ.

A proof is provided in Appendix A.5.

The bound for random projection regression depends on a convex combination of Σ and the constant diagonal matrix $\frac{\text{tr}(\Sigma)}{p_x} I_{p_x}$. When $k = 1$, nearly all weight is put on the latter. When $k = p_x$, all weight is put on Σ and the bound reduces to the exact expression for the unbiased estimator provided in (4).

3.2. MSFE under a finite number of draws

The derived bounds are valid for forecasts that depend on the expectation over the random matrix R. In practice, we need to approximate this expectation by averaging over a finite number of draws of the matrix R. 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 $\hat{y}_{T+1,S} = N^{-1} \sum_{i=1}^{N} \hat{y}_{T+1,R_i}$, 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 $\rho_S(p_w, k)$ the asymptotic MSFE based on $\hat{y}_{T+1,S}$. Denote by $\rho(p_w, k)$ the asymptotic MSFE based on $\hat{y}_{T+1}$ as in (14). When $N = O(p_x \log p_x)$,*

$$\rho_S(p_w, k) = (1 + \epsilon)\rho(p_w, k), \tag{17}$$

for an arbitrarily small constant $\epsilon > 0$.

A proof is provided in Appendix A.6.

Theorem 2 provides a theoretical justification of the results obtained in Elliott et al. (2013, 2015), who found no loss in accuracy when using a relatively small number of random subsets instead of all available subsets.

3.3. Comparison between the random subspace methods

The derived bounds identify cases where the random subspace methods are expected to perform well, and where one of the two methods is preferred over the other. Consider the case where all predictors are subject to dimension reduction, so that Lemmas 1 and 3 apply with $p_w = 0$ and $\Sigma = \Sigma_x$. Conform common practice in macroeconomic forecasting, we assume that the predictors are standardized, such that $[\Sigma_x]_{ii} = 1$ for $i = 1, \ldots, p_x$.

Decompose the covariance matrix $\Sigma_x = H \Lambda H'$, where the orthogonal matrix H contains the eigenvectors of Σ_x and Λ is a diagonal matrix with eigenvalues sorted in decreasing order. The bounds for both random subspace methods reduce to

$$\rho(0, k) \leq \sigma^2 k + \sum_{i=1}^{p_x} \lambda_i \left(1 - \frac{k}{p_x} \frac{\lambda_i}{w_r(\lambda_i - 1) + 1}\right) \left(\sum_{j=1}^{p_x} [\beta_{x,0}]_j h_{ji}\right)^2, \tag{18}$$

where $\lambda_i = [\Lambda]_{ii}$, $h_{ji} = [H]_{ji}$, and w_r is given for random subset regression by w_s in Lemma 1, and for random projection regression by w_p in Lemma 3.

Interpretation of the bound in (18) is further simplified by rewriting the data generating process in (1) to

$$y_{t+1} = f_t' \gamma + \varepsilon_{t+1}, \tag{19}$$

where $f_t' = x_t' H$ is a vector of principal components corresponding to the eigenvalues of Σ_x in decreasing order. Since the random subspace methods are applied to the predictors x_t, and $f_t' \gamma = x_t' H \gamma$, the random subspace methods estimate the parameter vector $\beta_x = H\gamma$. In line with (13), we assume that the factors are weakly related to the dependent variable and define $\gamma = \gamma_0 / \sqrt{T}$, where $\gamma_0 = (\gamma_{0,1}, \ldots, \gamma_{0,p_x})'$ is a fixed $p_x \times 1$ vector. The bound in (18) reduces to

$$\rho(0, k) \leq \sigma^2 k + \sum_{i=1}^{p_x} \gamma_{0,i}^2 \lambda_i \left(1 - \frac{k}{p_x} \frac{\lambda_i}{w_r(\lambda_i - 1) + 1}\right). \tag{20}$$

The second term of (20) represents the forecast bias. This bias term is a function of the weight w_r, which equals $(p_x - 1)^{-1}(k-1)$ for random subset regression and $[(p_x+2)(p_x-1)]^{-1}[p_x(k+1)-2]$ for random projection regression. Since the data is standardized, the average eigenvalue $\bar{\lambda} = p_x^{-1} \sum_{i=1}^{p_x} \lambda_i = 1$. In settings with p_x and $k < p_x$ sufficiently large, the weight w_r is close to k/p_x for both random subspace methods, and $w_r \lambda_i \gg 1$ for each $\lambda_i \gg \bar{\lambda}$. Then, the term in parenthesis in (20) is approximately equal to $1 - k/(p_x w_r)$ and, since w_r is close to k/p_x, the bias term corresponding to eigenvalues $\lambda_i \gg \bar{\lambda}$ will be small.

In macroeconomic forecasting, it is generally assumed that the dependent variable is well explained by a small number of principal components associated with the largest eigenvalues of the covariance matrix of the predictors. This implies that the nonzero coefficients $\gamma_{0,i}$ are associated with eigenvalues that are large relative to the average eigenvalue. The previous discussion suggests that in this case the nonzero bias terms in (20) are small, which means that the forecast variance can be reduced by setting $k < p_x$ without inducing a large bias in the forecast.

Furthermore, random subset regression is expected to outperform random projection regression in macroeconomic forecasting. The bound for the two random subspace methods differs only through the weight w_r. The difference between the weight w_r under random projection regression and random subset regression is given by $w_p - w_s = 2[(p_x + 2)(p_x - 1)]^{-1}(p_x - k) > 0$. This implies that a nonzero coefficient $\gamma_{0,i}$ associated with an eigenvalue $\lambda_i > 1$, induces less bias when using random subset regression. In contrast, if $\gamma_{0,i}$ is nonzero and $\lambda_i < 1$, the bias is smaller when using random projection regression. The effect of the weight difference is small when the eigenvalues are approximately equal to one.

4. Numerical evaluation

To verify whether the intuition gained from the bounds is reflected in the asymptotic MSFE, we numerically evaluate the asymptotic MSFE given by (14) in the setting of Section 3.3. We consider the principal component regression model in (19), where a small number of principal components drives the dependent variable. These principal components are extracted from the correlation matrix of the 130 indicators in the FRED-MD data used in Section 5. The number of nonzero coefficients is varied over $p_f = \{2, 5\}$. As Elliott et al. (2015), we consider local-to-zero coefficients $\gamma_i = \gamma_{0,i}/\sqrt{T}$, where $\gamma_{0,i} = c\lambda_i^{-1/2}$ for $i = 1, \ldots, p_f$. The multiplication with $\lambda_i^{-1/2}$ ensures that each principal component contributes equally to the signal. The parameter c determines the signal strength $\eta = \beta_{x,0}'\Sigma_x\beta_{x,0} = \gamma_0'\Lambda\gamma_0 = c^2 \cdot p_f$. We vary $\eta = \{p_x, p_f\}$, corresponding to a strong and a weak signal. The error variance $\sigma^2 = 1$.

For each setting, we calculate the bound for random subset regression and random projection regression in (20). Alongside the bounds we calculate the asymptotic MSFE using simulation. The proof of Lemma 7 in Appendix A.6 shows that the asymptotic MSFE in (14) can be rewritten as

$$\rho(0, k) = \sigma^2 \mathrm{tr}(\mathrm{E}_R[V]^2) + \gamma_0' H' \Sigma_x^{1/2} \mathrm{E}_R[U]^2 \Sigma_x^{1/2} H \gamma_0, \tag{21}$$

where $V = \Sigma_x^{1/2} R(R'\Sigma_x R)^{-1} R' \Sigma_x^{1/2}$, $U = I_{p_x} - V$, and H is the matrix of eigenvectors such that $\Sigma_x = H\Lambda H'$. We evaluate the expectations by averaging over $N = 10{,}000$ draws of the random matrix R. Note that the random subspace methods are applied to the original predictors instead of transforming these to the principal components, see the discussion following (19).

As a benchmark, we include the MSFE under principal component regression. This method extracts the principal components from Σ_x and then estimates (19) using the first k principal components. The MSFE equals

$$\rho^{PCR}(0, k) = \sigma^2 k + \mathrm{I}[k < p_f] \sum_{i=k+1}^{p_f} \gamma_{0,i}^2 \lambda_i, \tag{22}$$

with $\mathrm{I}[\cdot]$ an indicator function that is equal to one if its argument is true and zero otherwise. Note that in this set-up, there is no estimation uncertainty in the principal components, as there would be in a factor model in finite sample.

Fig. 1 depicts the bounds for the random subspace methods for different values of the subspace dimension k, together with the asymptotic MSFE, relative to that of OLS using all p_x predictors. The number of principal components is set to $p_f = 2$ in the upper panels, and $p_f = 5$ in the lower panels. We also plot the exact asymptotic MSFE achieved by principal component regression.

In the left panels, the signal is strong. The bounds on the asymptotic MSFE indicate that random subset regression should outperform random projection regression. This is also reflected in the asymptotic MSFE. The relative MSFE of RS to RP, both evaluated at the subspace dimension that yields the lowest MSFE, equals 0.865 when $p_f = 2$, and 0.882 when $p_f = 5$. The optimal principal component regression model selects the correct number of principal components, and hence, achieves an asymptotic MSFE equal to p_f. In case of a strong signal, we see that this method is preferred over the random subspace methods.

In the right panels of Fig. 1, we show the results for a weak signal. The reduction in signal lowers the bias incurred by using a small subspace dimension, and consequently we find a lower optimal subspace dimension. The difference between the random subspace methods is less apparent in this setting. The relative MSFE of RS to RP, both evaluated at the subspace dimension that yields the lowest MSFE, equals 0.968 when $p_f = 2$, and 0.985 when $p_f = 5$. The derived bounds are slightly above the asymptotic MSFE obtained by principal component regression. However, both random subspace methods achieve a lower asymptotic MSFE for each choice of k.

5. Empirical application

5.1. Data and methods

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), housing (3), consumption, orders, and inventories (4), money and credit (5), interest rate and exchange rates (6), prices (7), and stock market (8). The data and transformations are described in detail by McCracken and Ng (2016). We find a small number of missing values, which are recursively replaced by the value in the previous time period of that variable.

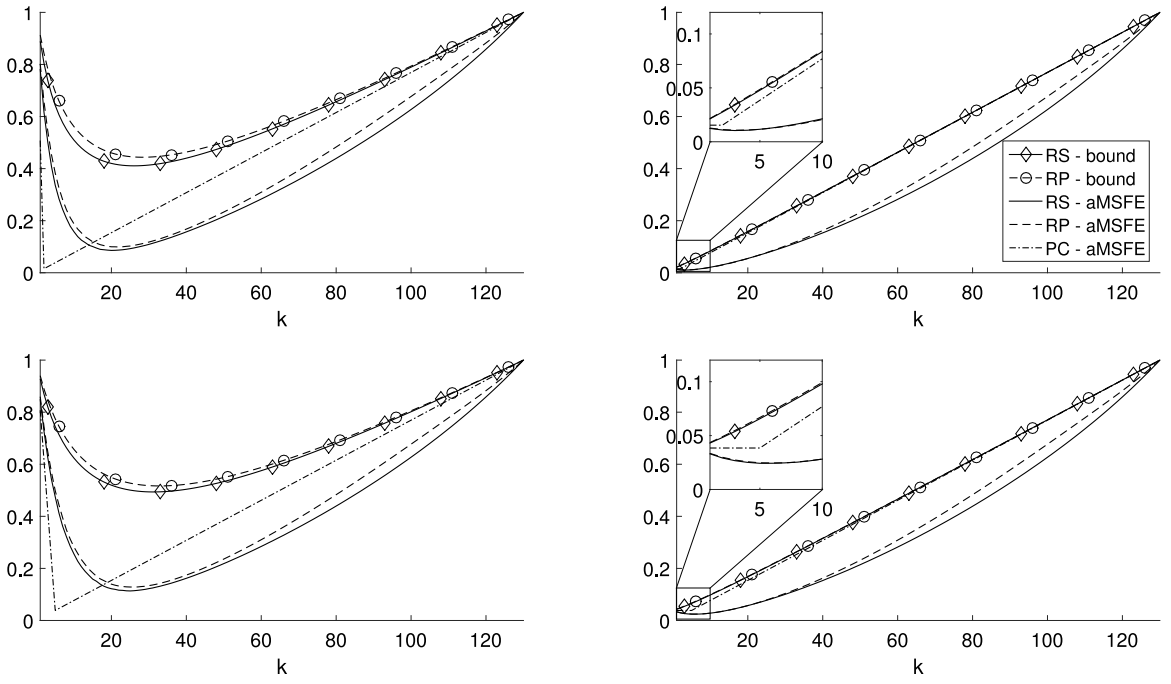

Fig. 1. Upper bounds asymptotic MSFE random subspace methods. *Note*: The figure shows the asymptotic MSFE in (21) as a function of the subspace dimension k for random subset regression (RS, solid line) and random projection regression (RP, dashed line) relative to (4) with $\sigma^2 = 1$, $p_w = 0$ and $p_x = 130$. The bounds in (20) are shown with markers. The dash-dotted line shows the asymptotic MSFE for principal component regression (PC) in (22). In the upper panel, the number of relevant principal components $p_f = 2$, the lower panel takes $p_f = 5$. The signal strength equals p_x on the left, and p_f on the right.

We use an expanding window to produce 420 forecasts, from January 1980 to December 2014. As Stock and Watson (2012), we use each of the 130 series as the dependent variable to be forecast. We follow Bai and Ng (2008) by allowing six lags of the dependent variable and evaluating the forecast performance relative to an AR(4) model. The dependent variable y_{t+1} in (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 lags of y_{t+1}, and all 129 remaining variables in the database. Including lags of x_t yields qualitatively similar results. We standardize the predictors in each estimation window.

We apply dimension reduction or regularization to the predictors in x_t using six different methods: random projection regression (RP), random subset regression (RS), principal component regression (PC), partial least squares (PL), ridge regression (RI), and lasso (LA). The subspace dimension, number of factors, or the penalty parameters of these methods are recursively selected on past predictive performance, using a burn-in period of 60 observations. The random subspace methods average $N = 1{,}000$ forecasts to obtain one prediction, and we vary the subspace dimension k over $\{0, \ldots, 100\}$.

Principal component regression and partial least squares achieve dimension reduction by using a small number of factors as predictors. We implement principal component regression by estimating k factors of dimension $(T+1) \times 1$ by applying principal component analysis to the standardized matrix of predictors $X = [x_0, \ldots, x_T]'$. This follows the estimation procedure outlined by Stock and Watson (2002b). To implement partial least squares, we follow the static approach of Groen and Kapetanios (2016). Denote the factors from one of the two methods by $\hat{f}_i$. We then generate the forecast as $\hat{y}_{T+1} = w_T'\hat{\beta}_w^F + \sum_{i=1}^k \hat{f}_{T,i}\hat{\beta}_{f,i}^F$, where $\{\hat{\beta}_w^F, \hat{\beta}_{f,i}^F\}$ are OLS estimates. The value of k is varied over $\{0, \ldots, 100\}$.

Ridge and lasso regression achieve dimension reduction via regularization of the estimated coefficients. The one-step-ahead forecasts equal $\hat{y}_{T+1} = w_T'\hat{\beta}_w^P + x_T'\hat{\beta}_x^P$, with

$$(\hat{\beta}_w^P, \hat{\beta}_x^P) = \arg\min_{b_w, b_x} \left(\frac{1}{2T} \sum_{t=0}^{T-1} (y_{t+1} - w_t'b_w - x_t'b_x)^2 + \lambda P(b_x) \right). \tag{23}$$

The penalty term $P(b_x) = (1/2)\sum_{i=1}^{p_x} b_{x,i}^2$ in case of ridge regression (Hoerl and Kennard, 1970) and $P(b_x) = \sum_{i=1}^{p} |b_{x,i}|$ for the lasso (Tibshirani, 1996). The penalty parameter λ varies over the grid $\log(\lambda) = \{-15, -14.7, \ldots, 15\}$ for ridge regression and over $\log(\lambda) = \{-30, -29.7, \ldots, 0\}$ for lasso. We use the coordinate descent algorithm of Friedman et al. (2010) to estimate the parameters.

Table 1
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 of the 130 available series for which a method listed in a row outperforms a method listed in a column, as well as all methods (last column). Ties occur if $k = 0$ is selected by both methods throughout the evaluation period.

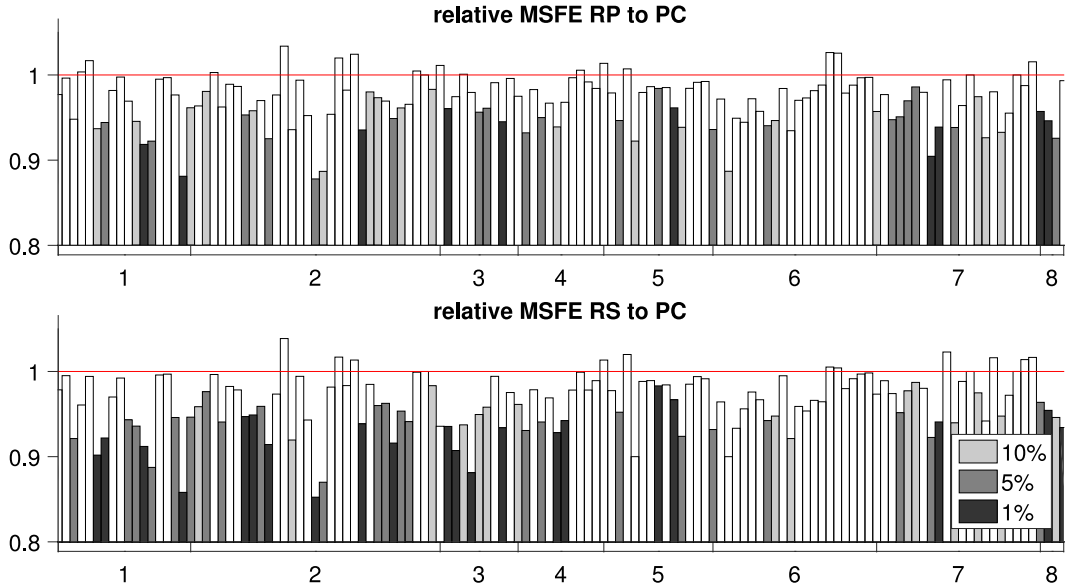

Fig. 2. FRED-MD: forecast accuracy relative to principal component regression. Note: this figure shows the MSFE of the forecasts for all series produced by random projection regression (upper panel) and random subset regression (lower panel), scaled by the MSFE of principal component regression. Series are grouped as in McCracken and Ng (2016). Values below one indicate the method is preferred over principal components. Colors indicate that the difference from one is significant at the 10% level (gray), 5% level (dark-gray), or 1% level (black), based on a two-sided Diebold–Mariano test.

5.2. Empirical results

The results show excellent forecast performance of the random subspace methods relative to the benchmark models. In line with the discussion in Section 3.3, we find a slightly higher accuracy for random subset regression relative to random projection regression. Table 1 shows the percentage wins in terms of MSFE in bivariate comparisons of the forecasting methods. Random subset regression achieves the best results for at least 66% of the series. A close competitor is random projection, which itself is also more accurate than all five benchmarks for a majority of the series. Ridge regression appears closest to random subset regression, although random subset regression is more accurate for over 66% of the series. The last column of Table 1 reports the percentage of the series for which a method outperforms all other methods. Random subset regression is more accurate than all other methods for 40% of the series.

Fig. 2 shows the MSFE of random projection regression (upper panel) and random subset regression (lower panel) relative to principal component regression. Values below one favor the random subspace method and the color of the bar indicates significance of the differences between the methods as determined by a Diebold and Mariano (1995) test. Random projection regression shows the largest improvements in category 7 (prices), and random subset regression in category 1 (output and income) and 2 (labor market).

We analyze industrial production (INDP) and inflation (CPI), corresponding to the FRED-MD mnemonics INDPRO and CPIAUCSL respectively, in more detail. For industrial production $y_t = \log(\text{INDPRO}_t/\text{INDPRO}_{t-1})$ and for inflation $y_t = \log(\text{CPIAUCSL}_t/\text{CPIAUCSL}_{t-1}) - \log(\text{CPIAUCSL}_{t-1}/\text{CPIAUCSL}_{t-2})$. Table 2 shows the MSFE relative to the AR(4) model for all models under consideration with the relevant tuning parameters selected based on past predictive performance. Random subset regression is most accurate for INDP and random projection regression for CPI.

Table 2
FRED-MD: relative MSFE on industrial production and inflation.

	RP	RS	PC	PL	RI	LA
INDP	0.843	0.820	0.890	0.898	0.844	0.826
CPI	0.870	0.888	0.962	0.872	0.901	0.897

Note: this table shows the MSFE relative to the AR(4) model for Industrial Production (INDP) and the Consumer Price Index (CPI) for different methods. The tuning parameters are selected based on past predictive performance.

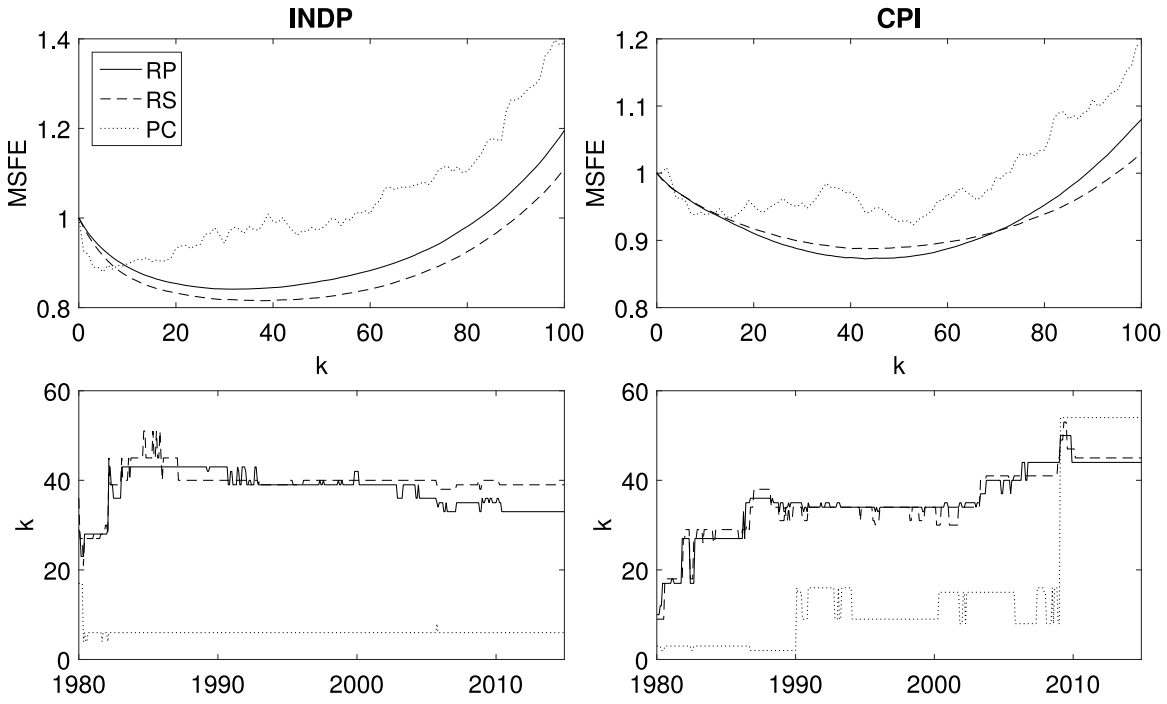

Fig. 3. FRED-MD: forecast accuracy and subspace dimension selection. Note: the upper panels show the MSFE for different values of the subspace dimension k relative to an AR(4) model. The lower panels show the recursive selection of the subspace dimension k. The different lines correspond to random projection (RP, solid), random subset (RS, dashed), and principal component regression (PC, dotted). The left panels correspond to industrial production (INDP) and the right panels to inflation (CPI).

In Fig. 3, we analyze how the selected subspace dimension affects forecast accuracy. The upper panels show how the MSFE depends on the subspace dimension k if the same dimension is selected throughout the forecast period. The forecast performance appears robust to the choice of the subspace dimension, with similar accuracy for values of k around the optimal choice.

We also show the relation between the MSFE and the number of principal components used by principal component regression. This shows that for industrial production, principal component regression optimally uses five components, while for inflation, it uses 53 components. The theoretical discussion in Section 3.3 then suggests this favors random subset regression on industrial production, while random projection is favored for inflation. This is indeed observed in the upper panels of Fig. 3.

The lower panels of Fig. 3 show the recursively selected subspace dimensions over time. The selected subspace dimension is relatively stable for the random subspace methods and, with the exception of the initial periods for CPI, close to the ex-post optimal value of k. Fig. 3 also shows the recursively selected number of principal components. For inflation, this number increases drastically around 2009. The large differences between the selected random subspace dimension and the selected number of principal components emphasize that these have a markedly different effect on the MSFE.

Finally, Fig. 4 shows the estimated coefficients used by both random subspace methods. We see that industrial production is related to variables from all indicator groups, while the inflation rate is best explained by indicators for money and credit (5) and prices (7). Although some differences in the estimated coefficients are observed, we find a correlation coefficient between the coefficients equal to 0.77 for industrial production and 0.82 for inflation.

6. Conclusion

Random subspace methods reduce the dimension of the predictor set by selecting subsets of the predictors, or by weighting the predictors using random weights. This paper derives bounds on the asymptotic mean squared forecast error

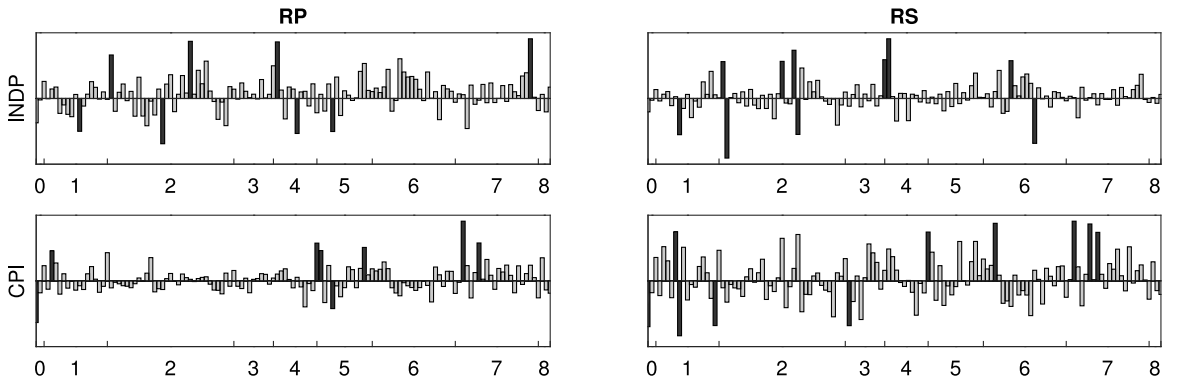

Fig. 4. FRED-MD: relative weight predictors in random subspace methods. Note: this figure shows the average coefficients of the predictors in x_t in random projection regression (RP) in the left column and random subset regression (RS) in the right column, estimated by $\mathrm{E}_R\left[R\hat{\beta}_{x,R}\right]$ for the optimal subspace dimension in the last estimation sample. Series are grouped as in McCracken and Ng (2016). The 'zero' group represents the first four lags of the dependent variable. The first row corresponds to industrial production (INDP) and the second to inflation (CPI). Dark colored bins indicate coefficients which differ two standard deviations from the average over all coefficients.

of these methods. The bounds show that these methods are expected to work well in macroeconomic forecasting, where the factors that explain most of the variation in the predictors drive the dependent variable. A numerical study and an empirical exercise support these findings.

Acknowledgments

We would like to thank Andreas Pick, Richard Paap, Michel van der Wel, Artūras Juodis, and Paul Bekker for helpful discussions. We also thank the Co-Editor, Associate Editor, and the referees for helpful comments. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix. Proofs

A.1. Independence between predictor and estimation error

Lemma 4. *Define $\hat{\beta}$ as in (3). For the model in (1) under Assumptions 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

$$\sqrt{T}(\hat{\beta}-\beta)=\left(\frac{1}{T}\sum_{t=0}^{T-1}z_t z_t'\right)^{-1}\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1}z_t\varepsilon_{t+1}. \tag{24}$$

We split $\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1}z_t\varepsilon_{t+1}$ into a part that is independent of z_T and one that is dependent of z_T, but negligible as $T\to\infty$.

$$\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1}z_t\varepsilon_{t+1}=\sqrt{\frac{T_1}{T}}\frac{1}{\sqrt{T_1}}\sum_{t=0}^{T_1}z_t\varepsilon_{t+1}+\sqrt{\frac{T-T_1}{T}}\frac{1}{\sqrt{T-T_1}}\sum_{t=T_1+1}^{T-1}z_t\varepsilon_{t+1}. \tag{25}$$

The following applies to all elements of the vectors in (25). By Assumption A4 and Chebyshev's inequality $\mathrm{P}(|z_{it}\varepsilon_{t+1}|\geq T^{\frac{1}{4}})\leq T^{-\frac{1}{2}}\Delta$. Using Bonferroni's inequality, we then have $\mathrm{P}(\max_{t=T_1+1,\ldots,T-1}|z_{it}\varepsilon_{t+1}|\geq T^{\frac{1}{4}})\leq T^{\frac{1}{2}-\alpha}\Delta$. Setting $\alpha>1/2$, we have that almost surely

$$\sqrt{\frac{T-T_1}{T}}\frac{1}{\sqrt{T-T_1}}\sum_{t=T_1+1}^{T-1}z_{it}\varepsilon_{t+1}\leq T^{-\frac{1}{2}+\frac{1}{4}+1-\alpha}. \tag{26}$$

Choosing $\alpha>3/4$,

$$\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1}z_{it}\varepsilon_{t+1}=\sqrt{\frac{T_1}{T}}\frac{1}{\sqrt{T_1}}\sum_{t=0}^{T_1}z_{it}\varepsilon_{t+1}+o_p(1). \tag{27}$$

Since under Assumptions A1–A7 a central limit theorem yields $\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1} z_t\varepsilon_{t+1} \sim N(0, \Sigma_z)$, 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 $T \to \infty$. By the same argument 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

The asymptotic MSFE is given by

$$\begin{aligned}\rho(p_w, k) &= \lim_{T\to\infty} T\mathrm{E}\left[\left(z_T'\beta - z_T'\mathrm{E}_R\left[S_R\hat{\beta}_R\right]\right)^2\right] \\ &= \lim_{T\to\infty} T\mathrm{E}\left[\mathrm{tr}\left\{z_T z_T'(\beta - \mathrm{E}_R[A_R]\hat{\beta})(\beta - \mathrm{E}_R[A_R]\hat{\beta})'\right\}\right],\end{aligned} \tag{28}$$

where we define $A_R \equiv S_R(S_R'Z'ZS_R)^{-1}S_R'Z'Z$. We now invoke the asymptotic independence of z_T and $\hat{\beta}$ established in Lemma 4, and use that $\mathrm{E}[z_T z_T'] = \Sigma_z$ to obtain

$$\begin{aligned}\rho(p_w, k) &= \lim_{T\to\infty} \mathrm{E}\left[T(\beta - \mathrm{E}_R[A_R]\hat{\beta})'\Sigma_z(\beta - \mathrm{E}_R[A_R]\hat{\beta})\right] \\ &= \lim_{T\to\infty} \mathrm{E}\left[(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])'Z'Z(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])\right],\end{aligned} \tag{29}$$

where we use $\mathrm{plim}_{T\to\infty} \frac{1}{T}Z'Z = \Sigma_z$ from (12).

Define the matrices $W = (w_0, \ldots, w_{T-1})'$ and $X = (x_0, \ldots, x_{T-1})'$, and the projection matrices $P_W = W(W'W)^{-1}W'$ and $M_W = I - P_W$. Write $\hat{\beta}_R = (\hat{\beta}_{w,R}', \hat{\beta}_{x,R}')'$. By the Frisch-Waugh-Lovell theorem,

$$\hat{\beta}_{x,R} = (R'X'M_WXR)^{-1}R'X'M_Wy, \quad \hat{\beta}_{w,R} = (W'W)^{-1}W'(y - XR\hat{\beta}_{x,R}) \tag{30}$$

Using that $Z = [W, X]$ and the definition of S_R provided in (8), we have

$$\begin{aligned}&(\beta - \mathrm{E}_R[S_R\hat{\beta}_R])'Z'Z(\beta - \mathrm{E}_R[S_R\hat{\beta}_R]) \\ &\quad = \varepsilon'P_W\varepsilon + (\beta_x - \mathrm{E}_R[R\hat{\beta}_{x,R}])'X'M_WX(\beta_x - \mathrm{E}_R[R\hat{\beta}_{x,R}]).\end{aligned} \tag{31}$$

The asymptotic MSFE then decomposes as

$$\rho(p_w, k) = \sigma^2 p_w + \lim_{T\to\infty} \mathrm{E}\left[(\beta_x - \mathrm{E}_R[R\hat{\beta}_{x,R}])'X'M_WX(\beta_x - \mathrm{E}_R[R\hat{\beta}_{x,R}])\right]. \tag{32}$$

By Jensen's inequality, this can be bounded by

$$\rho(p_w, k) \leq \sigma^2 p_w + \lim_{T\to\infty} \mathrm{E}^*\left[(\beta_x - R\hat{\beta}_{x,R})'X'M_WX(\beta_x - R\hat{\beta}_{x,R})\right], \tag{33}$$

where $\mathrm{E}^*[\cdot] = \mathrm{E}[\mathrm{E}_R[\cdot]]$.

The second term on the right-hand side of (33) can be rewritten as

$$\begin{aligned}&\lim_{T\to\infty} \mathrm{E}^*\left[(y - \varepsilon - XR\hat{\beta}_{x,R})'M_W(y - \varepsilon - XR\hat{\beta}_{x,R})\right] = \\ &\lim_{T\to\infty} \mathrm{E}^*\left[\varepsilon'M_W\varepsilon + (y - XR\hat{\beta}_{x,R})'M_W(y - XR\hat{\beta}_{x,R}) - 2\varepsilon'M_W(y - XR\hat{\beta}_{x,R})\right].\end{aligned} \tag{34}$$

To proceed, note that $\hat{\beta}_{x,R} = \arg\min_u (y - XRu)'M_W(y - XRu)$. Therefore, for an arbitrary $p_x \times 1$ vector v, (34) is upper bounded by

$$\begin{aligned}&\lim_{T\to\infty} \mathrm{E}^*\left[\varepsilon'M_W\varepsilon + (y - XRv)'M_W(y - XRv) - 2\varepsilon'M_W(y - XR\hat{\beta}_{x,R})\right] = \\ &\lim_{T\to\infty} \mathrm{E}^*\left[(\beta_x - Rv)'X'M_WX(\beta_x - Rv) + 2\varepsilon'M_W(XR\hat{\beta}_{x,R} - XRv)\right].\end{aligned} \tag{35}$$

Since we are free to choose v, we choose

$$v = \frac{1}{\sqrt{T}}R'u + (R'X'M_WXR)^{-1}R'X'M_W\varepsilon, \tag{36}$$

with u a fixed $p_x \times 1$ vector.

Substituting (36) into (35), we see that the cross terms from the first part of (35) cancel against the second part of (35). In total, we have

$$\lim_{T\to\infty} \mathrm{E}^* \left[\varepsilon' P_{M_W XR} \varepsilon + (\beta_{x,0} - RR'u)' \frac{1}{T} X' M_W X (\beta_{x,0} - RR'u) \right]. \tag{37}$$

where $P_{M_W XR} = M_W XR(R'X'M_W XR)^{-1} R'X'M_W$.

From (12), we know that $\frac{1}{\sigma^2}\varepsilon' P_{M_W XR}\varepsilon \xrightarrow{(d)} \chi^2(k)$ and $\Sigma = \mathrm{plim}_{T\to\infty} T^{-1} X'M_W X$. Then,

$$\rho(p_w, k) \leq \sigma^2(p_w + k) + \mathrm{E}_R \left[(\beta_{x,0} - RR'u)' \Sigma (\beta_{x,0} - RR'u) \right]. \tag{38}$$

The bound in (38) is valid for any choice of u. After taking the expectation with respect to R, we follow the approach by Thanei et al. (2017) for the in-sample MSE and minimize the bound with respect to u. Together with the fact that $\mathrm{E}_R[RR'] = (k/p_x) I_{p_x}$, this yields

$$\rho(p_w, k) \leq \sigma^2(p_w + k) + \beta_{x,0}' \Sigma \beta_{x,0} - \beta_{x,0}' \Sigma \left(\frac{p_x}{k} \mathrm{E}_R \left[RR' \Sigma RR' \right] \frac{p_x}{k} \right)^{-1} \Sigma \beta_{x,0}. \tag{39}$$

A.3. Proof of Lemma 1

Let $R \in \mathbb{R}^{p_x \times k}$ be a random selection matrix and Σ a positive definite matrix. 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 for $i \neq j$

$$[RR'\Sigma RR']_{ij} = \begin{cases} [\Sigma]_{ij} & \text{if } [RR']_{ii}[RR']_{jj} = 1, \\ 0 & \text{if } [RR']_{ii}[RR']_{jj} = 0. \end{cases} \tag{40}$$

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

$$\mathrm{E}_R \left[[RR'\Sigma RR']_{ii} \right] = \frac{k}{p_x} [\Sigma]_{ii}, \qquad \mathrm{E}_R \left[[RR'\Sigma RR']_{ij} \right] = \frac{k}{p_x} \frac{k-1}{p_x - 1} [\Sigma]_{ij}. \tag{41}$$

Then,

$$\mathrm{E}_R[RR'\Sigma RR'] = \frac{k}{p_x} D_\Sigma + \frac{k}{p_x} \frac{k-1}{p_x - 1} (\Sigma - D_\Sigma) = \frac{k}{p_x} \left(\frac{k-1}{p_x - 1} \Sigma + \frac{p_x - k}{p_x - 1} D_\Sigma \right) \quad .$$

A.4. Proof of Lemma 2

Decompose $R = QP$ with $Q = R(R'R)^{-1/2}$ and $P = (R'R)^{1/2}$. We have

$$\begin{aligned} S_R \hat{\beta}_R &= \begin{pmatrix} I_{p_w} & O \\ O & R \end{pmatrix} \begin{pmatrix} W'W & W'XR \\ R'X'W & R'X'XR \end{pmatrix}^{-1} \begin{pmatrix} W' \\ R'X' \end{pmatrix} y \\ &= \begin{pmatrix} (W'W)^{-1}W' - (W'W)^{-1}W'XH_R X'M_W \\ H_R X'M_W \end{pmatrix} y, \end{aligned} \tag{42}$$

where $H_R = R(R'X'M_W XR)^{-1}R'$. Using now that $R = QP$ with P a $k \times k$ invertible matrix, we see that $H_R = Q(Q'X'M_W XQ)^{-1}Q'$. Hence, $S_R \hat{\beta}_R = S_Q \hat{\beta}_Q$.

A.5. Proof of Lemma 3

Lemma 3 follows from Theorem 1, Lemma 2, and an analytic expression for the expectation appearing in Theorem 1 that is provided here. First, let R be a $p_x \times k$ matrix with independent standard normal entries. Consider again the decomposition $R = QP$, where $Q = R(R'R)^{-1/2}$ and $P = (R'R)^{1/2}$. For any orthogonal $p_x \times p_x$ matrix H_{p_x}, we have that

$$H_{p_x} R (R' H_{p_x}' H_{p_x} R)^{-1/2} = H_{p_x} Q. \tag{43}$$

Also, $H_{p_x} R$ has the same distribution as R by orthogonal invariance of the matrix variate normal distribution with identity covariance matrices. Combining this with (43) shows that $H_{p_x} Q$ has the same distribution as Q.

Use the invariance property together with the eigenvalue decomposition of $\Sigma = H \Lambda H'$, where H is a $p_x \times p_x$ orthogonal matrix, to rewrite

$$\begin{aligned} \mathrm{E}_Q[QQ'\Sigma QQ'] &= \mathrm{E}_Q[QQ'H\Lambda H'QQ'] \\ &= \mathrm{E}_Q[HH'QQ'H\Lambda H'QQ'HH'] \\ &= H \mathrm{E}_Q[QQ'\Lambda QQ']H'. \end{aligned} \tag{44}$$

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

$$\begin{aligned} m_{ii} &= \lambda_i \left(\sum_{j=1}^{k} q_{ij}^4 + \sum_{j \neq j'} q_{ij}^2 q_{ij'}^2 \right) + \sum_{l \neq i} \lambda_l \left(\sum_{j=1}^{k} q_{ij}^2 q_{lj}^2 + \sum_{j \neq j'} q_{ij} q_{lj} q_{ij'} q_{lj'} \right), \\ m_{ii'} &= \lambda_i \left(\sum_{j=1}^{k} q_{ij}^3 q_{i'j} + \sum_{j \neq j'} q_{ij}^2 q_{ij'} q_{i'j'} \right) + \lambda_{i'} \left(\sum_{j=1}^{k} q_{i'j}^3 q_{ij} + \sum_{j \neq j'} q_{i'j}^2 q_{i'j'} q_{ij'} \right) \\ &\quad + \sum_{l \neq \{i,i'\}} \lambda_l \left(\sum_{j=1}^{k} q_{ij} q_{i'j} q_{lj}^2 + \sum_{j \neq j'} q_{ij} q_{i'j'} q_{lj} q_{lj'} \right). \end{aligned} \tag{45}$$

From (45) it follows that we need the (mixed) moments of q_{ij} up to fourth order. The required non-zero moments are provided in the following lemma.

Lemma 5. *Let Q be a random orthogonal matrix. Denote the i, jth entry of Q by q_{ij}, and let i' and j' be indices such that $i \neq i'$, $j \neq j'$. The non-zero (mixed) moments of q_{ij} up to fourth-order are*

$$\begin{aligned} &E\left[q_{ij}^2\right] = \frac{1}{p_x}, \quad E\left[q_{ij}^4\right] = \frac{3}{p_x(p_x+2)}, \quad E\left[q_{ij}^2 q_{ij'}^2\right] = E\left[q_{ij}^2 q_{i'j}^2\right] = \frac{1}{p_x(p_x+2)}, \\ &E\left[q_{ij}^2 q_{i'j'}^2\right] = \frac{p_x+1}{p_x(p_x-1)(p_x+2)}, \quad E\left[q_{ij} q_{ij'} q_{i'j} q_{i'j'}\right] = \frac{-1}{p_x(p_x-1)(p_x+2)}. \end{aligned} \tag{46}$$

Proof. These expressions can be found using orthogonal Weingarten functions, see Example 2.1 and Section VII in Collins and Matsumoto (2009). Explicit calculations are available upon request.

Substituting the moments in Lemma 5 in the expectation of (45), we have

$$\begin{aligned} \mathrm{E}[m_{ii}] &= \frac{k}{p_x} \left(\frac{2+k}{p_x+2} \lambda_i + \frac{p_x - k}{(p_x+2)(p_x-1)} \sum_{l \neq i} \lambda_l \right) \\ &= \frac{k}{p_x} \left(\frac{p_x(k+1)-2}{(p_x+2)(p_x-1)} \lambda_i + \frac{p_x - k}{(p_x+2)(p_x-1)} \sum_{l=1}^{p_x} \lambda_l \right), \quad \mathrm{E}[m_{ii'}] = 0. \end{aligned} \tag{47}$$

Substituting this expression in (44), we arrive at

$$\mathrm{E}_Q[QQ'\Sigma QQ'] = \frac{k}{p_x} \left(\frac{p_x(k+1)-2}{(p_x+2)(p_x-1)} \Sigma + \frac{(p_x-k)p_x}{(p_x+2)(p_x-1)} \frac{\mathrm{tr}(\Sigma)}{p_x} I_{p_x} \right). \tag{48}$$

A.6. Proof of Theorem 2

The following matrices play an important role in the proof,

$$V = \Sigma^{1/2} R (R'\Sigma R)^{-1} R' \Sigma^{1/2} \quad \text{and} \quad U = I_{p_x} - V, \tag{49}$$

where R denotes a random selection or projection matrix.

A.6.1. Preliminary lemma's

Lemma 6. *Define $\tau = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$, and denote by $A \preceq B$ that $B - A$ is positive semidefinite. The matrices defined in (49) satisfy*

$$\frac{1}{\tau} \frac{k}{p_x} I_{p_x} \preceq E_R[P] \preceq \tau \frac{k}{p_x} I_{p_x}, \qquad \frac{1}{\tau} \frac{p_x - k}{p_x} I_{p_x} \preceq E_R[U] \preceq \tau \frac{p_x - k}{p_x} I_{p_x}. \tag{50}$$

Proof. We present the proof for $\mathrm{E}_R[V]$. The proof for $\mathrm{E}_R[U]$ follows.

$$\begin{aligned} & \lambda_{\min}(\Sigma) I_{p_x} \preceq && \Sigma && \preceq \lambda_{\max}(\Sigma) I_{p_x}, \\ \Rightarrow \quad & \lambda_{\min}(\Sigma) R'R \preceq && R'\Sigma R && \preceq \lambda_{\max}(\Sigma) R'R, \\ \Rightarrow \quad & \lambda_{\max}^{-1}(\Sigma)(R'R)^{-1} \preceq && (R'\Sigma R)^{-1} && \preceq \lambda_{\min}^{-1}(\Sigma)(R'R)^{-1}, \\ \Rightarrow \quad & \lambda_{\max}^{-1}(\Sigma) R(R'R)^{-1}R' \preceq && R(R'\Sigma R)^{-1}R' && \preceq \lambda_{\min}^{-1}(\Sigma) R(R'R)^{-1}R', \end{aligned}$$

$$\begin{aligned}
\Rightarrow \quad \lambda_{\max}^{-1}(\Sigma)\tfrac{k}{p_x} I_p \preceq \quad & \mathrm{E}_R[R(R'\Sigma R)^{-1}R'] \quad \preceq \lambda_{\min}^{-1}(\Sigma)\tfrac{k}{p_x} I_{p_x}, \\
\Rightarrow \quad \lambda_{\max}^{-1}(\Sigma)\tfrac{k}{p_x} \Sigma \preceq \quad & \mathrm{E}_R[V] \quad \preceq \lambda_{\min}^{-1}(\Sigma)\tfrac{k}{p_x} \Sigma, \\
\Rightarrow \quad \tau^{-1}\tfrac{k}{p_x} I_{p_x} \preceq \quad & \mathrm{E}_R[V] \quad \preceq \tau\tfrac{k}{p_x} I_{p_x}. \quad \blacksquare
\end{aligned}$$

Lemma 7. *For both random subset regression and random projection regression, a lower bound on the asymptotic MSFE is*

$$\rho(p_w, k) \geq \sigma^2 p_w + \tau^{-2}\left[((p_x - k)/p_x)^2 \beta_{x,0}' \Sigma \beta_{x,0} + \sigma^2 k^2/p_x\right]. \tag{51}$$

Proof. Using (30), and the convergence in distribution given by (12), (32) can be rewritten to

$$\begin{aligned}
\rho(p_w, k) &= \sigma^2 p_w + \lim_{T\to\infty} \mathrm{E}\left[(\beta_x - \mathrm{E}_R[R\hat{\beta}_{x,R}])' X' M_W X (\beta_x - \mathrm{E}_R[R\hat{\beta}_{x,R}])\right] \\
&= \sigma^2 p_w + \lim_{T\to\infty} \mathrm{E}[\varepsilon' M_W X A_T X' M_W X A_T X' M_W \varepsilon] \\
&\quad + \lim_{T\to\infty} \beta_x'(I - X' M_W X A_T) X' M_W X (I - A_T X' M_W X)\beta_x,
\end{aligned} \tag{52}$$

where $A_T = \mathrm{E}_R[R(R'X'M_W XR)^{-1}R']$. Using that $\operatorname{plim}_{T\to\infty} \frac{1}{T} X' M_W X = \Sigma$, we have $\operatorname{plim}_{T\to\infty} TA_T = \mathrm{E}_R[R(R'\Sigma R)^{-1}R'] \equiv A$. Under Assumption 8 and with U as in (49), the final term of (52) then equals

$$\beta_{x,0}'(I - \Sigma A)\Sigma(I - A\Sigma)\beta_{x,0} = \beta_{x,0}'\Sigma^{1/2}\mathrm{E}_R[U]^2\Sigma^{1/2}\beta_{x,0}. \tag{53}$$

Under Assumptions A1–A7, $T^{-1/2}X'M_W\varepsilon \xrightarrow{(d)} N(0, \sigma^2\Sigma)$. Then, with V as in (49), the second term of (52) equals $\sigma^2 \mathrm{tr}\left(\Sigma^{1/2}A\Sigma A\Sigma^{1/2}\right) = \sigma^2 \mathrm{tr}\left(\mathrm{E}_R[V]^2\right)$. Then

$$\rho(p_w, k) = \sigma^2 p_w + \sigma^2 \mathrm{tr}(\mathrm{E}_R[V]^2) + \beta_{x,0}'\Sigma^{1/2}\mathrm{E}_R[U]^2\Sigma^{1/2}\beta_{x,0}. \tag{54}$$

Applying Lemma 6 to (54) yields the required result. $\blacksquare$

Lemma 8 (*Ahlswede and Winter, 2002, Theorem 19*). *Let $(X_1, \ldots, X_N)$ be a sequence of independent $p_x \times p_x$ symmetric positive definite matrices with $\lambda_{\max}(X_i) \leq 1$ almost surely. Let $S_N = \sum_{i=1}^N X_i$ and $\Omega = \sum_{i=1}^N \lambda_{\max}(E[X_i])$, then for all $\epsilon \in (0, 1)$*

$$P\left(\lambda_{\max}(S_N - E[S_N]) \geq \epsilon\Omega\right) \leq 2p \exp(-\epsilon^2\Omega/4). \tag{55}$$

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).

A.6.2. Main proof of Theorem 2

The calculations leading to the MSFE expression in (54) can be repeated under a finite average over realizations of R, instead of an expectation. Define $V_E = \mathrm{E}_R[V]$, $U_E = \mathrm{E}_R[U]$, $V_S = \frac{1}{N}\sum_{i=1}^N V_i$, and $U_S = \frac{1}{N}\sum_{i=1}^N U_i$ where i indexes a particular draw of the random matrix R. Using (54), the difference in the MSFE under a sum and expectation is

$$\begin{aligned}
\Delta &= \beta_{x,0}\Sigma^{1/2}(U_E^2 - U_S^2)\Sigma^{1/2}\beta_{x,0} + \sigma^2 \mathrm{tr}(V_E^2 - V_S^2) \\
&\leq 2\beta_{x,0}\Sigma^{1/2}(U_S - U_E)U_E\Sigma^{1/2}\beta_{x,0} + 2\sigma^2 \mathrm{tr}\left((V_S - V_E)V_E\right) \\
&\leq 2\lambda_{\max}(U_S - U_E)\lambda_{\max}(U_E)\beta_{x,0}'\Sigma\beta_{x,0} + 2\sigma^2 p_x \lambda_{\max}(V_S - V_E)\lambda_{\max}(V_E) \\
&\leq 2\lambda_{\max}(U_S - U_E)\tau\frac{p_x - k}{p_x}\beta_{x,0}'\Sigma\beta_{x,0} + 2\sigma^2 p_x \lambda_{\max}(V_S - V_E)\tau\frac{k}{p_x},
\end{aligned} \tag{56}$$

where the second line is obtained by writing $U_S = U_S - U_E + U_E$, and similar for V_S. The third line follows from Cauchy–Schwarz, and the final line uses Lemma 6.

We now apply Lemma 8 to bound $\lambda_{\max}(V_S - V_E)$ and $\lambda_{\max}(U_S - U_E)$. 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 $\lambda_{\max}(X_i) = 1$. Since all realizations of R_i are independent, we meet the conditions of Lemma 8. Set $\Omega = N\lambda_{\max}(V_E)$, such that

$$\mathrm{P}\left(\lambda_{\max}(V_S - V_E) \geq \epsilon\lambda_{\max}(V_E)\right) \leq 2p_x \exp\left(-\epsilon^2 N\lambda_{\max}(V_E)/4\right). \tag{57}$$

By Lemma 6, this implies that

$$\mathrm{P}\left(\lambda_{\max}(V_S - V_E) \geq \epsilon\tau\frac{k}{p_x}\right) \leq 2p_x \exp\left(-\frac{\epsilon^2}{4}N\tau^{-1}\frac{k}{p_x}\right). \tag{58}$$

The right-hand side of (58) needs to be close to zero, which requires that for some $\delta \in (0, 1)$

$$2p_x \exp\left(-\frac{\epsilon^2}{4}N\tau^{-1}\frac{k}{p_x}\right) \leq \delta. \tag{59}$$

To satisfy this inequality we need to choose the number of samples

$$N \geq \frac{4\tau}{\epsilon^2}\frac{p_x}{k}\log\left(\frac{2p_x}{\delta}\right). \tag{60}$$

For this number of draws, $\lambda_{\max}(V_S - V_E) < \epsilon\tau(k/p_x)$ with large probability.

Analogous calculations show that we need to choose the number of samples

$$N \geq \frac{4\tau}{\epsilon^2}\frac{p_x}{p_x - k}\log\left(\frac{2p_x}{\delta}\right), \tag{61}$$

such that $\lambda_{\max}(U_S - U_E) < \epsilon\tau(p_x - k)/p_x$ with large probability. We therefore need $N = O(p_x \log p_x)$ to satisfy both (60) and (61). In that case,

$$\Delta \leq 2\epsilon\tau^2\left[\left(\frac{p_x - k}{p_x}\right)^2 \beta_{x,0}\Sigma\beta_{x,0} + \sigma^2\frac{k^2}{p_x}\right] \leq \eta\rho(p_w, k), \tag{62}$$

where $\eta = 2\epsilon\tau^4$, and we used the lower bound on the MSFE provided in Lemma 7.

References

Achlioptas, D., 2003. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. J. Comput. System Sci. 66 (4), 671–687.

Ahlswede, R., Winter, A., 2002. Strong converse for identification via quantum channels. IEEE Trans. Inform. Theory 48 (3), 569–579.

Ailon, N., Chazelle, B., 2009. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. SIAM J. Comput. 39 (1), 302–322.

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

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

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Breiman, L., 1999. Pasting small votes for classification in large databases and on-line. Mach. Learn. 36 (1–2), 85–103.

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

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

Collins, B., Matsumoto, S., 2009. On some properties of orthogonal Weingarten functions. J. Math. Phys. 50 (11), 113516.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econom. Statist. 13 (3), 253–263.

Elliott, G., Gargano, A., Timmermann, A., 2013. Complete subset regressions. J. Econometrics 177 (2), 357–373.

Elliott, G., Gargano, A., Timmermann, A., 2015. Complete subset regressions with large-dimensional sets of predictors. J. Econom. Dynam. Control 54, 86–110.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1.

Frieze, A., Kannan, R., Vempala, S., 2004. Fast Monte-Carlo algorithms for finding low-rank approximations. J. Assoc. Comput. Mach. 51 (6), 1025–1041.

Groen, J.J., Kapetanios, G., 2016. Revisiting useful approaches to data-rich macroeconomic forecasting. Comput. Statist. Data Anal. 100, 221–239.

Guhaniyogi, R., Dunson, D.B., 2015. Bayesian compressed regression. J. Amer. Statist. Assoc. 110 (512), 1500–1514.

Hansen, B.E., 2008. Least-squares forecast averaging. J. Econometrics 146 (2), 342–350.

Hansen, B.E., 2010. Averaging estimators for autoregressions with a near unit root. J. Econometrics 158 (1), 142–155.

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

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

Inoue, A., Kilian, L., 2008. How useful is bagging in forecasting economic time series? A case study of US consumer price inflation. J. Amer. Statist. Assoc. 103 (482), 511–522.

Johnson, W.B., Lindenstrauss, J., 1984. Extensions of Lipschitz mappings into a Hilbert space. Contemp. Math. 26 (189–206), 1.

Kabán, A., 2014. New bounds on compressive linear least squares regression. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, PMLR, vol. 33. pp. 448–456.

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

Lu, Y., Dhillon, P., Foster, D.P., Ungar, L., 2013. Faster ridge regression via the subsampled randomized hadamard transform. In: Advances in Neural Information Processing Systems. pp. 369–377.

Ma, P., Mahoney, M.W., Yu, B., 2015. A statistical perspective on algorithmic leveraging. J. Mach. Learn. Res. 16 (1), 861–911.

Mahoney, M.W., Drineas, P., 2009. Cur matrix decompositions for improved data analysis. Proc. Natl. Acad. Sci. 106 (3), 697–702.

Maillard, O., Munos, R., 2009. Compressed least-squares regression. In: Advances in Neural Information Processing Systems, vol. 22. pp. 1213–1221.

McCracken, M.W., Ng, S., 2016. FRED-MD: A monthly database for macroeconomic research. J. Bus. Econom. Statist. 34 (4), 574–589.

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

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

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

Stock, J.H., Watson, M.W., 2002a. Forecasting using principal components from a large number of predictors. J. Amer. Statist. Assoc. 97 (460), 1167–1179.

Stock, J.H., Watson, M.W., 2002b. Macroeconomic forecasting using diffusion indexes. J. Bus. Econom. Statist. 20 (2), 147–162.

Stock, J.H., Watson, M.W., 2006. Forecasting with many predictors. Handbook Econ. Forecasting 1, 515–554.

Stock, J.H., Watson, M.W., 2012. Generalized shrinkage methods for forecasting using many predictors. J. Bus. Econom. Statist. 30 (4), 481–493.

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

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

White, H., 1984. Asymptotic Theory for Econometricians. Academic Press.

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