University of Groningen

Subspace Methods

Boot, Tom; Nibbering, Didier

Published in:
Macroeconomic Forecasting in the Era of Big Data

DOI:
[10.1007/978-3-030-31150-6_9](https://doi.org/10.1007/978-3-030-31150-6_9)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Boot, T., & Nibbering, D. (2020). Subspace Methods. In P. Fuleky (Ed.), *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice* (pp. 267-291). (Advanced Studies in Theoretical and Applied Econometrics; Vol. 52). Springer. https://doi.org/10.1007/978-3-030-31150-6_9

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Chapter 9

Subspace Methods

Tom Boot and Didier Nibbering

9.1 Introduction

With a limited number of observations and a large number of variables, dimension reduction is necessary to obtain accurate macroeconomic forecasts. This can be done by identifying a small set of variables that appear most relevant to the variable of interest by variable selection (Ng, 2013) or shrinkage methods as the lasso (Tibshirani, 1996). Alternatively, one can transform the high dimensional data set into a small number of factors that capture most of the available information, as in factor models by Stock and Watson (2002, 2006) and Bai and Ng (2006, 2008), among others. In both cases, the low measurement frequency of macroeconomic data and the weak forecast relations make it a challenging task to find an accurate low-dimensional representation of the high-dimensional data. In fact, the selection methods themselves can add considerable noise to the forecast, potentially offsetting some of their gains.

Instead of trying to find the most informative part of the data, subspace methods reduce the dimension of the data in a data-oblivious fashion. For example, they draw a small set of variables at random, based on which a forecast can be constructed. Repeating this many times, and combining the forecasts using a simple average, has been found to be highly effective when constructing macroeconomic forecasts. The exact way in which these methods achieve such high forecast accuracy is not

T. Boot (✉)

Department of Economics, Econometrics and Finance, University of Groningen, Groningen, The Netherlands

e-mail: t.boot@rug.nl

D. Nibbering

Department of Econometrics and Business Statistics, Monash University, Clayton, VIC, Australia

e-mail: didier.nibbering@monash.edu

© Springer Nature Switzerland AG 2020

P. Fuleky (ed.), *Macroeconomic Forecasting in the Era of Big Data*,

Advanced Studies in Theoretical and Applied Econometrics 52,

https://doi.org/10.1007/978-3-030-31150-6_9

267

yet fully understood. Upper bounds on the mean squared forecast error indicate that high correlations in the data limit the omitted variables bias, while the small subspace dimension greatly reduces the forecast variance.

In this chapter, we provide an introduction to four different subspace methods: complete subset regression (CSR), random subset regression (RS), random projection regression (RP), and compressed regression (CR). The first is introduced by Elliott, Gargano, and Timmermann (2013) who propose to average over individual forecasts constructed from all possible linear regression models with a fixed number of predictors. Since using all available subsets is infeasible in high-dimensional settings, Elliott, Gargano, and Timmermann (2015) take arbitrary subsets as an approximation. The latter approach is further analyzed by Boot and Nibbering (2019), who discuss two different approaches to construct a random subspace in a macroeconomic forecasting setting. Random subset regression uses randomly selected subsets of predictors to estimate many low-dimensional approximations to the original model. Random projection regression forms a low-dimensional subspace by averaging over predictors using random weights drawn from a standard normal distribution. Koop, Korobilis, and Pettenuzzo (2019) exploit the computational gains from using sparse random projection matrices in a Bayesian compressed regression.

While applications of random subspace methods to economic forecasting are very recent, the ideas underlying random subspace methods have been used in a number of applications in the statistics and machine learning literature. Bootstrap aggregation methods aggregate the forecasts of models fitted on subsets of the data to a final prediction. Breiman (1996) and Breiman (1999) use bootstrap aggregation to draw subsets of observations instead of predictors, a method known as bagging. Feature bagging is a form of bootstrap aggregation where one draws subsets of predictors instead of observations. Within computer science, the performance of decision trees is often improved by training trees using random subsets of the available observations or variables at each split point, and then averaging the predictions from each of these trees. Random subset regression is a simple form of feature bagging aiming to reduce the dimension of the set of predictors in a linear regression model. Random projection regression and compressed regression are part of a large machine learning literature on constructing a new set of predictors (Frieze, Kannan, and Vempala, 2004; Mahoney and Drineas, 2009). The justification for the use of random projections is often derived from the Johnson–Lindenstrauss lemma (Johnson and Lindenstrauss, 1984). In a linear regression model, Kabán (2014) shows that the results can be sharpened without invoking the Johnson–Lindenstrauss lemma.

This chapter is organized as follows: Sect. 9.2 establishes notation and sets the forecasting framework. Section 9.3 briefly discusses two fundamental ideas behind subspace methods: forecast combinations and factor models. Section 9.4 introduces the random subspace methods. Section 9.5 surveys the performance of the random subspace methods in empirical applications reported in the literature. Theoretical guarantees on the forecast performance of random subspace methods are reviewed in Sect. 9.6. Section 9.7 illustrates the use of these methods in two

empirical applications to macroeconomic forecasting with many predictors, and Sect. 9.8 concludes with recommendations for future research.

9.2 Notation

We denote by y_{t+1} the macroeconomic variable of interest. To forecast this variable, we make use of a set of p_w must-have predictors $\mathbf{w}_t$, such as a constant and lags of y_{t+1} , and a set of p_x possibly relevant predictors $\mathbf{x}_t$. The case of interest is where p_x is large, so that including all predictors in a single model leads to inefficient forecasts. When it is not clear which variables to include in $\mathbf{w}_t$, this set can be left empty. Both $\mathbf{x}_t$ and $\mathbf{w}_t$ can include lags of y_{t+1} .

We focus on the linear model

$$y_{t+1} = \mathbf{w}_t' \boldsymbol{\beta}_w + \mathbf{x}_t' \boldsymbol{\beta}_x + \varepsilon_{t+1}, \quad (9.1)$$

where the forecast error is denoted by ε_{t+1} and the time index t runs from $t = 0, \dots, T$. The predictors $\mathbf{z}_t = (\mathbf{w}_t', \mathbf{x}_t')'$, with $t = 0, \dots, T-1$, are used in the estimation of the $p \times 1$ parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_w', \boldsymbol{\beta}_x')'$, and $\mathbf{z}_T = (\mathbf{w}_T', \mathbf{x}_T')'$ is only used for the construction of point forecasts $\hat{y}_{T+1}$ for y_{T+1} . Theoretically, the linearity assumption is not too restrictive, as in many cases we can approximate a nonlinear function $f(\mathbf{x}_t, \boldsymbol{\beta})$ using a first order Taylor expansion.

9.3 Two Different Approaches to Macroeconomic Forecasting

9.3.1 Forecast Combinations

A consistent finding in macroeconomic forecasting is that averaging forecasts from multiple forecasting models increases prediction accuracy (Timmermann, 2006). When the forecasts from the individual models are unbiased, combining multiple forecasts lowers the variance, and hence the mean squared forecast error. When individual models are misspecified, averaging might lower the bias if the misspecification error cancels.

A linear forecast combination pools over N individual forecasts $\hat{y}_{T+1,i}$ to obtain a forecast $\hat{y}_{T+1}$ for y_{T+1} by

$$\hat{y}_{T+1} = \sum_{i=1}^N \omega_i \hat{y}_{T+1,i},$$

where ω_i is the weight on $\hat{y}_{T+1,i}$.

Ideally, different models are combined using a weighted average, with weights that reflect the reliability of each forecast. In practice, estimation of combination weights introduces additional variance in the forecast. This can offset a large part of the variance reduction by combination. It turns out that taking a simple average over different models is generally hard to beat (Claeskens, Magnus, Vasnev, and Wang, 2016; Elliott and Timmermann, 2013). This is especially true if all models perform roughly the same, in which case equal weighting is close to being optimal.

The literature on forecast combinations generally starts by assuming that a number of different forecasts are available, but is silent on where these forecasts actually come from. The subspace methods discussed in this paper generate a sequence of forecasts that can then be combined to yield an accurate forecast.

9.3.2 *Principal Component Analysis, Diffusion Indices, Factor Models*

Factor models reflect the idea that a small number of latent components drive most variation in macroeconomic series. We observe a vector of regressors x_t that satisfies

$$x_t = \Lambda' f_t + e_t,$$

where f_t is a $r \times 1$ vector, Λ is a $r \times p$ matrix of loadings, and e_t is an idiosyncratic component. The factors are driving the dependent variable according to

$$y_{t+1} = w_t' \beta_w + f_t' \beta_f + \varepsilon_{t+1}.$$

It is generally assumed that the factors are orthonormal, such that $\frac{1}{T} \sum_{t=1}^T f_t f_t' = I_r$. Taking a singular value decomposition of x_t , we have

$$x_t = U S v_t = \Lambda' f_t + e_t.$$

If the idiosyncratic component $e_t = 0$, the factors equal the left singular vectors (up to a rotation), and the model essentially applies a linear dimension reduction by replacing x_t as

$$x_t' \rightarrow x_t' U S^{-1} = f_t'. \quad (9.2)$$

It turns out that this is still a good strategy if $e_t \neq 0$, as the left singular vectors are consistent for the space spanned by the unknown factors Stock and Watson (2002). Forecasts based on (estimated) factors are generally precise, and often an interpretation of the factors can be derived from the loadings.

Random subspace methods apply a linear dimension reduction step as in (9.2), but instead of multiplication with a data dependent matrix that is explicitly designed

to find factors, it replaces US^{-1} by a random matrix $\mathbf{R}$. Intuitively, suppose that all predictors strongly correlate with a single factor that drives the dependent variable. Then including a randomly selected subset of predictors, where $\mathbf{R}$ is a random permutation matrix, or a set of randomly weighted averages of the predictors, where $\mathbf{R}$ contains independent standard normally distributed elements, will not lead to a large omitted variable bias. When the dimension of these sets is small, there can however be a substantial reduction in variance compared to the case where one includes all available predictors. The theoretical results later in this chapter also point in this direction. To avoid that the random selection/weighting induces additional variance itself, we can average over forecasts obtained under many different random selections/weights.

9.4 Subspace Methods

Subspace methods can be seen as a combination of the ideas outlined in the previous section. First, a linear dimension reduction step as in a factor model is performed to obtain a low-dimensional approximation to the data. That is, we post-multiply the regressors x_t with a $p_x \times k$ matrix $\mathbf{R}_i$. The subspace dimension $k < p_x$ equals the number of predictors in the subspace. A commonly used deterministic matrix $\mathbf{R}_i$ contains the principal component loadings corresponding to the k largest eigenvalues of the sample covariance matrix of the predictors. If the elements of $\mathbf{R}_i$ are drawn from a known probability distribution, then this opens up the possibility to generate many low-dimensional approximations to the data. Inspired by the forecast combination literature, the forecasts based on the individual approximations are then combined to improve the forecast accuracy.

Replacing the full regressor matrix with the regressors projected onto the subspace leads to the approximating model

$$y_{t+1} = \mathbf{w}'_t \boldsymbol{\beta}_{w,i} + \mathbf{x}'_t \mathbf{R}_i \boldsymbol{\beta}_{x,i} + u_{t+1}, \quad (9.3)$$

which can be rewritten to

$$y_{t+1} = \mathbf{z}'_t \mathbf{S}_i \boldsymbol{\beta}_i + u_{t+1}, \text{ with } \mathbf{S}_i = \begin{pmatrix} \mathbf{I}_{p_w} & \mathbf{O} \\ \mathbf{O} & \mathbf{R}_i \end{pmatrix},$$

with $\mathbf{z}_t$ as defined below (9.1). Instead of estimating $\boldsymbol{\beta}$ in the high-dimensional model, where all $p_w + p_x$ predictors are included, we estimate the $p_w + k$ parameters in $\boldsymbol{\beta}_i$ in this low-dimensional model.

Under the assumption that $T > (p_w + k)$, the least squares estimator of $\boldsymbol{\beta}_i$ is given by

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{S}'_i \mathbf{Z}' \mathbf{Z} \mathbf{S}_i)^{-1} \mathbf{S}'_i \mathbf{Z}' \mathbf{y}.$$

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

$$\hat{y}_{T+1,i} = \mathbf{z}'_T \mathbf{S}_i \hat{\boldsymbol{\beta}}_i.$$

Based on N different realizations of the dimension reduction matrix $\mathbf{R}_i$, N different forecasts are obtained. These forecasts are subsequently averaged to obtain the prediction

$$\hat{y}_{T+1} = \frac{1}{N} \sum_{i=1}^N \hat{y}_{T+1,i}.$$

As we see here, each forecast receives an equal weight.

We now discuss three subspace methods that all have recently been applied to economic data.

9.4.1 Complete Subset Regression

Complete subset regression (Elliott et al., 2013) is a subspace method in which the individual forecasts are constructed from all possible linear regression models with a fixed number of k predictors. The matrix $\mathbf{R}_i$ in (9.3) selects one of the N unique combinations of k predictors out of the p_x available predictors. The selection matrix $\mathbf{R}_i$ is data-independent and has $\binom{p_x}{k}$ possible realizations. The set of models following from all possible realizations of $\mathbf{R}_i$ is called a complete subset.

To construct a more formal definition of complete subset regression, define an index $l = 1, \dots, k$ with k the dimension of the subspace, and a scalar $c(l)$ such that $1 \leq c(l) \leq p_x$. Denote by $\mathbf{e}_{c(l)}$ a p_x -dimensional vector with its $c(l)$ -th entry equal to one, then complete subset regressions are based on the matrices $\mathbf{R}_i$ of the form

$$[\mathbf{e}_{c(1)}, \dots, \mathbf{e}_{c(k)}], \quad \mathbf{e}_{c(m)} \neq \mathbf{e}_{c(n)} \text{ if } m \neq n,$$

where $c(1), \dots, c(k)$ is one of the $N = \binom{p_x}{k}$ combinations of k predictors.

A special case of complete subset regressions are forecast combinations of univariate models for each available predictor. It is generally found however that using larger values of k leads to lower omitted variable bias at the expense of a relatively small increase in the variance.

Subspace Dimension

Optimal selection of the subspace dimension is a largely unexplored area. The value of k is the number regressors in each regression, but the variance of the averaged forecast generally does not scale linearly with k . For example, when

all regressors are orthonormal, the variance scales as $k \cdot \frac{k}{p}$ (see Sect. 9.6). An appropriate information criterion to select k is not yet available.

Elliott et al. (2013) show that under the assumption that y_{t+1} and $\mathbf{z}_t$ are independent and identically distributed, the value of k can be set to minimize the mean squared forecast error for given values of β . However, both the i.i.d. assumption and available information on the values in β do generally not hold in macroeconomic forecasting.

In practice, the subspace dimension can be recursively selected by evaluating the forecast accuracy of each dimension in the current information set and selecting the best performing k . Empirical results in Boot and Nibbering (2019) and Elliott et al. (2015) indicate that the performance is not heavily dependent on the precise value of k .

Weighting Schemes

Since there is little empirical evidence that alternative weighting schemes outperform equal weighted forecast combinations, random subspace methods use an equal weighted average. Since macroeconomic data typically has a small number of observations, estimating sophisticated weighting schemes is particularly difficult in this setting. Elliott et al. (2013) experimented with a weighting scheme that is based on the Bayesian Information Criterion (BIC). The weight of each model is proportional to the exponential of its BIC, which results in larger weights for models with a high likelihood value. They do not find structural improvements over using equal weights.

Limitations

When the number of available predictors is large, considering all combinations of predictors for a given k becomes infeasible. As a solution, we can average over forecasts from a smaller set of models. This smaller set of models can be selected by a stochastic search. Elliott et al. (2013, 2015) implement a Markov Chain and a Shotgun approach but find similar performance for randomly drawing a set of models. They draw with uniform probability a feasible number of models without replacement from the complete subspace regressions. Randomly drawing models is simple, fast, and assigns, similar to equal weighting, the same probability to each submodel. This brings us to a feasible alternative to complete subset regressions in macroeconomic forecasting settings.

9.4.2 Random Subset Regression

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 to reduce the forecast variance while maintaining most of the signal. Note that this method is slightly different from randomly drawing models from the set of complete subset regressions. Instead of drawing a set of models without replacement, random subset regression draws a set of predictors without replacement for each submodel.

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

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

The same formulation as for complete subset regression can be used to represent a general matrix $\mathbf{R}_i$. We have a p_x -dimensional vector $\mathbf{e}_{c(l)}$ with its $c(l)$ -th entry equal to one, where $l = 1, \dots, k$ is an index and $c(l)$ a scalar such that $1 \leq c(l) \leq p_x$. Then random subset regression is based on random matrices

$$[\mathbf{e}_{c(1)}, \dots, \mathbf{e}_{c(k)}], \quad \mathbf{e}_{c(m)} \neq \mathbf{e}_{c(n)} \text{ if } m \neq n.$$

The intuition why random subset regression is effective in macroeconomic forecasting is the following. One might generally worry that selecting predictors at random results in irrelevant predictors being present in each submodel. However, in macroeconomic data, the predictors are often highly correlated. This means that even if a predictor is irrelevant, it is most likely correlated with a predictor that is relevant. As such, even predictors that are irrelevant when all variables are included are relevant when only small sets of variables are included. With strong correlation present, the use of random subsets reduces the variance of the forecast by at least a factor $\frac{k}{p_x}$, while the increase in bias is only small.

9.4.3 Random Projection Regression

An alternative for using a random selection matrix is to use a random weighting matrix. Instead of selecting individual regressors, this gives a new set of predictors which are weighted averages of the original predictors. This can be easily implemented by drawing a matrix $\mathbf{R}_i$ of which each element is distributed independent

and identically following some probability distribution. The most common choice is

$$[\mathbf{R}_i]_{hj} \sim N(0, 1/\sqrt{k}), \quad (9.4)$$

with $h = 1, \dots, p_x$ and $j = 1, \dots, k$. Repeating the same procedure as for random subset regression, only now with the matrix $\mathbf{R}_i$ defined in (9.4) yields random projection regression.

Random projections are originally motivated by data sets that contain too many observations to allow a calculation of the covariance matrix $\frac{1}{T} \mathbf{X}' \mathbf{X}$ in a reasonable time. To reduce the dimension T , we can note that $E_R[\frac{1}{T} \mathbf{X}' \mathbf{R} \mathbf{R}' \mathbf{X}] = \frac{1}{T} \mathbf{X}' \mathbf{X}$. If $\frac{1}{T} \mathbf{X}' \mathbf{R} \mathbf{R}' \mathbf{X}$ is sufficiently close to its expectation, then we can calculate the covariance matrix accurately using the smaller matrix $\mathbf{X}' \mathbf{R}$. Such a result is generally shown through the Johnson–Lindenstrauss lemma (Achlioptas, 2003; Dasgupta and Gupta, 2003). If we use (9.4), this still requires the multiplication of $\mathbf{X}' \mathbf{R}$. Hence, subsequent work focused on showing the Johnson–Lindenstrauss lemma for matrices $\mathbf{R}$ that contain many zeros (Achlioptas, 2003; Li, Hastie, and Church, 2006).

Since we apply dimension reduction to the large number of predictors, instead of reducing a large number of observations, the above motivation for random projection does not align with our application here. The problem we aim to solve is more statistical in nature, with a large variance caused by the inclusion of many predictors, rather than computational. This direction has received substantially less attention. We discuss the relevant empirical literature in Sect. 9.5, and theoretical results in Sect. 9.6.

9.4.4 Compressed Regression

To speed up calculations, sparse random projection matrices have been proposed. The prime example is to choose

$$R_{ij} = \sqrt{s} \begin{cases} -1 & \text{with probability } \frac{1}{2s}, \\ 0 & \text{with probability } 1 - \frac{1}{s}, \\ +1 & \text{with probability } \frac{1}{2s}. \end{cases}$$

This specification was proposed by Achlioptas (2003), who shows that the Johnson–Lindenstrauss lemma applies to this matrix for $s = 3$. This was extended to $s = O(\sqrt{n})$ by Li et al. (2006). As s governs the sparsity of $\mathbf{R}$, setting s large can lead to significant computational gains, without loss in accuracy.

Guhaniyogi and Dunson (2015) use an alternative specification for the random matrix,

$$R_{ij} = \sqrt{s} \begin{cases} -1 & \text{with probability } \frac{1}{s^2}, \\ 0 & \text{with probability } 2(1 - \frac{1}{s})\frac{1}{s}, \\ +1 & \text{with probability } (1 - \frac{1}{s})^2, \end{cases}$$

where $1/s \sim U[0.1, 0.9]$, and subsequently the columns of R are orthonormalized using the Gram–Schmidt algorithm. The subspace dimension k is drawn uniformly over $[2 \log p_x, \min(n, p_x)]$. Guhaniyogi and Dunson (2015) apply the dimension reduction in a Bayesian framework, and use Bayesian model averaging to weight the forecasts. Under normal-inverse gamma priors, and a diffuse prior on the models, the weights can be analytically calculated. Furthermore, they provide a set of sufficient conditions for the predictive density under dimension reduction to converge to the true predictive density.

9.5 Empirical Applications of Subspace Methods

Subspace methods are relatively new to the economics literature. This section lists early applications in the field of macroeconomic forecasting, but also in microeconomics and finance. Since subspace methods have their roots in the machine learning, we also refer to some applications in this literature.

9.5.1 Macroeconomics

The first applications of subspace methods in the macroeconomic literature use random subset regression to forecast economic indicators with a large number of possible predictors. Elliott et al. (2015) forecast quarterly US unemployment, GDP growth, and inflation and find more accurate point forecasts by random subset regression than dynamic factor models or univariate regressions over multiple forecast horizons. Although the paper presents the subspace method as complete subset regressions, the methods randomly sample low-dimensional models to form a final forecast in practice.

Leroux, Kotchoni, and Stevanovic (2017) draw similar conclusions in an extensive forecasting exercise on monthly US data. The authors compare the forecast accuracy of a large number of different forecasting models on industrial production, employment, inflation, the stock market index, and exchange rates. They find that random subset regression shows often better performance in terms of mean squared forecast error than univariate models, factor-augmented regressions, dynamic factor models, and standard forecast combinations for industrial production and employ-

ment growth. The forecast accuracy for other series does not seem to benefit from any data-rich model.

Garcia, Medeiros, and Vasconcelos (2017) also confirm the strong performance of random subset regression in a real-time forecasting exercise for Brazilian inflation. They use several lasso-type models, target factors, random forests, Bayesian vector autoregressions, standard time series models, random subset regression, and expert forecasts to produce forecasts for twelve different forecast horizons. Apart from the first two horizons, complete subset regression dominates all other forecasting methods on mean squared forecasting error. The same holds for density forecasts based on bootstrap resampling and for forecasts of accumulated inflation.

Koop et al. (2019) use compressed regression in a Bayesian vector autoregressive model to forecast 7 monthly macroeconomic variables with up to 129 dependent variables and 13 lags. They compare the forecast performance to univariate autoregressive models, dynamic factor models, factor augmented vector autoregressions, and the Minnesota prior vector autoregressions. The forecast accuracy of the Bayesian compressed vector autoregressions is similar or better than either factor methods or large vector autoregression methods involving prior shrinkage, for 1–12 months ahead forecast horizons.

Pick and Carpay (2018) analyze in detail multi-step ahead forecasting performance on 14 variables using a large vector autoregression to which the random subspace methods are applied, as well as a number of popular dimension reduction methods. They find that the random subspace methods and the lasso algorithm provide the most accurate forecasts. This is in line with the findings for one-step-ahead forecasts in Boot and Nibbering (2019).

9.5.2 *Microeconomics*

Schneider and Gupta (2016) use random projection regression to forecast 1-week-ahead sales for existing and newly introduced tablet computers. They produce the forecast with a bag-of-words model based on consumer reviews. This model counts the number of occurrences of words in a particular piece of text. Since many words can occur, the number of parameters is equally large. Dimension reduction by random projection results in improved forecast accuracy compared to a baseline model that ignores the consumer reviews and a support vector machine based on the reviews.

Chiong and Shum (2018) use random projections to estimate aggregate discrete-choice models with large choice sets. However, they project the large number of different choice categories on a low-dimensional subspace instead of reducing the variable space. The paper focuses on parameter estimation in several simulations and in an application to supermarket scanner data, but does not perform a forecasting exercise.

9.5.3 Finance

Elliott et al. (2013) propose complete subset regressions and illustrate the methods in a forecasting exercise on the monthly US stock returns. They find an increase in forecast accuracy for the one-step-ahead forecasts for the excess return of the S&P 500 by complete subset regressions relative to standard forecast combination methods, univariate forecasting models, lasso, ridge regression, bagging, and Bayesian model averaging.

Meligkotsidou, Panopoulou, Vrontos and Vrontos (2019) extend the complete subset regressions framework to a quantile regression setting. After constructing the complete subset combination of quantile forecasts, they employ a recursive algorithm that selects the best complete subset for each predictive regression quantile. The method is applied to forecasting the quarterly S&P 500 equity premium. Complete quantile subset regression outperforms both the historical average benchmark and the complete subset regressions in forecast accuracy and economic value.

Gillen (2016) uses a strategy very similar to complete subset regressions to obtain efficient portfolios. He selects a random subset of securities and determines the optimal portfolio weights for these securities. By generating many different random subsets of securities, one obtains multiple weights for each security. These weights are averaged to obtain the final portfolio weight for each security. The expected out-of-sample performance of this strategy dominates both the $1/N$ strategy rule and sample-based optimization of the portfolio. Shen and Wang (2017) propose a similar approach and also conclude that the out-of-sample performance is superior to various competing strategies on diversified data sets.

9.5.4 Machine Learning

Random subspace methods have their roots in the machine learning literature. Although this literature commonly does not consider macroeconomic applications, it does focus on forecasting and big data. We briefly discuss some of the contributions in the machine learning literature to forecasting with random subspace methods.

Ho (1998) introduces random subspace methods in decision trees. Multiple trees are constructed by training on random subsets of the available explanatory variables. Decisions of the trees are combined in a decision forest by averaging the estimates of the class probabilities in each tree. Compared to decision trees that are trained on the full set of explanatory variables, the subspace method shows significant improvements in forecast accuracy in a collection of different data sets.

Bay (1998) applies the same idea to nearest neighbor classifiers. He combines multiple nearest neighbor classifiers that only use a random subset of the explanatory variables. The method outperforms standard nearest neighbor classifiers, k

nearest neighbor classifiers, and neighbor classifiers with forward and backward selection of explanatory variables.

Bryll, Gutierrez-Osuna, and Quek (2003) develop a wrapper method that can be used with combinations of any type of classifier trained on random subsets of explanatory variables. The method selects the size of the random subset by comparing classification accuracy of different sizes of subsets. Moreover, the method only uses the best performing subsets to form a final prediction. The proposed method improves performance for bagging and several single-classifier algorithms on an image recognition data set.

Bingham and Mannila (2001) apply random projection as a dimensionality reduction tool to image and text data. They find that randomly projecting the data on a low-dimensional subspace performs similarly to principal component regression. However, they only assess the distortion created by the dimension reduction tool, that is, the Euclidean distance between two dimensionality reduced data vectors relative to their Euclidean distance in the original high-dimensional space, and the computational costs.

Fradkin and Madigan (2003) evaluate the performance of random projections in classification. They apply classifiers as decision trees, nearest neighbor algorithms, and support vector machines to several data sets after dimension reduction by principal component analysis and random projections. They find in all cases that principal component analysis results in better predictive performance than random projections. However, the results are based on a single draw of the random projection matrix instead of averaging over multiple predictions based on different draws of the random matrix. Cannings and Samworth (2017) also evaluate the performance of random projections in high dimensional classification. They provide a framework in which any type of classifier is applied to a low-dimensional subspace constructed by random projections of the original explanatory variables. The final prediction is based on an average over random projections, but only over random projections that provide the lowest error. The random projections approach outperforms a collection of benchmark classifiers on a number of empirical data sets in terms of forecast accuracy.

Guhaniyogi and Dunson (2015) introduce Bayesian compressed regression with an application to data from a molecular epidemiology study. They find better performance in terms of mean square prediction error relative to benchmark models in predicting individual's sensitivity to DNA damage and individual's repair rate after DNA damage.

9.6 Theoretical Results: Forecast Accuracy

This section provides theoretical results on the performance of the random subspace methods for point forecasts $\hat{y}_{T+1}$ for y_{T+1} . We consider the forecasts by random subset regression and random projection regression, two subspace methods that are feasible in high-dimensional macroeconomic forecasting settings.

A large part of this section is based on Boot and Nibbering (2019) and Elliott et al. (2013). The results are based on the assumptions that the number of regressors p_x is large, but fixed. Then, asymptotic results are established in the limit where $T \rightarrow \infty$. Estimation uncertainty is introduced by scaling the coefficients with a factor $T^{-1/2}$, which excludes the possibility to estimate these coefficients consistently, even in the asymptotic limit.

9.6.1 Mean Squared Forecast Error

In this section, we analyze the asymptotic mean squared forecast error. The results are valid for a range of weakly dependent time series models. For any estimator $\tilde{\beta}$ for the true parameter β , the asymptotic MSFE is defined as

$$\rho = \lim_{T \rightarrow \infty} TE \left[\left(y_{T+1} - z'_T \tilde{\beta} \right)^2 - \sigma^2 \right] = \lim_{T \rightarrow \infty} TE \left[(z'_T \beta - z'_T \tilde{\beta})^2 \right].$$

The error variance σ^2 is subtracted as it arises from the error ε_{T+1} , which is unpredictable by any method. This expression is very similar to the one in Hansen (2010), who divides the expression by the error variance σ^2 .

Denote the ordinary least squares (OLS) estimator without dimension reduction as $\hat{\beta}$. This estimator is based on all predictors in $z_t = (\mathbf{w}'_t, \mathbf{x}'_t)'$, and yields the following forecast:

$$\hat{y}_{T+1}^{\text{OLS}} = z'_T \hat{\beta} = z'_T (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y},$$

where $\mathbf{y} = (y_1, \dots, y_T)'$, $\mathbf{Z} = (z_0, \dots, z_{T-1})'$. Under standard regularity assumptions that ensure the convergence of $\sqrt{T}(\hat{\beta} - \beta)$ to a mean zero, normally distributed random vector, the asymptotic mean squared forecast error is

$$\begin{aligned} \rho(p_w, p_x) &= \lim_{T \rightarrow \infty} TE \left[(\hat{\beta} - \beta)' z_T z'_T (\hat{\beta} - \beta) \right] \\ &= \sigma^2(p_w + p_x), \end{aligned} \tag{9.5}$$

using that $E[z_t z'_t] = \Sigma_z$ for $t = 0, \dots, T$. The fact that the MSFE increases with p_x is the motivation to consider k -dimensional subspaces with $k < p_x$. This will reduce the variance from $\sigma^2 p_x$ to (at most) $\sigma^2 k$, but it will also induce a squared bias term into the MSFE. The trade-off between reducing the variance and increasing the bias is determined by the subspace dimension k .

When the coefficients β_x are fixed, the bias in the asymptotic mean squared forecast error is infinite because of the multiplication with T . In practice, one would expect the bias and the variance to be roughly of the same order. To make sure that the asymptotic theory reflects this finite sample situation, a convenient assumption

is that the coefficients β_x are local-to-zero, that is, $\beta_x = \beta_{x,0}/\sqrt{T}$. This renders a finite bias in the asymptotic limit. The forecast under the random subspace methods is given by

$$\hat{y}_{T+1} = \sum_{i=1}^N [\hat{y}_{T+1, S_i}],$$

$$\text{with } \hat{y}_{T+1, S_i} = \mathbf{z}'_T \mathbf{S}_i \hat{\beta}_{S_i}, \quad \hat{\beta}_{S_i} = (\mathbf{S}'_i \mathbf{Z}' \mathbf{Z} \mathbf{S}_i)^{-1} \mathbf{S}'_i \mathbf{Z}' \mathbf{y}, \text{ and } \mathbf{S}_i = \begin{pmatrix} \mathbf{I}_{p_w} & \mathbf{O} \\ \mathbf{O} & \mathbf{R}_i \end{pmatrix}.$$

Boot and Nibbering (2019) show that the sum over i can be replaced by an expectation with a negligible effect on the MSFE when the number of draws of the random matrix $\mathbf{R}$ is as large as $N = O(p_x \log p_x)$.

Identity Covariance Matrix

We can get a sense of the bias variance trade-off by making the assumption that $\Sigma_x = \mathbf{I}_{p_x}$. In this case, the exact mean squared error for random subset regression and random projection regression is

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

This shows that although the subspace dimension is k , the variance is in fact even lower, namely of $k \frac{k}{p_x}$. The optimal value of k is given by

$$k^* = \frac{\eta}{1 + \eta} p_x, \quad \eta = \frac{\beta'_{x,0} \beta_{x,0}}{\sigma^2 \cdot p_x},$$

for which we have

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

Comparing this to (9.5), we see that the factor $\sigma^2 p_x$ is reduced by a factor $\eta/(1 + \eta)$ with η the signal-to-noise ratio per parameter. If the informational content in each predictor is therefore small, gains can be expected from the dimension reduction procedure.

Intuitively, one might expect that when strong correlations between variables are present, the random subspace methods are increasingly effective. In this case, omitting one variable, but including a strongly correlated one would prevent a large omitted variable bias. However, when the covariance matrix Σ_x is of general form,

there is no analytic expression for the MSFE. To circumvent this problem, there have been several studies that derive bounds on the in-sample mean squared error under i.i.d. observations, starting with Kabán (2014) and refined by Thanei, Heinze, and Meinshausen (2017). Boot and Nibbering (2019) provide expressions for the asymptotic mean squared forecast error under general assumptions that allow for a broad range of time series models. We discuss these results in some detail to explain part of the success of random subspace methods in macroeconomic forecasting settings.

9.6.2 Mean Squared Forecast Error Bounds

For random subset regression, Boot and Nibbering (2019) show that the asymptotic MSFE can be upper bounded by

$$\rho(p_w, k) \leq \sigma^2(p_w + k) + \beta'_{x,0} \Sigma \beta_{x,0} - \frac{k}{p_x} \beta'_{x,0} \Sigma [w_s \Sigma + (1 - w_s) \mathbf{D}_\Sigma]^{-1} \Sigma \beta_{x,0}.$$

For random projection, the upper bound is

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

where $w_s = \frac{k-1}{p_x-1}$, $\mathbf{D}_\Sigma$ is a diagonal matrix with $[\mathbf{D}_\Sigma]_{ii} = [\Sigma]_{ii}$, $w_p = \frac{p_x(k+1)-2}{(p_x+2)(p_x-1)}$, and $\text{tr}(\Sigma)$ denotes the trace of $\Sigma = \text{plim}_{T \rightarrow \infty} X' M_W X$, where $M_W = \mathbf{I}_T - W(W'W)^{-1}W'$.

The first term reflects the variance, while the second and third term reflect the bias induced by selecting/projection to a low-dimensional subspace. As expected, the variance is reduced from $\sigma^2 p_x$ to $\sigma^2 k$. This variance reduction comes at the expense of a bias term. For random subset regression the bias is a weighted average of the covariance matrix Σ and its diagonal elements $\mathbf{D}_\Sigma$. When $k = 1$, all weight is put on the diagonal matrix and any information on cross-correlations is lost in the low-dimensional subspace. When $k = p_x$, we obtain the OLS mean squared forecast error (9.5). The bound for random projection regression depends on a weighted average of Σ and the constant diagonal matrix $\frac{\text{tr}(\Sigma)}{p_x} \mathbf{I}_{p_x}$. When $k = 1$, nearly all weight is put on the diagonal matrix. When $k = p_x$, we again obtain the OLS mean squared forecast error (9.5).

To get further insight in the bounds, decompose the covariance matrix $\Sigma_x = \mathbf{V} \mathbf{\Lambda} \mathbf{V}'$, where the orthogonal matrix $\mathbf{V}$ contains the eigenvectors of Σ_x and $\mathbf{\Lambda}$ is a diagonal matrix with eigenvalues sorted in decreasing order. We assume that the predictors are standardized, such that $[\Sigma_x]_{ii} = 1$ for $i = 1, \dots, p_x$. The data

generating process (9.1) can be rewritten in terms of the principal components $\mathbf{f}'_t = \mathbf{x}'_t \mathbf{V}$ as

$$y_{t+1} = \mathbf{f}'_t \boldsymbol{\gamma} + \varepsilon_{t+1}.$$

The bounds for both random subspace methods reduce to

$$\rho(0, k) \leq \sigma^2 k + \sum_{i=1}^{p_f} \gamma_{0,i}^2 \lambda_i \left(1 - \frac{k}{p_x} \frac{\lambda_i}{w_r(\lambda_i - 1) + 1} \right), \quad (9.6)$$

with $w_r = w_s$ for random subset regression and $w_r = w_p$ for random projection regression.

In macroeconomic forecasting, it is generally assumed that a small number of principal components, associated with the largest eigenvalues of $\boldsymbol{\Sigma}_x$, drive the dependent variable. In other words, the nonzero $\gamma_{0,i}$ corresponds to $\lambda_i \gg 1$. 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 $\lambda_i \gg 1$. The bound in (9.6) shows that in this setting a subspace dimension $k < p_x$ reduces the variance without inducing a large bias in the forecast. However, when $\gamma_{0,i}$ is nonzero for $\lambda_i < 1$, both methods should set k close to p_x to avoid bias.

9.6.3 Theoretical Results in the Literature

Random projections are historically motivated by the Johnson–Lindenstrauss lemma (Johnson and Lindenstrauss, 1984), for which Dasgupta and Gupta (2003) provide a simple proof. The lemma shows that random projections reduce the dimensionality of a set of points in Euclidean space while approximately preserving pairwise distances. Achlioptas (2003) and Li et al. (2006) show that this lemma also holds for random projections with very low computational costs. Given that they preserve pairwise distances, random projections can then, for example, be used to perform k -means clustering in a low-dimensional space with provably accurate assignment of observations to clusters (Boutsidis, Zouzias, Mahoney, and Drineas, 2015). Vempala (2005) gives an overview of the theoretical results for random projections.

Upper bounds on the in-sample mean squared error under random projections have been established by Maillard and Munos (2009) based on the Johnson–Lindenstrauss lemma, and improved upon by Kabán (2014). Thanei et al. (2017) showed that further improvements can be made by including a minimization step in the derivation, which uniformly lowers the upper bound. Boot and Nibbering (2019) improve these bounds further by recognizing that the random projection matrix, which is not exactly orthogonal, can be replaced by an exactly orthogonal matrix. They also provide the corresponding bound for random subset regression. The bounds show that the bias resulting from components associated with large

eigenvectors is expected to be small. This is also the conclusion of Slawski et al. (2018), who use a different route to arrive at an upper bound, where the bias is shown to depend on the variation in the data associated with small eigenvalues. Elliott et al. (2013) provide an exact expression for the MSFE which contains the sum over all subsets and therefore does not lend itself for interpretation. Exact results for the mean squared forecast error for both random projections and random subset regression are not yet available.

In a Bayesian framework, Guhaniyogi and Dunson (2015) show that the predictive density under random projection converges to the true predictive density when the model is near sparse, i.e., the sum of the absolute coefficients remains finite when the sample size and the number of predictors increase.

9.7 Empirical Illustrations

This section illustrates the performance of the random subspace methods in forecasting macroeconomic indicators of the US economy. The results in the first application are taken from Boot and Nibbering (2019), where here we add the results for random compression. The second application provides new results for multi-step-ahead forecasts using the data from Stock and Watson (2002).

9.7.1 Empirical Application: FRED-MD

A detailed description of the data and methods can be found in Boot and Nibbering (2019). We forecast industrial production and consumer price index (as transformed by McCracken and Ng (2016)) from January 1980 to December 2014. The model is a linear AR(4) model, treating the remaining 129 series in the FRED-MD data set, as well as lags five and six, as potentially relevant predictors x_t . We compare the forecast performance as measured by mean squared forecast error of random subset regression (RS), random projection regression (RP), compressed regression (CR), principal component regression (PC), partial least squares (PL), ridge regression (RI), and lasso (LA). Compressed regression is implemented as described above, with the only exception that, like the other methods, the subspace dimension k is determined through historical forecast performance.

Results

Table 9.1 shows the MSFE of the selected forecasts relative to the AR(4) model, for forecasting the industrial production index (INDP) and inflation (CPI). The subspace dimension, number of factors, and regularization constants are determined based on past performance. Random subset regression is most accurate for forecast-

Table 9.1 FRED-MD: relative MSFE on industrial production and inflation

	RS	RP	CR	PC	PL	RI	LA
INDP	0.814	0.840	0.836	0.890	0.898	0.844	0.826
CPI	0.887	0.868	0.938	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. 9.1 Forecast accuracy for different subspace dimensions

ing the industrial production index and random projection regression for forecasting inflation.

Figure 9.1 shows how the MSFE depends on the subspace dimension k , and on the number of principal components, if the same dimension is selected throughout the forecast period. For industrial production a much smaller number of principal components is selected compared to CPI. The bounds on the MSFE of the random subspace methods discussed above suggest that random subset regression is more accurate if only principal components associated with the largest eigenvalues drive the dependent variable. The empirical result here, although based on only two series, agrees with this observation. We further note that the dependence on the subspace dimension is relatively mild.

9.7.2 Empirical Application: Stock and Watson (2002)

Stock and Watson (2002) study the performance of factor models on macroeconomic data for multi-step-ahead forecasts. Here, we repeat their analysis and compare the forecast error of the diffusion index models based on principal component analysis to random subset regression, random projection regression, and compressed regression.

Methods

Details of the data and appropriate transformations can be found in Stock and Watson (2002). The goal is to forecast a set of four real and four price variables. The real variables are: industrial production (IP), personal income (PI), manufacturing (MFG), and nonagricultural employment (NE). The price variables are consumer price index (CPI), consumption deflator (CD), CPI excluding food and energy (CFE), and the producer price index (PPI). For each of these variables, we consider $h = \{6, 12, 24\}$ -step ahead forecasts for the period 1970:1+ h , until 1998:12.

Stock and Watson (2002) consider three versions of a diffusion index model: a model with only diffusion indices as explanatory variables (DI), a model that adds a number of lags of the dependent variable to DI (DI-AR), and a model that also includes lags of the diffusion indices themselves (DI-AR, Lag). The first two models lower the dimension of the space spanned by the predictors by considering $\mathbf{f}'_t = \mathbf{x}'_t \mathbf{E}$ with $\mathbf{E}$ the matrix of eigenvector loadings of the correlation matrix of the predictors. The random subspace methods instead take $\mathbf{f}'_t = \mathbf{x}'_t \mathbf{R}_i$, with $\mathbf{R}_i$ a random matrix, and average over a large number of forecasts obtained from using different realizations $\mathbf{R}_i$. The relative forecast performance then allows us to compare whether information is extracted more efficiently from $\mathbf{x}_t$ by using diffusion indices or random subspaces. For the DI-AR, Lag model, the random subspace analogue is less clear. For example, for random subset regression, we could select a number of variables and also include lags of these variables. Alternatively, we could expand $\mathbf{x}_t$ by lags of the data, and select directly from this enlarged matrix. We do not pursue these options here.

The forecasts of the diffusion index models are based on

$$\hat{y}_{T+h|T}^h = \hat{\alpha}_h + \sum_{j=1}^m \hat{\beta}'_{hj} \hat{\mathbf{f}}_{T-j+1} + \sum_{j=1}^p \hat{\gamma}_{hj} y_{T-j+1},$$

where the factors $\hat{\mathbf{f}}_t$ are the first k principal components of the standardized matrix of predictors $\mathbf{X}$. The DI model takes $p = 0$, $m = 1$, and $0 \leq k \leq 12$. The DI-AR model takes $m = 1$, $0 \leq p \leq 6$, and $0 \leq k \leq 12$. The DI-AR, Lag model takes $0 \leq k \leq 4$, $1 \leq m \leq 3$, and $0 \leq p \leq 6$. The number of factors and lags are selected by BIC.

The number of lags and the subspace dimension for the subspace methods are based on past predictive performance, with a burn-in of 36 months. The maximum subspace dimension is set as $k_{\max} = 70$, which is the largest number feasible for the first forecasts.

Results

For the real variables, Table 9.2 reports the root mean squared forecast error for the AR model, and the diffusion index models DI, DI-AR, and DI-AR, Lag. For the

Table 9.2 Stock and Watson (2002): forecasting results real variables

Horizon	IP			PI			MS			NE		
	6	12	24	6	12	24	6	12	24	6	12	24
DI-AR, Lag	0.73	0.66	0.57	0.79	0.82	0.76	0.66	0.56	0.64	0.93	0.87	0.74
DI-AR	0.78	0.67	0.58	0.81	0.92	0.83	0.76	0.61	0.66	0.97	0.88	0.76
DI	0.74	0.59	0.58	0.81	0.92	0.83	0.68	0.57	0.66	0.95	0.84	0.75
RS-AR	0.59	0.55	0.72	0.72	0.76	0.90	0.61	0.53	0.82	0.76	0.73	0.86
RP-AR	0.61	0.57	0.64	0.72	0.77	0.90	0.62	0.56	0.79	0.77	0.72	0.83
CR-AR	0.65	0.59	0.67	0.73	0.77	0.92	0.66	0.56	0.80	0.78	0.75	0.82
RS	0.60	0.53	0.64	0.71	0.77	0.87	0.63	0.54	0.79	0.76	0.72	0.87
RP	0.60	0.56	0.64	0.71	0.77	0.89	0.64	0.56	0.78	0.76	0.73	0.84
CR	0.61	0.57	0.64	0.72	0.80	0.87	0.65	0.57	0.79	0.80	0.81	0.88
RMSE, AR	0.030	0.049	0.075	0.016	0.027	0.046	0.028	0.045	0.070	0.008	0.017	0.031

Note: This table shows the mean squared forecast error relative to the AR model for the real variables industrial production (IP), personal income (PI), manufacturing and sales (MS), nonagricultural employment (NE), for forecast horizons $h = \{6, 12, 24\}$, for a model with lags of the dependent variable and lags of the diffusion indices (DI-AR, Lag), with diffusion indices and lags of the dependent variable (DI-AR), and with diffusion indices only (DI). We also report the relative mean squared forecast error of the random subset regression (RS), random projection regression (RP), and compression regression (CR) with and without lags of the dependent variable

DI-AR and DI model, the table also reports the relative mean squared error of the three random subspace methods.

We see a clear pattern where the subspace methods outperform the diffusion indices for $h = \{6, 12\}$, but are outperformed for $h = 24$. We also see that the differences between the three random subspace methods are generally small. The ranking of the three methods on 6 months and 1 year horizons is relatively stable as: random subset regression, random projection regression, and random compression. For 2-year ahead forecasts random projection seems to be the most stable.

In Table 9.3, we report the results for the price variables. Here, the results are more diffuse. For CPI and PPI, the best diffusion index models outperform the random subspace methods at $h = \{6, 12\}$. For CD and CPI-FE there is no clear winner. Surprisingly, on the price variables, random subset regression is the weakest random subspace methods. The difference between random projection regression and random compression is generally small.

9.8 Discussion

This chapter examines random subspace methods in a macroeconomic forecasting framework. Since the number of variables in this setting approaches the number of observations, parameter uncertainty is an important source of forecast inaccuracy. Random subspace methods form a new class of dimension reduction methods that can potentially address these concerns. This chapter summarizes empirical and theoretical results on the use of random subspace methods in forecasting. There is much scope for expanding these results, and below we briefly discuss four possible directions for future research.

There are no formal results on the effect of instabilities on the forecast performance under random subspace methods. There is a wealth of empirical evidence for unstable forecasting relations in economic forecasting models. This leaves open the possibility that the success of random subspace methods in macroeconomic forecasting is due to the fact that they somehow reduce adversary effects of structural instabilities.

This chapter only focuses on point forecasts under mean squared error loss. The performance of random subspace methods under alternative loss functions for point forecasts is not well studied. Moving away from point forecasts, it would be interesting to see whether the combination of big data and subspace methods can contribute to improvements in density forecasting.

Theoretical results on random subspace forecasts are limited in two important ways. We lack easy to interpret exact expressions of the mean squared forecast error, even when the number of variables grows slowly compared to the number of observations. This leaves open the question whether these methods are in some sense optimal, or at least optimal for a particular class of data generating processes.

Another main case of interest is when the number of variables is much larger than the number of observations. Given that subspace methods can easily be implemented

Table 9.3 Stock and Watson (2002): forecasting results price inflation

Horizon	CPI			CD			CPI-FE			PPI		
	6	12	24	6	12	24	6	12	24	6	12	24
DI-AR, Lag	0.79	0.70	0.59	0.97	0.90	0.67	0.85	0.84	0.84	0.91	0.86	0.76
DI-AR	0.78	0.69	0.70	0.95	0.87	0.70	0.85	0.85	0.87	0.91	0.85	0.86
DI	1.59	1.30	1.07	1.64	1.34	1.07	1.74	1.57	1.46	2.41	2.43	2.11
RS-AR	0.80	0.77	0.67	0.93	0.89	0.65	0.86	0.79	0.91	0.95	0.90	0.79
RP-AR	0.79	0.75	0.64	0.93	0.86	0.65	0.85	0.82	0.78	0.94	0.87	0.76
CR-AR	0.80	0.73	0.67	0.91	0.87	0.67	0.85	0.78	0.79	0.94	0.88	0.80
RS	1.61	1.31	1.10	1.51	1.27	1.03	1.76	1.55	1.37	2.42	2.45	2.05
RP	1.44	1.17	0.95	1.39	1.16	0.92	1.44	1.31	1.18	2.05	1.92	1.58
CR	1.46	1.17	0.95	1.39	1.16	0.92	1.45	1.31	1.19	2.04	1.91	1.58
RMSE, AR	0.010	0.021	0.052	0.007	0.016	0.038	0.009	0.019	0.046	0.017	0.033	0.078

Note: This table shows the mean squared forecast error relative to the AR model for the price inflation variables consumer price index (CPI), consumption deflator (CD), consumer price index excluding food and energy (CPI-FE), producer price index (PPI). For further information, see the note following Table 9.2

in scenarios where $p_x \gg T$, theoretical results under the assumption that $p_x/T \rightarrow c$ with $c \in (0, \infty)$ or even $p_x/T \rightarrow \infty$ would be highly desirable. However, we are not aware of any results in this regime.

References

- Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4), 671–687.
- Bai, J., & Ng, S. (2006). Confidence intervals for diffusion interval forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4), 1133–1150.
- Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2), 304–317.
- Bay, S. D. (1998). Combining nearest neighbor classifiers through multiple feature subsets. In *Proceedings of the Fifteenth International Conference on Machine Learning ICML* (vol. 98, pp. 37–45). San Francisco: Morgan Kaufmann.
- Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 245–250). New York: ACM.
- Boot, T., & Nibbering, D. (2019). Forecasting using random subspace methods. *Journal of Econometrics*, 209(2), 391–406. <https://doi.org/10.1016/j.jeconom.2019.01.009>.
- Boutsidis, C., Zouzias, A., Mahoney, M. W., & Drineas, P. (2015). Randomized dimensionality reduction for k -means clustering. *IEEE Transactions on Information Theory*, 61(2), 1045–1062.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.
- Breiman, L. (1999). Pasting small votes for classification in large databases and on-line. *Machine Learning*, 36(1–2), 85–103.
- Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition*, 36(6), 1291–1302.
- Cannings, T. I., & Samworth, R. J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4), 959–1035.
- Chiong, K. X., & Shum, M. (2018). Random projection estimation of discrete-choice models with large choice sets. *Management Science*, 65, 1–457.
- Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: a simple theoretical explanation. *International Journal of Forecasting*, 32(3), 754–762.
- Dasgupta, S., & Gupta, A. (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1), 60–65.
- Elliott, G., Gargano, A., & Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, 177(2), 357–373.
- Elliott, G., Gargano, A., & Timmermann, A. (2015). Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control*, 54, 86–110.
- Elliott, G., & Timmermann, A. (2013). *Handbook of economic forecasting*. Amsterdam: Elsevier.
- Fradkin, D., & Madigan, D. (2003). Experiments with random projections for machine learning. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 517–522). New York: ACM.
- Frieze, A., Kannan, R., & Vempala, S. (2004). Fast Monte-Carlo algorithms for finding low-rank approximations. *Journal of the Association for Computing Machinery*, 51(6), 1025–1041.
- Garcia, M. G., Medeiros, M. C., & Vasconcelos, G. F. (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting*, 33(3), 679–693.
- Gillen, B. J. (2016). Subset optimization for asset allocation. Social Science Working Paper, 1421, California Institute of Technology, Pasadena

- Guhaniyogi, R., & Dunson, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512), 1500–1514.
- Hansen, B. E. (2010). Averaging estimators for autoregressions with a near unit root. *Journal of Econometrics*, 158(1), 142–155.
- Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832–844.
- Johnson, W. B., & 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 *Artificial intelligence and statistics* (pp. 448–456). Boston: Addison-Wesley.
- Koop, G., Korobilis, D., & Pettenuzzo, D. (2019). Bayesian compressed vector autoregressions. *Journal of Econometrics*, 210(1), 135–154.
- Leroux, M., Katchoni, R., & Stevanovic, D. (2017). Forecasting economic activity in data-rich environment. University of Paris Nanterre, EconomiX.
- Li, P., Hastie, T. J., & Church, K. W. (2006). Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 287–296). New York: ACM.
- Mahoney, M. W., & Drineas, P. (2009). CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3), 697–702.
- Maillard, O., & Munos, R. (2009). Compressed least-squares regression. In *Advances in neural information processing systems* (pp. 1213–1221). Cambridge: MIT Press.
- McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4), 574–589.
- Meligkotsidou, L., Panopoulou, E., Vrontos, I. D., & Vrontos, S. D. (2019). Out-of-sample equity premium prediction: A complete subset quantile regression approach. *European Journal of Finance*, 1–26.
- Ng, S. (2013). Variable selection in predictive regressions. *Handbook of Economic Forecasting*, 2(Part B), 752–789.
- Pick, A., & Carpay, M. (2018). Multi-step forecasting with large vector autoregressions. Working Paper.
- Schneider, M. J., & 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.
- Shen, W., & Wang, J. (2017). Portfolio selection via subset resampling. In *Thirty-First AAAI Conference on Artificial Intelligence* (pp. 1517–1523).
- Slawski, M. et al. (2018). On principal components regression, random projections, and column subsampling. *Electronic Journal of Statistics*, 12(2), 3673–3712.
- Stock, J. H., & 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., & Watson, M. W. (2006). Forecasting with many predictors. *Handbook of economic forecasting* (vol. 1, pp. 515–554). Amsterdam: Elsevier.
- Thanei, G.-A., Heinze, C., & Meinshausen, N. (2017). Random projections for large-scale regression. In *Big and complex data analysis* (pp. 51–68). Berlin: Springer.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1), 267–288.
- Timmermann, A. (2006). Forecast combinations. In *Handbook of economic forecasting* (vol. 1, pp. 135–196). Amsterdam: Elsevier.
- Vempala, S. S. (2005). *The random projection method. Series in discrete mathematics and theoretical computer science*. Providence: American Mathematical Society.