CENTRE FOR ECONOMETRIC ANALYSIS
CEA@Cass

Centre for
Econometric
Analysis

http://www.cass.city.ac.uk/cea/index.html

Cass Business School
Faculty of Finance
106 Bunhill Row
London EC1Y 8TZ

Robust Forecasting by Regularization

Dobrislav Dobrev and Ernst Schaumburg

CEA@Cass Working Paper Series

WP–CEA–03-2012

Robust Forecasting by Regularization

`Preliminary and Incomplete`

Dobrislav Dobrev[a], Ernst Schaumburg[b,*]

[a] *Dobrislav Dobrev: Federal Reserve Board of Governors, dobrislav.p.dobrev@frb.gov*
[b] *Ernst Schaumburg: Federal Reserve Bank of New York, ernst.schaumburg@gmail.com*

Abstract

The prediction of multivariate outcomes in a linear regression setting with a large number of potential regressors is a common problem in macroeconomic and financial forecasting. We exploit that the frequently encountered problem of nearly collinear regressors can be addressed using standard shrinkage type estimation. Moreover, independently of near collinearity issues, when the outcomes are correlated random variables, univariate forecasting is often sub-optimal and can be improved upon by shrinkage based on a canonical correlation analysis. In this paper, we consider a family of models for multivariate prediction that employ both types of shrinkage to identify a parsimonious set of common forecasting factors. The approach is designed to jointly forecast a vector of variables of interest based on a near collinear set of predictors. We illustrate its promising performance in applications to several standard forecasting problems in macroeconomics and finance relative to existing approaches. In particular, we show that a single factor model can almost double the predictability of one-month bond excess returns across a wide maturity range by using a set of predictors combining yield slopes and the maturity related cycles of Cieslak and Povala (2011).

Keywords:
Out-of-sample forecasting, regularization, reduced rank regression, ridge regression

1. Introduction

Let Y be a m dimensional vector of variables of interest that the econometrician wishes to predict using a vector, X, consisting of a large but finite number n random variables. In the time series context, $Y = Y_{t+h}$ and $X = X_t$, and X possibly contains lagged elements of Y itself.[1] The goal is to identify the best linear predictor in the mean squared error sense based on the multivariate regression:

$$\mathbf{Y} \quad = \quad \mathbf{X}\Theta + \mathbf{e}\ , \quad \Theta \in \mathbb{R}^{n\times m} \tag{1}$$

[*] We are grateful for comments from Christopher Sims and participants at the IF workshop at the Federal Reserve Board of Governors.
The views expressed herein are those of the authors and should not be interpreted as reflecting the views of the Federal Reserve Board of Governors, the Federal Reserve Bank of New York or the Federal Reserve System.

[1] Without loss of generality, we shall assume throughout that X, Y are zero mean.

04/25/2012

where $\mathbf{Y}, \mathbf{X}$ are the $(T \times m)$ and $(T \times n)$ matrices of stacked observations of outcomes, Y, and predictors, X, and $\mathbf{e}$ is a $T \times m$ matrix of residual terms.

Prediction of multivariate outcomes based on a multivariate regression (1) with a large number of non-orthogonal regressors is commonplace in macroeconomics and finance. Stock and Watson (2011), for instance, consider forecasting $m = 35$ macro aggregates and $m = 108$ disaggregate series using the latter as $n = 108$ predictors for $T = 195$ quarters of observations. Cieslak and Povala (2011) extend Cochrane and Piazzesi (2005) to forecast up to $m = 20$ bond excess returns using up to $n = 20$ predictors derived from lagged yields and inflation for $T = 468$ monthly observations. We shall study these two examples in greater detail below. In many such forecasting applications, alternatives to ordinary least squares (OLS) are preferable due to the common occurrence of one or more of the following three features of the problem:

First, when the number of predictors, n, is larger than the number of observations, T, OLS is infeasible. Even when $n < T$ but n is large, the sheer number of potential right hand side predictors leads to an in-sample over-fitting problem. One way to address this problem, as we shall in this paper, is to postulate that $\mathbf{X}$ contains a smaller number $k \ll n$ components, $\mathbf{Z}$ that predict $\mathbf{Y}$:

$$\mathbf{Y} \quad = \quad \mathbf{Z}B + \mathbf{e}\,, \;\; B \in \mathbb{R}^{k \times m} \tag{2}$$

In reality, all n dimensions of the data may of course contain useful information for predicting Y and the justification for focussing on $k \ll n$ components is therefore that the signal-to-noise ratio in the relationship between Y and the remaining $n-k$ components is so poor that it would degrade the forecasting performance of the model to include them. In practice, the dimension k is therefore a key "bandwidth" parameter to be chosen by the econometrician (and one for which a strong prior is often not available). When $\mathbf{Z}$ consists of k elements or k linear combinations of $\mathbf{X}$, this is known as the *variable* selection and *factor* selection problems respectively. In this paper we focus strictly on the factor selection problem.

Second, while near collinearity of the predictors necessarily occurs when $n \approx T$, it is a prominent feature of the problem in some financial datasets even when $n \ll T$, especially when series are connected by a (near) arbitrage relationship or (near) accounting identity. The ill-condition of the design matrix, $\mathbf{X}$, typically results in severe instability of the estimated relationship between Y and X and a poor out-of-sample forecasting performance. A general framework for addressing an ill-conditioned system (1) is *regularization*, which naturally leads to a shrinkage type estimator that we shall use extensively in this paper.

Finally, when the dimension of Y is $m \geq 2$ and the elements of Y are correlated variables, naïve OLS may be dominated by a shrinkage estimator that exploits the structure of the canonical covariates of Y and X.[2] In other words, forecasting multiple outcomes using a smaller number

[2] This situation arises when the Ys themselves exhibit a strong (predictable) factor structure, such as the level, slope and curvature of bond yields.

of common forecasting factors imposes discipline on the factor extraction problem. When the design matrix is also ill conditioned, the two types of shrinkage estimation may be combined to produce a robust forecasting model. A main contribution of this paper is the development of a family of estimators of Θ that apply standard regularization techniques (to deal with near collinarity) to reduced rank regression (in order to exploit covariance between outcomes) that provides the econometrician with a flexible framework for extracting *common* predictive factor structures in the data. The resulting forecasting models are called Regularized Reduced Rank Regression models, or simply RRRR.

We demonstrate that the proposed RRRR estimators perform very well across a range of applications to both the Stock and Watson (2011) macro data set as well as bond excess returns, and investigate a number of data driven methods for the choice of regularization threshold based on random matrix theory. We find that the method of regularization has a non-trivial impact on forecasting performance. In particular, we find that the commonly used Tihonov regularization performs noticeably worse in our macro application than the simpler spectral truncation method which is a natural extension of principal components regression (PCR) to the reduced rank framework. By contrast, the Tikhonov scheme does markedly better in the finance application.

In all our applications, the RRRR model is among the best performing and most parsimonious out-of-sample predictors. In particular, we find support for the Stock and Watson (2011) finding of roughly 5 important principal components among the 108 individual predictors they consider (our estimate varies across subsamples from 3-8 with a median of 5), but that the dimension of the most parsimonious predictor set is somewhat less, at 3-5. Thus RRRR provides a more parsimonious model for jointly forecasting the 35 Macro aggregates in the Stock and Watson (2011) data set.

In the case of the notoriously hard problem of forecasting 1-month bond excess returns, we investigate a number of different predictors considered in the literature, including maturity related inflation cycles (henceforth "cycles"), forward rates, forward slopes, and the current yield slopes. Across all specifications, the RRRR is consistently among the best performing methods, while parsimoniously relying on a single common forecasting factor to predict the entire curve of bond excess returns (1-month excess returns to holding bonds of maturity from 1 to 15 years), consistent with the presence of a strong factor structure in the cross-section of bond returns. In particular, we confirm a recent result by Cieslak and Povala (2011) which suggests that a single or two factor model based on cycles is useful for jointly predicting holding period returns. We are able to improve somewhat on this result by including individual cycles as predictors and letting RRRR extract a single predictive factor that captures the relevant information. Remarkably, the out-of-sample R-squared of the non-overlapping monthly forecasts can be almost doubled by including current slopes along with cycles, but due to the severe ill-condition, only the RRRR approach is fully able to take advantage of the extra information.

The remainder of the paper is structured as follows. In Section 2, we briefly review regularization as a general technique to deal with high dimensional predictor sets and near collinearity. In Section 3, we discuss how shrinkage estimation arises naturally in the context of a multivariate response Y. We then turn to developing the Regularized Reduced Rank Regression (RRRR) model in Section 4 and the issue of factor interpretability in Section 5. Data driven techniques for choosing the degree of regularization are discussed in Section 6 while Section 7 documents the efficacy of RRRR as a forecasting model in our application to the Stock and Watson (2011) macro data and bond return forecasting. We find promising performance compared to other commonly used techniques, although we stress that no one method is uniformly best across datasets and sample periods. Section 8 concludes.

2. Regularization and Shrinkage Estimation

In classical regression analysis, regularization is a particular method for shrinking the set of admissible predictors that essentially involves a delicate trade-off between over- and under-fitting of the data. In this section we introduce filter-factors and two regularization schemes with long histories in applied work that differ dramatically in their treatment of eigenvalues of "intermediate" size. The first method, Principal Components Regression (PCR), eliminates eigenvalues of $\mathbf{X}$ that fall below a chosen threshold while the second scheme, Tikhonov regularization, down-weights small eigenvalues depending on their size.[3] Since the "optimal" filter factors depend on the properties of the un-known noise, there is in general no ex-ante preferred scheme and the performance of each must be evaluated in applications.

Unless otherwise indicated, we shall for notational simplicity assume that $T > n$ and work with two matrix norms compatible with a mean squared error forecast objective. On the space of positive semidefinite (PSD) $n \times n$ matrices, S, we define $\|S\| = tr\{S\}$. On the space of real $n \times m$ matrices, A, we shall use the Frobenius norm, $\|A\| = tr\{A'A\}^{1/2}$. Throughout we use the notation $S_{XY} = \mathbf{X}'\mathbf{Y}/T$ for the sample covariance matrix of two generic data matrices $\mathbf{X}$ and $\mathbf{Y}$.

2.1. Related Literature

The RRRR framework involves the choice of two shrinkage parameters: the degree of regularization, which we denote by ρ, and the predictor dimension k. There is a vast literature dealing with each of these types of shrinkage both from the frequentist and Bayesian perspective.

In the extensive Bayesian forecasting literature the ill-condition of the system (1) is naturally dealt with by transforming the problem of determining a point estimate in $\mathbb{R}^{n\times m}$ into a well-posed extension on the larger space of distributions. The precision of the Gaussian prior on the

[3] Another popular regularization scheme, least absolute shrinkage and selection (LASSO), is not considered here as it does not allow for a closed form solution but instead involves a difficult numerical optimization problem. See Mol, Giannone, and Reichlin (2008) for a comprehensive comparative study of ridge and LASSO regression based forecasts in a univariate setting.

regression coefficients Θ can be interpreted as a regularization parameter. Of particular relevance to our setting is Doan, Litterman, and Sims (1984) who consider multivariate Bayesian VAR forecasting, Koop and Potter (2004) who consider Bayesian forecasting in dynamic factor models with many regressors, and Geweke (1996), who proposed Bayesian estimation of reduced rank regressions. Although Geweke (1996) proposes a Bayesian model selection approach to choosing k there is no mention of the choice of prior variance, ρ, as ill-conditioned design matrices are not his focus. Moreover, the parametrization of the Bayesian reduced rank regression is not in terms of an easily interpretable prior that can be understood as a regularization of the corresponding frequentist model.[4]

Another rich strand of the Bayesian literature, concerned with model selection procedures, attempts to pick a subset of predictor variables from the original n predictors of Y. In the Bayesian framework, one needs the marginal distribution of the data, the prior probabilities of each of the 2^n models and the ability to compute the posterior distribution of the parameters of interest for each model. In the context of linear regression, each of these components is available in closed form, as shown in Raftery, Madigan, and Hoeting (1997). The main problem is that the model space quickly gets too large , even for modest size n, and the estimation of posterior model probabilities and Bayesian model averaging must be based on a subset of models. The factor approach implied by reduced rank regression circumvents the curse of dimensionality at the cost of the potential loss of interpretability of the resulting factors which are linear combinations of many, typically disparate, regressors. In Section 5 we directly address this concern and suggest a practical approach for imposing a degree of interpretability on the factor structure.

In the frequentist forecasting literature, Principal Component Regression (PCR) is perhaps the most frequently used method for dealing with ill-conditioned systems. Similarly to RRRR, PCR achieves regularization via down-weighting (in fact eliminating) the influence of small eigenvalues of S_{XX} but differs from RRRR in that it does not incorporate any information from the cross-moment matrix, S_{XY}, in the factor selection. A prominent example of PCR in macroeconomic forecasting, is Stock and Watson (1998), who suggest forecasting key variables like inflation and output using factors extracted from an extensive set of macroeconomic time series and choosing the number of factors based on out-of-sample forecasting performance.[5]

Partial Least Squares (PLS), which is based on a singular value decomposition of S_{XY}, has a long history in chemometrics but has also been applied in economics and is closely related to the 3PRF model recently proposed by Kelly and Pruitt (2011). However, PLS is not in general a shrinkage technique (in terms of the eigenvalues of S_{XX}) and it therefore does not per se address ill-conditioned design matrices. By contrast, the degree of shrinkage involved in RRRR

[4] To be precise, the reduced rank regression coefficient is $\Theta = AB$ where Θ is of rank $k < n$. Geweke (1996) considers separate (independent) Gaussian priors on A and B which are hard to interpret as it is the product AB that has economic meaning.

[5] This is clearly different from exploiting the information in the S_{XY} matrix because the most important factors in explaining X may not be the most important forecasting factors.

is explicitly parametrized and interpretable in terms of Bayesian precision priors. Moreover, the RRRR estimators solve an explicit penalized least squares objective function involving the two explicit shrinkage parameters whereas it is not clear in which sense the 3PRF/PLS type estimators are optimal nor which explicit objective function is being optimized, thereby complicating its interpretation considerably.

2.2. Regularized Least Squares

The properties of the linear system (1) are completely determined by the singular value decomposition of the matrix $\mathbf{X}$:

$$\mathbf{X} = U\Sigma V' = \sum_{i=1}^{n} \sigma_i u_i v_i' \tag{3}$$

where $U = (u_1, \ldots, u_n) \in \mathbb{R}^{T\times n}, V = (v_1, \ldots, v_n) \in \mathbb{R}^{n\times n}$ are orthonormal matrices and $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$ is a diagonal matrix containing the singular values in decreasing order. We shall often need to decompose $\mathbf{X}$ into the contribution from the r largest singular values versus the contribution from the $n - r$ smallest singular values:

$$\mathbf{X} \;\; = \;\; U_r\Sigma_r V_r' + U_{n-r}\Sigma_{n-r}V_{n-r}' \tag{4}$$

where $U = [U_r \; U_{n-r}]$, $V = [V_r \; V_{n-r}]$, and $\Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & \Sigma_{n-r} \end{bmatrix}$

The matrix $\mathbf{X}$, and hence the system (1), is called *ill-conditioned* if the following two conditions are satisfied: a) The condition number σ_1/σ_n is large, and b) The sequence of singular values $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ decreases gradually to zero.[6] Figure A.1 shows the singular values for our two empirical applications, illustrating the ill-condition of $\mathbf{X}$ in each case, ranging from the moderate (the Macro application) to the extreme (the Finance application). It is also clear from the picture, that there is no visible "gap" in the spectrum which is what is explicitly or implicitly assumed in approximate factor models in order to asymptotically identify the "true" number of factors (c.f. Chamberlain et al (1987) and Bai and Ng (2002)).

A large condition number is indicative of potential instability in the estimated Φ in the sense that even a small change in the observed Y in certain directions may lead to a disproportionate change in the estimated relationship between Y and X. To see this, note that the OLS estimate is simply

$$\hat{\Theta}_{OLS} = \sum_{i=0}^{n} v_i \frac{u_i'\mathbf{Y}}{\sigma_i} = \Theta_0 + \sum_{i=0}^{n} v_i \frac{u_i'\mathbf{e}}{\sigma_i} \tag{5}$$

[6] The case where one or more eigenvalues are literally zero is easily handled by eliminating redundant variables. However, in many situations, the addition of additional predictors simply increases the number of small eigenvalues.

where Θ_0 is the true value. Thus a large condition number implies that the OLS estimate, $\hat{\Theta}_{OLS}$, is disproportionately sensitive to noise components that lie in the space spanned by the left singular vectors corresponding to the smallest singular values.[7] In the context of the forecasting relationship (1), an ill-conditioned design matrix $\mathbf{X}$ therefore in general translates into a poor out-of-sample performance of the estimated relationship since it usually cannot be guaranteed that $u_i'\mathbf{e}/\sigma_i$ remains uniformly small (e.g. if errors are Gaussian). In the simple case of spherical errors, where $E[\mathbf{e}'\mathbf{e}] = \kappa^2 I_m$, it is easy to see that the MSE of the OLS estimator is $E\|\hat{\Theta}_{OLS} - \Theta_0\|^2 = \kappa^2\, tr\{(\mathbf{X}'\mathbf{X})^{-1}\} = \kappa^2 \sum_{i=1}^{n} \sigma_i^{-2}$, thus illustrating the problem of ill-condition.

An effective approach to solving ill-conditioned systems of equations is via *regularization* of the equation (5):

$$\tilde{\Theta} \;=\; \sum_{i=1}^{n} f_i v_i \left(\frac{u_i'\mathbf{Y}}{\sigma_i}\right) \;,\quad \|\tilde{\Theta}\|_F^2 = \sum_{i=1}^{n} f_i^2 \left(\frac{u_i'\mathbf{Y}}{\sigma_i}\right)^2 \tag{6}$$

where the sequence of so called filter factors $\{f_i\}_{i=1}^n$ satisfies that $0 \leq f_i \leq 1$ and decrease sufficiently fast that $f_i/\sigma_i \approx 0$ for large i. Clearly, in the case of OLS, $f_i \equiv 1$ and the estimator is un-regularized. Most standard regularization schemes can be expressed via a specific choice of filter factors and as such can be seen as *shrinkage* estimators with respect to the rotated coordinate system determined by the columns of V since $\|\tilde{\Theta}\|_F \leq \|\hat{\Theta}_{OLS}\|_F$.

The econometrician wishing to apply regularization techniques is thus faced with the familiar trade-off between suppressing (possibly spurious) fine features of the data (associated with small eigenvalues and presumably a high noise-to-signal ratio in finite samples) in return for gaining robustness. To be precise, let Θ_0 denote the true value, $\tilde{\Theta}_\infty$ the limiting value of the shrinkage estimator as $T \to \infty$, and $\tilde{\Theta}$ the finite sample shrinkage estimate. In general $\tilde{\Theta} \to \tilde{\Theta}_\infty \neq \Theta_0$ as $T \to \infty$ and we have the bound

$$\underbrace{E\|\Theta_0 - \tilde{\Theta}\|}_{\substack{\text{root mean squared} \\ \text{shrinkage estimation error}}} \quad\leq\quad \underbrace{\|(\Theta_0 - \tilde{\Theta}_\infty\|}_{\text{bias due to regularization}} \quad+\quad \underbrace{E\|(\tilde{\Theta}_\infty - \tilde{\Theta}\|}_{\substack{\text{dampened volatility} \\ \text{due to regularization}}} \tag{7}$$

which will tend to compare favorably to OLS when the design matrix is ill-conditioned. The first term is the (deterministic) bias induced by the regularization term under the null, which is increasing in the degree of regularization. The second term is increasing as a function of the noise dispersion but decreasing in the degree of shrinkage due to the dampening effect of the regularization term, thus creating a trade-off.[8]

[7] If all eigenvalues happen to be small (or very large), it of course merely means that the problem is badly scaled.

[8] Note that, in the classical case where $n/T \to 0$, one can let the degree of regularization go to zero at a suitable rate (to ensure a bias of order $o_p(1/\sqrt{T})$), in order to restore asymptotic unbiasedness: $\tilde{\Theta}_\infty = \Theta_0$.

2.2.1. Tikhonov Regularization a.k.a. Ridge Regression

One of the most commonly used regularization techniques is *Tikhonov* regularization due to its ease of implementation and interpretation as a penalized least squares estimator. In the (multivariate) regression context Tikhonov regularization is also known as (multivariate) *Ridge Regression* and corresponds to penalizing the norm of the solution[9]

$$\min_{\tilde{\Theta}} \|\mathbf{Y} - \mathbf{X}\tilde{\Theta}\|^2 + \rho^2\|\tilde{\Theta}\|^2\,,\ \ \tilde{\Theta} \in \mathbb{R}^{n\times m},\ \rho \geq 0 \tag{8}$$

Solving the Lagrangian implies that $\tilde{\Theta} = (\mathbf{X}'\mathbf{X} + \rho^2 I_n)^{-1}\mathbf{X}'\mathbf{Y} = \sum_{i=1}^{n}\left[\frac{\sigma_i^2}{\sigma_i^2+\rho^2}\right] v_i\left(\frac{u_i'\mathbf{Y}}{\sigma_i}\right)$, corresponding to the specific family of filter factors $f_i = \sigma_i^2/(\sigma_i^2+\rho^2)$. Clearly a larger ρ implies greater down weighting of small singular values and leads to a smaller norm of $\tilde{\Theta}$ at the cost of a greater residual norm. In general, the bias-variance trade-off (7) in the Tikonov case is

$$\tilde{\Theta} - \Theta_0 \ \ = \ \ -\underbrace{\left[I_n - (S_{XX} + \rho^2 I_n)^{-1}S_{XX}\right]}_{\text{bias}}\Theta_0 + \underbrace{(S_{XX} + \rho^2 I_n)^{-1}S_{Xe}}_{\text{dampened error}}$$

where the last term is bounded in squared norm by $\sum(\sigma_i^2 + \rho^2)^{-2}\|S_{Xe}\|^2$, whereas a (tight) upper bound on the (squared) norm of the OLS error is much larger at $\sum \sigma_i^{-4}\|S_{Xe}\|^2$.

From the penalty term in (8) it is also immediately clear that scaling and rotation of the problem is not innocuous, e.g. dividing a regressor by 10 will generally result in a different solution. Care must therefore be taken in appropriate selection and scaling of regressors.

2.2.2. Spectral Truncation Regularization a.k.a. Principal Component Regression (PCR)

For a given regularization threshold ρ, such that $\sigma_r \geq \rho \geq \sigma_{r+1}$, PCR methods simply set $f_1 = \cdots = f_r \equiv 1$ and $f_{r+1} = \cdots = f_n \equiv 0$ so than any components of Y orthogonal to the last $n - r$ left singular vectors of $\mathbf{X}$ is ignored with the tacit assumption that these components are "noisy".[10] This type of regularization can be motivated under the null that X is driven by an r-dimensional factor structure:

$$\mathbf{X} = \mathbf{F}\Lambda + \mathbf{E}.$$

Let the singular value decomposition of $\mathbf{X}$ be given by (3)-(4), then the r principal factors are given by $\mathbf{F} = \mathbf{X}V_r\Sigma_r^{-1}$ and it is assumed that only the factors (and not $\mathbf{E}$) have forecasting

[9] The Tikhonov formulation is usually slightly more general:

$$\min_{\Theta} \|\mathbf{Y} - \mathbf{X}\Theta\|^2 + \rho^2\|R'\,vec(\Theta)\|^2\,,\ \ R \in \mathbb{R}^{p\times nm}, \Theta \in \mathbb{R}^{n\times m}$$

but only $R = I_m \otimes I_n$ is usually considered in statistics. In the case of Bayesian linear regression with a i.i.d. Gaussian prior, $\rho = \frac{\sigma_{\text{noise}}}{\sigma_{\text{prior}}}$, is the ratio of standard deviation of the noise to the standard deviation of the prior.

[10] Regularization methods like PCR, that restrict attention to components of Y that lie in a subspace of $\mathbf{X}$ are also known as "sub-space" methods in the numerical analysis literature. In engineering and physics, where the system (1) frequently arises as a (deterministic) discretization of integral equations, the PCR approach has a long history and is commonly known as *Truncated Singular Value* (TSVD) or *Spectral Cutoff* regularization.

power for $\mathbf{Y}$.

The regularized (via spectral truncation) estimator is obtained by replacing S_{XX}^{-1} by its generalized inverse $S_{XX}^{\dagger} = V_r \Sigma_r^{-2} V_r'$ in the expression for the OLS estimator:

$$\tilde{\Theta} \;=\; S_{XX}^{\dagger} S_X Y = V_r \Sigma_r^{-1} (S_{U_r Y}) \tag{9}$$

while the PCR estimator is

$$\tilde{\Theta}_{PCR} \;=\; S_{FF}^{-1} S_{FY} = S_{U_r Y} \tag{10}$$

and we thus have: $\mathbf{F}\tilde{\Theta}_{PCR} = \mathbf{X}(V_r \Sigma_r^{-1}) S_{U_r Y} = \mathbf{X}\tilde{\Theta}$, so that the two methods coincide.

The Stock and Watson (1998) DFM5 estimator is an example of PCR (with $r = 5$) which we shall consider as our benchmark in our empirical study below.

In general, the bias-variance trade-off (7) in the spectral truncation case is

$$\tilde{\Theta} - \Theta_0 \;=\; \underbrace{-\left[V\, diag(\overbrace{0, \ldots, 0}^{r}, \overbrace{1, \ldots, 1}^{n-r})\, V' \right] \Theta_0}_{\text{bias}} + \underbrace{V diag(\sigma_1^{-2}, \ldots, \sigma_r^{-2}, 0, \ldots, 0) V'\, S_{Xe}}_{\text{dampened error}}$$

where the last term is bounded in squared norm by $\sum_{i=1,\ldots,r} \sigma_i^{-4} \|S_{Xe}\|^2$, whereas a (tight) upper bound on the (squared) norm of the OLS error is $\sum_{i=1,\ldots,n} \sigma_i^{-4} \|S_{Xe}\|^2$.

Finally we note that in all PCR techniques, a judiciously chosen pre-scaling of the components of $\mathbf{X}$ is clearly crucial as it will affect both singular values and vectors.

3. Reduced Rank Regression and Shrinkage Estimation

Shrinkage estimation arises as a natural procedure in situations where one wishes to jointly predict multiple outcomes, as famously pointed out by Stein (1956) in the multivariate Gaussian context. While the Stein result does not rely on any correlation between the Y components, further improvement may be possible when outcomes are correlated. This will be the case if the Ys a driven by a common low dimensional factor structure, e.g. if each component is a noisy measurement of a single variable y^*.

To see this, we momentarily abstract from the issue of ill-conditioned $\mathbf{X}$ and focus on the information about the relationship between $\mathbf{Y}$ and $\mathbf{X}$ contained in the S_{XY} matrix. Similarly to the regularization analysis, the goal will be to select "strong" signals and dampen "weak" or "noisy" signals conveyed through this matrix. The main tool of this type of analysis is the classic canonical correlation analysis (CCA) of Anderson (1951) in which linear combinations of $\mathbf{X}$ and $\mathbf{Y}$ are identified that are maximally correlated.[11] The canonical correlations analysis

[11] Anderson (1951) in turn builds on the seminal works on canonical correlations by Hotelling (1933).

identifies transformations $\tau_X = [\tau_X^{(1)}, \ldots, \tau_X^{(m)}] \in \mathbb{R}^{n\times m}, \tau_Y = [\tau_Y^{(1)}, \ldots, \tau_Y^{(m)}] \in \mathbb{R}^{m\times m}$ such that

$$\tau_X' S_{XX} \tau_X = I_m \quad \tau_Y' S_{YY} \tau_Y = I_m, \quad \tau_X' S_{XY} \tau_Y = \text{diag}(\kappa_1, \ldots, \kappa_m) \tag{11}$$

where $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_m$ are the m ordered canonical correlations.[12] Thus $\tau_X^{(1)'} X$ represents the component of X that best predicts Y and the component of Y that it predicts is given by $\tau_Y^{(1)'} Y$. The second column, $\tau_X^{(2)}$ represents the component of X, orthogonal to the first, that has the second most explanatory power for Y, and so on, where the explanatory power of each component is given by the respective canonical correlations, κ_j. The canonical correlations analysis thus combines the information contained in both S_{XX} and S_{XY} to extract the optimal predictors of Y. In applications, small canonical correlations indicate that certain dimensions of X are only weakly related to Y and including them in the point estimate of Θ may entail an unattractive bias-variance trade-off. To see this, define $\tilde{\mathbf{X}} = \mathbf{X}\tau_X$ and $\tilde{\mathbf{Y}} = \mathbf{Y}\tau_Y$ then

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\,[\text{diag}(\kappa_1, \ldots, \kappa_m)] + \tilde{\mathbf{E}} \tag{12}$$

where the standard OLS estimate in (1) can be retrieved as $\Theta = \tau_X[\text{diag}(\kappa_1, \ldots, \kappa_m)]\tau_Y^{-1}$. Note that the canonical correlations are invariant to rotation and scaling of the original X and Y (as opposed to regular correlation analysis), but we stress that CCA *does not* in and by itself address ill-condition of the $\mathbf{X}$ (or $\mathbf{Y}$) data.

This leads to the natural idea of zeroing out the $(m - k)$ canonical correlations that fall below a certain threshold and replacing the OLS estimator by the shrinkage estimator $\tilde{\Theta} = \tau_X[\text{diag}(\kappa_1, \ldots, \kappa_k, 0, \ldots, 0)]\tau_Y^{-1}$. This estimator clearly has smaller norm than the OLS estimator and k controls the degree of shrinkage.[13]

Zeroing out of the $m - k$ smallest canonical correlations is the key idea behind the reduced rank regression (RRR) of Izenman (1975) who considers a multivariate regression with a large number of non-orthogonal regressors and the problem of replacing X by a lower dimensional set of orthogonal predictors in a way that minimizes the increase in the in-sample (weighted) mean squared fitting error. To this end, define the "factors" $Z = A'X$, where A is a $n \times k$ matrix with $k \ll n$ and identifying restrictions $A' S_{XX} A = I_{k\times k}$ and consider the (weighted) least squares problem

$$\min_{\{A,B\}} \quad \|(\mathbf{Y} - \mathbf{X}AB)W^{1/2}\|^2, \quad A \in \mathbb{R}^{n\times k},\, B \in \mathbb{R}^{k\times m} \tag{13}$$

In (13), the parameters A, B are chosen jointly to minimize the fitting error of Y and the parameter k controls the degree of "shrinkage" relative to the OLS estimator. For $k \geq m$ there

[12] For simplicity we assume that $n > m$ in what follows.

[13] To see this, note that (11) implies that $\tilde{\Theta} = [\tau_X diag(\underbrace{1, \ldots, 1}_{k}, \underbrace{0, \ldots, 0}_{m-k})\tau_X'\, S_{XX}]\,\Theta_{OLS}$, and the matrix in square brackets is clearly a projection onto a k dimensional sub-space.

is no shrinkage since AB is full rank and the model is simply OLS. For $k < m$, on the other hand, the reduced rank condition imposes discipline on the choice of factors by forcing a few factors to simultaneously fit multiple components of Y.

In practice, (13) is solved separately for A and B in two steps,

$$\min_{\{A\}} \quad \|\mathrm{Var}\left[W^{1/2}Y|A'X\right]\|^2 \tag{14}$$

$$\min_{\{B\}} \quad \|(\mathbf{Y} - (\mathbf{X}A)B)W^{1/2}\|^2 \tag{15}$$

in a fashion similar to PCR, except that in PCR the A parameter in the first step solves

$$\min_{\{A\}} \quad \|\mathrm{Var}\left[X|A'X\right]\|^2 \tag{16}$$

without taking into account to the outcomes, Y, of ultimate interest.

For a given choice of weighting matrix $W \in \mathbb{R}^{m\times m}$ (e.g. $W = S_{YY}^{-1}$) in (14), it is well known that the optimal A is found by solving the generalized eigenvalue problem[14]

$$|S_{XY}WS_{YX} - \lambda S_{XX}| = 0 \tag{17}$$

and setting A equal to the k eigenvectors belonging to the largest eigenvalues. The expression (17) is also known as a *matrix pencil* and it is well-known that when the matrix S_{XX} in (17) is singular or ill-conditioned, the solution to the generalized eigenvalue problem becomes unstable (c.f. Gantmacher (1960)). Thus reduced rank regression, while a proper shrinkage estimator in the sense of exploiting the correlation structure of the outcomes, remains susceptible to instability when regressors are nearly collinear. This serves as our motivation to introduce regularization into the reduced rank regression framework. Combining the two forms of shrinkage delivers the RRRR models that are the focus of this paper.

4. Regularized Reduced Rank Regression (RRRR) Models

Combining the two types of shrinkage estimation described in the preceding sections produces a forecasting model which is robust to near collinearity and at the same time exploits the correlation structure between the Y variables. In this section, we focus on the regularization of reduced rank estimators for a fixed choice of k and defer the discussion of shrinkage parameter selection until Section 6 below.

[14] Setting A equal to the first k columns of τ_X and exploiting the relations (12) we see that the objective (13) becomes

$$\arg\min_A \|(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\tilde{C})\tau_Y^{-1}W^{1/2}\|_F^2\,, \quad \text{where } \tilde{C} = diag(\kappa_1, \ldots, \kappa_k, 0, \ldots, 0)$$

which is equivalent to running the least squares regression (13) when the choice of weighting matrix is $W = S_{YY}^{-1}$.

4.1. Tikhonov Regularization of Reduced Rank Regression

In the context of the reduced rank regression (13), Tikhonov regularization involves modifying the objective function to include a term that penalizes "large" values of $\|AB\|$:

$$\min_{\{A,B\}} \|(\mathbf{Y} - \mathbf{X}AB)W^{\frac{1}{2}}\|^2 + \rho^2 \|R(AB)W^{\frac{1}{2}}\|^2, \text{ s.t. } A'S_{XX}A = I_k \tag{18}$$

where R in general is a $q \times n$ matrix which may be chosen to differentially penalize certain directions in the parameter space.[15] In the special case when $R = I_n, W = I_m$ and $k = m$, (18) specializes to

$$\min_{\{A,B\}} \|(\mathbf{Y} - \mathbf{X}\Theta)\|^2 + \rho^2 \|\Theta\|^2 \tag{19}$$

This is known as a (multivariate) Ridge Regression in the statistics literature (and denoted RR in our applications) in which the squared norm of the implied OLS coefficients are penalized. However, we stress that in many cases of interest in macroeconomics and finance, $k \ll m$ and that the technique is much more general than that. The following Proposition thus generalizes Ridge Regression to the reduced rank context:

Proposition 1 (Regularized Reduced Rank Regression)**.** *Let $W \in \mathbb{R}^{m\times m}$ be a symmetric positive semi definite weighting matrix, then the solution to the weighted regularized reduced rank regression (18) for a given choice of k, is given by $A^\star = \{c_1; \cdots ; c_k\}$ where $c_1, \ldots, c_k$ are the k eigenvectors corresponding to the k largest eigenvalues, $\lambda_1, \ldots, \lambda_k$ of the generalized eigenvalue problem*

$$|S_{XY}WS_{YX} - \lambda(S_{XX} + \rho^2 R'R)| = 0 \tag{20}$$

Note that the weighting matrix is applied to the regularization term in (18) as well since it is natural to scale the regularization term for the m^{th} equation proportionally to the scaling of the in-sample fitting errors of the m^{th} equation. This choice also has the benefit of preserving the structure of the problem.[16]

4.2. Spectral Truncation Regularization of Reduced Rank Regression

Spectral truncation of the Reduced Rank Regression can be seen as a natural extension of the PCR idea to the reduced rank context which takes into account the correlation structure in the S_{XY}. In a multivariate PCR framework, the parameter r (the considered number of principal

[15] More generally, the penalty term would be of the form $\|\tilde{R}\, vec(AB)\|$ but to maintain the simple structure of the problem, we restrict attention to terms of the form $\tilde{R} = (W^{1/2} \otimes R)$.

[16] Finally, we note that the issue of missing values (while not discussed explicitly in this paper), in practice should be handled effectively using an EM type algorithm to iterate on the generalized eigenvalue problem in a manner similar to Stock and Watson (2011) or the methods described in Troyanskaya et al (2001).

components of $\mathbf{X}$), plays the double role of both regularizing S_{XX} as well as being the number of common factors. In general, not all r factors that are important for explaining the cross-sectional variation in $\mathbf{X}$ need be important for forecasting $\mathbf{Y}$.[17] In this case, the econometrician would want to investigate whether a subset of $k \leq r$ factors suffice for predicting Y while still using the spectral cutoff r to regularize S_{XX}.

One way to think about extending PCR to the reduced rank context is in terms of a two step procedure: In the first step, r principal factors, $\mathbf{F}$, are extracted from the n regressors, $\mathbf{X}$. Second, a reduced rank regression of $\mathbf{Y}$ on $\mathbf{F}$ is run to extract $k \leq r$ forecasting factors. It turns out that this formulation has an elegant implementation in terms of a penalized one-step estimator of the form (20) as stated in the following Proposition:

Proposition 2 (Regularized Reduced Rank Regression via Spectral Truncation)**.**
Let the singular value decomposition of $\mathbf{X}$ *be given by (3)-(4) and let* $\mathbf{F} = \mathbf{X}V_r\Sigma_r^{-1}$ *be the* r *principal factors of* $\mathbf{X}$*. For* $k \leq r$*, if* $a \in \mathbb{R}^{r\times k}$ *is the matrix of the* k *principal eigenvectors of*

$$0 \;=\; |S_{FY}WS'_{FY} - \lambda S_{FF}| \tag{21}$$

then $A = V_r\Sigma_r^{-1}a \in \mathbb{R}^{n\times k}$ *spans the eigen space of the* k *principal eigenvalues of*

$$0 \;=\; |S^{\dagger}_{XX}S_{XY}WS'_{XY} - \lambda I_n| \tag{22}$$

where $S^{\dagger}_{XX} = V_r\Sigma_r^{-2}V_r'$ *is the regularized (via spectral truncation) inverse of* S_{XX}*. Moreover, (22) can be understood as a penalized estimator of the form (20) with* $R = V'_{n-r}$ *and* $\rho \to \infty$*.*

The theorem shows that the ad-hoc two-step approach can be motivated in terms of a limiting case of a penalized estimator which puts infinite penalty on directions in the parameter space spanned by the right singular vectors belonging to the $n-r$ smallest singular values, V_{n-r}. Thus the formulation (18) is general enough to encompass spectral truncation as an important limiting special case. The limiting nature of this argument makes the Bayesian interpretation of the spectral cut-off (and other sub-space methods) somewhat more delicate relative to the more smooth Tikhonov prior.

More generally, we note that any set of filter factors can be captured by the formulation (18) since, by setting $R = V\,diag(\rho_1, \ldots, \rho_n)\,V'$ in equation (20), we have the one-to-one correspondance $f_i = \sigma_i^2/(\sigma_i^2 + \rho_i^2)$ and the interpretation of each ρ_i is as the penalty applied to the parameter sub-space spanned by the i^{th} right singular vector of $\mathbf{X}$.

[17] As a simple example, consider the case where $\mathbf{X}$ consists of lagged $\mathbf{Y}$ and some factors are serially uncorrelated and therefore not useful as predictors.

5. Factor Interpretability and Zero Restrictions

An important criticism of many factor based forecasting models in applied work is the lack of interpretability of the extracted statistical factors. In this section, we show how to partially alleviate this shortcoming by imposing zero restrictions on the columns of the reduced rank coefficient matrix $A = [A_1, \dots, A_k]$. Specifically, suppose that one wishes to impose the sequence of constraints:

$$P_1' A_1 = 0, P_2' A_2 = 0, \dots, P_k' A_k = 0 \tag{23}$$

where each P_i is some $n \times f_i$ matrix. For instance, P_i might contain only ones and zeros corresponding to selecting which variables should be excluded from the i^{th} factor. It is straightforward to directly impose these orthogonality constraints on the generalized eigenvalue problem in an iterative fashion using the following Corollary.

Corollary 1 (Constrained Regularized Reduced Rank Regression)**.** *Consider the penalized reduced rank regression problem (18) subject to the constraint $P'A = 0$ for some $P \in \mathbb{R}^{f \times n}$. Let $P^{\perp} \in \mathbb{R}^{n \times (n-f)}$ be a basis for the orthogonal complement of P, then the objective of the regularized reduced rank regression subject to the orthogonality constraint is:*[18]

$$\min_{\{a,B\}} \|(\mathbf{Y} - \mathbf{X}P^{\perp}aB)\|^2 + \rho^2 \|RP^{\perp}aB\|^2 \text{ , s.t. } a'P^{\perp\prime}S_{XX}P^{\perp}a = I_k \tag{24}$$

where $a \in \mathbb{R}^{(n-f)\times k}$ and the optimal factors are given by $A = P^{\perp}a$. For a given choice of k, the optimal solution is obtained by setting $a^{\star}$ equal to the eigenvectors corresponding to the k largest eigenvalues of the $n - f$ dimensional generalized eigenvalue problem

$$|P^{\perp\prime}S_{XY}S_{XY}'P^{\perp} - \lambda\, P^{\perp\prime}(S_{XX} + \rho^2 R'R)P^{\perp}| = 0 \tag{25}$$

When $k = 1$, the Corollary applies directly to finding the principal factor, A_1 subject to the f_1 linear constraints $P_1'A_1 = 0$. More generally, when $k > 1$, the constraint (23) is imposed iteratively: Assume that the first $j < k$ factors have been found, and that the $(j+1)^{\text{st}}$ factor should satisfy $P_{j+1}'A_{j+1} = 0$ and be orthogonal to the subset the preceding factors $\{A_{i_1}, \dots, A_{i_p}\}$. This factor can then be found by applying Corollary 1 by setting $P = [\underbrace{S_{XX}A_{i_1}}_{n\times 1}, \dots, \underbrace{S_{XX}A_{i_p}}_{n\times 1}, \underbrace{P_{j+1}}_{n\times f_{j+1}}]$, for a total of $p + f_{j+1}$ constraints. This choice ensures orthogonality with the specified preceding p factors and that $P_{j+1}'A_{j+1} = 0$.[19]

Example 1. Consider a setting with N regressors, each of which can be classified as belonging

[18] E.g. if $P = UDV'$ is the singular value decomposition, then $P^{\perp}$ can be taken as the last $n - f$ columns of U.

[19] General constraints of the form $P'\,vec(A)$ which impose constraints across eigenvectors do not preserve the structure of the problem. We therefore restrict attention to the block diagonal case, $P = diag(P_1, \dots, P_k)$.

to one or more of 4 groups, denoted by $\{G_1, G_2, G_3, G_4\}$. The goal is to find the principal factor consisting of only variables from a single group.

Variable	Memberships
1	G_1, G_2
2	G_2, G_3, G_4
3	G_1, G_4
4	G_3, G_4
5	G_1, G_3, G_4
$\vdots$	$\vdots$
N	G_3

$$\Rightarrow \text{To select "}G_1\text{"-factor, set } P' = P'_{\{G_1^\perp\}} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ & \vdots & & & \vdots & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Here the matrix $P_{\{G_1^\perp\}}$ is a $n \times g_1$ matrix, where g_1 is the number of variables that are *not* members of group G_1. The $g_i \times n$ matrices $P_{\{G_i^\perp\}}$, $i = 2, \ldots, 4$ can be similarly defined.

Solving (25), once for each of the four choices of $P \in \{P_{\{G_1^\perp\}}, \ldots, P_{\{G_4^\perp\}}\}$ yields four candidate factors and the principal factor is identified as the one associated with the largest eigenvalue. Subsequent factors can then be extracted iteratively as described in the discussion following Corollary 1 with orthogonality between factors imposed only within groups.[20]

Example 2. It is common in many situations for the econometrician to have pre-selected a small number, f, of fixed (zero mean) regressors, $F \in \mathbb{R}^f$, that he wishes to augment with a small number of k predictors, $Z = A'X$, subject to the orthogonality condition $A'S_{XF} = 0_{k \times f}$. The fixed regressors, F, may of course consist in whole or part of variables from X or lagged values of Y that the econometrician is confident belong in the forecasting equations. This is analogous to the factor augmented regression proposed by Stock and Watson (2002) for macroeconomic forecasting exploiting the information contained in a large number of variables.

$$\mathbf{Y} \;=\; \mathbf{F}\Theta + \mathbf{X}(AB) + \tilde{\mathbf{e}}\ , \qquad \Theta \in \mathbb{R}^{m \times f},\ B \in \mathbb{R}^{k \times m} \tag{26}$$

The reduced rank regression with exogenous regressors is in practice most often implemented by a two-step approach in which (Y, X) are first orthogonalized on the exogenous regressors, F, and the robust reduced rank procedure above applied to the residuals. Note that the regularization of $\mathbf{X}$ will not affect the slopes on the fixed regressors by virtue of the orthogonality requirement. It is clear that setting $P = S_{XF}$ in Corollary 1 yields a solution to the fixed regressor case.

6. Data Driven Procedures for Determining the Degree of Regularization via Sub-Space Methods

For a given choice of model complexity, $k < m$, the regularized reduced rank regression introduced in Section 4 requires a choice of the regularization parameter ρ (or r for sub-space

[20] In many cases of interest, imposing orthogonality between factors belonging to separate groups would make little sense. E.g. if group 1 was labeled "Real activity" and Group 2 "Interest rates", one would fully expect the principal factors from each group to be (imperfectly) correlated.

methods). As alluded to earlier, the key challenge in determining a regularization parameter is the generally unknown properties of the noise that make it difficult to determine the optimal trade-off in (7). Loosely speaking, the goal is to reduce the influence of "small" singular values that are prone to be "noisy" without losing potentially valuable information contained in the regressors.

In this paper, we consider an approach to the choice of regularization threshold based on the classic theory of the spectrum of large random matrices with i.i.d. entries, which is particularly well suited for studying the spectral truncation approach. Another commonly used (ad-hoc) technique often used by practitioners is based on the cross-validation idea. However, in our setting the implementation of cross validation for selection of ρ (and possibly k) is complicated by both serial dependence in the data and a relatively modest sample side in relation to the number of parameters.[21] Finally, numerous Bayesian approaches to the choice of shrinkage parameters exist and are currently being explored in a separate paper.

6.1. Regularization based on random matrix theory

We focus here on methods based on random matrix theory which are appropriate when the regressors, X, can be described by a "signal+noise" model. The ideas presented here are along the lines of e.g. Patterson, Price, and Reich (2006) and Onatski (2010) among others, but specialized to our regularization context. To be specific, assume a factor structure in $\mathbf{X}$ that is helpful for forecasting $\mathbf{Y}$:

Assumption 1 (Static Factor Structure)**.** *Let $\mathbf{Y}$ be the $T \times m$ matrix of outcomes, and $\mathbf{X}$ the $T \times n$ matrix of regressors. Assume further that $\mathbf{X}$ contains $r \ll n$ unobserved common factors, $\mathbf{F}$, satisfying the identifying restriction $\mathbf{E}[(\mathbf{F}'\mathbf{F})] = I_r$ and that $k \leq r$ linear combinations of these, $\mathbf{Z} = \mathbf{F}\tilde{A}$, contain information about $\mathbf{Y}$ while the residual, $\mathbf{E}$, is uninformative:*

$$\begin{aligned} \mathbf{X} &= \mathbf{F}\Lambda + \mathbf{E}, \quad \Lambda \in \mathbb{R}^{r \times n} && (27) \\ \mathbf{Y} &= \mathbf{F}\tilde{A}B + \mathbf{e}\ , \quad \tilde{A} \in \mathbb{R}^{r \times k}, B \in \mathbb{R}^{k \times n} && (28) \end{aligned}$$

where $\mathbf{F}$, $\mathbf{E}$, $\mathbf{e}$ are mutually independent and $\tilde{A}B$ is an $r \times m$ matrix of reduced rank $k \leq r$.

In this setting, the factors, $\mathbf{F}$, extracted from (27) play the role of the regressors in a standard reduced rank regression (28). However, whereas the reduced rank model (**??**) is ill-conditioned, the two-step procedure (27)-(28) yields a numerically well-behaved lower dimensional system that is equivalent to the (one-step) RRRR estimator with spectral truncation given in Corollary 2.

The covariance structure of the factor model (27) is given by

$$S_{XX} = \Lambda'\Lambda + \Omega_n \tag{29}$$

[21] See e.g. Burman, Chow, and Nolan (1994) and Racine (2000) for a discussion of this issue in the context of h-block and hv-block cross-validation.

where $\Omega_n = \frac{1}{T}\mathbf{E}'\mathbf{E}$. In the classical factor model setting, where n is fixed while $T \to \infty$, the estimated eigenvalues of S_{XX} will tend to their population values (and satisfy a standard $\sqrt{T}$ CLT) so that the identification of factors versus noise components becomes trivial in the limit for a given parametric assumption about the noise. However, in large panels, when both $n, T \to \infty$, the task of identifying "significant" eigenvalues is much more difficult since in this case the spectrum can typically not be consistently estimated and there will often be a lack of a clear "gap" in the empirically observed spectrum, as observed in some of the empirical applications studied in Section 7 below (c.f. Figure A.1).

The large n, T asymptotics we shall work with requires more structure than simply that $\min(n, T) \to \infty$, as in Bai and Ng (2002), and involves keeping the ratio n/T fixed (at least in the limit). In applications, this distinction is of course unimportant. The ratio n/T turns out to play an important role in controlling the distribution of the spectrum of the covariance matrix of the observed data. Our maintained assumption throughout will be:

Assumption 2 (Large n, T limit)**.** *$n, T \to \infty$ with $n/T \to \gamma \in (0; 1)$.*[22]

The high level regularity assumptions we shall need for the identification of factors in (27) are quite strong, but necessary to pin down the behavior of the spectrum of the covariance matrix S_{XX}. Loosely speaking, the assumptions ensure that the noise is "orthogonal" to the signal and that the signal (as conveyed through the smallest eigenvalue of $\Lambda'\Lambda$ matrix) is sufficiently strong compared to the noise (as conveyed through the largest eigenvalue of Ω). In general, weaker assumptions about the signal will necessitate stronger assumptions about the noise and vice versa.

In order to discuss convergence of a sequence of real symmetric random matrices that live on spaces of increasing dimension, we recall the Spectral Representation Theorem for self-adjoint operators, which states that the behavior of such operators (up to a rotation) is completely described by their spectral density. For the $n \times n$ real symmetric random matrix Ω_n with eigenvalues $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_n$, the empirical spectral density (ESD) is defined as the measure on the real line with density (where $\delta(\cdot)$ is the Dirac delta)

$$\mu_{\Omega_n}(x) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \delta(x - \omega_i) \tag{30}$$

Clearly, μ_{Ω_n} is a random probability measure and we shall say that the sequence of random matrices $\{\Omega_n\}$ converges almost surely if the sequence of random measures μ_{Ω_n} converges weakly almost surely.

Assumption 3 (Simple Noise Structure)**.** *The matrix $\mathbf{E} = \{\tilde{e}_{ij}\} \in \mathbb{R}^{T \times n}$ consists of i.i.d. mean zero, variance 1 entries with $E|e_{ij}|^4 < \infty$.*

[22] Since the eigenvalues of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ are the same up to $\max(n, T) - \min(n, T)$ zero eigenvalues, the $\gamma > 1$ case is a trivial extension in which the spectrum has a point mass of $1 - 1/\gamma$ is placed at zero.

The i.i.d. assumption is of course very strong but note that we only require finite fourth rather than eighth moments as is common in the approximate factor model literature (c.f. e.g. Bai and Ng (2002)). As we shall see in Section 6.1.2 below, the assumption appears to lead to a reasonable description of the behavior of the data.

The following classic theorem provides the limiting behavior of the noise as a function of $\gamma = n/T$:

Theorem 1 (Marcenko and Pastur (1967))**.** *Under Assumptions 2 and 3, the limiting distribution of the spectrum of $\Omega_n = \frac{1}{T}\mathbf{E}'\mathbf{E}$ is given by the measure μ_Ω with support on the interval $[(1-\sqrt{\gamma})^2; (1+\sqrt{\gamma})^2]$ and density:*

$$d\mu_\Omega(x) = \frac{\sqrt{(x-b_-)(b_+-x)}}{2\pi\gamma x}\,dx \ , \ where\ b_\pm = (1 \pm \sqrt{\gamma})^2 \tag{31}$$

Theorem 1 tells us that, although the population eigenvalues of Ω are all equal to 1, the empirical spectrum will asymptotically be distributed over the interval $[(1-\sqrt{\gamma})^2; (1+\sqrt{\gamma})^2]$ and the largest eigenvalue of Ω satisfies $\omega_1 \xrightarrow{p} (1+\sqrt{\gamma})^2 > 1$ as $n, T \to \infty$. Even this simple setting therefore tells us that our ability to identify "large" eigenvalues indicating a potential factor structure in the data, depends on γ: A "true" outlying eigenvalue associated with a weak signal may be absorbed in the bulk of the spectrum (which in the worst case scenario spans the interval $[0; 4]$ when $n = T$).

A simple condition which ensures the correct identification (asymptotically) of the correct number of factors is that additional regressors add substantial new information about the factors in the sense that the eigenvalues of $\Lambda'\Lambda$ diverge as $n \to \infty$.

Assumption 4 (Strong Factor Signal)**.** *Let the r non-zero eigenvalues of $\Lambda'\Lambda$ be given by $\lambda_1 \geq \cdots \geq \lambda_r$, where $\lim_{n\to\infty} \lambda_r = +\infty$.*

Under the additional assumptions above, Weyl's inequality (c.f. Lemma 1) guarantees that exactly r eigenvalues of S_{XX} will diverge in the large n, T limit and therefore the dimension of the factor structure is identified:

Proposition 3 (Spectral Rank Identification)**.** *Under the assumptions of Theorem 1 and if additionally Assumption 4 is satisfied, then exactly r eigenvalues of S_{XX} diverge as $n, T \to \infty$ and, for $i \geq r$, $p\lim_{n,T\to\infty} \sigma_i \leq (1+\sqrt{\gamma})^2$.*

6.1.1. Application of Random Matrix Theory to Regularization

Under the assumption that a factor structure holds in the data, e.g. Assumption 1, the theory presented in the preceding section suggests that the spectral truncation parameter should be chosen to retain only eigenvalues which cannot be attributed to the presence of noise. In particular, under the assumption of i.i.d noise, any eigenvalue above $(1+\sqrt{\gamma})^2$ should be retained in the limit.

The largest eigenvalue of the noise covariance matrix, Ω_n converges to $(1+\sqrt{\gamma})^2$, but how fast is the convergence and what is the asymptotic distribution of the largest eigenvalue? Theorem 2, due to Johnstone (2001), shows that the suitably normalized largest eigenvalue has a limiting Tracy-Widom distribution, denoted TW_1, and shown in Figure A.4:

Theorem 2 (Johnstone (2001))**.** *Under Assumptions 2 and 3, the largest eigenvalue, ω_1, of the noise covariance matrix, Ω_n, satisfies the CLT*

$$\frac{\omega_1 - \mu_{n,T}}{\sigma_{n,T}} \xrightarrow{\mathcal{D}} TW_1 \tag{32}$$

where TW_1 is the Tracy-Widom distribution corresponding to the first β-ensemble (c.f. Figure A.4):

$$\mu_{n,T} = \frac{(\sqrt{n}+\sqrt{T-1})^2}{T} \approx (1+\sqrt{\gamma})^2 \tag{33}$$

$$\sigma_{n,t} = \frac{(\sqrt{n}+\sqrt{T-1})}{T}\left[\frac{1}{\sqrt{T-1}}+\frac{1}{\sqrt{n}}\right]^{1/3} \approx T^{-2/3}(1+\sqrt{\gamma})^{4/3}\gamma^{-1/6} \tag{34}$$

Based on the Theorem, we see that the convergence of the maximum eigenvalue of S_{XX} under the null of no factor structure is quite rapid, and a cutoff value for the spectral truncation can be chosen based on a suitable percentile of the limiting distribution. In the empirical section below, we denote this data driven version of spectral truncation regularization by “SMP”.

In finite samples, there may be significant uncertainty around this signal-noise cut-off and the Tikhonov scheme provides a way of down-weighting eigenvalues below the cut-off while maintaining a large weight on eigenvalues above the cut-off. To be specific, we chose ρ such that an eigenvalue at the chosen Tracy-Widom quantile receives a weight of $\frac{1}{2}$, and denote this data driven regularization scheme by “TMP” in the empirical applications below.

6.1.2. Empirical Spectra

In section 7 below we consider forecasting based on a number of macro economic and yield curve datasets with different panel sizes. The question we ask here is, how well do these panels conform to the asymptotic random matrix theory laid out above? In panel (a) of Figure A.5 we show the empirical spectrum of the panel of 108 macro time series considered in Stock and Watson (2011), corresponding to a panel size of $n = 108, T = 195$, along with the limiting distribution of Theorem 1.[23] Clearly, the match is quite poor, with the Stock and Watson (2011) data displaying several large eigenvalues. In particular, there are around 6 eigenvalues that are clearly separated from the bulk spectrum. The observed deviation from the asymptotic theory could of course be due to several reasons: A finite sample phenomenon, serial dependence, and cross sectional dependence. To investigate this further, panel (b) shows the effect of applying

[23] Note that $T = 195$ considered here corresponds to the full sample. In our rolling out-of-sample exercises we have $T = 100$.

an AR(12) filter to the X, so that any serial dependence is vastly reduced, while the cross-sectional dependence is mainly left intact. This apparently has little or no effect on the empirical spectrum and we therefore can rule out serial dependence as a likely cause. Finally, in panel (c) we generate 10,000 synthetic panels by *independently* reshuffling the time indices of each individual time series, thereby breaking both time series as well as cross-sectional dependence in the data. Here we see a perfect match with the asymptotic theory and therefore conclude that the observed deviation in panel (a) was primarily due to cross-sectional dependence (i.e. factor structure) and not a finite sample phenomenon, nor due to serial dependence in the data.

The yield curve data sets correspond to much smaller panels with $n = 15, T = 468$ for which the asymptotic theory might be expected to be less in accordance with the observed finite sample behavior. Panel (1a)-(2a) of Figures A.6&A.7 show the empirical spectra for the four bond excess return predictor panels. The deviations from the asymptotic theory are substantial with at least one large outlier. In panels (1b)-(2b) we show that serial correlation does not appear to be driving this result. Finally, in panels (1c)-(2c), we see that, reshuffling the time index of each series independently, almost restores adherence to the asymptotic theory although minor deviations persist, likely due to the relatively small panel size of only 15 eigenvalues.

7. Empirical Applications

We illustrate the empirical performance of the proposed family of regularized reduced rank regression (RRRR) models, relative to a number of existing alternative models, when applied to the following standard forecasting problems in macroeconomics and finance: (i) forecasting a large set of macroeconomic series as in Stock and Watson (2011); (ii) forecasting a small set of bond excess return series as in Cochrane and Piazzesi (2005) and Cieslak and Povala (2011). In each application we explicitly account for model parsimony (Occam's razor) as given by the number of forecasting factors used for predicting all m outcomes.

7.1. Model taxonomy

The two types of shrinkage employed in our RRRR modeling approach lead to a natural model taxonomy in terms of number of forecasting factors and regressor components. Our taxonomy table A.1 summarizes all models considered in the empirical illustrations.

First, Panel A in Table A.1 depicts models based on a fixed number of regressor components with the r-th row ($r = 1, 2, ..., n$) and k-th column ($k = 1, 2, ..., \min(r, m-1)$ and $k = m$) corresponding to models with r regressor components and k forecasting factors. In particular, for $k = 1, 2, ..., \min(r, m-1)$ we denote as RRRRk-PCr our regularized reduced rank regression model with k forecasting factors and r principal components obtained via the fixed spectral truncation cutoff r in section 4.2 above. As indicated on the main diagonal of the table, for $k = r$ this is simply equivalent to principal component regression with r factors, denoted PCR-r, while the bottom right corner of the table corresponding to $r = n$ and $k = m$ represents OLS.

Finally, in the last column of the table, for $k = m$, we consider alternative methods that do not impose a smaller common set of forecasting factors across the m outcomes. For $r = 1, 2, ..., n-1$ these comprise partial least squares with r automatic regressor components denoted as PLS-r, the three-pass regression filter with r automatic regressor components denoted as 3PRF-r, as well as a version of ridge regression using spectral truncation with r principle components denoted as RR-r.[24,25]

Next, Panel B in Table A.1 presents models relying on data driven regularization of the regressor components stemming from our random matrix theory results in section 6.1. In particular, in column k of the table, for $k < m$, RRRRk-SMP (row 1) and RRRRk-TMP (row 2) stand for our regularized reduced rank regression model with k forecasting factors in which the number of regressor components is determined by proposition 3, utilizing respectively spectral truncation (SMP) or Tikhonov (TMP) schemes as detailed is Section 6.1.1. We further impose the natural restriction that the chosen number of regressor components is not smaller than k, which is the minimum number of components required to span k forecasting factors of full rank.[26] Finally, the last column in Panel B of the table for $k = m$ displays the corresponding models with no reduction in the number of forecasting factors, denoted as RR-SMP (row 1) and RR-TMP (row 2), which stand for ridge regression with the respective data driven regularization approaches.

We rely on the above model taxonomy in our empirical illustrations and compare the forecasting performance of various RRRR and RR models to OLS, PCR, PLS and 3PRF as relevant alternatives. Our primary focus in what follows is on the more interesting set of parsimonious RRRR models with $k << m$, which allows us to study the extent to which just a few common factors may jointly be able to forecast multiple variables of interest.

7.2. Forecasting macroeconomic series

In our first empirical illustration we consider the 35 aggregate and 108 disaggregate quarterly U.S. macroeconomic series analyzed by Stock and Watson (2011), with a total of 195 quarterly observations from 1960:Q2 through 2008:Q4. After transforming and categorizing each series, we produce rolling out-of-sample one-step-ahead forecasts with rolling window size 100 quarters for various models in our taxonomy table A.1.[27] Following Stock and Watson (2011), we report distributions of relative RMSE by forecasting method relative to the PCR-5 benchmark. Table A.2 summarizes the results when forecasting the entire set of 143 macroeconomic variables univariately without imposing a common factor structure as in Stock and Watson (2011), while

[24] Note that RR-r can also be defined as RRRR1-PCr when applied to forecast each outcome univariately in isolation from the rest of the outcomes rather than jointly.

[25] The PLS and 3PRF estimators referred to throughout are implemented using the MATLAB code accompanying Kelly and Pruitt (2011) which also introduces the “automatic” regressor terminology.

[26] The same restriction explains the lower triangular structure in Panel A of Table A.1 for the RRRR models with a fixed number of regressor components.

[27] We thank Mark Watson for making the Gauss programs for replicating Stock and Watson (2011) available.

Table A.3 contains results for the more interesting case when forecasting the subset of 35 aggregate macroeconomic variables by imposing common forecasting factors. The predictors in both cases comprise the subset of 108 disaggregate macroeconomic series. The tables present both percentiles and empirical probabilities for intervals chosen to highlight any substantial downward/upward deviations from a ratio equal to 1, indicating better/worse performance relative to the PCR-5 benchmark.

As a natural starting point, Table A.2, Panel A reports results for the AR-4 and PCR-50 "naive" benchmark models considered also by Stock and Watson (2011). In particular, the obtained percentiles coincide with those reported by Stock and Watson (2011) for the same "naive" models.[28] Such exact match allows for meaningful comparison between the performance of the rest of the models we present in Table A.2, Panels B, C, D, E to the performance of the other shrinkage models considered by Stock and Watson (2011) but not implemented here.[29] Overall, our findings are in line with the main conclusion in Stock and Watson (2011), that PCR-5 is hard to outperform consistently across all 143 series. Moreover, any improvement in the left tail of the distribution is more than offset by a deterioration in the right tail, keeping the median roughly equal to 1 at best. The only notable competitor to PCR-5 appears to be our RR-SMP model exploiting the random matrix theory results in section 6.1. As evident from the first row in Panel C of Table A.2, RR-SMP attains a slightly better left tail without any significant distortion in the right tail. A closer look at the distribution of the spectral truncation cutoff implied by our MP (data driven) method reveals that it has a median of 5 and varies only mildly from 3 to 8 across different series and time windows. This provides a compelling rationale for why PCR-5 emerges as a hard to beat benchmark in Stock and Watson (2011), leaving only modest room for improvement by suitable data driven procedures for determining the degree of regularization. As such, our RR-SMP model appears to be the only viable competitor to PCR-5 in terms of overall performance across all macro series among the shrinkage methods considered in this paper and in Stock and Watson (2011), as well as the recently proposed 3PRF models and its closely related PLS counterparts. We attribute the success of RR-SMP to the reasonably good finite sample validity of our random matrix theory results for the considered macroeconomic series.

We next consider the more restricted problem of jointly forecasting all 35 aggregate macroeconomic series with a common smaller set of factors extracted from the 108 disaggregate series. Table A.3 presents results for the distribution of RMSE relative to the PCR-5 benchmark for models grouped by number of forecasting factors set to 1 (panel A), 3 (Panel B), 5 (Panel C), 7 (Panel D), and 35 (Panel E). It is quite striking to observe that now a viable competitor to PCR-5 is delivered by RRRR5-SMP, performing essentially on par with RR-SMP and outperforming 3PRF and PLS, none of which imposes common factor structure. By contrast, our approach to

[28] See table 5, panel (a) in Stock and Watson (2011) in whose notation PCR-50 is denoted as OLS.

[29] See again table 5, panel (a) in Stock and Watson (2011).

combine the two types of shrinkage in a way that disentangles the degree of regularization of the predictors from the number of factors that explain the outcomes offers a viable parsimonious alternative to PCR-5. This finding should be of great interest to empirical macro economists in the construction of VAR models.

Finally, it is interesting to observe that there appears to be marked difference in the out-of-sample forecasting performance of the spectral (SMP) and Tikhonov (TMP) regularization schemes in the considered data driven versions of our RRRR and RR models. The distribution of relative RMSE vis-a-vis the PCR-5 benchmark reported in tables A.2 and A.3 reveals that overall, across the considered large set of macroeconomic series, spectral truncation is generally more preferable than Tikhonov regularization. In this regard, our results can be related to Mol et al. (2008) who use ridge regression with Tikhonov regularization in a Bayesian framework to forecast industrial production and inflation and provide a set of comparisons indicating that different PCR benchmarks (and PCR-5 in particular) are hard to beat in terms of relative RMSE using ridge regression. Using the much larger set of macroeconomic series studied by Stock and Watson (2011), we find that a similar result holds for our data-driven RR-TMP and RRRR-TMP models relying on Tikhonov regularization. By contrast, the spectral truncation regularization that we utilize in our RR-SMP and RRRR-SMP models appears to offer a viable data-driven alternative to the PCR-5 benchmark.

7.3. Forecasting bond excess returns

There are numerous examples in the finance literature where it is natural to think that a small number of forecasting factors drive multiple outcomes and hence our RRRR models are a particularly relevant forecasting approach. As an illustration we consider forecasting bond excess returns, known to be largely driven by a single common forecasting factor constructed differently by Cochrane and Piazzesi (2005) from forward rates and more recently by Cieslak and Povala (2011) from maturity-related inflation cycles. For the period 1972-2010 we produce rolling out-of-sample forecasts with rolling window size 120 months for five different sets of predictors: (i) cycles (table A.4); (ii) forwards (table A.5); (iii) forward slopes (table A.6); (iv) yield curve slopes (table A.7); (v) the union of cycles and yield curve slopes (table A.8).[30] Although there are only about 15 predictors, the design matrix, $\mathbf{X}$, is extremely ill-conditioned as shown in Figure A.1, thus necessitating the use of regularization.

For each set of predictors constructed from zero-coupon bonds with maturities from 1 to 15 years we forecast monthly bond excess returns for maturities ranging from 2 to 15 years and report out-of-sample R^2 by forecasting method relative to a rolling average benchmark. Our data source is the commonly used Gürkaynak, Sack, and Wright (2006) set of zero coupon yields (GSW), maintained and made publicly available by the Federal Reserve Board. As noted by Gürkaynak et al. (2006), the short end of the yield curve for maturities below 1 year is not

[30] Results for other possible combinations of predictors are available upon request.

reliably interpolated. Therefore, we construct forwards and cycles without utilizing GSW data for maturities shorter than 1 year, while in our set of yield curve slopes we instead opt to include the 1-month and 3-month T-bill rate from the CRSP Fama Risk-Free Rates Database.[31] The 1-month T-bill rate plays the role of the risk free rate that we use to construct monthly bond excess returns. Thus, the part of our empirical analysis based on forwards, forward slopes, and cycles complements Cochrane and Piazzesi (2005) and Cieslak and Povala (2011) by considering non-overlapping monthly bond excess returns in a rolling out-of-sample forecast exercise rather than in-sample analysis of 12-month overlapping bond excess returns. Moreover, using our RRRR methods we document non-trivial predictive power of the yield curve slopes (even more so when combined with cycles) for the monthly non-overlapping excess returns in our sample.

Our main findings from the bond data analysis can be summarized as follows. First, our regularized reduced rank regression models imposing common forecasting factors are always among the best performing models for each set of predictors. Second, we document that the predictive power of yield curve slopes (table A.7) is as strong as the predictive power of cycles (table A.4), while forward slopes (table A.6) and forwards (table A.5) in particular have markedly lower predictive power. Third, and most important of all, we document that combining yield curve slopes and cycles as predictors almost doubles the out-of-sample predictive power of the regressions for the longest maturities and our RRRR1-PC5 regularized reduced rank regression model clearly outperforms the rest of the methods in this case (table A.8), while RRRR1-SMP remains a close competitor among the data-driven methods for choosing the degree of regularization. Overall, our results make a strong case for using our regularized reduced rank models for forecasting bond excess returns which enable the extraction of predictive information from the combination of multiple (possibly extremely ill-conditioned) predictor sets.

Comparing the spectral (SMP) and Tikhonov (TMP) regularization schemes across the macro and bond applications, it can be observed that no one scheme uniformly dominates in terms of forecasting performance. Instead the appropriate choice appears to depend on the spectral properties of the data and (likely) the panel size. In particular, in the macro data (large n), eigenvalues tend to be relatively closely spaced around the MP cutoff and SMP clearly out-performs TMP. Comparing the filter factors (c.f. Figures A.2-A.3), the Tikhonov scheme assigns non-trivial weight to a great many (possibly noisy) eigenvalues while the spectral truncation scheme leads to a much simpler factor structure of the regularized regressors. By contrast, TMP outperforms SMP in the bond data applications (small n), where the spacing of eigenvalues around the MP cut-off tends to be sparse leading the SMP scheme to pick just 1 or 2 factors. One possible interpretation of the performance of TMP relative to SMP is therefore that a few of the eigenvalues just below the MP threshold (which would receive positive weight under TMP) contain valuable predictive information. This is consistent with the observed good performance of the less conservative fixed truncation rules such as RRRR1-PC5 in the case of

[31] Note that the corresponding monthly forward rates still cannot be constructed without interpolation.

the combined yield slopes and cycles predictor set.

8. Conclusion

We have proposed the Regularized Reduced Rank Regression (RRRR) forecasting model as a robust method for jointly forecasting multiple outcomes in situations with many predictors or nearly collinear predictors. The RRRR model combines two distinct types of shrinkage estimation (in terms of the singular values of S_{XX} and the canonical correlations) and can be derived from a penalized reduced rank regression model as the solution to a standard generalized eigenvalue problem. Analogous to the ridge regression, the penalized RRRR estimate has a natural Bayesian interpretation in terms of a Gaussian precision prior on the regression slopes. Moreover, in a purely frequentist setting, we have shown how to motivate the choice of regularization parameter using classical results from random matrix theory in the large n, T limit.

A key advantage of RRRR models over existing univariate techniques is the extraction of common predictive factors that jointly forecast the outcomes of interest. This is particularly pertinent when $\mathbf{Y}$ itself contains a strong factor structure that is forecastable. Compared to principal component regression (PCR), RRRR produces a more parsimonious forecasting model whenever some important factors in $\mathbf{X}$ are irrelevant for forecasting $\mathbf{Y}$, as clearly seen in our application to forecasting bond excess returns.

While factor models provide a convenient solution to the curse of dimensionality faced by variable selection methods, a common concern is the interpretability of purely "statistical" factors. We show how to alleviate this problem when the econometrician is able to assign (possibly non-exclusive) "group"-memberships to individual variables. In this case, a set of linear restrictions can be imposed on the factor extraction problem to ensure that each factor involves only variables that share a common group characteristic. While the total number of required forecasting factors may increase due to these restrictions, the factor interpretability is restored.

In our applications to out-of-sample forecasting of macro economic time series and bond excess returns, we find that the regularized reduced rank regression (RRRR) models are robust and offer an attractive alternative to principal component regression (PCR). In particular, they deliver more parsimonious (lower dimensional) forecasting models than competing methods when jointly predicting multiple outcomes that share a common factor structure (e.g. bond excess returns). Moreover, we show that a single factor model can almost double the predictability of one-month bond excess returns across a wide maturity range by using a set of predictors combining yield slopes and the maturity related cycles of Cieslak and Povala (2011). Furthermore, the data driven version of our models based on spectral truncation offers a formal justification why the Stock and Watson (2011) choice of five principal components is often the most suitable one when forecasting large sets of macro variables. However, we stress that no one model appears to be uniformly best in terms of out-of-sample performance across datasets and subsamples.

References

Anderson, T. W., 1951. Estimating linear restrictions on regression coefficients for multivariate normal distributions. The Annals of Mathematical Statistics 22 (3), pp. 327–351.

Bai, J., Ng, S., January 2002. Determining the number of factors in approximate factor models. Econometrica 70 (1), 191–221.

Burman, P., Chow, E., Nolan, D., 1994. A cross-validatory method for dependent data. Biometrika 81 (2), pp. 351–358.

Cieslak, A., Povala, P., 2011. Understanding bond risk premia. Tech. rep., Northwestern University.

Cochrane, J. H., Piazzesi, M., 2005. Bond risk premia. The American Economic Review 95 (1), pp. 138–160.

Doan, T., Litterman, R., Sims, C., 1984. Forecasting and conditional projection using realistic prior distributions. Econometric Reviews 3 (1), 1–100.

Gantmacher, F., 1960. Theory of Matrices. Chelsea, New York.

Geweke, J., 1996. Bayesian reduced rank regression in econometrics. Journal of Econometrics 75 (1), 121 – 146.

Gürkaynak, R. S., Sack, B., Wright, J. H., October 2006. The u.s. treasury yield curve: 1961 to the present. Tech. Rep. 28, Federal Reserve Board Finance and Economics Discussion Series.

Izenman, A. J., 1975. Reduced-rank regression for the multivariate linear model. Journal of Multivariate Analysis 5 (2), 248 – 264.

Johnstone, I. M., 2001. On the distribution of the largest eigenvalue in principal components analysis. The Annals of Statistics 29 (2), pp. 295–327.

Kelly, B., Pruitt, S., 2011. The three-pass regression filter: A new approach to forecasting using many predictors. Tech. rep., Chicago Booth.

Koop, G., Potter, S., 2004. Forecasting in dynamic factor models using bayesian model averaging. Econometrics Journal 7 (2), 550–565.

Marcenko, V. A., Pastur, L. A., 1967. Distribution of eigenvalues for some sets of random matrices. Math. USSR Sbornik 1 (4), 457–484.

Mol, C. D., Giannone, D., Reichlin, L., 2008. Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? Journal of Econometrics 146 (2), 318 – 328, ¡ce:title¿Honoring the research contributions of Charles R. Nelson¡/ce:title¿.

Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. The Review of Economics and Statistics 92 (4), pp. 1004–1016.

Patterson, N., Price, A. L., Reich, D., 12 2006. Population structure and eigenanalysis. PLoS Genet 2 (12), e190.

Racine, J., 2000. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. Journal of Econometrics 99 (1), 39 – 61.

Raftery, A. E., Madigan, D., Hoeting, J. A., 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92 (437), pp. 179–191.
URL `http://www.jstor.org/stable/2291462`

Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. Third Berkeley Symp. on Math. Statist. and Prob. 1, 197–206.

Stock, J., Watson, M., 2002. Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics 20, 147–162.

Stock, J. H., Watson, M. W., August 1998. Diffusion indexes. Working Paper 6702, National Bureau of Economic Research.
URL `http://www.nber.org/papers/w6702`

Stock, J. H., Watson, M. W., February 2011. Generalized shrinkage methods for forecasting using many predictors. Working paper, Princeton University.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R. B., 2001. Missing value estimation methods for dna microarrays. Bioinformatics 17 (6), 520–525.

Appendix A. Figures and Tables

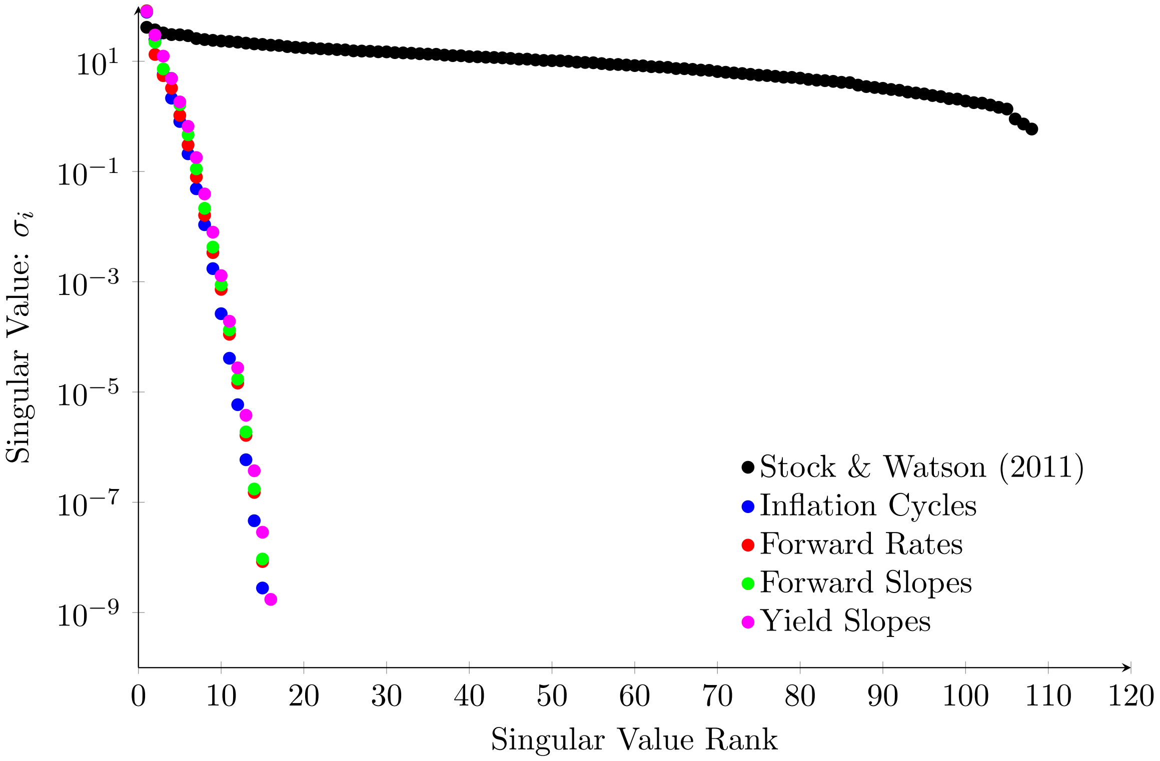


FIGURE A.1: The singular values of the Stock and Watson (2011) macro data and the four sets of bond excess return predictors considered: The Cieslak and Povala (2011) inflation cycles, the Forward rates, Forward slopes (with respect to the 1 month rate), the current yield slopes (with respect to the 1 month rate). The macro data contains 108 individual time series while the bond excess return predictors consist of 15 series each (corresponding to maturities of 1 through 15 years).

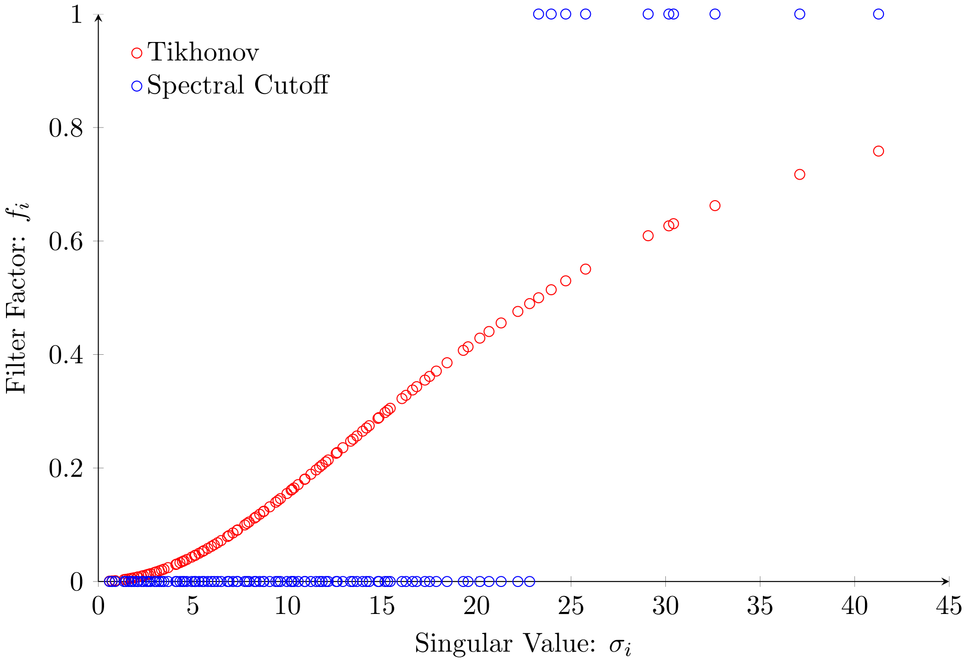


FIGURE A.2: The filter factors f_i as a function of the size of the singular value σ_i of the Stock and Watson (2011) macro data for the two regularization schemes considered. In each case the regularization parameter is set to $\rho = \sigma_{10}$, the tenth largest singular value.

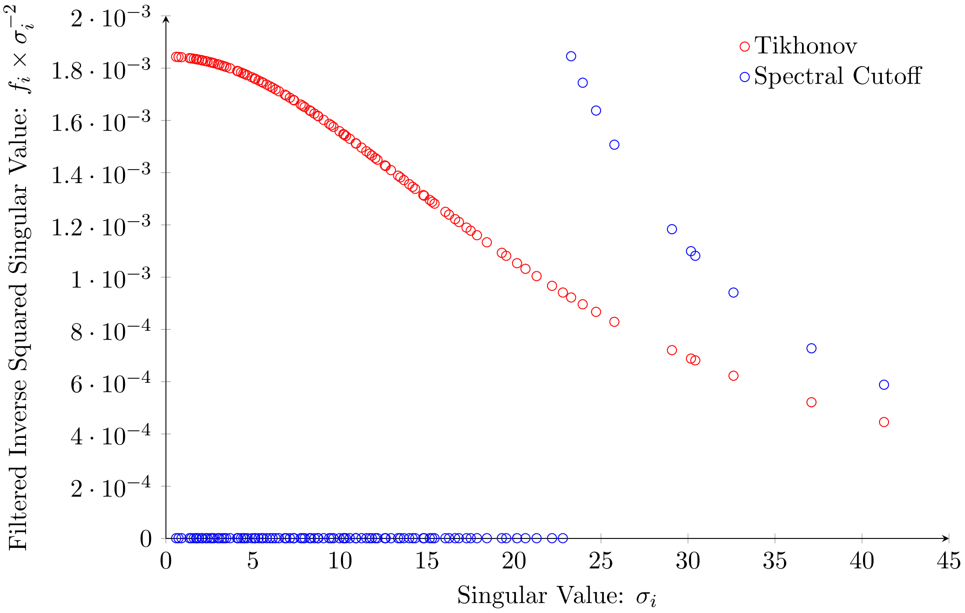


FIGURE A.3: The filtered reciprocal singular values of the Stock and Watson (2011) dataset of 108 macroeconomic variables. The spectral truncation filter works by setting all singular values of $\mathbf{X}$ that fall below a given cut-off level to zero while the Tikhonov scheme down weights small singular values. In each case the regularization parameter is set to $\rho = \sigma_{10}$, the tenth largest singular value.

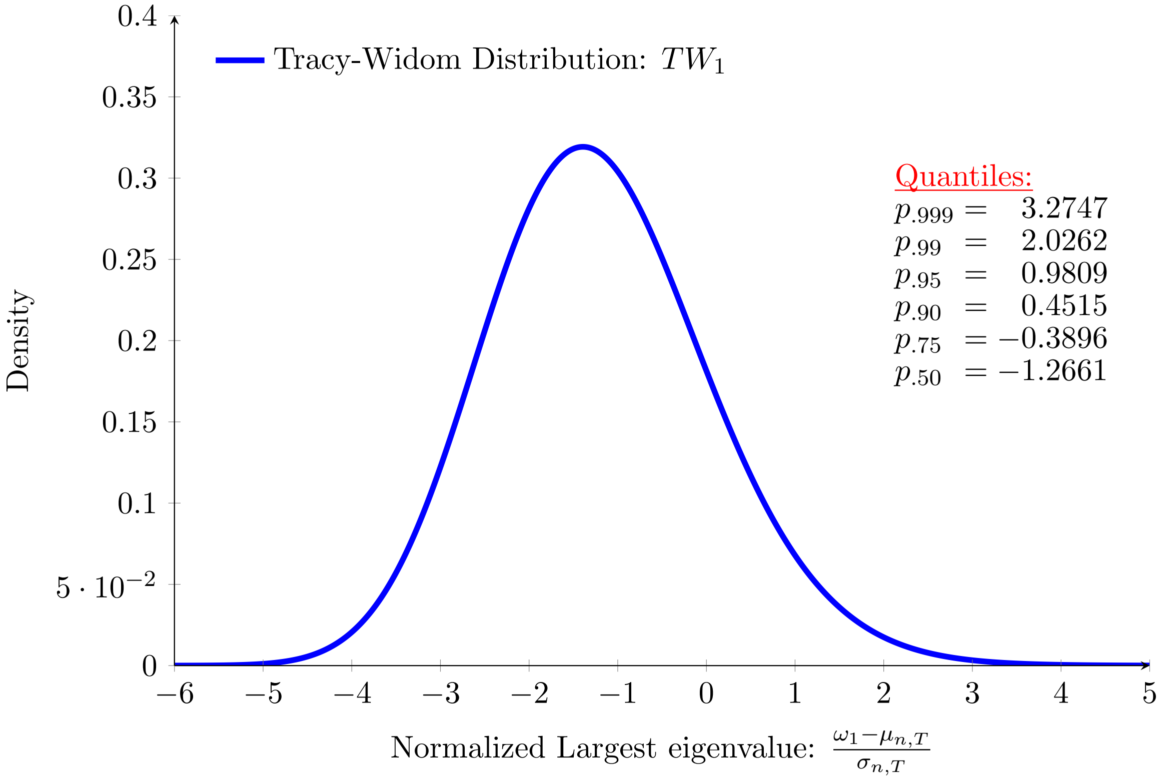
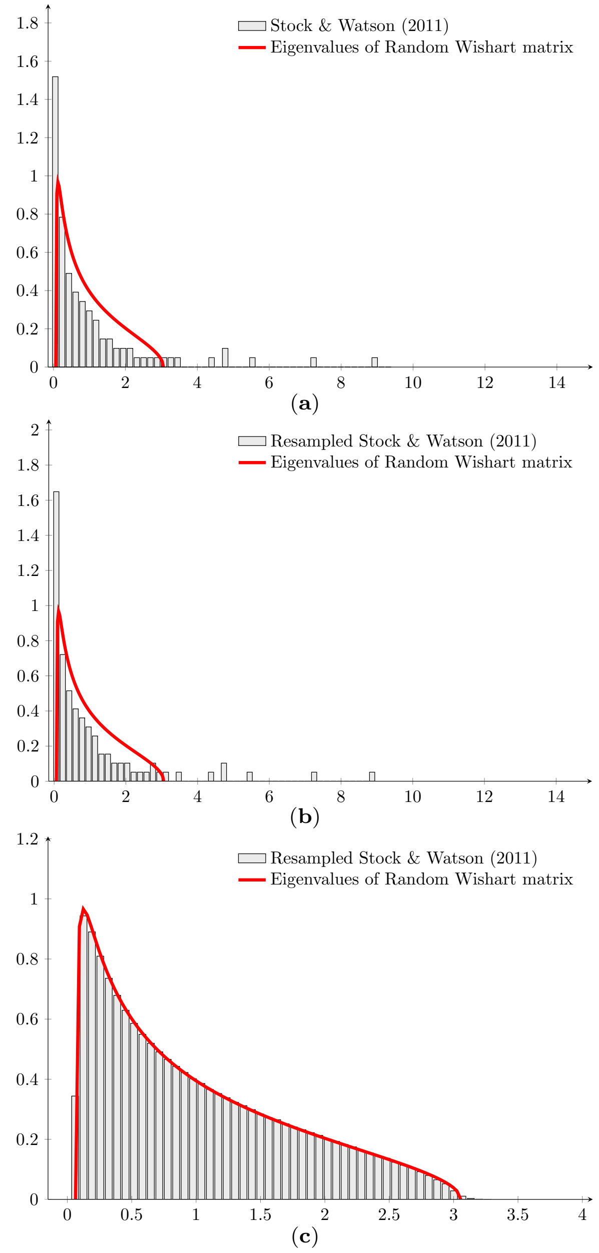


FIGURE A.4: The limiting Tracy-Widom distribution corresponding to the first β-ensemble (Gaussian Orthogonal Ensemble, c.f. Johnstone (2001)), for the normalized largest eigenvalue of the noise covariance matrix. The TW_1 distribution function is not known in closed form but given by $TW_1(s) = \exp\left\{-\frac{1}{2}\int_s^\infty q(x)\,dx\right\}$, where $q(\cdot)$ satisfies the Painleve type II equations: $q'' = xq + 2q^3$ with boundary condition $\lim_{x\to\infty}[q(x) - Ai(x)] = 0$ and $Ai(\cdot)$ is the Airy function. The solution can be found numerically to any desired accuracy using an ODE solver.

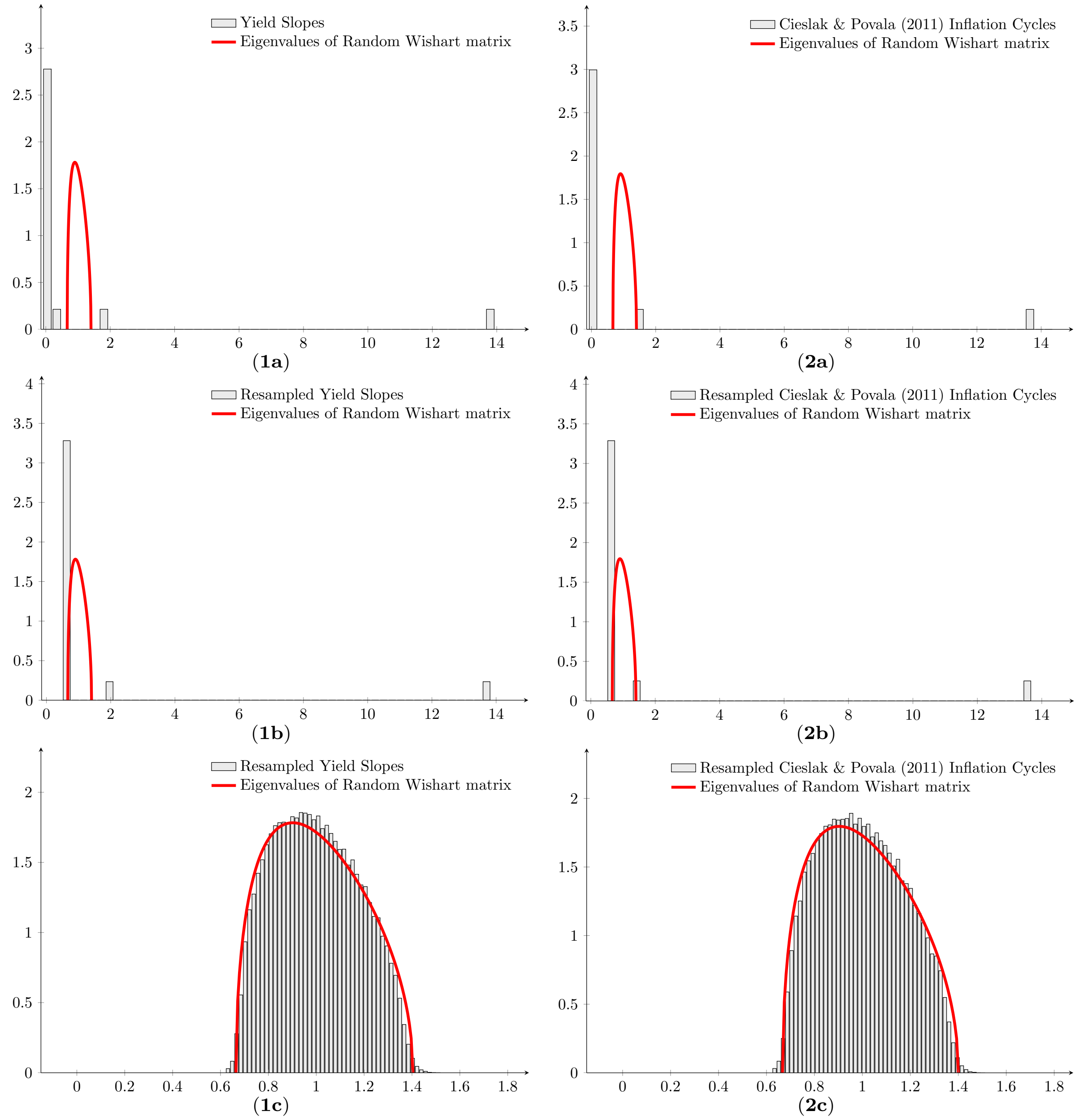


(a)

(b)

(c)

FIGURE A.5: **Eigenvalues of the Stock and Watson (2011) S_{XX} matrix.** In each panel the red curve shows the asymptotic distribution of the eigenvalues of the covariance matrix of a panel of i.i.d. N(0,1) random variables with $N/T = 108/198$ as in the Stock and Watson (2011) dataset. **Panel a:** The empirical distribution of the 108 eigenvalues of the S_{XX} matrix. **Panel (b):** The eigenvalue distribution of S_{XX} after applying an AR(12) filter to eliminate the effect of autocorrelation in the data while preserving the cross-sectional dependence. **Panel (c):** The eigenvalue distribution of S_{XX} for 10,000 resampled versions of the data in which the observation time indices have been scrambled independently for each series to eliminate the effect of both autocorrelation and cross-sectional dependence in the data.



(**1a**) (**2a**)

(**1b**) (**2b**)

(**1c**) (**2c**)

FIGURE A.6: **Eigenvalues of the S_{XX} matrix for the inflation cycle and yield slope data.** In each panel the red curve shows the asymptotic distribution of the eigenvalues of the covariance matrix of a panel of i.i.d. N(0,1) random variables with $N/T = 15/468$ as in the yield slope and inflation cycle datasets. **Panels 1a&2a:** The empirical distribution of the 15 eigenvalues of the S_{XX} matrix. **Panels 1b&2b:** The eigenvalue distribution of S_{XX} after applying an AR(12) filter to eliminate the effect of autocorrelation in the data while preserving the cross-sectional dependence. **Panels 1c&2c:** The eigenvalue distribution of S_{XX} for 10,000 resampled versions of the data in which the observation time indices have been scrambled independently for each series to eliminate the effect of both autocorrelation and cross-sectional dependence in the data.

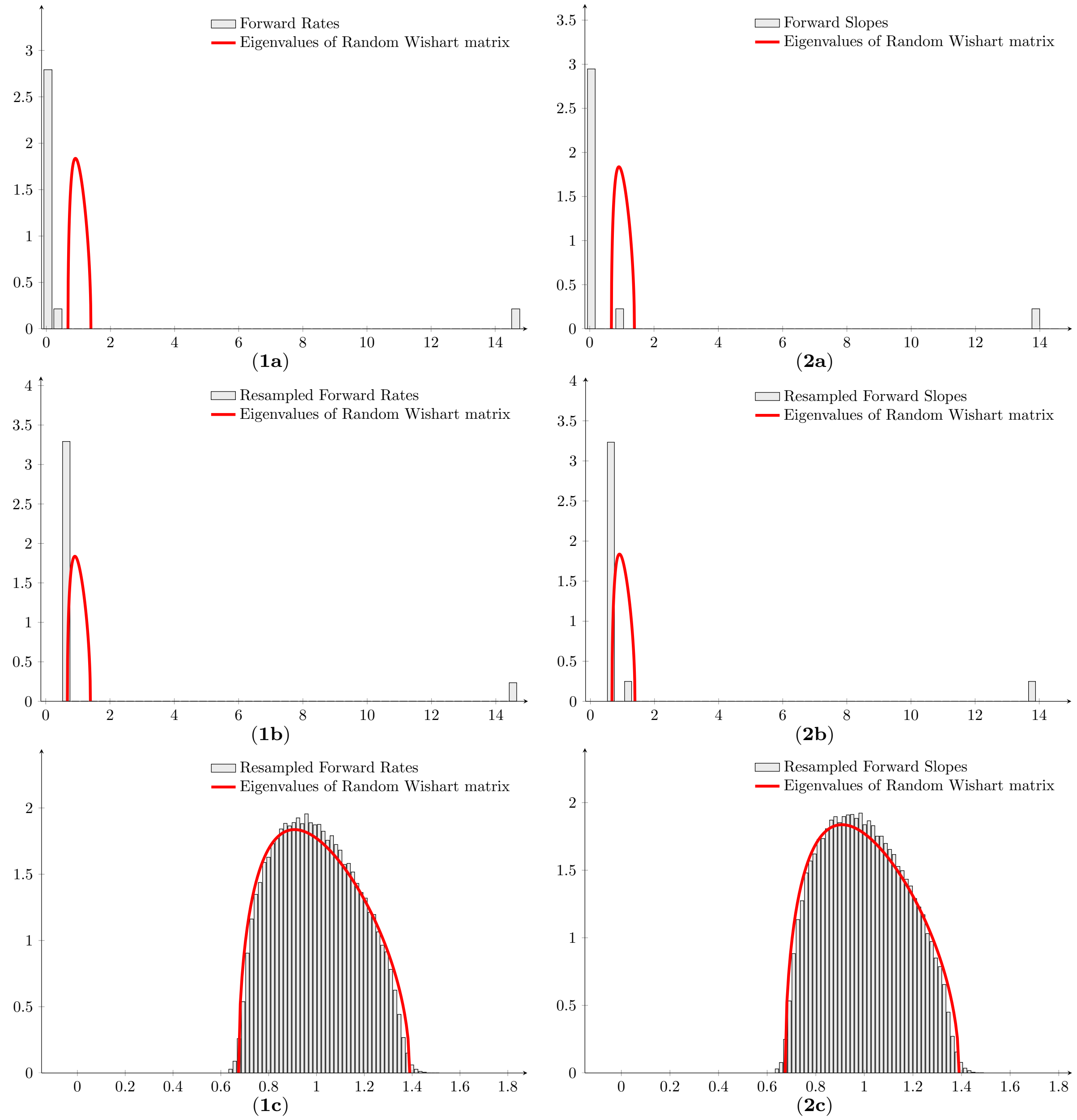
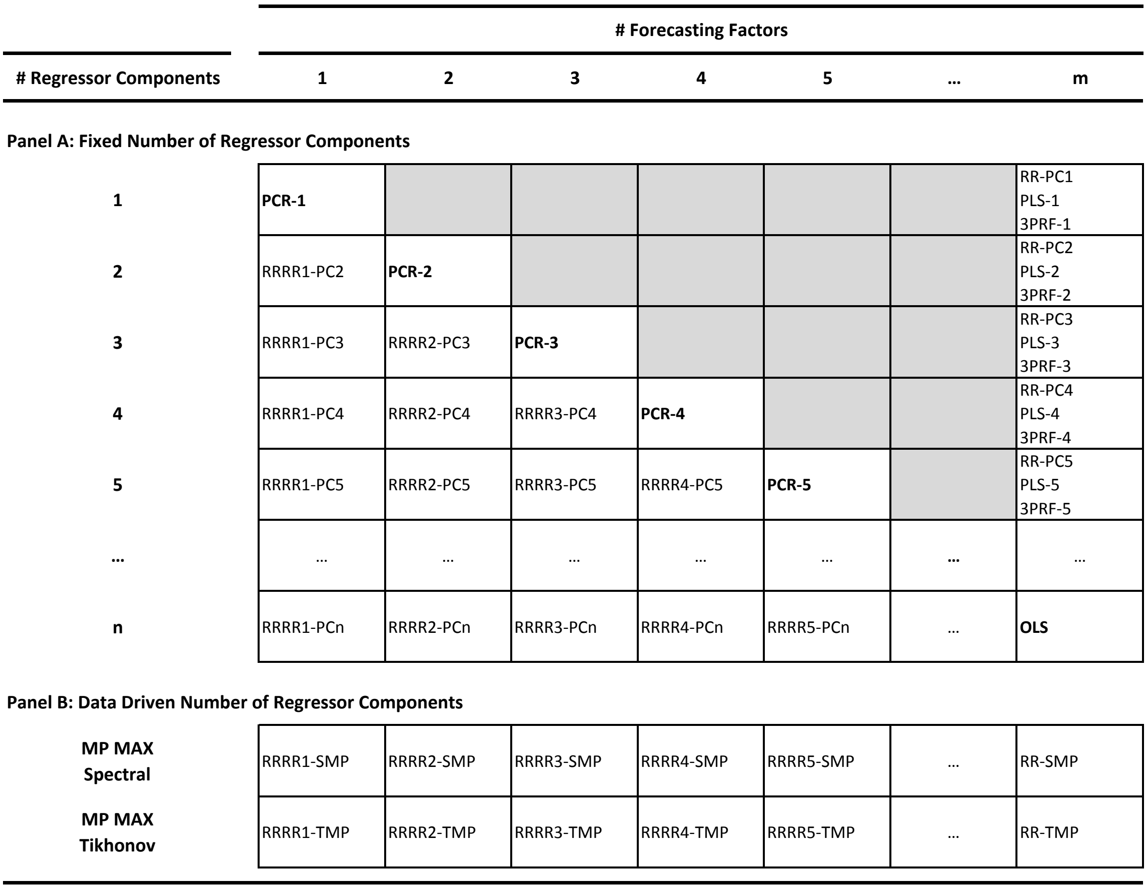


Figure A.7: **Eigenvalues of the S_{XX} matrix for the forward and forward slope data.** In each panel the red curve shows the asymptotic distribution of the eigenvalues of the covariance matrix of a panel of i.i.d. N(0,1) random variables with $N/T = 15/468$ as in the yield slope and inflation cycle datasets. **Panels 1a&2a:** The empirical distribution of the 15 eigenvalues of the S_{XX} matrix. **Panels 1b&2b:** The eigenvalue distribution of S_{XX} after applying an AR(12) filter to eliminate the effect of autocorrelation in the data while preserving the cross-sectional dependence. **Panels 1c&2c:** The eigenvalue distribution of S_{XX} for 10,000 resampled versions of the data in which the observation time indices have been scrambled independently for each series to eliminate the effect of both autocorrelation and cross-sectional dependence in the data.

Table A.1: **Taxonomy of forecasting models.** We present a taxonomy of forecasting models for any number of forecasting factors $1, 2, ..., M$ and any number of regressor components $1, 2, ..., N$. Panel A presents methods based on a fixed number of regressor components. Panel B presents methods based on a data driven number of regressor components.



	# Forecasting Factors						
# Regressor Components	**1**	**2**	**3**	**4**	**5**	**…**	**m**
Panel A: Fixed Number of Regressor Components							
1	**PCR-1**						RR-PC1 PLS-1 3PRF-1
2	RRRR1-PC2	**PCR-2**					RR-PC2 PLS-2 3PRF-2
3	RRRR1-PC3	RRRR2-PC3	**PCR-3**				RR-PC3 PLS-3 3PRF-3
4	RRRR1-PC4	RRRR2-PC4	RRRR3-PC4	**PCR-4**			RR-PC4 PLS-4 3PRF-4
5	RRRR1-PC5	RRRR2-PC5	RRRR3-PC5	RRRR4-PC5	**PCR-5**		RR-PC5 PLS-5 3PRF-5
…	…	…	…	…	…	**…**	…
n	RRRR1-PCn	RRRR2-PCn	RRRR3-PCn	RRRR4-PCn	RRRR5-PCn	…	**OLS**
Panel B: Data Driven Number of Regressor Components							
MP MAX Spectral	RRRR1-SMP	RRRR2-SMP	RRRR3-SMP	RRRR4-SMP	RRRR5-SMP	…	RR-SMP
MP MAX Tikhonov	RRRR1-TMP	RRRR2-TMP	RRRR3-TMP	RRRR4-TMP	RRRR5-TMP	…	RR-TMP

Table A.2: **Distributions of relative RMSE by forecasting method for a set of 143 macroeconomic variables from Stock & Watson (2011).** For rolling out-of-sample forecasts with rolling window size 100 quarters we report quantiles (left half of the table) and relative frequencies (right half of the table) of the empirical distributions of RMSE relative to PCR-5 by forecasting method for the set of 143 macroeconomic variables in Stock & Watson (2011). The predictors comprise 108 non-aggregate macroeconomic variables transformed in accordance with Stock & Watson (2011). Panel A represents replication check of the results for two naive benchmark models found also in Stock & Watson (2011). Panels B, C, D, and E present results for a number of competing methods described in the text and our model taxonomy table A.1.

Relative RMSE to PCR-5	**Percentiles**					**Empirical Distribution**				
Models	**5**	**25**	**50**	**75**	**95**	**<0.90**	**0.90-0.97**	**0.97-1.03**	**1.03-1.10**	**>1.10**
Panel A: Naïve benchmark models										
AR-4	0.918	0.979	1.007	1.041	1.144	0.014	0.189	0.490	0.182	0.126
PCR-50	0.968	1.061	1.110	1.179	1.281	0.007	0.056	0.091	0.273	0.573
Panel B: PCR models										
PCR-1	0.929	0.975	0.995	1.034	1.114	0.035	0.189	0.517	0.175	0.084
PCR-2	0.930	0.975	0.993	1.010	1.057	0.014	0.189	0.664	0.133	0.000
PCR-3	0.954	0.982	0.992	1.008	1.029	0.000	0.126	0.832	0.042	0.000
PCR-4	0.981	0.990	0.999	1.008	1.027	0.000	0.035	0.916	0.049	0.000
PCR-5	1.000	1.000	1.000	1.000	1.000	0.000	0.000	1.000	0.000	0.000
PCR-6	0.976	0.993	1.002	1.009	1.020	0.000	0.042	0.937	0.021	0.000
PCR-7	0.973	0.995	1.005	1.017	1.042	0.000	0.021	0.846	0.133	0.000
Panel C: RR models										
RR-SMP	0.977	0.990	0.996	1.003	1.013	0.000	0.028	0.965	0.007	0.000
RR-TMP	0.975	1.026	1.069	1.111	1.187	0.000	0.042	0.252	0.413	0.294
Panel D: PLS models										
PLS-1	0.950	0.987	1.009	1.035	1.087	0.000	0.133	0.594	0.224	0.049
PLS-2	0.976	1.038	1.082	1.130	1.271	0.000	0.021	0.196	0.406	0.378
PLS-3	1.019	1.088	1.153	1.234	1.422	0.000	0.000	0.063	0.217	0.720
PLS-4	1.046	1.143	1.228	1.324	1.609	0.000	0.000	0.028	0.098	0.874
PLS-5	1.086	1.207	1.301	1.428	1.733	0.000	0.000	0.007	0.063	0.930
PLS-6	1.123	1.261	1.363	1.519	1.841	0.000	0.000	0.000	0.035	0.965
PLS-7	1.130	1.309	1.420	1.606	1.906	0.000	0.000	0.000	0.007	0.993
Panel E: 3PRF models										
3PRF-1	0.947	0.980	1.002	1.026	1.081	0.000	0.147	0.629	0.203	0.021
3PRF-2	0.979	1.020	1.060	1.103	1.239	0.000	0.035	0.273	0.427	0.266
3PRF-3	1.010	1.080	1.144	1.229	1.424	0.000	0.007	0.084	0.231	0.678
3PRF-4	1.035	1.135	1.225	1.323	1.601	0.000	0.000	0.049	0.091	0.860
3PRF-5	1.070	1.198	1.302	1.426	1.726	0.000	0.000	0.007	0.063	0.930
3PRF-6	1.126	1.258	1.368	1.514	1.853	0.000	0.000	0.000	0.042	0.958
3PRF-7	1.140	1.307	1.420	1.585	1.914	0.000	0.000	0.007	0.021	0.972

Table A.3: **Distributions of relative RMSE by forecasting method for a set of 35 aggregate macroeconomic variables from Stock & Watson (2011).** For rolling out-of-sample forecasts with rolling window size 100 quarters we report quantiles (left half of the table) and relative frequencies (right half of the table) of the empirical distributions of RMSE relative to PCR-5 by forecasting method for the subset of 35 aggregate macroeconomic variables in Stock & Watson (2011). The predictors comprise the remaining 108 non-aggregate macroeconomic variables. Panels A, B, C and D present results for models with, respectively, 1, 3, 5 and 7 forecasting factors. Panel E presents results for models that do not impose common forecasting factor structure across the 35 macroeconomic aggregates. Description of the models can be found in the text and in our model taxonomy table A.1.

Relative RMSE to PCR-5	**Percentiles**					**Empirical Distribution**				
Models	**5**	**25**	**50**	**75**	**95**	**<0.90**	**0.90-0.97**	**0.97-1.03**	**1.03-1.10**	**>1.10**
Panel A: Models with 1 forecasting factor										
PCR-1	0.590	0.951	1.000	1.036	1.145	0.086	0.229	0.371	0.171	0.143
RRRR1-PC2	0.561	0.935	0.999	1.013	1.087	0.086	0.200	0.543	0.143	0.029
RRRR1-PC4	0.546	0.939	1.003	1.019	1.189	0.086	0.200	0.514	0.114	0.086
RRRR1-PC6	0.535	0.950	1.003	1.026	1.201	0.057	0.229	0.486	0.143	0.086
RRRR1-PC8	0.526	0.953	1.005	1.038	1.231	0.057	0.257	0.400	0.171	0.114
RRRR1-PC10	0.523	0.940	0.997	1.023	1.188	0.114	0.200	0.514	0.086	0.086
RRRR1-PC12	0.521	0.947	0.997	1.024	1.205	0.114	0.200	0.457	0.114	0.114
RRRR1-SMP	0.534	0.939	1.004	1.028	1.241	0.086	0.229	0.457	0.143	0.086
RRRR1-TMP	0.534	0.954	1.001	1.033	1.165	0.114	0.171	0.429	0.143	0.143
Panel B: Models with 3 forecasting factors										
PCR-3	0.681	0.972	0.988	1.006	1.033	0.057	0.143	0.743	0.057	0.000
RRRR3-PC4	0.495	0.971	0.987	1.006	1.031	0.057	0.171	0.714	0.057	0.000
RRRR3-PC6	0.500	0.990	0.998	1.026	1.082	0.086	0.057	0.629	0.229	0.000
RRRR3-PC8	0.482	0.982	0.996	1.054	1.131	0.086	0.029	0.600	0.171	0.114
RRRR3-PC10	0.436	0.983	1.000	1.026	1.116	0.086	0.057	0.686	0.086	0.086
RRRR3-PC12	0.432	0.990	1.014	1.039	1.140	0.086	0.057	0.543	0.200	0.114
RRRR3-SMP	0.467	0.987	0.995	1.012	1.037	0.086	0.057	0.771	0.086	0.000
RRRR3-TMP	0.445	0.993	1.042	1.096	1.148	0.057	0.114	0.314	0.314	0.200
Panel C: Models with 5 forecasting factors										
PCR-5	1.000	1.000	1.000	1.000	1.000	0.000	0.000	1.000	0.000	0.000
RRRR5-PC6	0.476	0.983	0.993	1.003	1.022	0.057	0.114	0.829	0.000	0.000
RRRR5-PC8	0.469	0.991	0.998	1.020	1.050	0.057	0.086	0.686	0.171	0.000
RRRR5-PC10	0.407	0.965	0.988	1.006	1.035	0.057	0.257	0.600	0.086	0.000
RRRR5-PC12	0.397	0.980	0.997	1.010	1.066	0.057	0.057	0.800	0.057	0.029
RRRR5-SMP	0.470	0.989	0.997	1.000	1.013	0.057	0.029	0.914	0.000	0.000
RRRR5-TMP	0.380	0.997	1.045	1.127	1.180	0.057	0.000	0.400	0.286	0.257
Panel D: Models with 7 forecasting factors										
PCR-7	0.969	0.996	1.004	1.029	1.332	0.000	0.057	0.714	0.171	0.057
RRRR7-PC8	0.470	0.987	1.002	1.026	1.037	0.057	0.029	0.714	0.200	0.000
RRRR7-PC10	0.407	0.964	0.992	1.010	1.035	0.057	0.200	0.686	0.057	0.000
RRRR7-PC12	0.397	0.972	0.999	1.022	1.060	0.057	0.114	0.686	0.143	0.000
RRRR7-SMP	0.473	0.985	1.000	1.012	1.037	0.057	0.086	0.743	0.114	0.000
RRRR7-TMP	0.376	1.003	1.045	1.125	1.172	0.057	0.000	0.400	0.257	0.286
Panel E: Models with 35 forecasting factors										
RR-SMP	0.467	0.981	0.995	1.007	1.017	0.057	0.057	0.886	0.000	0.000
RR-TMP	0.330	0.983	1.052	1.113	1.170	0.057	0.114	0.257	0.314	0.257
PLS-1	0.472	0.965	0.995	1.016	1.046	0.114	0.171	0.571	0.143	0.000
PLS-2	0.351	0.986	1.036	1.086	1.170	0.057	0.029	0.343	0.343	0.229
PLS-3	0.291	1.052	1.153	1.277	1.375	0.057	0.029	0.114	0.086	0.714
PLS-5	0.230	1.243	1.354	1.524	1.777	0.057	0.000	0.029	0.057	0.857
PLS-7	0.246	1.354	1.514	1.785	2.131	0.057	0.000	0.000	0.057	0.886
3PRF-1	0.455	0.965	1.007	1.036	1.092	0.086	0.171	0.457	0.286	0.000
3PRF-2	0.360	1.016	1.074	1.114	1.203	0.057	0.029	0.200	0.371	0.343
3PRF-3	0.309	1.061	1.170	1.297	1.411	0.057	0.029	0.057	0.171	0.686
3PRF-5	0.245	1.241	1.369	1.545	1.790	0.057	0.000	0.029	0.057	0.857
3PRF-7	0.255	1.360	1.528	1.800	2.146	0.057	0.000	0.000	0.029	0.914

Table A.4: **Out-of-sample R^2 by forecasting method for monthly bond excess returns predicted by the maturity-related cycles of Cieslak & Povala (2011).** For rolling out-of-sample forecasts with rolling window size 120 months we report out-of-sample R^2 by forecasting method relative to a rolling average benchmark. We forecast monthly excess returns of bonds ranging from 2 to 15 years of maturity. The risk-free rate is taken to be the 1-month T-bill rate from the CRSP Fama Risk-Free Rates Database. The set of predictors includes the maturity-related cycles of Cieslak & Povala (2011) for GSW yields from 1 to 15 years. The sample period is 1972-2010. Panel A presents results for commonly used simple benchmark models. Panels B and C present results for competing models with, respectively, 1 and 2 forecasting factors. Panel D presents results for models that do not impose common forecasting factor structure across the 14 bond excess return series. Description of the models can be found in the text and in our model taxonomy table A.1.

Out-of-sample R^2	**Bond Excess Returns**													
Models	**$rx^{(2)}$**	**$rx^{(3)}$**	**$rx^{(4)}$**	**$rx^{(5)}$**	**$rx^{(6)}$**	**$rx^{(7)}$**	**$rx^{(8)}$**	**$rx^{(9)}$**	**$rx^{(10)}$**	**$rx^{(11)}$**	**$rx^{(12)}$**	**$rx^{(13)}$**	**$rx^{(14)}$**	**$rx^{(15)}$**
Panel A: Naïve benchmark models														
Rolling Average	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Random Walk	-0.514	-0.585	-0.640	-0.686	-0.728	-0.765	-0.797	-0.824	-0.846	-0.862	-0.874	-0.881	-0.883	-0.882
Panel B: Models with 1 forecasting factor														
PCR-1	0.015	0.019	0.023	0.026	0.028	0.030	0.031	0.032	0.033	0.034	0.034	0.035	0.036	0.037
RRRR1-PC2	0.021	0.026	0.030	0.034	0.038	0.041	0.044	0.047	0.050	0.052	0.053	0.054	0.055	0.056
RRRR1-PC3	0.039	0.032	0.031	0.033	0.035	0.038	0.041	0.044	0.047	0.049	0.051	0.052	0.053	0.055
RRRR1-PC4	0.037	0.031	0.032	0.035	0.040	0.044	0.048	0.051	0.054	0.056	0.057	0.058	0.059	0.060
RRRR1-PC5	-0.048	-0.037	-0.025	-0.014	-0.005	0.004	0.011	0.018	0.023	0.027	0.030	0.032	0.034	0.035
RRRR1-SMP	0.018	0.024	0.030	0.035	0.040	0.043	0.047	0.050	0.052	0.055	0.057	0.058	0.060	0.061
RRRR1-TMP	0.011	0.019	0.026	0.031	0.036	0.040	0.044	0.047	0.050	0.052	0.054	0.056	0.057	0.059
OLS with 1 cycle	0.037	0.041	0.046	0.050	0.054	0.057	0.060	0.062	0.064	0.065	0.066	0.067	0.068	0.069
Panel C: Models with 2 forecasting factors														
PCR-2	0.011	0.017	0.024	0.030	0.035	0.040	0.044	0.047	0.050	0.052	0.054	0.055	0.056	0.057
RRRR2-PC3	0.021	0.023	0.028	0.033	0.038	0.042	0.045	0.047	0.048	0.048	0.048	0.048	0.047	0.047
RRRR2-PC4	0.034	0.029	0.030	0.033	0.038	0.042	0.046	0.050	0.052	0.055	0.056	0.058	0.059	0.060
RRRR2-PC5	-0.051	-0.039	-0.025	-0.013	-0.003	0.004	0.009	0.012	0.014	0.016	0.016	0.017	0.017	0.018
RRRR2-SMP	0.006	0.014	0.022	0.029	0.034	0.039	0.044	0.047	0.050	0.052	0.054	0.055	0.056	0.056
RRRR2-TMP	0.019	0.027	0.034	0.040	0.044	0.048	0.051	0.053	0.055	0.057	0.059	0.060	0.061	0.062
OLS with 2 cycles	0.017	0.024	0.031	0.037	0.042	0.046	0.050	0.053	0.055	0.057	0.058	0.060	0.060	0.061
Panel D: Models with 14 forecasting factors														
RR-SMP	0.019	0.025	0.030	0.035	0.040	0.043	0.047	0.049	0.052	0.054	0.056	0.058	0.060	0.061
RR-TMP	0.010	0.018	0.025	0.031	0.036	0.040	0.044	0.047	0.050	0.052	0.054	0.055	0.057	0.058
PLS-1	0.014	0.020	0.025	0.029	0.032	0.035	0.037	0.039	0.040	0.042	0.043	0.045	0.046	0.047
PLS-2	0.022	0.024	0.029	0.033	0.037	0.041	0.045	0.048	0.050	0.051	0.053	0.056	0.057	0.057
PLS-3	0.019	0.024	0.035	0.042	0.044	0.045	0.045	0.045	0.047	0.052	0.056	0.060	0.061	0.061
PLS-4	-0.002	0.012	0.023	0.030	0.037	0.044	0.052	0.055	0.058	0.061	0.062	0.067	0.062	0.057
PLS-5	-0.095	-0.052	-0.027	-0.010	-0.014	-0.007	-0.001	0.000	0.001	0.004	0.007	0.013	0.019	0.026
3PRF-1	0.059	0.053	0.048	0.045	0.045	0.044	0.044	0.044	0.043	0.043	0.042	0.042	0.041	0.041
3PRF-2	0.040	0.035	0.038	0.046	0.048	0.051	0.054	0.054	0.056	0.057	0.058	0.059	0.059	0.060
3PRF-3	0.020	0.026	0.035	0.040	0.047	0.053	0.055	0.059	0.062	0.063	0.064	0.062	0.064	0.068
3PRF-4	-0.073	-0.046	-0.023	-0.014	-0.007	-0.001	0.003	0.005	0.007	0.009	0.010	0.012	0.015	0.019
3PRF-5	-0.109	-0.065	-0.046	-0.010	-0.001	0.001	-0.001	-0.003	-0.003	-0.013	-0.004	0.000	0.005	0.011
OLS with all cycles	-0.542	-0.430	-0.365	-0.328	-0.308	-0.300	-0.297	-0.298	-0.299	-0.298	-0.296	-0.293	-0.289	-0.284

Table A.5: **Out-of-sample R^2 by forecasting method for monthly bond excess returns predicted by forward rates.** For rolling out-of-sample forecasts with rolling window size 120 months we report out-of-sample R^2 by forecasting method relative to a rolling average benchmark. We forecast monthly excess returns of bonds ranging from 2 to 15 years of maturity. The risk-free rate is taken to be the 1-month T-bill rate from the CRSP Fama Risk-Free Rates Database. The set of predictors includes the GSW forward rates for maturities from 1 to 15 years. The sample period is 1972-2010. Panel A presents results for commonly used simple benchmark models. Panels B and C present results for competing models with, respectively, 1 and 2 forecasting factors. Panel D presents results for models that do not impose common forecasting factor structure across the 14 bond excess return series. Description of the models can be found in the text and in our model taxonomy table A.1.

Out-of-sample R^2	**Bond Excess Returns**													
Models	**$rx^{(2)}$**	**$rx^{(3)}$**	**$rx^{(4)}$**	**$rx^{(5)}$**	**$rx^{(6)}$**	**$rx^{(7)}$**	**$rx^{(8)}$**	**$rx^{(9)}$**	**$rx^{(10)}$**	**$rx^{(11)}$**	**$rx^{(12)}$**	**$rx^{(13)}$**	**$rx^{(14)}$**	**$rx^{(15)}$**
Panel A: Naïve benchmark models														
Rolling Average	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Random Walk	-0.514	-0.585	-0.640	-0.686	-0.728	-0.765	-0.797	-0.824	-0.846	-0.862	-0.874	-0.881	-0.883	-0.882
Panel B: Models with 1 forecasting factor														
PCR-1	-0.003	-0.010	-0.013	-0.014	-0.014	-0.013	-0.012	-0.011	-0.010	-0.009	-0.007	-0.006	-0.006	-0.005
RRRR1-PC2	-0.020	-0.024	-0.027	-0.029	-0.030	-0.031	-0.031	-0.031	-0.032	-0.032	-0.032	-0.032	-0.032	-0.032
RRRR1-PC3	-0.001	-0.002	-0.001	0.000	0.000	-0.001	-0.003	-0.006	-0.009	-0.012	-0.014	-0.017	-0.018	-0.020
RRRR1-PC4	-0.080	-0.066	-0.058	-0.053	-0.050	-0.048	-0.046	-0.045	-0.044	-0.043	-0.042	-0.042	-0.041	-0.040
RRRR1-PC5	-0.097	-0.084	-0.077	-0.073	-0.070	-0.067	-0.064	-0.062	-0.059	-0.056	-0.053	-0.049	-0.047	-0.044
RRRR1-SMP	-0.011	-0.020	-0.024	-0.026	-0.026	-0.026	-0.025	-0.023	-0.021	-0.020	-0.018	-0.017	-0.016	-0.014
RRRR1-TMP	0.001	-0.005	-0.008	-0.009	-0.010	-0.010	-0.010	-0.010	-0.010	-0.010	-0.010	-0.010	-0.010	-0.010
Panel C: Models with 2 forecasting factors														
PCR-2	-0.026	-0.034	-0.037	-0.038	-0.036	-0.034	-0.031	-0.028	-0.026	-0.023	-0.021	-0.019	-0.017	-0.016
RRRR2-PC3	-0.011	-0.010	-0.006	-0.002	0.000	-0.001	-0.004	-0.008	-0.012	-0.017	-0.020	-0.023	-0.026	-0.027
RRRR2-PC4	-0.066	-0.053	-0.045	-0.041	-0.041	-0.043	-0.048	-0.054	-0.060	-0.066	-0.070	-0.074	-0.076	-0.077
RRRR2-PC5	-0.083	-0.065	-0.054	-0.050	-0.051	-0.055	-0.062	-0.070	-0.078	-0.085	-0.091	-0.095	-0.098	-0.099
RRRR2-SMP	-0.027	-0.035	-0.039	-0.039	-0.038	-0.035	-0.033	-0.030	-0.027	-0.024	-0.022	-0.020	-0.019	-0.017
RRRR2-TMP	-0.001	-0.004	-0.001	0.004	0.007	0.009	0.009	0.008	0.006	0.004	0.002	0.000	-0.001	-0.002
Panel D: Models with 14 forecasting factors														
RR-SMP	-0.013	-0.023	-0.027	-0.028	-0.028	-0.027	-0.025	-0.023	-0.021	-0.019	-0.017	-0.015	-0.014	-0.013
RR-TMP	0.012	0.003	-0.006	-0.011	-0.012	-0.012	-0.012	-0.012	-0.012	-0.013	-0.015	-0.015	-0.011	-0.010
PLS-1	-0.002	-0.006	-0.019	-0.020	-0.019	-0.017	-0.011	-0.011	-0.010	-0.010	-0.010	-0.010	-0.005	-0.005
PLS-2	-0.051	-0.055	-0.056	-0.056	-0.054	-0.052	-0.052	-0.051	-0.051	-0.046	-0.043	-0.043	-0.041	-0.029
PLS-3	-0.086	-0.058	-0.032	-0.018	-0.012	-0.017	-0.017	-0.018	-0.017	-0.026	-0.038	-0.046	-0.051	-0.057
PLS-4	-0.102	-0.068	-0.049	-0.045	-0.044	-0.047	-0.055	-0.063	-0.073	-0.077	-0.080	-0.084	-0.085	-0.080
PLS-5	-0.128	-0.093	-0.081	-0.073	-0.066	-0.067	-0.072	-0.082	-0.092	-0.098	-0.096	-0.100	-0.102	-0.102
3PRF-1	-0.054	-0.043	-0.033	-0.026	-0.020	-0.019	-0.019	-0.020	-0.020	-0.020	-0.020	-0.022	-0.021	-0.011
3PRF-2	-0.088	-0.044	-0.014	0.001	-0.001	0.004	0.012	0.014	0.009	0.001	-0.007	-0.014	-0.024	-0.032
3PRF-3	-0.121	-0.082	-0.054	-0.040	-0.039	-0.037	-0.038	-0.043	-0.049	-0.053	-0.056	-0.059	-0.058	-0.059
3PRF-4	-0.118	-0.076	-0.056	-0.054	-0.053	-0.049	-0.055	-0.065	-0.067	-0.067	-0.067	-0.066	-0.064	-0.063
3PRF-5	-0.160	-0.124	-0.102	-0.089	-0.084	-0.081	-0.084	-0.085	-0.087	-0.086	-0.080	-0.081	-0.082	-0.082
OLS	-0.567	-0.465	-0.413	-0.387	-0.377	-0.378	-0.385	-0.395	-0.404	-0.412	-0.417	-0.421	-0.424	-0.427

Table A.6: **Out-of-sample R^2 by forecasting method for monthly bond excess returns predicted by forward slopes.** For rolling out-of-sample forecasts with rolling window size 120 months we report out-of-sample R^2 by forecasting method relative to a rolling average benchmark. We forecast monthly excess returns of bonds ranging from 2 to 15 years of maturity. The risk-free rate is taken to be the 1-month T-bill rate from the CRSP Fama Risk-Free Rates Database. The set of predictors includes the GSW forward slopes for maturities from 1 to 15 years. The sample period is 1972-2010. Panel A presents results for commonly used simple benchmark models. Panels B and C present results for competing models with, respectively, 1 and 2 forecasting factors. Panel D presents results for models that do not impose common forecasting factor structure across the 14 bond excess return series. Description of the models can be found in the text and in our model taxonomy table A.1.

Out-of-sample R^2	**Bond Excess Returns**													
Models	**$rx^{(2)}$**	**$rx^{(3)}$**	**$rx^{(4)}$**	**$rx^{(5)}$**	**$rx^{(6)}$**	**$rx^{(7)}$**	**$rx^{(8)}$**	**$rx^{(9)}$**	**$rx^{(10)}$**	**$rx^{(11)}$**	**$rx^{(12)}$**	**$rx^{(13)}$**	**$rx^{(14)}$**	**$rx^{(15)}$**
Panel A: Naïve benchmark models														
Rolling Average	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Random Walk	-0.514	-0.585	-0.640	-0.686	-0.728	-0.765	-0.797	-0.824	-0.846	-0.862	-0.874	-0.881	-0.883	-0.882
Panel B: Models with 1 forecasting factor														
PCR-1	0.005	0.006	0.008	0.010	0.012	0.014	0.015	0.016	0.017	0.018	0.018	0.019	0.019	0.019
RRRR1-PC2	0.046	0.035	0.029	0.026	0.023	0.021	0.019	0.017	0.014	0.012	0.009	0.007	0.004	0.002
RRRR1-PC3	0.069	0.058	0.054	0.052	0.050	0.047	0.044	0.041	0.037	0.033	0.029	0.026	0.022	0.019
RRRR1-PC4	-0.017	-0.009	0.002	0.010	0.017	0.021	0.023	0.024	0.023	0.023	0.021	0.020	0.018	0.017
RRRR1-PC5	-0.047	-0.041	-0.031	-0.022	-0.015	-0.010	-0.007	-0.004	-0.003	-0.002	-0.002	-0.002	-0.002	-0.003
RRRR1-SMP	0.006	0.006	0.008	0.011	0.013	0.014	0.016	0.017	0.018	0.018	0.019	0.019	0.020	0.020
RRRR1-TMP	0.044	0.039	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.036	0.035	0.035	0.034
Panel C: Models with 2 forecasting factors														
PCR-2	0.014	0.011	0.011	0.013	0.015	0.017	0.018	0.018	0.018	0.017	0.017	0.016	0.015	0.015
RRRR2-PC3	0.059	0.048	0.044	0.042	0.041	0.040	0.039	0.038	0.037	0.037	0.037	0.037	0.037	0.038
RRRR2-PC4	-0.007	-0.001	0.008	0.015	0.019	0.020	0.019	0.016	0.013	0.009	0.006	0.003	0.000	-0.003
RRRR2-PC5	-0.041	-0.030	-0.017	-0.010	-0.007	-0.008	-0.011	-0.015	-0.020	-0.024	-0.028	-0.032	-0.035	-0.037
RRRR2-SMP	0.016	0.012	0.012	0.014	0.016	0.018	0.019	0.019	0.019	0.019	0.018	0.017	0.016	0.016
RRRR2-TMP	0.044	0.039	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.036	0.035	0.035	0.034
Panel D: Models with 14 forecasting factors														
RR-SMP	0.006	0.006	0.008	0.011	0.013	0.014	0.016	0.017	0.018	0.018	0.019	0.019	0.020	0.020
RR-TMP	0.075	0.065	0.060	0.055	0.049	0.044	0.040	0.036	0.033	0.032	0.030	0.029	0.029	0.029
PLS-1	0.021	0.019	0.033	0.018	0.020	0.019	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020
PLS-2	0.024	0.025	0.026	0.027	0.027	0.027	0.028	0.028	0.028	0.029	0.030	0.031	0.033	0.035
PLS-3	0.016	0.021	0.026	0.029	0.029	0.028	0.027	0.026	0.027	0.028	0.030	0.032	0.033	0.032
PLS-4	-0.054	-0.024	-0.003	0.008	0.013	0.014	0.012	0.000	-0.005	-0.012	-0.020	-0.024	-0.024	-0.018
PLS-5	-0.089	-0.046	-0.026	-0.025	-0.030	-0.032	-0.036	-0.028	-0.028	-0.030	-0.032	-0.038	-0.041	-0.031
3PRF-1	-0.020	-0.019	-0.014	-0.013	-0.013	-0.013	-0.012	-0.013	-0.016	-0.019	-0.022	-0.024	-0.023	-0.023
3PRF-2	-0.046	-0.034	-0.022	-0.016	-0.011	-0.011	-0.012	-0.014	-0.016	-0.015	-0.013	-0.010	-0.008	-0.008
3PRF-3	-0.098	-0.063	-0.038	-0.025	-0.018	-0.020	-0.019	-0.026	-0.031	-0.037	-0.043	-0.050	-0.055	-0.056
3PRF-4	-0.110	-0.072	-0.049	-0.029	-0.029	-0.031	-0.033	-0.047	-0.061	-0.065	-0.071	-0.074	-0.080	-0.075
3PRF-5	-0.123	-0.089	-0.067	-0.055	-0.048	-0.050	-0.057	-0.064	-0.071	-0.076	-0.087	-0.084	-0.082	-0.081
OLS	-0.510	-0.397	-0.335	-0.301	-0.287	-0.284	-0.289	-0.298	-0.306	-0.314	-0.319	-0.321	-0.323	-0.323

TABLE A.7: **Out-of-sample R^2 by forecasting method for monthly bond excess returns predicted by yield curve slopes.** For rolling out-of-sample forecasts with rolling window size 120 months we report out-of-sample R^2 by forecasting method relative to a rolling average benchmark. We forecast monthly excess returns of bonds ranging from 2 to 15 years of maturity. The risk-free rate is taken to be the 1-month T-bill rate from the CRSP Fama Risk-Free Rates Database. The set of predictors includes the yield curve slopes for the 1-month and 3-month T-bill rates from the CRSP Fama Risk-Free Rates Database and the GSW yields for maturities from 1 to 15 years. The sample period is 1972-2010. Panel A presents results for commonly used simple benchmark models. Panels B and C present results for competing models with, respectively, 1 and 2 forecasting factors. Panel D presents results for models that do not impose common forecasting factor structure across the 14 bond excess return series. Description of the models can be found in the text and in our model taxonomy table A.1.

Out-of-sample R^2	**Bond Excess Returns**													
Models	**$rx^{(2)}$**	**$rx^{(3)}$**	**$rx^{(4)}$**	**$rx^{(5)}$**	**$rx^{(6)}$**	**$rx^{(7)}$**	**$rx^{(8)}$**	**$rx^{(9)}$**	**$rx^{(10)}$**	**$rx^{(11)}$**	**$rx^{(12)}$**	**$rx^{(13)}$**	**$rx^{(14)}$**	**$rx^{(15)}$**
Panel A: Naïve benchmark models														
Rolling Average	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Random Walk	-0.514	-0.585	-0.640	-0.686	-0.728	-0.765	-0.797	-0.824	-0.846	-0.862	-0.874	-0.881	-0.883	-0.882
Panel B: Models with 1 forecasting factor														
PCR-1	0.019	0.019	0.020	0.022	0.024	0.025	0.026	0.027	0.027	0.027	0.028	0.028	0.028	0.028
RRRR1-PC2	0.012	0.010	0.010	0.012	0.013	0.015	0.016	0.017	0.019	0.020	0.020	0.021	0.022	0.022
RRRR1-PC3	0.053	0.039	0.034	0.032	0.030	0.030	0.029	0.029	0.029	0.029	0.029	0.030	0.031	0.032
RRRR1-PC4	0.030	0.016	0.012	0.012	0.013	0.015	0.017	0.019	0.021	0.023	0.025	0.027	0.029	0.032
RRRR1-PC5	0.059	0.051	0.053	0.056	0.059	0.060	0.061	0.060	0.059	0.058	0.056	0.055	0.054	0.053
RRRR1-SMP	0.019	0.019	0.019	0.021	0.022	0.023	0.023	0.024	0.024	0.024	0.025	0.025	0.025	0.025
RRRR1-TMP	0.045	0.038	0.036	0.036	0.036	0.037	0.037	0.038	0.038	0.038	0.039	0.039	0.039	0.040
Panel C: Models with 2 forecasting factors														
PCR-2	0.009	0.007	0.009	0.011	0.013	0.015	0.017	0.017	0.018	0.018	0.018	0.018	0.018	0.018
RRRR2-PC3	0.030	0.024	0.024	0.025	0.027	0.029	0.030	0.031	0.031	0.032	0.033	0.033	0.034	0.035
RRRR2-PC4	0.034	0.019	0.014	0.013	0.014	0.015	0.016	0.017	0.018	0.019	0.021	0.023	0.025	0.027
RRRR2-PC5	0.071	0.063	0.063	0.064	0.063	0.060	0.056	0.050	0.045	0.040	0.036	0.034	0.032	0.031
RRRR2-SMP	0.012	0.010	0.011	0.014	0.016	0.017	0.018	0.019	0.020	0.020	0.020	0.020	0.019	0.019
RRRR2-TMP	0.039	0.030	0.028	0.028	0.029	0.030	0.031	0.032	0.033	0.034	0.035	0.035	0.036	0.037
Panel D: Models with 14 forecasting factors														
RR-SMP	0.018	0.016	0.017	0.019	0.020	0.022	0.023	0.024	0.024	0.025	0.025	0.026	0.026	0.026
RR-TMP	0.074	0.058	0.050	0.045	0.041	0.039	0.037	0.036	0.036	0.035	0.035	0.035	0.035	0.036
PLS-1	0.054	0.032	0.028	0.027	0.027	0.027	0.028	0.029	0.029	0.030	0.030	0.030	0.031	0.031
PLS-2	0.005	0.002	0.004	0.014	0.021	0.014	0.026	0.029	0.029	0.028	0.026	0.026	0.025	0.024
PLS-3	0.031	0.026	0.026	0.025	0.025	0.015	0.011	0.012	0.013	0.015	0.018	0.018	0.019	0.021
PLS-4	0.006	0.001	0.002	0.009	0.019	0.026	0.028	0.051	0.034	0.032	0.033	0.034	0.038	0.037
PLS-5	-0.025	0.002	0.027	0.042	0.047	0.044	0.036	0.027	0.017	0.005	-0.002	0.004	0.015	0.026
3PRF-1	-0.032	-0.001	0.010	0.019	0.024	0.022	0.017	0.012	0.007	0.003	-0.001	-0.002	-0.002	-0.002
3PRF-2	0.023	0.021	0.022	0.024	0.021	0.017	0.018	0.017	0.017	0.016	0.016	0.019	0.021	0.023
3PRF-3	0.006	0.001	0.002	0.009	0.014	0.015	0.014	0.012	0.010	0.010	0.010	0.012	0.015	0.021
3PRF-4	-0.025	-0.002	0.015	0.030	0.038	0.035	0.026	0.015	0.003	-0.003	0.001	0.011	0.017	0.022
3PRF-5	-0.047	-0.016	0.001	0.008	0.011	0.007	0.001	-0.009	-0.018	-0.026	-0.025	-0.022	-0.020	-0.016
OLS	-0.549	-0.424	-0.361	-0.330	-0.318	-0.316	-0.320	-0.327	-0.333	-0.337	-0.340	-0.340	-0.338	-0.336

Table A.8: **Out-of-sample R^2 by forecasting method for monthly bond excess returns predicted by the combined set of yield curve slopes and corresponding maturity-related cycles of Cieslak & Povala (2011).** For rolling out-of-sample forecasts with rolling window size 120 months we report out-of-sample R^2 by forecasting method relative to a rolling average benchmark. We forecast monthly excess returns of bonds ranging from 2 to 15 years of maturity. The risk-free rate is taken to be the 1-month T-bill rate from the CRSP Fama Risk-Free Rates Database. The set of predictors is given by the yield curve slopes for the 1-month and 3-month T-bill rates from the CRSP Fama Risk-Free Rates Database and the GSW yields for maturities from 1 to 15 years in combination with the maturity-related cycles of Cieslak & Povala (2011) for GSW yields from 1 to 15 years. The sample period is 1972-2010. Panel A presents results for commonly used simple benchmark models. Panels B and C present results for competing models with, respectively, 1 and 2 forecasting factors. Panel D presents results for models that do not impose common forecasting factor structure across the 14 bond excess return series. Description of the models can be found in the text and in our model taxonomy table A.1.

Out-of-sample R^2	**Bond Excess Returns**													
Models	**$rx^{(2)}$**	**$rx^{(3)}$**	**$rx^{(4)}$**	**$rx^{(5)}$**	**$rx^{(6)}$**	**$rx^{(7)}$**	**$rx^{(8)}$**	**$rx^{(9)}$**	**$rx^{(10)}$**	**$rx^{(11)}$**	**$rx^{(12)}$**	**$rx^{(13)}$**	**$rx^{(14)}$**	**$rx^{(15)}$**
Panel A: Naïve benchmark models														
Rolling Average	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Random Walk	-0.514	-0.585	-0.640	-0.686	-0.728	-0.765	-0.797	-0.824	-0.846	-0.862	-0.874	-0.881	-0.883	-0.882
Panel B: Models with 1 forecasting factor														
PCR-1	-0.019	-0.011	-0.006	-0.002	0.000	0.002	0.003	0.003	0.004	0.004	0.004	0.004	0.004	0.005
RRRR1-PC2	0.034	0.036	0.040	0.043	0.047	0.051	0.054	0.056	0.059	0.061	0.062	0.063	0.065	0.065
RRRR1-PC3	0.012	0.017	0.024	0.030	0.035	0.041	0.045	0.050	0.053	0.057	0.059	0.062	0.064	0.065
RRRR1-PC4	0.027	0.021	0.022	0.025	0.029	0.034	0.038	0.042	0.045	0.048	0.051	0.053	0.056	0.058
RRRR1-PC5	0.057	0.046	0.045	0.048	0.053	0.059	0.065	0.071	0.077	0.082	0.087	0.091	0.095	0.099
RRRR1-SMP	0.034	0.036	0.040	0.043	0.047	0.051	0.054	0.056	0.059	0.061	0.062	0.063	0.065	0.065
RRRR1-TMP	0.027	0.029	0.034	0.039	0.044	0.049	0.054	0.057	0.061	0.064	0.066	0.068	0.070	0.071
Panel C: Models with 2 forecasting factors														
PCR-2	0.023	0.028	0.034	0.040	0.045	0.049	0.053	0.056	0.058	0.060	0.062	0.063	0.064	0.065
RRRR2-PC3	0.017	0.022	0.028	0.034	0.039	0.043	0.047	0.050	0.052	0.054	0.055	0.056	0.057	0.058
RRRR2-PC4	0.004	0.008	0.017	0.025	0.031	0.037	0.041	0.044	0.046	0.048	0.049	0.050	0.051	0.051
RRRR2-PC5	0.033	0.030	0.035	0.042	0.050	0.059	0.066	0.073	0.079	0.084	0.089	0.093	0.096	0.099
RRRR2-SMP	0.020	0.026	0.033	0.040	0.045	0.050	0.054	0.057	0.059	0.061	0.063	0.065	0.066	0.067
RRRR2-TMP	-0.012	0.003	0.016	0.028	0.038	0.045	0.051	0.056	0.059	0.062	0.064	0.065	0.066	0.067
Panel D: Models with 14 forecasting factors														
RR-SMP	0.020	0.026	0.033	0.040	0.045	0.050	0.054	0.057	0.059	0.061	0.063	0.065	0.066	0.067
RR-TMP	0.015	0.023	0.032	0.039	0.045	0.050	0.054	0.058	0.060	0.063	0.064	0.066	0.067	0.068
PLS-1	-0.003	0.011	0.023	0.033	0.042	0.048	0.053	0.057	0.060	0.062	0.064	0.066	0.067	0.068
PLS-2	0.019	0.027	0.034	0.044	0.041	0.046	0.052	0.054	0.052	0.055	0.058	0.060	0.063	0.065
PLS-3	0.011	0.016	0.025	0.032	0.039	0.044	0.047	0.048	0.053	0.055	0.058	0.062	0.061	0.063
PLS-4	-0.008	-0.006	0.000	0.012	0.022	0.026	0.035	0.035	0.041	0.052	0.057	0.059	0.062	0.066
PLS-5	0.019	0.023	0.027	0.037	0.035	0.039	0.046	0.051	0.054	0.056	0.059	0.061	0.067	0.069
3PRF-1	-0.028	-0.013	-0.002	0.013	0.024	0.018	0.014	0.009	0.001	-0.002	-0.003	-0.002	-0.001	0.001
3PRF-2	0.062	0.063	0.066	0.067	0.064	0.060	0.061	0.061	0.057	0.056	0.056	0.053	0.052	0.051
3PRF-3	0.050	0.036	0.037	0.046	0.055	0.058	0.062	0.065	0.066	0.068	0.068	0.069	0.069	0.069
3PRF-4	0.024	0.033	0.037	0.044	0.051	0.055	0.057	0.063	0.066	0.072	0.080	0.083	0.084	0.086
3PRF-5	0.019	0.031	0.042	0.048	0.052	0.050	0.053	0.056	0.055	0.051	0.055	0.057	0.061	0.066
OLS	-0.597	-0.478	-0.414	-0.381	-0.364	-0.358	-0.357	-0.357	-0.358	-0.357	-0.354	-0.349	-0.343	-0.336

Appendix B. Proofs

Appendix B.1. Random Matrix Theory Proofs

Since we shall work with sequences of matrices of increasing size, it will be convenient to work with the spectral norm for bounded linear operators in addition to the trace norm (Frobenious norm). For any square $n \times n$ matrix A as

$$|||A||| = \max_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \alpha_{\max}$$

where $\lambda_{\max}$ is the largest eigenvalue of the square matrix A and $\|x\|_2$ is the usual l^2 norm for vectors. The spectral norm is sub-multiplicative (i.e. $|||AB||| \leq |||A|||\,|||B|||$) and consistent (the norm definition is identical for any $n > 0$).

Another useful lemma is Weyl's inequality for Hermitian matrices.

Lemma 1 (Weyl's inequality)**.** *Let S, T, U be Hermitian $n \times n$ matrices with ordered eigenvalues $\{\sigma_1 \geq \ldots \geq \sigma_n\}$, $\{\tau_1 \geq \ldots \geq \tau_n\}$, and $\{\upsilon_1 \geq \ldots \geq \upsilon_n\}$ respectively. If $S = T + U$, then*

$$\tau_i + \upsilon_n \leq \sigma_i \leq \tau_i + \upsilon_1 \tag{B.1}$$

Proof of Proposition 3. Apply Weyl's lemma (Lemma 1) with $S = S_{XX}$, $T = \Lambda'\Lambda$, and $U = \Omega_n$. Since the eigenvalues of $Omega_n$ are bounded by Theorem 1, and $\tau_1, \ldots, \tau_r$ tend to infinity by Assumption 4, it is clear that $\sigma_1, \ldots, \sigma_r$ will also diverge. Next, since $\tau_{r+1} = \cdots = \tau_n = 0$, tells us that $\forall i > r$, the limiting value of σ_i is bounded between $(1 - \sqrt{\gamma})^2$ and $(1 + \sqrt{\gamma})^2$. □

Appendix B.2. Proofs of RRRR results

Lemma 2. *Let Γ, Λ be positive semidefinite $n \times n$ matrices and Λ be invertible, then*

$$A^\star = \arg \max_{\{A \in \mathbb{R}^{n \times k} : A'\Lambda A = I_{k \times k}\}} tr\{A'\Gamma A\,(A'\Lambda A)^{-1}\} \tag{B.2}$$

is given by the k eigenvectors belonging to the k largest eigenvalues from the generalized eigenvalue problem

$$|\Gamma - \lambda\,\Lambda| = 0 \tag{B.3}$$

Proof. Follows from the fact that if (λ_i, c_i) is an eigenvalue-eigenvector pair of (B.3), then $\lambda_i \Lambda c_i = \Gamma c_i$, and $C = (c_1, \ldots, c_n)$ is a basis for $\mathbb{R}^n$ where for $i \neq j$, $c_i'\Lambda c_j = 0$. The first order condition with respect to A in (B.2) yields

$$[(A'\Gamma A)(A'\Lambda A)^{-1}]\,A'\Lambda = A'\Gamma$$

Note that the term in brackets above is simply our objective whose trace we wish to maximize. Since the trace operator only involves the diagonal elements, the proof now proceeds by induction. Suppose $k = 1$, then the term in square brackets above is a scalar, and clearly is maximized when A is the eigenvector associated with the largest eigenvalue of (B.3). Next, given the first $k - 1$ columns of A, it is now trivial to see that the objective is maximized by setting the k^{th} column equal to the eigenvector belonging to the k^{th} largest eigenvalue. □

Proof of Proposition 1. The optimal A solves

$$\begin{aligned} & \min_A tr\{W^{1/2}S_{YY}W^{1/2} - W^{1/2}S_{YX}A(A'[S_{XX} - \rho^2 R'R]A)^{-1}A'S_{XY}W^{1/2}\} \\ = & \max_A tr\{A'S_{XY}WS_{YX}A[A'(S_{XX} - \rho^2 R'R]A)^{-1}\} \end{aligned}$$

The statement of the proposition now follows from Lemma 2 with $\Gamma = S_{XY}WS_{YX}$ and $\Lambda = S_{XX} - \rho^2 R'R$. □

Proof of Corollary 1. Let A be restricted to be of the form $A = P^{\perp}a$ for some $a \in \mathbb{R}^{(n-f)\times k}$. The optimal a then solves

$$\max_a tr\{(P^{\perp}a)'S_{XY}S_{YX}(P^{\perp}a)\left((P^{\perp}a)'[S_{XX} - \rho^2 R'R](P^{\perp}a)\right)^{-1}\}$$

The main result of the proposition now follows directly from Lemma 2 with the $(n-f)\times(n-f)$ matrices $\Gamma = P^{\perp\prime}S_{XY}S_{YX}P^{\perp}$ and $\Lambda = P^{\perp\prime}[S_{XX} - \rho^2 R'R]P^{\perp}$ □

Proof of Corollary 2. Let the singular value decomposition of $\mathbf{X}$ be given by (3)-(4), then the principal components of $\mathbf{X}$ are given by $\mathbf{F} = \mathbf{X}V_r\Sigma_r^{-1}$. The optimal loadings on $\mathbf{F}$ in the two-step approach are given by the matrix $a \in \mathbb{R}^{r\times k}$ consisting of the k principal eigenvectors of

$$0 = |S_{FY}S_{FY}' - \lambda S_{FF}| \Rightarrow 0 = |S_{U_rY}S_{U_rY}' - \lambda I_r| \tag{B.4}$$

and the resulting loading on $\mathbf{X}$ is therefore given by $\tilde{A} = V_r\Sigma_r^{-1}a$.

The regularized (via spectral truncation) reduced rank factor loadings, A, solve the generalized eigenvalue problem

$$0 = |V_r\Sigma_r^{-2}V_r'S_{XY}S_{XY}' - \lambda I_n| \Rightarrow 0 = |V_r\Sigma_r^{-1}S_{U_rY}[S_{U_rY}'\Sigma_rV_r' + S_{U_{n-r}Y}'\Sigma_{n-r}V_{n-r}'] - \lambda I_n| \tag{B.5}$$

To see if the two solutions are identical, we take an eigenvalue-eigenvector pair (λ_i, a_i) of (**??**) and check whether $\tilde{A}_i = V_r\Sigma_r^{-1}a_i$ is an eigenvalue of (B.5) corresponding to the eigenvalue λ_i:

$$\begin{aligned} V_r\Sigma_r^{-1}S_{U_rY}[S_{U_rY}'\Sigma_rV_r' + S_{U_{n-r}Y}'\Sigma_{n-r}V_{n-r}']\tilde{A}_i &= V_r\Sigma_r^{-1}S_{U_rY}[S_{U_rY}'\Sigma_rV_r' + S_{U_{n-r}Y}a_i && \text{(B.6)} \\ &= \lambda_iV_r\Sigma_r^{-1}a_i = \lambda_i\tilde{A}_i && \text{(B.7)} \end{aligned}$$

where the last equality follows from the fact that (λ_i, a_i) is an eigenvalue-eigenvector pair for (B.4). □