

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 high-dimensional 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 with the ability to enforce factor interpretability via variable grouping constraints implied by economic theory. As an important special case, our approach generalizes principal component regression by applying reduced rank rather than linear regression to the principal components of the regressors, thereby disentangling the forecasting factors driving the outcomes from the factor structure in the 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, factor interpretability

*We are grateful for comments from Luca Benzoni, Domenico Giannone (discussant), Seth Pruitt, Kevin Sheppard (discussant), Christopher Sims, Mark Watson, Kenneth West, and participants at the Seventh ECB Workshop on Forecasting Techniques - New Directions for Forecasting, Frankfurt, May 4-5, 2012, the Financial Econometrics Conference, Toulouse, May 11-12, 2012, the North American Summer Meeting of the Econometric Society, Evanston, June 28 - July 1, 2012, the NBER Summer Institute Workshop on Forecasting and Empirical Methods in Macroeconomics and Finance, Boston, July 10-13, 2012, the NBER-NSF Time Series Conference, Texas A&M University, October 26-27, 2012, and seminar participants at the Federal Reserve Board of Governors, the Federal Reserve Bank of Chicago, the Federal Reserve Bank of New York, the Cass Business School, Universitat Pompeu Fabra, and the Toulouse School of Economics.

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. 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:

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

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, noting that there appears to be scope for disentangling the forecasting factors driving the outcomes from the factor structure in the predictors.² 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} :

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

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

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

²For example, Cochrane and Piazzesi (2005) find that bond excess returns are driven by a single forecasting factor extracted from bond yields, which in turn are known to have three or higher dimensional factor structure.

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 .³ In other words, forecasting multiple outcomes using a smaller number 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 collinearity) to standard dimension reduction techniques (in order to exploit covariance between \mathbf{X} and \mathbf{Y}) that provides the econometrician with a flexible framework for extracting *common* predictive factor structures in the data. The common thread of these dimension reduction techniques is that they all solve a constrained maximization problem which can be formulated as a generalized eigenvalue problem involving ill-conditioned matrices to which regularization can be applied.

The current paper focusses on an important sub-class of these forecasting models, 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 fixed and data driven methods for the choice of regularization threshold. 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.

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

³This situation arises when the Y s themselves exhibit a strong (predictable) factor structure, such as the level, slope and curvature of bond yields.

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 6. Data driven techniques for choosing the degree of regularization are discussed in Section 7 while Section 8 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 9 concludes.

1.1. 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 regression coefficients Θ can be interpreted as a regularization parameter. Of particular relevance to our multivariate setting is Doan, Litterman, and Sims (1984) who consider 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 on Θ that can be understood as a regularization of the corresponding frequentist model.⁴ Carriero, Kapetanios, and Marcellino (2011) apply the Geweke (1996) model to the Stock and Watson macro variables and find that the dimension reduction of RRR in combination with Bayesian shrinkage is beneficial. This finding consistent with the results of this paper (which employs what amounts to different Bayesian priors) for both macro variables as well as bond returns. However, we note that the computation of the estimator in this paper is trivial even for very large panels (i.e. solving for the largest eigenvalues of a single generalized eigenvalue problem is near instantaneous even for $n \sim 1000$) whereas the computational effort of the Bayesian reduced rank regression is considerable for even modest $n \sim 50$ as pointed out by Carriero et al. (2011).

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 6 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

The Partial Least Squares (PLS) of Wold (1966), is explicitly designed to exploit the information contained in S_{XY} and has a long history in chemometrics. More recently, PLS has also

⁴To 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.

⁵This 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.

been applied in economics and is closely related to the 3PRF model recently proposed by Kelly and Pruitt (2011b) and applied to forecasting of stock returns and dividend growth in Kelly and Pruitt (2011a). While the PLS approach effectively sidesteps the issue of ill-conditioned S_{XX} (in fact one of the original motivations for its introduction), it is not in general a shrinkage technique and differs from the estimators considered in this paper that can be cast as an explicit penalized least squares objective function involving two distinct shrinkage parameters.

Finally, a recent paper by Chen, Chan, and Stenseth (2012) studying micro array gene expression data suggests what amounts to regularizing reduced rank regression with a LASSO type penalty, similar to that used by Mol, Giannone, and Reichlin (2008) in a univariate forecasting context. The model of Chen et al. (2012) can be seen as an alternative to the RRRR methods proposed here albeit with a non-smooth penalty and without the advantage of a closed form solution. The main advantage of the LASSO specification is the sparsity of the resulting factors in each subsample, but with the common drawback of instability across subsamples as pointed out by Mol et al. (2008). The comparison to the RRRR methods of this paper for forecasting macro economic and financial time series will be the subject of future work.

2. Regularization and Latent Factor Extraction

For given data \mathbf{Y}, \mathbf{X} , on T observations of an m dimensional vector of outcomes, Y and an n dimensional vector of predictors, X , the classical latent factor models posit that a set of $k \ll n$ linear combinations $A'X$ contain the useful information for forecasting Y in the sense that $E[Y|A'X] \approx E[Y|X]$. The use of “ \approx ” refers to the fact all dimensions of X may contain *some* additional information about Y , but that any forecasting factors beyond the first k have a unfavorable signal to noise ration and adding them would lead to a deterioration in the mean squared forecast error.

A number of classical latent factor models discussed below were developed under the explicit assumption that $T \rightarrow \infty$ for fixed m, n , and therefore not suited for large panels without further restrictions imposed. Regularization provides one such coherent scheme for imposing such restrictions.

2.1. Latent Predictive Factor Extraction Techniques

Consider the problem of capturing the covariation between X and Y through a set of $k \leq \min(n, m)$ maximally covarying linear combinations $\alpha'X$ and $\beta'Y$ where $\alpha = \{\alpha_1, \dots, \alpha_k\} \in \mathbb{R}^{n \times k}$ and $\beta = \{\beta_1, \dots, \beta_k\} \in \mathbb{R}^{m \times k}$. In general, the weights (α_i, β_i) can be solved for recursively such that they are normalized and satisfy a set of orthogonality constraints to ensure a unique solution (see e.g. Burnham, Viveros, and MacGregor (1996)):

$$\max_{\{\alpha_i, \beta_i\}} \alpha'_i S_{XY} \beta_i, \quad \text{s.t.} \quad \alpha'_i M_1 \alpha_i = 1, \beta'_i M_2 \beta_i = 1, \forall j < i : \alpha'_i M_3 \alpha_j = 0 \quad (3)$$

Several of the most popular dimension reduction techniques, including Hotelling’s canonical correlation analysis (CCA), the Reduced Rank Regression (RRR) of Izenman (1975), and the SIMPLS version (due to de Jong (1993)) of the PLS estimator of Wold (1966) fall within the framework in (3), as shown in Table 1. Thus all these techniques are quite closely related and share the objective of summarizing the information about the co-movement of X and Y contained in the S_{XY} matrix. This is in stark contrast to the popular principal components regression analysis (PCR) which instead solves for a set of $k < n$ weights α so that the resulting factors $\alpha'X$ summarize the variation in \mathbf{X} :

$$\max_{\{\alpha_i\}} \alpha_i' S_{XX} \alpha_i, \quad \text{s.t.} \quad \alpha_i' M_1 \alpha_i = 1, \quad \forall j < i : \alpha_i' M_3 \alpha_j = 0 \quad (4)$$

TABLE 1: Linear restrictions associated with each dimension reduction technique.

	CCA	RRR	SIMPLS	PLS [†]	PCR [‡]
M_1	S_{XX}	S_{XX}	I_n	$I_n - P_i$	S_{XX}
M_2	S_{YY}	I_m	I_m	I_m	.
M_3	S_{XX}	S_{XX}	S_{XX}	S_{XX}	S_{XX}

[†]The original Wold (1966) PLS specification has a non-constant M_1 matrix due to the iterative deflation of \mathbf{X} . Here P_i is the projection onto the space of $i - 1$ previously extracted components. [‡]The PCR objective is different and does not involve β .

In the classical analysis, m, n are fixed and $T \rightarrow \infty$, so that M_1, M_2 can be assumed to be well-behaved in the limit and (3) can be analyzed using the classical results on constrained eigenvalue problems in e.g. Rao (1964). In practice, however, problems arise when either M_1 or M_2 are ill-conditioned.⁶ To see the cause of the problem, consider the case of CCA and RRR. In this case one may concentrate out β in (4) to obtain α as the eigenvectors corresponding to the k principal eigenvalues of the generalized eigenvalue problem:

$$|S_{XY} M_2^{-1} S'_{XY} - \lambda M_1| = 0 \quad (5)$$

The expression (5) is also known as a *matrix pencil* and it is well-known from the theory of singular matrix pencils that when the matrices M_1 and/or M_2 are ill-conditioned, the solution to the generalized eigenvalue problem becomes extremely unstable (c.f. Gantmacher (1960)). By contrast, SIMPLS, PLS and PCR do not face this problem since M_1 and M_2 well behaved by definition (for PCR there is no M_2).

A natural approach to address this problem afflicting CCA and RRR is to regularize the two matrices, i.e. replace them by perturbed versions that have bounded inverses, and different regularization schemes will differ in the way this perturbation is carried out. In Vinod (1976), it

⁶This was recognized early on in the canonical correlations literature by Vinod (1976) but to our knowledge the use of regularized CCA never received much attention in the subsequent econometric literature on large panels.

was proposed to apply a ridge penalty in the CCA setting to extracting a single canonical variate. However, recognizing the common structure of (5), it is clear that the same regularization idea applies to CCA, RRR and SIMPLS more generally.⁷

In this paper, we choose to focus on the RRR case rather than CCA, so that our main concern is the ill-condition of S_{XX} . The reason for this is twofold. First, we are usually ultimately interested in the forecast of specific quantities with clear economic interpretation and not a rotated set of statistical “portfolios”. Finally, in our implementation of RRRR, we allow for a weighting matrix, W , to be applied which will play the role of M_2^{-1} , but specified to be well-behaved (typically equal to I_m or a diagonal matrix containing the inverse variances of Y in applications). Since one is free to let W equal a regularized inverse of S_{YY} , regularized CCA is a special case of the RRRR considered in this paper.

2.2. Regularization and Filter Factors

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, known either as Spectral Truncation or 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.⁸ 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 $\|S\| = \text{tr}\{S\}$. On the space of real $n \times m$ matrices, A , we shall use the Frobenius norm, $\|A\| = \text{tr}\{A'A\}^{1/2}$. 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} .

2.2.1. Regularized Least Squares

To understand the rational behind different regularization schemes, it is instructive to consider the baseline linear regression case. The properties of the linear system (1) are completely

⁷Albeit structurally different, we conjecture that PLS might well benefit from some degree of “pre-regularization” in certain situations.

⁸Another 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 non-trivial numerical optimization problem. See Mol et al. (2008) for a comprehensive comparative study of ridge and LASSO regression based forecasts in a univariate setting.

determined by the singular value decomposition of the matrix \mathbf{X} :

$$\mathbf{X} = U\Sigma V' = \sum_{i=1}^n \sigma_i u_i v_i' \quad (6)$$

where $U = (u_1, \dots, u_n) \in \mathbb{R}^{T \times n}$, $V = (v_1, \dots, v_n) \in \mathbb{R}^{n \times n}$ are orthonormal matrices and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ 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:

$$\mathbf{X} = U_r \Sigma_r V_r' + U_{n-r} \Sigma_{n-r} V_{n-r}' \quad (7)$$

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 $\sigma_1 \geq \dots \geq \sigma_n \geq 0$ decreases gradually to zero.⁹ Figure A.2 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 Θ 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

$$\hat{\Theta}_{OLS} = \sum_{i=0}^n v_i \frac{u_i' \mathbf{Y}}{\sigma_i} = \Theta_0 + \sum_{i=0}^n v_i \frac{u_i' \mathbf{e}}{\sigma_i} \quad (8)$$

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.¹⁰ 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

⁹The case where one or more eigenvalues are literally zero is easily handled by eliminating redundant variables. However, in many situations, the inclusion of additional predictors simply increases the number of small eigenvalues.

¹⁰If all eigenvalues happen to be small (or very large), it of course merely means that the problem is badly scaled.

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 $E\|\hat{\Theta}_{OLS} - \Theta_0\|^2 = \kappa^2 \text{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\} = \kappa^2 \sum_{i=1}^n \sigma_i^{-2}$, thus illustrating the problem of ill-condition.

An effective approach to solving ill-conditioned systems of equations is via *regularization* of the equation (8):

$$\tilde{\Theta} = \sum_{i=1}^n f_i v_i \left(\frac{u'_i \mathbf{Y}}{\sigma_i} \right), \quad \|\tilde{\Theta}\|_F^2 = \sum_{i=1}^n f_i^2 \left(\frac{u'_i \mathbf{Y}}{\sigma_i} \right)^2 \quad (9)$$

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 $T \rightarrow \infty$, and $\tilde{\Theta}$ the finite sample shrinkage estimate. In general $\tilde{\Theta} \rightarrow \tilde{\Theta}_\infty \neq \Theta_0$ as $T \rightarrow \infty$ and we have the bound

$$\underbrace{E\|\Theta_0 - \tilde{\Theta}\|}_{\text{root mean squared shrinkage estimation error}} \leq \underbrace{\|(\Theta_0 - \tilde{\Theta}_\infty)\|}_{\text{bias due to regularization}} + \underbrace{E\|(\tilde{\Theta}_\infty - \tilde{\Theta})\|}_{\text{dampened volatility due to regularization}} \quad (10)$$

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.¹¹

In general, the “optimal” filter factors depend on the *unknown* “signal-to-noise” ratio, which follows from

$$\tilde{\Theta} - \Theta_0 = \underbrace{\sum_{i=1}^n (f_i - 1) v_i (v'_i \Theta_0)}_{\text{Bias}} + \underbrace{\sum_{i=1}^n f_i v_i \left(\frac{u'_i \mathbf{e}}{\sigma_i} \right)}_{\text{Dampened variance}}$$

¹¹Note that, in the classical case where $n/T \rightarrow 0$, 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: $\tilde{\Theta}_\infty = \Theta_0$.

↓

$$E\|\tilde{\Theta} - \Theta_0\|^2 = \sum_{i=1}^n (f_i - 1)^2 (v_i' \Theta_0 \Theta_0' v_i) + f_i^2 \left(\frac{u_i' E[\mathbf{e}' \mathbf{e}] u_i}{\sigma_i^2} \right)$$

The infeasible optimal filter factors are uniquely determined and given by

$$f_i = \frac{v_i' \Theta_0 \Theta_0' v_i}{v_i' \Theta_0 \Theta_0' v_i + \frac{u_i' E[\mathbf{e}' \mathbf{e}] u_i}{\sigma_i^2}} = \frac{\text{“signal”}}{\text{“signal”} + \sigma_i^{-2} \text{“noise”}} \quad (11)$$

2.2.2. Tikhonov Regularization a.k.a. Ridge Regression

If nothing is known ex-ante about the noise, one may assume that the noise-to-signal ratio is identical in each direction indicated by the singular vectors of \mathbf{X} , and given by a constant ρ^2 , which leads to optimal filter factors of the form

$$f_i = \frac{1}{1 + \rho^2 / \sigma_i^2} = \frac{\sigma_i^2}{\sigma_i^2 + \rho^2} \quad (12)$$

The filter factors (12) correspond to one of the most commonly used regularization techniques, *Tikhonov* regularization, which has a natural interpretation as a penalized least squares estimator known as the *Ridge Regression* estimator:¹²

$$\min_{\tilde{\Theta}} \|\mathbf{Y} - \mathbf{X}\tilde{\Theta}\|^2 + \rho^2 \|\tilde{\Theta}\|^2, \quad \tilde{\Theta} \in \mathbb{R}^{n \times m}, \rho \geq 0 \quad (13)$$

Solving the Lagrangian implies that $\tilde{\Theta} = (\mathbf{X}'\mathbf{X} + \rho^2 I_n)^{-1} \mathbf{X}'\mathbf{Y} = \sum_{i=1}^n \left[\frac{\sigma_i^2}{\sigma_i^2 + \rho^2} \right] v_i \left(\frac{u_i' \mathbf{Y}}{\sigma_i} \right)$. Comparing with (12), we see that the Tikhonov scheme is optimal if and only if the signal to noise ratio is constant for all i and equals $1/\rho^2$. Clearly, more noisy data implies a larger optimal ρ which in turn 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 (10) in the Tikonov case is

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

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 $\sum \sigma_i^{-4} \|S_{Xe}\|^2$.

¹²The Tikhonov formulation is usually slightly more general:

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

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, $1/\rho^2 = \frac{\sigma_{\text{prior}}^2}{\sigma_{\text{noise}}^2}$, is the ratio of standard deviation of the noise to the standard deviation of the prior.

From the penalty term in (13) 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.

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

If the noise-to-signal ratio is assumed negligible in the directions indicated by the singular vectors corresponding to the first r singular values (i.e. $\rho \approx 0$) but infinite (i.e. $\rho \approx \infty$) for the remaining $n - r$ singular vectors, the optimal filtering scheme sets

$$f_1 = \dots = f_r \equiv 1 \text{ and } f_{r+1} = \dots = f_n \equiv 0 \quad (14)$$

so that any components of Y orthogonal to the last $n - r$ left singular vectors of \mathbf{X} are ignored. This scheme is known as spectral truncation (because small singular values are zeroed out) or principal components regression (PCR) because it can be interpreted as a regression in which \mathbf{X} is replaced by its first r singular vectors.¹³

This type of regularization can be motivated in large panels under the null that X is driven by an r -dimensional factor structure that is informative for Y :

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

and that the largest r eigenvalues of $\Lambda'\Lambda$ diverge as n, T grow large.¹⁴ Let the singular value decomposition of \mathbf{X} be given by (6)-(7), 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 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:

$$\tilde{\Theta} = S_{XX}^\dagger S_{XY} = V_r\Sigma_r^{-1}(S_{U_rY}) \quad (15)$$

while the PCR estimator is

$$\tilde{\Theta}_{PCR} = S_{FF}^{-1}S_{FY} = S_{U_rY} \quad (16)$$

and we thus have: $\mathbf{F}\tilde{\Theta}_{PCR} = \mathbf{X}(V_r\Sigma_r^{-1})S_{U_rY} = \mathbf{X}\tilde{\Theta}$, 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.

¹³Regularization 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.

¹⁴This is the identifying assumption of e.g. Bai&Ng (2002).

In general, the bias-variance trade-off (10) in the spectral truncation case is

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

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

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, Y , and these are driven by a common low dimensional latent factor structure, F :

$$Y_{t+1} = CF_t + \varepsilon_{t+1} \quad (17)$$

$$X_t = \Lambda F_t + \xi_t \quad (18)$$

The assumption in (18) is that we have available (a possible large) set of n observable predictors \mathbf{X} driven by r latent factors, F_t . The $m \times 1$ vector of predictable outcomes Y_{t+1} in turn is driven by a set of $k \leq r$ forecasting factors spanned by F_t , i.e. $0 < \text{rank}(C) = k \leq r$.

Under these assumptions, equation (18) can be inverted to substitute out for the latent factors in (17) to yield:

$$Y_{t+1} = (C\Lambda^-) X_t + \tilde{\varepsilon}_{t+1} \quad (19)$$

where Λ^- is a $r \times n$ left inverse of Λ . Under the assumptions made above, the matrix $(C\Lambda^-)$ is a $m \times n$ matrix of reduced rank equal to $k = \text{rank}(C) \leq r$. We denote k , the dimension of the forecastable component of Y , the *model complexity* or simply the *number of forecasting factors*.

From (19) one sees that the multivariate latent factor model naturally yields a reduced rank regression where the reduced rank equals the dimension of the predictable factor space. This model is the classical reduced rank regression of Anderson (1963) and Izenman (1975).

Going forward we shall parametrize (19) matrix as $(C\Lambda^-) = (AB)'$ where $A \in \mathbb{R}^{n \times k}$ and $B \in \mathbb{R}^{m \times k}$. In matrix form we then have

$$\mathbf{Y} = \mathbf{X}AB + E \quad \text{and} \quad (\hat{A}, \hat{B}) = \arg \max_{\{A, B\}} \|(\mathbf{Y} - \mathbf{X}AB)W^{1/2}\|^2 \quad (20)$$

where W is a non-singular weighting matrix. For a given choice of k in (20), 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 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 .

For a given choice of weighting matrix $W \in \mathbb{R}^{m \times m}$ (e.g. a diagonal matrix of inverse variances in our applications) in (20), the optimal A is found by solving the generalized eigenvalue problem

$$|S_{XY}WS_{YX} - \lambda S_{XX}| = 0 \quad (21)$$

and setting A equal to the k eigenvectors belonging to the largest eigenvalues. Thus reduced rank regression, while a proper shrinkage estimator, remains susceptible to instability when regressors are nearly collinear (e.g. when $n \sim T$), and *cannot* be directly applied in a large panel setting. 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 covariance structure between X and 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 the choice of number of forecasting factors until Section 7 below.

4.1. Tikhonov Regularization of Reduced Rank Regression

In the context of the reduced rank regression (20), Tikhonov regularization involves modifying the objective function to include a term that penalizes “large” values of $\|AB\|$:

$$\min_{\{A,B\}} \|(\mathbf{Y} - \mathbf{X}AB)W^{\frac{1}{2}}\|^2 + \rho^2 \|R(AB)W^{\frac{1}{2}}\|^2 \quad (22)$$

where R in general is a $q \times n$ matrix which may be chosen to differentially penalize certain directions in the parameter space.^{15,16,17} In the special case when $R = I_n, W = I_m$ and $k = m$,

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

¹⁶Note that the weighting matrix is applied to the regularization term in (22) 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.

¹⁷We 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).

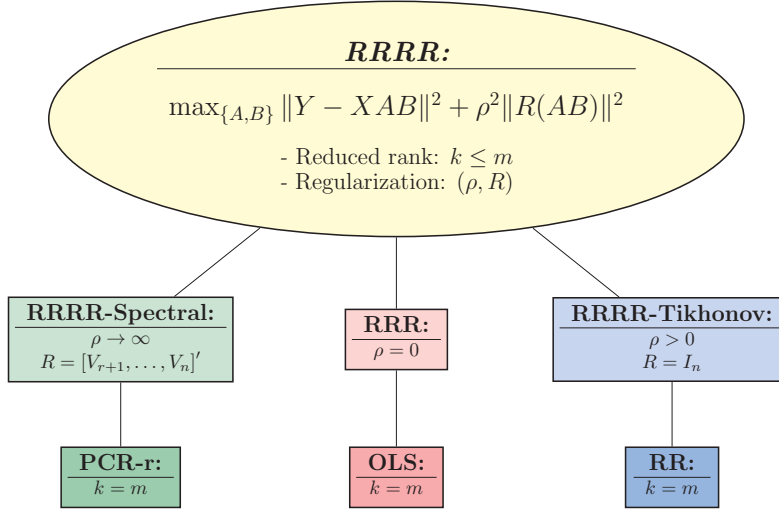


FIGURE 1: Special cases of the regularization scheme considered in this paper.

(22) specializes to

$$\min_{\{A,B\}} \|(\mathbf{Y} - \mathbf{X}\Theta)\|^2 + \rho^2 \|\Theta\|^2 \quad (23)$$

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 OLS coefficient is 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 (22) for a given choice of k , is given by $A^* = \{c_1; \dots; c_k\}$ where c_1, \dots, c_k are the k eigenvectors corresponding to the k largest eigenvalues, $\lambda_1, \dots, \lambda_k$ of the generalized eigenvalue problem*

$$|S_{XY}WS_{YX} - \lambda(S_{XX} + \rho^2 R'R)| = 0 \quad (24)$$

The most common form of the Tikhonov scheme encountered in practice sets $R = I_n$, implying a down weighting of $\sigma_i^2/(\sigma_i^2 + \rho^2)$ of the component of A that lies in the space spanned by the i^{th} right singular vector of \mathbf{X} .¹⁸ We also note that (22) can be given a Bayesian interpretation, albeit in terms of a non-standard prior on the subspace of $m \times n$ matrices of reduced rank k .

¹⁸To see this, set $M_1 = S_{XX} + \rho^2 I_n$ in equation (3)

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 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} , e.g. when $k = \text{rank}(C) < r$ in (17)-(18). 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} .

A natural 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, $\hat{\mathbf{F}}$, are extracted from the n regressors, \mathbf{X} . Second, a reduced rank regression of \mathbf{Y} on $\hat{\mathbf{F}}$ 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 (24) 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 (6)-(7) and let $\mathbf{F} = \mathbf{X}V_r\Sigma_r^{-1}$ be the r principal factors of \mathbf{X} . For $k \leq r$, if $\mathbf{a} \in \mathbb{R}^{r \times k}$ is the matrix of the k principal eigenvectors of

$$0 = |S_{FY}WS'_{FY} - \lambda S_{FF}| \quad (25)$$

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

$$0 = |S_{XX}^\dagger S_{XY}WS'_{XY} - \lambda I_n| \quad (26)$$

where $S_{XX}^\dagger = V_r\Sigma_r^{-2}V_r'$ is the regularized (via spectral truncation) inverse of S_{XX} . Moreover, (26) can be understood as a penalized estimator of the form (24) with $R = V'_{n-r}$ and $\rho \rightarrow \infty$.

The theorem shows that the 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 (22) 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, since they essentially involve an improper prior on the subspace of $\mathbb{R}^{m \times n}$ spanned by the first r right singular vectors, V_r , of \mathbf{X} .

More generally, we note that any set of filter factors can be captured by the formulation (22) since, by setting $R = V \text{diag}(\rho_1, \dots, \rho_n) V'$ in equation (24), we have the one-to-one correspondence $f_i = \sigma_i^2 / (\sigma_i^2 + \rho_i^2)$ 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} .

5. Large n, T Asymptotics and Consistency of the Spectral Truncation Scheme

The setup (17)-(18) is standard in the dynamic factor modeling literature. In particular, Bai and Ng (2002) provide sufficient conditions for consistent identification of the number of factors, r , and Bai (2003) provides the asymptotic distribution of the estimated factors, $F_t = (V_r \Sigma_r^{-1})' X_t$.

Importantly, Stock and Watson (2002) and Bai and Ng (2006) give rate conditions on n and T (e.g. $n, T \rightarrow \infty$ with $n^2/T \rightarrow \infty$) such that the first step factor estimate \hat{F}_t can be taken as given in a second stage regression, i.e. the asymptotic distribution of the second stage OLS coefficients is unaffected, and one is therefore justified in running the reduced rank regression:

$$Y_{t+1} = B\tilde{A}'\hat{F}_t + \tilde{\varepsilon}_t$$

where $\tilde{A} \in \mathbb{R}^{r \times k}$ and $B \in \mathbb{R}^{m \times k}$ so that $\tilde{A}B'$ is a $r \times m$ matrix of reduced rank k . From (27) it is then immediately clear that \tilde{A}, B can be consistently estimated as $T \rightarrow \infty$ for fixed values of r, m .

In the standard PCR approach, a consistent estimate of the regularized OLS estimator $\tilde{\Theta}$ is obtained in the second step from the regression

$$Y_{t+1} = \tilde{\Theta}'\hat{F}_t + \tilde{\varepsilon}_t$$

Under a asymptotic normality assumption, one may use a likelihood ratio test for whether $\tilde{\Theta}$ has reduced rank k and thus satisfies (27), see e.g. Anderson (1963).

6. 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. The problem arises when a single factor loads on disparate variables (e.g. GDP and exchange rates), rendering economic interpretation of the linear combination difficult. In this section, we show how to partially alleviate this shortcoming by requiring the econometrician to *ex-ante* assign each individual regressor to one or more groups. The requirement for each group is that variables belonging to the group should be “alike” in the sense that linear combinations of variables sharing a group membership can be given economic meaning (e.g. “Real Activity”, “Prices”, and “Interest Rates”). In this section, we show how factors may be extracted subject to the constraint that each factor loads only on variables that share a common group membership, so that a factor may be interpreted as e.g. a “Real Activity” factor. Computationally, this can be achieved by imposing zero restrictions on the columns of the reduced rank coefficient matrix A :

Corollary 1 (Constrained Regularized Reduced Rank Regression). *Consider the penalized reduced rank regression problem (22) subject to the constraint $P'A = 0$ for some $P \in \mathbb{R}^{n \times f}$.*

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:¹⁹

$$\min_{\{a, B\}} \|(\mathbf{Y} - \mathbf{X}P^\perp aB)\|^2 + \rho^2 \|RP^\perp aB\|^2, \text{ s.t. } a'P^\perp S_{XX}P^\perp a = I_k \quad (27)$$

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^* equal to the eigenvectors corresponding to the k largest eigenvalues of the $n - f$ dimensional generalized eigenvalue problem²⁰

$$|P^{\perp'} S_{XY} S'_{XY} P^\perp - \lambda P^{\perp'} (S_{XX} + \rho^2 R'R) P^\perp| = 0 \quad (29)$$

Consider a setting with N regressors, each of which can be classified as belonging to one or more of, say, 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 To select “ G_1 ”-factor, set $P'_{\{G_1^\perp\}} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots & & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$

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 and any “group 1 factor” must by definition be orthogonal to $P_{\{G_1^\perp\}}$. The remaining $g_i \times n$ matrices, $P_{\{G_i^\perp\}}$, $i = 2, \dots, 4$ can be similarly defined.

The principal factor is found by solving (28), once for each of the four choices of $P \in \{P_{\{G_1^\perp\}}, \dots, P_{\{G_4^\perp\}}\}$, yielding four candidate factors, one from each group. The principal factor is identified as the group factor associated with the largest eigenvalue, i.e. yielding the greatest explanatory power for \mathbf{Y} .

Subsequent factors are then be extracted iteratively as follows. To compute the $j + 1^{\text{st}}$ factor, first replace \mathbf{Y} by the residuals from regressing \mathbf{Y} on the preceding j factors and extract the principal factor as above. This yields the $j + 1^{\text{st}}$ factor as the group factor with the most explanatory power for the component of \mathbf{Y} not already explained by the preceding j factors.

Note that a feature of the factors extracted in this fashion is that they are *not* orthogonal. This is in many ways natural, as one would not wish to impose that, e.g. “Real Activity” factors

¹⁹E.g. if $P = UDV'$ is the singular value decomposition, then P^\perp can be taken as the last $n - f$ columns of U .

²⁰An alternative approach is to consider penalizing only directions that lie in the admissible space (i.e. satisfying $P'A = 0$), which leads to the eigenvalue problem

$$|P^{\perp'} S_{XY} S'_{XY} P^\perp - \lambda (P^{\perp'} S_{XX} P^\perp + \rho^2 R'R)| = 0 \quad (28)$$

should be orthogonal to “Price” factors. Moreover, even factors from the same group need not be orthogonal, which is consistent with the fact that there seldom is an economic rationale for structural economic factors being orthogonal in practice.

7. 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 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 (10). 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.

For the spectral truncation scheme, several data driven approaches to the choice of regularization threshold based on the classic theory of the spectrum of large random matrices with i.i.d. entries exist. These approaches are appropriate when the regressors, X , can be described by a “signal+noise” model (17)-(18). The version used in this paper, denoted RRRR-SMP, is based on the results of Marcenko and Pastur (1967) and Johnstone (2001). In particular the cutoff is chosen to equal the 95th percentile of the asymptotic distribution of the largest eigenvalue of an i.i.d. Gaussian panel with the same limiting ratio n/T .²¹ Alternatively, the information criteria of Bai and Ng (2002) could also be applied to consistently pick the number of significant factors driving X .

In the case of the Tikhonov scheme (with $R = I_n$), the motivation is the existence of weak factors in X and no formal frequentist scheme for selecting ρ without detailed specification of the properties of the noise. A commonly used (ad-hoc) technique by practitioners is based on the *cross-validation idea*. However, in our setting the implementation of cross validation for selection of ρ (and in principle k) is complicated by both serial dependence in the data and a relatively modest sample size in relation to the number of parameters and we do not pursue this method further in the present paper.²²

Finally, numerous Bayesian approaches to both the choice of ρ (by applying a hierarchical prior with a hyper-parameter controlling the precision of the prior on ρ) and k (essentially using Bayesian model selection or alternatively Bayesian model averaging). We leave this for future work.

²¹See e.g. Patterson, Price, and Reich (2006) and Onatski (2010) among others for examples of applications and the theory.

²²See e.g. Burman, Chow, and Nolan (1994) and Racine (2000) for a discussion of this issue in the context of h -block and $h\nu$ -block cross-validation.

8. 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.

8.1. 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.2 summarizes all models considered in the empirical illustrations.

First, Panel A in Table A.2 depicts models based on a fixed number of regressor components with the r -th row ($r = 1, 2, \dots, n$) and k -th column ($k = 1, 2, \dots, \min(r, m - 1)$ and $k = m$) corresponding to models with r regressor components and k forecasting factors. In particular, for $k = 1, 2, \dots, \min(r, m - 1)$ 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 $r = 1, 2, \dots, n - 1$ 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.^{23,24}

Next, Panel B in Table A.2 presents models relying on data driven regularization of the regressor components stemming random matrix theory. 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 the 10^{-3} percentile of the asymptotic distribution of the largest eigenvalue under the Wishart null in the theory of Marcenko and Pastur (1967) and Johnstone (2001) as illustrated in Figure A.5. We further impose the natural restriction that the chosen number

²³Note 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.

²⁴The PLS and 3PRF estimators referred to throughout are implemented using the MATLAB code accompanying Kelly and Pruitt (2011b) which also introduces the “automatic” regressor terminology.

of regressor components is not smaller than k , which is the minimum number of components required to span k forecasting factors of full rank.²⁵ 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 $k \ll m$, which allows us to study the extent to which just a few common factors may jointly be able to forecast multiple variables of interest.

8.2. Forecasting in Latent Factor Models: A Monte Carlo Study

We investigate the performance of alternative RRRRk-PCr specifications with fixed number r of regressor components and fixed number of forecasting factors $k \leq r$ in a factor model with a spiked population spectrum where spectral truncation, by construction, is the “correct” regularization technique. This controlled setting allows us to more clearly understand the effect of regularization thresholds and grouping on forecasting performance. In particular, we simulate the system:

$$Y_{t+1} = C F_t + \varepsilon_{t+1}, \quad m \geq r \quad (30)$$

$$X_t = \Lambda F_t + \xi_t, \quad n \gg r \quad (31)$$

in the baseline case where the number of forecasting factors is $k = \text{rank}(C) = 1$ and the number of regressor factors is $r = 5$, with the r non-zero eigenvalues of $\Lambda\Lambda'$ calibrated to be similar to the singular values observed in the Stock & Watson (2011) macro data. The predictors X are split into two groups: a small first group containing 10% of the predictors that are driven by the strongest two factors in the system, one of which is the forecasting factor relevant for Y , and a large second group containing the remaining 90% of the predictors that are driven by the weakest three of the five latent factors. This setting represents the case of a small informative group of predictors among many candidate ones. As such, it allows us to assess the ability of different methods to extract a single relevant forecasting factor hidden in a large set of both relevant and irrelevant predictors in a case where a one-factor model is optimal under the null.

We simulate twenty different panels with $T \in \{100, 250\}$ observations, $m \in \{5, 25, 50, 100, 250\}$ outcomes and $n \in \{125, 250\}$ predictors to explore the effects of each of these sample design parameters on the efficacy of RRRR. For each panel, we carry out 10,000 independent replications of the model (31) and compare the performance of different grouped and non-grouped RRRR

²⁵The same restriction explains the lower triangular structure in Panel A of Table A.2 for the RRRR models with a fixed number of regressor components.

specifications to the PCR benchmark in terms of the in-sample minimum angle between the estimated factor space and the true forecasting factor as well as the out-of-sample MSE relative to the infeasible best forecast.²⁶

Under the null (31), the model has 5 factors in the regressor components such that the optimal regularization threshold involves keeping the 5 largest principal components and extracting from these a single forecasting factor, i.e. RRRR1-PC5. This is borne out in Figure A.7 which shows the relative MSE of various RRRR (Panel (a)) and GRRR models (Panel (b)) as well as PCR, relative to the infeasible best forecast. When less than 5 regressor components are retained, Panel (a) shows that all RRRR models as well as PCR suffer dramatically from the omission of vital information contained in the latent forecasting factor. The minimum MSE is obtained when exactly 5 regressor components are retained with a gradual deterioration occurring as more spurious components are included. The deterioration is more evident for PCR which does not benefit from the second stage reduced rank regression which exploits the $m = 5$ outcomes to filter out noisy components. In fact, the most aggressive rank reduction, RRRR1, is clearly beneficial in this case as there is one true forecasting factor, with a gradual deterioration as more spurious forecasting factors are included. However, the RRRR models up to order 5 are uniformly better than the PCR for each level of regularization.

In Panel (b) of Figure A.7, we see the effect of imposing grouping. Clearly imposing a grouping structure that correctly brackets the true factor as belonging to one of the groups provides additional information. This is reflected in uniformly lower MSEs for the grouped RRRR methods. In particular, even as the number of regressor components increases, the GRRRR1 estimator is barely affected by the additional noise as the correct group continues to be selected for the principal factor.

To gain a clearer intuition for these findings, we next analyze in more detail the robustness of the optimal RRRR models (RRRR1 and GRRRR1) versus PCR in the same setting with one true forecasting factor. Each of the methods attempt to estimate the true space spanned by the latent forecasting factor and a natural metric for judging their in-sample efficacy in doing so is to compare the angle between the true forecasting factor space (which is one dimensional in this case) and the estimated factor space which is of fixed dimension k for the RRRRk-PCr/GRRRk-PCr models and of increasing dimension r for the PCR-r models, as the number of regressor components r increases. In Panel (a) of Figure A.8 we see that going from $r = 1$ to $r = 5$ as expected leads to a dramatic reduction in the minimum angle, indicating that all methods improve their ability to capture the true forecasting factor space. However, the RRRR1/GRRRR1 methods do so while preserving the one-dimensional nature of the estimated factor space, while PCR requires a 5-dimensional space. For $r = 5$, the parsimonious one-dimensional space estimated by RRRR1 is only slightly worse in terms of minimal angle than

²⁶For 3PRF/PLS we limit the number of replications to 1,000 due to the considerably larger computational burden of these methods.

for PCR5 while GRRRR1 does slightly better due to a group specification consistent with the true factor structure. As r increases, the angle for PCR mechanically drops as the dimension of the factor space increases. For RRRR1, there is a slight deterioration as the one-dimensional space gets contaminated by noise, which the second step reduced rank regression only partially is able to filter out. The GRRRR1 on the other hand is barely affected since the principal factor continues to be selected from the correct group. A similar pattern is seen in the out-of-sample forecasting performance in Panel (a) of Figure A.9, where GRRRR1 clearly dominates, followed by RRRR1 which benefits from its parsimony compared to PCR, despite the larger in-sample distance angle between the estimated and true factor space.

As the number of outcomes, m , increases, Panel (b) of Figures A.8 and A.9 show that RRRR1 improves markedly due to the increased ability of the second step reduced rank regression to identify the forecasting component contained in the noisy regressor components. GRRRR1 on the other hand exhibits a minimal added benefit since the imposed group structure already reduces the noise in the admissible factor space, even for small m .

Increasing the number of regressors has a beneficial impact on RRRR1 and PCR, as shown in Panel (b) of Figure A.10. However, the impact is limited for PCR as is to be expected based on the results of Bai and Ng (2002) as the ability to identify the true regressor components in the first step depends on $\min(n, T)$ and $n > T$ holds already in our baseline case. However, the GRRRR1 does benefit substantially, as the larger n tends to increase $\min(n_i, T)$ where n_i is the number of regressors that belongs to group i . The improvement for RRRR1 is much smaller as is to be expected. The in-sample ability to better estimate the forecasting factor space is also reflected in the out-of-sample MSEs in Figure A.11. Moreover, the parsimony of RRRR1 and GRRRR1 is beneficial in out-of-sample forecasting and leads to uniformly better performance compared to PCR.

Finally we consider the effect of increasing T in Figures A.12 and A.13. All methods improve their ability to correctly identify the forecasting factor space, with RRRR1 seeing the largest benefit in its ability to filter out noise in the second step reduced rank regression. For GRRRR1 the benefit is smaller as the group structure already provides substantial ability to filter out noise. For all methods, the increase in $\min(n, T)$ is modest, so that the improvement in the ability to identify regressor components is limited, as reflected in an almost unchanged PCR performance.

The robustness of RRRR to an incorrectly specified number of forecasting factors k and number of regressor components r in our Monte Carlo study suggests a simple recipe when the true model is unknown: one can apply a suitable set of reduced rank regressions (rather than linear regression) to some chosen principal components of the regressors. Moreover, we see a potentially substantial benefit from correctly imposing a group structure on the problem, with the added benefit of interpretability of the resulting factors. This allows the econometrician to ex-ante impose some discipline on the statistical factor extraction problem dictated by eco-

conomic theory. However, we stress that “incorrect” grouping may lead to a substantial loss of performance, and thus the grouping of regressors must be carried out carefully. Our empirical illustrations on real data further demonstrate the effectiveness of applying reduced rank rather than linear regression to the principal components of the available regressors.

8.3. 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.2.²⁷ Following Stock and Watson (2011), we report distributions of relative RMSE by forecasting method relative to the PCR-5 benchmark. Table A.3 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 Table A.4 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.3, 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.3, 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 Marcenko and Pastur (1967) and Johnstone (2001). As evident from the first row in Panel C of Table A.3, 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

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

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

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

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.4 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 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.

It is interesting to observe also 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.3 and A.4 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.

Finally, simple grouping of the available predictors into two disjoint groups of macro and financial variables in accordance with the categorization by Stock and Watson (2011) allows us to shed light on the economic interpretation of the extracted forecasting factors as revealed by RRRR with the corresponding variable grouping constraints. In particular, we find that while such coarse grouping of the variables does not lead to much improvement or deterioration of the best performing RRRR5-PC10 model, it reveals that the strongest among the extracted

five forecasting factors is composed by macro variables almost 100% of the time, while the weakest is composed by finance variables more than 60% of the time. More detailed analysis of the economic factor dynamics and composition implied by grouped RRRR is the subject of an ongoing study.

8.4. 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.5); (ii) forwards (table A.6); (iii) forward slopes (table A.7); (iv) yield curve slopes (table A.8); (v) the union of cycles and yield curve slopes (table A.9).³⁰ Although there are only about 15 predictors, the design matrix, \mathbf{X} , is extremely ill-conditioned as shown in Figure A.2, 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 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

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

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

predictive power of yield curve slopes (table A.8) is as strong as the predictive power of cycles (table A.5), while forward slopes (table A.7) and forwards (table A.6) 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.9), 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.3-A.4), 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 the combined yield slopes and cycles predictor set.

9. 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 Bayesian interpretation in terms of a precision prior on the regression slopes, albeit with a non-standard prior. Moreover, in a purely frequentist setting, we have shown how to motivate the choice of regularization scheme in terms of assumptions about the signal-to-noise ratio of certain dimensions of the data. In the case of spectral truncation regularization, one may also appeal to the existing literature on using random matrix theory for noise filtering 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 (partly) forecastable. Compared to principal component regression (PCR), RRRR often 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. In effect, as an important special case, our RRRR approach generalizes principal component regression by applying reduced rank rather than linear regression to the principal components of the regressors, thereby disentangling the forecasting factors driving the outcomes from the factor structure in the predictors. Moreover, the computational burden is no greater than PCR for large n, T and, in our empirical investigation, there seems to be limited (if any) downside to applying a reduced rank rather than linear regression in the second step of a principal component regression analysis.

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., 1963. Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics* 34 (1), pp. 122–148.
- 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.
- Burnham, A. J., Viveros, R., MacGregor, J. F., 1996. Frameworks for latent variable multivariate regression. *Journal of Chemometrics* 10 (1), 31–45.
- Carriero, A., Kapetanios, G., Marcellino, M., 2011. Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics* 26 (5), 735–761.
URL <http://dx.doi.org/10.1002/jae.1150>
- Chen, K., Chan, K.-S., Stenseth, N. C., 2012. Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74 (2), 203–221.
- 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., 2011a. Market expectations in the cross section of present values. Working Paper 11-08, Chicago Booth School of Business.
- Kelly, B., Pruitt, S., 2011b. 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, [jce: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>
- Rao, C. R., 1964. The use and interpretation of principal component analysis in applied research. *Sankhya: The*

- Indian Journal of Statistics, Series A 26 (4), pp. 329–358.
URL <http://www.jstor.org/stable/25049339>
- Stock, J., Watson, M., 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 1167–1179.
- 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.
- Vinod, H., 1976. Canonical ridge and econometrics of joint production. *Journal of Econometrics* 4 (2), 147 – 166.
- Wold, H., 1966. Estimation of principal components and related models by iterative least squares. In: Krishnaiah, P. R. (Ed.), *Multivariate analysis*. Academic Press, New York, pp. 391–420.

Appendix A. Figures and Tables

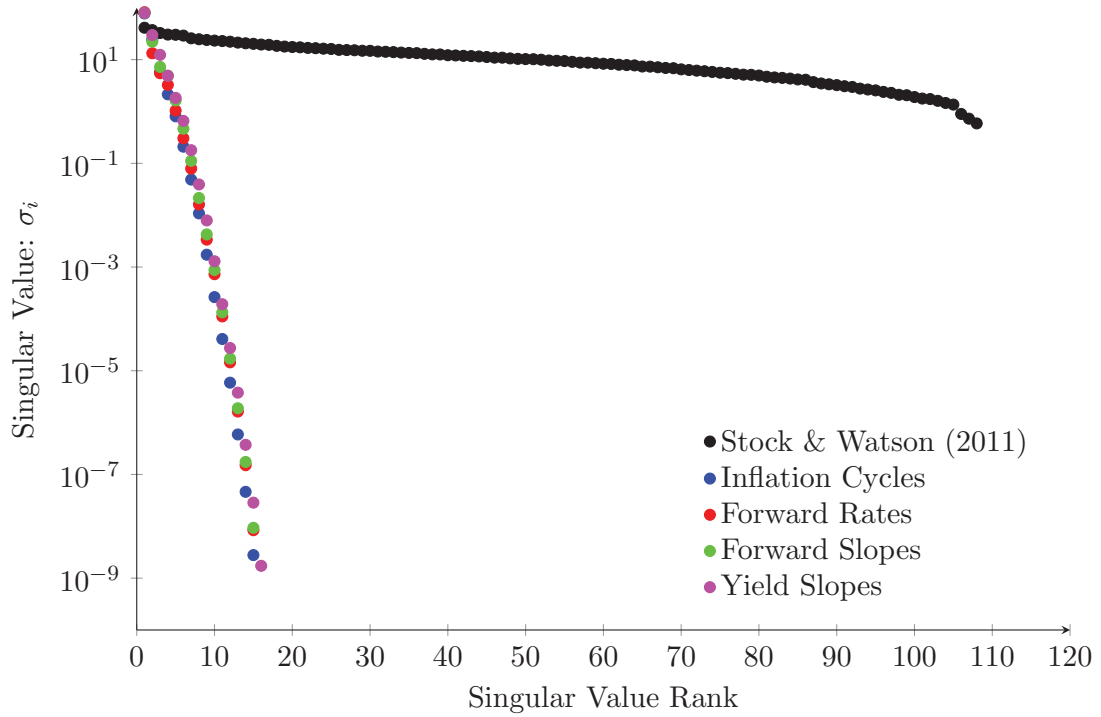


FIGURE A.2: The singular values of the Stock and Watson (2011) macro data and the four sets of bond excess return predictors considered: The Cieslak and Povala (2011) inflation cycles, the Forward rates, Forward slopes (with respect to the 1 month rate), the current yield slopes (with respect to the 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).

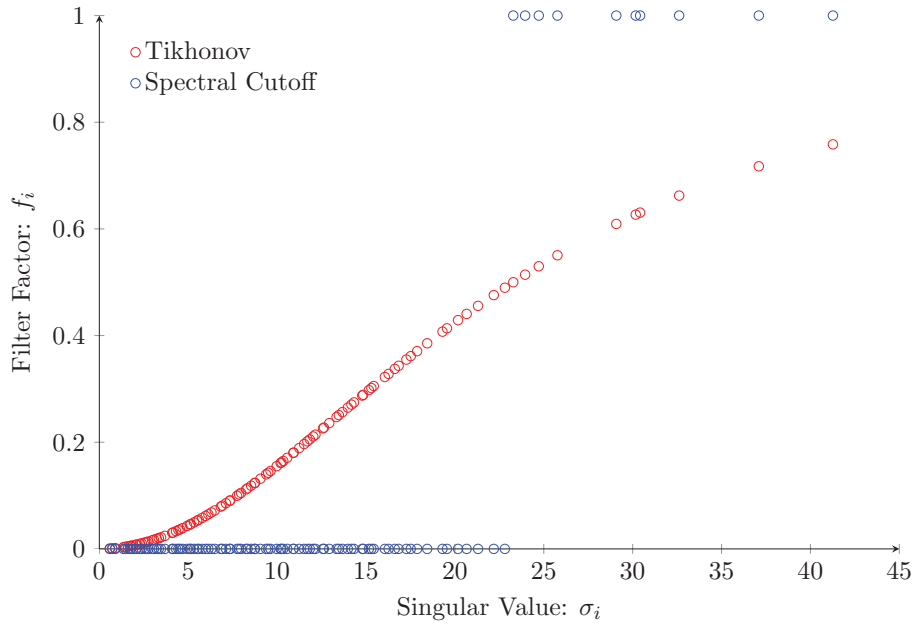


FIGURE A.3: The filter factors f_i as a function of the size of the singular value σ_i of the Stock and Watson (2011) macro data for the two regularization schemes considered. In each case the regularization parameter is set to $\rho = \sigma_{10}$, the tenth largest singular value.

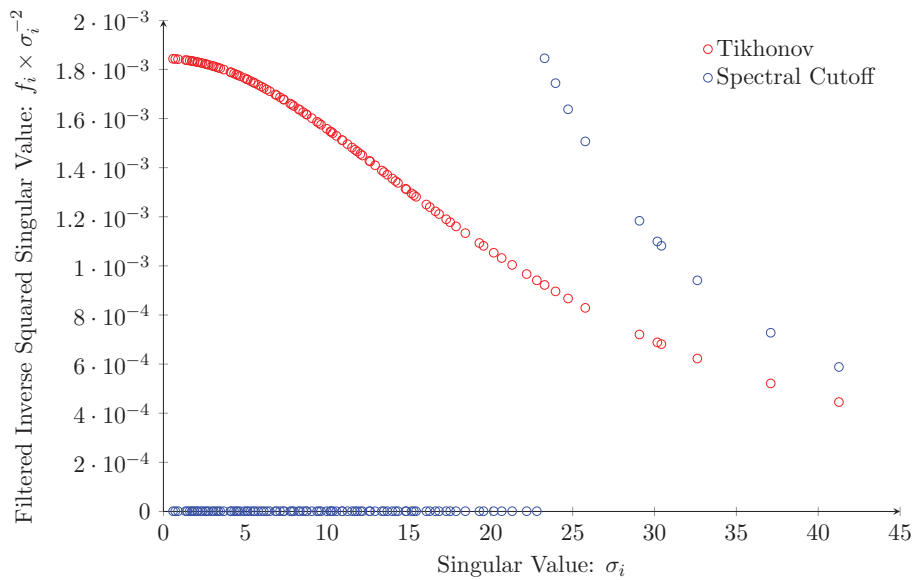


FIGURE A.4: The filtered reciprocal singular values of the Stock and Watson (2011) dataset of 108 macroeconomic variables. The spectral truncation filter works by setting all singular values of \mathbf{X} that fall below a given cut-off level to zero while the Tikhonov scheme down weights small singular values. In each case the regularization parameter is set to $\rho = \sigma_{10}$, the tenth largest singular value.

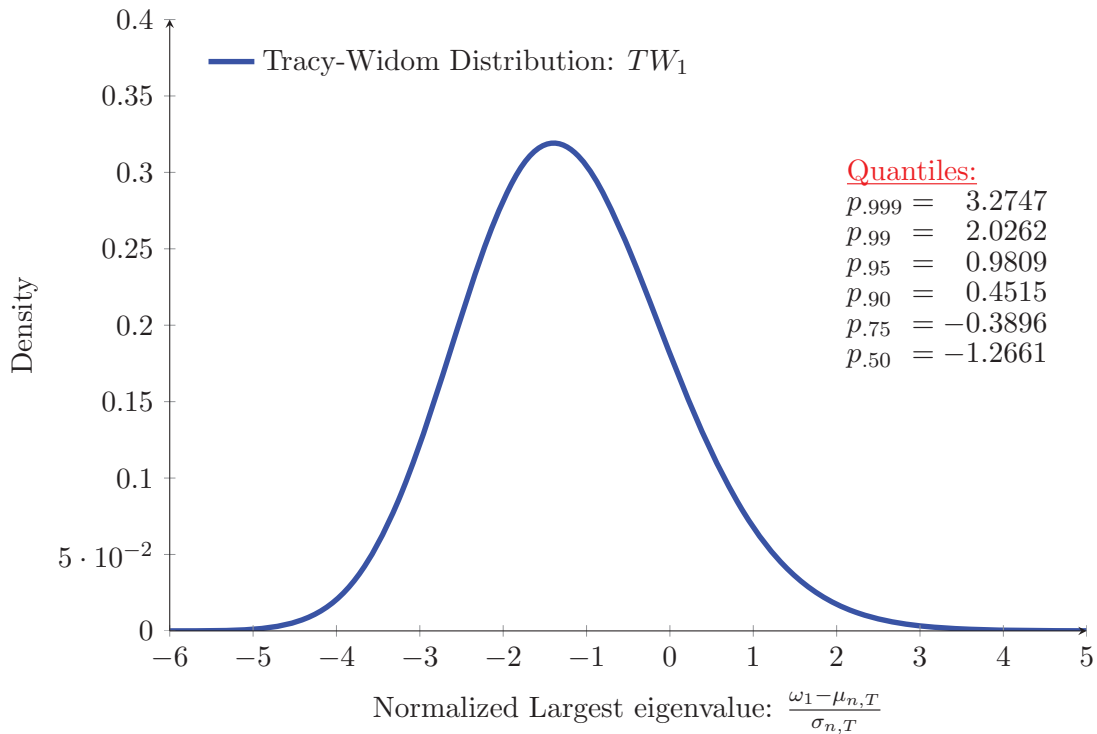


FIGURE A.5: The limiting Tracy-Widom distribution corresponding to the first β -ensemble (Gaussian Orthogonal Ensemble, c.f. Johnstone (2001)), for the normalized largest eigenvalue of the noise covariance matrix. The TW_1 distribution function is not known in closed form but given by $TW_1(s) = \exp \left\{ -\frac{1}{2} \int_s^\infty q(x) dx \right\}$, where $q(\cdot)$ satisfies the Painlevé type II equations: $q'' = xq + 2q^3$ with boundary condition $\lim_{x \rightarrow \infty} [q(x) - Ai(x)] = 0$ and $Ai(\cdot)$ is the Airy function. The solution can be found numerically to any desired accuracy using an ODE solver.

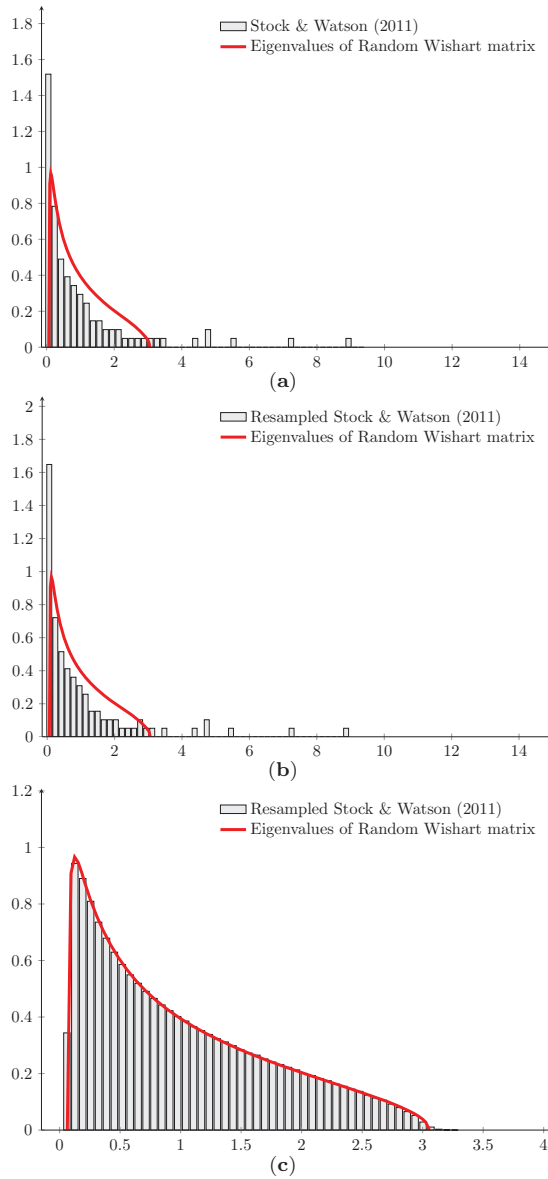


FIGURE A.6: **Eigenvalues of the Stock and Watson (2011) S_{XX} matrix.** In each panel the red curve shows the asymptotic distribution of the eigenvalues of the covariance matrix of a panel of i.i.d. $N(0,1)$ random variables with $N/T = 108/198$ as in the Stock and Watson (2011) dataset. **Panel a:** 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.

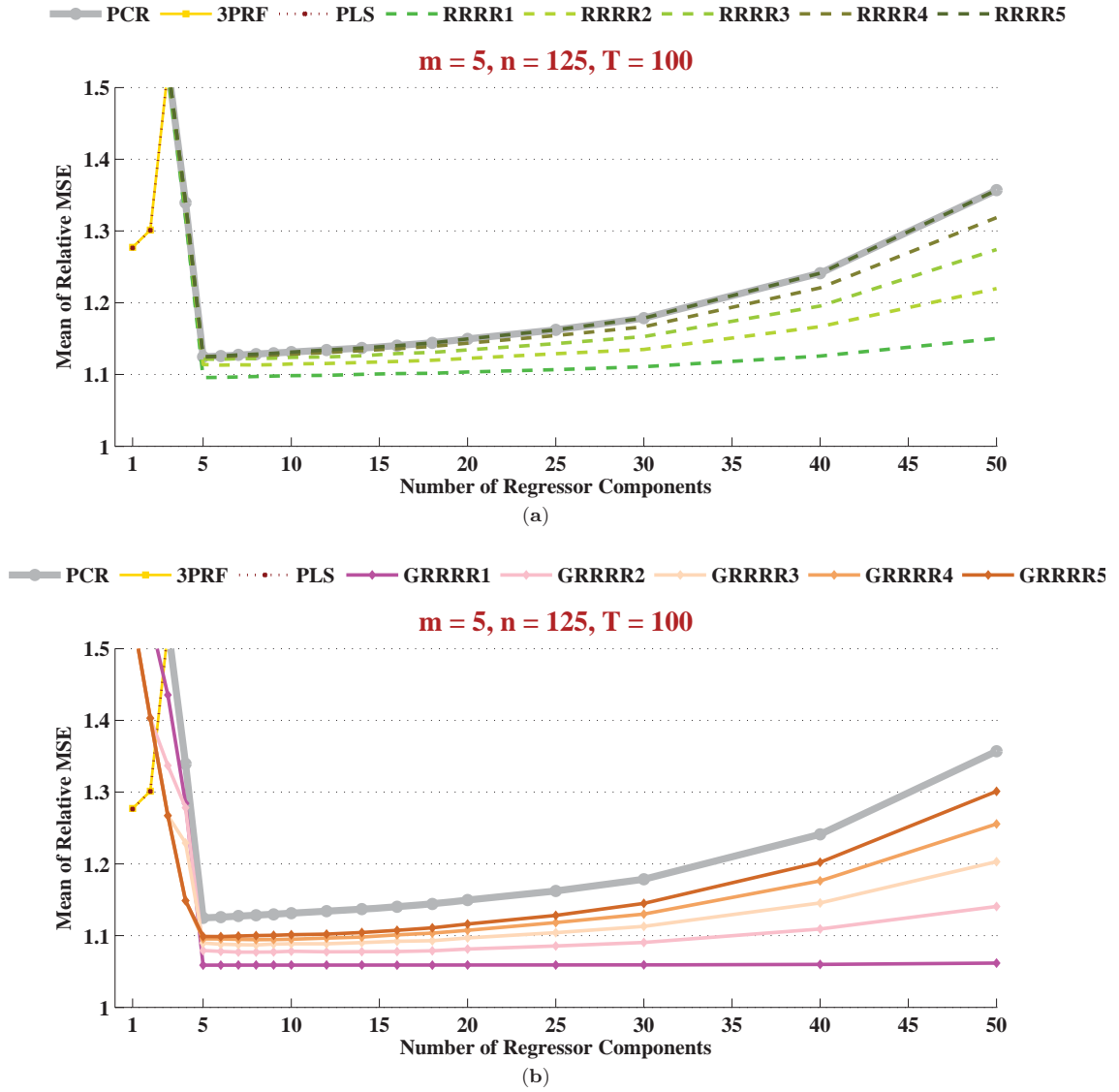


FIGURE A.7: Relative MSE plots with and without variable grouping constraints for RRRR factor interpretability in Monte Carlo datasets. In each panel we plot relative MSE as a function of number of regressor components for competing methods described in the text and our model taxonomy table A.2 and Monte Carlo datasets obtained as described in section 8.2 with sample size $T = 100$ observations, $m = 5$ outcomes, and $n = 125$ predictors. **Panel (a):** Unconstrained RRRR. **Panel (b):** RRRR with variable grouping constraints for factor interpretability.

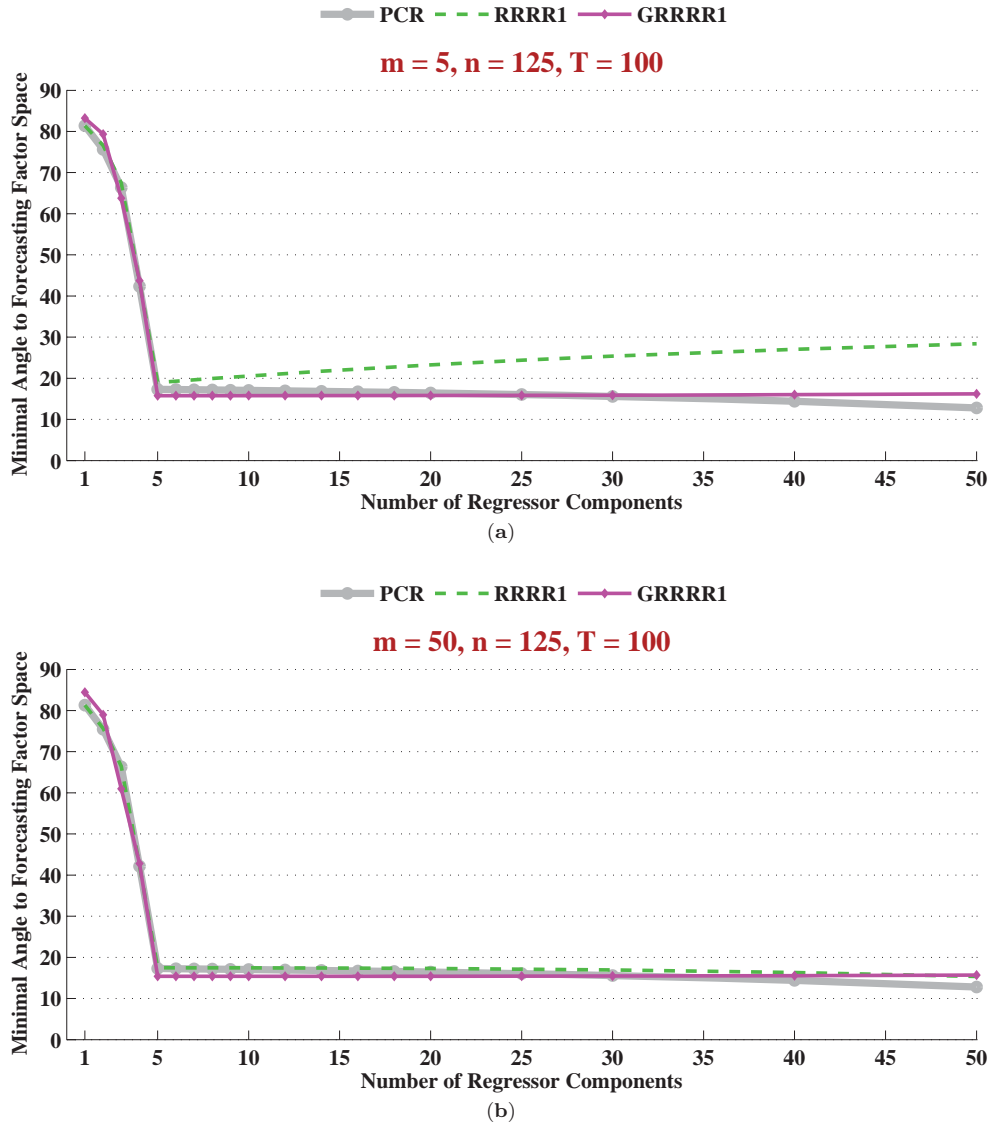
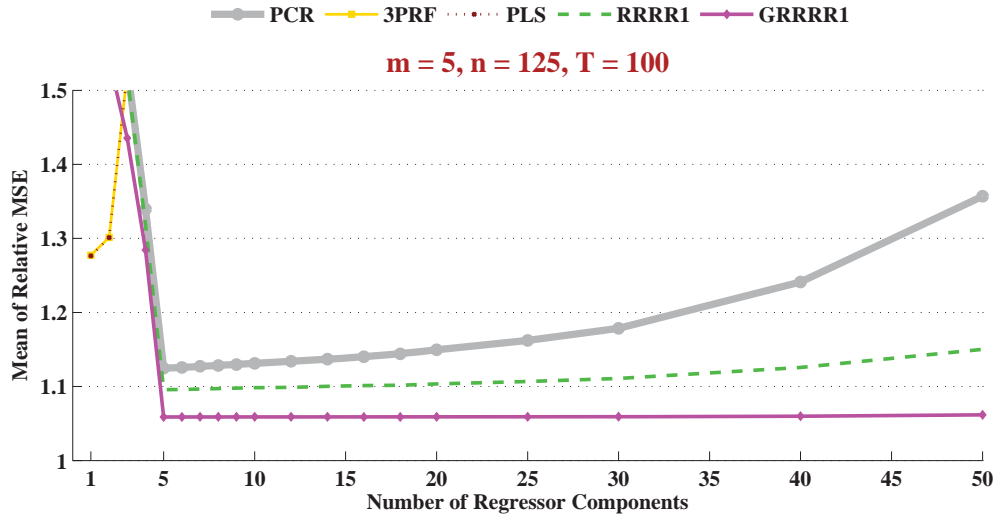
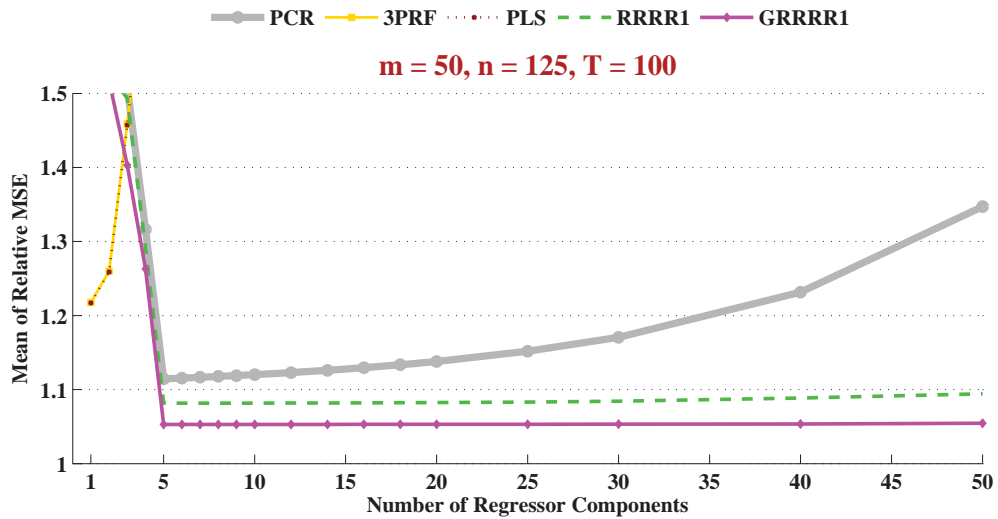


FIGURE A.8: Plot of minimal angles to the forecasting factor space for small versus large number m of outcomes Y in Monte Carlo datasets. In each panel we plot computed minimal angles to the forecasting factor space as a function of number of regressor components for competing methods described in the text and our model taxonomy table A.2 and Monte Carlo datasets obtained as described in section 8.2. **Panel (a):** Monte Carlo dataset of size $T = 100$ observations, $m = 5$ outcomes, and $n = 125$ predictors. **Panel (b):** Monte Carlo dataset of size $T = 100$ observations, $m = 50$ outcomes, and $n = 125$ predictors.

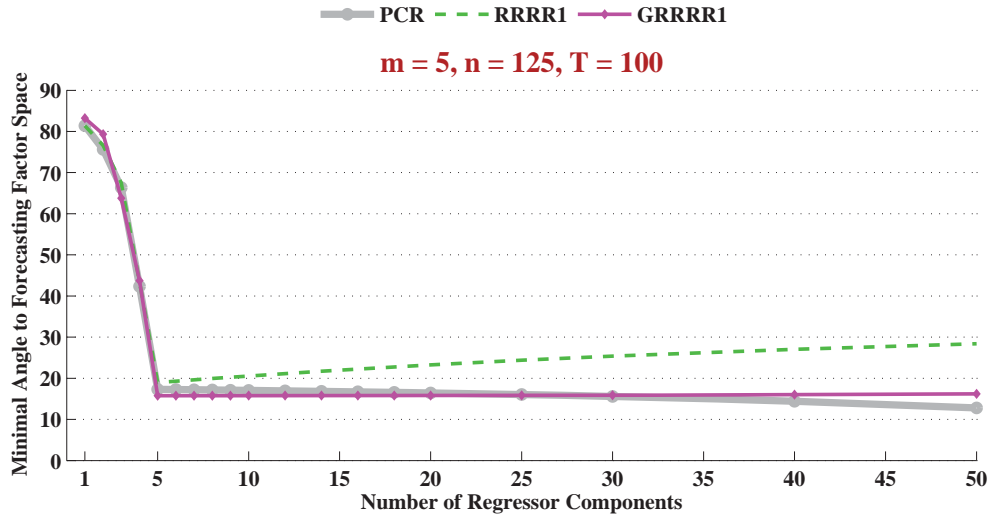


(a)

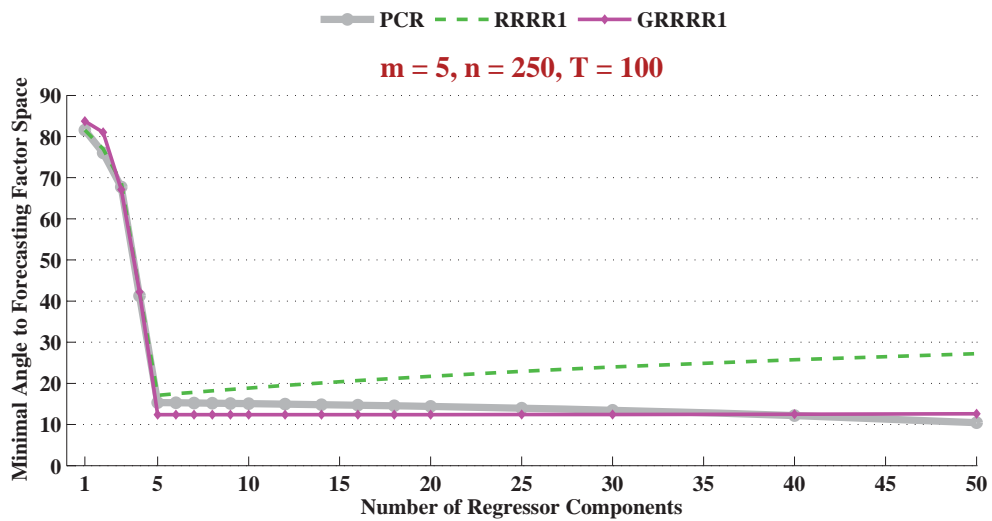


(b)

FIGURE A.9: **Relative MSE plots for small versus large number m of outcomes Y in Monte Carlo datasets.** In each panel we plot relative MSE as a function of number of regressor components for competing methods described in the text and our model taxonomy table A.2 and Monte Carlo datasets obtained as described in section 8.2. **Panel (a):** Monte Carlo dataset of size $T = 100$ observations, $m = 5$ outcomes, and $n = 125$ predictors. **Panel (b):** Monte Carlo dataset of size $T = 100$ observations, $m = 50$ outcomes, and $n = 125$ predictors.



(a)



(b)

FIGURE A.10: Plot of minimal angles to the forecasting factor space for small versus large number n of predictors X in Monte Carlo datasets. In each panel we plot computed minimal angles to the forecasting factor space as a function of number of regressor components for competing methods described in the text and our model taxonomy table A.2 and Monte Carlo datasets obtained as described in section 8.2. **Panel (a)**: Monte Carlo dataset of size $T = 100$ observations, $m = 5$ outcomes, and $n = 125$ predictors. **Panel (b)**: Monte Carlo dataset of size $T = 100$ observations, $m = 5$ outcomes, and $n = 250$ predictors.

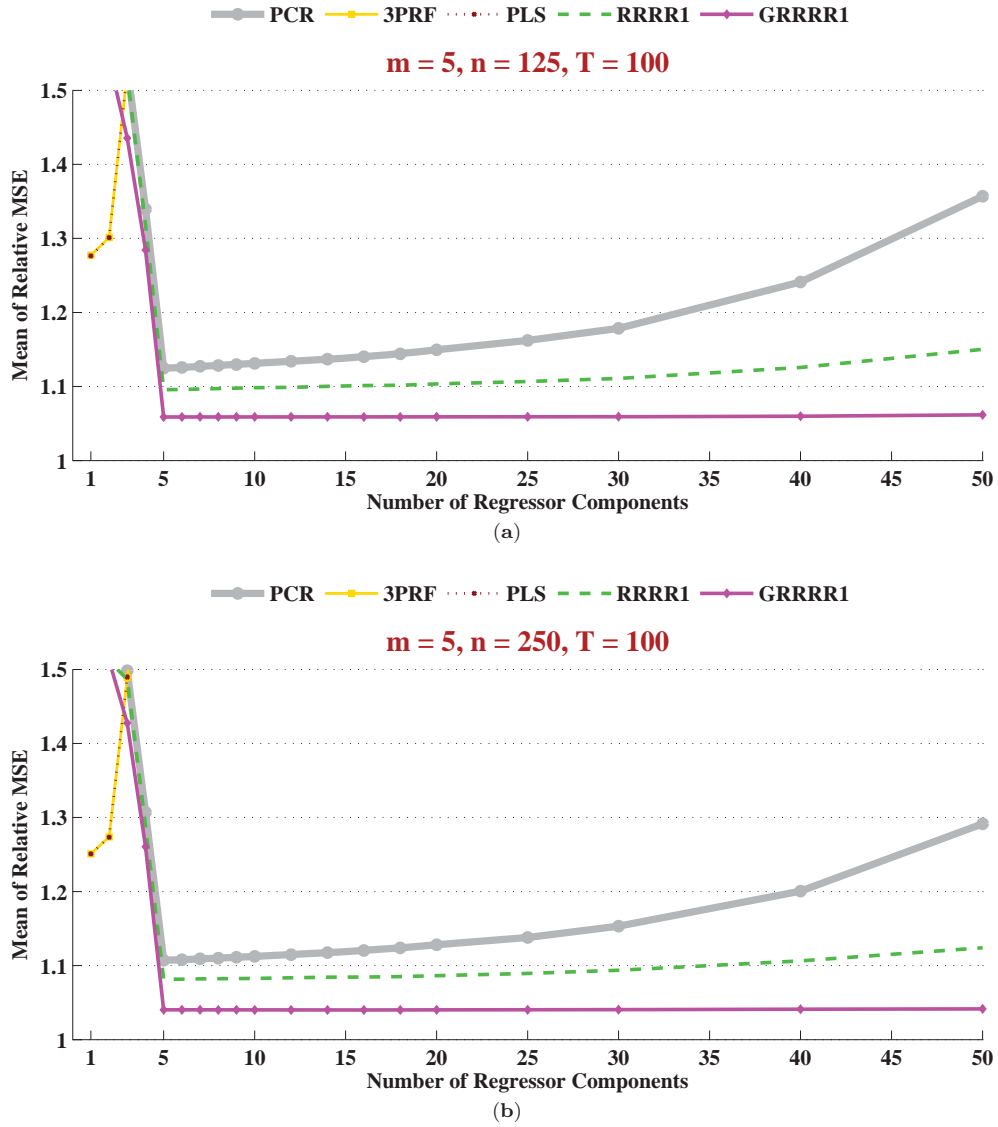


FIGURE A.11: **Relative MSE plots for small versus large number n of predictors X in Monte Carlo datasets.** In each panel we plot relative MSE as a function of number of regressor components for competing methods described in the text and our model taxonomy table A.2 and Monte Carlo datasets obtained as described in section 8.2. **Panel (a):** Monte Carlo dataset of size $T = 100$ observations, $m = 5$ outcomes, and $n = 125$ predictors. **Panel (b):** Monte Carlo dataset of size $T = 100$ observations, $m = 5$ outcomes, and $n = 250$ predictors.

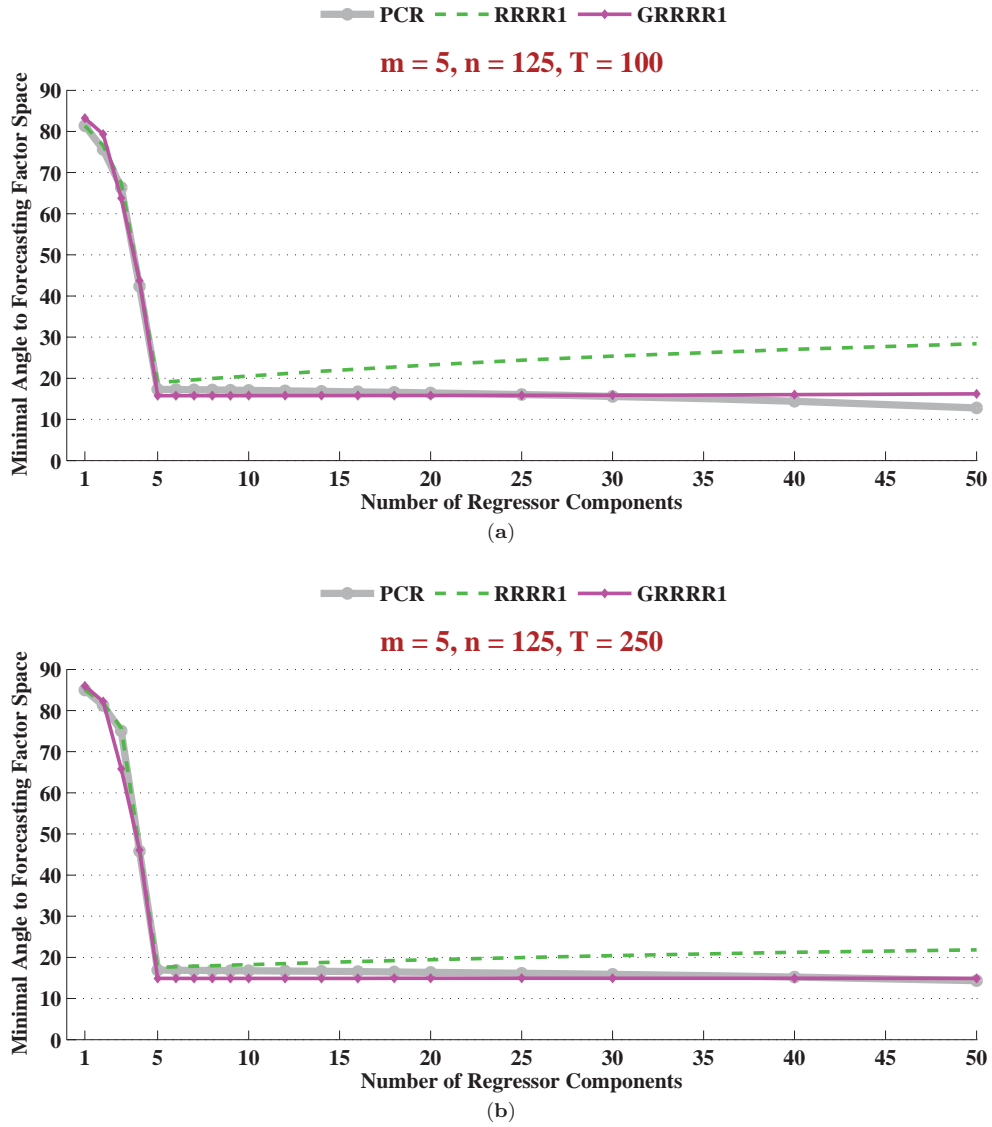


FIGURE A.12: Plot of minimal angles to the forecasting factor space for small versus large sample size T in Monte Carlo datasets. In each panel we plot computed minimal angles to the forecasting factor space as a function of number of regressor components for competing methods described in the text and our model taxonomy table A.2 and Monte Carlo datasets obtained as described in section 8.2. **Panel (a)**: Monte Carlo dataset of size $T = 100$ observations, $m = 5$ outcomes, and $n = 125$ predictors. **Panel (b)**: Monte Carlo dataset of size $T = 250$ observations, $m = 5$ outcomes, and $n = 125$ predictors.

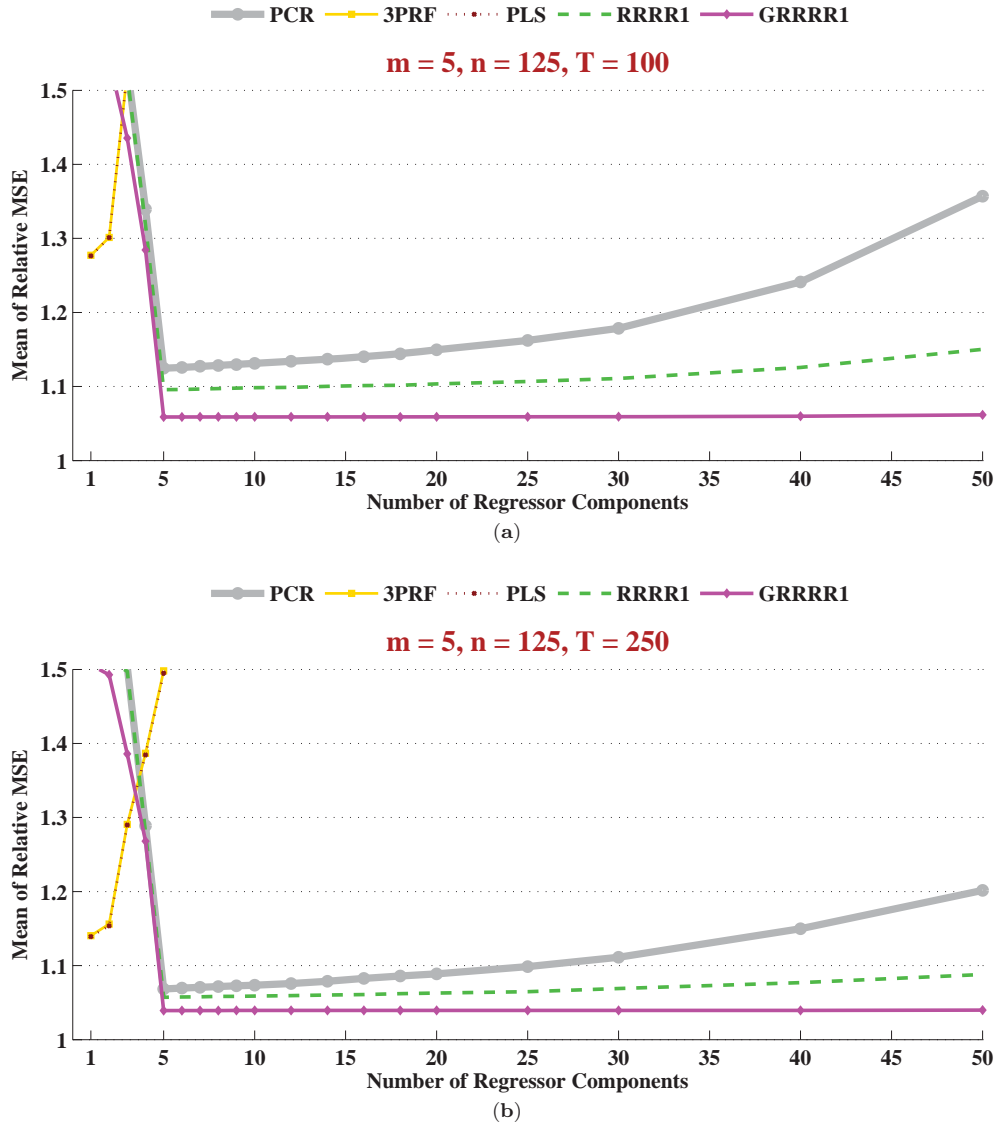


FIGURE A.13: **Relative MSE plots for small versus large sample size T in Monte Carlo datasets.** In each panel we plot relative MSE as a function of number of regressor components for competing methods described in the text and our model taxonomy table A.2 and Monte Carlo datasets obtained as described in section 8.2. **Panel (a):** Monte Carlo dataset of size $T = 100$ observations, $m = 5$ outcomes, and $n = 125$ predictors. **Panel (b):** Monte Carlo dataset of size $T = 250$ observations, $m = 5$ outcomes, and $n = 125$ predictors.

TABLE A.2: **Taxonomy of forecasting models.** We present a taxonomy of forecasting models for any number of forecasting factors $1, 2, \dots, M$ and any number of regressor components $1, 2, \dots, N$. 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.3: **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.2.

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: Naive 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.4: **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.2.

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.5: **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.2.

Out-of-sample R^2	Bond Excess Returns														
	Models	R^2 ⁽²⁾	R^2 ⁽³⁾	R^2 ⁽⁴⁾	R^2 ⁽⁵⁾	R^2 ⁽⁶⁾	R^2 ⁽⁷⁾	R^2 ⁽⁸⁾	R^2 ⁽⁹⁾	R^2 ⁽¹⁰⁾	R^2 ⁽¹¹⁾	R^2 ⁽¹²⁾	R^2 ⁽¹³⁾	R^2 ⁽¹⁴⁾	R^2 ⁽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	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.6: **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.2.

Out-of-sample R^2	Bond Excess Returns														
	Models	$r_x^{(2)}$	$r_x^{(3)}$	$r_x^{(4)}$	$r_x^{(5)}$	$r_x^{(6)}$	$r_x^{(7)}$	$r_x^{(8)}$	$r_x^{(9)}$	$r_x^{(10)}$	$r_x^{(11)}$	$r_x^{(12)}$	$r_x^{(13)}$	$r_x^{(14)}$	$r_x^{(15)}$
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	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.7: **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.2.

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	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	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.8: **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.2.

Out-of-sample R^2	Bond Excess Returns														
	Models	$\Gamma^X(2)$	$\Gamma^X(3)$	$\Gamma^X(4)$	$\Gamma^X(5)$	$\Gamma^X(6)$	$\Gamma^X(7)$	$\Gamma^X(8)$	$\Gamma^X(9)$	$\Gamma^X(10)$	$\Gamma^X(11)$	$\Gamma^X(12)$	$\Gamma^X(13)$	$\Gamma^X(14)$	$\Gamma^X(15)$
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	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	0.027	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.9: **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.2.

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	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	0.004	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	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	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	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	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	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	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	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	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	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	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	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	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	0.066	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	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	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	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	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	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	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	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	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	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	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	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	-0.336

Appendix B. Proofs

Appendix B.1. Proofs of RRRR results

Lemma 1. *Let Γ, Λ be positive semidefinite $n \times n$ matrices and Λ be invertible, then*

$$A^* = \arg \max_{\{A \in \mathbb{R}^{n \times k} : A' \Lambda A = I_{k \times k}\}} \text{tr}\{A' \Gamma A (A' \Lambda A)^{-1}\} \quad (\text{B.1})$$

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

$$|\Gamma - \lambda \Lambda| = 0 \quad (\text{B.2})$$

PROOF. Follows from the fact that if (λ_i, c_i) is an eigenvalue-eigenvector pair of (B.2), then $\lambda_i \Lambda c_i = \Gamma c_i$, and $C = (c_1, \dots, c_n)$ 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.1) 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.2). 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. \square

PROOF OF PROPOSITION 1. The optimal A solves

$$\begin{aligned} & \min_A \text{tr}\{W^{1/2} S_{Y Y} W^{1/2} - W^{1/2} S_{Y X} A (A' [S_{X X} - \rho^2 R' R] A)^{-1} A' S_{X Y} W^{1/2}\} \\ & = \max_A \text{tr}\{A' S_{X Y} W S_{Y X} A [A' (S_{X X} - \rho^2 R' R) A]^{-1}\} \end{aligned}$$

The statement of the proposition now follows from Lemma 1 with $\Gamma = S_{X Y} W S_{Y X}$ and $\Lambda = S_{X X} - \rho^2 R' R$. \square

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

$$\max_a \text{tr}\{(P^\perp a)' S_{X Y} S_{Y X} (P^\perp a) ((P^\perp a)' [S_{X X} - \rho^2 R' R] (P^\perp a))^{-1}\}$$

The main result of the proposition now follows directly from Lemma 1 with the $(n - f) \times (n - f)$ matrices $\Gamma = P^{\perp'} S_{X Y} S_{Y X} P^\perp$ and $\Lambda = P^{\perp'} [S_{X X} - \rho^2 R' R] P^\perp$. \square

PROOF OF COROLLARY 2. Let the singular value decomposition of \mathbf{X} be given by (6)-(7), 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

$$0 = |S_{F Y} S'_{F Y} - \lambda S_{F F}| \Rightarrow 0 = |S_{U_r Y} S'_{U_r Y} - \lambda I_r| \quad (\text{B.3})$$

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

$$0 = |V_r \Sigma_r^{-2} V_r' S_{XY} S_{XY}' - \lambda I_n| \Rightarrow 0 = |V_r \Sigma_r^{-1} S_{U_r Y} [S_{U_r Y}' \Sigma_r V_r' + S_{U_{n-r} Y}' \Sigma_{n-r} V_{n-r}'] - \lambda I_n| \quad (\text{B.4})$$

To see if the two solutions are identical, we take an eigenvalue-eigenvector pair (λ_i, a_i) of (B.3) and check whether $\tilde{A}_i = V_r \Sigma_r^{-1} a_i$ is an eigenvalue of (B.4) corresponding to the eigenvalue λ_i :

$$V_r \Sigma_r^{-1} S_{U_r Y} [S_{U_r Y}' \Sigma_r V_r' + S_{U_{n-r} Y}' \Sigma_{n-r} V_{n-r}'] \tilde{A}_i = V_r \Sigma_r^{-1} S_{U_r Y} [S_{U_r Y}' \Sigma_r V_r' + S_{U_{n-r} Y} a_i] \quad (\text{B.5})$$

$$= \lambda_i V_r \Sigma_r^{-1} a_i = \lambda_i \tilde{A}_i \quad (\text{B.6})$$

where the last equality follows from the fact that (λ_i, a_i) is an eigenvalue-eigenvector pair for (B.3). \square