Optimal Predictive Factor: Learning about the Market Risk Premium from a Large Set of Predictors

Fuwei Jiang* Kunpeng Li† Guoshi Tong‡ and Guofu Zhou§

September 2024

Abstract

In the age of big data, PCA and PLS are widely used in finance for dimension deduction to identify a few predictive factors. In this paper, we make a surprising discovery that the dimension deduction can achieve the optimal lower bound of one in an equivalent model. We propose a supervised learning method to find the optimal predictive factor, which is the best linear combination of a large set of predictors. Our approach outperforms alternative dimension reduction techniques, such as PCA and PLS, theoretically. Just as an efficient portfolio highlights which assets are crucial in the pricing kernel, our optimal predictive factor pinpoints the most significant combination of predictors in forecasting. When applied to predicting the market risk premium, our method outperforms empirically not only both the PCA and PLS, but also all those state of are machine learning methods use by Dong, Li, Rapach, and Zhou (2022). Furthermore, our method reveals a set of novel predictors. Additionally, we identify the optimal predictive factors for marker volatility, bond excess return and macroeconomic aggregate, and find our method continues to perform the best.

Key Words: Partial least square, Principal components analysis, Supervised dimension reduction, Expectation-Maximization method

JEL: C31; C33.

Jiang: jfuwei@gmail.com (Xiamen University); Li: likp.07@gmail.com (Capital University of Economics and Business); Tong: gstong@fudan.edu.cn (Fudan University); Zhou: zhou@wustl.edu (Washington University in St. Louis)

1 Introduction

In today's era of big data, market participants, policymakers, and economists have access to a vast array of predictive information, enabling them to forecast key economic indicators such as the equity risk premium, interest rates, and inflation. While this abundance of data offers valuable insights, it also presents a 'curse of dimensionality', as traditional multivariate regression models are prone to overfitting. Therefore, dimension reduction is essential for effectively utilizing this wealth of information, with principal component analysis (PCA) and partial least squares (PLS) being the primary tools. Both PCA and PLS extract factors from a large set of predictors and use a few of these factors for prediction instead of all the predictors, say Kelly and Pruitt (2013, 2015); Light, Maslov, and Rytchkov (2017); Huang, Jiang, Tu, and Zhou (2015); Gu, Kelly, and Xiu (2020); Bianchi, Büchner, and Tamoni (2021); Dong et al. (2022), among others. The general approach involve identifying multiple, typically the first few, factors and then use them for prediction. Since it is often the case of big data or a large number of predictors in applications, both PCA and PLS are popular in finance.

In this paper, we propose a novel method to obtain a single optimal predictive factor in a predictive regression, defined as the best linear combination of a large set of predictors. Our approach assumes the same data-generating process with a factor structure as the basis for PCA and PLS. However, unlike PCA and PLS, our method is a one-step process that directly estimates the optimal weight for each predictor. Economically, this is a surprising and important discovery as both the PCA and PLS can be replaced by a method reducing dimension to one. Theoretically, due to much fewer parameters to be estimated, we show that the forecasting errors of our single factor are significantly smaller than those of multiple PCA and PLS factors.

Our single optimal predictive factor is conceptually similar to the well-known optimal portfolio. While PCA and PLS can extract a set of factors from asset returns, the optimal portfolio of all assets can be approximated from the optimal portfolio of these extracted factors. However, the optimal portfolio of all assets is a linear combination of these assets. If the optimal weight for each asset can be estimated, the optimal portfolio can be determined in one step. Our approach suggests that such a one-step procedure is feasible for predictions, whereas achieving this in high-dimensional asset pricing might still be an open challenge.

We prove that, specifically, regardless of the number of latent factors in the factor model of predictors, there always exists an equivalent one-relevant-factor model as long as the link between latent factors and the predictive target is linear. This observation allows us to extract a single optimal factor for prediction, and thus to achieve the maximum level of dimension reduction. To estimate the single optimal factor, we consider an augmented factor model that

include both predictors variables and the target. As the target loads only on the single relevant factor, the augmented model needs to be restricted with few zero coefficients on the loading matrix. We estimate such restricted model through maximum likelihood, with the aid of expectation maximization (EM) algorithm, see Bai and Li (2012). The resultant estimator is termed supervised EM (SEM), as the direction of dimension reduction is supervised by the zero coefficients. We derive its asymptotic properties and also provide simulations to investigate its finite sample performance. Both the asymptotic theory and simulation evidence confirm that SEM extracts the optional predictive factor correctly to minimize the mean-squared forecasting error.

For comparison, our study also derive the asymptotic properties of PLS under the factor structure. Since Wold (1966, 1975), PLS has been widely used across various scientific disciplines, such as in management by Bork, Møller, and Pedersen (2020) and in accounting by Lu and Nikolaev (2022). In finance, its use has grown significantly since the work of Kelly and Pruitt (2013, 2015). Based on the seminal works by Bai and Ng (2002), Bai (2003) demonstrates that PCA could consistently estimate all latent factors up to a rotation. However, we show that the target relevant factor can not be consistently estimated by PLS. The root cause lies in that identification condition underlying a target relevant factor model differs from those in the standard large dimensional factor analysis. We also confirm these results through simulation evidence.

We apply the novel SEM estimator to the prediction of US equity risk premium, a central and widely studied topic in asset pricing. We extract the optimal predictive factor from a list of 50 macro variables, investor sentiment proxies, categorical economic policy uncertainty measures, and survey of professional forecasters on expected macro conditions. In an out-of-sample forecasting analysis, we find the SEM factor frequently place large weights on dividend yield, uncertainty of fiscal policy and expected inflation. In comparison, the first PLS factor emphasize more on general economic policy uncertainty, expected real DGP growth and recession probability; and the first PCA factor has less extreme weights. In terms of statistical accuracy, we find that the out-of-sample R^2_{OS} of our SEM more than double those of the multiple PCA or PLS factors, and the out performance has been predominant over the majority of evaluation period. Furthermore, the better return forecast accuracy also carries over to asset allocation decisions to a mean variance investor. The certainty equivalent gains by using our SEM forecasts could exceed the best among PCA and PLS forecasts by 1.4% a year.

We also apply the SEM method to predict US aggregate equity market volatility, long term Treasury bond excess return and industrial production growth. Again, we compare the accuracy of forecasts by a single SEM factor against the PCA and PLS forecasts. Additionally, we also compare with a list of forecast combination and machine learning methods to aggregate information from the set of predictors. Particularly, we include the state of art methods, such

as predictor average, simple combination, elastic net regression and combination with elastic net, which have been used by Dong et al. (2022) in their forecasts. Our evidence is still in support of the forecsts by the single SEM factor. While combination and machine learning tend to do well in predicting industrial production growth, the out-of-sample R^2_{OS} by SEM remains the highest, suggesting its robustness across a variety of forecast applications.

Our paper contributes to the literature on dimension reduction in finance. One stream of work has developed various extensions of PCA for identifying asset pricing factors or conducting asset pricing tests, see e.g. Kelly, Pruitt, and Su (2019); Lettau and Pelger (2020a,b); Giglio and Xiu (2021); Pelger and Xiong (2022a,b), among others. The other stream extends PCA, PLS and forecast combination for purpose of time series prediction with a large number of variables, including Bai and Ng (2008); Lin, Wu, and Zhou (2018); Huang, Jiang, Li, Tong, and Zhou (2022); Giglio, Xiu, and Zhang (2023). Our work complements these studies by proposing a novel approach to reduce the dimension of predictors to a single factor. Since our SEM method can consistently recover the target relevant factor, it can potentially be applied in many forecasting scenes to yield improved results.

Our paper also contributes to the literature on the predictability of the market risk premium. Predicting aggregate equity market excess return by macro variables has a long tradition in both academic research and industry practice. Welch and Goyal (2008) collect a list of macro variables studied in previous literature and question their genuine predictability out-of-sample. More than a decade later, Goyal, Welch, and Zafirov (2023) update both the sample period and list of predictors, and find that there is still a lack of out-of-sample predictability. These results on weak to no-predictability stimulate a substantial number of responses, including Campbell and Thompson (2008); Rapach, Strauss, and Zhou (2010); Cochrane (2011); Henkel, Martin, and Nardari (2011); Pettenuzzo, Timmermann, and Valkanov (2014); Kelly, Malamud, and Zhou (2024), among others. Rapach and Zhou (2013, 2020) provide survey on recent progress. An important theme in this literature is to develop more effective ways to filter noise and aggregate predictive information from many variables. Our study follows along this line and proposes a single factor to summarize the predictive information. Our evidence suggests stock return predictability is still strong, both statistically and economically, once better model has been employed.

The rest of this article is structured as follows. In Section 2, we propose the novel SEM method to estimate the single optimal predictive factor. There, we establish its asymptotic properties and compare with PLS and PCA. Section 3 explores the empirical performance of SEM in predicting US equity risk premium. Also provided are additional results on the prediction of stock market volatility, Treasury bond excess return and IP growth. Final Section concludes.

2 Optimal Prediction

In this section, we discuss the methodology of our paper. First, we explain intuitively the differences between PCA, PLS and our single factor approach. Second, we show that for any linear factor predicting model, there always exists an equivalent one-relevant-factor model. Then, we discuss the identification issue and propose an SEM estimator to extract the single relevant factor. Finally, we derive the asymptotic consistency of SEM and compare with those of PLS and PCA.

2.1 The model

Following Stock and Watson (2002a), Bai and Ng (2002); Bai (2003) and as well as Kelly and Pruitt (2013, 2015), we assume the following general predictive model for predicting a target variable y_t,

$$\begin{aligned} y_t &= \alpha + \beta' \boldsymbol{g}_t + \varepsilon_t, \\ x_{it} &= \nu_i + \phi_i' \boldsymbol{g}_t + \psi_i' \boldsymbol{h}_t + e_{it}, \end{aligned} \tag{1}$$

where $\boldsymbol{g}_t$, $r_1 \times 1$, are true but latent factors that predict y_t, whereas the observable predictors are proxies of the true factors, and are related to $\boldsymbol{g}_t$ by the second equation of model (1) in which $\boldsymbol{h}_t$ may be interpreted as systematic noise and e_{it} the idiosyncratic one. The parameters β, ϕ_i and ψ_i are r_1- and r_2-dimensional vectors, respectively.

The PCA approach is to apply the standard PCA to the predictor matrix $\boldsymbol{X} = \{x_{it}\}$, $N \times T$, that summarizes all the predictors. The optimization problem is to solve

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \mathbb{V}\mathrm{ar}(\boldsymbol{X}\boldsymbol{\alpha}) \\ \text{subject to:} \quad &\|\boldsymbol{\alpha}\| = 1, \quad \boldsymbol{\alpha}' \boldsymbol{S}_x \widehat{\zeta}_l = 0, \quad k = 1, 2, \ldots, k-1, \end{aligned} \tag{2}$$

where $\boldsymbol{S}_x$ is the sample covariance of $\boldsymbol{X}$. Under fairly general regularity conditions, Bai and Ng (2002); Bai (2003) show that the PCA factors recover the true latent factors up to a rotation.

In the PLS method, the kth PLS direction $\widehat{\varphi}_k$ solves

$$\max_{\boldsymbol{\alpha}} \mathbb{C}\text{ov}^2(\boldsymbol{Y}, \boldsymbol{X}\boldsymbol{\alpha}) \tag{3}$$

$$\text{subject to:} \quad \|\boldsymbol{\alpha}\| = 1, \quad \boldsymbol{\alpha}' \boldsymbol{S}_x \widehat{\varphi}_l = 0, \quad k = 1, 2, \ldots, k-1.$$

Since the target information $\boldsymbol{Y}$ is used, PLS is a supervised learning, whereas the PCA is not. Their difference is thus clear.

To see the difference between PCA and our approach, consider the case of two latent factors and three predictors. Then the predictive equation (1) is

$$\begin{aligned} y_t &= \alpha + \beta_1 g_1 + \beta_2 g_2 + \varepsilon_t \\ &= \alpha + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon_t \\ &= \alpha + \gamma(c_1 x_1 + c_2 x_2 + c_3 x_3) + \varepsilon_t, \end{aligned}$$

where the second equality follows from the PCA analysis and the three is a simply re-write, which shows that the optimal PCA factor reduction to 2 is observational equivalent to a factor of one, the predictive factor

$$P_f = c_1 x_1 + c_2 x_2 + c_3 x_3.$$

Our SEM method developed later is to estimate the c's directly in a high dimension set-up, without estimating the number of factors, r_1, or the PCA projection coefficients. Intuitively, due to reduced number of parameters, the estimation errors must be smaller, and so the result will converge faster, which is indeed the case as shown below by the asymptotic theory.

The above intuition has a close analogue in asset pricing. Let R_j be the asset returns in an economy of which there are J assets. A fundamental question in asset pricing is to see factors, $f_1, \ldots, f_K$, that prices the assets,

$$E[R_j m] = 1, \quad j = 1, 2, \ldots, J,$$

where m is the discount factor,

$$m = 1 - \lambda_1 f_1 - \cdots - \lambda_K f_K,$$

where λ_k's are the loadings on the factors. Kozak, Nagel, and Santosh (2020), among others, use PCA to extract the factors first to reduce the dimensionality, and then to estimate the λ_ks. In contrast, our approach here would suggest to estimation the optimal weights on the assets directly, even in high dimension, yielding a potentially a discount factor with the minimum estimation error.

2.2 Optimal dimension reduction

We first reparameterize the model for ease of analysis. Suppose that $\boldsymbol{Q}$ is an $r_1 \times r_1$ invertible matrix such that

$$\boldsymbol{Q}' = \left[\frac{1}{\|\boldsymbol{\beta}\|}\boldsymbol{\beta}, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_{r_1}\right]$$

, then the second equation of Model (1) can be written as

$$x_{it} = \boldsymbol{\phi}_i'\boldsymbol{Q}^{-1}\boldsymbol{Q}\boldsymbol{g}_t + \boldsymbol{\psi}_i'\boldsymbol{h}_t + e_{it} = \boldsymbol{\phi}_i'\boldsymbol{Q}^{-1}\begin{bmatrix} \frac{1}{\|\boldsymbol{\beta}\|}\boldsymbol{\beta}'\boldsymbol{g}_t \\ \boldsymbol{q}_2'\boldsymbol{g}_t \\ \vdots \\ \boldsymbol{q}_{r_1}'\boldsymbol{g}_t \end{bmatrix} + \boldsymbol{\psi}_i'\boldsymbol{h}_t + e_{it}.$$

Let ℓ_k be the kth column of the r_1-dimensional identity matrix. Now we redefine the three scalar parameters g_t^*, β^* and ϕ_i^* as

$$g_t^* = \frac{1}{\|\boldsymbol{\beta}\|}\boldsymbol{\beta}'\boldsymbol{g}_t, \quad \beta^* = \|\boldsymbol{\beta}\|, \quad \boldsymbol{\phi}_i^* = \boldsymbol{\phi}_i'\boldsymbol{Q}^{-1}\ell_1,$$

and redefine $(r_1 + r_2 - 1)$ dimensional vectors $\boldsymbol{\psi}_i^*$ and $\boldsymbol{h}_t^*$ as

$$\boldsymbol{h}_{t,k}^* = \begin{cases} \boldsymbol{q}_{k+1}'\boldsymbol{g}_t & \text{if } k = 1, 2, \ldots, r_1 - 1, \\ \boldsymbol{h}_{t,k-r_1+1} & \text{if } k = r_1, r_1+1, \ldots, r_1+r_2-1, \end{cases}$$

$$\boldsymbol{\psi}_{i,k}^* = \begin{cases} \boldsymbol{\phi}_i'\boldsymbol{Q}^{-1}\ell_{k+1} & \text{if } k = 1, 2, \ldots, r_1 - 1, \\ \boldsymbol{\psi}_{i,k-r_1+1} & \text{if } k = r_1, r_1+1, \ldots, r_1+r_2-1, \end{cases}$$

where $\boldsymbol{h}_{t,k}$ is the kth element of $\boldsymbol{h}_t$ and $\boldsymbol{\psi}_{i,k}, \boldsymbol{\psi}^*_{i,k}, \boldsymbol{h}^*_{t,k}$ are defined similarly. With the above notations, Model (1) can now be equivalently written as

$$y_t = \alpha + \beta^* g^*_t + \varepsilon_t, \tag{4}$$

$$x_{it} = \nu_i + \phi^*_i g^*_t + \boldsymbol{\psi}^{*\prime}_i \boldsymbol{h}^*_t + e_{it}.$$

Compared with Model (1), Model (4) has a striking feature that the relevant factor is only one, which is captured by g^*_t. This result does not depend on the dimensionality of the original g_t. In other words, whether the forecasting target is related to the full factors or partial factors, it can always be written into Model (4). This feature is presented formally into the following proposition.

Proposition 1. *For any linear factors predicting model, there always exists an equivalent one-relevant-factor Model (4).*

Hereafter, we proceed our analysis based on Model (4). For notational simplicity, we drop off asterisk from the symbols and therefore the model now is

$$y_t = \alpha + \beta g_t + \varepsilon_t \tag{5}$$

$$x_{it} = \nu_i + \phi_i g_t + \boldsymbol{\psi}'_i \boldsymbol{h}_t + e_{it} = \mu_i + \boldsymbol{\lambda}'_i \boldsymbol{f}_t + e_{it} \tag{6}$$

where $\boldsymbol{\lambda}_i = (\phi_i, \boldsymbol{\psi}'_i)'$ and $\boldsymbol{f}_t = (g_t, \boldsymbol{h}'_t)'$. As seen, all the information useful for prediction contained in $\boldsymbol{X}$ is summarized in the relevant scalar g_t.

We next discuss the identification issue related to model (5)-(6), which provides the motivation of this paper, and also serves the base of the subsequent theoretical analysis. It is well known in the literature that the loadings $\boldsymbol{\lambda}_i$ and the factors $\boldsymbol{f}_t$ can only be identified up to a rotation in a standard factor model $x_{it} = \nu_i + \boldsymbol{\lambda}'_i \boldsymbol{f}_t + e_{it}$. More specifically, for any invertible matrix $\boldsymbol{R}$, the model can be equivalently written as $x_{it} = \nu_i + \boldsymbol{\lambda}'_i \boldsymbol{R} \boldsymbol{R}^{-1} \boldsymbol{f}_t + e_{it}$. So if we define $\boldsymbol{\lambda}^\dagger_i = \boldsymbol{R}' \boldsymbol{\lambda}_i$ and $\boldsymbol{f}^\dagger_t = \boldsymbol{R}^{-1} \boldsymbol{f}_t$, we obtain another set of parameters $(\boldsymbol{\lambda}^\dagger_i, \boldsymbol{f}^\dagger_t)$ corresponding to the same observable data $\boldsymbol{X} = [x_{it}]_{N \times T}$. Basically, since the rotational matrix $\boldsymbol{R}$ has r^2 free parameters, we need to impose r^2 restrictions to fully identify them. We refer readers to Bai and Li (2012, 2016) for five popular identification strategies in a factor model.

We emphasize, however, that the identification issue here is quite different from the standard factor model. The key difference is that the factor g_t is identifiable in model (5)-(6), while g_t is subject to rotational indeterminacy in a standard factor model. To see that g_t is identifiable in the

former case, we note that model (5)-(6) implies that $\boldsymbol{G} = (g_1, g_2, \ldots, g_T)'$ can be asymptotically viewed as the projection values of the observed data $\boldsymbol{Y} = (y_1, y_2, \ldots, y_T)'$ on the factor space $\boldsymbol{P_F}$ where $\boldsymbol{P_F} = \boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}'$ with $\boldsymbol{F} = (\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_T)'$ in the sense that $\frac{\|\boldsymbol{G} - \boldsymbol{P_F}\boldsymbol{Y}\|}{\|\boldsymbol{Y}\|} \xrightarrow{p} 0$. Put differently, the factor $\boldsymbol{G}$ can be recovered by $\boldsymbol{P_F}\boldsymbol{Y}$. As the factor space $\boldsymbol{P_F}$ is invariant to rotational indeterminacy and can be recovered from the data $\boldsymbol{X}$, we see that $\boldsymbol{G}$ can be recovered from the datasets $\boldsymbol{X}$ and $\boldsymbol{Y}$, and therefore is identifiable. The identification of $\boldsymbol{G}$ naturally begs the question: How many restrictions are actually needed to fully identify the model. This question is answered by the following proposition.

Proposition 2. *For model (5)-(6), we only need $r^2 - r + 1$ restrictions to achieve identification.*

The proof of Proposition 2 is given in the appendix. Proposition 2 states that model (5)-(6) is in short of $(r-1)$ restrictions compared with the r^2 restrictions in standard factor model. In the classical identification conditions $\boldsymbol{S}_f = \frac{1}{T}\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_r$ and $\boldsymbol{\Pi}_{\boldsymbol{\Lambda}\boldsymbol{\Lambda}} = \frac{1}{N}\boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \boldsymbol{D}$ where $\boldsymbol{D}$ is a diagonal matrix with distinct diagonal elements, there are two sets of orthogonal restrictions in factor models $\boldsymbol{S}_{\boldsymbol{GH}} = \frac{1}{T}\boldsymbol{G}'\boldsymbol{H} = 0$ and $\boldsymbol{\Pi}_{\boldsymbol{\Phi}\boldsymbol{\Psi}} = \frac{1}{N}\boldsymbol{\Psi}'\boldsymbol{\Phi} = 0$, which both incur $(r-1)$ restrictions. When r^2 restrictions are needed in analysis, we can impose these two sets of orthogonal restrictions simultaneously. However, since only $r^2 - r + 1$ restrictions are needed now, we can only choose one set of restrictions to impose. This result, which plays an important role in the next section, is presented exclusively in the following corollary.

Corollary 1. *For model (5)-(6), we cannot impose $\boldsymbol{S}_{\boldsymbol{GH}} = \frac{1}{T}\boldsymbol{G}'\boldsymbol{H} = 0$ and $\boldsymbol{\Pi}_{\boldsymbol{\Phi}\boldsymbol{\Psi}} = \frac{1}{N}\boldsymbol{\Phi}'\boldsymbol{\Psi} = 0$ simultaneously.*

In the following analysis, we impose $\boldsymbol{S}_{\boldsymbol{GH}} = 0$ and let $\boldsymbol{\Pi}_{\boldsymbol{\Phi}\boldsymbol{\Psi}}$ be unrestricted.

2.3 Estimation through supervised EM algorithm

The basic idea of our method is very simple. Since the forecasting target $\boldsymbol{Y}$ contains all the useful information by equation (5), instead of using $\boldsymbol{X}$ to extract factors, we combine the data $\boldsymbol{Y}$ and $\boldsymbol{X}$ into $\boldsymbol{Z}$ and extract the factors based on $\boldsymbol{Z}$. Inclusion of $\boldsymbol{Y}$ into the dataset has two advantages. First, since $\boldsymbol{Y}$ only contains the relevant factor, inclusion of $\boldsymbol{Y}$ helps to separate the relevant factor from the remaining irrelevant ones, and therefore amount to supervising the direction of dimension reduction. Second, $\boldsymbol{Y}$ also contains useful information, inclusion of $\boldsymbol{Y}$ can improve the quality of estimation. Once we recover $\boldsymbol{G}$ from $\boldsymbol{X}$ and $\boldsymbol{Y}$, we can easily compute the associated weights, and then use these weights to construct the factor next period to perform forecasting.

Let $\boldsymbol{f}_t = (g_t, \boldsymbol{h}_t')'$ and $\boldsymbol{Z}_t$, $\boldsymbol{B}$, $\boldsymbol{\mu}$, $\boldsymbol{U}$ be defined as

$$\boldsymbol{Z}_t = \begin{bmatrix} y_t \\ x_{1t} \\ \vdots \\ x_{Nt} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \alpha \\ \nu_1 \\ \vdots \\ \nu_N \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \beta & \mathbf{0}_{r-1}' \\ \phi_1 & \psi_1' \\ \vdots & \vdots \\ \phi_N & \psi_N' \end{bmatrix}, \quad \boldsymbol{U}_t = \begin{bmatrix} \varepsilon_t \\ e_{1t} \\ \vdots \\ e_{Nt} \end{bmatrix}.$$

Now we can rewrite the model into the following matrix form

$$\boldsymbol{Z}_t = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{U}_t. \tag{7}$$

As seen in the above equation (7), a key difference of our model from the standard one is that there exists zero restrictions in the first row of $\boldsymbol{B}$. These zero elements restrict rotations of the factors $\boldsymbol{f}_t$ and play an important role on the supervised dimension reduction. Our estimation strategy amounts to estimating $\boldsymbol{B}$ with explicitly accommodation of these zero restrictions. To this end, we propose the following algorithm, which is essentially based on the idea of the expectation-maximization (EM) algorithm proposed by Dempster, Laird and Rubin (1977). Since zero restrictions are explicitly treated and used to supervise the direction of dimension reduction, we call our algorithm the supervised EM algorithm (abbreviated by SEM).

Algorithm 1 Forecasting based on one SEM direction

1. Using data $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_T]$ to construct the initial values $\boldsymbol{\Theta}^{(1)} = (\boldsymbol{B}^{(1)}, \boldsymbol{\Omega}_u^{(1)})$, which are given in Algorithm 2 below.

2. Based on the estimator at the sth iteration $\boldsymbol{\Theta}^{(s)} = (\boldsymbol{B}^{(s)}, \boldsymbol{\Omega}_u^{(s)})$,

   - Calculate the vector

$$\boldsymbol{P}_t = \left[\boldsymbol{B}^{(s)\prime}(\boldsymbol{\Omega}_u^{(s)})^{-1}\boldsymbol{B}^{(s)} + \boldsymbol{I}_r\right]^{-1} \boldsymbol{B}^{(s)\prime}(\boldsymbol{\Omega}_u^{(s)})^{-1}\boldsymbol{Z}_t,$$

   and the matrix

$$\boldsymbol{Q}_t = \boldsymbol{P}_t\boldsymbol{P}_t' + \left[\boldsymbol{B}^{(s)\prime}(\boldsymbol{\Omega}_u^{(s)})^{-1}\boldsymbol{B}^{(s)} + \boldsymbol{I}_r\right]^{-1}.$$

- Update the parameters as

$$\beta^{(s+1)} = \sum_{t=1}^{T} y_t \boldsymbol{P}_{t,1} (\sum_{t=1}^{T} \boldsymbol{Q}_{t,11})^{-1} \quad \text{and} \quad \boldsymbol{\Lambda}^{(s+1)} = \sum_{t=1}^{T} \boldsymbol{X}_t \boldsymbol{P}_t' (\sum_{t=1}^{T} \boldsymbol{Q}_t)^{-1},$$

where $\boldsymbol{P}_{t,1}$ is the first element of $\boldsymbol{P}_t$ and $\boldsymbol{Q}_{t,11}$ is the (1,1)th element of $\boldsymbol{Q}_t$. Then construct $\boldsymbol{B}^{(s+1)} = [\beta^{(s+1)}, 0_{r-1}' | \boldsymbol{\Lambda}^{(s+1)}]$ and calculate

$$\boldsymbol{\Omega}_u^{(s+1)} = \text{diag}(\boldsymbol{S}_z - \boldsymbol{B}^{(s+1)} \boldsymbol{B}^{(s)\prime}),$$

where $\text{diag}(\cdot)$ denotes the operator which sets the non-diagonal elements of its input to zeros.

3. Iterate $\boldsymbol{\Theta}^{(s)} \to \boldsymbol{\Theta}^{(s+1)}$ until $\|\boldsymbol{B}^{(s+1)} - \boldsymbol{B}^{(s)}\| + \|\boldsymbol{\Omega}_u^{(s+1)} - \boldsymbol{\Omega}_u^{(s)}\| \leq c$ where c is some prespecified tolerance condition. Let $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\Omega}}_u$ be the final estimators.

4. Construct the SEM weight $\boldsymbol{w}_{\text{SEM}}$ as

$$\boldsymbol{w}_{\text{SEM}} = \widehat{\boldsymbol{\Omega}}_e^{-1} \widehat{\boldsymbol{\Lambda}} (\widehat{\boldsymbol{\Lambda}}' \widehat{\boldsymbol{\Omega}}_e^{-1} \widehat{\boldsymbol{\Lambda}} + \boldsymbol{I}_r)^{-1} \ell_1,$$

where $\widehat{\boldsymbol{\Lambda}}$ is the matrix obtained by deleting the first row of $\widehat{\boldsymbol{B}}$, and $\widehat{\boldsymbol{\Omega}}_e$ is the lower-right $N \times N$ submatrix of $\widehat{\boldsymbol{\Omega}}_u$, and ℓ_1 is the first column of r-dimensional identity matrix.

5. Construct the forecasting value of y_{T+1} as $\widehat{y}_{T+1} = \widehat{\alpha} + \widehat{\beta}(\boldsymbol{w}_{\text{SEM}}' \boldsymbol{X}_{T+1})$, where $\widehat{\alpha} = \frac{1}{T} \sum_{t=1}^{T} y_t$ and $\hat{\beta}$ is the (1,1)th element of $\widehat{\boldsymbol{B}}$.

The above SEM forecasting procedure is based on one SEM direction. For the linear forecasting model, we can show that one SEM direction is sufficient for forecasting, see Theorem 1 below. However, in real data applications, the relationship between the target and the factors may not be exactly linear. This explains that the existing literature usually extracts more than one direction to perform the forecasting, see Kelly and Pruitt (2013) among others. Our procedure can be easily adapted to allow multiple directions. Suppose that we have obtained the weight for the first direction $\boldsymbol{w}_{\text{SEM}}^{(1)}$, then calculate the factor $\widehat{\boldsymbol{G}} = \boldsymbol{X}' \boldsymbol{w}_{\text{SEM}}^{(1)}$ with $\boldsymbol{X} = (\underline{\boldsymbol{X}}_1, \underline{\boldsymbol{X}}_2, \ldots, \underline{\boldsymbol{X}}_T)$. We next treat $\underline{\boldsymbol{X}}_i M_{\widehat{\boldsymbol{G}}}$ as the new $\underline{\boldsymbol{X}}_i$ for each i, and treat $\boldsymbol{Y} M_{\widehat{\boldsymbol{G}}}$ as the new $\boldsymbol{Y}$, and $r-1 \to r$, then repeat steps 1 and 4 in Algorithm 1 to obtain the second direction $\boldsymbol{w}_{\text{SEM}}^{(2)}$. This procedure is repeated and any values of directions are available. The procedure can be illuminated by the following interpretation. Suppose that the link function of y_t and $\boldsymbol{f}_t$ is $y_t = g_1(f_{t1}) + g_2(f_{t2}) +$

$\cdots + g_r(f_{tr}) + e_t$, and also suppose that the variation of $g_i(f_{ti})$ decreases dramatically over i. If we use linear model to approximate $g_1(f_{t1})$, since the variation of the remaining part is much smaller than $g_1(f_{1t})$, then the forecasting equation can be viewed as a one-regressor linear regression and therefore our algorithms applies. Now if the linear approximation error for $g_1(f_{t1})$ is much smaller relative to $g_2(f_{t2})$, given that the variations of $g_3(f_{t3}), \cdots, g_r(f_{tr})$ is much smaller, we can apply again algorithm 1 to obtain the second direction if $g_2(f_{t2})$ is approximated by another linear function. The similar explanations go for the next directions.

For the initial values of Algorithm 1, we suggest the following method to make iterations converge quickly.

Algorithm 2 Initial values for the SEM algorithm

1. For each i, run regression of $\underline{\boldsymbol{X}}_i$ on $\boldsymbol{Y}$ to obtain $\widetilde{\boldsymbol{\Phi}}$.
2. Calculate the estimator $\boldsymbol{H}^{(1)}$ by $\sqrt{T}$ times of the $(r-1)$ largest eigenvectors of the matrix $\boldsymbol{X}'\boldsymbol{M}_{\widetilde{\boldsymbol{\Phi}}}\boldsymbol{X}$.
3. Calculate the estimator $\boldsymbol{G}^{(1)}$ by $\sqrt{T}$ times of the largest eigenvector of the matrix $\boldsymbol{M}_{\boldsymbol{H}^{(1)}}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{M}_{\boldsymbol{H}^{(1)}}$.
4. Run regression of $\boldsymbol{Y}$ on $\mathbb{1}_T$ and $\boldsymbol{G}^{(1)}$ to obtain $\beta^{(1)}$, the coefficient of $\boldsymbol{G}^{(1)}$.
5. Calculate $\boldsymbol{\Lambda}^{(1)} = \boldsymbol{X}\boldsymbol{F}^{(1)}/T$ with $\boldsymbol{F}^{(1)} = [\boldsymbol{G}^{(1)}, \boldsymbol{H}^{(1)}]$ and construct $\boldsymbol{B}^{(1)}$ accordingly with $\beta^{(1)}$ and $\boldsymbol{\Lambda}^{(1)}$.
6. Calculate
$$\boldsymbol{\Omega}_u^{(1)} = \text{diag}\left(\frac{1}{T}(\boldsymbol{Z} - \boldsymbol{B}^{(1)}\boldsymbol{F}^{(1)\prime})(\boldsymbol{Z} - \boldsymbol{B}^{(1)}\boldsymbol{F}^{(1)\prime})'\right).$$
7. Output $\boldsymbol{B}^{(1)}$ and $\boldsymbol{\Omega}_u^{(1)}$.

2.4 Asymptotic properties

In this section, we present the asymptotic result of the SEM estimators. Our analysis is built on the maximum likelihood (ML) analysis of large dimensional factor model proposed by Bai and Li (2012, 2016). However, the current analysis differs from theirs because of the presence of zero restrictions in the first row of $\boldsymbol{B}$. As will be seen, these zero restrictions would change the asymptotic behavior of the ML estimators.

Now we start our analysis. By assuming that $\boldsymbol{U}_t \sim \mathscr{N}(\boldsymbol{0}_{\bar{N}}, \boldsymbol{\Omega}_u)$ and $\boldsymbol{f}_t \sim \mathscr{N}(\boldsymbol{0}_r, \boldsymbol{I}_r)$ where $\bar{N} = N+1$ is the dimension of the vector $\boldsymbol{Z}_t$, we see that $\boldsymbol{\Omega}_z \equiv \mathbb{V}\text{ar}(\boldsymbol{Z}_t) = \boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Omega}_u$. So we can write out the quasi-likelihood function of model (7) as

$$\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{B}, \boldsymbol{\Omega}_u) = -\frac{1}{2N}\ln|\boldsymbol{\Omega}_z| - \frac{1}{2NT}\sum_{t=1}^{T}(\boldsymbol{Z}_t - \boldsymbol{\mu})'\boldsymbol{\Omega}_z^{-1}(\boldsymbol{Z}_t - \boldsymbol{\mu}). \tag{8}$$

It is easy to see that $\boldsymbol{\mu}$ maximizes the above function as $\bar{\boldsymbol{Z}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{Z}_t$, this gives the concentrated objective function as

$$\mathscr{L}(\boldsymbol{B}, \boldsymbol{\Omega}_u) = -\frac{1}{2N}\ln|\boldsymbol{\Omega}_z| - \frac{1}{2NT}\sum_{t=1}^{T}(\boldsymbol{Z}_t - \bar{\boldsymbol{Z}})'\boldsymbol{\Omega}_z^{-1}(\boldsymbol{Z}_t - \bar{\boldsymbol{Z}}) = -\frac{1}{2N}\ln|\boldsymbol{\Omega}_z| - \frac{1}{2N}\text{tr}(\boldsymbol{S}_z\boldsymbol{\Omega}_z^{-1}),$$

where $\boldsymbol{S}_z = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{Z}_t - \bar{\boldsymbol{Z}})(\boldsymbol{Z}_t - \bar{\boldsymbol{Z}})'$ is the sample covariance of the vector $\boldsymbol{Z}_t$.

Now we impose the identification conditions to fully identify the parameters. We note that although the MSFE of the SEM method is invariant to the rotational indeterminacy, we still would like to show the results under the full identification scenario because this treatment lets readers see clearly that the identification here is different from the one in the standard model. In addition, this treatment makes our results more intuitive to understand. We consider the following identification conditions.

Identification conditions:

IC1. $\sum_{t=1}^{T}\boldsymbol{f}_t = 0$

IC2. $\boldsymbol{S}_f = \frac{1}{T}\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_r$

IC3. $\frac{1}{N}\boldsymbol{\Psi}'\boldsymbol{\Omega}_e^{-1}\boldsymbol{\Psi} = \boldsymbol{D}$, which is a diagonal matrix with distinct diagonal elements.

We note that because of presence of the intercept $\boldsymbol{\mu}$ in equation (7), we have to impose $\sum_{t=1}^{T}\boldsymbol{f}_t = 0$ otherwise μ cannot be identified. For IC2 and IC3, they are imposed to eliminate the rotational indeterminacy. Note that IC2 provides $r(r+1)/2$ restrictions and IC3 provides $(r-1)(r-2)/2$ restrictions. So the total restrictions is equal to $r^2 - r + 1$, as required by Proposition 2. We also note that the above identification strategy has the so-called sign problem. If one multiplies -1 on some column of $\boldsymbol{B}$, and also multiple -1 on the corresponding column of $\boldsymbol{F}$, we obtain the same model. For this reason, following Bai and Li (2012), we assume that $\boldsymbol{B}$ and its estimator $\widehat{\boldsymbol{B}}$ have the same column signs. This implicit assumption will be maintained throughout our theoretical analysis.

For the purpose of analysis, we impose the following assumption.

Assumption A. The factors $\{\boldsymbol{f}_t\}$ are a sequence of fixed constants such that $\boldsymbol{S}_f = \frac{1}{T}\boldsymbol{F}'\boldsymbol{F}$ is a strictly positive definite matrix.

Assumption B. The errors satisfy

B.1 $\boldsymbol{U}_t = (\varepsilon_t, e_{1t}, \ldots, e_{N,t})'$ is independent and identically distributed over time such that $\mathbb{E}(\boldsymbol{U}_t) = 0$. We use $\boldsymbol{\Omega}_u = \mathbb{E}(\boldsymbol{U}_t\boldsymbol{U}_t')$ to denote the variance of $\boldsymbol{U}_t$.

B.2 ε_t is independent with e_{it} for any i. In addition, e_{it} is independent with e_{jt} for any i and j.

B.3 $\max\left(\mathbb{E}(\varepsilon_t^4), \max\limits_{1\le i\le N}\mathbb{E}(e_{i,t}^4)\right) \le C$.

Assumption C. The loadings $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_N)'$ are fixed values satisfying $\max\limits_{1\le i\le N}\|\boldsymbol{\lambda}_i\| \le C$. In addition, β is a fixed value such that $|\beta| \le C$.

Assumption D. The parameters σ_ε^2 and $\sigma_i = \mathbb{E}(e_{it}^2)$ for $i = 1, 2, \ldots, N$ are estimated in a compact set $[C^{-1}, C]$ for $C > 1$.

Following the literature, we use the mean square forecast error (MSFE) as the evaluation criterion, which is defined as

$$\text{MSFE} = \mathbb{E}\Big[(y_{T+1} - \widehat{y}_{T+1|T})^2\Big]. \tag{9}$$

We have the following theorem on the SEM estimators.

Theorem 1. *Under Assumptions A-D, together with the identification conditions, as $N, T \to \infty$ and $N/T \to \kappa$,*

- $\sqrt{T}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma_\varepsilon^2)$;
- *for each t,*

$$\sqrt{N}(\boldsymbol{w}_{\text{SEM}}'\boldsymbol{X}_{T+1} - g_{T+1}) \xrightarrow{d} \mathcal{N}\Big(0, \lim_{N\to\infty}\ell_1'\Big(\frac{1}{N}\boldsymbol{\Lambda}'\boldsymbol{\Omega}_e^{-1}\boldsymbol{\Lambda}\Big)^{-1}\ell_1 + \kappa\frac{1}{\beta^2}\boldsymbol{h}_{T+1}'\boldsymbol{h}_{T+1}\Big).$$

Given the above result, we have

$$\text{MSFE}_{\text{SEM}} = \sigma_\varepsilon^2 + \frac{1}{T}\boldsymbol{f}_{T+1}'\boldsymbol{f}_{T+1}\sigma_\varepsilon^2 + \overline{\beta}'(\boldsymbol{\Lambda}'\boldsymbol{\Omega}_e^{-1}\boldsymbol{\Lambda})^{-1}\overline{\beta} + o_p(\frac{1}{N} + \frac{1}{T}),$$

where $\overline{\beta} = \beta\ell_1$.

Remark 1. We have emphasized that the MSFE is invariant to the rotational indeterminacy. If the identification conditions are not imposed, the MSFE would admit the expression

$$\text{MSFE}_{\text{SEM}} = \sigma_\varepsilon^2 + \boldsymbol{f}_{T+1}'(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{f}_{T+1}\sigma_\varepsilon^2 + \overline{\beta}'(\boldsymbol{\Lambda}'\boldsymbol{\Omega}_e^{-1}\boldsymbol{\Lambda})^{-1}\overline{\beta} + o_p(\frac{1}{N}+\frac{1}{T}).$$

2.5 Comparison with PLS and PCA

The partial least squares (PLS) was introduced by the Swedish statistician Herman Wold and has been historically defined in terms of the iterative algorithms NIAPLS (Wold (1975)) and SIMPLS (De Jong, 1993). PLS seeks directions that have high variance and have high correlation with the response, in contrast to principal components method which seeks only the high variance. Due to this supervised learning feature, PLS has received growing attention in the forecasting literature. The representative studies include Kelly and Pruitt (2013); Huang et al. (2015), among others. In the PLS method, the kth PLS direction $\widehat{\boldsymbol{\varphi}}_k$ solves

$$\max_{\boldsymbol{\alpha}} \mathbb{C}\text{ov}^2(\boldsymbol{Y}, \boldsymbol{X}\boldsymbol{\alpha}) \tag{10}$$

$$\text{subject to:} \quad \|\boldsymbol{\alpha}\| = 1, \quad \boldsymbol{\alpha}'\boldsymbol{S}_x\widehat{\boldsymbol{\varphi}}_l = 0, \quad k = 1, 2, \ldots, k-1.$$

where $\boldsymbol{S}_x$ is the sample covariance of $\boldsymbol{X}$. The PLS direction can be iteratively obtained by the following algorithm

Algorithm 3 The PLS algorithm

1. Standardize each predictor $\boldsymbol{x}_j$ and $\boldsymbol{Y}$ to have mean zero and variance one. Set $\underline{\boldsymbol{X}}_j^{(0)} = \underline{\boldsymbol{X}}_j$ and $\boldsymbol{Y}^{(0)} = \bar{y}\mathbb{1}_T$.
2. For $l = 1, 2, \cdots, k$, do the following:
   (a) For each i, run a time series regression of $\underline{\boldsymbol{X}}_i^{(l-1)}$ on $\boldsymbol{Y}$ obtain the slope estimator $\widehat{a}_i^{(l)}$ and construct the N-dimensional vector $\widehat{\boldsymbol{a}}$ by $\widehat{\boldsymbol{a}}^{(l)} = (\widehat{a}_1^{(l)}, \widehat{a}_2^{(l)}, \ldots, \widehat{a}_N^{(l)})'$.
   (b) For each t, run a cross sectional regression of $\boldsymbol{X}_t^{(l-1)}$ on $\widehat{\boldsymbol{a}}^{(l)}$ to obtain the slope estimator $\widehat{b}_t^{(l)}$ and construct the T-dimensional vector $\widehat{\boldsymbol{b}}^{(l)}$ by $\widehat{\boldsymbol{b}}^{(l)} = (\widehat{b}_1^{(l)}, \widehat{b}_2^{(l)}, \ldots, \widehat{b}_T^{(l)})'$.
   (c) Calculate $\widehat{\boldsymbol{Y}}^{(l)} = \widehat{\boldsymbol{Y}}^{(l)} + \boldsymbol{P}_{\widehat{\boldsymbol{b}}^{(l)}}\boldsymbol{Y}$ and $\underline{\boldsymbol{X}}_j^{(l)} = \boldsymbol{M}_{\widehat{\boldsymbol{b}}^{(l)}}\underline{\boldsymbol{X}}_j^{(j-1)}$.
3. Output the PLS-fitted value $\boldsymbol{Y}^{(k)}$.

The above algorithm is adapted from Algorithm 3.3 of Hastie et al. (2009) and is widely used in the asset pricing literature. We emphasize that the PLS method aims to iteratively solve (10), an optimization problem which has nothing to do with the factor structure of $\boldsymbol{X}$. Said another way, one actually extracts the PLS factors based on (10) or Algorithm 3 without any knowledge of internal structure of $\boldsymbol{X}$. However, the factor-structure assumption seems ubiquitous in empirical asset pricing (Kozak et al., 2008) and is deeply rooted in the modern asset pricing theory. Many researchers may be reluctant to drop off the factor structure assumption for some reasons such as the ease of exposition, the routine of similar studies, and so forth. This bears a question: what the performance of the PLS method would be if the predictors have factor structure. At the first glance, one may quickly draw the conclusion that the performance of the PLS method will be unaffected based on intuition established by the principal component analysis (PCA). In the PCA, the kth direction $\widehat{\boldsymbol{\xi}}_k$ solves

$$\max_{\boldsymbol{\alpha}} \mathbb{V}\text{ar}(\boldsymbol{X}\boldsymbol{\alpha}) \tag{11}$$
$$\text{subject to:} \quad \|\boldsymbol{\alpha}\| = 1, \quad \boldsymbol{\alpha}' \boldsymbol{S}_x \widehat{\boldsymbol{\xi}}_l = 0, \quad k = 1, 2, \ldots, k-1.$$

As seen, the optimization problem in (11) also has nothing to do with the factor structure assumption. However, it has been eventually proved by the large-dimensional factor analysis literature such as Bai (2003) and Bai and Ng (2002) that if the predictors are equipped with the factor structure by $x_{it} = \mu_i + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + e_{it}$, the PCA-based factors $\widehat{\boldsymbol{f}}_t^{\text{PCA}}$ is a consistent estimator of $\boldsymbol{f}_t$ in the sense that $\frac{1}{T}\sum_{t=1}^{T} \|\widehat{\boldsymbol{f}}_t^{\text{PCA}} - \boldsymbol{R}\boldsymbol{f}_t\|^2 = O_p(\frac{1}{N} + \frac{1}{T})$ for some invertible matrix $\boldsymbol{R}$, had some regularity assumptions been imposed. In a nutshell, the PCA method continues to work in the scenario that the predictors have factor structure.

Unfortunately, we emphasize that similar phenomenon does not occur in the PLS case. We will prove this result strictly. Since the above algorithm does not explicitly deliver the weights to generate the PLS factors, we propose the following algorithm to facilitate the analysis of the asymptotic properties of the PLS-based forecasting.

Algorithm 4 Forecasting based on the PLS algorithm

1. Standardize each predictor $\boldsymbol{x}_j$ to have mean zero and variance one. Let $\boldsymbol{X} = [x_{it}]_{N\times T}$.

2. Calculate $\boldsymbol{w}_{\text{PLS}}^{(1)}$, $\boldsymbol{W}^{(1)}$ and $\boldsymbol{V}^{(1)}$ by

$$\boldsymbol{w}_{\text{PLS}}^{(1)} = \boldsymbol{XY}\frac{\boldsymbol{Y}'\boldsymbol{X}'\boldsymbol{XY}}{\boldsymbol{Y}'(\boldsymbol{X}'\boldsymbol{X})^2\boldsymbol{Y}}, \qquad \boldsymbol{W}^{(1)} = \frac{1}{\boldsymbol{Y}'(\boldsymbol{X}'\boldsymbol{X})^2\boldsymbol{Y}}\boldsymbol{XYY}'\boldsymbol{X}'\boldsymbol{X},$$
$$\boldsymbol{V}^{(1)} = \left(\boldsymbol{I}_N - \boldsymbol{W}^{(1)}\boldsymbol{X}'\right)\left(\boldsymbol{I}_N - \boldsymbol{W}^{(1)}\boldsymbol{X}'\right)'.$$

If $k = 1$ then go to Step 3, otherwise set $j = 2$ and do the following iteratively until $j = k$

(a) Calculate

$$\boldsymbol{w}_{\text{PLS}}^{(j)} = \frac{\boldsymbol{Y}'\boldsymbol{X}'\boldsymbol{V}^{(j-1)}\boldsymbol{XY}}{\boldsymbol{Y}'(\boldsymbol{X}'\boldsymbol{V}^{(j-1)}\boldsymbol{X})^2\boldsymbol{Y}}\boldsymbol{V}^{(j-1)}\boldsymbol{XY},$$

(b) Calculate

$$\boldsymbol{W}^{(j)} = \frac{1}{\boldsymbol{Y}'(\boldsymbol{X}'\boldsymbol{V}^{(j-1)}\boldsymbol{X})^2\boldsymbol{Y}}\boldsymbol{V}^{(j-1)}\boldsymbol{XYY}'\boldsymbol{X}'\boldsymbol{V}^{(j-1)}\boldsymbol{X},$$

and

$$\boldsymbol{V}^{(j)} = \left[\boldsymbol{I}_N - \left(\sum_{l=1}^{j}\boldsymbol{W}^{(l)}\right)\boldsymbol{X}'\right]\left[\boldsymbol{I}_N - \left(\sum_{l=1}^{j}\boldsymbol{W}^{(l)}\right)\boldsymbol{X}'\right]'.$$

(c) $j+1 \to j$.

3. Then the PLS forecast for y_{T+1} is equal to $\widehat{y}_{T+1} = \bar{y} + \sum_{j=1}^{k}\widetilde{\beta}_j \boldsymbol{w}_{\text{PLS}}^{(j)\prime}\boldsymbol{x}_{T+1}$, where the coefficients $\widetilde{\beta}_j$ (j=1,..., k) are obtained by running regression of $\dot{y}_t = y_t - \bar{y}$ on $\boldsymbol{w}_{\text{PLS}}^{(j)\prime}\boldsymbol{x}_t$ for $j = 1, 2, \ldots, k$

Before presenting the theoretical results, we provide some remarks on the PCA analysis. First, the PCA analysis is a highly nonlinear analysis. This causes problems on deriving an explicit formulae of the MSFE. To facilitate theoretical analysis, we consider a shrinking framework. To make the point clear, note that for predictor x_{it}, it admits the following expression

$$x_{it} = \nu_i + \phi_i g_t + \psi_i'\boldsymbol{h}_t + e_{it} = \nu_i + \boldsymbol{\lambda}_i'\boldsymbol{f}_t + e_{it}$$

with $\boldsymbol{\lambda}_i = (\boldsymbol{\phi}_i, \boldsymbol{\psi}_i')'$ and $\boldsymbol{f}_t = (g_t, \boldsymbol{h}_t)'$. With this definition, we see that

$$\boldsymbol{\Pi}_{\lambda\lambda} \triangleq \begin{bmatrix} \boldsymbol{\Pi}_{\phi\phi} & \boldsymbol{\Pi}_{\phi\psi} \\ \boldsymbol{\Pi}_{\psi\phi} & \boldsymbol{\Pi}_{\psi\psi} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Pi}_{\phi\phi} & 0 \\ 0 & \boldsymbol{\Pi}_{\psi\psi} \end{bmatrix} + \begin{bmatrix} 0 & \boldsymbol{\Pi}_{\phi\psi} \\ \boldsymbol{\Pi}_{\psi\phi} & \mathbf{0} \end{bmatrix} = \boldsymbol{A} + \boldsymbol{V}.$$

where $\boldsymbol{A}$ and $\boldsymbol{V}$ are implicitly defined in the above equation. In our analysis, we assume that $\boldsymbol{V}$ shrinks to zero with an arbitrarily slow rate. To reflect this assumption on the symbol, we use $d\boldsymbol{A}$ instead of $\boldsymbol{V}$ to denote this shrinking value. The shrinking assumption has two advantages. First, it allows us to use the perturbation theory to analyze the effect of $d\boldsymbol{A}$ on $\boldsymbol{A}$, which is important to derive the final expression of MSFE. Second, it allows us to discard some high-order small terms and make our analysis essentially a first-order linear one. In practice, matrix $\boldsymbol{V}$ is certainly viewed as a fixed value and does not shrink to zero. However, we can treat the fixed value as the limit of the shrinking case. For example, let $C_{NT} = C(NT)^{-\nu}$. In this case, C_{NT} shrinks to zero as N and T grow large. But as $\nu \to 0$, we see $C_{NT} \to C$ for fixed N and T. Then our analysis can be justified by providing a good approximation to the limiting case.

Second, we assume that the value of $\boldsymbol{\Pi}_{\phi\phi}$ is larger than all the eigenvalues of the matrix $\boldsymbol{\Pi}_{\psi\psi}$. With this assumption, the first eigenvector extracted from matrix $\boldsymbol{A}$ would be a consistent estimator of the relevant factor. This assumption is natural as we would like to see the performance of the PCA in the setup most favorable to it. For this reason, we only consider the MSFE based on one PCA factor. The MSFE based on two PCA factors can be derived similarly but we note that since the second factor without perturbation is orthogonal to the relevant factor g_t, it is easy to see that a perturbed second factor has no additional power for forecasting.

Third, without loss of generality, we assume that $\boldsymbol{\Phi}_{\psi\psi}$ is a diagonal matrix. This assumption is a standard normalization of factor analysis due to the well-known rotational indeterminacy in $\boldsymbol{h}_t$.

We therefore have the following theorem on the first PLS and PCA factors, and the MSFEs based on these factors.

Theorem 2. *Under Assumptions A-C, together with the restriction $\boldsymbol{S}_{\boldsymbol{GH}} = \mathbf{0}$, we have that, for the first PLS factor*

$$\boldsymbol{w}_{\mathrm{PLS}}^{(1)\prime} \boldsymbol{X}_t \propto \boldsymbol{\Pi}_{\boldsymbol{\Phi}\boldsymbol{\Phi}} g_t + \boldsymbol{\Pi}_{\boldsymbol{\Phi}\boldsymbol{\Psi}} \boldsymbol{h}_t.$$

and for the first PCA factor, if $\min(\sqrt{N}, \sqrt{T}) \|\boldsymbol{\Pi}_{\boldsymbol{\Phi}\boldsymbol{\Psi}}\|^2 \to \infty$,

$$\boldsymbol{w}_{\mathrm{PCA}}^{(1)\prime} \boldsymbol{X}_t \propto g_t + \boldsymbol{\Pi}_{\boldsymbol{\Phi}\boldsymbol{\Psi}}^{*} \boldsymbol{h}_t$$

where $\mathbf{\Pi}_{\boldsymbol{xy}} = \frac{1}{N}\boldsymbol{x}'\boldsymbol{y}$ *for* $\boldsymbol{x}$ *or* $\boldsymbol{y} = \mathbf{\Phi}$ *or* $\mathbf{\Psi}$ *and* $\mathbf{\Pi}^*_{\mathbf{\Phi\Psi}} = [\mathbf{\Pi}^*_{\phi\psi_1}, \mathbf{\Pi}^*_{\phi\psi_2}, \ldots, \mathbf{\Pi}^*_{\phi\psi_{r-1}}]$ *is calcualted according to* $\mathbf{\Pi}_{\mathbf{\Phi\Psi}} = [\mathbf{\Pi}_{\phi\psi_1}, \mathbf{\Pi}_{\phi\psi_2}, \ldots, \mathbf{\Pi}_{\phi\psi_{r-1}}]$ *and*

$$\mathbf{\Pi}^*_{\phi\psi_j} = \frac{1}{\mathbf{\Pi}_{\phi\phi} - \mathbf{\Pi}_{\psi_j\psi_j}} \mathbf{\Pi}_{\phi\psi_j}.$$

Given the above result, we have that the mean square forecasting error (MSFE) based on the first PLS and PCA factor admits

$$\begin{aligned}
\text{MSFE}^{(1)}_{\text{PLS}} &= \beta^2 \left[\frac{\|\mathbf{\Pi}_{\mathbf{\Phi\Psi}}\|^2 g_{T+1} - \mathbf{\Pi}_{\mathbf{\Phi\Phi}}\mathbf{\Pi}_{\mathbf{\Phi\Psi}}\boldsymbol{h}_{T+1}}{\mathbf{\Pi}^2_{\mathbf{\Phi\Phi}} + \|\mathbf{\Pi}_{\mathbf{\Phi\Psi}}\|^2} \right]^2 + \sigma^2_\varepsilon + o_p(1) \\
\text{MSFE}^{(1)}_{\text{PCA}} &= \beta^2 \left[\frac{\|\mathbf{\Pi}^*_{\mathbf{\Phi\Psi}}\|^2 g_{T+1} - \mathbf{\Pi}^*_{\mathbf{\Phi\Psi}}\boldsymbol{h}_{T+1}}{1 + \|\mathbf{\Pi}^*_{\mathbf{\Phi\Psi}}\|^2} \right]^2 + \sigma^2_\varepsilon + \|\mathbf{\Pi}^*_{\mathbf{\Phi\Psi}}\|^2 o_p(1).
\end{aligned}$$

For the second PLS factor,

$$\boldsymbol{w}^{(2)\prime}_{\text{PLS}} \boldsymbol{X}_t \propto \widetilde{\mathbf{\Pi}}_{\mathbf{\Phi\Phi}} g_t + \widetilde{\mathbf{\Pi}}_{\mathbf{\Phi\Psi}} \boldsymbol{h}_t.$$

where $\widetilde{\mathbf{\Pi}}_{\mathbf{\Phi\Phi}} = \mathbf{\Pi}_{\mathbf{\Phi\Psi}}\mathbf{\Pi}_{\mathbf{\Psi\Psi}}\mathbf{\Pi}_{\mathbf{\Psi\Phi}}\mathbf{\Pi}_{\mathbf{\Phi\Phi}} - (\mathbf{\Pi}_{\mathbf{\Phi\Psi}}\mathbf{\Pi}_{\mathbf{\Psi\Phi}})^2$ *and*

$$\widetilde{\mathbf{\Pi}}_{\mathbf{\Phi\Psi}} = \mathbf{\Pi}_{\mathbf{\Phi\Psi}}\mathbf{\Pi}_{\mathbf{\Psi\Phi}}\mathbf{\Pi}_{\mathbf{\Phi\Phi}}\mathbf{\Pi}_{\mathbf{\Phi\Psi}} + \mathbf{\Pi}_{\mathbf{\Phi\Psi}}\mathbf{\Pi}_{\mathbf{\Psi\Psi}}\mathbf{\Pi}_{\mathbf{\Psi\Phi}}\mathbf{\Pi}_{\mathbf{\Phi\Psi}} - \mathbf{\Pi}^2_{\mathbf{\Phi\Phi}}\mathbf{\Pi}_{\mathbf{\Phi\Psi}}\mathbf{\Pi}_{\mathbf{\Psi\Psi}} - \mathbf{\Pi}_{\mathbf{\Phi\Psi}}\mathbf{\Pi}_{\mathbf{\Psi\Phi}}\mathbf{\Pi}_{\mathbf{\Phi\Psi}}\mathbf{\Pi}_{\mathbf{\Psi\Psi}},$$

and the MSFE based on the first two PLS factors admits

$$\begin{aligned}
\text{MSFE}^{(2)}_{\text{PLS}} = \beta^2 \Bigg[&\Big(1 - \frac{\mathbf{\Pi}^2_{\mathbf{\Phi\Phi}}}{\mathbf{\Pi}^2_{\mathbf{\Phi\Phi}} + \|\mathbf{\Pi}_{\mathbf{\Phi\Psi}}\|^2} - \frac{\widetilde{\mathbf{\Pi}}^2_{\mathbf{\Phi\Phi}}}{\widetilde{\mathbf{\Pi}}^2_{\mathbf{\Phi\Phi}} + \|\widetilde{\mathbf{\Pi}}_{\mathbf{\Phi\Psi}}\|^2}\Big) g_{T+1} \\
&- \Big(\frac{\mathbf{\Pi}_{\mathbf{\Phi\Phi}}\mathbf{\Pi}_{\mathbf{\Phi\Psi}}}{\mathbf{\Pi}^2_{\mathbf{\Phi\Phi}} + \|\mathbf{\Pi}_{\mathbf{\Phi\Psi}}\|^2} + \frac{\widetilde{\mathbf{\Pi}}_{\mathbf{\Phi\Phi}}\widetilde{\mathbf{\Pi}}_{\mathbf{\Phi\Psi}}}{\widetilde{\mathbf{\Pi}}^2_{\mathbf{\Phi\Phi}} + \|\widetilde{\mathbf{\Pi}}_{\mathbf{\Phi\Psi}}\|^2} \Big) \boldsymbol{h}_{T+1} \Bigg]^2 + \sigma^2_\varepsilon + o_p(1).
\end{aligned}$$

As emphasized in the previous section, we generally cannot impose $\mathbf{\Pi}_{\mathbf{\Phi\Psi}} = \mathbf{0}$ if we have already impose $\boldsymbol{S}_{\boldsymbol{GH}} = \mathbf{0}$. Given $\mathbf{\Pi}_{\mathbf{\Phi\Psi}} \neq \mathbf{0}$, Theorem 2 implies that both the PLS and the PCA does not recover the relevant factor g_t since they contains the irrelevant component $\boldsymbol{h}_t$. This undesirable feature eventually transmits to the MSFE. As seen in Theorem 2, the MSFE based on the first PLS factor and the PCA factor are larger than σ^2_ε. In other words, the PLS-based forecasting under the factor structure assumption of predictors is generally inconsistent. So is the PCA-based forecasting. The only excepted case is that the data happens to satisfy $\mathbf{\Pi}_{\mathbf{\Phi\Psi}} = \mathbf{0}$. In the next subsection, the inconsistency of the PLS-based and the PCA-based forecasting can be seen

clearly by simulations.

2.6 Simulation evidence

In this section, we conduct Monte Carlo simulations to assess the finite sample forecast performance of SEM, and compare against PCA and PLS factors.

Our simulation design is based on a latent factor models with a total of $r = 5$ factors, $f_{1t}, ..., f_{5t}$. The forecast target is then set to be $y_{t+1} = f_{1t} + f_{2t} + \varepsilon_{t+1}$, so that only the first two factors are target relevant. We choose this setup of a relatively large r to demonstrate that, even in the presence of multiple target relevant latent factors and more irrelevant factors, our algorithm can precisely recover the single forecasting factor to achieve dimension reduction.

Remaining assumptions are all standard. All factors are drawn independently from normal distribution of zero mean and unit variance, i.e., $f_{jt} \sim N(0,1)$ for $j = 1...5$. The forecast errors are also independent and normally distributed $\varepsilon_{t+1} \sim N(0,1)$, for $t = 1...T$. As a result, the infeasible best forecast has a lowest mean squared forecast error (MSFE) of 1. Meanwhile, there are a total of N observable predictors $X_{i,t}$ $(i = 1, ..., N)$, which load on both the target relevant and irrelevant factors $f_{1t}...f_{5t}$. Predictors loadings are drawn independently from standard uniform distribution $U[0,1]$. In addition, each predictor $X_{i,t}$ also contains an idiosyncratic noise term u_{it}, which is generated from a normal distribution with zero mean and standard deviation of σ_i. We draw σ_i randomly from a uniform distribution with support of $U[0,2]$. This allows us to take into account cross sectional heteroskedastic among these idiosyncratic noises.

Table 1 presents the simulation results under various specifications of sample size (N,T) on the number of predictors N and time length T. For a particular sample size, we simulate observations on N predictors and the target for a total time length of $T + 50$. Then in each simulated sample, we use the first T observations for initial parameter estimation and the remaining 50 observations for out-of-sample evaluation. In computing MSFE, we rely on a rolling window scheme so that parameter estimation always involve T observations in each out-of-sample forecast. We then report the median of mean squared forecast error (MSFE) over 500 repetitions.

Unanimously across different sample size specification of (T,N), we find that SEM forecasts produce lower MSFE than PC forecasts by 1, 2 and 3 PC factors. The reductions in MSFE are sizable and can reach from about 10% to 15%. More interestingly, even if we include all 5 factors in PC forecasts, a single SEM factor can still deliver notable reduction of MSFE that are larger than 5%. Though, as (T,N) increases, both forecasts tend to have MSFEs that are approaching the infeasible best of 1, the out-performance by SEM remain evident even for relatively large T and N.

As for PLS forecasts, we find forecasts with 2 PLS factors deliver relatively small MSFE that out-perform 3 PC factors forecasts. However, they still under-perform the single factor SEM forecasts. Including 3 factors (or more) does not seem to further improve the forecast performance and only increase forecast errors. Overall, our simulation results confirm the theoretical predictions above: a single SEM factor can effectively achieve dimension reduction and minimize forecast error.

3 Empirical Analysis

In this section, we apply the above novel supervised EM method to extract the optimal predictive factor, termed as SEM factor, in forecasting equity risk premium. We examine the composition of our optimal predictive factor, and contrast with those of PCA and PLS factors. We then assess the predictive accuracy and explore the asset allocation benefit of having access to this optimal predictive factor. Finally, we extend from predicting equity market excess return to predicting stock market volatility, bond market return and macroeconomic aggregate. Our results demonstrate the pivotal role of estimating optimal predictive factor in a broad range of applications.

3.1 Data

To uncover US equity market premium, we focus on the excess aggregate equity market return, computed as the continuously compounded log return on the S&P 500 index (including dividends) minus the risk-free rate.

We include a total of 50 financial and macroeconomic variables as predictors, all of which have been developed in previous literature to characterize the state of economy and market sentiment. They contain various types of information from macro and price data; newspaper textual information; to survey of experts. Some of the variables collectively have already spurred extensive research and application of many dimension reduction methods, including PCA and PLS, in predicting financial market return. Below, we describe the sources of these variables. A detailed list of their names and abbreviations are provided in the Appendix.

1. 25 macro variables, based on a combination of the lists of predictors compiled initially by Welch and Goyal (2008) and subsequently updated by Goyal et al. (2023) with more variables to predict aggregate stock market return. Each of the variable, such as log dividend price ratio, has been proposed in published article but Goyal et al. (2023) raised concerns on their out-of-sample predictability. Nevertheless, subsequent studies, including Campbell

and Thompson (2008); Rapach et al. (2010); Kelly and Pruitt (2013); Pettenuzzo et al. (2014); Kelly et al. (2024) among others, have demonstrate that these variables combined or equipped with economic restrictions still contain useful information on equity market premium. The updated data on these variables are available from Amit Goyal's website at monthly frequency.

2. 5 investor sentiment measures, from Baker and Wurgler (2006, 2007). These measures include equity share in new issues; closed end fund discount rate; number of IPOs; first day returns of IPOs and dividend premium. Huang et al. (2015) showed that when these measures are aligned for purpose of explaining returns, the extracted relevant factor could deliver strong equity market return predictive power. Moreover, the return predictive information contained goes above and beyond those by the first principal component of these measures and by the macro variables in Welch and Goyal (2008). The updated data on these sentiment measures are available from Jeffrey Wurgler's website at monthly frequency.

3. 12 categorical economic policy uncertainty (EPU) measures, from Baker, Bloom, and Davis (2016). These are indexes based on newspaper coverage frequency derived from using the Access World News database of over 2,000 US newspapers. The categorical policy terms include: monetary policy; fiscal policy; tax; government spending; healthcare; national security; entitlement program; regulation; financial regulation; trade policy; soverign debt and currency crisis, as well as the overall economic policy. Baker, Bloom, Davis, and Kost (2019); Baker, Bloom, Davis, and Sammon (2021), among others, have used these indexes to explain stock price movements. The updated data on these indexes are available on their policy-uncertainty website at monthly frequency.

4. 8 macroeconomic expectation measures, based on survey of experts from Survey of Professional Forecasters (SPF). We collect survey by individual forecasters and then compute mean consensus nowcasts on the following 8 macroeconomic variables: 1) growth rate of real GDP; 2) growth rate of nominal GDP; 3) growth rate of industrial production; 4) recession probability; 5) civilian unemployment rate; 6) growth rate of corporate profit; 7) growth rate of housing starts; 8) growth rate of GDP price index (Inflation). Such survey data has the merit of being forward looking and has been used to predict stock and bond market return by Eriksen (2017) and Deng, Wang, and Zhou (2024) among others. The updated data on SPF survey are available from Philadelphia Fed at a quarterly frequency.

We work with a data range of Jan 1985 to Dec 2023, since that is when the indexes of economic policy uncertainty begin. We further transform the quarterly survey data into monthly observations

using the same manner of Huang, Li, and Wang (2021), such that all predictors data are now available at monthly frequency.

3.2 Predictive regression model

Consider the following predictive regression model for equity market premium:

$$r^m_{t+h} = \alpha + \beta' \boldsymbol{g}^J_t + \varepsilon_{t+h}, \;\; J = SEM, PC, PLS, \tag{12}$$

where r^m_{t+h} is the excess aggregate equity market return over a period of time t to $t+h$; $\boldsymbol{g}^{SEM}_t$ is our single optimal predictive factor extracted by the novel supervised EM algorithm; and $\boldsymbol{g}^{PC}_t$, $\boldsymbol{g}^{PLS}_t$ are the PCA and PLS factors, respectively. The above equation is simply a particular case of our general factor predictive model (1) in Section 2, with the target being equity market return. We start from predicting return at monthly horizon ($h = 1$ month), but will also examine longer horizon prediction (h up to 1 year) in later sections.

We focus on the out-of-sample return predictability of our single optimal predictive factor, and compare against PCA and PLS factors. Specifically, we follow Welch and Goyal (2008); Goyal et al. (2023), among many others, to estimate the predictive regression model and generate return prediction recursively using an expanding window scheme. In particular, at each time t, the out-of-sample return forecast $\hat{r}^m_{t+h}$ given all predictor information up to t is given by,

$$\hat{r}^m_{t+h} = \hat{\alpha}_t + \hat{\beta}'_t \hat{\boldsymbol{g}}^J_t, \;\; J = SEM, PC, PLS, \tag{13}$$

where $\hat{\alpha}_t, \hat{\beta}$ are coefficients estimates of regressing past returns $\{r^m_{s+h}\}_{s=1}^{t-h}$ on a constant and the estimated predictive factor $\{\hat{\boldsymbol{g}}^J_s\}_{s=1}^{t-h}$ considered. Importantly, the optimal predictive factor, as well as PCA and PLS factors, are also estimated recursively at each date with information up to period t. Since SEM estimation also depends on a parameter r, that is the number of latent factors underlying the predictors. We also require r to be estimated recursively with its value pined down through the information criteria of Bai and Ng (2002).

We split the data into two parts. The first $p = 180$ months (15 years) of observations from 1985:01 to 1999:12 are set as the initial estimation window, so that the first monthly return forecast pertains to 2000:01. As the estimation window expands recursively each month, we then make monthly return forecasts for 2000:02, 2000:03, ..., 2023:12. Hence, the remaining 288 (24 years) of observations from 2000:01 to 2023:12 are out-of-sample forecasts periods.

We evaluate the accuracy of our return forecasts $\hat{r}^m_{t+h}$ by optimal predictive factor and PCA, PLS factors against a benchmark return forecast. As a literature convention, we use the historical average (prevailing mean) of excess aggregate market return computed as,

$$\bar{r}^m_{t+h} = \frac{1}{t-h}\sum_{s=1}^{t-h}\{r^m_{s+h}\}. \tag{14}$$

Welch and Goyal (2008); Goyal et al. (2023) argue that historical average return already offers a stringent out-of-sample benchmark hard to beat by many individual macro variables. It is then common to report the Campbell and Thompson (2008) R^2_{OS}, computed as

$$R^2_{OS} = 1 - \frac{\sum_p^{T-h}(r^m_{t+h} - \hat{r}^m_{t+h})^2}{\sum_p^{T-h}(r^m_{t+h} - \bar{r}^m_{t+h})^2} \tag{15}$$

which measures the percentage reduction in mean squared forecast error (MSFE) of our competing predictive model against the historical average return benchmark.

We expect a positive $R^2_{OS} > 0$ by our optimal predictive factor, which implies our return forecasts can generate a lower MSFE than the benchmark. Clark and West (2007) rely on a MSFE-adjusted statistic to test the null hypothesis of $H_0: R^2_{OS} \leq 0$ against a one sided upper tail alternative of $H_A: R^2_{OS} > 0$. The CW testing procedure can be implemented through a standard t-test in a simple regression from,

$$(r^m_{t+h} - \bar{r}^m_{t+h})^2 - (r^m_{t+h} - \hat{r}^m_{t+h})^2 + (\bar{r}^m_{t+h} - \hat{r}^m_{t+h})^2 = c + \varepsilon_{t+h}, \tag{16}$$

where the left hand side is the reduction in squared forecast error by the competing model against the historical mean benchmark plus an adjustment term. The adjustment reflects the fact that competing predictive model nests the historical mean benchmark, so it inherently produces more noisy forecast and hence lower R^2_{OS} under the null hypothesis of no predictability. The t-test corresponding to the estimate of constant c on the right hand side is then equivalent to testing H_0 against H_A.

3.3 Optimal predictive factor

We start by understanding the loadings of the single optimal predictive factor, SEM, and contrast with that of the first PCA and PLS factor. Figure 1 plots the weights of each individual predictor on the recursively estimated predictive factors of SEM, PC, and PLS, separately in three panels, over the out-of-sample period. We re-scale each factor by its standard deviation, so that the magnitudes of predictors weights can be compared across the factors.

We observe from Figure 1 that loadings pattern do share some similarity across the SEM, first PCA and PLS factors. As such, the three factors will be correlated. However, the weights of first PCA factor are a lot more sparse, with heavier loading on expected real and nominal GDP growth, fiscal policy uncertainty, output gap and dividend yield. The rest of predictors have rather trivial loadings. In contrasts, weights by PLS and SEM factors are denser and more extreme. In particular, both factors place sizable weights across all categories of predictors. The absolute magnitudes of weights on expected economic conditions and some macro variables are also a lot higher than that of first PCA factor. Nonetheless, weights by SEM factor still differ substantial from PLS especially for predictors like expected inflation, fiscal policy uncertainty and dividend yield.

Figure 2 further plots the forecasted equity market return based on the recursively constructed optimal predictive factor SEM, against the return forecasts by the first PCA and PLS factor over the out-of-sample period. As expected, we observe that the three forecasts are correlated with each other. Among these three, the return forecasts by the first PCA factor (dotted line) are relatively smooth and flat compared to the realized equity market return (solid line in light blue). In comparison, the return forecasts by first PLS factors (dashed line) turns less smooth and pick up some of short run return movement. By contrast, forecasted return by SEM factor (solid line in green) are much more volatile, and captures a lot of large return spikes, both positive and negative ones. Except for some difference around 2011, the blue (fixed r) and green (recursive r) solid lines closely track each other, indicating the robustness of SEM forecasts to the hyper-parameter r. These results suggest that with more diverse predictors weights of larger magnitudes, the constructed optimal predictive factor of SEM indeed create more expected variations in market return than PCA and PLS factors.

3.4 Statistical accuracy and link to economy

We next evaluate the accuracy of out-of-sample return forecasts by the optimal predictive factor of SEM. Here, we compare against not only the return forecast by the first PCA and PLS factors, but also those by the first two, three or more PCA and PLS factors. We estimate the number of

PCA factors again through the information criteria of Bai and Ng (2002), and rely on Bayesian information criteria to select number of PLS factors, both recursively over the out-of-sample period.

Table 2 presents the out-of-sample R^2_{OS}s of each factor model based return forecasts, along with the MSPE-adjusted statistic to judge their statistical significance. Panel B reveals, while the first PC factor produces a positive R^2_{OS} of 0.44%, it's not statistically significant according to the MSFE-adjusted statistic. Adding a second PCA factors raises the R^2_{OS} to 1.26% and thus delivers a lower MSFE than the historical average benchmark. The out-performance by the two PCA factors forecasts is also statistical significant according to the MSPE-adjusted statistic. This finding confirms our argument that, though explaining the most variance of the predictors, the first PCA factor is not supervised by the variance of stock return. Therefore, a single PCA factor does not necessarily serve the purpose of prediction. The next few rows in Panel B indicates that adding a third PC factor or recursively estimate the number of PCA factors to use deteriorate the forecast performance. This is probably due to estimating more factors that introduce more estimation error and propagate into the the return forecasts in the out-of-sample environment.

Panel C turns to the forecasts by PLS factors. By taking into account the covariance with stock return, the first PLS factor now generates a positive and significant R^2_{OS} of 1.07% and thus less forecast error than historical average benchmark. While adding a second PLS factor slightly reduce the R^2_{OS} to 1%, including up to three PLS factors could double the R^2_{OS} to 1.92%. Theses magnitude of R^2_{OS}s by PLS factors are comparable or larger than the two PC factor forecasts. Recursively estimating the number of PLS factors does not seem to further improve the forecast performance, though it still beat the best PCA factors forecasts. The MSFE-adjusted statistic also confirm the significance of PLS factors. These evidence suggests a single or first few PLS factor are more powerful than PCA in capturing equity risk premium information.

In contrast, Panel A shows the optimal predictive factor of SEM exhibits much stronger out-of-sample predictive ability for aggregate stock market return. Its R^2_{OS} could reach 5.50% when the parameter r of SEM is estimated recursively over time. The magnitude of R^2_{OS} more than double that of the best PLS factors forecasts, and the MSPE-adjusted statistics are also much larger, affirming their superior forecast accuracy against the historical average benchmark are significance. Importantly, this improvement forecast performance is achieved by a single factor, suggesting the efficiency of SEM algorithm in extracting the predictive information from a large number of predictors.

Aside from the overall R^2_{OS}, we further investigate the out-of-sample predictability over different stages of the business cycle to better understand its driving force. Accordingly, we compute R^2_{OS} separately for NBER expansions and recession episodes during the out-of-sample

period. Column "Expansion" and "Recession" in Table 2 shows that return predictability is higher in recession for PCA and PLS factors forecasts. For instance, R^2_{OS}s by the two PCA factors forecasts are 2.03% in recession, but only 0.97% in expansion. Likewise, while the overall R^2_{OS} by three PLS factors is 1.92%, it's 2.64% in recession and 1.65% in expansion. These documented higher R^2_{OS} in recession periods are consistent with the findings in literature, e.g. Rapach et al. (2010), suggesting the role of a time varying risk premium channel. In comparison, the optimal predictive factor of SEM further improve the R^2_{OS} in recession to 4.37%, a more than 60% increase than PLS forecasts. And at the same time, the R^2_{OS} by SEM in expansion reaches as high as 5.94%. As such, the improvement in forecast accuracy is driven by better forecast in both recession and expansion. The results suggest that SEM factor has captured predictive information beyond the traditional ones reflected in business cycle condition.

To further diagnose the forecast performance over time, Figure 3 plots the time series of difference between cumulative squared prediction error (CSPE) by the benchmark return forecasts of historical average and the CSPE by the predictive regressions forecasts based on SEM, PCA and PLS factors. At any particular out-of-sample period, an increase of difference in CSPE and thus upward slope of the curve indicates the factor model forecasts outperform the historical average benchmark in statistical accuracy, while the opposite is true when the curve is downward sloping. Therefore, this Figure provides an informative graphical device to assess the consistency of out- or under-performance of the model forecast against the benchmark forecast over time.

The solid line in Figure 3 demonstrate that the single optimal predictive factor consistently out-perform the historical average benchmark. The curve has been predominantly upward sloping over the out-of-sample period, except for the few years around 2005. This suggests the superior forecast accuracy by single SEM factor steps from the majority of the sample and is not driven by any specific episode. In particular, the CDSPE line not only exhibits steep upward slopes in NBER recessions, despite some deterioration in the later of the 2008 recession, but also witnesses very steady increase in the expansion periods. The results are consistent with our earlier finding in Table 2 that R^2_{OS} stay sizable and positive in both recession and expansion. For comparison, the dotted and dashed lines present the CDSPE by PCA and PLS factors forecasts. Though getting upward sloped in certain episodes, the curves stay relatively flat or even negatively sloped for extended period of time especially during expansion. The CDSPE line by SEM factor therefore dominates those by PCA and PLS factors forecasts, affirming the consistency of its superior predictive ability.

3.5 Long horizon prediction

While our focus is on equity risk premium prediction at monthly horizon by the optimal predictive factor, it is also of interest to investigate its prediction on market excess return at longer horizon. In the literature, Cochrane (2011) show that stock return predictability by dividend price ratio strengthens in terms of in-sample R^2 as the forecast horizon extends longer. Moreover, the phenomena of long horizon predictability is mechanically linked to short horizon predictability and the persistence of predictors. Deng et al. (2024) document similar finding using expected macro condition from survey of professional forecasters as predictors. However, Huang et al. (2015) present evidence that predictive power by index of investor sentiment proxies peaks at around 9-month horizon. This is probably due to mispricing from investor sentiment been faded away over longer period of time.

In this subsection, we re-estimate the optimal predictive factor using longer horizon return as the supervising target. Table 3 records the R^2_{OS}s of predicting 3-month, 6-month, and 12-month ahead stock market excess return, respectively, by the corresponding single optimal predictive factor against the first few PC and PLS factors. Three sets of observations follow. First, the predictability by all factor forecasts have been raised substantially at longer forecast horizon. While R^2_{OS} by PCA and PLS factors barely exceeds 2% at monthly forecast horizon, they shoot up to a range of around 3% to 10% at longer horizon. Second, for PCA and PLS factors, the predictive power increase steadily as we extend the forecast horizon. For example, R^2_{OS}s by PCA factors were generally small and insignificant at monthly horizon, but it goes up gradually to about 3% for 3-month , 5% for 6-month, and 8% for 1-year horizon. PLS forecasts also display a similar patter with R^2_{OS} raised up to about 11% at 1-year horizon. Interestingly, single factor also tends to perform comparable to multiple factors at longer forecast horizons. Third, the SEM factor still possesses the strongest predictive power at longer horizon. The R^2_{OS} increase to 13.47% at 3-month horizon, about double that of the PCA and PLS forecasts. R^2_{OS} then decreases a bit to 12.67% at 6-month horizon, which still out-perform PC and PLS forecasts by a large margin. At 1-year forecast horizon, with R^2_{OS} of 10.54%, performance of SEM forecast beats PCA but is on par with PLS. Overall, though predictive R^2_{OS} seem to peak at 3-months horizon, optimal predictive factor remain valid in efficiently extracting predictive information for long horizon equity risk premium.

3.6 Asset allocation

We next investigate the economic value of accessing the optimal predictive factor in making portfolio allocation decisions. Particularly, we consider a mean variance investor who recursively

allocate her wealth between aggregate equity market portfolio and a risk free asset of Treasury bill using the out-of-sample market return forecasts. At each time t, the investor optimally allocation

$$w_t = \frac{1}{\gamma}\frac{\hat{R}^m_{t+h}}{\hat{\sigma}^2_{t+h}} \tag{17}$$

proportion of wealth to the aggregate equity market portfolio and $1 - w_t$ to the risk free asset. γ denotes the investor's relative risk aversion; $\hat{R}^m_{t+h}$ and $\hat{\sigma}^2_{t+h}$ are the out-of-sample predictions of expected mean and variance of excess simple equity market return. The return forecasts are based on optimal predictive factor, as well as the PC and PLS factors; variance forecasts are built as the sample variance of past realized monthly return over a 5 year moving window. Following Campbell and Thompson (2008), portfolio weights w_t are restricted to stand in between 0 to 1.5, excluding short sell and limiting the leverage ratio of 50%.

We then measure the performance of investor's portfolio by computing the Sharpe ratio (SR) and certainty equivalent return (CER) as

$$SR_p = \frac{\mu_p}{\sigma_p}, \;\; CER_p = \mu_p - 0.5\gamma\sigma_p^2, \tag{18}$$

where μ_p and σ_p are the sample mean and variance of the excess return of investor's portfolio over the out-of-sample portfolio allocation period. We report the Sharpe ratio as well as the CER gain as the difference in CER by investor who uses the optimal predictive factor, or PC and PLS factors, in return forecasts and who uses the historical average return. The Sharpe ratio are multiplied by $\sqrt{12}$ to be annualized and the difference in CER are multiplied by 12, so that the CER gain represents the annualized management fees an investor would willing to pay to have access to the optimal predictive factor instead of the historical average return forecasts. We measure performance of portfolio decision based on risk aversion coefficients of 3. Additionally, we also take into account both the case of no transaction cost and a 50 basis point transaction when investor re-balance the portfolio over the out-of-sample period.

Table 4 shows that PC forecasts generally produce low economic gains for a mean variance investor, consistent with their low R^2_{OS}s. The CER gains are close to 0 when we use a single PC factor, and reach about 1% if we use multiple PC factors. The net of transaction cost CER gains are even lower, at about 0.3% to 0.5%. The Sharpe ratios stay around 0.20 to 0.27, depending on the number of PCA factors in use. PLS forecasts perform better economically, with CER gains ranging from 3.29% to 6.40%, and Sharpe ratios from 0.38 to 0.57. The improvements in economic value align with the better statistical accuracy by PLS factors. Compared with other factor models, the

optimal predictive factor of SEM again stands out in terms of economic value. The CER gain reaches as high as 7.8% and remains economically large at 6.38% net of transaction costs. This means an investor would be willing to pay more than 6% per year to get access to return forecasts by SEM factor as opposed to using historical mean. The resultant Sharpe ratio of 0.65 is also well above those by the PC and PLS forecasts.

3.7 Comparison with alternative methods

The analysis above have focused on dimension reduction through factor models to extract predictive information from a large set of predictors. However, the literature has also proposed alternative methods to aggregate information from many predictors. These methods include forecast combination and linear machine learning methods such as LASSO and Elastic Net (see e.g. Bates and Granger (1969); Rapach et al. (2010); Rapach, Strauss, and Zhou (2013); Gu et al. (2020); Rapach and Zhou (2020)). As opposed to factor model, they essentially rely on shrinkage to consolidate information from many predictors while guarding against over-fitting.

In this subsection, we apply such shrinkage methods to our predictors set and compare their forecast performance with our optimal predictive factor. Following Dong et al. (2022), we consider four approaches. The first is predictor average, where we take cross sectional average of all predictors and run predictive regression using this average as the new explanatory variable. The second is a simple forecast combination, where we start by running a set of uni-variate predictive regression of stock return on each individual predictor and then take the arithmetic mean of the uni-variate forecasts. The third is Elastic Net (Enet), where run multi-variate predictive regression with a penalty term that shrink the coefficients toward zero. The penalty term includes both an l_1 (LASSO) component that performs variable selection, and an l_2 (Ridge) component that allows correlated variables to be included simultaneously. The fourth is combination Enet (C-Enet), where we run ENet regression of actual return on return forecasts by individual predictor first and then take mean of the selected return forecasts. This offers a refinement to the simple combination, which could become too conservative when number of predictors is large.

Panel D of Table 2 and 3 presents the out-of-sample forecast results by these alternative methods. For monthly forecast horizon, we find C-Enet delivers best performance among the four methods followed by predictor average. The R^2_{OS} by C-Enet reaches 1.98%, exceeding that by PLS factors, but it still under-perform SEM factor by more than 50%. At a forecast horizon of 3-month to 6-month, three out of four methods perform reasonably well. R^2_{OS}s by predictor average, Enet and C-Enet goes up to about 5%, in between those by PCA and PLS factors. But as before, they

are way below the level of above 12% by SEM. At longer horizon of 12-month, however, C-ENet performs particularly well, with R^2_{OS} reaching above 10%. The magnitude is comparable to both multiple PLS and the single SEM factor. Panel D of Table 4 further shows the economic value of alternative forecasts. The CER gain by C-Enet stands out among the four alternative methods and is as high as 5.2%, but 2.6% lower than SEM factor. It also drop by more than 50% to the level of 2.5% when we take into account transaction cost. Similar pattern for the Sharpe ratio can also be observed. In sum, our evidence indicates the shrinkage methods also work in our return forecasting context, but their predictive power are still dominated by the optimal predictive factor of SEM.

3.8 Stock volatility, bond return and macroeconomic predictions

In this subsection, we demonstrate the pivotal role of optimal predictive factor not only in forecasting aggregate stock market return, but also in forecasting stock market volatility, Treasury bond excess return and key macroeconomic quantity.

Whereas understanding the movement of first moment of stock market excess return, i.e. equity risk premium, has been the focus of a lot of literature, the second moment, i.e. volatility, is more variable. Theoretically, aggregate stock market volatility could be affected by macroeconomic condition and sentiment for a few reasons. In particular, the counter-cyclicality of volatility of aggregate cash flow as well as asymmetric response of stock return to change in economic condition over the business cycle could both play an important role (Bansal and Yaron (2004); Mele (2007)). Accordingly, we re-estimate the optimal predictive factor from the same set of 50 predictors that cover macro variables, economic policy uncertainty, investor sentiment and expected economic conditions, to conduct out-of-sample forecast of the realized volatility of S&P 500 index return.

Additionally, since the seminal works of Stock and Watson (2002a,b), a voluminous of literature has studied the prediction of macroeconomic aggregates, using factors extracted from a large number of dis-aggregated economic indicators as predictors. Ludvigson and Ng (2009, 2011); Gargano, Pettenuzzo, and Timmermann (2019), among others, further show that bond returns in US Treasury market can be predicted significantly by such macro factors. Their results provide important insights on the link between economic fundamentals and bond market risk premium. As such, we also examine the out-of-sample predictability of US Treasury bond excess return as well as industrial production growth by applying our novel predictive factor model.

We measure realized stock market variance at month t by summing up daily squared market

excess return within the month, as well as their auto-covariance following the practice of French, Schwert, and Stambaugh (1987). Denote r_j to be the daily excess return at day j in month t and N_t as the number of days in month t, the realized variance is therefore,

$$RV_t = \sum_{1}^{N_t} (r_j - \bar{r}_t)^2 + 2 \sum_{1}^{N_t - 1} (r_j - \bar{r}_t)(r_{j+1} - \bar{r}_t),$$

where $\bar{r}_t$ is the average daily market excess return within the month. We predict the logarithm of this realized stock market variance at a horizon of one month, that is we make out-of-smaple forecast on $Log(RV_{t+1/12})$ with $h = 1/12$ year.

The bond excess return are computed based on the constant maturity Treasury yield curve data of Liu and Wu (2021). This dataset overcomes some limitations of the yield data by Fama and Bliss (1987); Gürkaynak, Sack, and Wright (2007) and provides better fits of observed Treasury yields. Denote $y_t(n)$ as the Treasury yield at time t of a maturity n years, the log price of the n year Treasury bond then relates to the yield through $p_t(n) = -ny_t(n)$. The holding period return of buying an n year bond and sell it a month later would be $r_{t+1/12}(n) = p_{t+1/12}(n - 1/12) - p_t(n)$, and the excess return is $rx_{t+1/12}(n) = r_{t+1/12}(n - 1/12) - y_t(1/12)$. Following the practice of Cochrane and Piazzesi (2005), we predict the average bond excess return across the maturity of 2-year to 5-years, $\bar{rx}_{t+1/12}(n) = 1/4 \sum_{n=2}^{5} rx_{t+1/12}(n)$. Industrial production data are from the FRED database of St. Louis Fed. We compute monthly growth rate of industrial production as the difference of the logarithm of industrial production, $log(IP_{t+1/12}/IP_t)$.

In forecasting bond excess return and industrial production growth, we rely on the monthly FRED-MD database of McCracken and Ng (2016), maintained and updated regularly on the website of St Louis Fed. This macro dataset contains monthly release of a large number of economic indicators, which cover a broad range of categories that include: output and income; labor market; housing; consumption, orders and inventories; money and credit; interest and exchange rates; prices; stock market. We adhere to the guidelines in McCracken and Ng (2016) to apply various transformations to the underlying series, which ensures that the variables we use are stationary and based on economic reasoning. We end up with a balanced panel of 122 indicators. While we used the recent vintage, data on all indicators are lagged by two months to account for publication lags and potential data revisions.

Table 5 presents the out-of-sample R^2_{OS}s and the Clark and West (2007) MSFE-adjusted statistics for predicting monthly stock market volatility, monthly bond excess return and monthly growth rate of industrial production. The out-of-sample periods are over 2000:01 to 2023:12, 1976:06 to 2021:12 and 1975:01 to 2022:12, respectively. To account for the persistence in market

volatility and industrial production growth, we use an AR(1) forecasts by their lagged value as the benchmark. We compare the forecast performance by the optimal predictive factor of SEM against those by PC and PLS factors, as well as alternative methods based on shrinkage. For market volatility prediction, we observe that two PC factors and Enet forecasts both generate R^2_{OS} above 2%, while SEM factor has an R^2_{OS} of 4.12% that is statistically significant. Turning to the bond return forecast, out-performance by SEM is even more evident. R^2_{OS} by SEM reaches 15.28%, while the two PLS factors as the best among all other methods has an R^2_{OS} of 7.76%. Industrial production growth is predictable by many approaches. Multiple PC or PLS factors and the combination Enet method could all bring an R^2_{OS} of above 12%. SEM factor still attains an even higher R^2_{OS} of 15.56%. In sum, our evidence confirm the effectiveness of optimal predictive factor in a variety of forecast applications beyond equity risk premium prediction.

4 Conclusions

In this article, we propose a novel method to estimate a single optimal predictive factor in a predictive regression. The single factor can be estimated in one step as the best linear combination of a large set of predictors. The estimation relies on a maximum likelihood approach implemented through expectation maximization algorithm. As such, the single factor is termed supervised EM, SEM. The single SEM factor outperforms existing and popular dimension reduction techniques of PCA and PLS, which typically involve multiple factors.

Theoretically,our approach is a surprising discovery that, despite the number of relevant factors, there always exists a single relevant factor equivalent model. We then establish the asymptotic consistency of SEM in recovering the single relevant factor. In contrast, we show that PLS cannot consistently estimate the relevant factor. The root cause lies in that identification condition underlying a target relevant factor model is different to those in standard large dimensional factor analysis. We also provide simulation evidence to support our arguments.

Empirically, we extract a single factor for equity risk premium prediction from a large set of macro variables, investor sentiment proxies, economic policy uncertainty measures and survey based expected macro condition. We identify a single best combination of the predictors and show that the single factor delivers the best forecast accuracy of equity market excess return in an out-of-sample environment. We further apply the SEM method to predicting stock market volatility, Treasury bond excess return, and industrial production growth. The SEM method outperforms not only both the PCA and PLS factors, but also the state of art methods of forecast combination and machine learning as used in Dong et al. (2022). Since big data and high dimension are common in

finance, wide application of the SEM are expected.

References

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.

Bai, J., Li, K., 2012. Statistical analysis of factor models of high dimension. The Annals of Statistics 40, 436–465.

Bai, J., Li, K., 2016. Maximum likelihood estimation and inference for approximate factor models of high dimension. Review of Economics and Statistics 98, 298–309.

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

Bai, J., Ng, S., 2008. Forecasting economic time series using targeted predictors. Journal of Econometrics 146, 304–317.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. The journal of Finance 61, 1645–1680.

Baker, M., Wurgler, J., 2007. Investor sentiment in the stock market. Journal of economic perspectives 21, 129–151.

Baker, S. R., Bloom, N., Davis, S. J., 2016. Measuring economic policy uncertainty. The quarterly journal of economics 131, 1593–1636.

Baker, S. R., Bloom, N., Davis, S. J., Kost, K. J., 2019. Policy news and stock market volatility. Tech. rep., National Bureau of Economic Research.

Baker, S. R., Bloom, N., Davis, S. J., Sammon, M., 2021. What triggers stock market jumps? Tech. rep., National Bureau of Economic Research.

Bansal, R., Yaron, A., 2004. Risks for the long run: A potential resolution of asset pricing puzzles. The journal of Finance 59, 1481–1509.

Bates, J. M., Granger, C. W., 1969. The combination of forecasts. Journal of the operational research society 20, 451–468.

Bianchi, D., Büchner, M., Tamoni, A., 2021. Bond risk premiums with machine learning. The Review of Financial Studies 34, 1046–1089.

Bork, L., Møller, S. V., Pedersen, T. Q., 2020. A new index of housing sentiment. Management Science 66, 1563–1583.

Campbell, J. Y., Thompson, S. B., 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? The Review of Financial Studies 21, 1509–1531.

Clark, T. E., West, K. D., 2007. Approximately normal tests for equal predictive accuracy in nested models. Journal of econometrics 138, 291–311.

Cochrane, J. H., 2011. Presidential address: Discount rates. The Journal of finance 66, 1047–1108.

Cochrane, J. H., Piazzesi, M., 2005. Bond risk premia. American economic review 95, 138–160.

Deng, Y., Wang, Y., Zhou, T., 2024. Macroeconomic expectations and expected returns. Journal of Financial and Quantitative Analysis pp. 1–70.

Dong, X., Li, Y., Rapach, D. E., Zhou, G., 2022. Anomalies and the expected market return. The Journal of Finance 77, 639–681.

Eriksen, J. N., 2017. Expected business conditions and bond risk premia. Journal of Financial and Quantitative Analysis 52, 1667–1703.

Fama, E. F., Bliss, R. R., 1987. The information in long-maturity forward rates. The American Economic Review pp. 680–692.

French, K., Schwert, G. W., Stambaugh, R. F., 1987. Expected stock returns and volatility. Journal of Financial Economics 19, 3–30.

Gargano, A., Pettenuzzo, D., Timmermann, A., 2019. Bond return predictability: Economic value and links to the macroeconomy. Management Science 65, 459–954.

Giglio, S., Xiu, D., 2021. Asset pricing with omitted factors. Journal of Political Economy 129, 1947–1990.

Giglio, S., Xiu, D., Zhang, D., 2023. Prediction when factors are weak. University of Chicago, Becker Friedman Institute for Economics Working Paper .

Goyal, A., Welch, I., Zafirov, A., 2023. A comprehensive look at the empirical performance of equity premium prediction ii. Tech. rep., Forthcoming at REview of Financial Studies.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

Gürkaynak, R. S., Sack, B., Wright, J. H., 2007. The us treasury yield curve: 1961 to the present. Journal of monetary Economics 54, 2291–2304.

Henkel, S. J., Martin, J. S., Nardari, F., 2011. Time-varying short-horizon predictability. Journal of financial economics 99, 560–580.

Huang, D., Jiang, F., Li, K., Tong, G., Zhou, G., 2022. Scaled pca: A new approach to dimension reduction. Management Science 68, 1678–1695.

Huang, D., Jiang, F., Tu, J., Zhou, G., 2015. Investor sentiment aligned: A powerful predictor of stock returns. The Review of Financial Studies 28, 791–837.

Huang, D., Li, J., Wang, L., 2021. Are disagreements agreeable? evidence from information aggregation. Journal of Financial Economics 141, 83–101.

Kelly, B., Malamud, S., Zhou, K., 2024. The virtue of complexity in return prediction. The Journal of Finance 79, 459–503.

Kelly, B., Pruitt, S., 2013. Market expectations in the cross-section of present values. The Journal of Finance 68, 1721–1756.

Kelly, B., Pruitt, S., 2015. The three-pass regression filter: A new approach to forecasting using many predictors. Journal of Econometrics 186, 294–316.

Kelly, B. T., Pruitt, S., Su, Y., 2019. Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics 134, 501–524.

Kozak, S., Nagel, S., Santosh, S., 2020. Shrinking the cross-section. Journal of Financial Economics 135, 271–292.

Lettau, M., Pelger, M., 2020a. Estimating latent asset-pricing factors. Journal of Econometrics 218, 1–31.

Lettau, M., Pelger, M., 2020b. Factors that fit the time series and cross-section of stock returns. The Review of Financial Studies 33, 2274–2325.

Light, N., Maslov, D., Rytchkov, O., 2017. Aggregation of information about the cross section of stock returns: A latent variable approach. The Review of Financial Studies 30, 1339–1381.

Lin, H., Wu, C., Zhou, G., 2018. Forecasting corporate bond returns with a large set of predictors: An iterated combination approach. Management Science 64, 4218–4238.

Liu, Y., Wu, J. C., 2021. Reconstructing the yield curve. Journal of Financial Economics 142, 1395–1425.

Lu, Y., Nikolaev, V. V., 2022. Expected loan loss provisioning: An empirical model. The Accounting Review 97, 319–346.

Ludvigson, S. C., Ng, S., 2009. Macro factors in bond risk premia. Review of Financial Studies 22, 5027–5067.

Ludvigson, S. C., Ng, S., 2011. A factor analysis of bond risk premia. In: Ullah, A., Giles, D. E. A. (eds.), *Handbook of Empirical Economics and Finance*, Chapman and Hall, pp. 313–372.

McCracken, M. W., Ng, S., 2016. Fred-md: A monthly database for macroeconomic research. Journal of Business & Economic Statistics 34, 574–589.

Mele, A., 2007. Asymmetric stock market volatility and the cyclical behavior of expected returns. Journal of financial economics 86, 446–478.

Pelger, M., Xiong, R., 2022a. Interpretable sparse proximate factors for large dimensions. Journal of Business & Economic Statistics 40, 1642–1664.

Pelger, M., Xiong, R., 2022b. State-varying factor models of large dimensions. Journal of Business & Economic Statistics 40, 1315–1333.

Pettenuzzo, D., Timmermann, A., Valkanov, R., 2014. Forecasting stock returns under economic constraints. Journal of Financial Economics 114, 517–553.

Rapach, D., Zhou, G., 2013. Forecasting stock returns. In: *Handbook of economic forecasting*, Elsevier, vol. 2, pp. 328–383.

Rapach, D. E., Strauss, J. K., Zhou, G., 2010. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. The Review of Financial Studies 23, 821–862.

Rapach, D. E., Strauss, J. K., Zhou, G., 2013. International stock return predictability: What is the role of the united states? The Journal of Finance 68, 1633–1662.

Rapach, D. E., Zhou, G., 2020. Time-series and cross-sectional stock return forecasting: New machine learning methods. Machine learning for asset management: New developments and financial applications pp. 1–33.

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

Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. The Review of Financial Studies 21, 1455–1508.

Wold, H., 1966. Estimation of principal components and related models by iterative least squares. Multivariate analysis pp. 391–420.

Wold, H., 1975. Path models with latent variables: The nipals approach. In: *Quantitative sociology*, Elsevier, pp. 307–357.

Table 1: Forecasting performance in Simulation

T	N	PC1	PC2	PC3	PC-all	PLS1	PLS2	PLS3	SEM
50	50	1.352	1.302	1.274	1.254	1.302	1.277	1.501	1.142
50	100	1.354	1.288	1.248	1.194	1.304	1.235	1.445	1.146
50	200	1.344	1.254	1.200	1.129	1.29	1.152	1.312	1.098
100	50	1.298	1.248	1.186	1.124	1.257	1.151	1.313	1.052
100	100	1.354	1.281	1.225	1.119	1.304	1.151	1.345	1.085
100	200	1.293	1.224	1.165	1.067	1.249	1.098	1.322	1.049
200	50	1.354	1.283	1.231	1.119	1.307	1.127	1.191	1.064
200	100	1.306	1.237	1.168	1.062	1.266	1.078	1.188	1.032
300	200	1.308	1.235	1.174	1.047	1.267	1.059	1.182	1.036

Note: This table reports the simulated mean squared forecast errors (MSFEs) by a single SEM factor, and compare with multiple PCA and PLS factors. Simulation is based on $r = 5$ latent factors $f_{1t}...f_{5t}$ that drive a total of N predictors, and the forecast target is $y_{t+1} = f_{1t} + f_{2t} + \varepsilon_{t+1}$. We consider a series of sample size specifications, $(T,N) = \{(50,50),(50,100),(50,200),(100,50),(100,100),(100,200),(200,50).(200,100),(300,200)\}$. For each (T,N), we simulate a total of $T+50$ observations on target and N predictors. Factors are generated independently from a standard normal distribution and loadings from a standard uniform distribution. Predictors also contain idiosyncratic errors u_{it} that follow normal distribution of $N(0,\sigma_i)$. These idiosyncratic errors are set to be cross-sectional heterogeneous with $\sigma_i \sim U[0,2]$. For each simulated sample, we then use the first T observations for parameter training and the rest 50 for out-of-sample evaluation. We rely on a rolling window scheme so that each forecast always involve T observations. Reported are the median MSFEs of each forecasting method on the basis of 500 simulations. Columns PC1-PC3 and PCall denote the PC regression forecasts with 1,2,3 and all $r = 5$ factors; Columns PLS1-PLS3 denote the PLS forecasts with 1 to 3 factors; and Column SEM denotes the forecasts by a single SEM factor.

Table 2: Performance in Forecasting Market Returns

	Overall		Expansion		Recession	
Methods	R^2_{OS} (%)	CW-test	R^2_{OS} (%)	CW-test	R^2_{OS} (%)	CW-test
Panel A: SEM						
SEM1 (recursive r)	5.50	4.05***	5.94	4.08***	4.37	1.27
Panel B: PCA						
PC1	0.44	0.96	1.61	2.62***	-2.58	-1.01
PC1+PC2	1.26	2.36***	0.97	1.76**	2.03	1.72**
PC1-PC3	0.57	1.37*	0.08	0.84	1.85	1.64*
PC1-PCr (recursive r)	0.07	0.86	-0.13	0.70	0.58	0.51
Panel C: PLS						
PLS1	1.07	2.00**	1.01	1.82**	1.21	0.83
PLS1+PLS2	1.00	1.83**	1.30	1.92**	0.23	0.28
PLS1-PLS3	1.92	2.45***	1.65	2.09**	2.64	1.30*
PLS1-PLSk (recursive k)	1.82	2.33***	2.10	2.27**	1.09	0.64
Panel D: Other methods						
Predictor avg	0.99	1.77**	0.51	1.20	2.25	1.57*
Combination	0.36	1.16	0.50	2.07**	-0.02	0.06
Enet	0.47	1.23	0.56	1.28*	0.23	0.22
C-Enet	1.98	2.38***	1.66	1.96**	2.82	1.41*

Note: This table reports the R^2_{OS} and Clark and West (2007) MSFE-adjusted statistics for predicting monthly market returns. The MSFE-adjusted statistic is used for testing $H_0 : R^2_{OS} \leq 0$ against $H_A : R^2_{OS} > 0$. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. The tuning parameter r is determined by the information criteria of Bai and Ng (2002), and the tuning parameter k is determined by BIC criteria. We apply the constraint of Pettenuzzo et al. (2014) to the methods subject to overfitting. The out-of-sample period is over 2000:01-2023:12. Expansion and recession correspond to NBER-dated expansion and recession periods

Table 3: Forecasting longer-horizon market returns

	h = 3 months		h = 6 months		h = 12 months	
Methods	R^2_{OS} (%)	CW-test	R^2_{OS} (%)	CW-test	R^2_{OS} (%)	CW-test
Panel A: SEM						
SEM1 (recursive r)	13.47	6.16***	12.67	6.26***	10.54	5.26***
Panel B: PCA						
PC1	2.95	2.51***	4.74	4.20***	8.38	4.90***
PC1+PC2	3.22	3.72***	4.69	3.85***	8.67	4.89***
PC1-PC3	2.2	2.78***	3.61	3.34***	5.75	4.73***
PC1-PCr (recursive r)	3.45	2.98***	5.37	3.62***	7.22	4.84***
Panel C: PLS						
PLS1	5.75	4.39***	6.90	4.43***	11.22	5.33***
PLS1+PLS2	4.88	4.00***	6.49	4.61***	8.74	4.66***
PLS1-PLS3	7.27	5.02***	9.24	5.53***	10.08	5.15***
PLS1-PLSk (recursive k)	6.24	4.63***	8.65	5.26***	7.41	4.53***
Panel D: Other methods						
Predictor avg	4.86	3.98***	5.87	4.68***	5.80	5.54***
Combination	1.08	2.14**	1.27	1.99**	1.79	3.48***
Enet	4.75	3.82***	4.12	2.99***	6.17	3.34***
C-Enet	4.35	3.78***	5.55	4.03***	10.22	4.83***

Note: This table reports the R^2_{OS} and Clark and West (2007) MSFE-adjusted statistics for predicting longer-horizon market returns. The MSFE-adjusted statistic is used for testing $H_0 : R^2_{OS} \leq 0$ against $H_A : R^2_{OS} > 0$. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. The tuning parameter r is determined by the information criteria of Bai and Ng (2002), and the tuning parameter k is determined by BIC criteria. We apply the constraint of Pettenuzzo et al. (2014) to the methods subject to overfitting. The out-of-sample period is over 2000:01-2023:12.

Table 4: Asset allocation

	No transaction cost		50bps transaction cost	
Methods	CER gain (%)	Sharpe ratio	CER gain (%)	Sharpe ratio
Panel A: SEM				
SEM1 (recursive r)	7.80	0.65	6.38	0.55
Panel B: PCA				
PC1	0.12	0.20	-0.37	0.16
PC1+PC2	1.26	0.26	0.46	0.20
PC1-PC3	0.99	0.26	0.28	0.21
PC1-PCr (recursive r)	1.17	0.27	0.35	0.21
Panel C: PLS				
PLS1	3.29	0.38	1.82	0.28
PLS1+PLS2	4.45	0.44	2.60	0.32
PLS1-PLS3	6.40	0.57	4.17	0.40
PLS1-PLSk (recursive k)	4.89	0.46	2.15	0.27
Panel D: Other methods				
Predictor avg	1.71	0.27	0.32	0.18
Combination	0.47	0.22	0.24	0.19
Enet	3.06	0.35	1.04	0.22
C-Enet	5.20	0.48	2.50	0.30

Note: This table reports the portfolio performance measures for a mean-variance investor with a risk aversion coefficient (γ) of 3, who hold the market portfolio or allocates monthly between stocks and risk-free bills using the out-of-sample forecasts of the excess market returns. CER gain is the annualized certainty equivalent return gain for the investor. Monthly Sharpe ratio is the mean portfolio return in excess of the risk-free rate divided by its standard deviation. We report the annualized Sharpe ratio which is calculated by multiplying monthly Sharpe ratio by $\sqrt{12}$. The out-of-sample evaluation period is over 2000:01-2023:12.

Table 5: Forecasting stock market volatility, bond returns, and industrial production

	Market volatility		Bond return		Industrial production	
Methods	R^2_{OS} (%)	CW-test	R^2_{OS} (%)	CW-test	R^2_{OS} (%)	CW-test
Panel A: SEM						
SEM1 (recursive r)	4.12	3.18***	15.28	8.00***	15.56	7.23***
Panel B: PCA						
PC1	1.41	2.28**	-0.14	0.94	10.92	5.62***
PC1+PC2	2.26	2.78***	2.03	2.71***	10.62	5.41***
PC1-PC3	1.21	2.50***	3.79	3.77***	12.74	5.79***
PC1-PCr (recursive r)	0.67	2.10**	3.52	4.41***	13.50	5.97***
Panel C: PLS						
PLS1	1.25	2.60***	6.55	6.18***	7.99	5.40***
PLS1+PLS2	0.17	2.86***	7.76	7.60***	13.61	6.35***
PLS1-PLS3	1.04	3.55***	-5.71	6.62***	10.37	6.24***
PLS1-PLSk (recursive k)	1.35	3.73***	-11.59	6.42***	6.61	5.72***
Panel D: Other methods						
Predictor avg	1.23	2.17**	-0.56	-0.18	8.75	5.59***
Combination	1.47	3.50***	1.10	6.14***	2.51	6.55**
Enet	2.17	3.22***	5.62	5.63***	10.33	5.88***
C-Enet	-2.98	3.24***	2.77	4.57***	12.06	5.82***

Note: This table reports the R^2_{OS} and Clark and West (2007) MSFE-adjusted statistics for predicting monthly stock market volatility, monthly bond returns (average of monthly excess returns of 2-year, 3-year, 4-year and 5-year US treasury bonds), and monthly growth of industrial production. The MSFE-adjusted statistic is used for testing $H_0: R^2_{OS} \leq 0$ against $H_A: R^2_{OS} > 0$. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. We apply 50 stock return predictors on volatility prediction, and use 122 FRED-MD variables to forecast bond returns and industrial production. The tuning parameter r is determined by the information criteria of Bai and Ng (2002), and the tuning parameter k is determined by BIC criteria. For stock market volatility and industrial production growth, we use AR(1) as benchmark, and employ $y_{t+1} = \alpha + \beta' \boldsymbol{F}_t + \gamma y_t + \varepsilon_{t+1}$ as the predictive regression of SEM, PCA, PLS, and predictor avg, where $\boldsymbol{F}_t$ is the factor(s). We combine the forecasts with historical average equally for the methods subject to overfitting. The out-of-sample periods are over 2000:01-2023:12, 1976:06-2021:12, and 1975:01-2022:12 for stock market volatility, bond return, and industrial production, respectively.

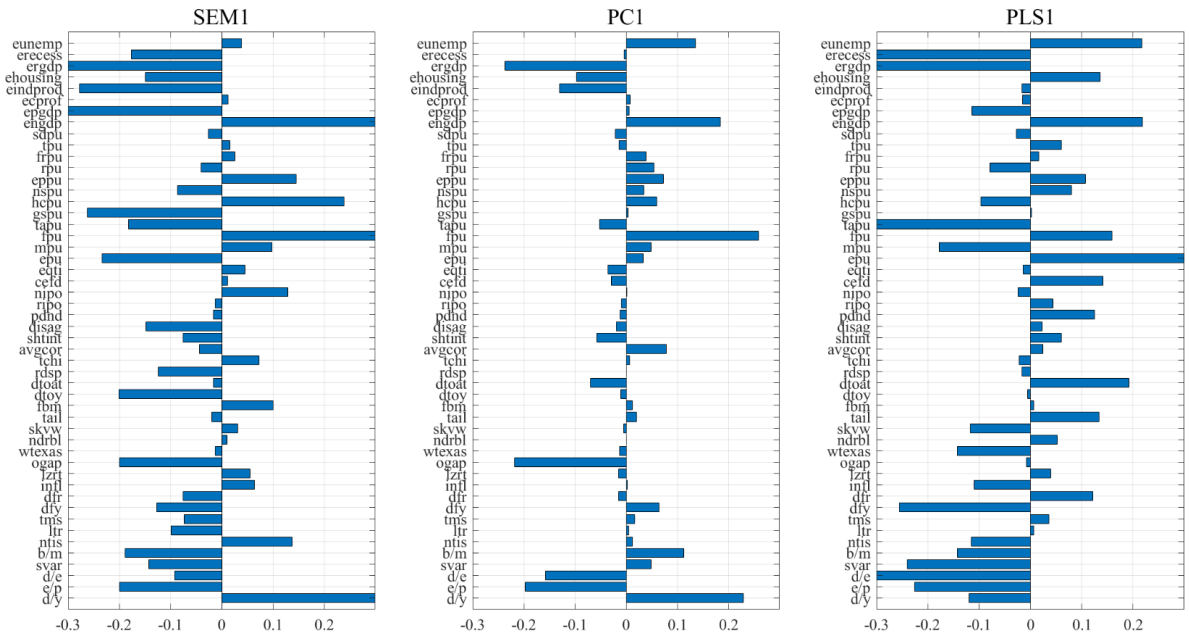


Figure 1: Factor loading

Note: This figure depicts the weights of the 50 stock return predictors for the first SEM, PC, and PLS factors. We recursively estimate the factors based on information up to the period of forecast formation period t alone and store the factors' last values, and then estimate the weights by regressing the obtained time series on the 50 predictors during out-of-sample period. To facilitate easier comparison, all the predictors are standardized by their corresponding standard deviation.

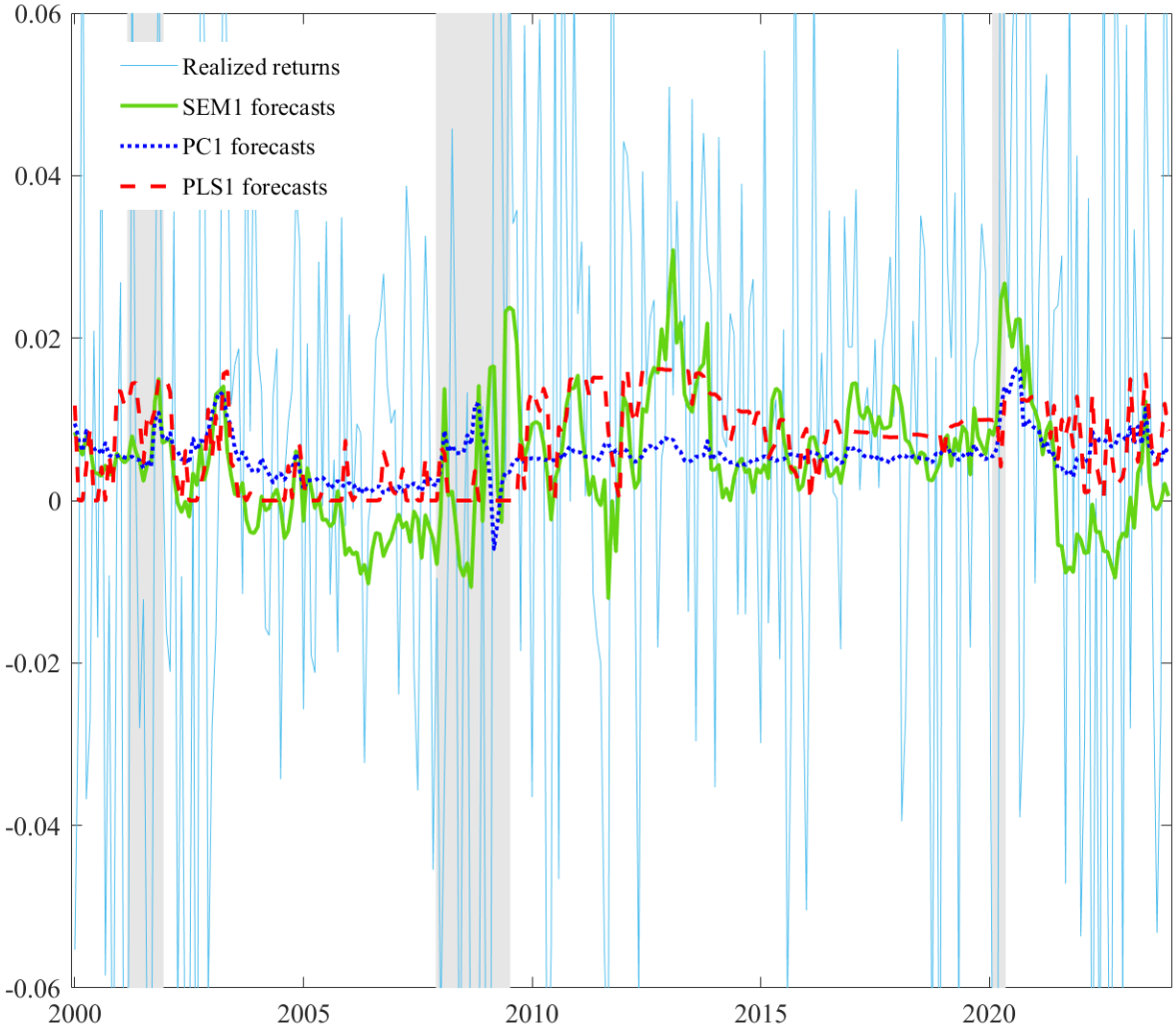


Figure 2: Stock market returns and the out-of-sample forecasts

Note: The green solid line depicts the out-of-sample predictive regression forecast for excess market return based on the recursively constructed SEM1 factor. The dotted line depicts the out-of-sample predictive regression forecast for excess market return based on the recursively constructed PC1 factor. The dashed line depicts the out-of-sample predictive regression forecast for excess market return based on the recursively constructed PLS1 factor. The factors and excess market return forecasts are estimated recursively based on information up to the period of forecast formation period t alone. The vertical bars correspond to NBER-dated recessions.

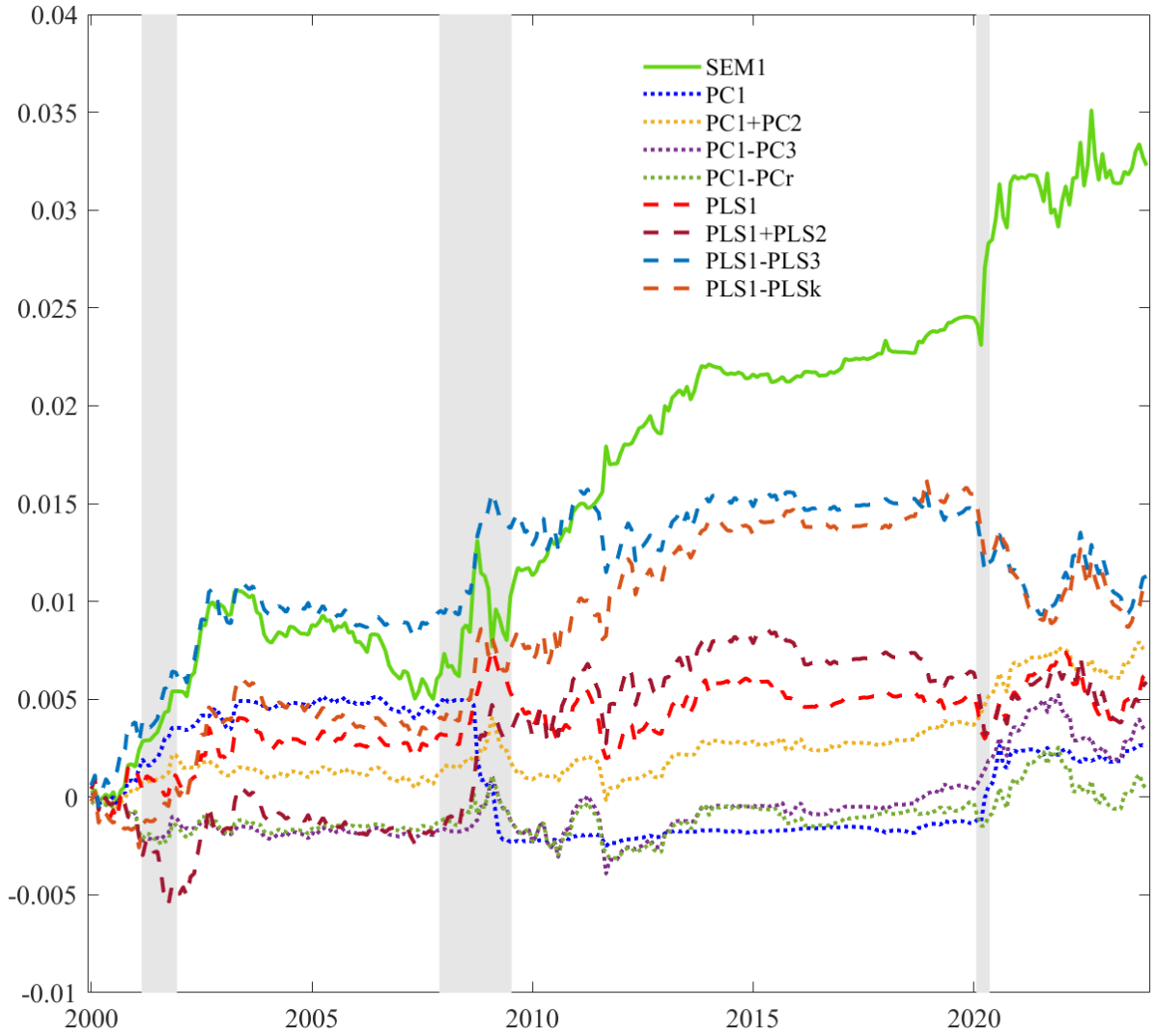


Figure 3: The cumulative difference in the squared prediction error (CDSPE)

Note: The green solid line depicts the difference between the cumulative squared prediction error (CSPE) for the historical average benchmark and the CSPE for the out-of-sample predictive regression forecast based on the recursively constructed SEM1 factor. The dotted line depicts the difference between the CSPE for the historical average benchmark and the CSPE for the out-of-sample predictive regression forecast based on the recursively constructed PC factors. The dashed line depicts the difference between the CSPE for the historical average benchmark and the CSPE for the out-of-sample predictive regression forecast based on the recursively constructed PLS factors. The factors and regression coefficients are estimated recursively based on information up to the period of forecast formation period t alone. The vertical bars correspond to NBER-dated recessions.

Appendix A: Predictors Dataset

Table A.1: Stock return predictors

Variable	Acronym	Variable	Acronym	Variable	Acronym
Macro and financial condition measures		Single factor from B/M cross-section	fbm	Financial regulation policy uncertainty	frpu
Dividend yield	d/y	Difference to the 52-week high of the Dow Jones index	dtoy	Trade policy uncertainty	tpu
Earning-price ratio	e/p	Difference to the lifetime high of the Dow Jones index	dtoat	Sovereign debt policy uncertainty	sdpu
Dividend-payout ratio	d/e	Return dispersion	rdsp	***Investor sentiment and subjected expectation measures***	
Sock variance	svar	Technical indicator	tchi	Dividend premium	pdnd
Book-to-market ratio	b/m	Average correlation of daily stock returns	avgcor	First-day returns of IPOs	ripo
Net equity expansion	ntis	Short stock interest	shtint	Number of IPOs	nipo
Long-term return	ltr	Analyst forecast disagreements	disag	Close-end fund discount rate	cefd
Term spread	tms	***Economic policy uncertainty measures***		Equity share in new issues	eqti
Default yield spread	dfy	Economic policy uncertainty	epu	Pofessional forecasters' expectation on nominal GDP	engdp
Default return spread	dfr	Monetary policy uncertainty	mpu	Pofessional forecasters' expectation on GDP price index	epgdp
Inflation	infl	Fiscal policy uncertainty	fpu	Pofessional forecasters' expectation on nominal corporate profits	ecprof
Aggregate illiquidity	lzrt	Tax policy uncertainty	tapu	Pofessional forecasters' expectation on industrial production index	eindprod
Output gap of industrial production	ogap	Government spending policy uncertainty	gspu	Pofessional forecasters' expectation on housing starts	ehousing
Oil price changes	wtexas	Health care policy uncertainty	hcpu	Pofessional forecasters' expectation on real GDP	ergdp
New orders to shipments of durable goods	ndrbl	National security policy uncertainty	nspu	Pofessional forecasters' expectation on recession probability	erecess
Average skewness	skvw	Entitlement programs policy uncertainty	eppu	Pofessional forecasters' expectation on unemployment rate	eunemp
Tail risk from cross-section	tail	Regulation policy uncertainty	rpu		