

Article

# Using the Entire Yield Curve in Forecasting Output and Inflation

Eric Hillebrand <sup>1</sup>, Huiyu Huang <sup>2</sup>, Tae-Hwy Lee <sup>3,\*</sup> and Canlin Li <sup>4</sup><sup>1</sup> Aarhus University and CREATES, Fuglesangs Allé 4, 8210 Aarhus V, Denmark; ehillebrand@creates.au.dk<sup>2</sup> ICBC Credit Suisse Asset Management, Beijing 100033, China; huang.huiyu@icbccs.com.cn<sup>3</sup> Department of Economics, University of California, Riverside, CA 92521, USA<sup>4</sup> Monetary and Financial Market Analysis Section, Division of Monetary Affairs, Federal Reserve Board, Washington, DC 20551, USA; canlin.li@frb.gov

\* Correspondence: taelee@ucr.edu

Received: 17 June 2018; Accepted: 21 August 2018; Published: 29 August 2018



**Abstract:** In forecasting a variable (forecast target) using many predictors, a factor model with principal components (PC) is often used. When the predictors are the yield curve (a set of many yields), the Nelson–Siegel (NS) factor model is used in place of the PC factors. These PC or NS factors are combining information (CI) in the predictors (yields). However, these CI factors are not “supervised” for a specific forecast target in that they are constructed by using only the predictors but not using a particular forecast target. In order to “supervise” factors for a forecast target, we follow Chan et al. (1999) and Stock and Watson (2004) to compute PC or NS factors of many forecasts (not of the predictors), with each of the many forecasts being computed using one predictor at a time. These PC or NS factors of forecasts are combining forecasts (CF). The CF factors are supervised for a specific forecast target. We demonstrate the advantage of the supervised CF factor models over the unsupervised CI factor models via simple numerical examples and Monte Carlo simulation. In out-of-sample forecasting of monthly US output growth and inflation, it is found that the CF factor models outperform the CI factor models especially at longer forecast horizons.

**Keywords:** level, slope, and curvature of the yield curve; Nelson–Siegel factors; supervised factor models; combining forecasts; principal components

**JEL Classification:** C5; E4; G1

## 1. Introduction

The predictive power of the yield curve for macroeconomic variables has been documented in the literature for a long time. Many different points on the yield curve have been used and various methodologies have been examined. For example, [Stock and Watson \(1989\)](#) find that two interest rate spreads, the difference between the six-month commercial paper rate and the six-month Treasury bill rate, and the difference between the ten-year and one-year Treasury bond rates, are good predictors of real activity, thus contributing to their index of leading indicators. [Bernanke \(1990\)](#), [Friedman and Kuttner \(1993\)](#), [Estrella and Hardouvelis \(1991\)](#), and [Kozicki \(1997\)](#), among many others, have investigated a variety of yields and yield spreads individually on their ability to forecast macroeconomic variables. [Hamilton and Kim \(2002\)](#) as well as [Diebold et al. \(2005\)](#) provide a brief summary of this line of research and the link between the yield curve and macroeconomic variables.

Various macroeconomic models for exploring the yield curve information for real activity prediction are proposed. [Ang and Piazzesi \(2003\)](#) and [Piazzesi \(2005\)](#) study the role of macroeconomic variables in an arbitrage-free affine yield curve model. [Estrella \(2005\)](#) constructs an analytical rational

expectations model to investigate the reasons for the success of the slope of the yield curve (the spread between long-term and short-term government bond rates) in predicting real economic activity and inflation. The model in [Ang et al. \(2006\)](#), Piazzesi and Wei is an arbitrage-free dynamic model (using lags of GDP growth and yields as regressors) that characterizes expectations of GDP growth. [Rudebusch and Wu \(2008\)](#) provide an example of a macro-finance specification that employs more macroeconomic structure and includes both rational expectations and inertial elements.

[Stock and Watson \(1999, 2002\)](#) investigate forecasts of output growth and inflation using over a hundred of economic indicators, including many interest rates and yield spreads. [Stock and Watson \(2002, 2012\)](#) advocate methods that aim at solving the large- $N$  predictor problem, particularly those using principal components (PC). [Ang et al. \(2006\)](#) suggest the use of the short rate, the five-year to three-month yield spread, and lagged GDP growth in forecasting GDP growth out-of-sample. The choice of these two yield curve characteristics, as they argue, is because they have almost one-to-one correspondence with the first two principal components of the short rate and five yield spreads that account for 99.7% of quarterly yield curve variation.

Alternatively to the PC factor approach on the large- $N$  predictor information set, [Diebold and Li \(2006\)](#) propose the [Nelson and Siegel \(1987\)](#) (NS) factors for the large- $N$  yields. They use a modified three-factor NS model to capture the dynamics of the yield curve and show that the three NS factors may be interpreted as level, slope, and curvature. [Diebold et al. \(2006\)](#) examine the correlations between NS yield factors and macroeconomic variables. They find that the level factor is highly correlated with inflation and that the slope factor is highly correlated with real activity. For more on the yield curve background and the three characteristics of the yield curve, see [Litterman and Scheinkman \(1991\)](#) and [Diebold and Li \(2006\)](#).

In this paper, we utilize the yield curve information for prediction of macro-economic variables. Using a large number of yield curve points with different maturities yields a large- $N$  problem in the predictive regression. The PC factors or the NS factors of the yield curve may be used to reduce the large dimension of the predictors. However, the PC and NS factors of the yield curve are not supervised for a specific variable to forecast. These factors simply combine information (CI) of many predictors (yields) without having to look at a forecast target. Hence, the conventional CI factor models (using factors of the predictors) are unsupervised for any forecast target.

Our goal in this paper is to consider factor models where the factors are computed with a particular forecast target in mind. Specifically, we consider the PC or NS factors of forecasts (not of predictors), with each of the forecasts formed using one predictor at a time. (It could be generalized to make each forecast from using more than one predictor, e.g., a subset of the  $N$  predictors, in which case there can be as many as  $2^N$  forecasts to combine.) These factors will combine the forecasts (CF). The PC factors of forecasts are combined forecasts using the combining weights that solves a singular value problem for a set of forecasts, while the NS factors of forecasts are combined forecasts using the combining weights obtained from orthogonal polynomials that emulate the shape of a yield curve (in level, slope, and curvature). The PC or NS factors of the many forecasts are supervised for a forecasting target. The main idea of the CF-factor model is to focus on the space spanned by *forecasts* rather than the space spanned by *predictors*. The factorization of forecasts (CF-factor model) can substantially improve forecasting performance compared to the factorization of predictors (CI-factor model). This is because the CF-factor model takes the forecast target into the factorization, while the conventional CI-factor model is blind to the forecast target because the factorization uses only information on predictors.

For both CI and CF schemes, the NS factor model can be relevant only when the yield curve is used as predictors while the PC factor model can be used in general. The NS factors are specific to the yield curve factors such as level, slope, and curvature factors. When the predictors are from the points on the yield curve, the NS factor models proposed here is nearly the same as the PC factors. Given the similarity of NS and PC and the generality of PC, we begin the paper with the PC models to understand the mechanism of the supervision in CF-factor models. We demonstrate how the supervised CF factor models outperform the unsupervised CI factor model, under the presence of

many predictors (50 points on the yield curve at each time). The empirical work shows that there are potentially big gains in the CF-factor models. In out-of-sample forecasting of U.S. monthly output growth and inflation, it is found that the CF factor models (CF-NS and CF-PC) are substantially better than the conventional CI factors models (CI-NS and CI-PC). The advantage of supervised factors is even greater for longer forecast horizons.

The paper is organized as follows: in Section 2, we describe the CI and CF frameworks and principal component approaches for their estimation, present theoretical results about supervision, and an example to provide intuition. Section 3 provides simulations of supervision under different noise, predictor correlation, and predictor persistence conditions. In Section 4, we introduce the NS component approaches for the CI and CF frameworks. In Section 5, we show the out-of-sample performance of the proposed methods in forecasting U.S. monthly output growth and inflation. Section 6 presents the conclusions.

## 2. Supervising Factors

### 2.1. Factor Models

Let  $y_{t+h}$  denote the variable to be forecast (output growth or inflation) using yield curve information stamped at time  $t$ , where  $h$  denotes the forecast horizon. The predictor vector  $x_t$  contains information about the yield curve at various maturities:  $x_t := (x_{1t}, x_{2t}, \dots, x_{Nt})'$ , where  $x_{it} := x_t(\tau_i)$  denotes the yield at time  $t$  with maturity  $\tau_i$  ( $i = 1, 2, \dots, N$ ).

Consider the CI model when  $N$  is large

$$y_{t+h} = (1 \ x_t')\alpha + \varepsilon_{t+h}, \quad (t = 1, 2, \dots, T) \quad (1)$$

for which the forecast at time  $T$  is

$$\hat{y}_{T+h}^{\text{CI-OLS}} = (1 \ x_T')\hat{\alpha}, \quad (2)$$

with  $\hat{\alpha}$  estimated by OLS using the information up to time  $T$ . A problem is that here the mean-squared forecast error (MSFE) is of order  $O\left(\frac{N}{T}\right)$  increasing with  $N$ .<sup>1</sup> A solution to this problem is to reduce the dimension either by selecting a subset of the  $N$  predictors, e.g., by Lasso type regression (Tibshirani 1996) or by using factor models of, e.g., Stock and Watson (2002). In this paper, we focus on using the factor model rather than selecting a subset of the  $N$  predictors.<sup>2</sup>

#### 2.1.1. CI-Factor Model

The conventional factor model is the CI factor model for  $x_t$  of the form

$$x_t = \Lambda_{\text{CI}} f_{\text{CI},t} + v_{\text{CI},t}, \quad (t = 1, \dots, T), \quad (3)$$

where  $\Lambda_{\text{CI}}$  is  $N \times k_{\text{CI}}$  and  $f_{\text{CI},t}$  is  $k_{\text{CI}} \times 1$ . The estimated factor loadings  $\hat{\Lambda}_{\text{CI}}$  are obtained either by following Stock and Watson (2002) and Bai (2003), or by following Nelson and Siegel (1987) and Diebold and Li (2006). The latter approach is discussed in Section 4. The factors are then estimated by

$$\hat{f}_{\text{CI},t} = \hat{\Lambda}_{\text{CI}}' x_t. \quad (4)$$

As this model computes the factors from all  $N$  predictors of  $x_t$  directly, it will be called "CI-factor". The forecast  $\hat{y}_{T+h} = (1 \ \hat{f}_{\text{CI},T}')\hat{\alpha}_{\text{CI}}$  can be formed using  $\hat{\alpha}_{\text{CI}}$  estimated at time  $T$  from the regression

$$y_t = (1 \ \hat{f}_{\text{CI},t-h}')\alpha_{\text{CI}} + u_{\text{CI},t}, \quad (t = h+1, \dots, T). \quad (5)$$

<sup>1</sup> This is explained in Bai and Ng 2008; Huang and Lee 2010; Stock and Watson 2002.

<sup>2</sup> Bai and Ng (2008) consider CI factor models with a selected subset (targeted predictors).

In matrix form, we write the factor model (3) and (5) for the vector of forecast target observations  $y$  and for the  $T \times N$  matrix of predictors  $X$  as follows:<sup>3</sup>

$$X = F_{CI}\Lambda'_{CI} + v_{CI}, \quad (6)$$

$$y = F_{CI}\alpha_{CI} + u_{CI}, \quad (7)$$

where  $y$  is the  $T \times 1$  vector of observations,  $F_{CI}$  is a  $T \times k_{CI}$  matrix of factors,  $\Lambda_{CI}$  is an  $N \times k_{CI}$  matrix of factor loadings,  $\alpha_{CI}$  is a  $k_{CI} \times 1$  parameter vector,  $v_{CI}$  is a  $T \times N$  random matrix, and  $u_{CI}$  is a  $T \times 1$  vector of random errors.

**Remark 1. (No supervision in CI-factor model):** Consider the joint density of  $(y_{t+h}, x_t)$

$$D(y_{t+h}, x_t; \theta) = D_1(y_{t+h}|x_t; \theta)D_2(x_t; \theta), \quad (8)$$

where  $D_1$  is the conditional density of  $y_{t+h}$  given  $x_t$ , and  $D_2$  is the marginal density of  $x_t$ . The CI-factor model assumes a situation where the joint density operates a “cut” in the terminology of [Barndorff-Nielsen \(1978\)](#) and [Engle et al. \(1983\)](#), such that

$$D(y_{t+h}, x_t; \theta) = D_1(y_{t+h}|x_t; \theta_1)D_2(x_t; \theta_2), \quad (9)$$

where  $\theta = (\theta_1 \theta_2)'$ , and  $\theta_1 = \alpha$ ,  $\theta_2 = (F, \Lambda)'$  are “variation-free”. Under this situation, the forecasting equation in (5) is obtained from the conditional model  $D_1$  and the factor equation in (3) is solely obtained from the marginal model  $D_2$  of the predictors. The computation of the factors is entirely from the marginal model  $D_2$  that is blind to the forecast target  $y_{t+h}$ .

While the CI factor analysis of a large predictor matrix  $X$  solves the dimensionality problem, it computes the factors using information in  $X$  only, without accounting for the variable  $y$  to be forecast, and therefore the factors are not supervised for the forecast target. Our goal in this paper is to improve this approach by accounting for the forecast target in the computation of the factors. The procedure will be called *supervision*.

There are some attempts in the literature to supervise factor computation for a given forecast target. For example, [Bair et al. \(2006\)](#) and [Bai and Ng \(2008\)](#) consider factors of selected predictors that are informative for a specified forecast target; [Zou et al. \(2006\)](#) consider sparse loadings of principal components; [De Jong \(1993\)](#) and [Groen and Kapetanios \(2016\)](#) consider partial least squares regression; [De Jong and Kiers \(1992\)](#) consider principal covariate regression; [Armah and Swanson \(2010\)](#) select variables for factor proxies that have the maximum predictive power for the variable being forecast; and some weighted principal components have been used to downweight noisier series.

In this paper, we consider the CF-factor model that computes factors from forecasts rather than from predictors. This approach has been proposed in [Chan et al. \(1999\)](#) and in [Stock and Watson \(2004\)](#), there labeled “principal component forecast combination”. We will refer to this approach as CF-PC (combining forecasts principal components). The details are as follows.

---

<sup>3</sup> The suppressed time stamp of  $y$  and  $X$  captures the  $h$ -lag relation for the forecast horizon and we treat the data centered so that we do not include a constant term explicitly in the regression for notational simplicity.

### 2.1.2. CF-Factor Model

The forecasts from a CF-factor model are computed in two steps. The first step is to estimate the factors of the individual forecasts. Let the individual forecasts be formed by regressing the forecast target  $y_{t+h}$  using the  $i$ th individual predictor  $x_{it}$ :

$$\hat{y}_{T+h}^{(i)} := a_{i,T} + b_{i,T}x_{iT} \quad (i = 1, 2, \dots, N). \quad (10)$$

Stack the  $N$  individual forecasts into a vector  $\hat{\mathbf{y}}_{t+h} := (\hat{y}_{t+h}^{(1)}, \hat{y}_{t+h}^{(2)}, \dots, \hat{y}_{t+h}^{(N)})'$  and consider a factor model of  $\hat{\mathbf{y}}_{t+h}$ :

$$\hat{\mathbf{y}}_{t+h} = \Lambda_{CF}f_{CF,t+h} + v_{CF,t+h}. \quad (11)$$

The CF-factor is estimated from

$$\hat{f}_{CF,t+h} := \hat{\Lambda}'_{CF}\hat{\mathbf{y}}_{t+h}. \quad (12)$$

The second step is to estimate the forecasting equation (for which the estimated CF-factors from the first step are used as regressors)<sup>4</sup>

$$y_{t+h} = \hat{f}'_{CF,t+h}\alpha_{CF} + u_{CF,t+h}. \quad (13)$$

Then, the CF-factor forecast at time  $T$  is

$$\hat{y}_{T+h}^{CF} = \hat{f}'_{CF,T+h}\hat{\alpha}_{CF}, \quad (14)$$

where  $\hat{\alpha}_{CF}$  is estimated. See (Chan et al. 1999; Huang and Lee 2010; Stock and Watson 2004).

To write the CF-factor model in matrix form, we assume for notational simplicity that the data has been centered so that we do not include a constant term. We regress  $y$  on the columns  $x_i$  of  $X$ ,  $i = 1, \dots, N$ , one at a time, and write the fitted values in (10) as

$$\hat{y}^{(i)} = x_i(x_i'x_i)^{-1}x_i'y =: x_ib_i. \quad (15)$$

Collect the fitted values in the matrix

$$\hat{Y} = [\hat{y}^{(1)} \hat{y}^{(2)} \dots \hat{y}^{(N)}] := XB \in \mathbb{R}^{T \times N}, \quad (16)$$

where  $B = \text{diag}(b_1, \dots, b_N) \in \mathbb{R}^{N \times N}$  is a diagonal matrix containing the regression coefficients. We call  $B$  the *supervision matrix*. Then, the CF-factor model is

$$\hat{Y} = F_{CF}\Lambda'_{CF} + v_{CF}, \quad (17)$$

$$y = F_{CF}\alpha_{CF} + u_{CF}, \quad (18)$$

where  $F_{CF}$  is a  $T \times k_{CF}$  matrix of factors of  $\hat{Y} = XB$ ,  $\Lambda_{CF}$  is an  $N \times k_{CF}$  matrix of factor loadings,  $\alpha_{CF}$  is an  $k_{CF} \times 1$  parameter vector,  $v_{CF}$  is a  $T \times N$  random matrix, and  $u_{CF}$  is a  $T \times 1$  vector of random errors. In the rest of the paper, the subscripts CI and CF may be omitted for simplicity.

We use principal components (PC) as discussed in Stock and Watson (2002), Bai (2003), and Bai and Ng (2006). For the specific case of yield curve data, we use NS components as discussed in Nelson and Siegel (1987) and Diebold and Li (2006). We use both CF and CI approaches together with PC factors and NS factors. Our goal is to show that forecasts using supervised factor models (CF-PC and CF-NS) are better than forecasts from conventional unsupervised factor models (CI-PC and CI-NS).

<sup>4</sup> Given the dependent nature of macroeconomic and financial time series, the forecasting equation can be extended to allow the supervision to be based on the relation between  $y_t$  and some predictors after controlling for lagged dependent variables and to allow the dynamic factor structure, which we leave for future work.

We show analytically and in simulations how supervision works to improve factor computation with respect to a specified forecast target. In Section 5, we present empirical evidence.

**Remark 2. (Estimation of B):** The CF-factor model in (17) and (18) with  $B = I_N$  (identity matrix) is a special case when there is no supervision. In this case, the CF-factor model collapses to the CI-factor model. If  $B$  were consistently estimated by minimizing the forecast error loss, then the CF-factor model with the “optimal”  $B$  would outperform the CI-factor model. However, as the dimension of the supervision matrix  $B$  grows with  $N^2$ ,  $B$  is an “incidental parameter” matrix and can not be estimated consistently. See [Neyman and Scott \(1948\)](#) and [Lancaster \(2000\)](#). Any estimation error in  $B$  translates into forecast error in the CF-factor model. Whether there is any virtue in considering Bayesian methods of estimating  $B$ , while still avoiding this problem, is left for future research. Instead, in this paper, we circumvent this difficulty by imposing that  $B = \text{diag}(b_1, \dots, b_N)$  be a diagonal matrix and by estimating the diagonal elements  $b_i$ 's from the ordinary least squares regression in (10) or (15) with one predictor  $x_i$  at a time. The supervision matrix  $B$  can be non-diagonal in general. As imposing the diagonality on  $B$  may be restrictive, it would be an interesting empirical question to examine if the CF-factor forecast with this restriction and the estimation strategy of  $B$  can still outperform the CI-factor forecast with  $B = I_N$ . Our empirical results in Section 5 (Table 1) support this simple estimation strategy for the diagonal matrix  $B$ , in favor of the CF-factor model.

**Remark 3. (Combining forecasts with many predictors):** It is generally believed that it is difficult to estimate the forecast combination weights when  $N$  is large. Therefore, the equal weights  $\left(\frac{1}{N}\right)$  have been widely used instead of estimating weights.<sup>5</sup> It is often found in the literature that equally-weighted combined forecasts are often the best. [Stock and Watson \(2004\)](#) call this the “forecast combination puzzle”. See also [Timmermann \(2006\)](#). [Smith and Wallis \(2009\)](#) explore a possible explanation of the forecast combination puzzle and conclude that it is due to estimation error of the combining weights.

Now, we note that, in the CF-factor model described above, we can consistently estimate the combining weights. From the CF-factor forecast (14) and the estimated factor (12),

$$\hat{y}_{T+h} = \hat{f}'_{CF,T+h} \hat{\alpha}_{CF} = (\hat{y}'_{T+h} \hat{\Lambda}_{CF}) \hat{\alpha}_{CF} := \hat{y}'_{T+h} \hat{w}, \quad (19)$$

where

$$\hat{w} := \hat{\Lambda}_{CF} \hat{\alpha}_{CF} \quad (20)$$

is estimated consistently as long as  $\hat{\Lambda}_{CF}$  and  $\hat{\alpha}_{CF}$  are estimated consistently.

## 2.2. Singular Value Decomposition

In this section, we formalize the concept of supervision and explain how it improves factor extraction. We compare the two different approaches CI-PC (Combining Information—Principal Components) and CF-PC (Combining Forecasts—Principal Components) in a linear forecast problem of the time series  $y$  given predictor data  $X$ . We explain the advantage of the CF-PC approach over CI-PC in Section 2.3 and give some examples in Section 2.4. We explore the advantage of supervision in simulations in Section 3.2. As an alternative to PC factors, we propose the use of NS factors in Section 4.

**Principal components of predictors X (CI-PC):** Let  $X \in \mathbb{R}^{T \times N}$  be a matrix of regressors and let

$$X = R\Sigma W' \in \mathbb{R}^{T \times N} \quad (21)$$

<sup>5</sup> An exception is [Wright \(2009\)](#), who uses Bayesian model averaging (BMA) for pseudo out-of-sample prediction of U.S. inflation, and finds that it generally gives more accurate forecasts than simple equal-weighted averaging. He uses  $N = 107$  predictors.

be the singular value decomposition of  $X$ , with  $\Sigma \in \mathbb{R}^{T \times N}$  diagonal rectangular, that is, diagonal square matrix padded with zero rows below the square if  $\min(T, N) = N$  or padded with zero columns next to the square if  $\min(T, N) = T$ ,  $R \in \mathbb{R}^{T \times T}$ , and  $W \in \mathbb{R}^{N \times N}$  is unitary. Write

$$X'X = W\Sigma'R'RW' = W\Sigma'\Sigma W', \tag{22}$$

where  $\Sigma'\Sigma := \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$  is diagonal and square. Therefore,  $W$  contains the eigenvectors of  $X'X$ . For a matrix  $A \in \mathbb{R}^{T \times N}$ , denote by  $A_k \in \mathbb{R}^{T \times k}$  the matrix consisting of the first  $k \leq N$  columns of  $A$ . Then,  $W_k$  is the matrix containing the singular vectors corresponding to the  $k = k_{\text{CI}}$  largest singular values  $(\sigma_1, \dots, \sigma_k)$ . The first  $k$  principal components are given by

$$F_{\text{CI}} := XW_k = R\Sigma W'W_k = R\Sigma \begin{bmatrix} I_k \\ \mathbf{0} \end{bmatrix} = R\Sigma_k = R_k\Sigma_{kk}, \tag{23}$$

where  $I_k$  is the  $k \times k$  identity matrix,  $\mathbf{0}$  is an  $(N - k) \times k$  matrix of zeros, and  $\Sigma_{kk}$  is the  $k \times k$  upper-left diagonal block of  $\Sigma$ . Note that the first  $k$  principal components  $F_{\text{CI}}$  of  $X$  are constant multiples of columns of  $R_k$  as  $\Sigma_{kk}$  is diagonal. The projection (forecast) of  $y$  onto  $F_{\text{CI}}$  is given by

$$\begin{aligned} \hat{y}_{\text{CI-PC}} &:= F_{\text{CI}}(F_{\text{CI}}'F_{\text{CI}})^{-1}F_{\text{CI}}'y = XW_k(W_k'X'XW_k)^{-1}W_k'X'y \\ &= R_k\Sigma_{kk}(\Sigma_{kk}'R_k'R_k\Sigma_{kk})^{-1}\Sigma_{kk}'R_k'y = R_k(R_k'R_k)^{-1}R_k'y = R_kR_k'y, \end{aligned} \tag{24}$$

as  $R_k'R_k = I_k$ . Therefore, the CI forecast,  $\hat{y}_{\text{CI-PC}}$ , is the projection of  $y$  onto  $R_k$ . The CI forecast error and the CI sum of squared error (SSE) are

$$y - \hat{y}_{\text{CI-PC}} = y - R_kR_k'y = (I_T - R_kR_k')y, \tag{25}$$

$$SSE_{\text{CI-PC}} = \|y - \hat{y}_{\text{CI-PC}}\|^2 = y'(I_T - R_kR_k')y, \tag{26}$$

as  $(I_T - R_kR_k')$  is symmetric idempotent.

Bai (2003) shows that, under general assumptions on the factor and error structure,  $F_{\text{CI}}$  is a consistent and asymptotically normal estimator of  $F_{\text{CI}}H$ , where  $H$  is an invertible  $k \times k$  matrix.<sup>6</sup> This identification problem is also clear from Equation (24), and it conveniently allows us to identify the principal components  $F_{\text{CI}} = R_k\Sigma_{kk}$  as  $F_{\text{CI}} = R_k$  since  $\Sigma_{kk}$  is diagonal. The principal components are scalar multiples of the first  $k$  columns of  $R$ . Bai's result shows that principal components can be estimated consistently only up to linear combinations. Bai and Ng (2006) show that the parameter vector  $\alpha$  in the forecast equation can be estimated consistently for  $\alpha'H^{-1}$  with an asymptotically normal distribution.

**Principal components of forecasts  $\hat{Y}$  (CF-PC):** To generate forecasts in a CF-factor scheme, we regress  $y$  on the columns  $x_i$  of  $X$ ,  $i = 1, \dots, N$ , one at a time, and calculate the fitted values of (15). Collect the fitted values in the matrix as in (16), with  $B = \text{diag}(b_1, \dots, b_N)$  containing the regression coefficients in its diagonal. Compute the singular value decomposition of  $\hat{Y}$ :

$$\hat{Y} = S\Theta V', \tag{27}$$

---

<sup>6</sup> In order for the objects in Bai's (2003) analysis to converge, he introduces scaling such that the singular values are the eigenvalues of the matrix  $X'X/T$ . Then, the singular vectors are multiplied by  $\sqrt{T}$ . In our notation, the singular value decomposition becomes  $X = \sqrt{T}R \frac{\Sigma}{\sqrt{T}} W'$ .

with  $\Theta \in \mathbb{R}^{T \times N}$  is diagonal rectangular, and  $S \in \mathbb{R}^{T \times T}, V \in \mathbb{R}^{N \times N}$  unitary. Pick the first  $k = k_{CF}$  principal components of  $\hat{Y}$ ,

$$F_{CF} := \hat{Y}V_k = S\Theta V'V_k = S\Theta \begin{bmatrix} I_k \\ \mathbf{0} \end{bmatrix} = S\Theta_k = S_k\Theta_{kk}, \tag{28}$$

where  $V_k$  is the  $N \times k$  matrix of the singular vectors corresponding to the  $k$  largest singular values  $(\theta_1, \dots, \theta_k)$  and  $\Theta_{kk}$  is the  $k \times k$  upper-left diagonal block of  $\Theta$ . Again, we can identify the estimated  $k$  principal components of  $\hat{Y}$  with  $F_{CF} = S_k$ , where  $F_{CF}$  is the  $T \times k_{CF}$  matrix of factors of  $\hat{Y}$ . The projection (forecast) of  $y$  onto  $F_{CF}$  is given by:

$$\begin{aligned} \hat{y}_{CF-PC} &:= F_{CF}(F'_{CF}F_{CF})^{-1}F'_{CF}y = \hat{Y}V_k(V'_k\hat{Y}'\hat{Y}V_k)^{-1}V'_k\hat{Y}'y \\ &= S_k\Theta_{kk}(\Theta'_{kk}S'_kS_k\Theta_{kk})^{-1}\Theta'_{kk}S'_ky = S_k(S'_kS_k)^{-1}S'_ky = S_kS'_ky \end{aligned} \tag{29}$$

as  $S'_kS_k = I_k$ . The CF forecast,  $\hat{y}_{CF-PC}$ , is the projection of  $y$  onto  $S_k$ . The CF forecast error and the CF SSE are

$$y - \hat{y}_{CF-PC} = y - S_kS'_ky = (I_T - S_kS'_k)y, \tag{30}$$

$$SSE_{CF-PC} = \|y - \hat{y}_{CF-PC}\|^2 = y'(I_T - S_kS'_k)y, \tag{31}$$

as  $(I_T - S_kS'_k)$  is symmetric idempotent.

### 2.3. Supervision

In this sub-section, we explain the advantage of CF-PC over CI-PC in factor computation. We call the advantage “supervision”, which is defined as follows:

**Definition 1.** (Supervision). *The advantage of CF-PC over CI-PC, called supervision, is the selection of principal components according to their contribution to variation in  $y$ , as opposed to selection of principal components according to their contribution to variation in the columns of  $X$ . This is achieved by selecting principal components from a matrix of forecasts of  $y$ .*

We use the following measures of supervision of CF-PC in comparison with CI-PC.

**Definition 2.** (Absolute Supervision). *Absolute supervision is the difference of the sums of squared errors (SSE) of CI-PC and CF-PC:*

$$s_{abs}(X, y, k_{CI}, k_{CF}) := \|y - \hat{y}_{CI-PC}\|^2 - \|y - \hat{y}_{CF-PC}\|^2 = y'(S_{k_{CF}}S'_{k_{CF}} - R_{k_{CI}}R'_{k_{CI}})y. \tag{32}$$

**Definition 3.** (Relative Supervision). *Relative supervision is the ratio of the sums of squared errors of CI-PC over CF-PC:*

$$s_{rel}(X, y, k_{CI}, k_{CF}) := \frac{\|y - \hat{y}_{CI-PC}\|^2}{\|y - \hat{y}_{CF-PC}\|^2} = \frac{y'(I_T - R_{k_{CI}}R'_{k_{CI}})y}{y'(I_T - S_{k_{CF}}S'_{k_{CF}})y}. \tag{33}$$

**Remark 4.** *When  $k_{CI} = k_{CF} = N$ , there is no room for supervision*

$$s_{abs}(X, y, N, N) = y'(SS' - RR')y = y'(I_T - I_T)y = 0 \tag{34}$$

*because  $SS' = RR' = I_T$ . Relative supervision is defined only for  $k_{CF} < N$ .*

For the sake of simplifying the notation and presentation, we consider the same number of factors in CI and CF factor models with  $k_{CI} = k_{CF} = k$  for the rest of the paper.



**Remark 5.**  $S_k$  is a block of a basis change matrix that in the expression  $y'S_k$  returns the first  $k$  coordinates of  $y$  with respect to the new basis. This new basis is the one with respect to which the mapping  $\hat{Y}\hat{Y}' = XBBX' = S\Theta\Theta'S'$  becomes diagonal, with singular values in descending order such that the first  $k$  columns of  $S$  correspond to the  $k$  largest singular values. Therefore,  $y'S_kS_k'y$  is the sum of the squares of these coordinates. Broadly speaking, the  $S_k$  are the  $k$  largest components of  $y$  in the sense of  $\hat{Y}$  and its construction from the single regression coefficients. Thus,  $y'S_kS_k'y$  is the sum of the squares of the  $k$  coefficients in  $y$  that contributes most to the variation in the columns of  $\hat{Y}$ .

Analogously,  $R_k$  is a block of a basis change matrix that for  $y'R_k$  returns the first  $k$  coordinates of  $y$  with respect to the basis that diagonalizes the mapping  $XX' = R\Sigma\Sigma'R'$ . Therefore,  $y'R_kR_k'y$  is the sum of squares of the  $k$  coordinates of  $y$  selected according to their contribution to variation in the columns of  $X$ .

We emphasize the factors that explain most of the variation of the columns of  $X$ , i.e., the eigenvectors associated with the largest eigenvalues of  $XX'$ , which are selected in the principal component analysis of  $X$ , may have little to do with the factors that explain most of the variation of  $y$ , however. The relation between  $X$  and  $y$  in the data-generating process can, at worst, completely reverse the order of principal components in the columns of  $X$  and in  $y$ . We demonstrate this in the following Example 1.

#### 2.4. Example 1

In this subsection, we give a small example to facilitate intuition for the supervision mechanics of CF-PC. Example 1 illustrates how the supervision of factor computation defined in Definition 1 operates. In Example 2 in the next section, we add randomness to Example 1 to explore the effect of stochasticity in a well-understood problem.

Let

$$X = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 1/5 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{35}$$

with  $T = 6$  and  $N = 5$ . The singular value decomposition of  $X = R\Sigma W$  is

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{5} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{36}$$

Let

$$y = (1, 2, 3, 4, 5, 0)'. \tag{37}$$

Then, the diagonal matrix  $B$  that contains the coefficients of  $y$  w.r.t. each column of  $X$  is

$$B = \text{diag}(4, 9, 1, 25, 16), \tag{38}$$

and

$$\hat{Y} := XB = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{39}$$

The singular value decomposition of  $\hat{Y} = XB = S\Theta V$  is

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}. \tag{40}$$

We set  $k_{CI} = k_{CF} = k$  and compare CI-PC and CF-PC with the same number of principal components. Recall from (23) that  $F_{CI} = R\Sigma_k$  and from (28) that  $F_{CF} = S\Theta_k$ . The absolute supervision and relative supervision, defined in (32) and (33), are computed for each  $k$ :

	$s_{abs}(X, y, k_{CI}, k_{CF})$	$s_{rel}(X, y, k_{CI}, k_{CF})$
$k = 1$	24	1.8
$k = 2$	36	3.6
$k = 3$	36	8.2
$k = 4$	24	25.0
$k = 5$	0	N/A

See Appendix A for the calculation. The absolute supervision is all positive and the relative supervision is larger than 1 for all  $k < N$ .

As noted in Remarks 1 and 5, the relation between  $X$  and  $y$  is crucial. In this example, the magnitude of the components in  $y$  is reversed from the order in  $X$ . For  $X$ , the ordering of the columns of  $X$  with respect to the largest eigenvalues of  $XX'$  is  $\{3, 1, 2, 5, 4\}$ . For  $y$ , the ordering of the columns of  $X$  with respect to the largest eigenvalues of  $\hat{Y}\hat{Y}'$  is  $\{4, 5, 2, 1, 3\}$ . For example, consider the case  $k = 2$ , i.e., we choose two out of five factors in the principal component analysis. CI-PC, the analysis of  $X$ , will pick the columns 3 and 1 of  $X$ , that is, the vectors  $(1, 0, 0, 0, 0, 0)'$  and  $(0, 1/2, 0, 0, 0, 0)'$ . These correspond to the two largest singular values 1 and  $1/2$  of  $X$ . CF-PC, the analysis of  $\hat{Y}$ , will pick columns 4 and 5 of  $X$ , that is, the vectors  $(0, 0, 0, 0, 1/5, 0)'$  and  $(0, 0, 0, 1/4, 0, 0)'$ . These correspond to the two largest singular values 5 and 4 of  $\hat{Y}$ . The regression coefficients in  $B = \text{diag}(4, 9, 1, 25, 16)$  de-emphasize columns 3 and 1 of  $X$  and emphasize columns 4 and 5 of  $X$ .

### 3. Monte Carlo

There are several simplifications in the construction of Example 1, which we relax by the following extensions:

- (a) Adding randomness makes the estimation of the regression coefficients in  $B$  a statistical problem. The sampling errors influence the selection of the components of  $\hat{Y}$ .
- (b) Adding correlation among regressors (columns of  $X$ ) introduces correlation among individual forecasts (columns of  $\hat{Y}$ ), increasing the effect of sampling error in the selection of the components of  $\hat{Y}$ .
- (c) Increasing  $N$  to realistic magnitudes, in particular in the presence of highly correlated regressors, will increase estimation error in the principal components due to collinearity.

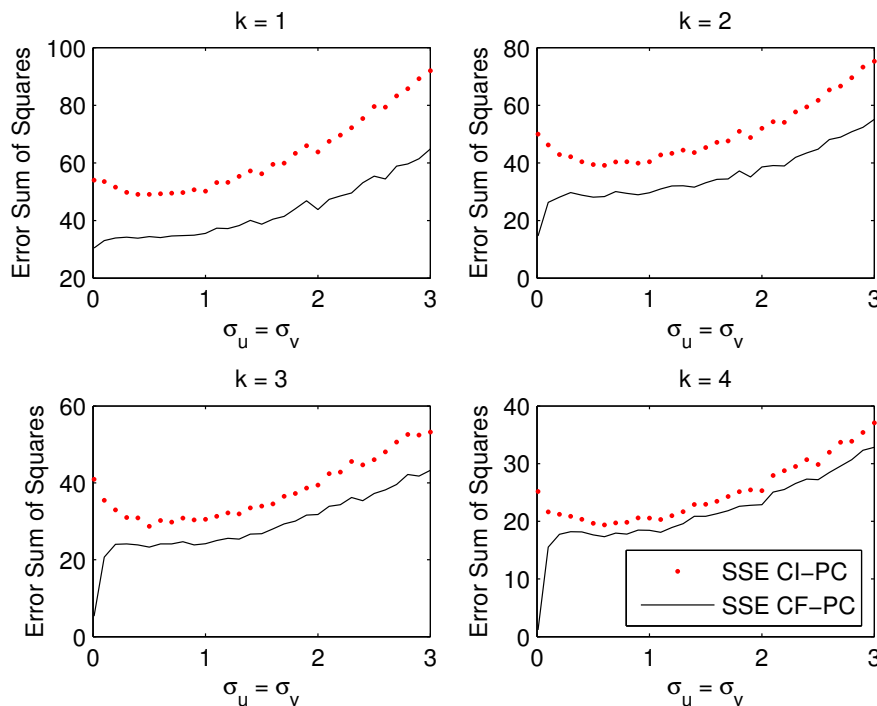
We address the first extension (a) in Example 2. All three extensions (a), (b), (c) are addressed in Example 3 of Section 3.2.

#### 3.1. Example 2

Consider adding some noise to  $X, y$  in Example 1. Let  $v$  be a  $T \times N$  matrix of independent random numbers, each entry distributed as  $N(0, \sigma_v^2)$ , and  $u$  be a vector of independent random numbers, each distributed as  $N(0, \sigma_u^2)$ . In this example, the new regressor matrix  $X$  is the sum of  $X$  in Example 1 and the noise term  $v$ , and the new  $y$  is the sum of  $y$  in Example 1 and the noise term  $u$ . For simplicity,

we set  $\sigma_v = \sigma_u$  in the simulations and let both range from 0.01 to 3. This covers a substantial range of randomness given the magnitude of the numbers in  $X$  and  $y$ . For each scenario of  $\sigma_v = \sigma_u$ , we generate 1000 random matrices  $v$  and random vectors  $u$  and calculate the Monte Carlo average of the sums of squared errors (SSE).

Figure 1 plots the Monte Carlo average of the SSEs for selection of  $k = 1$  to  $k = 4$  components. For standard deviations  $\sigma_v = \sigma_u$  close to zero, the sum of squared errors are as calculated in Example 1. As the noise increases, the advantage of CF over CI decreases but remains substantial, in particular for smaller numbers of principal components. For  $k = 5$  estimated components (not shown), the SSEs of CI-PC and CF-PC coincide because  $k = N$ .



**Figure 1.** For Example 2. Monte Carlo averages of the sum of squared errors (SSE) against a grid of standard deviations  $\sigma_u = \sigma_v$  ranging from 0.01 to 3 in factor and forecast equations, for a selection of  $k = 1$  to  $k = 4$  components. When the standard deviation is close to zero, the SSE are close to the ones reported in Example 1. With increasing noise, the advantage of CF over CI decreases but remains substantial, in particular for few components. For  $k = 5 = N$  (not shown), the SSE of CI-PC and CF-PC coincide, as shown in Remark 4.

### 3.2. Example 3

We consider the data-generating process (DGP)

$$X = F\Lambda' + v, \tag{41}$$

$$y = F\alpha + u, \tag{42}$$

where  $y$  is the  $T \times 1$  vector of observations,  $F$  is a  $T \times r$  matrix of factors,  $\Lambda$  is an  $N \times r$  matrix of factor loadings,  $\alpha$  is an  $r \times 1$  parameter vector,  $v$  is a  $T \times N$  random matrix, and  $u$  is a  $T \times 1$  vector of random errors. We set  $T = 200$ ,  $N = 50$  and consider  $r = 3$  data-generating factors.

Note that, under this DGP, the CI-PC model in Equations (6) and (7) is correctly specified if the correct number of factors is identified, i.e.,  $k_{CI} = r$ . Even under this DGP, however, an insufficient

number of factors,  $k_{CI} < r$ , can still result in an advantage of the CF-PC model over the CI-PC model. We will explore this question in this section.

**Factors and persistence:** For each run in the simulation, we generate the  $r$  factors in  $F$  as independent AR(1) processes with zero mean and a normally distributed error with mean zero and variance one:

$$F_{t,i} = \phi F_{t-1,i} + \varepsilon_{t,i}, \quad t = 2, \dots, T, \quad i = 1, \dots, r. \tag{43}$$

We consider a grid of 19 different AR(1) coefficients  $\phi$ , equidistant between 0 and 0.90. We consider  $r = 3$  data-generating factors and  $k \in \{1, 2, 3, 4\}$  estimated factors.

**Contemporaneous factor correlation:** Given a correlation coefficient  $\rho$  for adjacent regressors, the  $N \times r$  matrix  $\Lambda$  of factor loadings is obtained from the first  $r$  columns of an upper triangular matrix from a Cholesky decomposition of

$$\begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{N-1} \\ \rho & 1 & \rho & \dots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{N-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \dots & 1 \end{bmatrix}. \tag{44}$$

We consider a grid of 19 different values for  $\rho$ , equidistant between the points  $-0.998$  and  $0.998$ . In this setup, the 10th value is very close to  $\rho = 0$ . Then, the covariance matrix of the regressors is given by

$$\mathbb{E}X'X = \mathbb{E}[(\Lambda F' + v')(F\Lambda' + v)] = \Omega_F + \Omega_v, \tag{45}$$

where  $\Omega_F = \Lambda\Lambda'$  and  $\Omega_v = \mathbb{E}v'v$  is given by the identity matrix in our simulations. The relation  $\mathbb{E}F'F = I$  is due to the independence of the factors, but may be subject to substantial finite sample error, in particular for  $\phi$  close to one, for well-known reasons.

**Relation of  $X$  and  $y$ :** The  $r \times 1$  parameter vector  $\alpha$  is drawn randomly from a standard normal distribution for each run in the simulation. This allows  $\alpha$  to randomly shuffle which factors are important for  $y$ .

**Noise level:** We set  $\sigma_u = \sigma_v$  and let it range between 0.1 and 3 in steps of 0.1. We add the case of 0.01 that essentially corresponds to a deterministic factor model.

For a given number  $r = 3$  of data-generating factors, the simulation setup varies along the dimensions  $\phi$  (19 points),  $k$  (4 points),  $\rho$  (19 points),  $\sigma_u = \sigma_v$  (31 points). For every single scenario, we run 1000 simulations and calculate the SSEs of CI-PC and CF-PC, and the relative supervision  $s_{rel}(X, y, k, k)$ . Then, we take the Monte Carlo average of the SSEs and  $s_{rel}(X, y, k, k)$  over the 1000 simulations.<sup>7</sup>

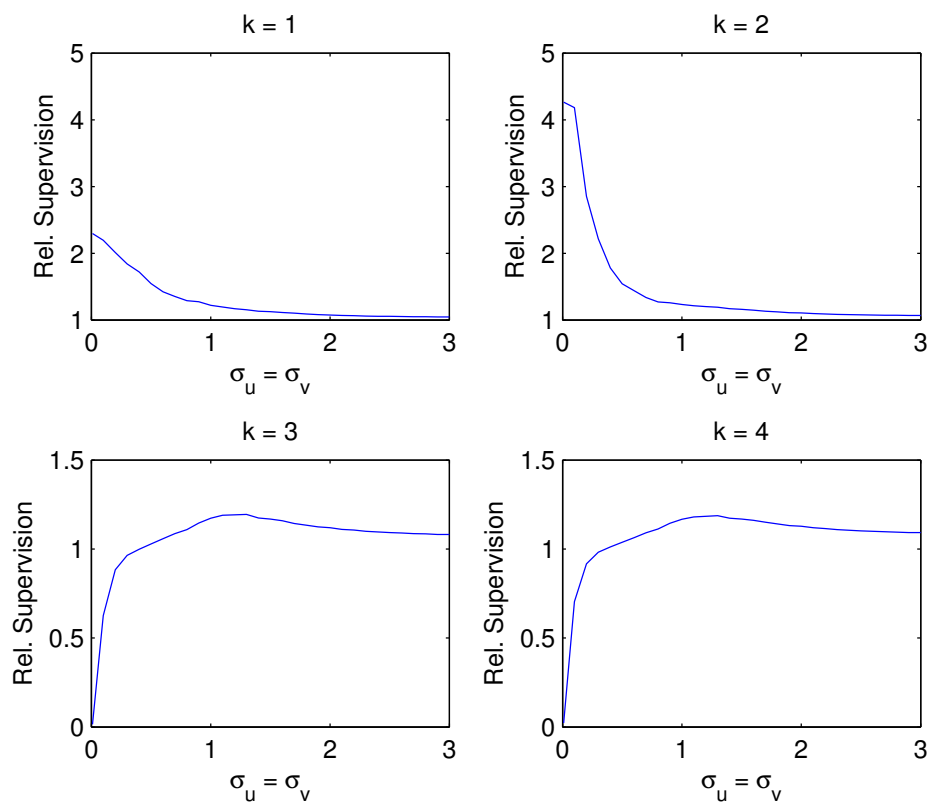
The Monte Carlo results are presented in Figures 2–4. Each figure contains four panels that plot the situation for  $k = 1, 2, 3, 4$  estimated number of factors. The main findings from the figures can be summarized as follows:

1. **Figure 2:** If the number of estimated factors  $k$  is below the true number  $r = 3$ , as shown in top panels, the supervision becomes smaller with increasing noise. If the correct number of factors or

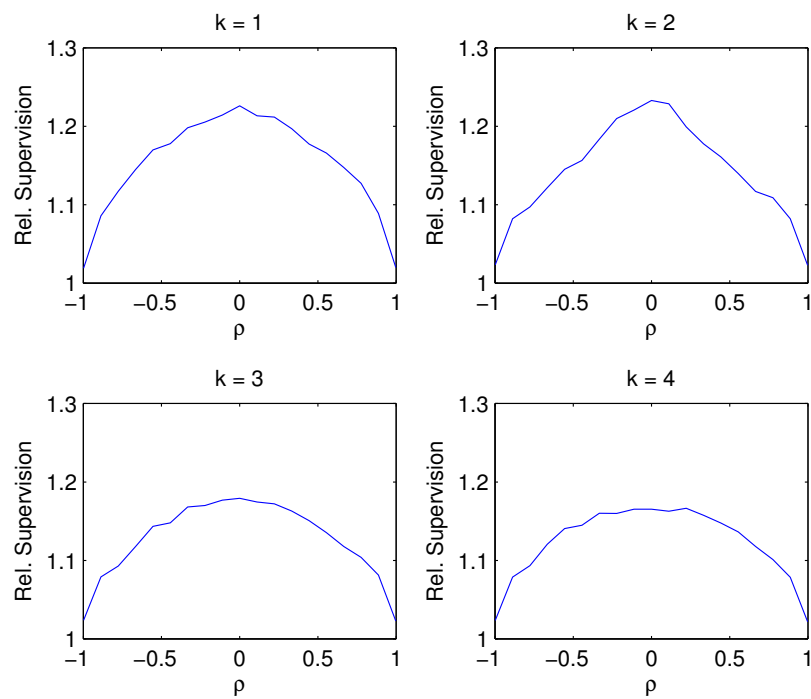
---

<sup>7</sup> In relation to the empirical application using the yield data in Section 5, we could have calibrated the simulation design to make the Monte Carlo more realistic for the empirical application in Section 5. Nevertheless, our Monte Carlo design covers wide ranges of the parameter values for the noise levels, correlation structures ( $\rho$  and  $\phi$ ) in the yield data. Figure 2 shows that the supervision is smaller with larger noise levels, which may be rather obvious intuitively. Figure 4 shows that the advantage of supervision when the factors are persistence, which depends on the number of factors  $k$  relative to the true number of factors  $r$ . Particularly interesting is Figure 3 which shows that the advantage of supervision is smaller when the contemporaneous correlation  $\rho$  between predictors is larger, which may be relevant for the yield data because the yields with different maturities may be moderately contemporaneously correlated. We thank a referee for pointing this out.

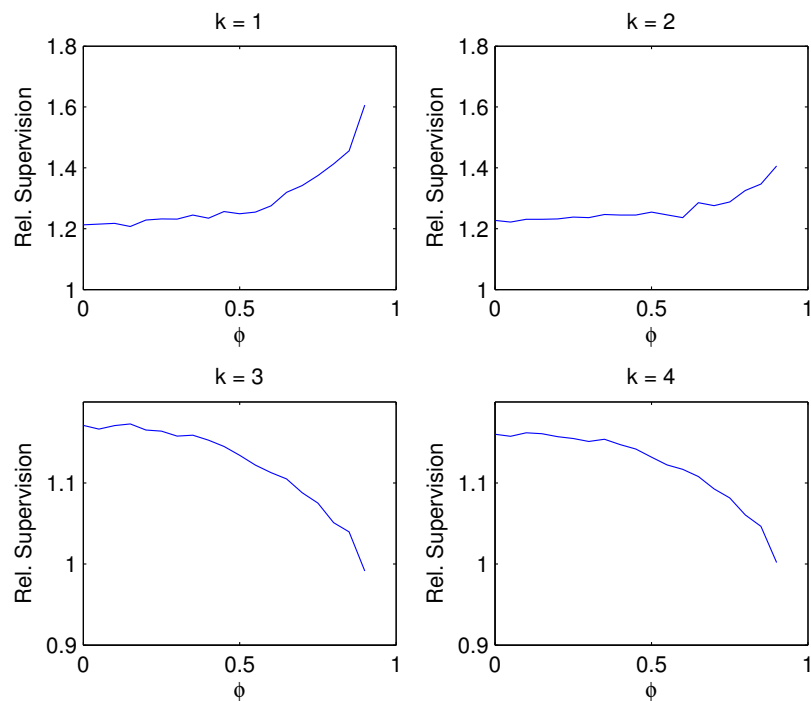
- more are estimated ( $k \geq r$ ), as in bottom panels, the advantage of supervision increases with the noise level  $\sigma_u = \sigma_v$ . Even in this case when the CI-PC is the correct model ( $k \geq r$ ), supervision becomes larger as the noise increases.
- Figure 3: The advantage of supervision is greatest when the contemporaneous correlation  $\rho$  between predictors is minimal. For almost perfect correlation, the advantage of supervision disappears. This is true regardless of whether the correct number of factors is estimated or not. Intuitively, for near-perfect factor correlation, the difference between those factors that explain variation in the columns of  $X$  and those that explain variation in  $\hat{Y}$  vanishes, and so supervision becomes meaningless.
  - Figure 4: If the correct number of factors or more are estimated ( $k \geq r$ ), the advantage of supervision decreases with factor persistence  $\phi$ . High persistence induces spurious contemporaneous correlation, and in this sense the situation is related to the result in No. 2. If the number of estimated factors is below the true number of factors ( $k < r$ ), however, the advantage of supervision increases with factor persistence.



**Figure 2.** Supervision dependent on noise. Relative supervision against a grid of standard deviations in factor and forecast equation  $\sigma_u = \sigma_v$ , ranging from 0.01 to 3, while the factor serial correlation is fixed at  $\phi = 0$  and the contemporaneous factor correlation is  $\rho = 0$ .



**Figure 3.** Supervision dependent on contemporaneous factor correlation  $\rho$ . Relative supervision against a grid of contemporaneous correlation coefficients  $\rho$  ranging from  $-0.998$  to  $0.998$ , while the factor serial correlation  $\phi$  is fixed at zero and the noise level is fixed at  $\sigma_u = \sigma_v = 1$ .



**Figure 4.** Supervision dependent on factor persistence  $\phi$ . Relative supervision against a grid of AR(1) coefficients  $\phi$  ranging from 0 to 0.9, while the noise level is fixed at  $\sigma_u = \sigma_v = 1$  and the contemporaneous regressor correlation is  $\rho = 0$ .

#### 4. Supervising Nelson–Siegel Factors

In the previous section, we have examined the factor model based on principal components. When the predictors are points on the yield curve, an alternative factor model can be constructed based on Nelson–Siegel (NS) components. We introduce two new factor models, CF-NS and CI-NS, by replacing principal components with NS components in CF-PC and CI-PC models. Like CI-PC, CI-NS is unsupervised. Like CF-PC, CF-NS is supervised for the particular forecast target of interest.

##### 4.1. Nelson–Siegel Components of the Yield Curve

As an alternative to using principal components in the factor model, one can apply the modified Nelson–Siegel (NS) three-factor framework of [Diebold and Li \(2006\)](#) to factorize the yield curve. [Nelson and Siegel \(1987\)](#) propose Laguerre polynomials  $L_n(z) = \frac{e^{-z}}{n!} \frac{d^n}{dz^n} (z^n e^{-z})$  with weight function  $w(z) = e^{-z}$  to model the instantaneous nominal forward rate (forward rate curve)

$$\begin{aligned} f_t(\tau) &= \beta_1 + (\beta_2 + \beta_3) \left( L_0(z) e^{-\theta\tau} \right) - \beta_3 \left( L_1(z) e^{-\theta\tau} \right) \\ &= \beta_1 + (\beta_2 + \beta_3) e^{-\theta\tau} - \beta_3 (1 - \theta\tau) e^{-\theta\tau} \\ &= \beta_1 + \beta_2 e^{-\theta\tau} + \beta_3 \theta\tau e^{-\theta\tau}, \end{aligned} \quad (46)$$

where  $z = \theta\tau$ ,  $L_0(z) = 1$ ,  $L_1(z) = 1 - \theta\tau$ , and  $\beta_j \in \mathbb{R}$  for all  $j$ . The decay parameter  $\theta$  may change over time, but we fixed  $\theta = 0.0609$  for all  $t$  following [Diebold and Li \(2006\)](#).<sup>8</sup>

Then, the continuously compounded zero-coupon nominal yield  $x_t(\tau)$  of the bond with maturity  $\tau$  months at time  $t$  is

$$x_t(\tau) = \frac{1}{\tau} \int_0^\tau f_t(s) ds = \beta_1 + \beta_2 \left( \frac{1 - e^{-\theta\tau}}{\theta\tau} \right) + \beta_3 \left( \frac{1 - e^{-\theta\tau}}{\theta\tau} - e^{-\theta\tau} \right). \quad (47)$$

Allowing the  $\beta_j$ 's to change over time and adding the approximation error  $v_{it}$ , we obtain the following approximate NS factor model for the yield curve for  $i = 1, \dots, N$ :

$$\begin{aligned} x_t(\tau_i) &= \beta_{1t} + \beta_{2t} \left( \frac{1 - e^{-\theta\tau_i}}{\theta\tau_i} \right) + \beta_{3t} \left( \frac{1 - e^{-\theta\tau_i}}{\theta\tau_i} - e^{-\theta\tau_i} \right) + v_{it} \\ &= \begin{bmatrix} 1 & \left( \frac{1 - e^{-\theta\tau_i}}{\theta\tau_i} \right) & \left( \frac{1 - e^{-\theta\tau_i}}{\theta\tau_i} - e^{-\theta\tau_i} \right) \end{bmatrix} \begin{bmatrix} \beta_{1t} \\ \beta_{2t} \\ \beta_{3t} \end{bmatrix} + v_{it} \\ &= \lambda'_i f_t + v_{it}, \end{aligned} \quad (48)$$

where  $f_t = (\beta_{1t}, \beta_{2t}, \beta_{3t})'$  are the three NS factors and  $\lambda'_i = \left[ 1 \left( \frac{1 - e^{-\theta\tau_i}}{\theta\tau_i} \right) \left( \frac{1 - e^{-\theta\tau_i}}{\theta\tau_i} - e^{-\theta\tau_i} \right) \right]$  are the factor loadings. Because  $x_t(\infty) = \beta_{1t}$ ,  $x_t(\infty) - x_t(0) = -\beta_{2t}$ , and  $[x_t(0) + x_t(\infty)] - 2x_t(\tau_m)$  with  $\tau_m = 24$  (say) is proportional to  $-\beta_{3t}$ , the three NS factors  $(\beta_{1t}, \beta_{2t}, \beta_{3t})'$  are associated with level, slope, and curvature of the yield curve.

<sup>8</sup> [Diebold and Li \(2006\)](#) show that fixing Nelson–Siegel decay parameter at  $\theta = 0.0609$  maximizes the curvature loading at the two-year bond maturity and allows better identifications of the three NS factors. They also show that allowing the  $\theta$  to be a free parameter does not improve the forecasting performance. Therefore, following their advice, we fix  $\theta = 0.0609$  and did not estimate it. A small  $\theta$  (for a slow decaying curve) fits the curve for long maturities better and a large  $\theta$  (for a fast decaying curve) fits the curve for short maturities better.

#### 4.2. CI-NS and CF-NS

##### 4.2.1. NS Components of Predictors X (CI-NS)

We have  $N$  predictors of yields  $x_t = (x_{1t}, x_{2t}, \dots, x_{Nt})'$  where  $x_{it} = x_t(\tau_i)$  denotes the yield to maturity  $\tau_i$  months at time  $t$ , ( $i = 1, 2, \dots, N$ ). Stacking  $x_{it}$  for  $i = 1, 2, \dots, N$ , (48) can be written as

$$x_t = \Lambda_{CI} f_{CI,t} + v_{CI,t}, \tag{49}$$

or

$$x_{it} = \lambda'_{CI,i} f_{CI,t} + v_{CI,it}, \tag{50}$$

where  $\lambda_i$  denotes the  $i$ -th row of

$$\Lambda_{CI} = \begin{pmatrix} 1 & \frac{1-e^{-\theta\tau_1}}{\theta\tau_1} & (\frac{1-e^{-\theta\tau_1}}{\theta\tau_1} - e^{-\theta\tau_1}) \\ \vdots & \vdots & \vdots \\ 1 & \frac{1-e^{-\theta\tau_N}}{\theta\tau_N} & (\frac{1-e^{-\theta\tau_N}}{\theta\tau_N} - e^{-\theta\tau_N}) \end{pmatrix}, \tag{51}$$

which is the  $N \times 3$  matrix of *known* factor loadings because we fix  $\theta = 0.0609$  following Diebold and Li (2006). The NS factors  $\hat{f}_{CI,t} = (\hat{\beta}_{1t}, \hat{\beta}_{2t}, \hat{\beta}_{3t})'$  are estimated from regressing  $x_{it}$  on  $\lambda'_{CI,i}$  (over  $i = 1, \dots, N$ ) by fitting the yield curve period by period for each  $t$ .

Then, we consider a linear forecast equation

$$y_t = (1 \hat{f}'_{CI,t-h}) \alpha_{CI} + u_{CI,t}, \quad t = h + 1, \dots, T, \tag{52}$$

in order to forecast  $y_{t+h}$  (such as output growth or inflation). We first estimate  $\hat{\alpha}_{CI}$  using the information up to time  $T$  and then form the forecast we call CI-NS by

$$\hat{y}_{T+h}^{CI-NS} = (1 \hat{f}'_{CI,T}) \hat{\alpha}_{CI}. \tag{53}$$

This method is comparable to CI-PC with number of factors fixed at  $k = 3$ . It differs from CI-PC, however, in that the three NS factors ( $\hat{\beta}_{1t}, \hat{\beta}_{2t}, \hat{\beta}_{3t}$ ) have intuitive interpretations as level, slope and curvature of the yield curve, while the first three principal components may not have a clear interpretation. In the empirical section, we also consider two alternative CI-NS forecasts by including only the level factor  $\hat{\beta}_{1t}$  (denoted CI-NS ( $k = 1$ )), and only the level and slope factors ( $\hat{\beta}_{1t}, \hat{\beta}_{2t}$ ) (denoted CI-NS ( $k = 2$ )) to see whether the level factor or the combination of level and slope factors have dominant contribution in forecasting output growth and inflation.

##### 4.2.2. NS Components of Forecasts $\hat{Y}$ (CF-NS)

While CI-NS solves the large- $N$  dimensionality problem by reducing the  $N$  yields to three factors  $\hat{f}_{CI,t} = (\hat{\beta}_{1t}, \hat{\beta}_{2t}, \hat{\beta}_{3t})'$ , it computes the factors entirely from yield curve information  $x_t$  only, without accounting for the variable  $y_{t+h}$  to be forecast. Similar in spirit to CF-PC, here we can improve CI-NS by supervising the factor computation, which we term as CF-NS.

The CF-NS forecast is based on the NS factors of  $\hat{y}_{t+h} := (\hat{y}_{t+h}^{(1)}, \hat{y}_{t+h}^{(2)}, \dots, \hat{y}_{t+h}^{(N)})'$ , a vector of the  $N$  individual forecasts as in (10) and (11),

$$\hat{y}_{t+h} = \Lambda_{CF} f_{CF,t+h} + v_{CF,t+h}, \tag{54}$$



with  $\Lambda_{CF} = \Lambda_{CI}$  in (51). Hence,  $\Lambda_{CI} = \Lambda_{CF} = \Lambda$  for the NS factor models. Note that, when the NS factors loadings are normalized to sum up to one, the three CF-NS factors

$$\begin{aligned} \hat{f}_{CF,t+h} &= \Lambda' \hat{y}_{t+h} \\ &= \left( \frac{1}{s_1} \sum_{i=1}^N \hat{y}_{T+h}^{(i)} \quad \frac{1}{s_2} \sum_{i=1}^N \left( \frac{1-e^{-\theta\tau_i}}{\theta\tau_i} \right) \hat{y}_{T+h}^{(i)} \quad \frac{1}{s_3} \sum_{i=1}^N \left( \frac{1-e^{-\theta\tau_i}}{\theta\tau_i} - e^{-\theta\tau_i} \right) \hat{y}_{T+h}^{(i)} \right)' \end{aligned} \tag{55}$$

are weighted individual forecasts with the three *normalized* NS loadings, with  $s_1 = N$ ,  $s_2 = \sum_{i=1}^N \left( \frac{1-e^{-\theta\tau_i}}{\theta\tau_i} \right)$ , and  $s_3 = \sum_{i=1}^N \left( \frac{1-e^{-\theta\tau_i}}{\theta\tau_i} - e^{-\theta\tau_i} \right)$ . The CF-NS forecast can be obtained from the forecasting equation

$$\begin{aligned} y_{t+h} &= \hat{f}'_{CF,t+h} \alpha_{CF} + u_{CF,t+h}, \\ \hat{y}_{T+h}^{CF-NS} &= \hat{f}'_{CF,T+h} \hat{\alpha}_{CF}, \end{aligned} \tag{56}$$

which is denoted CF-NS( $k = 3$ ). The parameter vector  $\hat{\alpha}_T$  is estimated using information up to time  $T$ . Using only the first factor or the first two factors, one can obtain the forecasts CF-NS( $k = 1$ ) and CF-NS( $k = 2$ ).

Note that, while the CF-PC method can be used for data of many kinds, the CF-NS method we propose is tailored to forecasting using the yield curve. It uses fixed factor loadings in  $\Lambda$  that are the NS exponential factor loadings for yield curve modeling, and hence avoids the estimation of factor loadings. In contrast, CF-PC needs to estimate  $\Lambda$ .

Also note that, by construction, CF-NS( $k = 1$ ) is the equally weighted combined forecast  $\frac{1}{N} \sum_{i=1}^N \hat{y}_{T+h}^{(i)}$ .

### 5. Forecasting Output Growth and Inflation

This section presents the empirical analysis where we describe the data, implement forecasting methods introduced in the previous sections on forecasting output growth and inflation, and analyze out-of-sample forecasting performances. This allows us to analyze the differences between output growth and inflation forecasting using the same yield curve information and to compare the strengths of different methods.

#### 5.1. Data

Let  $y_{t+h}$  denote the variable to be forecast (output growth or inflation) using yield information up to time  $t$ , where  $h$  denotes the forecast horizon. The predictor vector  $\mathbf{x}_t = (x_t(\tau_1), x_t(\tau_2), \dots, x_t(\tau_N))'$  contains the information about the yield curve at various maturities:  $x_t(\tau_i)$  denotes the zero coupon yield of maturity  $\tau_i$  months at time  $t$  ( $i = 1, 2, \dots, N$ ).

Two forecast targets, output growth and inflation, are constructed respectively as monthly growth rate of Personal Income (PI, seasonally adjusted annual rate) and monthly change in CPI (Consumer Price Index for all urban consumers: all items, seasonally adjusted) from 1970:01 to 2010:01. PI and CPI data are obtained from the web site of the Federal Reserve Bank of St. Louis (FRED2).

We apply the following data transformations. For the monthly growth rate of PI, we set  $y_{t+h} = 1200[(1/h) \ln(\text{PI}_{t+h}/\text{PI}_t)]$  as the forecast target (as used in [Ang et al. \(2006\)](#)). For the consumer price index (CPI), we set  $y_{t+h} = 1200[(1/h) \ln(\text{CPI}_{t+h}/\text{CPI}_t)]$  as the forecast target (as used in [Stock and Watson \(2007\)](#)).<sup>9</sup>

---

<sup>9</sup>  $y_{t+h} = 1200[(1/h) \ln(\text{CPI}_{t+h}/\text{CPI}_t) - \ln(\text{CPI}_t/\text{CPI}_{t-1})]$  is used in [Bai and Ng \(2008\)](#).

Our yield curve data consist of U.S. government bond prices, coupon rates, and coupon structures, as well as issue and redemption dates from 1970:01 to 2009:12.<sup>10</sup> We calculate zero-coupon bond yields using the unsmoothed [Fama and Bliss \(1987\)](#) approach. We measure bond yields on the second day of each month. We also apply several data filters designed to enhance data quality and focus attention on maturities with good liquidity. First, we exclude floating rate bonds, callable bonds and bonds extended beyond the original redemption date. Second, we exclude outlying bond prices less than 50 or greater than 130 because their price discounts/premium are too high and imply thin trading, and we exclude yields that differ greatly from yields at nearby maturities. Finally, we use only bonds with maturity greater than one month and less than fifteen years because other bonds are not actively traded. Indeed, to simplify our subsequent estimation, using linear interpolation we pool the bond yields into fixed maturities of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 33, 36, 39, 42, 45, 48, 51, 54, 57, 60, 63, 66, 72, 78, 84, 90, 96, 102, 108, and 120 months, where a month is defined as 30.4375 days.<sup>11</sup>

We examine some descriptive statistics (not reported for space) of the two forecast targets and yield curve level, slope, and curvature (empirical measures), over the full sample from 1970:01 to 2009:12 and the out-of-sample evaluation period from 1995:02 to 2010:01. We observe that both PI growth and CPI inflation become more moderate and less volatile from around the mid-1980s. This has become a stylized fact known as the “Great Moderation”. In particular, there is a substantial drop in persistency of CPI inflation. The volatility and persistency of the yield curve slope and curvature do not change much. The yield curve level, however, decreases and stabilizes.

In predicting macroeconomic variables using the term structure, yield spreads between yields with various maturities and the short rate are commonly used in the literature. One possible reason for this practice is that yield levels are treated as I(1) processes, so yield spreads will likely be I(0). Similarly, macroeconomic variables are typically assumed to be I(1) and transformed properly into I(0), so that, in using yield spreads to forecast macro targets, issues such as spurious regression are avoided. In this paper, however, we use yield levels (not spreads) to predict PI growth and CPI inflation (not change in inflation), for the following reasons. First, whether yields and inflation are I(1) or I(0) is still arguable. [Stock and Watson \(1999, 2012\)](#) use yield spreads and treat inflation as I(1), so they forecast change in inflation. [Inoue and Kilian \(2008\)](#), however, treat inflation as I(0). Since our target is forecasting inflation, not change in inflation, we will treat CPI inflation as well as yields as I(0) in our empirical analysis. Second, we emphasize real-time, out-of-sample forecasting performance more than in-sample concerns. As long as out-of-sample forecast performance is unaltered or even improved, we think the choice of treating the variables as I(1) or I(0) variables does not matter much.<sup>12</sup> Third, using yield levels will allow us to provide clearer interpretations for questions such as what part of the yield curve contributes the most towards predicting PI growth or CPI inflation, and how the different parts of the yield curve interact in the prediction, etc.

## 5.2. Out-of-Sample Forecasting

All forecasting models are estimated in a rolling window scheme with window size  $R = 300$  months ending at month  $t$  (starting at  $t - R + 1$ ). In the evaluation period from  $t = 1995:02$

<sup>10</sup> As a robust check, we apply our method to the original yield data of [Diebold and Li \(2006\)](#) and also to the sub-samples in our data set. The results are essentially the same as those summarized at the end of Section 5.

<sup>11</sup> It may be interesting to explore whether different maturity yields might have different effects on the forecast outcome. However, the present paper is focused on the comparison between CF and CI, rather than a detailed CI-only analysis, e.g., to find the best maturity yield for the forecast outcome. Nevertheless, our CI-NS model has reflected such effects as the three NS factors (level, slope, and curvature) are different combinations of bond maturities as shown in Equation (55). The different coefficients on the NS factors suggest that different bond maturities have different effects on the forecast outcome, as [Gogas et al. \(2015\)](#) has found.

<sup>12</sup> While not reported for space, we tried forecasting change in inflation and found forecasting inflation directly using all yield levels improves out-of-sample performances of most forecasting methods by a large margin.

to  $t = 2010:01$  (180 months), the first rolling sample to estimate models begins at 1970:02 and ends at 1995:01, the second rolling sample is for 1970:03–1995:02, the third 1970:04–1995:03, and so on. The out-of-sample evaluation period is from 1995:02 to 2010:01 (hence out-of-sample size  $P = 180$ ).<sup>13</sup> In all NS-related methods (CI and CF), we set  $\theta$ , the parameter that governs the exponential decay rate, at 0.0609 for reasons discussed in Diebold and Li (2006).<sup>14</sup> We compare  $h$ -months-ahead out-of-sample forecasting results of those methods introduced so far for  $h = 1, 3, 6, 12, 18, 24, 30, 36$  months ahead.

Figure 5 illustrates what economic contents these factors in CF-PC may bear. It shows that the first PC assigns about equal weights to all  $N = 50$  individual forecasts that use yields at various maturities (in months) so that it may be interpreted as the factor that captures the level of the yield curve; the second PC assigns roughly increasing weights so that it may be interpreted as the factor capturing the slope; and the third PC assigns roughly first decreasing then increasing weights, so that it may be interpreted as factor capturing curvature.

Tables 1 and 2 present the root mean squared forecast errors (RMSFE) of PC methods with  $k = 1, 2, 3, 4, 5$ , and of NS methods with  $k = 1, 2, 3$ , for PI growth (Table 1A) and for CPI inflation (Table 2A) forecasts using all 50 yield levels.<sup>15</sup> In Panel A of Tables 1 and 2, we report the Root Mean Squared Forecast Errors (RMSFE, which is the squared root of the MSFE of a model).<sup>16</sup> In Panel B of Tables 1 and 2, we report Relative Supervision of CI-PC vs. CF-PC and Relative Supervision of CI-NS vs. CF-NS, according to Definition 3, which is the ratio of the MSFEs of two CI and CF models. The relative supervision in Panel B can be obtained from RMSFEs in Panel A. For simplicity of presentation in Panel B, we present the relative supervision only with the same number of factors ( $k_{CI} = k_{CF}$  and  $k_{NS} = k_{NS}$ ).

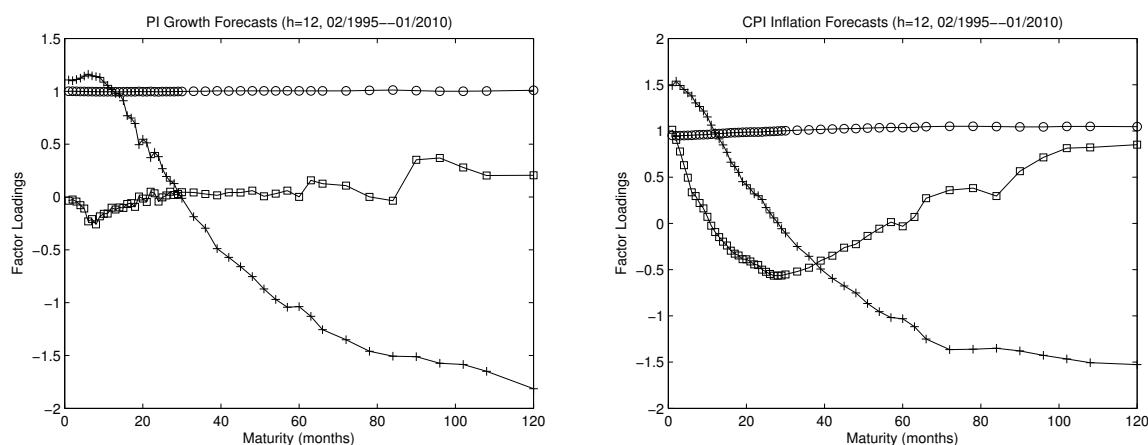


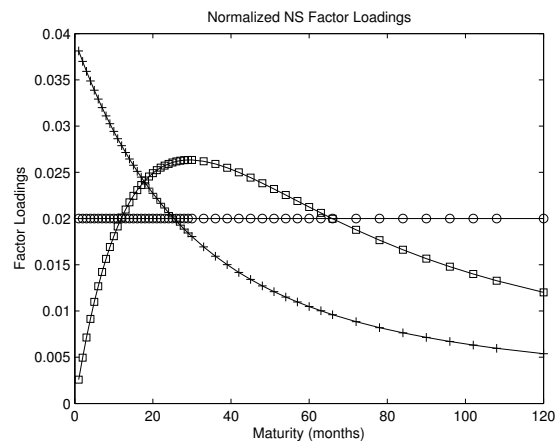
Figure 5. Cont.

<sup>13</sup> As a robust check, we have also tried with different sample splits for the estimation and prediction periods, i.e., the number of in-sample regression observations and the out-of-sample evaluation observations. We find that the results are similar.

<sup>14</sup> For different values of  $\theta$ , the performances of CI-NS and CF-NS change only marginally.

<sup>15</sup> While we report the results for  $k = 4, 5$  for CF-PC, we do not report for  $k = 4, 5$  for CF-NS. Svensson (1995) and Christensen et al. (2009) (CDR 2009) extend the three factor NS model to four or five factor NS models. CDR's dynamic generalized NS model has five factors with one level factor, two slope factors and two curvature factors. The Svensson and CDR extensions are useful to fit the yield curve at longer maturities ( $>10$  years). Because we only used yields with maturities  $\leq 10$  years, the second curvature factor loadings will look similar to the slope factor loadings and we will have collinearity problem. CDR use yields up to 30 years. The 4th and 5th factors have no clear economic interpretations and are hard to explain. For these reasons, we report results for  $k = 1, 2, 3$  for the CF-NS model.

<sup>16</sup> For the statistical significance of the loss-difference (see Definition 2), the asymptotic  $p$ -values of the Diebold–Mariano statistics are all very close to zero especially for larger values of the forecast horizon  $h$ .



**Figure 5.** Factor loadings of principal components and Nelson–Siegel factors. The first two panels: factor loadings of the first three principal components in CF-PC ( $k = 3$ ) averaged over the out-of-sample period (02/1995–01/2010), for both PI growth (first panel) and CPI inflation (second panel). The abscissa refers to the 50 individual forecasts that use yields at the 50 maturities (in months). The loading of the first principal component has the circle-symbol, the second the cross-symbol, and the third the square-symbol. The third panel: three normalized Nelson–Siegel (NS) exponential loadings in CF-NS that correspond to the three NS factors, respectively. The abscissa refers to the 50 individual forecasts that use yields at the 50 maturities (in months). The circled line denotes the first normalized NS factor loading  $1/N$ , the crossed line denotes the second normalized NS factor loading  $(1 - e^{-\theta\tau})/(\theta\tau)$ , divided by the sum, and the squared line denotes the third normalized NS factor loading  $(1 - e^{-\theta\tau})/(\theta\tau) - e^{-\theta\tau}$ , divided by the sum, where  $\tau$  denotes maturity and  $\theta$  is fixed at 0.0609.

We find that, in general, supervised factorization performs better. The CF schemes (CF-PC and CF-NS) perform substantially better than the CI schemes (CI-PC and CI-NS). Within the same CF or CI schemes, two alternative factorizations work similarly: CF-PC and CF-NS are about the same, and CI-PC and CI-NS are about the same. We summarize our findings from Figure 5 and Tables 1 and 2 as follows.

1. *Supervision is similar for CF-PC and CF-NS.* The factor loadings for CF-NS and for CF-PC are similar as shown in Figure 5. Panel (c) of the figure plots three normalized NS exponential loadings in CF-NS that correspond respectively to the three NS factors. Note that the factor loadings in CF-NS are pre-specified while those in CF-PC are estimated from the  $N$  individual forecasts. Nevertheless, their shapes in panel (a) look very similar to those of the CF-PC loadings in panels (a) and (b) (apart from the signs). Accordingly, out-of-sample forecasting performance of CF-PC and CF-NS are very similar as shown in Panel A of Tables 1 and 2.
2. *Supervision is substantial.* Supervised factor models perform better than unsupervised factor models in forecasting. Both CF-PC and CF-NS are much better than CI-PC and CI-NS models as shown in Panel B of Tables 1 and 2.
3. *Supervision is generally stronger for a longer forecast horizon  $h$ .* The advantage of CF-PC over CI-PC generally increases with forecast horizon  $h$ , as shown in Panel B of Tables 1 and 2.<sup>17</sup>

<sup>17</sup> We conducted a Monte Carlo (not reported), which are consistent with the empirical results that the supervision is stronger for a longer forecast horizon  $h$ .

4. We often get the best supervised predictions with a single factor ( $k = 1$ ) with the CF-factor models.<sup>18</sup> Since CF-NS( $k = 1$ ) is the equally weighted combined forecast as noted in Section 4.2.2, this is another case of the forecast combination puzzle discussed in Remark 3 that the equal-weighted forecast combination is hard to beat. Since CF-PC( $k = 1$ ) is numerically identical to CF-NS( $k = 1$ ) as shown in Figure 5, CF-PC( $k = 1$ ) is also effectively equally weighted forecast averaging.<sup>19</sup>

**Table 1.** Out-of-sample forecasting of personal income growth.

Panel A. Root Mean Squared Forecast Errors								
	$h = 1$	$h = 3$	$h = 6$	$h = 12$	$h = 18$	$h = 24$	$h = 30$	$h = 36$
CI-PC( $k = 1$ )	5.64	3.56	2.99	2.78	2.61	2.50	2.46	2.42
CI-PC( $k = 2$ )	5.67	3.64	3.12	3.00	2.81	2.66	2.55	2.45
CI-PC( $k = 3$ )	5.71	3.69	3.19	3.08	2.92	2.77	2.63	2.49
CI-PC( $k = 4$ )	5.72	3.76	3.23	3.12	2.93	2.77	2.58	2.36
CI-PC( $k = 5$ )	5.74	3.78	3.26	3.15	2.98	2.81	2.61	2.38
CI-NS( $k = 1$ )	5.84	3.84	3.28	3.06	2.86	2.69	2.53	2.41
CI-NS( $k = 2$ )	5.71	3.71	3.20	3.11	2.93	2.77	2.62	2.48
CI-NS( $k = 3$ )	5.72	3.69	3.19	3.09	2.93	2.78	2.63	2.47
CF-PC( $k = 1$ )	5.60	3.45	2.83	2.54	2.24	1.95	1.75	1.58
CF-PC( $k = 2$ )	5.56	3.43	2.83	2.62	2.31	1.93	1.76	1.61
CF-PC( $k = 3$ )	5.60	3.44	2.94	2.78	2.47	2.02	1.65	1.48
CF-PC( $k = 4$ )	5.63	3.60	3.08	2.83	2.39	1.97	1.67	1.45
CF-PC( $k = 5$ )	5.63	3.60	3.05	2.87	2.41	2.05	1.69	1.51
CF-NS( $k = 1$ )	5.60	3.45	2.83	2.54	2.24	1.95	1.75	1.58
CF-NS( $k = 2$ )	5.56	3.43	2.84	2.62	2.30	1.95	1.76	1.62
CF-NS( $k = 3$ )	5.59	3.44	2.94	2.79	2.47	2.02	1.64	1.48

<sup>18</sup> Figlewski and Urich (1983) talked about various constrained models in forming a combination of forecasts and examined when we need more than the simple averaging combined forecast. They discussed a sufficient condition when the simple average of forecasts is the optimal forecast combination: “Under the most extensive set of constraints, forecast errors are assumed to have zero mean and to be independent and identically distributed. In this case the optimal forecast is the simple average.” This corresponds to CF-PC( $k = 1$ ) and CF-NS( $k = 1$ ) when the first factor ( $k = 1$ ) in PC or NS is sufficient for the CF factor model. It is clearly the case in CF-NS as shown in Equation (55). One can show that the first PC (corresponding to the largest singular value) would also be the simple average. Hence, in terms of the CF-factor model, the forecast combination puzzle amounts to the fact that we often do not need the second PC factor. Interestingly, (Figlewski and Urich 1983, p. 696) continued to note the cases when the simple average is not optimal: “However, the hypothesis of independence among forecast errors is overwhelmingly rejected for our data—errors are highly positively correlated with one another.” On the other hand, they also noted other reasons why the simple average may still be preferred, as they wrote, “Because the estimated error structure was not completely stable over time, the models which adjusted for correlation did not achieve lower mean squared forecast error than the simple average in out-of-sample tests. Even so, we find...that forecasts from these models, while less accurate than the simple mean, do contain information which is not fully reflected in prices in the money market, and is therefore economically valuable.” We thank a referee for letting us know on this from Figlewski and Urich (1983).

<sup>19</sup> While the simple equally weighted forecast combination can be implemented without the use of PCA or without making reference to the NS model, it is important to note that the simple average combined forecast indeed corresponds the first CF-PC factor (CF-PC( $k = 1$ )) or the first CF-NS factor (CF-NS( $k = 1$ )). In view of Figlewski and Urich (1983), it will be useful to know when the first factor ( $k = 1$ ) is enough so that the simple average is good or when the higher order factors ( $k > 1$ ) may be necessary as they contain more information in addition to the first CF-factor. This is important in understanding the forecast combination puzzle. The forecast combination puzzle is about whether to include only the first CF factor or more.

Table 1. Cont.

Panel B. Relative Supervision $s_{rel}(X, y, k_{CI}, k_{CF})$								
	$h = 1$	$h = 3$	$h = 6$	$h = 12$	$h = 18$	$h = 24$	$h = 30$	$h = 36$
CI-PC( $k = 1$ ) vs. CF-PC( $k = 1$ )	1.01	1.06	1.12	1.20	1.36	1.64	1.98	2.35
CI-PC( $k = 2$ ) vs. CF-PC( $k = 2$ )	1.04	1.13	1.22	1.31	1.48	1.90	2.10	2.32
CI-PC( $k = 3$ ) vs. CF-PC( $k = 3$ )	1.04	1.15	1.18	1.23	1.40	1.88	2.54	2.83
CI-PC( $k = 4$ ) vs. CF-PC( $k = 4$ )	1.03	1.09	1.10	1.22	1.50	1.98	2.39	2.65
CI-PC( $k = 5$ ) vs. CF-PC( $k = 5$ )	1.04	1.10	1.14	1.20	1.53	1.88	2.39	2.48
CI-NS( $k = 1$ ) vs. CF-NS( $k = 1$ )	1.09	1.24	1.34	1.45	1.63	1.90	2.09	2.33
CI-NS( $k = 2$ ) vs. CF-NS( $k = 2$ )	1.05	1.17	1.27	1.41	1.62	2.02	2.22	2.34
CI-NS( $k = 3$ ) vs. CF-NS( $k = 3$ )	1.05	1.15	1.18	1.23	1.41	1.89	2.57	2.79

The forecast target is Output Growth  $y_{t+h} = 1200 \times \log(P_{t+h}/P_t) \div h$ . Out-of-sample forecasting period is 02/1995–01/2010. In Panel A, reported are the Root Mean Squared Forecast Errors (which is the squared root of the MSFE of a model). In Panel B, reported are Relative Supervision of CI-PC vs. CF-PC and Relative Supervision of CI-NS vs. CF-NS, according to Definition 3, which is the ratio of the MSFEs of the two models. For simplicity of presentation, we present the relative supervision in Panel B only with the same number of factors ( $k_{CI} = k_{CF} = k$  and  $k_{NS} = k_{NS} = k$ ).

Table 2. Out-of-sample forecasting of CPI inflation.

Panel A. Root Mean Squared Forecast Errors								
	$h = 1$	$h = 3$	$h = 6$	$h = 12$	$h = 18$	$h = 24$	$h = 30$	$h = 36$
CI-PC( $k = 1$ )	3.77	2.86	2.25	1.92	1.94	2.16	2.47	2.75
CI-PC( $k = 2$ )	4.21	3.45	2.96	2.76	2.77	2.84	2.96	3.08
CI-PC( $k = 3$ )	4.24	3.50	3.00	2.82	2.88	2.98	3.10	3.19
CI-PC( $k = 4$ )	4.31	3.57	3.05	2.87	2.91	3.00	3.12	3.18
CI-PC( $k = 5$ )	4.30	3.58	3.07	2.93	3.00	3.10	3.20	3.23
CI-NS( $k = 1$ )	3.95	3.12	2.62	2.48	2.60	2.79	2.97	3.10
CI-NS( $k = 2$ )	4.22	3.46	2.98	2.82	2.88	2.98	3.09	3.18
CI-NS( $k = 3$ )	4.24	3.50	3.01	2.83	2.89	2.99	3.11	3.20
CF-PC( $k = 1$ )	3.65	2.67	1.91	1.31	1.01	0.90	0.96	1.08
CF-PC( $k = 2$ )	3.66	2.70	1.93	1.35	1.10	1.05	1.11	1.19
CF-PC( $k = 3$ )	3.68	2.72	1.97	1.47	1.29	1.19	1.19	1.20
CF-PC( $k = 4$ )	3.74	2.80	2.01	1.47	1.22	1.14	1.15	1.17
CF-PC( $k = 5$ )	3.74	2.79	1.98	1.45	1.20	1.12	1.18	1.20
CF-NS( $k = 1$ )	3.65	2.68	1.91	1.31	1.02	0.90	0.96	1.08
CF-NS( $k = 2$ )	3.66	2.70	1.93	1.35	1.10	1.05	1.10	1.19
CF-NS( $k = 3$ )	3.68	2.73	1.97	1.47	1.29	1.20	1.19	1.20

Panel B. Relative Supervision $s_{rel}(X, y, k_{CI}, k_{CF})$								
	$h = 1$	$h = 3$	$h = 6$	$h = 12$	$h = 18$	$h = 24$	$h = 30$	$h = 36$
CI-PC( $k = 1$ ) vs. CF-PC( $k = 1$ )	1.07	1.15	1.39	2.15	3.69	5.76	6.62	6.48
CI-PC( $k = 2$ ) vs. CF-PC( $k = 2$ )	1.32	1.63	2.35	4.18	6.34	7.32	7.11	6.70
CI-PC( $k = 3$ ) vs. CF-PC( $k = 3$ )	1.33	1.66	2.32	3.68	4.98	6.27	6.79	7.07
CI-PC( $k = 4$ ) vs. CF-PC( $k = 4$ )	1.33	1.63	2.30	3.81	5.69	6.93	7.36	7.39
CI-PC( $k = 5$ ) vs. CF-PC( $k = 5$ )	1.32	1.65	2.40	4.08	6.25	7.66	7.35	7.25
CI-NS( $k = 1$ ) vs. CF-NS( $k = 1$ )	1.17	1.36	1.88	3.58	6.50	9.61	9.57	8.24
CI-NS( $k = 2$ ) vs. CF-NS( $k = 2$ )	1.33	1.64	2.38	4.36	6.85	8.05	7.89	7.14
CI-NS( $k = 3$ ) vs. CF-NS( $k = 3$ )	1.33	1.64	2.33	3.71	5.02	6.21	6.83	7.11

The forecast target is Inflation  $y_{t+h} = 1200 \times \log(CPI_{t+h}/CPI_t) \div h$ . Out-of-sample forecasting period is 02/1995–01/2010. In Panel A, reported are the Root Mean Squared Forecast Errors (which is the squared root of the MSFE of a model). In Panel B, reported are Relative Supervision of CI-PC vs. CF-PC and Relative Supervision of CI-NS vs. CF-NS, according to Definition 3, which is the ratio of the MSFEs of the two models. For simplicity of presentation, we present the relative supervision in Panel B only with the same number of factors ( $k_{CI} = k_{CF} = k$  and  $k_{NS} = k_{NS} = k$ ).

## 6. Conclusions

For forecasting in the presence of many predictors, it is often useful to reduce the dimension by a factor model (in a dense case) or by variable selection (in a sparse case). In this paper, we consider a factor model. In particular, we examine the supervised principal component analysis of Chan et al. (1999). The model is called CF-PC, as the principal components of many forecasts are the combined forecasts.

The CF-PC extracts factors from the space spanned by forecasts rather than from the space spanned by predictors. This factorization of the forecasts improves forecast performance compared to factor analysis of the predictors. We extend the CF-PC to CF-NS, which uses the NS factor model in place of the PC factor model, for the application where the predictors are the yield curve. While the yield curve is a functional data consisting of many different maturity points on a curve at each time, the NS factors can parsimoniously capture the shapes of the curve.

We have applied the CF-PC and CF-NS models in forecasting output growth and inflation using a large number of bond yields to examine if the supervised factorization improves forecast performance. In general, we have found that CF-PC and CF-NS perform substantially better than CI-PC and CI-NS, that the advantage of supervised factor models is even larger for longer forecast horizons, and that the two alternative factor models based on PC and NS factors are similar and perform similarly.

**Author Contributions:** All authors contributed equally to the paper.

**Acknowledgments:** We would like to thank the two referees for helpful comments. We also thank Jonathan Wright and seminar participants at FRB of San Francisco, FRB of St Louis, Federal Reserve Board (Washington DC), Bank of Korea, SETA meeting, Stanford Institute of Theoretical Economics (SITE), University of Cambridge, NCSU, UC Davis, UCR, UCSB, UCSD, USC, Purdue, LSU, Indiana, Drexel, OCC, WMU, and SNU, for useful discussions and comments. All errors are our own. E.H. acknowledges support from the Danish National Research Foundation. The views presented in this paper are solely those of the authors and do not necessarily represent those of ICBCCS, the Federal Reserve Board or their staff.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Calculation of Absolute and Relative Supervision in Example 1

Using  $R$  and  $\Sigma_k$  obtained from the SVD for CI in (36), and  $S$  and  $\Theta_k$  obtained from the SVD for CF in (40), we calculate the absolute supervision and relative supervision for each  $k$ . The CI factors are  $F_{CI} = R\Sigma_k$  from (23), and the CF factors  $F_{CF} = S\Theta_k$  from (28).

For  $k = 1$ ,

$$F_{CI} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad F_{CF} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 5 \\ 0 \end{bmatrix}, \quad (A1)$$

$$\hat{y}_{CI-PC} = R_1 R_1' y = (1, 0, 0, 0, 0, 0)', \quad (A2)$$

$$\hat{y}_{CF-PC} = S_1 S_1' y = (0, 0, 0, 0, 5, 0)', \quad (A3)$$

$$\|y - \hat{y}_{CI-PC}\|^2 = \|(1, 2, 3, 4, 5, 0)' - (1, 0, 0, 0, 0, 0)'\|^2 = 54, \quad (A4)$$

$$\|y - \hat{y}_{CF-PC}\|^2 = \|(1, 2, 3, 4, 5, 0)' - (0, 0, 0, 0, 5, 0)'\|^2 = 30. \quad (A5)$$

Hence,  $s_{abs}(X, y, 1, 1) = \|y - \hat{y}_{CI-PC}\|^2 - \|y - \hat{y}_{CF-PC}\|^2 = 54 - 30 = 24$ , and  $s_{rel}(X, y, 1, 1) = \|y - \hat{y}_{CI-PC}\|^2 / \|y - \hat{y}_{CF-PC}\|^2 = 54/30 = 1.8$ .

For  $k = 2$ ,

$$F_{CI} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad F_{CF} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 0 \\ 0 & 4 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 4 \\ 5 & 0 \\ 0 & 0 \end{bmatrix}, \quad (A6)$$

$$\hat{y}_{CI-PC} = R_2 R_2' y = (1, 2, 0, 0, 0, 0)', \quad (A7)$$

$$\hat{y}_{CF-PC} = S_2 S_2' y = (0, 0, 0, 4, 5, 0)', \quad (A8)$$

$$\|y - \hat{y}_{CI-PC}\|^2 = \|(1, 2, 3, 4, 5, 0)' - (1, 2, 0, 0, 0, 0)'\|^2 = 50, \quad (A9)$$

$$\|y - \hat{y}_{CF-PC}\|^2 = \|(1, 2, 3, 4, 5, 0)' - (0, 0, 0, 4, 5, 0)'\|^2 = 14. \quad (A10)$$

Hence,  $s_{abs}(X, y, 2, 2) = \|y - \hat{y}_{CI-PC}\|^2 - \|y - \hat{y}_{CF-PC}\|^2 = 50 - 14 = 36$ , and  $s_{rel}(X, y, 2, 2) = \|y - \hat{y}_{CI-PC}\|^2 / \|y - \hat{y}_{CF-PC}\|^2 = 50/14 = 3.6$ .

For  $k = 3$ ,

$$F_{CI} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad F_{CF} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \\ 0 & 4 & 0 \\ 5 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad (A11)$$

$$\hat{y}_{CI-PC} = R_3 R_3' y = (1, 2, 3, 0, 0, 0)', \quad (A12)$$

$$\hat{y}_{CF-PC} = S_3 S_3' y = (0, 0, 3, 4, 5, 0)', \quad (A13)$$

$$\|y - \hat{y}_{CI-PC}\|^2 = \|(1, 2, 3, 4, 5, 0)' - (1, 2, 3, 0, 0, 0)'\|^2 = 41, \quad (A14)$$

$$\|y - \hat{y}_{CF-PC}\|^2 = \|(1, 2, 3, 4, 5, 0)' - (0, 0, 3, 4, 5, 0)'\|^2 = 5. \quad (A15)$$

Hence,  $s_{abs}(X, y, 3, 3) = \|y - \hat{y}_{CI-PC}\|^2 - \|y - \hat{y}_{CF-PC}\|^2 = 41 - 5 = 36$ , and  $s_{rel}(X, y, 3, 3) = \|y - \hat{y}_{CI-PC}\|^2 / \|y - \hat{y}_{CF-PC}\|^2 = 41/5 = 8.2$ .

For  $k = 4$ ,

$$F_{CI} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad F_{CF} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 3 \\ 0 & 4 & 0 \\ 5 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad (A16)$$

$$\hat{y}_{CI-PC} = R_4 R_4' y = (1, 2, 3, 4, 0, 0)', \quad (A17)$$

$$\hat{y}_{CF-PC} = S_4 S_4' y = (0, 2, 3, 4, 5, 0)', \quad (A18)$$



$$\|y - \hat{y}_{\text{CI-PC}}\|^2 = \|(1, 2, 3, 4, 5, 0)' - (1, 2, 3, 4, 0, 0)'\|^2 = 25, \quad (\text{A19})$$

$$\|y - \hat{y}_{\text{CF-PC}}\|^2 = \|(1, 2, 3, 4, 5, 0)' - (0, 2, 3, 4, 5, 0)'\|^2 = 1. \quad (\text{A20})$$

Hence,  $s_{\text{abs}}(X, y, 4, 4) = \|y - \hat{y}_{\text{CI-PC}}\|^2 - \|y - \hat{y}_{\text{CF-PC}}\|^2 = 25 - 1 = 24$ , and  $s_{\text{rel}}(X, y, 4, 4) = \|y - \hat{y}_{\text{CI-PC}}\|^2 / \|y - \hat{y}_{\text{CF-PC}}\|^2 = 25/1 = 25$ .

For  $k = 5$ ,  $s_{\text{abs}}(X, y, 5, 5) = y'(SS' - RR')y = 0$  because

$$\|y - \hat{y}_{\text{CI-PC}}\|^2 = \|(1, 2, 3, 4, 5, 0)' - (1, 2, 3, 4, 5, 0)'\|^2 = 0, \quad (\text{A21})$$

$$\|y - \hat{y}_{\text{CF-PC}}\|^2 = \|(1, 2, 3, 4, 5, 0)' - (1, 2, 3, 4, 5, 0)'\|^2 = 0. \quad (\text{A22})$$

Hence, as noted in Remark 4,  $s_{\text{abs}}(X, y, 5, 5) = \|y - \hat{y}_{\text{CI-PC}}\|^2 - \|y - \hat{y}_{\text{CF-PC}}\|^2 = 0 - 0 = 0$ , and  $s_{\text{rel}}(X, y, 5, 5)$  is not defined for  $k = N = 5$ .

## References

- Ang, Andrew, and Monika Piazzesi. 2003. A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables. *Journal of Monetary Economics* 50: 745–87. [\[CrossRef\]](#)
- Ang, Andrew, Monika Piazzesi, and Min Wei. 2006. What Does the Yield Curve Tell Us about GDP Growth? *Journal of Econometrics* 131: 359–403. [\[CrossRef\]](#)
- Armah, Nii Ayi, and Norman R. Swanson. 2010. Seeing Inside the Black Box: Using Diffusion Index Methodology to Construct Factor Proxies in Large Scale Macroeconomic Time Series Environments. *Econometric Reviews* 29: 476–510. [\[CrossRef\]](#)
- Bai, Jushan. 2003. Inferential Theory for Factor Models of Large Dimensions. *Econometrica* 71: 135–71. [\[CrossRef\]](#)
- Bai, Jushan and Serena Ng. 2006. Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica* 74: 1133–50. [\[CrossRef\]](#)
- Bai, Jushan and Serena Ng. 2008. Forecasting Economic Time Series Using Targeted Predictors. *Journal of Econometrics* 146: 304–17. [\[CrossRef\]](#)
- Bair, Eric, Trevor Hastie, Debashis Paul, and Robert Tibshirani. 2006. Prediction by Supervised Principal Components. *Journal of the American Statistical Association* 101: 119–37. [\[CrossRef\]](#)
- Barndorff-Nielsen, Ole. 1978. *Information and Exponential Families in Statistical Theory*. New York: Wiley.
- Bernanke, Ben. 1990. On the Predictive Power of Interest Rates and Interest Rate Spreads, Federal Reserve Bank of Boston. *New England Economic Review* November/December: 51–68.
- Chan, Lewis, James Stock, and Mark Watson. 1999. A Dynamic Factor Model Framework for Forecast Combination. *Spanish Economic Review* 1: 91–121. [\[CrossRef\]](#)
- Christensen, Jens, Francis Diebold, and Glenn Rudebusch. 2009. An Arbitrage-free Generalized Nelson–Siegel Term Structure Model. *Econometrics Journal* 12: C33–C64. [\[CrossRef\]](#)
- De Jong, Sijmen. 1993. SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems* 18: 251–61. [\[CrossRef\]](#)
- De Jong, Sijmen and Henk Kiers. 1992. Principal Covariate Regression: Part I. Theory. *Chemometrics and Intelligent Laboratory Systems* 14: 155–64. [\[CrossRef\]](#)
- Diebold, Francis, and Canlin Li. 2006. Forecasting the Term Structure of Government Bond Yields. *Journal of Econometrics* 130: 337–64. [\[CrossRef\]](#)
- Diebold, Francis, Monika Piazzesi, and Glenn Rudebusch. 2005. Modeling Bond Yields in Finance and Macroeconomics. *American Economic Review* 95: 415–20. [\[CrossRef\]](#)
- Diebold, Francis, Glenn Rudebusch, and Boragan Aruoba. 2006. The Macroeconomy and the Yield Curve: A Dynamic Latent Factor Approach. *Journal of Econometrics* 131: 309–38. [\[CrossRef\]](#)
- Engle, Robert, David Hendry, and Jean-Francois Richard. 1983. Exogeneity. *Econometrica* 51: 277–304. [\[CrossRef\]](#)
- Estrella, Arturo. 2005. Why Does the Yield Curve Predict Output and Inflation? *The Economic Journal* 115: 722–44. [\[CrossRef\]](#)

- Estrella, Arturo, and Gikas Hardouvelis. 1991. The Term Structure as a Predictor of Real Economic Activity. *Journal of Finance* 46: 555–76. [\[CrossRef\]](#)
- Fama, Eugene, and Robert Bliss. 1987. The Information in Long-maturity Forward Rates. *American Economic Review* 77: 680–92.
- Figlewski, Stephen, and Thomas Ulrich. 1983. Optimal Aggregation of Money Supply Forecasts: Accuracy, Profitability and Market Efficiency. *Journal of Finance* 38: 695–710. [\[CrossRef\]](#)
- Friedman, Benjamin, and Kenneth Kuttner. 1993. Why Does the Paper-Bill Spread Predict Real Economic Activity? In *New Research on Business Cycles, Indicators and Forecasting*. Edited by James Stock and Mark Watson. Chicago: University of Chicago Press, pp. 213–54.
- Gogas, Periklis, Theophilos Papadimitriou, and Efthymia Chrysanthidou. 2015. Yield Curve Point Triplets in Recession Forecasting. *International Finance* 18: 207–26. [\[CrossRef\]](#)
- Groen, Jan, and George Kapetanios. 2016. Revisiting Useful Approaches to Data-Rich Macroeconomic Forecasting. *Computational Statistics & Data Analysis* 100: 221–39.
- Hamilton, James, and Dong Heon Kim. 2002. A Reexamination of the Predictability of Economic Activity Using the Yield Spread. *Journal of Money, Credit, and Banking* 34: 340–60. [\[CrossRef\]](#)
- Huang, Huiyu, and Tae-Hwy Lee. 2010. To Combine Forecasts or To Combine Information? *Econometric Reviews* 29: 534–70. [\[CrossRef\]](#)
- Inoue, Atsushi, and Lutz Kilian. 2008. How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation. *Journal of the American Statistical Association* 103: 511–22. [\[CrossRef\]](#)
- Kozicki, Sharon. 1997. Predicting Real Growth and Inflation with the Yield Spread, Federal Reserve Bank of Kansas City. *Economic Review* 82: 39–57.
- Lancaster, Tony. 2000. The incidental parameter problem since 1948. *Journal of Econometrics* 95: 391–413. [\[CrossRef\]](#)
- Litterman, Robert, and Jose Scheinkman. 1991. Common Factors Affecting Bond Returns. *Journal of Fixed Income* 1: 54–61. [\[CrossRef\]](#)
- Nelson, Charles, and Andrew Siegel. 1987. Parsimonious Modeling of Yield Curves. *Journal of Business* 60: 473–89. [\[CrossRef\]](#)
- Neyman, Jerzy, and Elizabeth Scott. 1948. Consistent Estimation from Partially Consistent Observations. *Econometrica* 16: 1–32. [\[CrossRef\]](#)
- Piazzesi, Monika. 2005. Bond Yields and the Federal Reserve. *Journal of Political Economy* 113: 311–44. [\[CrossRef\]](#)
- Rudebusch, Glenn, and Tao Wu. 2008. A Macro-Finance Model of the Term Structure, Monetary Policy, and the Economy. *Economic Journal* 118: 906–26. [\[CrossRef\]](#)
- Smith, Jeremy, and Kenneth Wallis. 2009. A Simple Explanation of the Forecast Combination Puzzle. *Oxford Bulletin of Economics and Statistics* 71: 331–55. [\[CrossRef\]](#)
- Stock, James, and Mark Watson. 1989. New Indexes of Coincident and Leading Indicators. In *NBER Macroeconomic Annual*. Edited by Olivier Blanchard and Stanley Fischer. Cambridge: MIT Press, vol. 4.
- Stock, James, and Mark Watson. 1999. Forecasting Inflation. *Journal of Monetary Economics* 44: 293–335. [\[CrossRef\]](#)
- Stock, James, and Mark Watson. 2002. Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association* 97: 1167–79. [\[CrossRef\]](#)
- Stock, James, and Mark Watson. 2004. Combination Forecasts of Output Growth in a Seven-country Data Set. *Journal of Forecasting* 23: 405–30. [\[CrossRef\]](#)
- Stock, James, and Mark Watson. 2007. Has Inflation Become Harder to Forecast? *Journal of Money, Credit, and Banking* 39: 3–34. [\[CrossRef\]](#)
- Stock, James, and Mark Watson. 2012. Generalized Shrinkage Methods for Forecasting Using Many Predictors. *Journal of Business and Economic Statistics* 30: 481–93. [\[CrossRef\]](#)
- Svensson, Lars. 1995. Estimating Forward Interest Rates with the Extended Nelson–Siegel Method. *Quarterly Review* 3: 13–26.
- Tibshirani, Robert. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B* 58: 267–88.
- Timmermann, Alan. 2006. Forecast Combinations. In *Handbook of Economic Forecasting*. Edited by Graham Elliott, Clive Granger and Alan Timmermann. Amsterdam: North-Holland, vol. 1, chp. 4.

Wright, Jonathan. 2009. Forecasting US Inflation by Bayesian Model Averaging. *Journal of Forecasting* 28: 131–44. [[CrossRef](#)]

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2006. Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics* 15: 262–86. [[CrossRef](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).