Forecasting the Yield Curve in a Data-Rich Environment

Peter Exterkate*

This version: August 1, 2008

Abstract

We investigate alternative methods for extracting macroeconomic information from a data-rich environment, with the objective of forecasting yield curves using the Nelson-Siegel model.

In particular, we consider (1) extracting principal components from grouped data, where the grouping follows from economic theory; (2) extracting principal covariates, which provide a trade-off between summarizing the information and constructing useful predictors, as in Heij et al. (2007); (3) selecting variables based on their joint predictive power, using the multiresponse sparse regression technique from Simila and Tikka (2006); and (4) selecting variables based on their individual predictive power, using a hard thresholding rule as proposed by Bai and Ng (2007).

Empirical results show that the first option results in accurate predictions for medium-term to long-term yields, while better forecasts for short-term yields are obtained using the second option. For predicting the medium-term part of the yield curve, the thresholding rule is also a useful tool.

*I would like to thank Christiaan Heij, Patrick Groenen, and Dick van Dijk for helpful comments on earlier versions of this paper.
1 Introduction

Forecasting yield curves is a notoriously difficult problem. Many attempts have failed to outperform the very naive benchmark of the no-change forecast obtained from a random walk model. An important branch of the relevant literature has focused on affine models, which postulate that yields evolve as affine functions of a limited number of latent risk factors. Duffee (2002) dismisses this entire class of models as being dominated by the simple random walk model in terms of forecast accuracy.

Recently, there has been renewed attention for the work of Nelson and Siegel (1987). Originally intended to model cross-sectional aspects of yield curves, the Nelson-Siegel model imposes a low-parametric and interpretable three-factor structure on the links between yields of different maturities. Diebold and Li (2006) find that a dynamic restatement of this model provides forecasts that outperform the random walk. Diebold et al. (2006) propose a way to include non-yield factors in a Nelson-Siegel-type model, and they find that the effect of macro aggregates on market yields is substantial.

The presence of a link between the macroeconomy and the yield curve should come as no surprise. Central banks around the world use short-term interest rates as their policy instruments, and it is widely recognized that their actions respond to macroeconomic aggregates such as inflation and real output; see Taylor (1993) for a discussion of these stylized facts. Because longer-term interest rates can be regarded as expected future short-term rates, discounted for risk premia, it is certainly plausible that the entire yield curve responds to macroeconomic shocks.

Such a link is also present in the reverse direction, in perhaps a more straightforward way. Economic agents respond to changing interest rates by altering their investment plans and by adjusting their inflation expectations. Diebold et al. (2006) find clear statistical evidence for the existence of this link, in both directions.

In today’s data-rich environment, a natural question is how to choose the macro factors to include in a model for forecasting yields. Diebold et al. (2006) simply use three major variables intended to represent the level of real economic activity, monetary policy, and inflation, respectively. Arguably, many more macro aggregates influence the evolution of the yield curve. However, including many macro factors leads to an enormous increase in the number of parameters to be estimated. For this reason, de Pooter et al. (2007) extract principal components from a large panel of macroeconomic aggregates and they indeed find an increase in forecast accuracy. Mönch (2008) finds similar results when augmenting an affine model with such principal components.

This study focuses on more advanced ways of incorporating macroeconomic information in the Nelson-Siegel framework. Simply extracting principal components from a pool containing any macroeconomic data that one can find, has two important drawbacks. Firstly, this procedure completely ignores economic theory. Like Diebold et al. (2006), an economist has prior ideas on which groups of economic aggregates may be useful in forecasting yields. Rather than “picking” some time series from these groups, however, it may be beneficial to use all the information in these groups.
To address this issue, one possible approach is to define groups of macroeconomic predictors based on economic theory, and then to extract principal components from these groups separately.

Secondly, and perhaps more fundamentally, in a principal components analysis the macro factors are created without any reference to the forecasting objective. That is, regardless of which series we wish to forecast, the constructed predictors will be the same. It can be expected that the accuracy of forecasts improves if the objective is explicitly used in constructing predictors. Three approaches to this end are pursued in this study:

- extracting principal covariates instead of principal components from groups of macroeconomic variables, as introduced by Heij et al. (2007);
- selecting macroeconomic variables using multiresponse sparse regression, a selection method proposed by Similä and Tikka (2006); and
- constructing macro factors using variables selected by a hard thresholding rule, a method proposed by Bai and Ng (2007).

These alternatives are compared on their ability to produce reliable predictions of the yield curve for different forecasting horizons. This comparison shows that principal components extracted from theory-based groups are very useful in forecasting medium-term to long-term yields, while short-term yields are more reliably predicted using principal covariates. The thresholding rule leads to improved forecasts for medium-term yields. Multiresponse sparse regression does not produce more accurate forecasts in the present context.

In the remainder of this article, we begin Section 2 with a description of the Nelson-Siegel model with the extensions by Diebold et al. (2006). Furthermore, the techniques of principal covariate regression, multiresponse sparse regression, and hard thresholding are introduced. Section 3 contains empirical results on forecasting U.S. zero-coupon yields. The last section concludes.
2 Methods

This section reviews the Nelson-Siegel model and the employed techniques of principal covariate regression, multiresponse sparse regression, and hard thresholding.

2.1 Nelson-Siegel model

Nelson and Siegel (1987) derive a very parsimonious model that fits observed yield curves surprisingly well. With minor modifications as explained in Diebold and Li (2006), their model is

\[ y_t(\tau) = \beta_{1t} + \beta_{2t} \left( \frac{1 - \exp(-\lambda_t \tau)}{\lambda_t \tau} \right) + \beta_{3t} \left( \frac{1 - \exp(-\lambda_t \tau)}{\lambda_t \tau} - \exp(-\lambda_t \tau) \right), \]

where \( y_t(\tau) \) is the yield at time \( t \) for maturity \( \tau \), \( \lambda_t \) is a parameter that governs the rate of decay, and the \( \beta_{it} \) are parameters that can be interpreted as level, slope, and curvature parameters respectively, following Diebold and Li (2006). If we consider a fixed set of maturities \( (\tau_1, \ldots, \tau_m) \) and denote the vector of observed yields at time \( t \) by \( y_t = (y_t(\tau_1), y_t(\tau_2), \ldots, y_t(\tau_m))' \) and the parameter vector by \( \beta_t = (\beta_{1t}, \beta_{2t}, \beta_{3t})' \), and if we add an error term to Equation (1), we find

\[ y_t = \Lambda_t \beta_t + \varepsilon_t, \]

where \( \Lambda_t \) depends on \( \lambda_t \) only.

Diebold and Li (2006) interpret Equation (2) as the measurement equation of a state space model. The state variables \( \beta_t \) are assumed to evolve according to a first-order vector autoregressive process:

\[ \beta_t - \mu = A(\beta_{t-1} - \mu) + \eta_t. \]

The disturbances \( \varepsilon_t \) and \( \eta_t \) are assumed to be white noise and mutually uncorrelated. Since there is no disturbance term in Equation (1), \( \varepsilon_t \) can be interpreted as a measurement error, given the good fit of the Nelson-Siegel model. Therefore, its covariance matrix is commonly assumed to be diagonal. The covariance matrix of \( \eta_t \) is unrestricted.

Typically, \( \lambda_t \) is assumed constant and its value is fixed by the researcher. However, a constant \( \lambda_t \) can also be estimated along with the other model parameters using the Kalman filter, and Koopman et al. (2007) propose ways of estimating certain types of time-varying \( \lambda_t \). To keep the present study focused, however, we decide to impose a time-invariant \( \lambda \), which we estimate along with the other model parameters.

Now assume that, at time \( t \), we have a large group of macroeconomic aggregates at our disposal, denoted \( x_t \). We wish to summarize this large amount of information
by much fewer factors $f_t$. These factors can be picked by the researcher as in [Diebold et al. (2006)], extracted using principal component analysis as in [de Pooter et al. (2007)], or constructed using any of the methods that are described below.

For now, assume that we have $f_t$ available. We follow [Diebold et al. (2006)] in their procedure for introducing this macro information into the Nelson-Siegel framework. The observation equation (2) remains unchanged; in the state equation (3), $f_t$ is appended to the state vector and the dimensions of $A$ and $\eta_t$ are increased as appropriate. Maximum likelihood estimation of the parameters, and forecasting, can still be done using the standard Kalman filter.

For convenience, the full model considered in this study is shown below:

$$
\begin{align*}
    y_t &= \Lambda \beta_t + \varepsilon_t, \\
    (\beta_t - \mu) &= A (\beta_{t-1} - \mu) + \eta_t, \\
    \varepsilon_t &\sim \mathcal{NID}(0, \text{diag}(\sigma_1^2, \ldots, \sigma_m^2)), \\
    \eta_t &\sim \mathcal{NID}(0, Q).
\end{align*}
$$

(4)

To facilitate the estimation of the parameters in this model, we use a reparameterization. Details on this technical point can be found in Appendix A.

Below, we describe methods to extract the factors $f_t$ from the macroeconomic variables $x_t$. To facilitate the discussion, define

$$
B^+ = \begin{pmatrix}
    \beta_2' \\
    \beta_3' \\
    \vdots \\
    \beta_T'
\end{pmatrix}, \\
B = \begin{pmatrix}
    \beta_1' \\
    \beta_2' \\
    \vdots \\
    \beta_{T-1}'
\end{pmatrix}, \\
X = \begin{pmatrix}
    x_1' \\
    x_2' \\
    \vdots \\
    x_{T-1}'
\end{pmatrix} \quad \text{and} \quad F = \begin{pmatrix}
    f_1' \\
    f_2' \\
    \vdots \\
    f_{T-1}'
\end{pmatrix},
$$

where $T$ is the length of the estimation sample. To rule out any scale effects, first normalize every column of $X$ (that is, the time series of all observations on each variable separately) to have mean zero and variance unity. In essence, all of the considered methods for constructing factors boil down to choosing a matrix $W$ and setting $F = XW$. That is, the factors that we use in Model (4) are linear combinations of observed macroeconomic aggregates. The following sections describe three methods for choosing these combinations. Common features are that they achieve a dimension reduction ($f_t$ contains much fewer elements than $x_t$) and that the constructed factors $f_{t-1}$, together with $\beta_{t-1}$, have predictive power for $\beta_t$.

One more technical comment is in place at this stage. Each of the factor construction methods described below requires the observability of $B$, while it is, of course, latent. We choose to overcome this problem by the following simple approach. First, we estimate Model (4) without any macro factors. Second, we use the estimates of $\beta_t$ (in Kalman filtering terminology: the smoothed state vectors) to form the response variables in constructing macro factors. Model (4) is then reestimated using these macro factors as $f_t$.

---

2 Including further lags of $\beta_t$ would increase the number of parameters in our model dramatically. Furthermore, [Diebold and Li (2006)] find that one lag is sufficient to describe the dynamics of $\beta_t$. 
Of course, this procedure could be iterated. However, to keep the computational burden at a manageable level, we do not pursue this approach.

2.2 Principal covariate regression

Heij et al. (2007) propose principal covariate regression (PCovR) as an alternative to principal component regression (PCR).

In PCR, principal components are first extracted from a group of predictors and then used as regressors. To obtain $p$ principal components, the predictors in the $n \times k$ matrix $X$ are “summarized” in $p \ll k$ factors by minimizing $||X - XUV||$ over the $k \times p$ matrix $U$ and the $p \times k$ matrix $V$. The desired factors are the columns of $XU$. The method then proceeds as a standard OLS regression of the (univariate) dependent variable $z^+$ on a constant and $XU$. That is, if we denote a vector of ones by $\iota$, the objective function $||z^+ - \alpha - XU\gamma||$ is minimized over the scalar $\alpha$ and the vector $\gamma$. (Here and henceforth, $||M||$ represents the Frobenius norm of a matrix, $(\sum_i \sum_j m_{ij}^2)^{1/2}$. Likewise, $||v||$ represents the $L_2$ norm of a vector, $(\sum_i v_i^2)^{1/2}$.)

Heij et al. (2007) argue that the failure to take the prediction objective into account when constructing the factors is a drawback of PCR. To overcome this problem, they combine the two steps of PCR into one objective function: in the same notation as above, they minimize

$$w \left( ||z^+ - \alpha - XU\gamma||^2 / ||z^+||^2 + (1 - w) ||X - XUV||^2 / ||X||^2 \right), \quad (5)$$

where $w \in [0, 1]$ is a tuning parameter that governs the weight placed on each of the two objectives. (Scale effects are ruled out through the normalizations by $||z^+||^2$ and $||X||^2$, respectively.) Thus, the aims of good prediction and adequate use of the data are balanced in the PCovR objective (5): setting $w$ at a higher value means that more weight is placed on predicting $z^+$, relative to summarizing $X$.

An obvious multivariate extension of Objective (5) to our problem is to minimize

$$w \left( ||B^+ - \iota\alpha' - B\Delta - XUT||^2 / ||B^+||^2 + (1 - w) ||X - XUV||^2 / ||X||^2 \right), \quad (6)$$

where $\alpha$ is now a vector and $\Delta$ and $\Gamma$ are matrices of coefficients, and then to use $XU$ as our macro factors. As the addition of $B$ as predictors makes direct minimization of (5) infeasible, our approach is in two steps: we first perform an auxiliary OLS regression in $B^+ = \iota\alpha' + B\Delta + E$. We save the residuals $R \equiv B^+ - \iota\hat{\alpha}' - B\hat{\Delta}$ and minimize

$$w \left( ||R - XUT||^2 / ||R||^2 + (1 - w) ||X - XUV||^2 / ||X||^2 \right). \quad (7)$$

Direct minimization of Objective (7) can be done using two singular value decompositions, as outlined in Heij et al. (2007). To operationalize this procedure, we need to specify a value for $w$. The high-dimensional nature of $X$ leads to overfitting if $w$
2 METHODS

...is chosen too large; see Heij et al. (2006) for a discussion of this aspect. The upper bound that they derive evaluates to \( w \leq 0.1086 \) for our problem. In order to make PCovR sufficiently different from standard PCR, we do not want to set \( w \) too small either; therefore, we fix \( w = 0.1 \).

2.3 Multiresponse sparse regression

Similä and Tikka (2006) propose the following method for selecting a few informative variables out of a large group. They refer to it as \( L_2 \)-MRSR, or “multiresponse sparse regression using the \( L_2 \) norm”.

Formally, we wish to find a vector \( \alpha \) and matrices \( \Delta \) and \( \Gamma \) such that (1) the difference \( B - \iota \alpha' - B^- \Delta - X \Gamma \) is “small” and (2) \( \Gamma \) has “many” zero rows, so that only a few relevant variables are selected. To this end, we employ a similar two-step procedure as for principal covariate regression: the residuals \( R \) from an OLS regression of \( B \) on a constant and \( B^- \) are used as responses in the \( L_2 \)-MRSR algorithm.

This MRSR algorithm is a multivariate extension of the Least Angle Regression methodology, the latter originating with Efron et al. (2004). The main idea is to “add” regressors one at a time, starting with the regressor that correlates most with the response variable. The regression coefficient is increased from zero, up to the point where the residual is equally correlated with the regressor chosen initially and one other regressor. The coefficients on both regressors are now simultaneously increased in such a way as to keep the residuals equally correlated with the two regressors until a third regressor shows equal correlation, et cetera. This procedure ensures that the coefficients follow a continuous path. It can be stopped at any desired point. Clearly, if we choose such an intermediate point “early” in the path, many coefficients will be equal to zero; hence, Least Angle Regression can be used as a variable selection method. It should be noted that this procedure closely approximates the more well-known Lasso method proposed by Tibshirani (1996); see Efron et al. (2004) for a discussion of this similarity.

In a multivariate setting, an extension of the correlation concept is required. Denote the fitted value of \( R \), based on the first \( k \) regressors chosen, by \( \bar{R}_k \) (with \( \bar{R}_0 = 0 \)), and denote the \( j \)-th column of \( X \) by \( x_{(j)} \). Following Similä and Tikka (2006), the role of the correlation in the description of the univariate case above is now played by \( \| (R - \bar{R}_k)' x_{(j)} \| \), where \( \| \cdot \| \) is again the \( L_2 \) vector norm. No other changes to the procedure are needed. An efficient algorithm to find the order in which variables are added is presented in Similä and Tikka (2006).

As we are not interested in the values of \( \alpha \), \( \Delta \) and \( \Gamma \) per se, we only use the \( L_2 \)-MRSR algorithm to select a predefined number of explanatory variables from a large panel of macroeconomic data. The selected predictors are used as macro factors \( f_t \) in Model (4).
2.4 Hard thresholding

Bai and Ng (2007) propose hard thresholding as a simple method for variable selection. In their univariate setting, consider forecasting a univariate response \( z^+ \) using its own lag \( z \) and the columns of \( X \). To select the most relevant columns, perform the OLS regressions

\[
z^+ = \alpha i + \delta z + \gamma_j x_{(j)} + \xi
\]

for all columns \( j = 1, 2, \ldots, k \). Labeling the \( t \) statistic corresponding to \( \gamma_j \) as \( t_j \), we select a prespecified number of variables from \( X \) for which \( |t_j| \) is largest.

Bai and Ng (2007) suggest that, to construct \( p \) predictors, \( p^* \) variables can be selected, with \( p^* > p \). The first \( p \) principal components extracted from this subset of selected variables are then used as predictors.

As a multivariate extension of Regressions (8), we perform multivariate OLS in

\[
B^+ = \iota \alpha' + B \Delta + x_{(j)} \gamma_j' + \Xi
\]

and compute the Wald statistic \( W_j \) for the test \( \gamma_j = 0 \). The variables with the largest \( W_j \) are selected, and we use principal components extracted from this set as macro factors in Model (4).

---

Note that this Wald test involves a zero restriction on one parameter in each of multiple (in our case, three) interdependent univariate regression equations. The computation of the Wald statistic in this nonstandard situation is outlined in Roy (1957).
3 Empirical results

3.1 Data

Our yield data are unsmoothed Fama-Bliss U.S. Treasury yields for seventeen maturities (ranging from three months to ten years), observed in 408 months (ranging from January 1970 to December 2003). Time series plots of a short-term (three months), a medium-term (two years), and a long-term (ten years) yield are shown in Figure 1. We observe that these yields vary over a wide range of values over the sample period. Moreover, although the yields for different maturities show a large degree of comovement, it can be seen that the spreads (that is, the differences between yields of different maturities quoted at the same date) are also highly variable.

Figure 1: Time series plots of U.S. zero-coupon yields for three selected maturities.

This variability of the yield curve is even more visible in Figure 2, where we show all seventeen yields in four selected months. A “typical” yield curve would be increasing and slightly concave, as is the case in January 1976. However, other shapes, such as humped (1979), decreasing (1981), or almost flat (1990) also occur frequently. As Nelson and Siegel (1987) show, their model can accurately approximate all these different shapes.

---

4For more information on the construction of these data, we refer to Bliss (1997). We thank Michiel de Pooter for making these data available to us.
To obtain predictors, we use monthly data over the same period on 132 major macroeconomic variables, transformed to stationarity. These variables are split into groups in a natural way (forming groups of, among others, “Price Indices”, “Real Output and Income”, and “Interest Rates and Spreads”).

To visualize the connections between the macroeconomy and the yield curve that motivates this field of research, see Figure 3. It appears that long-term yields correlate with a broad class of interest rates, as expected. Short-term yields are correlated with price indices. This connection, well-known in economics, was already hypothesized by Fisher (1930): the observed real interest rate equals a hypothetical nominal interest rate, plus expected inflation.

Also, large spreads can be associated with a high level of real activity. Diebold et al. (2006), who also note and exploit the connections mentioned here, argue that this correlation suggests that spreads are “intimately connected to the cyclical dynamics of the economy”.

---

5 A previous version of this data set is described in Stock and Watson (2002). They also describe the transformations to stationarity. We thank Cem Cakmakli for making a more recent version of these data available to us.
Figure 3: Relations between the macroeconomy and the yield curve.

Notes: This figure shows time series plots of the first principal component from each of three selected macroeconomic groups, and related yields and spreads. Top panel: Price Indices and a short yield; middle panel: Real Output and Income and a short-long spread; bottom panel: Interest Rates and Spreads and a long yield. Note that the location and scale of principal components is arbitrary. The location and scale shown here were chosen for ease of comparison. In our estimation procedure, all principal components are normalized to have mean zero and variance one.
3 EMPIRICAL RESULTS

3.2 Forecasting procedure

We estimate the parameters of Model (4) over a rolling window with a length of 120 months, using our sample from January 1970 to December 2003. At the end of each window (say, time \( t^* \)), predictions are made for times \( t^* + h \) with \( h = 1, 3, 6, 12 \). (Thus, for the forecast horizon \( h = 1 \), our first prediction is for January 1980. For \( h = 3 \), it is for March 1980, et cetera.)

If we denote the predicted yield for maturity \( \tau \) at time \( t + h \), predicted at time \( t \) using method \( M \), by \( \hat{y}_{t+h|t}(M, \tau) \), we can evaluate the goodness of fit by comparing relative sums of squared errors, defined as

\[
RSSE(M, h, \tau) = \frac{\sum_t \left[ \hat{y}_{t+h|t}(M, \tau) - y_{t+h}(\tau) \right]^2}{\sum_t \left[ y_{t+h}(\tau) - \bar{y}(\tau) \right]^2},
\]  

(10)

where \( \bar{y}(\tau) \) is the sample mean over the time period for which we evaluate the RSSE.

The forecasting methods that we consider differ only in the procedure that they employ for constructing macroeconomic factors; they are listed in Table 1.

The method called PCAB (“PCA - best”) requires some explanation. The forecast produced by this method is equal to that from a PCA method, with any of 1, 2, 3, 4 components. The “best” number of components is chosen by comparing the RSSEs, summed over all seventeen maturities \( \tau \), calculated over the estimation window. That is, to choose the number of components at time \( t^* \), the summations in Equation (10) run from \( t = t^* - 120 \) to \( t = t^* - 1 \). (By contrast, in comparing the goodness of fit, these summations run over the entire prediction period.) The idea is that a method that performed well in recent times can be expected to continue to perform well. Observe in particular that different numbers of principal components may be chosen for different forecast horizons.

Note that the groups of variables used in the PCAG and PCOVR methods are chosen such that each group contains one of DRA’s selected variables. This choice is intended to make our results comparable to those in Diebold et al. (2006); of course, the information from other groups could also be used in this manner.

3.3 Forecasting results

The values of \( RSSE(M, h, \tau) \) as defined in Equation (10), for three selected representative maturities (three months, two years and ten years, respectively), can be found in Tables 2—4.\footnote{For windows ending prior to 1990, not all these forecasts have been made. In these cases, we sum over all available forecasts. Thus, in December 1979 (at the end of the first estimation window), no PCAB forecast can be produced, because there are no previous forecasts available to evaluate the RSSEs. In January 1980, the best method is chosen on the basis of the quality of the forecasts made in December 1979 only, et cetera.} Results for maturity \( \tau = 24 \) months are also graphically represented\footnote{Results for the other maturities are available from the author upon request.}.
Table 1: Macro factor construction methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>The Nelson-Siegel model without any macro factors, as in Diebold and Li (2006)</td>
</tr>
<tr>
<td>DRA</td>
<td>Macro variables as in Diebold et al. (2006): capacity utilization, federal funds rate, and an inflation measure</td>
</tr>
<tr>
<td>PCA3</td>
<td>Extract three principal components from the full macro data set, as in de Pooter et al. (2007)</td>
</tr>
<tr>
<td>PCAB</td>
<td>Extract $n$ principal components from the full macro data set, where $n$ is the “best” of 1, 2, 3, 4, as clarified below Equation (10)</td>
</tr>
<tr>
<td>PCAG</td>
<td>One principal component from each of the three groups “Real Output and Income”, “Interest Rates and Spreads”, and “Price Indices”</td>
</tr>
<tr>
<td>PCOVR</td>
<td>One principal covariate from each of the three groups above</td>
</tr>
<tr>
<td>MRSR</td>
<td>The first three macro variables selected by $L_2$-MRSR</td>
</tr>
<tr>
<td>THRES</td>
<td>Three principal components extracted from the forty variables with largest $W_j$, as defined below Equation (9)</td>
</tr>
</tbody>
</table>

in Figures 7—10 in Appendix B, for the other maturities, the results are qualitatively similar.

For all methods, the general trend is as expected: forecasting is easier if the horizon is shorter or the maturity is longer, as can be seen by comparing the first rows in the tables. We observe that simple forecasts made without any macro information are already of very good quality. (For example, for horizon $h = 1$ and maturity $\tau = 3$ months, the reported RSSE is 0.04, meaning that the forecast accounts for 96% of the variability in the true yield.) It appears that the DRA and PCA3 methods perform worse than this benchmark. Even when picking the “best” number of principal components, the forecasts obtained using this method are still inferior to the standard Nelson-Siegel forecasts.

We now turn to the methods we consider in this study. We observe that MRSR also fails to outperform the no-macro model under all circumstances. This result is in line with the view expressed in the Introduction, that selecting only a few variables from the large available data set is wasteful. (Interestingly, MRSR does perform consistently better than DRA’s selected variables. This result shows that if variables are to be selected, this is perhaps best done in a data-driven way.)

---

8These results are not consistent with Diebold et al. (2006) and de Pooter et al. (2007), respectively. A possible explanation is that our rolling-window approach allows the model parameters to evolve more rapidly over time, which could be advantageous for the forecasting power of the low-parametric no-macro model. (Diebold et al. (2006) estimate the parameters based on the full sample, while de Pooter et al. (2007) use an expanding window.)
3 EMPIRICAL RESULTS

Table 2: Sums of squared forecast errors for maturity $\tau = 3$ months, 1980—2003.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td></td>
<td>0.04</td>
<td>0.15</td>
<td>0.24</td>
<td>0.42</td>
</tr>
<tr>
<td>DRA</td>
<td></td>
<td>2.79</td>
<td>2.60</td>
<td>2.62</td>
<td>2.16</td>
</tr>
<tr>
<td>PCA3</td>
<td></td>
<td>1.01</td>
<td>1.05</td>
<td>1.14</td>
<td>1.37</td>
</tr>
<tr>
<td>PCAB</td>
<td></td>
<td>1.07</td>
<td>1.06</td>
<td>1.10</td>
<td>1.27</td>
</tr>
<tr>
<td>PCAG</td>
<td></td>
<td>1.23</td>
<td>1.15</td>
<td>1.19</td>
<td>1.22</td>
</tr>
<tr>
<td>PCOVR</td>
<td></td>
<td><strong>0.93</strong></td>
<td><strong>0.98</strong></td>
<td>1.18</td>
<td>1.28</td>
</tr>
<tr>
<td>MRSR</td>
<td></td>
<td>2.06</td>
<td>1.56</td>
<td>1.65</td>
<td>1.56</td>
</tr>
<tr>
<td>THRES</td>
<td></td>
<td>1.37</td>
<td>1.25</td>
<td>1.46</td>
<td>1.56</td>
</tr>
</tbody>
</table>

Notes: In the first row, we report $RSSE(NO, h, \tau)$, as defined in Equation 10. For ease of comparison, relative values $RSSE(M, h, \tau)/RSSE(NO, h, \tau)$ follow in the remaining rows, for all other methods described in Table 1. For each horizon, the lowest relative RSSE is printed in bold. If a column contains no bold entry, this implies that none of the methods outperform the benchmark of the Nelson-Siegel model without macro factors.

Table 3: Sums of squared forecast errors for maturity $\tau = 24$ months, 1980—2003.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td></td>
<td>0.03</td>
<td>0.11</td>
<td>0.18</td>
<td>0.38</td>
</tr>
<tr>
<td>DRA</td>
<td></td>
<td>2.18</td>
<td>2.38</td>
<td>2.39</td>
<td>1.80</td>
</tr>
<tr>
<td>PCA3</td>
<td></td>
<td>1.02</td>
<td>1.04</td>
<td>1.08</td>
<td>1.27</td>
</tr>
<tr>
<td>PCAB</td>
<td></td>
<td>1.07</td>
<td>1.04</td>
<td>1.06</td>
<td>1.14</td>
</tr>
<tr>
<td>PCAG</td>
<td></td>
<td><strong>0.94</strong></td>
<td><strong>0.94</strong></td>
<td><strong>0.90</strong></td>
<td><strong>0.90</strong></td>
</tr>
<tr>
<td>PCOVR</td>
<td></td>
<td>1.02</td>
<td>1.05</td>
<td>1.23</td>
<td>1.21</td>
</tr>
<tr>
<td>MRSR</td>
<td></td>
<td>1.48</td>
<td>1.38</td>
<td>1.48</td>
<td>1.34</td>
</tr>
<tr>
<td>THRES</td>
<td></td>
<td>1.11</td>
<td><strong>0.93</strong></td>
<td>0.93</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Notes: see Table 2.

Table 4: Sums of squared forecast errors for maturity $\tau = 120$ months, 1980—2003.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td></td>
<td>0.03</td>
<td>0.08</td>
<td>0.15</td>
<td>0.32</td>
</tr>
<tr>
<td>DRA</td>
<td></td>
<td>1.89</td>
<td>2.30</td>
<td>2.29</td>
<td>1.81</td>
</tr>
<tr>
<td>PCA3</td>
<td></td>
<td>1.04</td>
<td>1.12</td>
<td>1.22</td>
<td>1.39</td>
</tr>
<tr>
<td>PCAB</td>
<td></td>
<td>1.08</td>
<td>1.10</td>
<td>1.14</td>
<td>1.25</td>
</tr>
<tr>
<td>PCAG</td>
<td></td>
<td>1.25</td>
<td>1.05</td>
<td><strong>0.88</strong></td>
<td><strong>0.81</strong></td>
</tr>
<tr>
<td>PCOVR</td>
<td></td>
<td>1.04</td>
<td>1.09</td>
<td>1.25</td>
<td>1.26</td>
</tr>
<tr>
<td>MRSR</td>
<td></td>
<td>1.09</td>
<td>1.08</td>
<td>1.12</td>
<td>1.08</td>
</tr>
<tr>
<td>THRES</td>
<td></td>
<td>1.17</td>
<td>1.12</td>
<td>1.12</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Notes: see Table 2.
Of the remaining methods, PCAG, PCOVR and THRES, none consistently outperforms the no-macro benchmark in all situations. However, each of these methods appears to have its strong points. For short maturities and short horizons, PCOVR produces the most reliable forecasts. For medium and long maturities, grouped PCA and (to a lesser degree, for medium-term yields) thresholding are clearly superior.

Looking at the graphed forecast error series, we note that all forecasting methods have more difficulties in forecasting the volatile yields in the 1980s than they have in later years. To see how this fact affects the results described above, we repeat the calculation of relative sums of squared errors for two subsamples, consisting of the years 1980—1989 and 1990—1999, respectively.

These RSSEs are reported in Tables 8—13 in Appendix B. The first notable feature is the difference in performance of the PCA3 and PCAB methods. In the volatile 1980s, the relatively naive method of principal components was clearly insufficient to produce reliable forecasts; on the other hand, in the following low-volatility decade, this procedure performs remarkably well in forecasting the short end of the yield curve.

The good performance of grouped PCA is present in both decades. Apparently, principal component analysis can quite consistently yield better results when given some theory-based “guidance”, especially for the relatively difficult task of forecasting over longer horizons.

For the variable selection methods DRA and MRSR, this subsample analysis does not change the results much. DRA clearly performs worse than all other methods in both decades. MRSR outperforms the no-macro forecasts in a few cases, but the differences in forecasting performance are only marginal.

Both PCOVR and thresholding appear to add more forecasting power to a yields-only model in the 1990s than they do in the 1980s. However, compared to all other methods, PCOVR retains its favorable status only in the 1980s. A tentative explanation is that during a period of high volatility, it is more difficult to decide on how to use the macroeconomic data. As a result, relatively simple procedures like PCA suffice in a low-volatility era; in more volatile times, more advanced methods are needed.

Another point of interest is that part of the squared errors may be due to bias. In some cases (for example, the PCA methods with horizon $h = 12$ months), the error plots in Appendix B suggest that this part is substantial. To quantify this effect, we report the bias $(1/T) \sum_t \left[ \hat{y}_{t+h|t} (M, \tau) - y_{t+h} (\tau) \right]$ in Tables 5—7.

We observe that in most cases, the forecast bias is positive. Over our forecasting period (1980—2003), yields are mostly decreasing. (See also Figure I.) It seems that the Nelson-Siegel model, even when augmented with macroeconomic factors, has some difficulties in anticipating these decreases. Forecasting a year ahead, for many methods this feature leads to the predicted short-term yield being on average a full percentage point too high. Again, like the squared errors, the bias becomes less severe when forecasting yields of longer maturities.

We note that the more advanced methods (PCAG, PCOVR, MRSR, and THRES) appear less susceptible to bias than the other methods. This result is intuitively plausible: a small bias correlates positively (although not perfectly) with a low squared error.
### 3 EMPIRICAL RESULTS

#### Table 5: Forecast bias for maturity $\tau = 3$ months, 1980–2003.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td></td>
<td>0.15</td>
<td>0.30</td>
<td>0.49</td>
<td>0.85</td>
</tr>
<tr>
<td>DRA</td>
<td></td>
<td>0.59</td>
<td>0.94</td>
<td>0.98</td>
<td>0.78</td>
</tr>
<tr>
<td>PCA3</td>
<td></td>
<td>0.09</td>
<td>0.27</td>
<td>0.58</td>
<td>1.16</td>
</tr>
<tr>
<td>PCAB</td>
<td></td>
<td>0.17</td>
<td>0.33</td>
<td>0.55</td>
<td>1.30</td>
</tr>
<tr>
<td>PCAG</td>
<td></td>
<td>0.20</td>
<td>0.40</td>
<td>0.59</td>
<td>0.91</td>
</tr>
<tr>
<td>PCOVR</td>
<td></td>
<td>0.11</td>
<td>0.20</td>
<td>0.30</td>
<td>0.47</td>
</tr>
<tr>
<td>MRSR</td>
<td></td>
<td>0.22</td>
<td>0.29</td>
<td>0.34</td>
<td>0.44</td>
</tr>
<tr>
<td>THRES</td>
<td></td>
<td>0.19</td>
<td>0.39</td>
<td>0.68</td>
<td>1.20</td>
</tr>
</tbody>
</table>

#### Table 6: Forecast bias for maturity $\tau = 24$ months, 1980–2003.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td></td>
<td>0.08</td>
<td>0.20</td>
<td>0.37</td>
<td>0.68</td>
</tr>
<tr>
<td>DRA</td>
<td></td>
<td>0.24</td>
<td>0.41</td>
<td>0.39</td>
<td>0.22</td>
</tr>
<tr>
<td>PCA3</td>
<td></td>
<td>0.05</td>
<td>0.20</td>
<td>0.46</td>
<td>0.95</td>
</tr>
<tr>
<td>PCAB</td>
<td></td>
<td>0.11</td>
<td>0.25</td>
<td>0.45</td>
<td>1.09</td>
</tr>
<tr>
<td>PCAG</td>
<td></td>
<td>−0.00</td>
<td>0.06</td>
<td>0.16</td>
<td>0.42</td>
</tr>
<tr>
<td>PCOVR</td>
<td></td>
<td>0.04</td>
<td>0.09</td>
<td>0.14</td>
<td>0.26</td>
</tr>
<tr>
<td>MRSR</td>
<td></td>
<td>0.10</td>
<td>0.16</td>
<td>0.20</td>
<td>0.28</td>
</tr>
<tr>
<td>THRES</td>
<td></td>
<td>−0.08</td>
<td>−0.01</td>
<td>0.12</td>
<td>0.45</td>
</tr>
</tbody>
</table>

#### Table 7: Forecast bias for maturity $\tau = 120$ months, 1980–2003.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td></td>
<td>0.05</td>
<td>0.12</td>
<td>0.23</td>
<td>0.45</td>
</tr>
<tr>
<td>DRA</td>
<td></td>
<td>0.12</td>
<td>0.18</td>
<td>0.14</td>
<td>−0.01</td>
</tr>
<tr>
<td>PCA3</td>
<td></td>
<td>0.05</td>
<td>0.17</td>
<td>0.35</td>
<td>0.70</td>
</tr>
<tr>
<td>PCAB</td>
<td></td>
<td>0.09</td>
<td>0.18</td>
<td>0.34</td>
<td>0.84</td>
</tr>
<tr>
<td>PCAG</td>
<td></td>
<td>−0.07</td>
<td>−0.09</td>
<td>−0.02</td>
<td>0.20</td>
</tr>
<tr>
<td>PCOVR</td>
<td></td>
<td>−0.03</td>
<td>−0.05</td>
<td>−0.06</td>
<td>−0.01</td>
</tr>
<tr>
<td>MRSR</td>
<td></td>
<td>0.08</td>
<td>0.09</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>THRES</td>
<td></td>
<td>−0.07</td>
<td>−0.16</td>
<td>−0.22</td>
<td>−0.18</td>
</tr>
</tbody>
</table>
As a final point of interest, we also examine the number of principal components selected by the PCAB method. For each of the forecast horizons under consideration, the selected numbers of components are depicted in Figure 4. We observe that the “best” number of components varies over the sample period, but this number is usually stable for prolonged periods. Interestingly, de Pooter et al. (2007) work with three principal components because they find similar results with more and worse results with less components. Our results suggest that these observations may only hold true on shorter periods. For example, it seems that one component is sufficient in the first half of the 1980s, while four components can also be useful from time to time. (Three components are selected, however, during most of the 1990s, when we find the best results for the PCA methods.)

Figure 4: Number of principal components for the PCAB method for each forecast horizon.
3.4 Parameter estimates

We show time series of the estimates of $\lambda$ and $\mu$ in Figures 5 and 6. In these plots, all estimates are dated in the last month of the estimation window.

Typical values of $\lambda$ found in the literature are around 0.06 or 0.08. Except for the MRSR method, our results are on average largely consistent with this view. However, the $\lambda$ estimates vary considerably over time, casting some doubt on studies where $\lambda$ is kept fixed over long time periods. (Examples are Diebold and Li (2006), Diebold et al. (2006), and de Pooter et al. (2007).)

Figure 5: Estimates of the $\lambda$ parameter in Model (4).

Note: No results for PCAB are shown, because the selected $\lambda$ estimate depends on the forecasting horizon; we show the estimates for PCA1—4 instead.

The $\mu$ parameter estimates do not seem to vary much, and their variability is even slightly smaller when macro factors are used than when they are ignored. (This is especially true for the MRSR and PCAG methods, and to a lesser extent for THRES and DRA.) This result has two implications. First, it lends credibility to the stationarity...
assumption: level, slope and curvature of the yield curve do indeed have a well-defined mean to which they revert. Second, it shows how macroeconomic information is useful in the Nelson-Siegel model: it is used to model deviations from the long-run average yield curve, rather than considering every deviation as a mean shift.⁹

Figure 6: Estimates of the $\mu$ parameters in Model (4).

Notes: The curves show the estimated means $\mu$ of level (solid), slope (dashed) and curvature (dotted) of the yield curve. No results for PCAB are shown, because the selected $\mu$ estimates depend on the forecasting horizon; we show the estimates for PCA1—4 instead.

For the other parameters in Model (4), our estimates are comparable to those reported by Diebold et al. (2006) and they do not vary significantly over the methods or over time. The standard deviations $\sigma_i$ are largest for the shortest maturities: on average just over 0.20. Their values are between 0.05 and 0.10 for medium-term yields, and again slightly higher for longer maturities (over five years): around 0.10.

⁹A similar comment is made by Diebold et al. (2006).
EMPIRICAL RESULTS

The transition matrix \( A \) is always estimated close to nonstationarity, with large own-lag coefficients for the level and slope parameters, illustrating the high persistence of yield curves. The blocks of \( A \) governing the macro-to-yields and yields-to-macro feedback relations contain coefficients that are clearly nonzero. Finally, the estimates of the state covariance matrix \( Q \) generally show a pattern with much larger shocks to the curvature factor than to the other factors. The off-diagonal elements take a wide range of values.

As an illustrative example, we provide the parameter estimates found by the PCOVR method over the first window (January 1970—December 1979). The vector of maturities \( \tau \) is included for interpretation of the vector of standard deviations \( \sigma \).

\[
\lambda = 0.08, \quad \mu = (\begin{array}{c}
7.45, -1.02, 0.10
\end{array})',
\]
\[
\sigma = (\begin{array}{cccccccc}
0.13, 0.07, 0.12, 0.12, 0.09, 0.07, 0.07, 0.07, 0.07, 0.07, 0.08, 0.10, 0.10, 0.12, 0.08, 0.06, 0.08, 0.11
\end{array})',
\]
\[
\tau = (3, 6, 9, 12, 15, 18, 21, 24, 30, 36, 48, 60, 72, 84, 96, 108, 120)',
\]
\[
A = \begin{pmatrix}
0.93 & 0.05 & 0.05 & -0.01 & 0.00 & 0.06 \\
0.16 & 0.80 & 0.00 & 0.19 & 0.03 & -0.33 \\
0.08 & 0.06 & 0.61 & 0.10 & -0.06 & 0.04 \\
-0.06 & 0.00 & 0.03 & 0.86 & -0.08 & 0.08 \\
0.10 & 0.02 & -0.02 & 0.26 & 0.68 & -0.20 \\
0.01 & -0.08 & 0.03 & -0.08 & -0.07 & 0.83
\end{pmatrix},
\]
\[
Q = \begin{pmatrix}
0.07 & -0.01 & -0.05 & -0.00 & 0.02 & -0.00 \\
-0.01 & 0.31 & -0.02 & -0.01 & -0.01 & -0.07 \\
-0.05 & -0.02 & 1.30 & 0.01 & -0.05 & -0.10 \\
-0.00 & -0.01 & 0.01 & 0.21 & 0.05 & 0.00 \\
-0.02 & -0.01 & -0.05 & 0.05 & 0.15 & 0.01 \\
-0.00 & -0.07 & -0.10 & 0.00 & 0.01 & 0.06
\end{pmatrix}.
\]

Inspecting these values, we see that the estimate of \( \lambda \) is in agreement with results commonly found in the literature. The same is true for \( \mu \). The level parameter represents the hypothetical yield of infinitely long maturity; it is estimated at almost 7.5%. The slope parameter is negative, which corresponds to an upward-sloping yield curve (this unfortunate notation is due to historical reasons). The curvature parameter is positive but close to zero, indicating a slightly concave yield curve. The estimate of \( \sigma \) generally resembles the pattern described above.

The transition matrix \( A \) has relatively large diagonal entries. Its eigenvalues are \( 0.9690, 0.9072 \pm 0.0860i, 0.7699 \) and \( 0.5794 \pm 0.0001i \). We observe that the first three eigenvalues are quite close to 1 in absolute value, indicating that the model is close to being nonstationary. (A further discussion of stationarity issues can be found in Appendix A.) Another important aspect of these eigenvalues is that four out of six are not real. For the latter pair, the imaginary parts are of negligible magnitudes; for
3 EMPIRICAL RESULTS

the first pair, however, the interpretation is interesting. Complex eigenvalues result in cyclical behavior — as expected in economic data. According to Formula (2.15) in Harvey (1993), the length of the cycle predicted by these eigenvalues would be

\[
\frac{2\pi}{\arctan (0.0860/0.9072)} \approx 66 \text{ months},
\]

which is consistent with the idea of a business cycle.

Finally, note that coefficients in the “feedback blocks” (the two blocks of size $3 \times 3$ that are not on the main diagonal) are generally small, but that some of them are nonetheless different from zero. As an example, the fourth entry in the second row is 0.19. This coefficient governs the feedback from the “Real Output” factor to the slope coefficient of the yield curve. Such feedback corresponds to one of the relations we identified in Section 3.1; it is visualized in the middle panel of Figure 3.

The estimated covariance matrix $Q$, finally, shows that the unexpected shocks to the curvature coefficient are largest (with a variance of 1.30). Also, the off-diagonal elements are generally much smaller than the diagonal elements. The interpretation of this result is that shocks to the different factors are almost independent in this case.\(^{10}\)

\(^{10}\)However, we prefer not to impose a diagonal structure on $Q$. Diebold et al. (2006) tested this restriction and clearly rejected it; moreover, much larger covariances are found in other estimation windows.
4 Conclusions

In this study, we investigated novel ways of incorporating macroeconomic information in the Nelson-Siegel framework for forecasting the yield curve. Comparing the predictive performance of these techniques with methods found in the recent literature, we find that the alternative methods perform better in many cases.

In particular, we find that dividing a pool of macroeconomic information in groups based on economic theory before extracting principal components helps the forecasting performance. This advantage is most clearly present in forecasting medium-term to long-term yields.

Principal covariate regression, instead of regular principal component analysis, is found useful for the traditionally difficult task of forecasting short-term yields. It is worth noting that this technique appears particularly useful in times of high volatility, when other, less advanced methods perform relatively poorly.

In hard thresholding, a sub-pool of the large macroeconomic data set is created based on the individual forecasting power of the variables, before principal components are extracted. This procedure works remarkably well in forecasting yields of medium maturities. Explaining this result is a question left for further research.

The final method investigated in this study is multiresponse sparse regression, where variables are selected based on their joint predictive power. Judging from our results, this method does not seem to be very useful in the particular context of the Nelson-Siegel model.

Our results for methods that were previously investigated in the literature differ from those reported in [Diebold et al.] (2006) and [de Pooter et al.] (2007). We argue that a possible cause for these differences lies in our rolling-window estimation technique. However, more research is needed to clarify these matters.

The results of this study open up interesting research avenues. The factor construction rules proposed here may also prove useful, for example, in augmented affine models for the yield curve, in the style of Mönch (2008). Besides, the importance of forecasting panels of variables in a data-rich environment is in no way limited to yield curve modeling.

To conclude, we find that the methods that produce the best forecasts are two modifications of principal component analysis. Grouped PCA guides the components to reflect economic theory, while principal covariate regression guides the components to reflect the predictive objective. Given our results, finding a way to combine these two guidance rules could be a fruitful research topic.
References


A Parameterization of the Nelson-Siegel model

To facilitate maximum likelihood estimation of the parameters in Model (4), we repa-
rameterize the model as follows:

1. To restrict λ to take only positive values, we optimize over log λ. To understand
this restriction, examine Equation (1): if λ is positive, lim_{τ→∞} y_t (τ) = β_{3t}. By
contrast, if λ < 0, this limit does not exist: as τ → ∞, y_t (τ) → +∞ if β_{3t} ≤ 0
and y_t (τ) → −∞ otherwise. The possibility of such an “explosive” yield curve
is economically uninteresting.

2. The mean vector μ is unrestricted.

3. We want the variances σ^2_i to be positive; thus, log σ_i is used.

4. Likewise, Q is parameterized in terms of the entries of its Cholesky factor Q^{1/2}
to assure a positive definite variance matrix.

5. The matrix A requires more work, because we wish to restrict each of its eigen-
values to lie inside the unit circle in the complex plane. To this end, we per-
form the eigenvalue decomposition A = PDP^{-1}, where D is a diagonal matrix
containing the eigenvalues and P is a full matrix containing the corresponding
eigenvectors.

Note that A has only real entries. This implies that for any complex eigen-
value d_i, the complex conjugate \overline{d_i} is also an eigenvalue. Thus, let us pair
the eigenvalues \((d_1, d_2), (d_3, d_4), \ldots\), such that any pair consists of either two
real numbers or two complex conjugates. (If A is of odd dimension, one eigen-
value will remain unpaired. However, as this eigenvalue is real, its parameter-
ization poses no problems: a standard transformation \(T_1: (-1, 1) \rightarrow \mathbb{R}\) with
\(T_1 (d) = \log \left( \frac{1+d}{1-d} \right) \) is used.)

For each pair of eigenvalues, we compute the sum s and the product p. As a result
of the pairing we chose, s and p are both real numbers. The two eigenvalues are
the roots of the quadratic equation \(d^2 - sd + p = 0\), and it is well known from the
analysis of univariate second-order autoregressive processes that stationarity
corresponds to restricting s and p to the triangle
\[
\{(s, p) : p < 1, p > -s - 1, p > s - 1\}.
\]

11 Otherwise, the vector autoregressive process becomes nonstationary. This feature would obviously
be undesirable from a computational point of view, as the factors would tend to “explode” over time, but
it is also economically not appealing: we use macroeconomic variables that have been transformed to
stationarity (see Section 3), and most economists would agree that an “average yield curve” does exist.

12 Given the eigenvalue-eigenvector pair \((d_i, v_i)\), \(Av_i = d_i v_i\) implies \(A\overline{v_i} = \overline{d_i} v_i\). As A is real, this
simplifies to \(A\overline{v_i} = \overline{d_i} v_i\), yielding the eigenvalue-eigenvector pair \((d_i, \overline{v_i})\).

13 See, for example, Equations (2.23) in Harvey’s 1993.
We can bijectively transform an \((s, p)\) pair to another pair of real numbers, such that the image of this triangle is the unit square, by using \(T_2(s, p) = \left( \frac{1+p}{2}, \frac{1+p-s}{2(1-p)} \right)\). Finally, each of these numbers is transformed to the entire real line in a straightforward way, using \(T_3 : (0, 1) \to \mathbb{R}\) with \(T_3(d) = \log \left( \frac{d}{1-d} \right)\).

Concerning the eigenvectors, only their directions are needed; we normalize the eigenvectors so that their first components are equal to one. For a real eigenvector, all components except the first are saved. For a complex eigenvector, note that its complex conjugate will be another eigenvector. (See Footnote[12]) Thus, saving the real and imaginary parts of one eigenvector allows us to retrieve also the other.

All in all, then, for an \(r \times r\) matrix, let us count the number of parameters we use to identify it. The \(r\) eigenvalues are paired and each pair is parameterized by two real numbers; in case of odd \(r\), the additional eigenvalue requires one additional parameter. Thus, we are using \(r\) parameters for \(r\) eigenvalues. For a real eigenvector of length \(r\), we save \(r - 1\) components. For a pair of complex eigenvectors, we save \(r - 1\) real parts and \(r - 1\) imaginary parts. Adding up, an \(r \times r\) matrix requires \(r + r (r - 1) = r^2\) parameters, as expected.
B  Additional tables and figures

In this appendix, we provide tables of sums of squared forecast errors on the subsamples 1980—1989 and 1990—1999. In addition, time series plots of the forecast errors for maturity $\tau = 24$ months are shown for the entire estimation period 1980—2003. These results are discussed in Section 3.3.

Table 8: Sums of squared forecast errors for maturity $\tau = 3$ months, 1980—1989.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>NO</td>
<td>0.12</td>
</tr>
<tr>
<td>DRA</td>
<td>2.16</td>
</tr>
<tr>
<td>PCA3</td>
<td>1.01</td>
</tr>
<tr>
<td>PCAB</td>
<td>1.09</td>
</tr>
<tr>
<td>PCAG</td>
<td>1.27</td>
</tr>
<tr>
<td>PCOVR</td>
<td><strong>0.92</strong></td>
</tr>
<tr>
<td>MRSR</td>
<td>1.47</td>
</tr>
<tr>
<td>THRES</td>
<td>1.35</td>
</tr>
</tbody>
</table>

Notes: see Table 2.

Table 9: Sums of squared forecast errors for maturity $\tau = 3$ months, 1990—1999.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>NO</td>
<td>0.06</td>
</tr>
<tr>
<td>DRA</td>
<td>7.33</td>
</tr>
<tr>
<td>PCA3</td>
<td><strong>0.56</strong></td>
</tr>
<tr>
<td>PCAB</td>
<td>0.84</td>
</tr>
<tr>
<td>PCAG</td>
<td>0.70</td>
</tr>
<tr>
<td>PCOVR</td>
<td>0.70</td>
</tr>
<tr>
<td>MRSR</td>
<td>5.90</td>
</tr>
<tr>
<td>THRES</td>
<td>1.62</td>
</tr>
</tbody>
</table>

Notes: see Table 2.
### Table 10: Sums of squared forecast errors for maturity $\tau = 24$ months, 1980—1989.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td></td>
<td>0.10</td>
<td>0.31</td>
<td>0.44</td>
<td>0.76</td>
</tr>
<tr>
<td>DRA</td>
<td></td>
<td>1.24</td>
<td>1.27</td>
<td>1.51</td>
<td>1.70</td>
</tr>
<tr>
<td>PCA3</td>
<td></td>
<td>1.02</td>
<td>1.04</td>
<td>1.09</td>
<td>1.28</td>
</tr>
<tr>
<td>PCAB</td>
<td></td>
<td>1.09</td>
<td>1.08</td>
<td>1.13</td>
<td>1.28</td>
</tr>
<tr>
<td>PCAG</td>
<td></td>
<td>0.93</td>
<td>0.94</td>
<td>0.96</td>
<td>0.99</td>
</tr>
<tr>
<td>PCOVR</td>
<td></td>
<td>0.97</td>
<td>0.99</td>
<td>1.29</td>
<td>1.41</td>
</tr>
<tr>
<td>MRSR</td>
<td></td>
<td>1.16</td>
<td>1.17</td>
<td>1.41</td>
<td>1.34</td>
</tr>
<tr>
<td>THRES</td>
<td></td>
<td>1.11</td>
<td>0.99</td>
<td>1.07</td>
<td>1.12</td>
</tr>
</tbody>
</table>

Notes: see Table 2.

### Table 11: Sums of squared forecast errors for maturity $\tau = 24$ months, 1990—1999.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td></td>
<td>0.07</td>
<td>0.34</td>
<td>0.76</td>
<td>1.73</td>
</tr>
<tr>
<td>DRA</td>
<td></td>
<td>5.22</td>
<td>4.43</td>
<td>3.38</td>
<td>2.08</td>
</tr>
<tr>
<td>PCA3</td>
<td></td>
<td>0.90</td>
<td>0.89</td>
<td>0.92</td>
<td>1.02</td>
</tr>
<tr>
<td>PCAB</td>
<td></td>
<td>1.02</td>
<td>0.93</td>
<td>1.04</td>
<td>1.06</td>
</tr>
<tr>
<td>PCAG</td>
<td></td>
<td>1.07</td>
<td>0.98</td>
<td>0.89</td>
<td>0.85</td>
</tr>
<tr>
<td>PCOVR</td>
<td></td>
<td>1.13</td>
<td>1.10</td>
<td>1.06</td>
<td>0.96</td>
</tr>
<tr>
<td>MRSR</td>
<td></td>
<td>2.62</td>
<td>1.72</td>
<td>1.30</td>
<td>0.98</td>
</tr>
<tr>
<td>THRES</td>
<td></td>
<td>1.25</td>
<td>0.75</td>
<td>0.70</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Notes: see Table 2.
### Table 12: Sums of squared forecast errors for maturity $\tau = 120$ months, 1980—1989.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td></td>
<td>0.07</td>
<td>0.22</td>
<td>0.39</td>
<td>0.79</td>
</tr>
<tr>
<td>DRA</td>
<td></td>
<td>1.10</td>
<td>1.25</td>
<td>1.51</td>
<td>1.67</td>
</tr>
<tr>
<td>PCA3</td>
<td></td>
<td>1.01</td>
<td>1.02</td>
<td>1.03</td>
<td>1.10</td>
</tr>
<tr>
<td>PCAB</td>
<td></td>
<td>1.08</td>
<td>1.08</td>
<td>1.07</td>
<td>1.10</td>
</tr>
<tr>
<td>PCAG</td>
<td></td>
<td>1.22</td>
<td>1.05</td>
<td><strong>0.93</strong></td>
<td><strong>0.86</strong></td>
</tr>
<tr>
<td>PCOVR</td>
<td></td>
<td>1.04</td>
<td>1.11</td>
<td>1.37</td>
<td>1.45</td>
</tr>
<tr>
<td>MRSR</td>
<td></td>
<td><strong>0.99</strong></td>
<td><strong>0.97</strong></td>
<td>1.03</td>
<td>1.00</td>
</tr>
<tr>
<td>THRES</td>
<td></td>
<td>1.16</td>
<td>1.13</td>
<td>1.18</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Notes: see Table 2.

### Table 13: Sums of squared forecast errors for maturity $\tau = 120$ months, 1990—1999.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast horizon $h$ (in months)</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td></td>
<td>0.07</td>
<td>0.24</td>
<td>0.49</td>
<td>1.08</td>
</tr>
<tr>
<td>DRA</td>
<td></td>
<td>1.33</td>
<td>1.53</td>
<td>1.45</td>
<td>1.11</td>
</tr>
<tr>
<td>PCA3</td>
<td></td>
<td>1.06</td>
<td>1.23</td>
<td>1.37</td>
<td>1.50</td>
</tr>
<tr>
<td>PCAB</td>
<td></td>
<td>1.07</td>
<td>1.15</td>
<td>1.29</td>
<td>1.53</td>
</tr>
<tr>
<td>PCAG</td>
<td></td>
<td>1.59</td>
<td>1.14</td>
<td><strong>0.82</strong></td>
<td><strong>0.71</strong></td>
</tr>
<tr>
<td>PCOVR</td>
<td></td>
<td>1.04</td>
<td><strong>1.00</strong></td>
<td>0.95</td>
<td>0.80</td>
</tr>
<tr>
<td>MRSR</td>
<td></td>
<td>1.41</td>
<td>1.30</td>
<td>1.17</td>
<td>0.96</td>
</tr>
<tr>
<td>THRES</td>
<td></td>
<td>1.31</td>
<td>1.16</td>
<td>1.04</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Notes: see Table 2.
Figure 7: Forecast errors $\hat{y}_{t+h|t}(\tau) - y_t(\tau)$, forecast horizon $h = 1$, maturity $\tau = 24$ months.
Figure 8: Forecast errors $\hat{y}_{t+h}(\tau) - y_t(\tau)$, forecast horizon $h = 3$, maturity $\tau = 24$ months.
Figure 9: Forecast errors $\hat{y}_{t+h|t}(\tau) - y_{t}(\tau)$, forecast horizon $h = 6$, maturity $\tau = 24$ months.
Figure 10: Forecast errors $\hat{y}_{t+h|t} - y_t(\tau)$, forecast horizon $h = 12$, maturity $\tau = 24$ months.