A Survey of Dynamic Nelson-Siegel Models, Diffusion Indexes, and Big Data Methods for Predicting Interest Rates*

Hal Pedersen[1] **and Norman R. Swanson**[2]

[1]**Conning, Inc. and** [2]**Rutgers University**

January 2019

Abstract

In this paper we survey a number of recent empirical findings regarding the usefulness of including diffusion indexes in dynamic Nelson-Siegel (DNS) type models used to predict the term structure of interest rates (see e.g., Diebold and Li (2007) and Diebold and Rudebusch (2013)). We also survey various empirical methods used in the construction of DNS models, and used to specify and estimate diffusion index augmented DNS models. In particular, we review (sparse) principal component analysis, factor augmented autoregression, and various dimension reduction, variable selection, machine learning, and shrinkage methods, such as the least absolute shrinkage operator (lasso), the elastic net, and independent component analysis, among others. Finally, we discuss the importance of using real-time data in contexts where datasets are subject to revision; and we compare and contrast the use of targeted versus un-targeted specification methods when including diffusion indexes in DNS type prediction models. Interestingly, as noted in Swanson and Xiong (2018a,b), the usefulness of diffusion indexes is crucially dependent upon whether real-time data are used or not. Specifically, when real-time data are used to estimate the weights in diffusion indexes, it is found that relatively few "data rich" models that use big data are preferred to simpler DNS models, post 2010. Instead, pure DNS models that rely only on historical interest rate data deliver mean square error "best" forecasts. However, when data are not real-time, diffusion indexes always have marginal predictive content for interest rates. Moreover, it is clear that in more volatile interest rate regimes, such as prior to 2010, machine learning and related methods have much to offer, regardless of the type of dataset used in their construction.

JEL Classification: C12, C22, C53.

Keywords: Real-time Forecasting, Dynamic Nelson-Siegel Model, Term Structure of Interest Rates, Real-time Dataset, FRED-MD, Real-Time Diffusion Index.

---

* Corresponding Author: Norman R. Swanson, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA, nswanson@econ.rutgers.edu. Hal Pedersen, Conning, 1 Financial Plaza, Hartford, CT 06103, USA, hal.pedersen@conning.com. The authors are grateful to Mingmian Cheng, Valentina Corradi, Eric Ghysels, Yuan Liao, Massimiliano Marcellino, Christian Schumacher, Greg Tkacz, and Xiye Yang for useful comments.

1 Introduction

The term structure of interest rates plays a important role in asset management. One reason for this is that interest rates contain important information for pricing interest rate contingent assets. As a consequence, industries ranging from banking and finance to insurance are interested in forecasting and simulating yields on government, municipal, and corporate bonds. In this paper, we survey two key recent models used for prediction and simulation of interest rates. These include: (1) dynamic Nelson-Siegel (DNS) type models (see e.g., Nelson and Siegel (1985), Svensson (1994), Diebold and Li (2006) and Diebold, Rudebusch and Aruoba (2006), Diebold and Rudebusch (2013), and Swanson and Xiong (2018a,b), and the references cited therein); and (2) factor augmented regression models (see e.g., Stock and Watson (2002a,b), Bai and Ng (2006), and Kim and Swanson (2014,2018). We also discuss hybrid models which combine (1) and (2), and summarize recent empirical findings in which such models are estimated using both "real-time" and "fully revised" datasets. The hybrid models that are referred to above are models in which diffusion indexes are included in DNS type models. This variety of model is particularly interesting because DNS models are often estimated using very small datasets that only include interest rate data, while diffusion indexes are typically constructed using so-called "big-data" (i.e., datasets with potentially hundreds of variables). In this sense, the hybrid models that we survey combine data-rich models with non-data-rich models.

When the objective is prediction (or simulation), as in this paper, an important data related issue related to the availability of the data used to estimate forecasting models arises. This issue involves whether or not to use real-time data in model specification and estimation. In order to understand what is meant by real-time data in this context, consider monthly inflation and interest rates. In the empirical findings surveyed in this paper, monthly interest rates are predicted. Now, historical interest rate data are never revised. Thus, it is clear which interest rate data to use when estimating DNS models that only include historical yields. However, historical inflation rates are regularly revised, and so if one specifies a hybrid DNS type model that includes diffusion indexes that are constructed using inflation (and other revised macroeconomic variables) one must account for the fact that inflation data are subject to revision. Why is this important? The reason is that a diffusion index prediction constructed for the calendar dated time period January 2019, say, using data "pulled" in December 2018, and hence using data available through December 2018, will potentially be different from a diffusion index prediction constructed for the same calendar dated time period (i.e., January 2019), using data that were "pulled" at any point in time after December 2018. Why? In 2019, inflation data for calendar periods prior to and including December 2018 may have been revised by the government, so the timing associated with the "pulling" of data becomes relevant. Namely, data for specific calendar dated time periods is revised periodically, and over time, as the government receives new survey and related information. This in turn results in a revision of the model used to construct diffusion index predictions for calendar dates January 2019 and earlier.

To accurately account for new releases of economic variables, the entire history of the diffusion index in question must potentially be revised each month, as new "estimates" of historical data become available. Now, since the diffusion index is one of the inputs into our hybrid DNS type prediction model of interest rates, real-time data matters. In practice, what this means is that analysts and researchers interested in constructing prediction models that reflect the type of prediction models used in industry must use real-time data, and cannot simply download a single set of explanatory variables for use in ex ante prediction experiments. Stated differently, assume that the objective of a researcher is to simulate the environment faced by an practitioner when constructing interest rate forecasts. Given that the practitioner constructs new forecasts on a regular basis, say monthly, they are faced with an entirely new historical dataset each month. For variables like inflation, money, GDP, and unemployment that are regularly revised, many calendar dated observations that were previously available have been revised. In principle, one can collect an entire time series for each calendar dated observation of a variable, where the elements of the time series denote different "revisions" of the same calendar dated observation, which are successively available, over time. Such datasets, in which case a matrix of data is available for each variable, rather than the usual (time series) vector, are referred to as real-time datasets. It should be stressed, however, that not all variables have this feature, some variables, like interest rates and various asset prices are never revised, and hence are considered "fully revised".

In the context of estimating (hybrid) DNS type prediction models, it is important to decide whether to carry out targeted or un-targeted prediction. Here, "targeting" refers to a specific approach used to construct diffusion indexes. Consider one of the approaches used to construct diffusion indexes in this paper. Namely consider principal components analysis (PCA). Now, diffusion indexes (or latent factors) constructed using PCA can be interpreted as weighted combinations of all of the variables in the large-scale dataset used in their construction. In un-targeted prediction, diffusion indexes included in regression models and DNS type models are usually chosen to be those that explain the maximal correlation across the entire dataset. When using PCA, this amounts to selecting the eigenvectors of a certain correlation matrix that are associated with the largest eigenvalues of a particular eigenvalue-eigenvector decomposition of said matrix. However, there is no guarantee that these eigenvectors have particularly useful information for forecasting interest rates, which are our "target" variable. Namely, achieving maximal correlation with the entire set of variables does not ensure maximal correlation (or predictive content) for a particular variable. In targeted prediction, the variables used in the construction of the diffusion indexes are carefully pre-selected, so as to maximize their predictive content for interest rates. Many methods for doing this are available in the machine learning, dimension reduction, shrinkage, and variable selection literatures. Some such methods that are discussed in this survey include the partial least squares, the lasso, the elastic net, ridge regression, and bootstrap aggregation. Additionally, as alternatives to PCA, we discuss sparse principal components analysis and independent components

analysis. These methods directly induce sparseness in the coefficient vectors used to construct diffusion indexes.

In the empirical part of this survey, we review various empirical findings from Swanson and Xiong (2018a,b). These findings are based on analyses that utilize real-time datasets, as well as analyses that assume all data are fully revised. In the latter case, this amounts to taking a "snapshot" of the data that are available at a particular point in time, and assuming that the historical elements of that data are never revised, even when carrying out real-time forecasting experiments in which DNS hybrid DNS, and factor augmented regression models are re-estimated at each point in time, prior to the construction of each new forecast. As alluded to above, the use of real-time data (or not) has broad implications when specifying and utilizing DNS type models with and without diffusion indexes for forecasting interest rates; and we shall see that empirical findings are indeed dependent upon which type of data are used in model specification and estimation. We also review findings indicating that the use of more sophisticated targeted prediction methods results in hybrid DNS-diffusion index prediction models that outperform pure DNS models prior to around 2010. However, the same is not true after 2010. In particular, after 2010, if real-time data are used in hybrid model construction, then hybrid models are (predictively) outperformed by pure DNS models. On the other hand, when non real-time data are used in a similar "post-2010" empirical exercise, hybrid DNS models still (predictively) outperform pure DNS models. This underscores the importance of carefully selecting the data to be used in DNS models, when simulating and predicting the term structure of interest rates.

The rest of the paper is organized as follows. Section 2 reviews dynamic Nelson Siegel models, and Section 3 discusses modeling with diffusion indexes. In Section 4, factor augmented autoregression is reviewed, and in Section 5, dimension reduction, variable selection, machine learning, and shrinkage methods are surveyed. Section 6 summarizes selected recent empirical evidence concerning the usefulness of diffusion indexes in DNS modelling. Concluding remarks are gathered in Section 7.

2 Dynamic Nelson Siegel Models

In this section, we discuss two representative DNS type models, drawing on the discussion of Swanson and Xiong (2018b). For a complete survey, see Diebold and Rudebusch (2013). Additionally, refer to Diebold and Li (2006) and Diebold, Rudebusch and Aruoba (2006).

First, we discuss the three factor dynamic Nelson-Siegel model. Motivated by rational expectation theory, Nelson and Siegel (1985) express spot interest rates in terms of instantaneous forward rates. Namely, the instantaneous forward interest rate of a bond with maturity m is denoted as $f(m)$, and the spot interest rate of a bond with maturity τ as $y(\tau)$. Then, the yield to maturity of a bond can be written

as the average of forward rates

$$y(\tau) = \frac{1}{\tau}\int_0^\tau f(m)dm.$$

Nelson and Siegel (1985) motivate the use of the following model of the forward rate that can generate monotonically increasing, humped, and occasionally S-shaped yield curves, a range of shapes for yield curves:

$$f(m) = \beta_1 + \beta_2 \cdot \exp(\frac{m}{\theta_t}) + \beta_3 \cdot [(\frac{m}{\theta_t})\exp(\frac{m}{\theta_t})],$$

where $\lambda_t = \frac{1}{\theta_t}$ is the so-called decay parameter, which must be estimated, is assumed fixed in this model, and is time varying in the dynamic version of the model discussed below. It is then easy to derive the following model for bond yields:

$$y(\tau) = \beta_1 + \beta_2 \cdot [\frac{1-\exp(-\frac{\tau}{\theta_t})}{\frac{\tau}{\theta_t}}] + \beta_3 \cdot [\frac{1-\exp(-\frac{\tau}{\theta_t})}{\frac{\tau}{\theta_t}} - \exp(-\frac{\tau}{\theta_t})].$$

In the above model, the latent factors (i.e., the "betas") are fixed. Diebold and Li (2006) generalize this model to allow for time-varying betas: $\beta_{1,t}$, $\beta_{2,t}$ and $\beta_{3,t}$. Their so-called Dynamic Nelson-Siegel (DNS) model is estimated using a two-step procedure. First, the rate of decay λ_t is set to a constant. Next, at each point in time, t, the yield cross section is linearly projected onto the set of factor loadings $(1, \frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau}, \frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau} - \exp(-\lambda_t\tau))$. In our experiments, various different dimensions are considered when specifying the yield cross section. Namely, we consider yield cross sections using 10, 12, and 30 different yield maturities. For example, with our 12-dimensional cross section, we estimate the latent factors by fitting the following regression:

$$\begin{pmatrix} y_t(\tau_1) \\ y_t(\tau_2) \\ y_t(\tau_3) \\ \vdots \\ y_t(\tau_{12}) \end{pmatrix}_{12\times 1} = \begin{pmatrix} 1 & \frac{1-\exp(-\lambda_t\tau_1)}{\lambda_t\tau_1} & \frac{1-\exp(-\lambda_t\tau_1)}{\lambda_t\tau_1} - \exp(-\lambda_t\tau_1) \\ 1 & \frac{1-\exp(-\lambda_t\tau_2)}{\lambda_t\tau_2} & \frac{1-\exp(-\lambda_t\tau_2)}{\lambda_t\tau_2} - \exp(-\lambda_t\tau_2) \\ 1 & \frac{1-\exp(-\lambda_t\tau_3)}{\lambda_t\tau_3} & \frac{1-\exp(-\lambda_t\tau_3)}{\lambda_t\tau_3} - \exp(-\lambda_t\tau_3) \\ \vdots & \vdots & \vdots \\ 1 & \frac{1-\exp(-\lambda_t\tau_{12})}{\lambda_t\tau_{12}} & \frac{1-\exp(-\lambda_t\tau_{12})}{\lambda_t\tau_{12}} - \exp(-\lambda_t\tau_{12}) \end{pmatrix}_{12\times 3} \begin{pmatrix} \beta_{1,t} \\ \beta_{2,t} \\ \beta_{3,t} \end{pmatrix}_{3\times 1}$$

The betas (i.e., $\hat{\beta}_{1,t}$, $\hat{\beta}_{2,t}$, and $\hat{\beta}_{3,t}$) are called the "level", "slope", and "curvature" factors. In particular, note that the loading on $\hat{\beta}_{1,t}$ is one, which is naturally interpreted as the "level" factor. The loading on $\hat{\beta}_{2,t}$ decreases as bond maturity increases, resulting in an increase of the "slope" of bond yield curve. Finally, the loading on the third latent factor, $\hat{\beta}_{3,t}$, starts from zero on the short end of yield curve, reaches its peak at some maturity in the middle, and gradually decays to zero as maturity goes to infinity. Figures 3 exhibits the three NS factors estimated with ordinary least squares for sample period 1988:8 - 2017:10.[1]

[1] An increase in the "level" component, β_{1t}, affects all yields equally, thus it determines the level of the yield curve. Also, as maturity τ goes to infinity, $\beta_{1t} = y_t(\infty)$ by definition. An increase in "slope" component β_{2t} affects short rates more than long rates, thereby changing the slope, or the so-called "term spread" of the yield curve. Finally, an increase in β_{3t}, the "curvature" component, will increase medium-term yields and have little effect on the short and long end of the curve.

In summary, the DNS model can be written as follows:

$$\hat{y}_t(\tau) = \hat{\beta}_{1,t} + \hat{\beta}_{2,t} \cdot [\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau}] + \hat{\beta}_{3,t} \cdot [\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau} - \exp(-\lambda_t\tau)]. \tag{2.1}$$

In order to construct predictions using the DNS model, we fit estimated factors to AR and VAR models, as follows.

$$\hat{\beta}_{i,t+1} = c_i + \gamma_i\hat{\beta}_{i,t} + \epsilon_t \quad i = 1,2,3 \quad \text{or,} \tag{2.2}$$

$$\hat{\boldsymbol{\beta}}_{t+1} = \mathbf{c} + \boldsymbol{\Gamma}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t, \tag{2.3}$$

where ϵ_t is a scalar stochastic disturbance term, $\boldsymbol{\epsilon}_t$ is a 3×1 vector of stochastic disturbance terms, and c_i, $\mathbf{c}$, γ_i, and $\boldsymbol{\Gamma}$, $i = 1, ..., 3$, are conformably defined constants, constant vectors and constant matrices. With these last two models, one con construct predictions of the $\hat{\beta}_{i,t}$, for $i = 1, ..., 3$, which can in turn be inserted into the above model of $\hat{y}_t(\tau)$ in order to generate predictions thereof. In all experiments in the sequel, rolling estimation is carried out when estimating the above models (and all other models), using windows of length 120 months, so that "real-time" predictions are constructed in all cases. Additionally, we consider two types of prediction models. In one, the decay parameter is fixed. In the other, the decay parameter is re-estimated prior to the construction of each new prediction.

Svensson (1994) extends the Nelson-Siegel Svensson (NSS) model by adding a fourth term. This additional term allows for a second "hump" shape in term structures. In particular, he discusses using the following four-factor model for fitting the instantaneous forward interest rate:

$$f(m) = \beta_1 + \beta_2 \cdot \exp(\frac{m}{\theta_{1,t}}) + \beta_3 \cdot [(\frac{m}{\theta_{1,t}}) \cdot \exp(\frac{m}{\theta_{1,t}})] + \beta_4 \cdot [(\frac{m}{\theta_{2,t}}) \cdot \exp(\frac{m}{\theta_{2,t}})].$$

Notice that in the above equation there are now two different decay parameters controlling the double-hump shape of the forward curve, called θ_1 and θ_2. Similar to the DNS model, we consider a dynamic version of the NSS model. Thus, we utilize the following variant of the DNS model (factor estimation and prediction construction is carried out using the DNS modeling approach discussed above).

$$\begin{aligned}\hat{y}_t(\tau) = \hat{\beta}_{1,t} + \hat{\beta}_{2,t} \cdot [\frac{1-\exp(-\lambda_{1,t}\tau)}{\lambda_{1,t}\tau}] + \hat{\beta}_{3,t} \cdot [\frac{1-\exp(-\lambda_{1,t}\tau)}{\lambda_{1,t}\tau} - \exp(-\lambda_{1,t}\tau)] \\ + \hat{\beta}_{4,t} \cdot [\frac{1-\exp(-\lambda_{1,t}\tau)}{\lambda_{2,t}\tau} - \exp(-\lambda_{2,t}\tau)],\end{aligned}$$

where we now have two decay parameters, as discussed above. These are called $\lambda_{1,t}$ and $\lambda_{2,t}$. As discussed in De Pooter (2007), the second hump in the NSS model is difficult to identify without imposing additional restrictions. We adopt his approach to solving this issue, which includes assumptions that the two humps

Therefore, the yield curve will becomes more hump shaped. As demonstrated in Diebold and Li (2006), the "level" factor can be approximated with the 10-year bond yield, the "slope" factor can be approximated with 10-year - 3-month bond yield spreads, and the "curvature" factor moves closely with two times the 2-year yield minus the sum of the 3-month and 10-year yields.

are at least one year apart, and that the second hump reaches its maximum for a maturity which is at least twelve months shorter than the first hump. Additionally, it is assumed that $\lambda_1 \neq \lambda_2$, in order to avoid multicollinearity. Figures 4A - B plots the four NSS factors estimated with static and dynamic decay parameters, $\lambda_{1,t}$ and $\lambda_{2,t}$. Figure 4C plots estimated rates of decay used in the construction of the four Nelson-Siegel-Svensson factors, where the rates of decay $(\lambda_{1,t}, \lambda_{2,t})$ are either set to fixed numbers, or estimated recursively using nonlinear least squares. See Section 3.1.2 for details on model estimation.

For further discussion of DNS models, see Diebold and Rudebusch (2013) and De Pooter (2007). For further discussion comparing arbitrage free dynamic latent factor and DNS models, see Ang and Piazzesi (2003), Diebold, Rudebusch and Aruoba (2006), Christensen, Diebold, and Rudebusch (2011), Duffie (2011), and the references cited therein.

3 Modelling with Diffusion Indexes

Continuing to draw on discussion contained in Swansion and Xiong (2018b), we now turn our attention to so-called diffusion indexes, which have been utilized in numerous recent empirical investigations of economic data (see e.g. Andreou, Gagliardini, Ghysels, and Rubin (2018), Boivin and Ng (2005), Cheng and Hansen (2015), Exerkate, van Dijk, Heij, and Groenen (2013), Ludvigson and Ng (2009), Schumacher (2007,2009), and the references cited therein). A key open question in this literature remains whether or not macroeconomic, financial and other non-yield information is useful in fitting and forecasting the yield curves. As Duffee (2013) points out that assuming yields follow a Markov implies that all information in fundamental economic variables, should already be embedded in yield cross sections. However, many of the aforementioned paper find that so-called "unspanned risks", as proxied for by additional economic variables and/or diffusion indexes contain useful predictive content for yields. For example, Ang and Piazzesi (2003) find that macroeconomic variables are significant for explaining Treasury security yield dynamics; Mönch (2008) shows that including diffusion indexes in an affine Gaussian term structure model results in improved predictive performance; and Diebold, Rudebusch, and Aruoba (2006) discover strong evidence in favor of causal linkages between macroeconomic variables and future yield curve dynamics.

In the above paragraph, we refer to diffusion indexes, which summarize information contained in (potentially) largescale economic datasets (big data). As discussed in the introduction, one important aspect of big data in our context is the use of so-called real-time data. Recently, McCracken and Ng (2016) and St. Louis Federal Reserve Bank's data desk created the FRED-MD, which is a large monthly real-time database that contains over 130 macro-variables and all revisions of all of these variables. The dataset contains variables summarizing economic output and income, labor markets, consumption, money and credit, housing, and stock market, for example. Moreover, they show that diffusion indexes extracted from their FRED-MD dataset contain the same predictive content as diffusion indexes constructed using the classic Stock and Watson dataset (Stock and Watson (2002a,b)). However, the FRED-MD is a real-

time database, while the Stock and Watson dataset contains only fully revised data. Several studies have revealed the importance of collecting and updating such real-time datasets including Diebold and Rudebusch (1991), Hamilton and Perez-Quiros (1996), Bernanke and Boivin (2003), and the papers cited therein.

One of our main objectives in this paper is examining whether diffusion indexes are useful for predicting yields when the data used to construct the indexes are purely "real-time", rather than fully revised. A natural approach for answering this question involves adopting a dynamic factor model framework resembling that used by Coroneo, Giannone and Modugno (2016) and many others. Namely, assume that yields curve factors, (which are the betas in the above discussion are are here called $F_{y,t}$), are driven by both past yield curve factors and macro factors (i.e., diffusion indexes), called $F_{x,t}$. Additionally, it is assumed that macroeconomic variables are driven only by $F_{x,t}$ only. In particular, consider the following model, as discussed in Swanson and Xiong (2018b):

$$\begin{pmatrix} F_{y,t+h} \\ x_t \end{pmatrix} = \begin{pmatrix} c_y \\ c_x \end{pmatrix} + \begin{bmatrix} \Gamma_y & \Gamma_x \\ 0 & \Gamma_{xx} \end{bmatrix} \begin{pmatrix} F_{y,t} \\ F_{x,t} \end{pmatrix} + \begin{pmatrix} e_{y,t+h} \\ e_{x,t} \end{pmatrix},$$

where c_y, c_x are vectors containing constant terms, h is the forecast horizon, Γ_y contains factor loadings on yield factors, Γ_{xx} contains factor loadings on the macro factors, and Γ_x summarizes the marginal effect of macro factors on yield factors. Additionally, $e_{y,t+h}$ and $e_{x,t}$ are idiosyncratic stochastic disturbance terms. In their paper, Coroneo, Giannone and Modungo (2016) use a so-called expectation conditional restricted maximization algorithm for model estimation, and measure the effect of "unspanned" macroeconomic variables (risks) on the yield curve. We use principal component analysis (PCA) for estimating our macro diffusion indexes (i.e., macro factors), following Stock and Watson (2002a,b), and consider various alternative models that utilize macro diffusion indexes. For instance, we examine whether adding macro diffusion indexes to our DNS and NSS models improves the predictive accuracy of these models. Of course, we also consider baseline DNS (or NSS) models that contain only yield factors. More concretely, $h-$step ahead predictions for yield factors are constructed using the following model:

$$\widehat{F}^f_{y,t+h} = \hat{c}_y + \hat{\Gamma}'_y \widehat{F}_{y,t}, \tag{3.1}$$

where $\widehat{F}_{y,t}$ is our estimated DNS (or NSS) latent factor (i.e. $\widehat{F}_{y,t}$ are our betas in the above discussion), $\widehat{F}^f_{y,t+h}$ is our prediction constructed by specifying simple AR(1) or VAR(1) models, $\hat{c}_y$ is an estimate of c_y, and $\hat{\Gamma}_y$ is an estimate of Γ_y. We additionally add the first r_x principle components from a PCA analysis of our real-time dataset, denoted as $\widehat{F}_{x,t}$, to the above prediction model, yielding:

$$\widehat{F}^f_{y,t+h} = \hat{c}_y + \hat{\Gamma}'_y \widehat{F}_{y,t} + \hat{\Gamma}'_x \widehat{F}_{x,t} \tag{3.2}$$

where $\hat{\Gamma}_x$ is an estimate of Γ_x. When predicting yields, in addition to utilizing DNS and NSS models, we also examine whether adding macro diffusion indexes to benchmark AR and VAR models improves

predictive accuracy. In particular, we consider the following model:

$$\begin{pmatrix} y_{t+h} \\ x_t \end{pmatrix} = \begin{pmatrix} c \\ c_x \end{pmatrix} + \begin{bmatrix} \Delta_y & \Delta_x \\ 0 & \Gamma_{xx} \end{bmatrix} \begin{pmatrix} y_t \\ F_{x,t} \end{pmatrix} + \begin{pmatrix} e_{t+h} \\ e_{x,t} \end{pmatrix},$$

where c is the vector containing constant terms, all coefficient matrices (i.e., Δ_y, Δ_x, and Γ_{xx}) are a conformably defined coefficient matrices, Δ_x summarizes the marginal effect of macro diffusion indexes on yields, and e_{t+h} is an idiosyncratic stochastic disturbance term. Summarizing, our focus of interest is on h-step ahead yield predictions constructed using the following model:

$$\hat{y}_{t+h}(\tau) = \hat{c}(\tau) + \hat{\delta}_y' y_t, \tag{3.3}$$

where $\hat{c}(\tau)$ is an estimate of $c(\tau)$, which is an element of c. Also, $\hat{\delta}_y$ is an estimate of δ_y, which is a row vector of Δ_y. y_t contains lags of $y_{t+1}(\tau)$ in autoregressive specifications, and contains lags of y_{t+1} in vector autoregressive specifications. We additionally add the macro diffusion indexes discussed above, F_t^x, to this model, yielding:

$$\hat{y}_{t+h}(\tau) = \hat{c}(\tau) + \hat{\delta}_y' y_t + \hat{\delta}_x' \widehat{F}_{x,t}, \tag{3.4}$$

where $\hat{\delta}_x$ is an estimate of δ_x, which is a row vector of Δ_x. For further discussion of diffusion indexes in macroeconomic forecasting, see Banerjee, Marcellino and Marsten (2008) Boivin and Ng (2005), and Kim and Swanson (2014).

In the following section, we review methods used to estimate diffusion indexes, which are denoted $\widehat{F}_{y,t}$ and $\widehat{F}_{x,t}$ in the above discussion. This is done in the context of constructing forecasts using factor augmented autoregressions. This class of models nests the model outlined above. Our discussion centers around both the specification of the models and their estimation. When discussing estimation, we focus on the use of "un-targeted" prediction, in which case diffusion indexes are extracted from a large dataset, and the diffusion indexes that are utilized in forecasting models are simply those that are the most information rich, in the sense that they explain the largest share of the overall covariance matrix of all of the variables in the dataset. In the context of PCA, this corresponds to using diffusion indexes which are the eigenvectors associated with the largest eigenvalues of an eigenvector-eigenvalue decomposition of the correlation matrix of the variables in the dataset. Needless to say, such an approach does not guarantee that a particular set of variables (i.e., interest rates) can be predicted with optimal precision using diffusion indexes selected this way.

The above arguments suggest that it may, in some cases, be preferable to use "targeted prediction", where "targeting" means that the variables used in forming diffusion indexes are selected in order to maximize their association with the actual variables that are being predicted. In this approach, one might use machine learning, say, in order to "pre-select" a set of relevant variables for inclusion in the analysis used to construct the diffusion indexes. Examples of variable selection type machine learning methods that can be used for targeted prediction include the elastic net, the lasso, and the non-negative

garrote. Before turning to a discussion of such methods, however, we first discuss factor augmented forecasting models and the construction of diffusion indexes.

3.1 Forecasting Using Factor Augmented Autoregressive Models

In addition to utilizing the models discussed in the previous 2 sections, the empirical analysis discussed in this paper draws on some of the most highly touted recent developments in forecasting concerning estimation and asymptotic properties of diffusion indexes based on PCA; and the use of diffusion indexes in the construction of forecasting models. Drawing from the discussion in Swanson and Xiong (2018a), we summarize key features of recent developments by considering static and dynamic factor models in order to motivate the use of diffusion indexes in forecasting. For further discussion, refer to Stock and Watson (2002a,b) and Armah and Swanson (2010a,b)

Let y_{t+h} be the scalar target forecast variable and X_t be an N-dimensional vector of predictor variables, for $t = 1, \ldots, T$. Assume that (y_{t+1}, X_t) has a dynamic factor model representation with $\overline{r}$ common dynamic factors, f_t, which can be written as:

$$y_{t+h} = \beta' W_t + \alpha(L) f_t + \varepsilon_{t+h} \tag{3.5}$$

and

$$x_{it} = \lambda_i(L) f_t + e_{it}, \tag{3.6}$$

for $i = 1, 2, \ldots, N$, where W_t is an $l \times 1$ vector of observable variables with $l << N$, including lags of y_t; $\alpha(L) = \sum_{j=0}^{q} \alpha_j L^j$ and $\lambda_i(L) = \sum_{j=0}^{q} \lambda_{ij} L^j$ are finite order lag polynomials in nonnegative powers of L; and $h > 0$ is the forecast horizon. It is important to note that this framework ensures that all variables in X_t can be expressed as a linear function of the dynamic factors (and an idiosyncratic shock, e_{it}). For a discussion of approximate factor models, in which this condition does not hold, refer to Carrasco and Rossi (2016). Next, wrte write (3.5) and (3.6) in static form as:

$$y_{t+h} = \beta' W_t + \alpha' F_t + \varepsilon_{t+h} \tag{3.7}$$

and

$$x_{it} = \Lambda_i' F_t + e_{it}, \tag{3.8}$$

where $F_t = (f_t', \ldots, f_{t-q}')'$ is an $r \times 1$ vector of static factors, with $r = (q+1)\overline{r}$, α is an $r \times 1$ vector, and $\Lambda_i = (\lambda_{i0}', \ldots, \lambda_{iq}')'$ is a vector of factor loadings on the static factors, where λ_{ij} is an $\overline{r} \times 1$ vector for $j = 0, \ldots, q$ and $\beta = (\beta_1, \ldots, \beta_l)'$. The model in (3.7) is called a factor augmented forecasting model (i.e. see Stock and Watson (2002a,b) and Bai and Ng (2007)). The static factor in (3.8) is thus named because the contemporaneous relationship between x_{it} and F_t. One major advantage of the static representation of the dynamic factor model is it enables one to utilize PCA to estimate the diffusion indexes (factors).

An important theoretical feature of the model in (3.7) is that consistent estimation of the factors in F_t, which can be achieved via simple application of PCA, allows for subsequent $\sqrt{T}$ consistent estimation of α and β in (3.7) using quasi-maximum likelihood, as long as $\sqrt{T}/N \to 0$, as $N, T \to \infty$. Thus, as shown in Bai and Ng (2006), F_t, when estimated using the PCA method outlined in Stock and Watson (2002a,b), can be treated as a vector of observed regressors, eschewing the need to address the generated regressors problem that often arises in applied econometrics. For a discussion of alternative methods for factor forecasting based on estimation of generalized dynamic factor (GDF) models, see Forni, Hallin, Lippi and Reichlin (2005) and Forni, Hallin, Lippi and Zaffaroni (2015). Note also that Boivin and Ng (2005) compare alternative factor based forecast methodologies, and conclude that when the dynamic structure is unknown, the static factor modeling approach of Stock and Watson performs favorably when compared with dynamic factor modeling.

In the literature on factor modeling, many additional question that arise in the context of diffusion index estimation have been addressed. For example, Bai and Ng (2006b) examine whether observable economic variables can serve as proxies for the underlying unobserved factors. In particular, they use a variety of statistics to determine whether a group of observed variables yields the same information as that contained in the latent factors. Armah and Swanson (2010) and Stock and Watson (2002a) also discuss methods for reducing the complexity of estimated diffusion indexes. Stock and Watson (1998,2009) demonstrate that when PCA is used in estimation, factors remain consistent even when there is some time variation in factor loadings and small amounts of data contamination, so long as the number of variables in the panel dataset or the number of predictors is very large (i.e., $N >> T$). The usefulness of factor augmented models that include cointegration restrictions is discussed in Banerjee, Marcellino and Marsten (2014). The importance of assessing and testing for structural breaks in these models is discussed in Banerjee, Marcellino and Marsten (2008), Stock and Watson (2009), and Chen, Dolado and Gonzalo (2014). Factor loading and parameter stability testing is addressed in Corradi and Swanson (2014), Breitung and Eickmeier (2011), Goncalves and Perron (2014), and Han and Inoue (2014). Finally, the empirical and theoretical properties of factor augmented VARMA models are investigated in Dufour and Stevanovic (2013).

Now, consider estimation of the factors appearing in (3.7). Drawing from the discussion in Armah and Swanson (2010a,b) and Swanson and Xiong (2018a), let k ($k < \min\{N, T\}$) be an arbitrary number of factors, Λ^k be $N \times k$ factor loadings matrix, $(\Lambda_1^k, \ldots, \Lambda_N^k)'$, and F^k be the $T \times k$ matrix of factors $(F_1^k, \ldots, F_T^k)'$. From (3.8), estimates of Λ_i^k and F_t^k are obtained by solving the optimization problem:

$$V(k) = \min_{\Lambda^k, F^k} (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \Lambda_i^{k\prime} F_t^k)^2. \tag{3.9}$$

Let $\widetilde{F}^k$ and $\widetilde{\Lambda}^k$ be the minimizers of equation (3.9). Since Λ^k and F^k are not separately identifiable, if $N > T$, a computationally expedient approach would be to concentrate out $\widetilde{\Lambda}^k$ and minimize (3.9) subject

to the normalization $F^{k\prime}F^k/T = I_k$. Minimizing (3.9) is equivalent to maximizing $tr[F^{k\prime}(XX')F^k]$. This optimization is solved by setting $\widetilde{F}^k$ to be the matrix of the k eigenvectors of XX' that correspond to the k largest eigenvalues of XX'. Note that $tr[\cdot]$ represents the matrix trace. Let $\widetilde{D}$ be a $k \times k$ diagonal matrix consisting of the k largest eigenvalues of XX'. The estimated factor matrix, denoted by $\widetilde{F}^k$, is $\sqrt{T}$ times the eigenvectors corresponding to the k largest eigenvalues of the $T \times T$ matrix XX'. Given $\widetilde{F}^k$ and the normalization $F^{k\prime}F^k/T = I_k$, $\widetilde{\Lambda}^{k\prime} = (\widetilde{F}^{k\prime}\widetilde{F}^k)^{-1}\widetilde{F}^{k\prime}X = \widetilde{F}^{k\prime}X/T$ is the corresponding factor loadings matrix.

The solution to the optimization problem in (3.9) is not unique. If $N < T$, it becomes computationally advantageous to concentrate out $\overline{F}^k$ and minimize (3.9) subject to $\overline{\Lambda}^{k\prime}\overline{\Lambda}^k/N = I_k$. This minimization is the same as maximizing $tr[\Lambda^{k\prime}X'X\Lambda^k]$, the solution of which is to set $\overline{\Lambda}^k$ equal to the eigenvectors of the $N \times N$ matrix $X'X$ that correspond to its k largest eigenvalues. One can thus estimate the factors as $\overline{F}^k = X'\overline{\Lambda}^k/N$. $\widetilde{F}^k$ and $\overline{F}^k$ span the same column spaces, hence for forecasting purposes, they can be used interchangeably. Given $\widetilde{F}^k$ and $\widetilde{\Lambda}^k$, let $\widehat{V}(k) = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \widetilde{\Lambda}_i^{k\prime}\widetilde{F}_t^k)^2$ be the sum of squared residuals from regressions of X_i on the k factors, $\forall i$. A penalty function for over fitting, $g(N,T)$, is chosen such that the loss function

$$IC(k) = \log(\widehat{V}(k)) + kg(N,T) \tag{3.10}$$

can consistently estimate r. Let $kmax$ be a bounded integer such that $r \leq k\max$. Bai and Ng (2002) propose three versions of the penalty function $g(N,T)$, namely, $g_1(N,T) = \left(\frac{N+T}{NT}\right)\log\left(\frac{NT}{N+T}\right)$, $g_2(N,T) = \left(\frac{N+T}{NT}\right)\log C_{NT}^2$, and $g_3(N,T) = \left(\frac{\log(C_{NT}^2)}{C_{NT}^2}\right)$, all of which lead to consistent estimation of r. Additional details on the estimation of r are contained in Bai and Ng (2002). Alternative methods for selecting r are discussed in Chen, Huang, and Tu (2010), Onatski (2015), Carrasco and Rossi (2016), and the references cited therein.

In the above discussion, we alluded to the fact that diffusion indexes can be quite complex. For example, diffusion indexes constructed using PCA are linear combinations of every variable in the dataset being used by the applied practitioner. We also mentioned that dimension reduction of a dataset can be accomplished prior to application of PCA, for example, by applying machine learning, shrinkage or other variable selection methods. When such methods are used, the variable selection undertaken can be "targeted" so that only variables useful for predicting the target variable are selected. These variables can then in turn be used to construct diffusion indexes that are potentially much less complex that those constructed by directly applying PCA. In the next section, we discuss a variety of such methods.

4 Machine Learning, Dimension Reduction, Shrinkage, and Variable Selection Methods

In this section, we briefly review select methods in machine learning, dimension reduction, shrinkage, and variable selection methods that are important in economics in general, and in particular for our discussion of the usefulness of "big data" and diffusion indexes in DNS model based forecasting.[2] For forecasters, a key objective when predicting economic variables (such as interest rates) using big data is to remove redundant and irrelevant information from datasets. This is particularly important if the objective is targeted prediction. This problem has historically been be tackled via step-wise regression and ridge regression. However, variables are typically highly correlated in time series applications. Hence, statistical significance tests used in many regression type algorithms suffer from severe size distortion issues. Ghysels, Hill, and Motegi (2017) address this issue by examining multiple parsimonious regressions, each with one key regressor, while jointly accounting for sequential testing problems.

A second solution to the dimension reduction problem with correlated regressors is the use of partial least squares (PLS), which was originally proposed by Herman Wold in the mid 1960s. Broadly speaking, PLS is a latent variable approach to modeling the covariance structure between two sets of variables. One set might be a target variable or variables to be predicted (say Y), while the other might be a very large set of correlated predictor variables, say X. More precisely, the model underlying PLS has

$$\begin{aligned} Y &= F_1 L_1 + E_1 \\ X &= F_2 L_2 + E_2, \end{aligned}$$

where F_1 and F_2 are projection matrices of X and Y; and $L_1$and L_2 are so-called factor loading matrices that operate on the latent factors F_1 and F_2. Additionally, the error terms, E_1 and E_2 are assumed to be identically and independently distributed, and all matrices are conformably defined, given the dimensions of X and Y. In this setup, the decompositions of X and Y maximize the covariance between the latent factors F_1 and F_2.

A third solution uses principle components analysis (PCA), in which latent factors (often called diffusion indexes) are again estimated, but this time via use of an eigenvalue-eigenvector decomposition of the covariance or correlation matrix of the data, for example. Just as in PLS, the objective is to "explain" the data" using a reduced set of (latent) explanatory variables, with the idea being that the useful information in a large set of predictors is often contained in a (much smaller) set of latent factors, which are themselves simply linear combinations of the original variables. A key difference between PCA and PLS is that PLS directly attempts to account for correlation between the target variable and the

[2]For various discussions of big data and its uses in economics, see the 2015 issue of the *Journal of Econometrics* entitled **High Dimensional Problems in Econometrics**.

predictors, while PCA is "unsupervised", in the sense that correlation with any given target variable is not emphasized in the construction of the latent factors. Rather, overall explanation of the entire dataset is the focus of PCA. Needless to say, this particular feature of PCA is of potential concern when targeting (predicting) a specific variable or variables. For this reason, many supervised versions of PCA have been developed. For example, Carrasco and Rossi (2016) use cross validation methods to supervise PCA, while Bai and Ng (2008) consider targeted forecasting using subsets of X (see also Armah and Swanson (2010a,b)) and Cheng, Swanson, And Yang (2017). Given its ease of application as well as recent empirical evidence on its usefulness, PCA (which is the oldest of the methods discussed in this paper; see Spearman (1904) and the discussion in Swanson (2016) for further details), has received the most attention in economics recently, and hence will be discussed in considerably more detail below.

Penalized regression or shrinkage methods, which reduce or shrink redundant or irrelevant variables are also important in big data analysis. Key examples include ridge regression, the lasso, and the elastic net. When viewed through the lens of multivariate regression analysis, all of these methods involve shrinking the magnitude of coefficients in regression models. When the "penalty functions" are carefully designed, and when the "regularization parameters" used to regulate the strength of the penalties in these functions are of sufficient magnitude, then substantial dimension reduction can be achieved. For example, when shrinkage is used in conjunction with PCA, factor loading matrices can be induced to be sparse, in the sense that certain coefficients in the linear combinations of the predictor variables are identically zero. This nice feature imposes parsimony on the number of variables used to form latent factors in PCA, whereas under standard PCA; all predictors receive non-zero weight in each latent factor. Just as in the case of PLS, the number of predictors may be greater than the number of observations in the dataset being analyzed using PCA.

To fix ideas, let's consider the "original" shrinkage estimator. Namely, consider ridge regression and its associated estimator. Assume that we are interested in the following regression model:

$$Y = X\theta + \varepsilon,$$

where Y contains data on a single variable, there are many (possibly highly correlated) variables represented in the data matrix, X, and ε is an error term. Later, we shall introduce the ridge estimator slightly differently, but for now, note that the ridge estimator can be expressed as:

$$\widehat{\theta}_{ridge} = (X'X + \lambda I)^{-1}X'Y.$$

The "ridge" down the diagonal in this estimator is equivalent to adding a penalty of $\lambda \sum_{i=1}^{N} \widehat{\theta}_i^2$ to the usual residual sum of squares term that is minimized in least squares estimation of the above regression model, where N is the number of predictors in X. Here, as $\lambda \to 0$, $\widehat{\theta}_{ridge} \to \widehat{\theta}_{ols}$, and as $\lambda \to \infty$, $\widehat{\theta}_{ridge} \to 0$. Evidently, applying the ridge penalty shrinks parameter estimates towards zero, which increase bias and reduces estimator variance. One very important feature of ridge regression is that invertibility problems

associated with $X'X$ when the number of predictors is too large relative to the number of observations are no longer an issue, and there is always a unique solution (i.e., $\widehat{\theta}_{ridge}$). Other shrinkage estimators that shall be discussed in the sequel include one where the penalty function is $\lambda \sum_{i=1}^{N} \left|\widehat{\theta}_i\right|$ (the lasso) and another that combines both of the above penalty functions (the elastic net).

Another shrinkage estimator is based on bootstrap aggregation (bagging), and was introduced by Breiman (1996). Stock and Watson (2012) note that predictions of Y, at a point in time, $T+1$, conditional on information available up through period T, say $y^f_{T+1|T}$ can be constructed as follows:

$$y^f_{T+1|T} = \sum_{i=1}^{N} \psi(\lambda t_{\widehat{\theta}(i)}) \widehat{\theta}(i) X_T(i),$$

where $X_T(i)$ is the datum on the i^{th} variable in X for period T, $\widehat{\theta}(i)$ is the least squares estimator from regressing $X_{T-1}(i)$ on Y_T, and $\psi(\lambda t_{\widehat{\theta}(i)})$ is a regularized (through λ) function of the t-statistic associated with the aforementioned regression.[3] For bagging $\lambda = 1$, while various Bayesian predictors, including Bayesian model averaging and empirical Bayes can also be formulated in this manner, by setting λ appropriately. Interestingly, Hirano and Wright (2017) show that forecasting models constructed using out-of-sample or split sample schemes perform well only when combined with other methods, such as bagging. Broadly speaking, their results offer a glimpse into the benefits of using state of the art (asymptotic) statistical analysis in order to examine new methods that combine conventional out-of-sample approaches to model selection and estimation with algorithmic approaches such as bagging. In their paper, they show that out-of-sample schemes so regularly used for model selection (and estimation are inefficient when applied in the conventional manner. This finding is reversed when bagging or other risk reduction methods are combined with conventional out-of-sample schemes, however.

As discussed earlier, ongoing research efforts in the study of factor augmented forecasting models include the analysis of problems associated with the "selection" of diffusion indexes that are most useful for predicting y_{t+1}. For example, see Bai and Ng (2008,2009) and Schumacher (2009), who discuss using targeted predictors based on quadratic principal components and thresholding rules for variable subset selection to estimate diffusion indexes. Armah and Swanson (2010a,b) also discuss this issue. Further, Carrasco and Rossi (2016) propose cross validation methods for selecting the "best" diffusion index for use in forecasting). A related area of research, which is the subject of this subsection, is the development of alternative diffusion index estimators, important examples of which use shrinkage methods in order to impose sparseness on the factor loadings used in the construction of diffusion indexes. Two of the

[3] In their setup, Stock and Watson (2012) assume that the predictors are zero mean random orthonormal variables. Also, Y_t is assumed to be zero mean, and the underlying model is assumed to be:

$$Y_t = \theta' X_{t-1} + \varepsilon_t,$$

where ε_t is an error term with fixed variance.

many interesting new estimators in this context include sparse principal components analysis (SPCA) and independent component analysis (ICA).

Zou, Hastie, and Tibshirani (2006) note that diffusion indexes estimated using PCA are linear combinations of all underlying predictor variables, and factor loadings are hence all nonzero, which adversely affects the parsimony of forecasting models, a property known to be important in time series forecasting. Moreover, they stress that diffusion indexes are thus difficult to interpret. In light of this, they propose SPCA, in which the least absolute shrinkage selection operator (lasso) or the related shrinkage estimator called the elastic net is utilized in order to construct principal components with sparse loadings. This is done this by first reformulating PCA as a regression type optimization problem, and then by using a lasso (elastic net) on the coefficients in a suitably constrained regression model.

Before further discussing SPCA, it is worth noting that the lasso and elastic net are important techniques for big data analysis in and of themselves, and are related to the venerable ridge regression estimator. Using the above notation, say that

$$y_t = X_t'\theta + \varepsilon_t.$$

Here, penalized (shrinkage type) regression is carried out as follows: For the ridge estimator, construct:

$$\widehat{\theta}_{ridge} = \arg\min_{\theta} \left\{ \left\| y - \Sigma_{i=1}^{N} X_i \theta_i \right\|^2 + \lambda_2 \Sigma_{i=1}^{N} \theta_i^2 \right\},$$

where y is the $Tx1$ target variable, $X = [X_1, ..., X_N]$, $i = 1, ..., N$ is the TxN predictor matrix, with $X_i = (X_{1,i}, ..., X_{T,i})'$, and $\lambda > 0$ is the tuning parameter. Notice that this is an alternative formulation of $\widehat{\theta}_{ridge}$ to that given earlier. The more recently developed lasso and the elastic net estimators involve imposition of L_1 (lasso) and L_1+L_2 $-norm$ penalties on parameter magnitudes, and are formulated as:

$$\widehat{\theta}_{lasso} = \arg\min_{\theta} \left\{ \left\| y - \Sigma_{i=1}^{N} X_i \theta_i \right\|^2 + \lambda_1 \Sigma_{i=1}^{N} |\theta_i| \right\},$$

and

$$\widehat{\theta}_{elastic\ net} = (1 + \lambda_2) \arg\min_{\theta} \left\{ \left\| y - \Sigma_{i=1}^{N} X_i \theta_i \right\|^2 + \lambda_1 \Sigma_{j=1}^{N} |\theta_j| + \lambda_2 \Sigma_{j=1}^{N} \theta_j^2 \right\}.$$

Interestingly, SPCA follows directly by formulating PCA as a regression-type optimization problem, and then by subsequently imposing lasso (elastic net) constraints on the regression coefficients in the optimization problem. Put simply, factor loading can be recovered by regressing principal components on the N variables in X_t, as shown in Zou, Hastie, and Tibshirani (2006). Here, imposition of the L_2 $-norm$ penalty in ridge regression allows for $N > T$. Moreover, when the lasso or elastic net is utilized in this context, then large enough λ_1 yields sparse $\widehat{\theta}$. In this sense, SPCA is a natural data reduction method. Since the important paper by Zou et al., many authors have proposed modifications to SPCA, as discussed in Kim and Swanson (2017).

Broadly speaking, the lasso and elastic net constitute two of the most important penalized regression methods currently available, in which all predictor variables are retained in a model, but are constrained (regularized) by shrinking them towards zero. For important descriptions of these methods, see Tibshirani (1996), Zou and Hastie (2005), and Zou (2006).

All of the above penalized regression or shrinkage type methods are examples of machine learning. Other machine learning algorithms have also recently been explored in economics. Two examples are bagging and boosting. Bagging (also called bootstrap aggregation) involves first drawing bootstrap samples from an in-sample training dataset, and then constructing predictions, which are later combined. This algorithm is discussed above. Boosting is another so-called machine learning ensemble meta-algorithm algorithm that utilizes a supervised and user-determined set of functions or *learners* (e.g., least square estimators), and uses the set repeatedly on filtered data, which are typically outputs from previous iterations of the learning algorithm. Broadly speaking, boosting isolates which variables, from amongst a large group of variables, are useful for predicting a target variable. More specifically, boosting estimates an unknown function (e.g., the conditional mean) using sequential step-wise forward regression, with learners that may not only be least squares estimators, but may also be smoothing splines and kernel regressions, for example. For further discussion of boosting, see Freund and Schapire (1997), Bai and Ng (2009), Kim and Swanson (2014), and the references therein.

Two further examples include the non-negative garrote (see Breiman (1995) and Yuan and Lin (2007)) and least angle regression (see Efron, Hastie, Johnstone and Tibshirani (2004) and Bai and Ng (2008)), both of which are closely related to the elastic net.

Returning to the main subject of this section, we now discuss independent component analysis, which is predicated on the idea of "opening" the black box in which principal components often reside, and is an alternative to PCA and SPCA. ICA is used in many applications, from brain imaging to stock price return modeling. In all cases, there is a large set of observed individual signals, and it is assumed that each signal depends on several factors, which are unobserved. In this sense, the motivation is exactly the same as that used to justify PCA.

The starting point for ICA is the very simple assumption that the components, say F, are statistically independent in equation (3.7). This assumption is potentially much stronger than the orthogonality imposed under PCA. The key issue in ICA is the measurement of the "level" of independence between components. More specifically, ICA begins with statistically independent (and unobserved) source data, S, which are mixed according to an unknown "mixing matrix", Ω; and X, which is observed, is a mixture of S, weighted by Ω. For simplicity, we assume that the unknown mixing matrix, Ω, is square, although

this assumption can be relaxed. Thus, it is assumed that $X = S\Omega$. Stated differently, assume that:

$$\begin{aligned} X_1 &= \omega_{11}S_1 + \cdots + \omega_{1N}S_N \\ X_2 &= \omega_{21}S_1 + \cdots + \omega_{2N}S_N \\ &\vdots \\ X_N &= \omega_{1N}S_1 + \cdots + \omega_{NN}S_N, \end{aligned} \tag{4.1}$$

where ω_{ij} is the (i,j) element of Ω. Since Ω and S are unobserved, one must estimate the "demixing matrix", Ψ, which transforms the observed X into the independent components, F. That is, $F = X\Psi$, or $F = S\Omega\Psi$. As detailed in Kim and Swanson (2017), if Ω is square, then so is Ψ, and $\Psi = \Omega^{-1}$, so that F is exactly the same as S, and perfect separation occurs. In general, it is only possible to find Ψ such that $\Omega\Psi = PD$, where P is a permutation matrix and D is a diagonal scaling matrix. The independent components, F are latent variables, and are analogous to the principal components discussed in the case of PCA. In summary, upon estimation of Ω and S, it is feasible to estimate the demixing matrix Ψ, and the independent components, F. However (4.1) is not identified unless several assumptions are made. The first assumption is that the sources, S, are statistically independent. Since various sources of information (for example, consumer's behavior, political decisions, etc.) may have an impact on the values of macroeconomic variables, this assumption is not strong. The second assumption is that the signals are stationary. For further details, see Tong, Liu, Soon, Huan (1991). ICA maps the N components of X into the rank N matrix, F. However, we can simply construct factors using up to r $(< N)$ components, without loss of generality, for comparability with PCA. Alternatively, one might carry out ICA using r principal components, hence further filtering diffusion indexes constructed using PCA in order to obtain statistically independent variants thereof (see Stone (2004) for further details). In general, the above model would be more realistic if there were noise terms added. See Hyvärinen and Oja (2000) for a detailed discussion of the noise-free model, and Hyvärinen (1998,1999) for a discussion of the model with noise added.

For a detailed comparison of ICA with PCA, see Kim and Swanson (2016), who note that the main difference between ICA and PCA is in the properties of the factors obtained. Principal components are uncorrelated and have descending variance so that they are naturally ordered in terms of their variances. While setting the diffusion index in equation (3.5) equal to the highest variance (correlation) principal components may well not equate with the specification of the indexes that are most useful for forecasting a given variable, say y_t, it is certainly the case that components explaining the largest share of the variance are often assumed to be the "relevant" ones. For simplicity, consider two observables, $X = (X_1, X_2)$. PCA finds a matrix which transforms X into uncorrelated components $F = (F_1, F_2)$, such that the uncorrelated components have a joint probability density function, $p_F(F)$ with:

$$E(F_1F_2) = E(F_1)\,E(F_2). \tag{4.2}$$

On the other hand, ICA finds a demixing matrix which transforms the observed $X = (X_1, X_2)$ into independent components $F^* = (F_1^*, F_2^*)$, such that the independent components have a joint pdf $p_{F^*}(F^*)$ with:

$$E\left[F_1^{*p} F_2^{*q}\right] = E\left[F_1^{*p}\right] E\left[F_2^{*q}\right], \tag{4.3}$$

for every positive integer value of p and q. Evidently, ICA is more restrictive, and it should thus not be surprising that implementation is much more difficult than PCA, in which estimation is much simpler, since it just involves finding a linear transformation of components which are uncorrelated. Moreover, there is no natural ordering of latent factors in ICA. This is perhaps a blessing in disguise. Namely, as stated above, there is no a priori reason why the ordinal (correlation) ranking of diffusion indexes corresponds to a ranking of their usefulness for predicting y_t (see Kim and Swanson (2014), Bai and Ng (2008) and Carrasco and Rossi (2016) for further discussion of this issue).

Even given all of the recent progress in the area, much remains to be done. There are a vast number of estimators and algorithms than can be utilized for machine learning (we have touched in our discussion on only a very few of these). In the end, what will probably differentiate the "good methods" from the "not so good" is their ability to properly marry the latest tools in statistical inference with the latest algorithmic techniques. For example, step-wise methods now often rely on learning functions and thresholding variables (such as t-statistics) centered around conditional mean type prediction, while there is a clearly a need to fully incorporate conditional or predictive density type prediction in new methods. As another example, recall our earlier discussion on the use of asymptotic analysis to examine the combination of conventional out-of-sample schemes with bootstrap aggregation. Many of these sorts of analyses remain to be done in the context of combining conventional forecasting approaches with state of the art dimension reduction, machine learning, and penalized regression algorithms.

5 Survey of Select Recent Empirical Findings

5.1 Setup

In this section, we survey recent empirical findings reported in Swanson and Xiong (2018a,b). In these two papers, the authors compare interest rate predictions from a variety of different models, including:

(i) Benchmark time series models:

$$y_{t+h}(\tau) = c(\tau) + \delta_y' W_t + \varepsilon_{t+h}, \tag{5.1}$$

where τ denotes the maturity of a bond (bill) for which the scalar, $y_{t+h}(\tau)$, measures the annual yield. Additionally, W_t may contain lags of $y_t(\tau)$ as well as lags of additional explanatory variables, δ_y is a conformably defined coefficient vector, and $c(\tau)$ is a constant term. In all models, up to 5 lags of $y_t(\tau)$

are included, with the number of lags selected using the Schwarz information criterion (SIC). In addition to estimating AR(SIC) and VAR(SIC) models, straw-man AR(1) and VAR(1) models are estimated. Finally, in experiments where VAR models are estimated, W_t includes five bond yields with maturities 3 months, 1 year, 3 years, 5 years, and 10 years.

(ii) DNS and DNSS models:

These models include the dynamic Nelson-Siegel (DNS) and dynamic Nelson-Siegel-Svensson (DNSS) models discussed above. As outlined in Xiong and Swanson (2018b), forecasting model estimation in this context is carried out by estimating latent factors using the following regression:

$$y_t(\tau) = \beta_{1,t} + \beta_{2,t} \cdot [\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau}] + \beta_{3,t} \cdot [\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau} - \exp(-\lambda_t\tau)] + \varepsilon_t, \tag{5.2}$$

where ε_t is a stochastic disturbance term. Forecasts of y_{t+h} are constructed using the model:

$$\widehat{y}_{t+h}(\tau) = \widehat{\beta}^f_{1,t+h} + \widehat{\beta}^f_{2,t+h} \cdot [\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau}] + \widehat{\beta}^f_{3,t+h} \cdot [\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau} - \exp(-\lambda_t\tau)], \tag{5.3}$$

where $y_{t+h}(\tau)$ is a scalar, and $\widehat{\beta}^f_{1,t+h}$, $\widehat{\beta}^f_{2,t+h}$, and $\widehat{\beta}^f_{3,t+h}$ are predictions constructed by specifying AR and VAR models for $\widehat{\beta}_{1,t}$, $\widehat{\beta}_{2,t}$, and $\widehat{\beta}_{3,t}$.

Analogously, estimates of the DNSS factors (i.e. $\beta_{1,t}$, $\beta_{2,t}$, $\beta_{3,t}$, and $\beta_{4,t}$) are constructed at each point in time by regressing $(1, \frac{1-\exp(-\lambda_{1,t}\tau)}{\lambda_{1,t}\tau}, \frac{1-\exp(-\lambda_{1,t}\tau)}{\lambda_{1,t}\tau} - \exp(-\lambda_{1,t}\tau), \frac{1-\exp(-\lambda_{2,t}\tau)}{\lambda_{2,t}\tau} - \exp(-\lambda_{2,t}\tau))$ on $y_t(\tau)$. In this case, thus, the regression model is:

$$\begin{aligned} y_t(\tau) = \beta_{1,t} + \beta_{2,t} \cdot [\frac{1-\exp(-\lambda_{1,t}\tau)}{\lambda_{1,t}\tau}] + \beta_{3,t} \cdot [\frac{1-\exp(-\lambda_{1,t}\tau)}{\lambda_{1,t}\tau} - \exp(-\lambda_{1,t}\tau)] \\ + \beta_{4,t} \cdot [\frac{1-\exp(-\lambda_{2,t}\tau)}{\lambda_{2,t}\tau} - \exp(-\lambda_{2,t}\tau)] + \varepsilon_t. \end{aligned} \tag{5.4}$$

Forecasts of $y_{t+h}(\tau)$ are constructed using:

$$\begin{aligned} \widehat{y}_{t+h}(\tau) = \widehat{\beta}^f_{1,t+h} + \widehat{\beta}^f_{2,t+h} \cdot [\frac{1-\exp(-\lambda_{1,t}\tau)}{\lambda_{1,t}\tau}] + \widehat{\beta}^f_{3,t+h} \cdot [\frac{1-\exp(-\lambda_{1,t}\tau)}{\lambda_{1,t}\tau} - \exp(-\lambda_{1,t}\tau)] \\ + \widehat{\beta}^f_{4,t+h} \cdot [\frac{1-\exp(-\lambda_{2,t}\tau)}{\lambda_{2,t}\tau} - \exp(-\lambda_{2,t}\tau)] \end{aligned} \tag{5.5}$$

where $y_{t+h}(\tau)$ is a scalar, and $\widehat{\beta}^f_{1,t+h}$, $\widehat{\beta}^f_{2,t+h}$, $\widehat{\beta}^f_{3,t+h}$, and $\widehat{\beta}^f_{4,t+h}$ are predictions constructed by specifying AR and VAR models. For complete details, refer to Swanson and Xiong (2018b).

(iii) Hybrid DNS and DNSS models with diffusion indexes:

All of the above models are also estimated with latent factors (i.e., diffusion indexes) added as additional regressors. In particular, for the above benchmark time series models, predictions are constructed using

$$y_{t+h}(\tau) = c(\tau) + \delta'_y W_t + \delta'_x F^x_t + \varepsilon_{t+h}, \tag{5.6}$$

where F^x_t includes either 1, 2 or 3 diffusion indexes, and W_t is defined as above, yielding AR and VAR variants of these models. Here, $c(\tau)$ is a constant term, and δ_y and δ_x are conformably defined

vectors of coefficients. In these models, diffusion indexes, (i.e., F_t^x) are estimated using recursive PCA with both fully revised (see Swanson and Xiong (2018a)) or real-time (see Swanson and Xiong (2018b)) macroeconomic datasets.[4]

When constructing DNS type prediction models, diffusion indexes are included by augmenting the models used to predict $\widehat{\beta}_{i,t+h}^f$, for $i = 1, 2, 3, 4$. Namely, for AR based forecasts, the following prediction models were used:

$$\widehat{\beta}_{i,t+h}^f = \hat{c}_i + \hat{\gamma}_{y,i}' \widehat{\beta}_{i,t} + \hat{\gamma}_{x,i}' F_t^x, \quad \text{for } i = 1, 2, 3, 4$$

where F_t^x includes either 1, 2 or 3 latent factors. All other terms are conformably defined and analogous to our above discussion. We also construct forecasts using the following VAR(1) variant of this model:

$$\widehat{\boldsymbol{\beta}}_{t+h}^f = \hat{c}_y + \hat{\Gamma}_y \widehat{\boldsymbol{\beta}}_t + \hat{\Gamma}_x F_t^x,$$

where $\widehat{\boldsymbol{\beta}}_{t+h}^f = (\widehat{\beta}_{1,t+h}^f, \widehat{\beta}_{2,t+h}^f, \widehat{\beta}_{3,t+h}^f, \widehat{\beta}_{4,t+h}^f)'$, $\hat{c}$ is 4×1 vector, $\hat{\Gamma}_y = (\widehat{\boldsymbol{\gamma}}_1, \widehat{\boldsymbol{\gamma}}_2, \widehat{\boldsymbol{\gamma}}_3, \widehat{\boldsymbol{\gamma}}_4)$, $\widehat{\boldsymbol{\gamma}}_j$ is a 4×1 vector, for $j = 1, 2, 3, 4$, and $\hat{\Gamma}_x$ is a conformably defined matrix of constants.[5]

When comparing the predictive performance of the models detailed below, Swanson and Xiong (2018a,b) report mean square forecast errors (MSFEs), defined as:

$$\text{MSFE}_h(\tau) = \sum_{t=1}^{P} (\hat{y}_{t+h}(\tau) - y_{t+h}(\tau))^2 \tag{5.7}$$

where $\hat{y}_{t+h}(\tau)$ is the h-step-ahead forecast of the Treasury bond yield, with maturity τ. Here, P is the number of ex ante predictions. As alluded to above, all model parameters are estimated with maximum likelihood and PCA; and parameters are updated prior to the construction of each forecast using a rolling window of 120 months of historical data. For an analysis of the use of rolling versus recursive and alternative windowing techniques in the context of forecasting, see Clark and McCracken(2009) and Hansen and Timmermann (2012) and Rossi and Inuoe (2012).

In addition to using un-targeted PCA, Swanson and Xiong (2018b) construct real-time diffusion indexes by implementing machine learning and related techniques to first select a subset of the 130 macroeconomic variables (see discussion below) in their dataset. These "variable subsets" are selected using both the elastic net and the least absolute shrinkage operator, in which ten-fold cross validation is used, in real-time, to estimate tuning parameters in the operators. Then, the "variable subsets" are inputted into PCA in order to construct diffusion indexes (i.e. targeted PCA is carried out).

[4]Different lag specifications were examined in the aforementioned papers, empirical results using only one lag in the above specification were reported on, as one lag yields mean square forecast error "best" models.

[5]For DNS models, only the first three diffusion indexes are used in the above AR and VAR forecasting equations (i.e., $i = 1, 2, 3$), while for DNSS models four diffusion indexes are used in the above AR and VAR forecasting equations (i.e., $i = 1, 2, 3, 4$).

Before turning to a discussion of empirical findings, it is worth noting that the term structure data used are monthly U.S. zero-coupon (end of month) yield curve data reported by the Federal Reserve Board (see *https://www.quandl.com/data/FED/SVENY-US-Treasury-Zero-Coupon-Yield-Curve* and Gürkaynak, Sack and Wright (GSW: 2006)). Swanson and Xiong (2018a,b) utilize GSW monthly data for the August 1988 through October 2017, which contains data on 1 to 30 years maturity bond yields. In addition to GSW zero-yields, 3- and 6-months T-bill yields are utilized in order to "fill-out" the short end of the yield curve. Hence, the authors analyze a panel of dataset containing $N = 32$ dimensional yields and $T = 351$ monthly observations. When constructing betas, the authors consider subsets of either 10, 12, or 30 yields.

Real-time macroeconomic factors (i.e., diffusion indexes) are constructed using PCA, as mentioned above. The dataset used for this is the FRED-MD dataset, which is a real-time monthly database of over 130 macroeconomic time series that covers categories ranging from output and income, to labor market, prices, and interest rates. The FRED-MD dataset is developed and maintained by the Federal Reserve Bank of St. Louis. For details, see McCracken and Ng (2016), and for access to the dataset, visit *https://research.stlouisfed.org/econ/mccracken/fred-databases/*. A key advantage of FRED-MD is that all time series are updated monthly by the Federal Reserve Bank of St. Louis. Thus, researchers have truly real-time data available for conducting forecasting experiments, in which all vintages (revisions) of all variables are available. Use of such data ensures that future information cannot inadvertently be used to revise data from prior periods, which is a serious potential problem with non-real-time or fully revised data. Refer to Stark (2010) for further discussion of real-time datasets. Moreover, fully revised datasets "mix" vintages of observations, in the sense that the most recent observation in a fully revised dataset is a so-called "first release", while earlier calendar dated observations have possibly been revised and re-released many times. Macroeconomic factors are also constructed using fully revised data, resulting in an alternative set of predictions that are not truly real-time.

5.2 Empirical Findings

Key empirical findings can be summarized as follows. First, many of the models that are in the top three "MSFE-best" specifications found by Swanson and Xiong (2018b) include diffusion indexes. Although the majority of these models include only 1 diffusion index, some also include 2 or 3 indexes. Moreover, many of these "top" models are DNS and DNSS specifications. This pattern is quite prevalent across 1-month, 3-month, and 12-month ahead forecast horizons, and for bond maturities including 1, 3, 5, and 10 years. The pattern is least prevalent during the period from 2011-2017, and is more prevalent during periods from 2001-2005 and 2006-2010. Summarizing, Xiong and Swanson (2018b) uncover strong evidence that DNS and DNSS models are useful, as previously found by many authors, and that the usefulness of these models can often be improved by including real-time diffusion indexes.

As just stated, this result is not ubiquitous, however. For example, in the third subsample (post Great Recession), diffusion indexes are only in one MSFE-best model, indicating a deterioration in predictive gains associated with using diffusion indexes, post Great Recession. However, they are still the 2nd and 3rd "best" models, for numerous yield maturities and forecast horizons. Still, it should be stressed that the usefulness of diffusion indexes appears to be somewhat sample dependent. This is not surprising, given the low variability interest rate regime predominating after the Great Recession. For further discussion in the recent literature suggesting that DNS type model performance has deteriorated in recent post credit crisis years, see Altavilla, Giacomini and Ragusa (2017), Diebold, and Rudebusch (2013), and Mönch (2008).

Second, Swanson and Xiong (2018b) present evidence that targeted PCA yields diffusion indexes with more predictive content that those constructed using on un-targeted PCA, for various sample periods and yield maturities; but particularly when $\tau = 10$ years. For example, for the 2001-2005 period, DNSS models with targeted diffusion indexes constructed using the elastic net yield the top 3 MSFE performing models for both 1-month and 3-month ahead predictions. For this sample period, forecasting 1-year and 3-year maturities 1-month ahead also benefits from using shrinkage operators, as in these cases DNS models with diffusion indexes constructed using the elastic net and least absolute shrinkage operator are MSFE-best. However, it should be stressed that there are also many instances where models with diffusion indexes constructed using un-targeted PCA are MSFE-best. Moreover, there are various maturity/forecast horizon/sample period permutations where benchmark time series models are MSFE-best.

Third, when considering forecast combinations consisting of equal weighted averages of various forecasting models, it is found by Swanson and Xiong (2018b) that models with real-time diffusion indexes often significantly outperform a random walk benchmark, across all five maturities and three forecast horizons. The exception to this finding is during the sample period from 2011-2017, where there is a deterioration in predictive gains associated with using diffusion indexes. Essentially, the mechanism that ties unspanned risks to the term-structure, pre-2011, seems to break down in the extremely low interest rate regime post 2010.

Fourth, Xiong and Swanson (2018b) find that combinations that utilize the average of all non-diffusion index type models are MSFE-best in the majority of maturity/forecast horizon/sample period permutations, even beating combinations that include models incorporating real-time diffusion indexes. This result is different from that reported in Swanson and Xiong (2018a), where diffusion indexes appear in almost all MSFE-best combinations. Why is this? The main reason appears to be that Swanson and Xiong (2018a) carry out experiments using a fully revised macroeconomic dataset rather than a real-time macroeconomic dataset. Thus, the use of fully revised data to construct diffusion indexes may have an important confounding effect upon results obtained when carrying out real-time prediction experiments.

Summarizing, DNS and DNSS type forecasting models are predictively accurate, and generally out-

perform benchmark time series models as well as pure diffusion index models. Moreover, in many cases, augmenting DNS and DNSS models to include diffusion indexes constructed either with PCA or targeted PCA yields even more precise predictions.

6 Concluding Remarks

In this paper, we survey recent methods used for predicting the term structure of interest rates using dynamic Nelson-Siegel (DNS), dynamic Nelson-Siegel Svensson (DNSS), and various econometric models. We also survey methods for constructing diffusion indexes using principal component analysis (PCA), as well as various dimension reduction, variable selection, machine learning, and shrinkage methods, all in the context of "mining" big data. We then discuss how diffusion indexes can be used to construct "hybrid" DNS and DNSS prediction models that exhibit good forecasting properties. Finally, we review select recent empirical findings regarding the use of hybrid DNS and DNSS type prediction models that include diffusion indexes constructed using both un-targeted (e.g. PCA) and targeted (e.g. elastic net) methods of variable selection. It is noted that there are many time periods during the last twenty years during which DNS and DNSS models that include diffusion indexes constructed either with PCA or targeted PCA yields result in predictions of the term structure of interest rates that are more precise, in a mean square forecast error sense, than pure DNS/DNSS type models, pure diffusion index models, and various benchmark linear econometric models. However, in recent years, pure DNS/DNSS models yield more precise predictions of the term structure of interest rates than any of the alternative models that we examine.

7 References

Altavilla, C., R. Giacomini and G. Ragusa (2017), Anchoring the Yield Curve Using Survey Expectations, *Journal of Applied Econometrics*, 32, 1055-1068.

Andreou, E., P. Gagliardini, E. Ghysels, and M. Rubin (2018), Is Industrial Production Still the Dominant Factor for the US Economy? Evidence from a New Class of Mixed Frequency (Group) Factor Models, Swiss Finance Institute Research Paper No. 16-11.

Ang, A. and M. Piazzesi (2003), A No-Arbitrage Vector Autoregression of Term Structure Dynamics With Macroeconomic and Latent Variables, *Journal of Monetary Economics*, 50, 745-787.

Bai, J. and S. Ng (2008), Forecasting Economic Time Series Using Targeted Predictors, *Journal of Econometrics*, 146, 304-317.

Bai, J. and S. Ng (2009), Boosting Diffusion Indices, *Journal of Applied Econometrics*, 24, 607-629.

Banerjee, A., M. Marcellino and I. Marsten (2008), Forecasting Macroeconomic Variables Using Diffusion Indexes in Short Samples with Structural Change, in **Forecasting in the Presence of Structural Breaks and Model Uncertainty**, pp. 149-194, Emerald Group Publishing, New York.

Bernanke, B. S. and Boivin, J. (2003), Monetary Policy in a Data-rich Environment, *Journal of Monetary Economics*, 50, 525-546.

BIS (2005), Zero-Coupon Yield Curves: Technical Documentation, Working Paper, Bank of International Settlements.

Boivin, J. and S. Ng (2005), Understanding and Comparing Factor Based Macroeconomic Forecasts, *International Journal of Central Banking*, 1, 117-152.

Carrasco, M. and B. Rossi, (2016), In-Sample Inference and Forecasting in Misspecified Factor Models, *Journal of Business and Economic Statistics*, 34, 313-338.

Cheng, X. and Hansen, B. E. (2015), Forecasting with Factor-Augmented Regression: A Frequentist Model Averaging Approach, *Journal of Econometrics*, 186, 280-293.

Christensen, J.H.E., F.X. Diebold, and G.D. Rudebusch (2011), The Affine Arbitrage Free Class of Nelson-Siegel Term Structure Models, *Journal of Econometrics*, 164, 4-20.

Clark, T. E. and McCracken, M. W. (2009), Improving Forecast Accuracy by Combining Recursive and Rolling Forecast, *International Economic Review*, 50, 363-395.

Corradi, V. and N.R. Swanson (2006), Predictive Density Evaluation, in G. Elliot, C. W. J. Granger, and A. Timmermann, (eds.), **Handbook of Economic Forecasting**, Volume 1, pp. 197-284, Elsevier, Amsterdam.

Corradi, V. and N.R. Swanson (2014), Testing for Structural Stability of Factor Augmented Forecasting Models, *Journal of Econometrics*, 182, 2014, 100-118.

Coroneo, L., D. Giannone and M. Modugno (2016), Unspanned Macroeconomic Factors in the Yield Curve, *Journal of Business and Economic Statistics*, 34, 472-485.

De Pooter, M., (2007), Examining the Nelson-Siegel Class of Term Structure Models: In-Sample Fit versus Out-of-Sample Forecasting Performance, *Tinbergen Institute Discussion Paper*, No. 07-043/4.

Diebold, F.X. and Li, C. (2006), Forecasting the Term Structure of Government Bond Yields, *Journal of Econometrics*, 130, 337-364.

Diebold, F.X. and R.S. Mariano (1995), Comparing Predictive Accuracy, *Journal of Business and Economic Statistics* 13, 253-263.

Diebold, F. X. and Rudebusch, G. D. (1991), Forecasting Output with the Composite Leading Index: A Real-Time Analysis, *Journal of the American Statistical Association*, 86(415), 603-610.

Diebold, F.X., G.D. Rudebusch, and S. B. Aruoba (2006), The Macroeconomy and the Yield Curve: A Dynamic Latent Factor Approach, *Journal of Econometrics*, 131, 309–338.

Diebold, F.X. and G.D. Rudebusch (2013), **Yield Curve Modeling and Forecasting: The Dynamic Nelson-Siegel Approach**, Princeton University Press: Princeton.

Duffee, G.R. (2011), Information In (and Not In) the Term Structure, *Review of Financial Studies*, 24, 2895-2934.

Duffee, G. R. (2013), Forecasting Interest Rates, in **Handbook of Economic Forecasting**, Volume 2, pp. 385-426, Elsevier, Amsterdam.

Exterkate, P., van Dijk, D., Heij, C., and Groenen, P.J. (2013), Forecasting the Yield Curve in a DataRich Environment Using the FactorAugmented NelsonSiegel Model, *Journal of Forecasting*, 32(3), 193-214.

Ghysels, E. and M. Marcellino (2018), **Applied Economic Forecasting using Time Series Models**, Oxford University Press, London.

Gürkaynak, R.S., B. Sack, and J.H. Wright (2006), The U.S. Treasury Yield Curve: 1961 to the Present, Working Paper, Federal Reserve Bank - Finance and Economics Discussion Series 2006-28.

Gürkaynak, R.S. and J.H. Wright (2012), Macroeconomics and the Term Structure, *Journal of Economic Literature*, 50, 331-67.

Hamilton, J. D. and G. Perez-Quiros (1996), What Do the Leading Indicators Lead?, *Journal of Business*, 69, 27-49.

Hamilton, J.D. and J.C. Wu (2012), The Effectiveness of Alternative Monetary Policy Tools in a Zero Lower Bound Environment, *Journal of Money, Credit and Banking*, 44, 3-46.

Hansen, P.R. and A. Timmermann (2012), Choice of Sample Split in Out-of-Sample Forecast Evaluation, Working Paper, Rady School of Management, University of California, San Diego.

Hirano, K. and J.H. Wright (2017), Forecasting With Model Uncertainty: Representations and Risk Reduction, *Econometrica*, 85, 617-643.

Kim, H.H. and N.R. Swanson, (2014), Forecasting Financial and Macroeconomic Variables Using Data Reduction Methods: New Empirical Evidence, *Journal of Econometrics*, 178, 352-367.

Ludvigson, S. C. and S. Ng (2009), Macro Factors in Bond Risk Premia, *Review of Financial Studies*,

22, 5027-5067.

McCracken, M.W. (2000), Robust Out-of-Sample Inference, *Journal of Econometrics*, 99, 195-223.

McCracken, M.W. and S. Ng (2016), Fred-MD: A Monthly Database for Macroeconomic Research, *Journal of Business and Economic Statistics*, 34, 574-589.

Mönch, E. (2008), Forecasting the Yield Curve in a Data-Rich Environment: A No-Arbitrage Factor-Augmented VAR Approach, *Journal of Econometrics*, 146, 26-43.

Nelson, C. and A. Siegel (1985), Parsimonious Nodeling of Yield Curves for US Treasury Bills, Working Paper 1594, National Bureau of Economic Statistics.

Nelson, C. and A. Siegel (1987), Parsimonious Modeling of Yield Curves, *Journal of Business*, 60, 473-489.

Rossi, B. and S. Sekhposyan (2011), Understanding Models' Forecasting Performance, *Journal of Econometrics*, 164, 158-172.

Rossi, B. and A. Inoue (2012), Out-of-Sample Forecast Tests Robust to the Choice of Window Size, *Journal of Business and Economic Statistics*, 30, 432-453.

Rudebusch, G. D. and T. Wu (2008), A Macro-Finance Model of the Term Structure, Monetary Policy and the Economy, *The Economic Journal*, 118, 906-926.

Schumacher, C. (2007), Forecasting German GDP Using Alternative Factor Models Based on Large Datasets, *Journal of Forecasting*, 26, 271-302.

Schumacher, C. (2009), Factor Forecasting Using International Targeted Predictors: The Case of German GDP, *Economics Letters*, 107, 95-98.

Stark, T. (2010), Realistic evaluation of real-time forecasts in the Survey of Professional Forecasters, Special Report, Federal Reserve Bank of Philadelphia Research Division.

Stock, J.H. and M.W. Watson (2002a), Macroeconomic Forecasting Using Diffusion Indexes, *Journal of Business and Economic Statistics*, 20, 147-162.

Stock, J.H. and M.W. Watson (2002b), Forecasting Using Principal Components from a Large Number of Predictors, *Journal of the American Statistical Association*, 97, 1167-1179.

Stock, J. H. and M.W. Watson (2012), Generalized Shrinkage Methods for Forecasting Using Many Predictors, *Journal of Business and Economic Statistics* 30, 481-493.

Svensson, L. E. (1994), Estimating and Interpreting Forward Interest Rates: Sweden 1992-1994, Working Paper 4871, National Bureau of Economic Research.

Swanson, N.R. and Xiong, W. (2018a), Big Data Analytics In Economics: What Have We Learned So Far, And Where Should We Go From Here?, *Canadian Journal of Economics*, 3, 695-746.

Swanson, N.R. and Xiong, W. (2018b), Predicting Interest Rates Using Shrinkage Methods, Real-Time Diffusion Indexes, and Model Combinations, Working Paper, Rutgers University.