Chapter 10

FORECASTING WITH MANY PREDICTORS[*]

JAMES H. STOCK

Department of Economics, Harvard University and the National Bureau of Economic Research

MARK W. WATSON

Woodrow Wilson School and Department of Economics, Princeton University and
the National Bureau of Economic Research

Contents

[*] We thank Jean Boivin, Serena Ng, Lucrezia Reichlin, Charles Whiteman and Jonathan Wright for helpful comments. This research was funded in part by NSF grant SBR-0214131.

Handbook of Economic Forecasting, Volume 1
Edited by Graham Elliott, Clive W.J. Granger and Allan Timmermann

DOI: 10.1016/S1574-0706(05)01010-4

Abstract

Historically, time series forecasts of economic variables have used only a handful of predictor variables, while forecasts based on a large number of predictors have been the province of judgmental forecasts and large structural econometric models. The past decade, however, has seen considerable progress in the development of time series forecasting methods that exploit many predictors, and this chapter surveys these methods. The first group of methods considered is forecast combination (forecast pooling), in which a single forecast is produced from a panel of many forecasts. The second group of methods is based on dynamic factor models, in which the comovements among a large number of economic variables are treated as arising from a small number of unobserved sources, or factors. In a dynamic factor model, estimates of the factors (which become increasingly precise as the number of series increases) can be used to forecast individual economic variables. The third group of methods is Bayesian model averaging, in which the forecasts from very many models, which differ in their constituent variables, are averaged based on the posterior probability assigned to each model. The chapter also discusses empirical Bayes methods, in which the hyperparameters of the priors are estimated. An empirical illustration applies these different methods to the problem of forecasting the growth rate of the U.S. index of industrial production with 130 predictor variables.

Keywords

forecast combining, dynamic factor models, principal components analysis, Bayesian model averaging, empirical Bayes forecasts, shrinkage forecasts

JEL classification: C32, C53, E17

1. Introduction

1.1. Many predictors: Opportunities and challenges

Academic work on macroeconomic modeling and economic forecasting historically has focused on models with only a handful of variables. In contrast, economists in business and government, whose job is to track the swings of the economy and to make forecasts that inform decision-makers in real time, have long examined a large number of variables. In the U.S., for example, literally thousands of potentially relevant time series are available on a monthly or quarterly basis. The fact that practitioners use many series when making their forecasts – despite the lack of academic guidance about how to proceed – suggests that these series have information content beyond that contained in the major macroeconomic aggregates. But if so, what are the best ways to extract this information and to use it for real-time forecasting?

This chapter surveys theoretical and empirical research on methods for forecasting economic time series variables using many predictors, where "many" can number from scores to hundreds or, perhaps, even more than one thousand. Improvements in computing and electronic data availability over the past ten years have finally made it practical to conduct research in this area, and the result has been the rapid development of a substantial body of theory and applications. This work already has had practical impact – economic indexes and forecasts based on many-predictor methods currently are being produced in real time both in the U.S. and in Europe – and research on promising new methods and applications continues.

Forecasting with many predictors provides the opportunity to exploit a much richer base of information than is conventionally used for time series forecasting. Another, less obvious (and less researched) opportunity is that using many predictors might provide some robustness against the structural instability that plagues low-dimensional forecasting. But these opportunities bring substantial challenges. Most notably, with many predictors come many parameters, which raises the specter of overwhelming the information in the data with estimation error. For example, suppose you have twenty years of monthly data on a series of interest, along with 100 predictors. A benchmark procedure might be using ordinary least squares (OLS) to estimate a regression with these 100 regressors. But this benchmark procedure is a poor choice. Formally, if the number of regressors is proportional to the sample size, the OLS forecasts are not first-order efficient, that is, they do not converge to the infeasible optimal forecast. Indeed, a forecaster who only used OLS would be driven to adopt a principle of parsimony so that his forecasts are not overwhelmed by estimation noise. Evidently, a key aspect of many-predictor forecasting is imposing enough structure so that estimation error is controlled (is asymptotically negligible) yet useful information is still extracted. Said differently, the challenge of many-predictor forecasting is to turn dimensionality from a curse into a blessing.

1.2. Coverage of this chapter

This chapter surveys methods for forecasting a single variable using many (n) predictors. Some of these methods extend techniques originally developed for the case that n is small. Small-n methods covered in other chapters in this Handbook are summarized only briefly before presenting their large-n extensions. We only consider linear forecasts, that is, forecasts that are linear in the predictors, because this has been the focus of almost all large-n research on economic forecasting to date.

We focus on methods that can exploit many predictors, where n is of the same order as the sample size. Consequently, we do not examine some methods that have been applied to moderately many variables, a score or so, but not more. In particular, we do not discuss vector autoregressive (VAR) models with moderately many variables [see Leeper, Sims and Zha (1996) for an application with $n = 18$]. Neither do we discuss complex model reduction/variable selection methods, such as is implemented in PC-GETS [see Hendry and Krolzig (1999) for an application with $n = 18$].

Much of the research on linear modeling when n is large has been undertaken by statisticians and biostatisticians, and is motivated by such diverse problems as predicting disease onset in individuals, modeling the effects of air pollution, and signal compression using wavelets. We survey these methodological developments as they pertain to economic forecasting, however we do not discuss empirical applications outside economics. Moreover, because our focus is on methods for forecasting, our discussion of empirical applications of large-n methods to macroeconomic problems other than forecasting is terse.

The chapter is organized by forecasting method. Section 2 establishes notation and reviews the pitfalls of standard forecasting methods when n is large. Section 3 focuses on forecast combining, also known as forecast pooling. Section 4 surveys dynamic factor models and forecasts based on principal components. Bayesian model averaging and Bayesian model selection are reviewed in Section 5, and empirical Bayes methods are surveyed in Section 6. Section 7 illustrates the use of these methods in an application to forecasting the Index of Industrial Production in the United States, and Section 8 concludes.

2. The forecasting environment and pitfalls of standard forecasting methods

This section presents the notation and assumptions used in this survey, then reviews some key shortcomings of the standard tools of OLS regression and information criterion model selection when there are many predictors.

2.1. Notation and assumptions

Let Y_t be the variable to be forecasted and let X_t be the $n \times 1$ vector of predictor variables. The h-step ahead value of the variable to be forecasted is denoted by Y_{t+h}^h.

For example, in Section 7 we consider forecasts of 3- and 6-month growth of the Index of Industrial Production. Let IP_t denote the value of the index in month t. Then the h-month growth of the index, at an annual rate of growth, is

$$Y^h_{t+h} = (1200/h)\ln(IP_{t+h}/IP_t), \tag{1}$$

where the factor $1200/h$ converts monthly decimal growth to annual percentage growth.

A forecast of Y^h_{t+h} at period t is denoted by $Y^h_{t+h|t}$, where the subscript $|t$ indicates that the forecast is made using data through date t. If there are multiple forecasts, as in forecast combining, the individual forecasts are denoted $Y^h_{i,t+h|t}$, where i runs over the m available forecasts.

The many-predictor literature has focused on the case that both X_t and Y_t are integrated of order zero (are $I(0)$). In practice this is implemented by suitable preliminary transformations arrived at by a combination of statistical pretests and expert judgment. In the case of *IP*, for example, unit root tests suggest that the logarithm of *IP* is well modeled as having a unit root, so that the appropriate transformation of *IP* is taking the log first difference (or, for h-step ahead forecasts, the hth difference of the logarithms, as in (1)).

Many of the formal theoretical results in the literature assume that X_t and Y_t have a stationary distribution, ruling out time variation. Unless stated otherwise, this assumption is maintained here, and we will highlight exceptions in which results admit some types of time variation. This limitation reflects a tension between the formal theoretical results and the hope that large-n forecasts might be robust to time variation.

Throughout, we assume that X_t has been standardized to have sample mean zero and sample variance one. This standardization is conventional in principal components analysis and matters mainly for that application, in which different forecasts would be produced were the predictors scaled using a different method, or were they left in their native units.

2.2. *Pitfalls of using standard forecasting methods when n is large*

OLS regression Consider the linear regression model

$$Y_{t+1} = \beta' X_t + \varepsilon_t, \tag{2}$$

where β is the $n \times 1$ coefficient vector and ε_t is an error term. Suppose for the moment that the regressors X_t have mean zero and are orthogonal with $T^{-1}\sum_{t=1}^{T} X_t X_t' = I_n$ (the $n \times n$ identity matrix), and that the regression error is i.i.d. $\mathrm{N}(0, \sigma^2_\varepsilon)$ and is independent of $\{X_t\}$. Then the OLS estimator of the ith coefficient, $\hat{\beta}_i$, is normally distributed, unbiased, has variance σ^2_ε/T, and is distributed independently of the other OLS coefficients. The forecast based on the OLS coefficients is $x'\hat{\beta}$, where x is the $n \times 1$ vector of values of the predictors used in the forecast. Assuming that x and $\hat{\beta}$ are independently distributed, conditional on x the forecast is distributed $\mathrm{N}(x'\beta, (x'x)\sigma^2_\varepsilon/T)$. Because $T^{-1}\sum_{t=1}^{T} X_t X_t' = I_n$, a typical value of X_t is $O_p(1)$, so a typical x vector used to construct a forecast will have norm of order $x'x = O_p(n)$. Thus let $x'x = cn$, where c is a constant. It follows that the forecast $x'\hat{\beta}$ is distributed $\mathrm{N}(x'\beta, c\sigma_\varepsilon^2(n/T))$. Thus, the forecast – which is unbiased under these assumptions – has a forecast error variance that is proportional to n/T. If n is small relative to T, then $E(x'\hat{\beta} - x'\beta)^2$ is small and OLS estimation error is negligible. If, however, n is large relative to T, then the contribution of OLS estimation error to the forecast does not vanish, no matter how large the sample size.

Although these calculations were done under the assumption of normal errors and strictly exogenous regressors, the general finding – that the contribution of OLS estimation error to the mean squared forecast error does not vanish as the sample size increases if n is proportional to T – holds more generally. Moreover, it is straightforward to devise examples in which the mean squared error of the OLS forecast using all the X's exceeds the mean squared error of using no X's at all; in other words, if n is large, using OLS can be (much) worse than simply forecasting Y by its unconditional mean.

These observations do not doom the quest for using information in many predictors to improve upon low-dimensional models; they simply point out that forecasts should not be made using the OLS estimator $\hat{\beta}$ when n is large. As Stein (1955) pointed out, under quadratic risk ($E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)]$), the OLS estimator is not admissible. James and Stein (1960) provided a shrinkage estimator that dominates the OLS estimator. Efron and Morris (1973) showed this estimator to be related to empirical Bayes estimators, an approach surveyed in Section 6 below.

Information criteria Reliance on information criteria, such as the Akaike information criterion (AIC) or Bayes information criterion (BIC), to select regressors poses two difficulties when n is large. The first is practical: when n is large, the number of models to evaluate is too large to enumerate, so finding the model that minimizes an information criterion is not computationally straightforward (however the methods discussed in Section 5 can be used). The second is substantive: the asymptotic theory of information criteria generally assumes that the number of models is fixed or grows at a very slow rate [e.g., Hannan and Deistler (1988)]. When n is of the same order as the sample size, as in the applications of interest, using model selection criteria can reduce the forecast error variance, relative to OLS, but in theory the methods described in the following sections are able to reduce this forecast error variance further. In fact, under certain assumptions those forecasts (unlike ones based on information criteria) can achieve first-order optimality, that is, they are as efficient as the infeasible forecasts based on the unknown parameter vector β.

3. Forecast combination

Forecast combination, also known as forecast pooling, is the combination of two or more individual forecasts from a panel of forecasts to produce a single, pooled forecast. The theory of combining forecasts was originally developed by Bates and Granger

(1969) for pooling forecasts from separate forecasters, whose forecasts may or may not be based on statistical models. In the context of forecasting using many predictors, the n individual forecasts comprising the panel are model-based forecasts based on n individual forecasting models, where each model uses a different predictor or set of predictors.

This section begins with a brief review of the forecast combination framework; for a more detailed treatment, see Chapter 4 in this Handbook by Timmermann. We then turn to various schemes for evaluating the combining weights that are appropriate when n – here, the number of forecasts to be combined – is large. The section concludes with a discussion of the main empirical findings in the literature.

3.1. Forecast combining setup and notation

Let $\{Y^h_{i,t+h|t},\ i = 1, \dots, n\}$ denote the panel of n forecasts. We focus on the case in which the n forecasts are based on the n individual predictors. For example, in the empirical work, $Y^h_{i,t+h|t}$ is the forecast of Y^h_{t+h} constructed using an autoregressive distributed lag (ADL) model involving lagged values of the ith element of X_t, although nothing in this subsection requires the individual forecast to have this structure.

We consider linear forecast combination, so that the pooled forecast is

$$Y^h_{t+h|t} = w_0 + \sum_{i=1}^{n} w_{it} Y^h_{i,t+h|t}, \tag{3}$$

where w_{it} is the weight on the ith forecast in period t.

As shown by Bates and Granger (1969), the weights in (3) that minimize the means squared forecast error are those given by the population projection of Y^h_{t+h} onto a constant and the individual forecasts. Often the constant is omitted, and in this case the constraint $\sum_{i=1}^{n} w_{it} = 1$ is imposed so that $Y^h_{t+h|t}$ is unbiased when each of the constituent forecasts is unbiased. As long as no one forecast is generated by the "true" model, the optimal combination forecast places weight on multiple forecasts. The minimum MSFE combining weights will be time-varying if the covariance matrices of $(Y^h_{t+h|t}, \{Y^h_{i,t+h|t}\})$ change over time.

In practice, these optimal weights are infeasible because these covariance matrices are unknown. Granger and Ramanathan (1984) suggested estimating the combining weights by OLS (or by restricted least squares if the constraints $w_{0t} = 0$ and $\sum_{i=1}^{n} w_{it} = 1$ are imposed). When n is large, however, one would expect regression estimates of the combining weights to perform poorly, simply because estimating a large number of parameters can introduce considerable sampling uncertainty. In fact, if n is proportional to the sample size, the OLS estimators are not consistent and combining using the OLS estimators does not achieve forecasts that are asymptotically first-order optimal. As a result, research on combining with large n has focused on methods which impose additional structure on the combining weights.

Forecast combining and structural shifts Compared with research on combination forecasting in a stationary environment, there has been little theoretical work on forecast combination when the individual models are nonstationary in the sense that they exhibit unstable parameters. One notable contribution is Hendry and Clements (2002), who examine simple mean combination forecasts when the individual models omit relevant variables and these variables are subject to out-of-sample mean shifts, which in turn induce intercept shifts in the individual misspecified forecasting models. Their calculations suggest that, for plausible ranges of parameter values, combining forecasts can offset the instability in the individual forecasts and in effect serves as an intercept correction.

3.2. *Large-n forecast combining methods*[1]

Simple combination forecasts Simple combination forecasts report a measure of the center of the distribution of the panel of forecasts. The equal-weighted, or average, forecast sets $w_{it} = 1/n$. Simple combination forecasts that are less sensitive to outliers than the average forecast are the median and the trimmed mean of the panel of forecasts.

Discounted MSFE weights Discounted MSFE forecasts compute the combination forecast as a weighted average of the individual forecasts, where the weights depend inversely on the historical performance of each individual forecast [cf. Diebold and Pauly (1987); Miller, Clemen and Winkler (1992) use discounted Bates–Granger (1969) weights]. The weight on the ith forecast depends inversely on its discounted MSFE:

$$w_{it} = m_{it}^{-1} \Big/ \sum_{j=1}^{n} m_{jt}^{-1}, \quad \text{where } m_{it} = \sum_{s=T_0}^{t-h} \rho^{t-h-s} \left(Y_{s+h}^h - \hat{Y}_{i,s+h|s}^h\right)^2, \tag{4}$$

where ρ is the discount factor.

Shrinkage forecasts Shrinkage forecasts entail shrinking the weights towards a value imposed a priori which is typically equal weighting. For example, Diebold and Pauly (1990) suggest shrinkage combining weights of the form

$$w_{it} = \lambda \hat{w}_{it} + (1 - \lambda)(1/n), \tag{5}$$

where $\hat{w}_{it}$ is the ith estimated coefficient from a recursive OLS regression of Y_{s+h}^h on $\hat{Y}_{1,s+h|s}^h, \ldots, \hat{Y}_{n,s+h|s}^h$ for $s = T_0, \ldots, t - h$ (no intercept), where T_0 is the first date for the forecast combining regressions and where λ controls the amount of shrinkage towards equal weighting. Shrinkage forecasts can be interpreted as a partial implementation of Bayesian model averaging (see Section 5).

[1] This discussion draws on Stock and Watson (2004a).

Time-varying parameter weights Time-varying parameter (TVP) weighting allows the weights to evolve as a stochastic process, thereby adapting to possible changes in the underlying covariances. For example, the weights can be modeled as evolving according to the random walk, $w_{it} = w_{it+1} + \eta_{it}$, where η_{it} is a disturbance that is serially uncorrelated, uncorrelated across i, and uncorrelated with the disturbance in the forecasting equation. Under these assumptions, the TVP combining weights can be estimated using the Kalman filter. This method is used by Sessions and Chatterjee (1989) and by LeSage and Magura (1992). LeSage and Magura (1992) also extend it to mixture models of the errors, but that extension did not improve upon the simpler Kalman filter approach in their empirical application.

A practical difficulty that arises with TVP combining is the determination of the magnitude of the time variation, that is, the variance of η_{it}. In principle, this variance can be estimated, however estimation of $\text{var}(\eta_{it})$ is difficult even when there are few regressors [cf. Stock and Watson (1998)].

Data requirements for these methods An important practical consideration is that these methods have different data requirements. The simple combination methods use only the contemporaneous forecasts, so forecasts can enter and leave the panel of forecasts. In contrast, methods that weight the constituent forecasts based on their historical performance require a historical track record for each forecast. The discounted MSFE methods can be implemented if there is historical forecast data, but the forecasts are available over differing subsamples (as would be the case if the individual X variables become available at different dates). In contrast, the TVP and shrinkage methods require a complete historical panel of forecasts, with all forecasts available at all dates.

3.3. Survey of the empirical literature

There is a vast empirical literature on forecast combining, and there are also a number of simulation studies that compare the performance of combining methods in controlled experiments. These studies are surveyed by Clemen (1989), Diebold and Lopez (1996), Newbold and Harvey (2002), and in Chapter 4 of this Handbook by Timmermann. Almost all of this literature considers the case that the number of forecasts to be combined is small, so these studies do not fall under the large-n brief of this survey. Still, there are two themes in this literature that are worth noting. First, combining methods typically outperform individual forecasts in the panel, often by a wide margin. Second, simple combining methods – the mean, trimmed mean, or median – often perform as well as or better than more sophisticated regression methods. This stylized fact has been called the “forecast combining puzzle”, since extant statistical theories of combining methods suggest that in general it should be possible to improve upon simple combination forecasts.

The few forecast combining studies that consider large panels of forecasts include Figlewski (1983), Figlewski and Urich (1983), Chan, Stock and Watson (1999), Stock and Watson (2003, 2004a), Kitchen and Monaco (2003), and Aiolfi and Timmermann (2004). The studies by Figlewski (1983) and Figlewski and Urich (1983) use static factor models for forecast combining; they found that the factor model forecasts improved equal-weighted averages in one instance ($n = 33$ price forecasts) but not in another ($n = 20$ money supply forecasts). Further discussion of these papers is deferred to Section 4. Stock and Watson (2003, 2004b) examined pooled forecasts of output growth and inflation based on panels of up to 43 predictors for each of the G7 countries, where each forecast was based on an autoregressive distributed lag model with an individual X_t. They found that several combination methods consistently improved upon autoregressive forecasts; as in the studies with small n, simple combining methods performed well, in some cases producing the lowest mean squared forecast error. Kitchen and Monaco (2003) summarize the real time forecasting system used at the U.S. Treasury Department, which forecasts the current quarter's value of GDP by combining ADL forecasts made using 30 monthly predictors, where the combination weights depend on relative historical forecasting performance. They report substantial improvement over a benchmark AR model over the 1995–2003 sample period. Their system has the virtue of readily permitting within-quarter updating based on recently released data. Aiolfi and Timmermann (2004) consider time-varying combining weights which are nonlinear functions of the data. For example, they allow for instability by recursively sorting forecasts into reliable and unreliable categories, then computing combination forecasts with categories. Using the Stock–Watson (2003) data set, they report some improvements over simple combination forecasts.

4. Dynamic factor models and principal components analysis

Factor analysis and principal components analysis (PCA) are two longstanding methods for summarizing the main sources of variation and covariation among n variables. For a thorough treatment for the classical case that n is small, see Anderson (1984). These methods were originally developed for independently distributed random vectors. Factor models were extended to dynamic factor models by Geweke (1977), and PCA was extended to dynamic principal components analysis by Brillinger (1964).

This section discusses the use of these methods for forecasting with many predictors. Early applications of dynamic factor models (DFMs) to macroeconomic data suggested that a small number of factors can account for much of the observed variation of major economic aggregates [Sargent and Sims (1977), Stock and Watson (1989, 1991), Sargent (1989)]. If so, and if a forecaster were able to obtain accurate and precise estimates of these factors, then the task of forecasting using many predictors could be simplified substantially by using the estimated dynamic factors for forecasting, instead of using all n series themselves. As is discussed below, in theory the performance of estimators of the factors typically improves as n increases. Moreover, although factor analysis and PCA differ when n is small, their differences diminish as n increases; in fact, PCA (or dynamic PCA) can be used to construct consistent estimators of the factors in DFMs. These observations have spurred considerable recent interest in economic forecasting using the twin methods of DFMs and PCA.

This section begins by introducing the DFM, then turns to algorithms for estimation of the dynamic factors and for forecasting using these estimated factors. The section concludes with a brief review of the empirical literature on large-n forecasting with DFMs.

4.1. The dynamic factor model

The premise of the dynamic factor model is that the covariation among economic time series variables at leads and lags can be traced to a few underlying unobserved series, or factors. The disturbances to these factors might represent the major aggregate shocks to the economy, such as demand or supply shocks. Accordingly, DFMs express observed time series as a distributed lag of a small number of unobserved common factors, plus an idiosyncratic disturbance that itself might be serially correlated:

$$X_{it} = \lambda_i(\mathrm{L})' f_t + u_{it}, \quad i = 1, \ldots, n, \tag{6}$$

where f_t is the $q \times 1$ vector of unobserved factors, $\lambda_i(\mathrm{L})$ is a $q \times 1$ vector lag polynomial, called the "dynamic factor loadings", and u_{it} is the idiosyncratic disturbance. The factors and idiosyncratic disturbances are assumed to be uncorrelated at all leads and lags, that is, $E(f_t u_{is}) = 0$ for all i, s.

The unobserved factors are modeled (explicitly or implicitly) as following a linear dynamic process

$$\Gamma(\mathrm{L}) f_t = \eta_t, \tag{7}$$

where $\Gamma(\mathrm{L})$ is a matrix lag polynomial and η_t is a $q \times 1$ disturbance vector.

The DFM implies that the spectral density matrix of X_t can be written as the sum of two parts, one arising from the factors and the other arising from the idiosyncratic disturbance. Because F_t and u_t are uncorrelated at all leads and lags, the spectral density matrix of X_{it} at frequency ω is

$$S_{XX}(\omega) = \lambda\bigl(\mathrm{e}^{\mathrm{i}\omega}\bigr) S_{ff}(\omega) \lambda\bigl(\mathrm{e}^{-\mathrm{i}\omega}\bigr)' + S_{uu}(\omega), \tag{8}$$

where $\lambda(z) = [\lambda_1(z) \ldots \lambda_n(z)]'$ and $S_{ff}(\omega)$ and $S_{uu}(\omega)$ are the spectral density matrices of f_t and u_t at frequency ω. This decomposition, which is due to Geweke (1977), is the frequency-domain counterpart of the variance decomposition of classical factor models.

In classical factor analysis, the factors are identified only up to multiplication by a nonsingular $q \times q$ matrix. In dynamic factor analysis, the factors are identified only up to multiplication by a nonsingular $q \times q$ matrix lag polynomial. This ambiguity can be resolved by imposing identifying restrictions, e.g., restrictions on the dynamic factor loadings and on $\Gamma(\mathrm{L})$. As in classical factor analysis, this identification problem makes it difficult to interpret the dynamic factors, but it is inconsequential for linear forecasting because all that is desired is the linear combination of the factors that produces the minimum mean squared forecast error.

Treatment of Y_t The variable to be forecasted, Y_t, can be handled in two different ways. The first is to include Y_t in the X_t vector and model it as part of the system (6) and (7). This approach is used when n is small and the DFM is estimated parametrically, as is discussed in Section 4.3. When n is large, however, computationally efficient nonparametric methods can be used to estimate the factors, in which case it is useful to treat the forecasting equation for Y_t as a single equation, not as a system.

The single forecasting equation for Y_t can be derived from (6). Augment X_t in that expression by Y_t, so that $Y_t = \lambda_Y(\mathrm{L})' f_t + u_{Yt}$, where $\{u_{Yt}\}$ is distributed independently of $\{f_t\}$ and $\{u_{it}\}$, $i = 1, \ldots, n$. Further suppose that u_{Yt} follows the autoregression, $\delta_Y(\mathrm{L})u_{Yt} = \nu_{Yt}$. Then $\delta_Y(\mathrm{L})Y_{t+1} = \delta_Y(\mathrm{L})'\lambda_Y(\mathrm{L}) f_{t+1} + \nu_{t+1}$ or $Y_{t+1} = \delta_Y(\mathrm{L})\lambda_Y(\mathrm{L})' f_{t+1} + \gamma(\mathrm{L})Y_t + \nu_{t+1}$, where $\gamma(\mathrm{L}) = \mathrm{L}^{-1}(1 - \delta_Y(\mathrm{L}))$. Thus $E[Y_{t+1} \mid X_t, Y_t, f_t, X_{t-1}, Y_{t-1}, f_{t-1}, \ldots] = E[\delta_Y(\mathrm{L})\lambda_Y(\mathrm{L})' f_{t+1} + \gamma(\mathrm{L})Y_t + \nu_{t+1} \mid Y_t, f_t, Y_{t-1}, f_{t-1}, \ldots] = \beta(\mathrm{L}) f_t + \gamma(\mathrm{L})Y_t$, where $\beta(\mathrm{L}) f_t = E[\delta_Y(\mathrm{L})\lambda_Y(\mathrm{L})' f_{t+1} \mid f_t, f_{t-1}, \ldots]$. Setting $Z_t = Y_t$, we thus have

$$Y_{t+1} = \beta(\mathrm{L}) f_t + \gamma(\mathrm{L})' Z_t + \varepsilon_{t+1}, \tag{9}$$

where $\varepsilon_{t+1} = \nu_{Yt+1} + (\delta_Y(\mathrm{L})\lambda_Y(\mathrm{L})' f_{t+1} - E[\delta_Y(\mathrm{L})\lambda_Y(\mathrm{L})' f_{t+1} \mid f_t, f_{t-1}, \ldots])$ has conditional mean zero given X_t, f_t, Y_t and their lags. We use the notation Z_t rather than Y_t for the regressor in (9) to generalize the equation somewhat so that observable predictors other than lagged Y_t can be included in the regression, for example, Z_t might include an observable variable that, in the forecaster's judgment, might be valuable for forecasting Y_{t+1} despite the inclusion of the factors and lags of the dependent variable.

Exact vs. approximate DFMs Chamberlain and Rothschild (1983) introduced a useful distinction between exact and approximate DFMs. In the *exact DFM*, the idiosyncratic terms are mutually uncorrelated, that is,

$$E(u_{it}u_{jt}) = 0 \quad \text{for } i \neq j. \tag{10}$$

The *approximate DFM* relaxes this assumption and allows for a limited amount of correlation among the idiosyncratic terms. The precise technical condition varies from paper to paper, but in general the condition limits the contribution of the idiosyncratic covariances to the total covariance of X as n gets large. For example, Stock and Watson (2002a) require that the average absolute covariances satisfy

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| E(u_{it}u_{jt}) \right| < \infty. \tag{11}$$

There are two general approaches to the estimation of the dynamic factors, the first employing parametric estimation using an exact DFM and the second employing nonparametric methods, either PCA or dynamic PCA. We address these in turn.

4.2. DFM estimation by maximum likelihood

The initial applications of the DFM by Geweke's (1977) and Sargent and Sims (1977) focused on testing the restrictions implied by the exact DFM on the spectrum of X_t, that is, that its spectral density matrix has the factor structure (8), where S_{uu} is diagonal. If n is sufficiently larger than q (for example, if $q = 1$ and $n \geqslant 3$), the null hypothesis of an unrestricted spectral density matrix can be tested against the alternative of a DFM by testing the factor restrictions using an estimator of $S_{XX}(\omega)$. For fixed n, this estimator is asymptotically normal under the null hypothesis and the Wald test statistic has a chi-squared distribution. Although Sargent and Sims (1977) found evidence in favor of a reduced number of factors, their methods did not yield estimates of the factors and thus could not be used for forecasting.

With sufficient additional structure to ensure identification, the parameters of the DFM (6), (7) and (9) can be estimated by maximum likelihood, where the likelihood is computed using the Kalman filter, and the dynamic factors can be estimated using the Kalman smoother [Engle and Watson (1981), Stock and Watson (1989, 1991)]. Specifically, suppose that Y_t is included in X_t. Then make the following assumptions:

(1) the idiosyncratic terms follow a finite order AR model, $\delta_i(\mathrm{L})u_{it} = \nu_{it}$;
(2) $(\nu_{1t}, \ldots, \nu_{nt}, \eta_{1t}, \ldots, \eta_{qt})$ are i.i.d. normal and mutually independent;
(3) $\Gamma(\mathrm{L})$ has finite order with $\Gamma_0 = I_r$;
(4) $\lambda_i(\mathrm{L})$ is a lag polynomial of degree p; and
(5) $[\lambda'_{10} \ldots \lambda'_{q0}]' = I_q$.

Under these assumptions, the Gaussian likelihood can be constructed using the Kalman filter, and the parameters can be estimated by maximizing this likelihood.

One-step ahead forecasts Using the MLEs of the parameter vector, the time series of factors can be estimated using the Kalman smoother. Let $f_{t|T}$ and $u_{it|T}$, $i = 1, \ldots, n$, respectively denote the Kalman smoother estimates of the unobserved factors and idiosyncratic terms using the full data through time T. Suppose that the variable of interest is the final element of X_t. Then the one-step ahead forecast of the variable of interest at time $T + 1$ is $Y_{T+1|T} = X_{nT+1|T} = \hat{\lambda}_n(\mathrm{L})' f_{T|T} + u_{nT|T}$, where $\hat{\lambda}_n(\mathrm{L})$ is the MLE of $\lambda_n(\mathrm{L})$.[2]

h-step ahead forecasts Multistep ahead forecasts can be computed using either the iterated or the direct method. The iterated h-step ahead forecast is computed by solving the full DFM forward, which is done using the Kalman filter. The direct h-step ahead forecast is computed by projecting Y^h_{t+h} onto the estimated factors and observables, that is, by estimating $\beta_h(\mathrm{L})$ and $\gamma_h(\mathrm{L})$ in the equation

$$Y^h_{t+h} = \beta_h(\mathrm{L})' f_{t|t} + \gamma_h(\mathrm{L}) Y_t + \varepsilon^h_{t+h} \tag{12}$$

[2] Peña and Poncela (2004) provide an interpretation of forecasts based on the exact DFM as shrinkage forecasts.

(where $\mathrm{L}^i f_{t/t} = f_{t-i/t}$) using data through period $T - h$. Consistent estimates of $\beta_h(\mathrm{L})$ and $\gamma_h(\mathrm{L})$ can be obtained by OLS because the signal extraction error $f_{t-i} - f_{t-i/t}$ is uncorrelated with $f_{t-j/t}$ and Y_{t-j} for $j \geqslant 0$. The forecast for period $T + h$ is then $\hat{\beta}_h(\mathrm{L})' f_{T|T} + \hat{\gamma}_h(\mathrm{L}) Y_T$. The direct method suffers from the usual potential inefficiency of direct forecasts arising from the inefficient estimation of $\beta_h(\mathrm{L})$ and $\gamma_h(\mathrm{L})$, instead of basing the projections on the MLEs.

Successes and limitations Maximum likelihood has been used successfully to estimate the parameters of low-dimensional DFMs, which in turn have been used to estimate the factors and (among other things) to construct indexes of coincident and leading economic indicators. For example, Stock and Watson (1991) use this approach (with $n = 4$) to rationalize the U.S. Index of Coincident Indicators, previously maintained by the U.S. Department of Commerce and now produced the Conference Board. The method has also been used to construct regional indexes of coincident indexes, see Clayton-Matthews and Crone (2003). (For further discussion of DFMs and indexes of coincident and leading indicators, see Chapter 16 by Marcellino in this Handbook.) Quah and Sargent (1993) estimated a larger system ($n = 60$) by MLE. However, the underlying assumption of an exact factor model is a strong one. Moreover, the computational demands of maximizing the likelihood over the many parameters that arise when n is large are significant. Fortunately, when n is large, other methods are available for the consistent estimation of the factors in approximate DFMs.

4.3. DFM estimation by principal components analysis

If the lag polynomials $\lambda_i(\mathrm{L})$ and $\beta(\mathrm{L})$ have finite order p, then (6) and (9) can be written

$$X_t = \Lambda F_t + u_t, \tag{13}$$

$$Y_{t+1} = \beta' F_t + \gamma(\mathrm{L})' Z_t + \varepsilon_{t+1}, \tag{14}$$

where $F_t = [f_t' f_{t-1}' \ldots f_{t-p+1}']'$, $u_t = [u_{1t} \ldots u_{nt}]$, Λ is a matrix consisting of zeros and the coefficients of $\lambda_i(\mathrm{L})$, and β is a vector of parameters composed of the elements of $\beta(\mathrm{L})$. If the number of lags in β exceeds the number of lags in Λ, then the term $\beta' F_t$ in (14) can be replaced by a distributed lag of F_t.

Equations (13) and (14) rewrite the DFM as a static factor model, in which there are r static factors consisting of the current and lagged values of the q dynamic factors, where $r \leqslant pq$ (r will be strictly less than pq if one or more lagged dynamic factors are redundant). The representation (13) and (14) is called the static representation of the DFM.

Because F_t and u_t are uncorrelated at all leads and lags, the covariance matrix of X_t, Σ_{XX}, is the sum of two parts, one arising from the common factors and the other arising from the idiosyncratic disturbance:

$$\Sigma_{XX} = \Lambda \Sigma_{FF} \Lambda' + \Sigma_{uu}, \tag{15}$$

where Σ_{FF} and Σ_{uu} are the variance matrices of F_t and u_t. This is the usual variance decomposition of classical factor analysis.

When n is small, the standard methods of estimation of exact static factor models are to estimate Λ and Σ_{uu} by Gaussian maximum likelihood estimation or by method of moments [Anderson (1984)]. However, when n is large simpler methods are available. Under the assumptions that the eigenvalues of Σ_{uu} are $O(1)$ and $\Lambda'\Lambda$ is $O(n)$, the first r eigenvalues of Σ_{XX} are $O(N)$ and the remaining eigenvalues are $O(1)$. This suggests that the first r principal components of X can serve as estimators of Λ, which could in turn be used to estimate F_t. In fact, if Λ were known, then F_t could be estimated by $(\Lambda'\Lambda)^{-1}\Lambda'X_t$: by (13), $(\Lambda'\Lambda)^{-1}\Lambda'X_t = F_t + (\Lambda'\Lambda)^{-1}\Lambda'u_t$. Under the two assumptions, $\operatorname{var}[(\Lambda'\Lambda)^{-1}\Lambda'u_t] = (\Lambda'\Lambda)^{-1}\Lambda'\Sigma_{uu}\Lambda(\Lambda'\Lambda)^{-1} = O(1/n)$, so that if Λ were known, F_t could be estimated precisely if n is sufficiently large.

More formally, by analogy to regression we can consider estimation of Λ and F_t by solving the nonlinear least-squares problem

$$\min_{F_1,\dots,F_T,\Lambda} T^{-1}\sum_{t=1}^{T}(X_t - \Lambda F_t)'(X_t - \Lambda F_t) \tag{16}$$

subject to $\Lambda'\Lambda = I_r$. Note that this method treats $F_1, \dots, F_T$ as fixed parameters to be estimated.[3] The first order conditions for maximizing (16) with respect to F_t shows that the estimators satisfy $\hat{F}_t = (\hat{\Lambda}'\hat{\Lambda})^{-1}\hat{\Lambda}'X_t$. Substituting this into the objective function yields the concentrated objective function, $T^{-1}\sum_{t=1}^{T} X_t'[I - \Lambda(\Lambda'\Lambda)^{-1}\Lambda]X_t$. Minimizing the concentrated objective function is equivalent to maximizing $\operatorname{tr}\{(\Lambda'\Lambda)^{-1/2\prime}\Lambda'\hat{\Sigma}_{XX}\Lambda(\Lambda'\Lambda)^{-1/2}\}$, where $\hat{\Sigma}_{XX} = T^{-1}\sum_{t=1}^{T} X_tX_t'$. This in turn is equivalent to maximizing $\Lambda'\hat{\Sigma}_{XX}\Lambda$ subject to $\Lambda'\Lambda = I_r$, the solution to which is to set $\hat{\Lambda}$ to be the first r eigenvectors of $\hat{\Sigma}_{XX}$. The resulting estimator of the factors is $\hat{F}_t = \hat{\Lambda}'X_t$, which is the vector consisting of the first r principal components of X_t. The matrix $T^{-1}\sum_{t=1}^{T}\hat{F}_t\hat{F}_t'$ is diagonal with diagonal elements that equal the largest r ordered eigenvalues of $\hat{\Sigma}_{XX}$. The estimators $\{\hat{F}_t\}$ could be rescaled so that $T^{-1}\sum_{t=1}^{T}\hat{F}_t\hat{F}_t' = I_r$, however this is unnecessary if the only purpose is forecasting. We will refer to $\{\hat{F}_t\}$ as the PCA estimator of the factors in the static representation of the DFM.

PCA: large-n theoretical results Connor and Korajczyk (1986) show that the PCA estimators of the space spanned by the factors are pointwise consistent for T fixed and $n \to \infty$ in the approximate factor model, but do not provide formal arguments for n, $T \to \infty$. Ding and Hwang (1999) provide consistency results for PCA estimation of

[3] When $F_1, \dots, F_T$ are treated as parameters to be estimated, the Gaussian likelihood for the classical factor model is unbounded, so the maximum likelihood estimator is undefined [see Anderson (1984)]. This difficulty does not arise in the least-squares problem (16), which has a global minimum (subject to the identification conditions discussed in this and the previous sections).

the classic exact factor model as n, $T \to \infty$, and Stock and Watson (2002a) show that, in the static form of the DFM, the space of the dynamic factors is consistently estimated by the principal components estimator as n, $T \to \infty$, with no further conditions on the relative rates of n or T. In addition, estimation of the coefficients of the forecasting equation by OLS, using the estimated factors as regressors, produces consistent estimates of $\beta(\mathrm{L})$ and $\gamma(\mathrm{L})$ and, consequently, forecasts that are first-order efficient, that is, they achieve the mean squared forecast error of the infeasible forecast based on the true coefficients and factors. Bai (2003) shows that the PCA estimator of the common component is asymptotically normal, converging at a rate of $\min(n^{1/2}, T^{1/2})$, even if u_t is serially correlated and/or heteroskedastic.

Some theory also exists, also under strong conditions, concerning the distribution of the largest eigenvalues of the sample covariance matrix of X_t. If n and T are fixed and X_t is i.i.d. $\mathrm{N}(0, \Sigma_{XX})$, then the principal components are distributed as those of a noncentral Wishart; see James (1964) and Anderson (1984). If n is fixed, $T \to \infty$, and the eigenvalues of Σ_{XX} are distinct, then the principal components are asymptotically normally distributed (they are continuous functions of $\hat{\Sigma}_{XX}$, which is itself asymptotically normally distributed). Johnstone (2001) [extended by El Karoui (2003)] shows that the largest eigenvalues of $\hat{\Sigma}_{XX}$ satisfy the Tracy–Widom law if n, $T \to \infty$, however these results apply to unscaled X_{it} (not divided by its sample standard deviation).

Weighted principal components Suppose for the moment that u_t is i.i.d. $\mathrm{N}(0, \Sigma_{uu})$ and that Σ_{uu} is known. Then by analogy to regression, one could modify (16) and consider the nonlinear generalized least-squares (GLS) problem

$$\min_{F_1,\ldots,F_T,\Lambda} \sum_{t=1}^{T} (X_t - \Lambda F_t)' \Sigma_{uu}^{-1} (X_t - \Lambda F_t). \tag{17}$$

Evidently the weighting schemes in (16) and (17) differ. Because (17) corresponds to GLS when Σ_{uu} is known, there could be efficiency gains by using the estimator that solves (17) instead of the PCA estimator.

In applications, Σ_{uu} is unknown, so minimizing (17) is infeasible. However, Boivin and Ng (2003) and Forni et al. (2003b) have proposed feasible versions of (17). We shall call these weighted PCA estimators since they involve alternative weighting schemes in place of simply weighting by the inverse sample variances as does the PCA estimator (recall the notational convention that X_t has been standardized to have sample variance one). Jones (2001) proposed a weighted factor estimation algorithm which is closely related to weighted PCA estimation when n is large.

Because the exact factor model posits that Σ_{uu} is diagonal, a natural approach is to replace Σ_{uu} in (17) with an estimator that is diagonal, where the diagonal elements are estimators of the variance of the individual u_{it}'s. This approach is taken by Jones (2001) and Boivin and Ng (2003). Boivin and Ng (2003) consider several diagonal weighting schemes, including schemes that drop series that are highly correlated with others. One simple two-step weighting method, which Boivin and Ng (2003) found worked well in their empirical application to U.S. data, entails estimating the diagonal elements of Σ_{uu}

by the sample variances of the residuals from a preliminary regression of X_{it} onto a relatively large number of factors estimated by PCA.

Forni et al. (2003b) also consider two-step weighted PCA, where they estimated Σ_{uu} in (17) by the difference between $\hat{\Sigma}_{XX}$ and an estimator of the covariance matrix of the common component, where the latter estimator is based on a preliminary dynamic principal components analysis (dynamic PCA is discussed below). They consider both diagonal and nondiagonal estimators of Σ_{uu}. Like Boivin and Ng (2003), they find that weighted PCA can improve upon conventional PCA, with the gains depending on the particulars of the stochastic processes under study.

The weighted minimization problem (17) was motivated by the assumption that u_t is i.i.d. N(0, Σ_{uu}). In general, however, u_t will be serially correlated, in which case GLS entails an adjustment for this serial correlation. Stock and Watson (2005) propose an extension of weighted PCA in which a low-order autoregressive structure is assumed for u_t. Specifically, suppose that the diagonal filter $D(\mathrm{L})$ whitens u_t so that $D(\mathrm{L})u_t \equiv \tilde{u}_t$ is serially uncorrelated. Then the generalization of (17) is

$$\min_{D(\mathrm{L}), \tilde{F}_1, \ldots, \tilde{F}_T, \Lambda} \sum_{t=1}^{T} \left[D(\mathrm{L})X_t - \Lambda \tilde{F}_t\right]' \Sigma_{\tilde{u}\tilde{u}}^{-1} \left[D(\mathrm{L})X_t - \Lambda \tilde{F}_t\right], \tag{18}$$

where $\tilde{F}_t = D(\mathrm{L})F_t$ and $\Sigma_{\tilde{u}\tilde{u}} = E\tilde{u}_t\tilde{u}_t'$. Stock and Watson (2005) implement this with $\Sigma_{\tilde{u}\tilde{u}} = I_n$, so that the estimated factors are the principal components of the filtered series $D(\mathrm{L})X_t$. Estimation of $D(\mathrm{L})$ and $\{\tilde{F}_t\}$ can be done sequentially, iterating to convergence.

Factor estimation under model instability There are some theoretical results on the properties of PCA factor estimates when there is parameter instability. Stock and Watson (2002a) show that the PCA factor estimates are consistent even if there is some temporal instability in the factor loadings, as long as the temporal instability is sufficiently dissimilar from one series to the next. More broadly, because the precision of the factor estimates improves with n, it might be possible to compensate for short panels, which would be appropriate if there is parameter instability, by increasing the number of predictors. More work is needed on the properties of PCA and dynamic PCA estimators under model instability.

Determination of the number of factors At least two statistical methods are available for the determination of the number of factors when n is large. The first is to use model selection methods to estimate the number of factors that belong in the forecasting equation (14). Given an upper bound on the dimension and lags of F_t, Stock and Watson (2002a) show that this can be accomplished using an information criterion. Although the rate requirements for the information criteria in Stock and Watson (2002a) technically rule out the BIC, simulation results suggest that the BIC can perform well in the sample sizes typically found in macroeconomic forecasting applications.

The second approach is to estimate the number of factors entering the full DFM. Bai and Ng (2002) prove that the dimension of F_t can be estimated consistently for approximate DFMs that can be written in static form, using suitable information criteria which they provide. In principle, these two methods are complementary: a full set of factors could be chosen using the Bai–Ng method, and model selection could then be applied to the Y_t equation to select a subset of these for forecasting purposes.

h-step ahead forecasts Direct h-step ahead forecasts are produced by regressing Y_{t+h}^h against $\hat{F}_t$ and, possibly, lags of $\hat{F}_t$ and Y_t, then forecasting Y_{t+h}^h.

Iterated h-step ahead forecasts require specifying a subsidiary model of the dynamic process followed by F_t, which has heretofore not been required in the principal components method. One approach, proposed by Bernanke, Boivin and Eliasz (2005) models (Y_t, F_t) jointly as a VAR, which they term a factor-augmented VAR (FAVAR). They estimate this FAVAR using the PCA estimates of $\{F_t\}$. Although they use the estimated model for impulse response analysis, it could be used for forecasting by iterating the estimated FAVAR h steps ahead.

In a second approach to iterated multistep forecasts, Forni et al. (2003b) and Giannoni, Reichlin and Sala (2004) developed a modification of the FAVAR approach in which the shocks in the F_t equation in the VAR have reduced dimension. The motivation for this further restriction is that F_t contains lags of f_t. The resulting h-step forecasts are made by iterating the system forward using the Kalman filter.

4.4. DFM estimation by dynamic principal components analysis

The method of dynamic principal components was introduced by Brillinger (1964) and is described in detail in Brillinger's (1981) textbook. Static principal components entails finding the closest approximation to the covariance matrix of X_t among all covariance matrices of a given reduced rank. In contrast, dynamic principal components entails finding the closest approximation to the spectrum of X_t among all spectral density matrices of a given reduced rank.

Brillinger's (1981) estimation algorithm generalizes static PCA to the frequency domain. First, the spectral density of X_t is estimated using a consistent spectral density estimator, $\hat{S}_{XX}(\omega)$, at frequency ω. Next, the eigenvectors corresponding to the largest q eigenvalues of this (Hermitian) matrix are computed. The inverse Fourier transform of these eigenvectors yields estimators of the principal component time series using formulas given in Brillinger (1981, Chapter 9).

Forni et al. (2000, 2004) study the properties of this algorithm and the estimator of the common component of X_{it} in a DFM, $\lambda_i(\mathrm{L})' f_t$, when n is large. The advantages of this method, relative to parametric maximum likelihood, are that it allows for an approximate dynamic factor structure, and it does not require high-dimensional maximization when n is large. The advantage of this method, relative to static principal components, is that it admits a richer lag structure than the finite-order lag structure that led to (13).

Brillinger (1981) summarizes distributional results for dynamic PCA for the case that n is fixed and $T \to \infty$ (as in classic PCA, estimators are asymptotically normal because they are continuous functions of $\hat{S}_{XX}(\omega)$, which is asymptotically normal).

Forni et al. (2000) show that dynamic PCA provides pointwise consistent estimation of the common component as n and T both increase, and Forni et al. (2004) further show that this consistency holds if $n, T \to \infty$ and $n/T \to 0$. The latter condition suggests that some caution should be exercised in applications in which n is large relative to T, although further evidence on this is needed.

The time-domain estimates of the dynamic common components series are based on two-sided filters, so their implementation entails trimming the data at the start and end of the sample. Because dynamic PCA does not yield an estimator of the common component at the end of the sample, this method cannot be used for forecasting, although it can be used for historical analysis or [as is done by Forni et al. (2003b)] to provide a weighting matrix for subsequent use in weighted (static) PCA. Because the focus of this chapter is on forecasting, not historical analysis, we do not discuss dynamic principal components further.

4.5. DFM estimation by Bayes methods

Another approach to DFM estimation is to use Bayes methods. The difficulty with maximum likelihood estimation of the DFM when n is large is not that it is difficult to compute the likelihood, which can be evaluated fairly rapidly using the Kalman filter, but rather that it requires maximizing over a very large parameter vector. From a computational perspective, this suggests that perhaps averaging the likelihood with respect to some weighting function will be computationally more tractable than maximizing it; that is, Bayes methods might be offer substantial computational gains.

Otrok and Whiteman (1998), Kim and Nelson (1998), and Kose, Otrok and Whiteman (2003) develop Markov Chain Monte Carlo (MCMC) methods for sampling from the posterior distribution of dynamic factor models. The focus of these papers was inference about the parameters, historical episodes, and implied model dynamics, not forecasting. These methods also can be used for forecast construction (see Otrok, Silos and Whiteman (2003) and Chapter 1 by Geweke and Whiteman in this Handbook), however to date not enough is known to say whether this approach provides an improvement over PCA-type methods when n is large.

4.6. Survey of the empirical literature

There have been several empirical studies that have used estimated dynamic factors for forecasting. In two prescient but little-noticed papers, Figlewski (1983) ($n = 33$) and Figlewski and Urich (1983) ($n = 20$) considered combining forecasts from a panel of forecasts using a static factor model. Figlewski (1983) pointed out that, if forecasters are unbiased, then the factor model implied that the average forecast would converge in probability to the unobserved factor as n increases. Because some forecasters are better than others, the optimal factor-model combination (which should be close to but not equal to the largest weighted principle component) differs from equal weighting. In an application to a panel of $n = 33$ forecasters who participated in the Livingston price survey, with $T = 65$ survey dates, Figlewski (1983) found that using the optimal static factor model combination outperformed the simple weighted average. When Figlewski and Urich (1983) applied this methodology to a panel of $n = 20$ weekly forecasts of the money supply, however, they were unable to improve upon the simple weighted average forecast.

Recent studies on large-model forecasting have used pseudo-out-of-sample forecast methods (that is, recursive or rolling forecasts) to evaluate and to compare forecasts. Stock and Watson (1999) considered factor forecasts for U.S. inflation, where the factors were estimated by PCA from a panel of up to 147 monthly predictors. They found that the forecasts based on a single real factor generally had lower pseudo-out-of-sample forecast error than benchmark autoregressions and traditional Phillips-curve forecasts. Stock and Watson (2002b) found substantial forecasting improvements for real variables using dynamic factors estimated by PCA from a panel of up to 215 U.S. monthly predictors, a finding confirmed by Bernanke and Boivin (2003). Boivin and Ng (2003) compared forecasts using PCA and weighted PCA estimators of the factors, also for U.S. monthly data ($n = 147$). They found that weighted PCA forecasts tended to outperform PCA forecasts for real variables but not nominal variables.

There also have been applications of these methods to non-U.S. data. Forni et al. (2003b) focused on forecasting Euro-wide industrial production and inflation (HICP) using a short monthly data set (1987:2–2001:3) with very many predictors ($n = 447$). They considered both PCA and weighted PCA forecasts, where the weighted principal components were constructed using the dynamic PCA weighting method of Forni et al. (2003a). The PCA and weighted PCA forecasts performed similarly, and both exhibited modest improvements over the AR benchmark. Brisson, Campbell and Galbraith (2002) examined the performance factor-based forecasts of Canadian GDP and investment growth using two panels, one consisting of only Canadian data ($n = 66$) and one with both Canadian and U.S. data ($n = 133$), where the factors were estimated by PCA. They find that the factor-based forecasts improve substantially over benchmark models (autoregressions and some small time series models), but perform less well than the real-time OECD forecasts of these series. Using data for the UK, Artis, Banerjee and Marcellino (2001) found that 6 factors (estimated by PCA) explain 50% of the variation in their panel of 80 variables, and that factor-based forecasts could make substantial forecasting improvements for real variables, especially at longer horizons.

Practical implementation of DFM forecasting requires making many modeling decisions, notably to use PCA or weighted PCA, how to construct the weights if weighted PCA weights is used, and how to specify the forecasting equation. Existing theory provides limited guidance on these choices. Forni et al. (2003b) and Boivin and Ng (2005) provide simulation and empirical evidence comparing various DFM forecasting methods, and we provide some additional empirical comparisons are provided in Section 7 below.

DFM-based methods also have been used to construct real-time indexes of economic activity based on large cross sections. Two such indexes are now being produced and publicly released in real time. In the U.S., the Federal Reserve Bank of Chicago publishes the monthly Chicago Fed National Activity Index (CFNAI), where the index is the single factor estimated by PCA from a panel of 85 monthly real activity variables [Federal Reserve Bank of Chicago (undated)]. In Europe, the Centre for Economic Policy Research (CEPR) in London publishes the monthly European Coincident Index (EuroCOIN), where the index is the single dynamic factor estimated by weighted PCA from a panel of nearly 1000 economic time series for Eurozone countries [Altissimo et al. (2001)].

These methods also have been used for nonforecasting purposes, which we mention briefly although these are not the focus of this survey. Following Connor and Korajczyk (1986, 1988), there have been many applications in finance that use (static) factor model methods to estimate unobserved factors and, among other things, to test whether those unobserved factors are consistent with the arbitrage pricing theory; see Jones (2001) for a recent contribution and additional references. Forni and Reichlin (1998), Bernanke and Boivin (2003), Favero and Marcellino (2001), Bernanke, Boivin and Eliasz (2005), Giannoni, Reichlin and Sala (2002, 2004) and Forni et al. (2005) used estimated factors in an attempt better to approximate the true economic shocks and thereby to obtain improved estimates of impulse responses as variables. Another application, pursued by Favero and Marcellino (2001) and Favero, Marcellino and Neglia (2002), is to use lags of the estimated factors as instrumental variables, reflecting the hope that the factors might be stronger instruments than lagged observed variables. Kapetanios and Marcellino (2002) and Favero, Marcellino and Neglia (2002) compared PCA and dynamic PCA estimators of the dynamic factors. Generally speaking, the results are mixed, with neither method clearly dominating the other. A point stressed by Favero, Marcellino and Neglia (2002) is that the dynamic PCA methods estimate the factors by a two-sided filter, which makes it problematic, or even unsuitable, for applications in which strict timing is important, such as using the estimated factors in VARs or as instrumental variables. More research is needed before clear recommendation about which procedure is best for such applications.

5. Bayesian model averaging

Bayesian model averaging (BMA) can be thought of as a Bayesian approach to combination forecasting. In forecast combining, the forecast is a weighted average of the individual forecasts, where the weights can depend on some measure of the historical accuracy of the individual forecasts. This is also true for BMA, however in BMA the weights are computed as formal posterior probabilities that the models are correct. In addition, the individual forecasts in BMA are model-based and are the posterior means of the variable to be forecast, conditional on the selected model. Thus BMA extends forecast combining to a fully Bayesian setting, where the forecasts themselves are optimal Bayes forecasts, given the model (and some parametric priors). Importantly, recent research on BMA methods also has tackled the difficult computational problem in which the individual models can contain arbitrary subsets of the predictors X_t. Even if n is moderate, there are more models than can be computed exhaustively, yet by cleverly sampling the most likely models, BMA numerical methods are able to provide good approximations to the optimal combined posterior mean forecast.

The basic paradigm for BMA was laid out by Leamer (1978). In an early contribution in macroeconomic forecasting, Min and Zellner (1993) used BMA to forecast annual output growth in a panel of 18 countries, averaging over four different models. The area of BMA has been very active recently, mainly occurring outside economics. Work on BMA through the 1990s is surveyed by Hoeting et al. (1999) and their discussants, and Chapter 1 by Geweke and Whiteman in this Handbook contains a thorough discussion of Bayesian forecasting methods. In this section, we focus on BMA methods specifically developed for linear prediction with large n. This is the focus of Fernandez, Ley and Steel (2001a) [their application in Fernandez, Ley and Steel (2001b) is to growth regressions], and we draw heavily on their work in the next section.

This section first sets out the basic BMA setup, then turns to a discussion of the few empirical applications to date of BMA to economic forecasting with many predictors.

5.1. Fundamentals of Bayesian model averaging

In standard Bayesian analysis, the parameters of a given model are treated as random, distributed according to a prior distribution. In BMA, the binary variable indicating whether a given model is true also is treated as random and distributed according to some prior distribution.

Specifically, suppose that the distribution of Y_{t+1} conditional on X_t is given by one of K models, denoted by $M_1, \ldots, M_K$. We focus on the case that all the models are linear, so they differ by which subset of predictors X_t are contained in the model. Thus M_k specifies the list of indexes of X_t contained in model k. Let $\pi(M_k)$ denote the prior probability that the data are generated by model k, and let D_t denote the data set through date t. Then the predictive probability density for Y_{T+1} is

$$f(Y_{T+1} \mid D_T) = \sum_{k=1}^{K} f_k(Y_{T+1} \mid D_T) \Pr(M_k \mid D_T), \tag{19}$$

where $f_k(Y_{T+1} \mid D_T)$ is the predictive density of Y_{T+1} for model k and $\Pr(M_k \mid D_T)$ is the posterior probability of model k. This posterior probability is given by

$$\Pr(M_k \mid D_T) = \frac{\Pr(D_T \mid M_k)\pi(M_k)}{\sum_{i=1}^{K} \Pr(D_T \mid M_i)\pi(M_i)}, \tag{20}$$

where $\Pr(D_T \mid M_k)$ is given by

$$\Pr(D_T \mid M_k) = \int \Pr(D_T \mid \theta_k, M_k)\pi(\theta_k \mid M_k)\,\mathrm{d}\theta_k, \tag{21}$$

where θ_k is the vector of parameters in model k and $\pi(\theta_k \mid M_k)$ is the prior for the parameters in model k.

Under squared error loss, the optimal Bayes forecast is the posterior mean of Y_{T+1}, which we denote by $\tilde{Y}_{T+1|T}$. It follows from (19) that this posterior mean is

$$\tilde{Y}_{T+1|T} = \sum_{k=1}^{K} \Pr(M_k \mid D_T) \tilde{Y}_{M_k,T+1|T}, \tag{22}$$

where $\tilde{Y}_{M_k,T+1|T}$ is the posterior mean of Y_{T+1} for model M_k.

Comparison of (22) and (3) shows that BMA can be thought of as an extension of the Bates–Granger (1969) forecast combining setup, where the weights are determined by the posterior probabilities over the models, the forecasts are posterior means, and, because the individual forecasts are already conditional means, given the model, there is no constant term ($w_0 = 0$ in (3)).

These simple expressions mask considerable computational difficulties. If the set of models is allowed to be all possible subsets of the predictors X_t, then there are $K = 2^n$ possible models. Even with $n = 30$, this is several orders of magnitude more than is feasible to compute exhaustively. Thus the computational objective is to approximate the summation (22) while only evaluating a small subset of models. Achieving this objective requires a judicious choice of prior distributions and using appropriate numerical simulation methods.

Choice of priors Implementation of BMA requires choosing two sets of priors, the prior distribution of the parameters given the model and the prior probability of the model. In principle, the researcher could have prior beliefs about the values of specific parameters in specific models. In practice, however, given the large number of models this is rarely the case. In addition, given the large number of models to evaluate, there is a premium on using priors that are computationally convenient. These considerations lead to the use of priors that impose little prior information and that lead to posteriors (21) that are easy to evaluate quickly.

Fernandez, Ley and Steel (2001a) conducted a study of various priors that might usefully be applied in linear models with economic data and large n. Based on theoretical consideration and simulation results, they propose a benchmark set of priors for BMA in the linear model with large n. Let the kth model be

$$Y_{t+1} = X_t^{(k)\prime} \beta_k + Z_t' \gamma + \varepsilon_t, \tag{23}$$

where $X_t^{(k)}$ is the vector of predictors appearing in model k, Z_t is a vector of variables to be included in all models, β_k and γ are coefficient vectors, and ε_t is the error term. The analysis is simplified if the model-specific regressors $X_t^{(k)}$ are orthogonal to the common regressor Z_t, and this assumption is adopted throughout this section by taking $X_t^{(k)}$ to be the residuals from the projection of the original set of predictors onto Z_t. In applications to economic forecasting, because of serial correlation in Y_t, Z_t might include lagged values of Y that potentially appear in each model.

Following the rest of the literature on BMA in the linear model [cf. Hoeting et al. (1999)], Fernandez, Ley and Steel (2001a) assume that $\{X_t^{(k)}, Z_t\}$ is strictly exogenous and ε_t is i.i.d. N$(0, \sigma^2)$. In the notation of (21), $\theta_k = [\beta_k' \, \gamma' \, \sigma]'$. They suggest using conjugate priors, an uninformative prior for γ and σ^2 and Zellner's (1986) g-prior for β_k:

$$\pi(\gamma, \sigma \mid M_k) \propto 1/\sigma, \tag{24}$$

$$\pi(\beta_k \mid \sigma, M_k) = \mathrm{N}\left(0, \sigma^2\left(g \sum_{t=1}^{T} X_t^{(k)} X_t^{(k)\prime}\right)^{-1}\right). \tag{25}$$

With the priors (24) and (25), the conditional marginal likelihood $\Pr(D_T \mid M_k)$ in (21) is

$$\begin{aligned}&\Pr(Y_1, \ldots, Y_T \mid M_k)\\ &\quad = \text{const} \times a(g)^{\frac{1}{2}\#M_k}\left[a(g) SSR^R + \left(1 - a(g)\right) SSR_k^U\right]^{-\frac{1}{2}\mathrm{df}^R},\end{aligned} \tag{26}$$

where $a(g) = g/(1+g)$, SSR^R is the sum of squared residuals of Y from the restricted OLS regression of Y_{t+1} on Z_t, SSR_k^U is the sum of squared residuals from the OLS regression of Y onto $(X_t^{(k)}, Z_t)$, $\#M_k$ is the dimension of $X_t^{(k)}$, df^R is the degrees of freedom of the restricted regression, and the constant is the same from one model to the next [see Raftery, Madigan and Hoeting (1997) and Fernandez, Ley and Steel (2001a)].

The prior model probability, $\pi(M_k)$, also needs to be specified. One choice for this prior is a multinomial distribution, where the probability is determined by the prior probability that an individual variable enters the model; see, for example, Koop and Potter (2004). If all the variables are deemed equally likely to enter and whether one variable enters the model is treated as independent of whether any other variable enters, then the prior probability for all models is the same and the term $\pi(\theta_k)$ drops out of the expressions. In this case, (22), (20) and (26) imply that

$$\begin{aligned}&\tilde{Y}_{T+1|T} = \sum_{k=1}^{K} w_k \tilde{Y}_{M_k, T+1|T},\\ &\quad \text{where } w_k = \frac{a(g)^{\frac{1}{2}\#M_k}[1 + g^{-1} SSR_k^U / SSR^R]^{-\frac{1}{2}\mathrm{df}^R}}{\sum_{i=1}^{K} a(g)^{\frac{1}{2}\#M_i}[1 + g^{-1} SSR_i^U / SSR^R]^{-\frac{1}{2}\mathrm{df}^R}}.\end{aligned} \tag{27}$$

Three aspects of (27) bear emphasis. First, this expression links BMA and forecast combining: for the linear model with the g-prior and in which each model is given equal prior probability, the BMA forecast as a weighted average of the (Bayes) forecasts from the individual models, where the weighting factor depends on the reduction in the sum of squared residuals of model M_k, relative to the benchmark model that includes only Z_t.

Second, the weights in (27) (and the posterior (26)) penalize models with more parameters through the exponent $\#M_k/2$. This arises directly from the g-prior calculations and appears even though the derivation here places equal weight on all models. A further penalty could be placed on large models by letting $\pi(M_k)$ depend on $\#M_k$.

Third, the weights are based on the posterior (marginal likelihood) (26), which is conditional on $\{X_t^{(k)}, Z_t\}$. Conditioning on $\{X_t^{(k)}, Z_t\}$ is justified by the assumption that the regressors are strictly exogenous, an assumption we return to below.

The foregoing expressions depend upon the hyperparameter g. The choice of g determines the amount of shrinkage appears in the Bayes estimator of β_k, with higher values of g corresponding to greater shrinkage. Based on their simulation study, Fernandez, Ley and Steel (2001a) suggest $g = 1/\min(T, n^2)$. Alternatively, empirical Bayes methods could be used to estimate the value of g that provides the BMA forecasts with the best performance.

Computation of posterior over models If n exceeds 20 or 25, there are too many models to enumerate and the population summations in (27) cannot be evaluated directly. Instead, numerical algorithms have been developed to provide precise, yet numerically efficient, estimates of this the summation.

In principle, one could approximate the population mean in (27) by drawing a random sample of models, evaluating the weights and the posterior means for each forecast, and evaluating (27) using the sample averages, so the summations run over sampled models. In many applications, however, a large fraction of models might have posterior probability near zero, so this method is computationally inefficient. For this reason, a number of methods have been developed that permit accurate estimation of (27) using a relatively small sample of models. The key to these algorithms is cleverly deciding which models to sample with high probability. Clyde (1999a, 1999b) provides a survey of these methods. Two closely related methods are the stochastic search variable selection (SSVS) methods of George and McCulloch (1993, 1997) [also see Geweke (1996)] and the Markov chain Monte Carlo model composition (MC^3) algorithm of Madigan and York (1995); we briefly summarize the latter.

The MC^3 sampling scheme starts with a given model, say M_k. One of the n elements of X_t is chosen at random; a new model, $M_{k'}$, is defined by dropping that regressor if it appears in M_k, or adding it to M_k if it does not. The sampler moves from model M_k to $M_{k'}$ with probability $\min(1, B_{k,k'})$, where $B_{k,k'}$ is the Bayes ratio comparing the two models (which, with the g-prior, is computed using (26)). Following Fernandez, Ley and Steel (2001a), the summation (27) is estimated using the summands for the visited models.

Orthogonalized regressors The computational problem simplifies greatly if the regressors are orthogonal. For example, Koop and Potter (2004) transform X_t to its principal components, but in contrast to the DFM methods discussed in Section 3, all or a large number of the components are kept. This approach can be seen as an extension of the DFM methods in Section 4, where BIC or AIC model selection is replaced by BMA, where nonzero prior probability is placed on the higher principal components entering as predictors. In this sense, it is plausible to model the prior probability of the kth principle component entering as a declining function of k.

Computational details for BMA in linear models with orthogonal regressors and a g-prior are given in Clyde (1999a) and Clyde, Desimone and Parmigiani (1996). [As

Clyde, Desimone and Parmigiani (1996) point out, the method of orthogonalization is irrelevant when a g-prior is used, so weighted principal components can be used instead of standard PCA.] Let γ_j be a binary random variable indicating whether regressor j is in the model, and treat γ_j as independently (but not necessarily identically) distributed with prior probability $\pi_j = \Pr(\gamma_j = 1)$. Suppose that σ_ε^2 is known. Because the regressors are exogenous and the errors are normally distributed, the OLS estimators $\{\hat{\beta}_j\}$ are sufficient statistics. Because the regressors are orthogonal, γ_j, β_j and $\hat{\beta}_j$ are jointly independently distributed over j. Consequently, the posterior mean of β_j depends on the data only through $\hat{\beta}_j$ and is given by

$$E\bigl(\beta_j \mid \hat{\beta}_j, \sigma_\varepsilon^2\bigr) = a(g)\hat{\beta}_j \times \Pr\bigl(\gamma_j = 1 \mid \hat{\beta}_j, \sigma_\varepsilon^2\bigr), \tag{28}$$

where g is the g-prior parameter [Clyde (1999a, 1999b)]. Thus the weights in the BMA forecast can be computed analytically, eliminating the need for a stochastic sampling scheme to approximate (27). The expression (28) treats σ_ε^2 as known. The full BMA estimator can be computed by integrating over σ_ε^2, alternatively one could use a plug-in estimator of σ_ε^2 as suggested by Clyde (1999a, 1999b).

Bayesian model selection Bayesian model selection entails selecting the model with the highest posterior probability and using that model as the basis for forecasting; see the reviews by George (1999) and Chipman, George and McCulloch (2001). With suitable choice of priors, BMA can yield Bayesian model selection. For example, Fernandez, Ley and Steel (2001a) provide conditions on the choice of g as a function of k and T that produce consistent Bayesian model selection, in the sense that the posterior probability of the true model tends to one (the asymptotics hold the number of models K fixed as $T \to \infty$). In particular they show that, if $g = 1/T$ and the number of models K is held fixed, then the g-prior BMA method outlined above, with a flat prior over models, is asymptotically equivalent to model selection using the BIC.

Like other forms of model selection, Bayesian model selection might be expected to perform best when the number of models is small relative to the sample size. In the applications of interest in this survey, the number of models is very large and Bayesian model selection would be expected to share the problems of model selection more generally.

Extension to h-step ahead forecasts The algorithm outlined above does not extend to iterated multiperiod forecasts because the analysis is conditional on X and Z (models for X and Z are never estimated). Although the algorithm can be used to produce multiperiod forecasts, its derivation is inapplicable because the error term ε_t in (23) is modeled as i.i.d., whereas it would be MA$(h-1)$ if the dependent variable were Y_{t+h}^h, and the likelihood calculations leading to (27) no longer would be valid.

In principle, BMA could be extended to multiperiod forecasts by calculating the posterior using the correct likelihood with the MA$(h-1)$ error term, however the simplicity of the g-prior development would be lost and in any event this extension seems not to be in the literature. Instead, one could apply the formulas in (27), simply replacing Y_{t+1}

with Y_{t+h}^{h}; this approach is taken by Koop and Potter (2004), and although the formal BMA interpretation is lost the expressions provide an intuitively appealing alternative to the forecast combining methods of Section 3, in which only a single X appears in each model.

Extension to endogenous regressors Although the general theory of BMA does not require strict exogeneity, the calculations based on the g-prior leading to the average forecast (27) assume that $\{X_t, Z_t\}$ are strictly exogenous. This assumption is clearly false in a macro forecasting application. In practice, Z_t (if present) consists of lagged values of Y_t and one or two key variables that the forecaster "knows" to belong in the forecasting equation. Alternatively, if the regressor space has been orthogonalized, Z_t could consist of lagged Y_t and the first few one or two factors. In either case, Z is not strictly exogenous. In macroeconomic applications, X_t is not strictly exogenous either. For example, a typical application is forecasting output growth using many interest rates, measures of real activity, measures of wage and price inflation, etc.; these are predetermined and thus are valid predictors but X has a future path that is codetermined with output growth, so X is not strictly exogenous.

It is not clear how serious this critique is. On the one hand, the model-based posteriors leading to (27) evidently are not the true posteriors $\Pr(M_k \mid D_T)$ (the likelihood is fundamentally misspecified), so the elegant decision theoretic conclusion that BMA combining is the optimal Bayes predictor does not apply. On the other hand, the weights in (27) are simple and have considerable intuitive appeal as a competitor to forecast combining. Moreover, BMA methods provide computational tools for combining many models in which multiple predictors enter; this constitutes a major extension of forecast combining as discussed in Section 3, in which there were only n models, each containing a single predictor. From this perspective, BMA can be seen as a potentially useful extension of forecast combining, despite the inapplicability of the underlying theory.

5.2. *Survey of the empirical literature*

Aside from the contribution by Min and Zellner (1993), which used BMA methods to combine forecasts from one linear and one nonlinear model, the applications of BMA to economic forecasting have been quite recent.

Most of the applications have been to forecasting financial variables. Avramov (2002) applied BMA to the problem of forecasting monthly and quarterly returns on six different portfolios of U.S. stocks using $n = 14$ traditional predictors (the dividend yield, the default risk spread, the 90-day Treasury bill rate, etc.). Avramov (2002) finds that the BMA forecasts produce RMSFEs that are approximately two percent smaller than the random walk (efficient market) benchmark, in contrast to conventional information criteria forecasts, which have higher RMSFEs than the random walk benchmark. Cremers (2002) undertook a similar study with $n = 14$ predictors [there is partial overlap between Avramov's (2002) and Cremers' (2002) predictors] and found improvements in in-sample fit and pseudo-out-of-sample forecasting performance comparable to those found by Avramov (2002). Wright (2003) focuses on the problem of forecasting four exchange rates using $n = 10$ predictors, for a variety of values of g. For two of the currencies he studies, he finds pseudo-out-of-sample MSFE improvements of as much as 15% at longer horizons, relative to the random walk benchmark; for the other two currencies he studies, the improvements are much smaller or nonexistent. In all three of these studies, n has been sufficiently small that the authors were able to evaluate all possible models and simulation methods were not needed to evaluate (27).

We are aware of only two applications of BMA to forecasting macroeconomic aggregates. Koop and Potter (2004) focused on forecasting GDP and the change of inflation using $n = 142$ quarterly predictors, which they orthogonalized by transforming to principal components. They explored a number of different priors and found that priors that focused attention on the set of principal components that explained 99.9% of the variance of X provided the best results. Koop and Potter (2004) concluded that the BMA forecasts improve on benchmark AR(2) forecasts and on forecasts that used BIC-selected factors (although this evidence is weaker) at short horizons, but not at longer horizons. Wright (2004) considers forecasts of quarterly U.S. inflation using $n = 93$ predictors; he used the g-prior methodology above, except that he only considered models with one predictor, so there are only a total of n models under consideration. Despite ruling out models with multiple predictors, he found that BMA can improve upon the equal-weighted combination forecasts.

6. Empirical Bayes methods

The discussion of BMA in the previous section treats the priors as reflecting subjectively held a priori beliefs of the forecaster or client. Over time, however, different forecasters using the same BMA framework but different priors will produce different forecasts, and some of those forecasts will be better than others: the data can inform the choice of "priors" so that the priors chosen will perform well for forecasting. For example, in the context of the BMA model with prior probability π of including a variable and a g-prior for the coefficient conditional upon inclusion, the hyperparameters π and g both can be chosen, or estimated, based on the data.

This idea of using Bayes methods with an estimated, rather than subjective, prior distribution is the central idea of empirical Bayes estimation. In the many-predictor problem, because there are n predictors, one obtains many observations on the empirical distribution of the regression coefficients; this empirical distribution can in turn be used to find the prior (to estimate the prior) that comes as close as possible to producing a marginal distribution that matches the empirical distribution.

The method of empirical Bayes estimation dates to Robbins (1955, 1964), who introduced nonparametric empirical Bayes methods. Maritz and Lwin (1989), Carlin and Louis (1996), and Lehmann and Casella (1998, Section 4.6) provide monograph and textbook treatments of empirical Bayes methods. Recent contributions to the theory of empirical Bayes estimation in the linear model with orthogonal regressors include

George and Foster (2000) and Zhang (2003, 2005). For an early application of empirical Bayes methods to economic forecasting using VARs, see Doan, Litterman and Sims (1984).

This section lays out the basic structure of empirical Bayes estimation, as applied to the large-n linear forecasting problem. We focus on the case of orthogonalized regressors (the regressors are the principle components or weighted principle components). We defer discussion of empirical experience with large-n empirical Bayes macroeconomic forecasting to Section 7.

6.1. Empirical Bayes methods for large-n linear forecasting

The empirical Bayes model consists of the regression equation for the variable to be forecasted plus a specification of the priors. Throughout this section we focus on estimation with n orthogonalized regressors. In the empirical applications these regressors will be the factors, estimated by PCA, so we denote these regressors by the $n \times 1$ vector F_t, which we assume have been normalized so that $T^{-1}\sum_{t=1}^{T} F_t F_t' = I_n$. We assume that $n < T$ so all the principal components are nonzero; otherwise, n in this section would be replaced by $n' = \min(n, T)$. The starting point is the linear model

$$Y_{t+1} = \beta' F_t + \varepsilon_{t+1}, \tag{29}$$

where $\{F_t\}$ is treated as strictly exogenous. The vector of coefficients β is treated as being drawn from a prior distribution. Because the regressors are orthogonal, it is convenient to adopt a prior in which the elements of β are independently (although not necessarily identically) distributed, so that β_i has the prior distribution G_i, $i = 1, \ldots, n$.

If the forecaster has a squared error loss function, then the Bayes risk of the forecast is minimized by using the Bayes estimator of β, which is the posterior mean. Suppose that the errors are i.i.d. $\mathrm{N}(0, \sigma_\varepsilon^2)$, and for the moment suppose that σ_ε^2 is known. Conditional on β, the centered OLS estimators, $\{\hat{\beta}_i - \beta_i\}$, are i.i.d. $\mathrm{N}(0, \sigma_\varepsilon^2/T)$; denote this conditional pdf by ϕ. Under these assumptions, the Bayes estimator of β_i is

$$\hat{\beta}_i^B = \frac{\int x\phi(\hat{\beta}_i - x)\,\mathrm{d}G_i(x)}{\int \phi(\hat{\beta}_i - x)\,\mathrm{d}G_i(x)} = \hat{\beta}_i + \sigma_\varepsilon^2 \ell_i(\hat{\beta}_i), \tag{30}$$

where $\ell_i(x) = \mathrm{d}\ln(m_i(x))/\mathrm{d}x$, where $m_i(x) = \int \phi(x - \beta)\,\mathrm{d}G_i(\beta)$ is the marginal distribution of $\hat{\beta}_i$. The second expression in (30) is convenient because it represents the Bayes estimator as a function of the OLS estimator, σ_ε^2, and the score of the marginal distribution [see, for example, Maritz and Lwin (1989)].

Although the Bayes estimator minimizes the Bayes risk and is admissible, from a frequentist perspective it (and the Bayes forecast based on the predictive density) can have poor properties if the prior places most of its mass away from the true parameter value. The empirical Bayes solution to this criticism is to treat the prior as an unknown distribution to be estimated. To be concrete, suppose that the prior is the same for all i, that is, $G_i = G$ for all i. Then $\{\hat{\beta}_i\}$ constitute n i.i.d. draws from the marginal distribution m, which in turn depends on the prior G. Because the conditional distribution ϕ is known, this permits inference about G. In turn, the estimator of G can be used in (30) to compute the empirical Bayes estimator. The estimation of the prior can be done either parametrically or nonparametrically.

Parametric empirical Bayes The parametric empirical Bayes approach entails specifying a parametric prior distribution, $G_i(X;\theta)$, where θ is an unknown parameter vector that is common to all the priors. Then the marginal distribution of $\hat{\beta}_i$ is $m_i(x;\theta) = \int \phi(x-\beta)\,\mathrm{d}G_i(\beta;\theta)$. If $G_i = G$ for all i, then there are n i.i.d. observations on $\hat{\beta}_i$ from the marginal $m(x;\theta)$, and inference can proceed by maximum likelihood or by method of moments.

In the application at hand, where the regressors are the principal components, one might specify a prior with a spread that declines with i following some parametric structure. In this case, $\{\hat{\beta}_i\}$ constitute n independent draws from a heteroskedastic marginal distribution with parameterized heteroskedasticity, which again permits estimation of θ. Although the discussion has assumed that σ_ε^2 is known, it can be estimated consistently if $n, T \to \infty$ as long as $n/T \to \text{const} < 1$.

As a leading case, one could adopt the conjugate g-prior. An alternative approach to parameterizing G_i is to adopt a hierarchical prior. Clyde and George (2000) take this approach for wavelet transforms, as applied to signal compression, where the prior is allowed to vary depending on the wavelet level.

Nonparametric empirical Bayes The nonparametric empirical Bayes approach treats the prior as an unknown distribution. Suppose that the prior is the same (G) for all i, so that $\ell_i = \ell$ for all i. Then the second expression in (30) suggests the estimator

$$\hat{\beta}_i^{\mathrm{NEB}} = \hat{\beta}_i + \sigma_\varepsilon^2 \hat{\ell}(\hat{\beta}_i), \tag{31}$$

where $\hat{\ell}$ is an estimator of ℓ.

The virtue of the estimator (31) is that it does not require direct estimation of G; for this reason, Maritz and Lwin (1989) refer to it as a simple empirical Bayes estimator. Instead, the estimator (31) only requires estimation of the derivative of the log of the marginal likelihood, $\ell(x) = \mathrm{d}\ln(m_i(x))/\mathrm{d}x = (\mathrm{d}m(x)/\mathrm{d}x)/m(x)$. Nonparametric estimation of the score of i.i.d. random variables arises in other applications in statistics, in particular adaptive estimation, and has been extensively studied. Going into the details would take us beyond the scope of this survey, so instead the reader is referred to Maritz and Lwin (1989), Carlin and Louis (1996), and Bickel et al. (1993).

Optimality results Robbins (1955) considered nonparametric empirical Bayes estimation in the context of the compound decision problem, in which there are samples from each of n units, where the draws for the ith unit are from the same distribution, conditional on some parameters, and these parameters in turn obey some distribution G. The distribution G can be formally treated either as a prior, or simply as an unknown distribution describing the population of parameters across the different units. In this setting, given G, the estimator of the parameters that minimizes the Bayes risk is the

Bayes estimator. Robbins (1955, 1964) showed that it is possible to construct empirical Bayes estimators that are asymptotically optimal, that is, empirical Bayes estimators that achieve the Bayes risk based on the infeasible Bayes estimator using the true unknown distribution G as the number of units tends to infinity.

At a formal level, if $n/T \to c$, $0 < c < 1$, and if the true parameters β_i are in a $1/n^{1/2}$ neighborhood of zero, then the linear model with orthogonal regressors has a similar mathematical structure to the compound decision problem. Knox, Stock and Watson (2001) provide results about the asymptotic optimality of the parametric and nonparametric empirical Bayes estimators. They also provide conditions under which the empirical Bayes estimator (with a common prior G) is, asymptotically, the minimum risk equivariant estimator under the group that permutes the indexes of the regressors.

Extension to lagged endogenous regressors As in the methods of Sections 3–5, in practice it can be desirable to extend the linear regression model to include an additional set of regressors, Z_t, that the researcher has confidence belong in the model; the leading case is when Z_t consists of lags of Y_t. The key difference between Z_t and F_t is associated with the degree of certainty about the coefficients: Z_t are variables that the researcher believes to belong in the model with potentially large coefficients, whereas F_t is viewed as having potentially small coefficients. In principle a separate prior could be specified for the coefficients on Z_t. By analogy to the treatment in BMA, however, a simpler approach is to replace X_t and Y_{t+1} in the foregoing with the residuals from initial regressions of X_t and Y_{t+1} onto Z_t. The principal components F_t then can be computed using these residuals.

Extensions to endogenous regressors and multiperiod forecasts Like BMA, the theory for empirical Bayes estimation in the linear model was developed assuming that $\{X_t, Z_t\}$ are strictly exogenous. As was discussed in Section 5, this assumption is implausible in the macroeconomic forecasting. We are unaware of work that has extended empirical Bayes methods to the large-n linear forecasting model with regressors that are predetermined but not strictly exogenous.

7. Empirical illustration

This section illustrates the performance of these methods in an application to forecasting the growth rate of U.S. industrial production using $n = 130$ predictors. The results in this section are taken from Stock and Watson (2004a), which presents results for additional methods and for forecasts of other series.

7.1. Forecasting methods

The forecasting methods consist of univariate benchmark forecasts, and five categories of multivariate forecasts using all the predictors. All multistep ahead forecasts (including the univariate forecasts) were computed by the direct method, that is, using a single noniterated equation with dependent variable being the h-period growth in industrial production, Y^h_{t+h}, as defined in (1). All models include an intercept.

Univariate forecasts The benchmark model is an AR, with lag length selected by AIC (maximum lag $= 12$). Results are also presented for an AR(4).

OLS The OLS forecast is based on the OLS regression of Y^h_{t+h} onto X_t and four lags of Y_t.

Combination forecasts Three combination forecasts are reported. The first is the simple mean of the 130 forecasts based on autoregressive distributed lag (ADL) models with four lags each of X_t and Y_t. The second combination forecast is a weighted average, where the weights are computed using the expression implied by g-prior BMA, specifically, the weights are given by w_{it} in (27) with $g = 1$, where in this case the number of models K equals n [this second method is similar to one of several used by Wright (2004)].

DFM Three DFM forecasts are reported. Each is based on the regression of Y^h_{t+h} onto the first three factors and four lags of Y_t. The forecasts differ by the method of computing the factors. The first, denoted PCA(3, 4), estimates the factors by PCA. The second, denoted diagonal-weighted PCA(3, 4), estimates the factors by weighted PCA, where the weight matrix Σ_{uu} is diagonal, with diagonal element $\Sigma_{uu,ii}$ estimated by the difference between the corresponding diagonal elements of the sample covariance matrix of X_t and the dynamic principal components estimator of the covariance matrix of the common components, as proposed by Forni et al. (2003b). The third DFM forecast, denoted weighted PCA(3, 4) is similarly constructed, but also estimates the off-diagonal elements of Σ_{uu} analogously to the diagonal elements.

BMA Three BMA forecasts are reported. The first is BMA as outlined in section with correlated X's and $g = 1/T$. The second two are BMA using orthogonal factors computed using the formulas in Clyde (1999a) following Koop and Potter (2004), for two values of g, $g = 1/T$ and $g = 1$.

Empirical Bayes Two parametric empirical Bayes forecasts are reported. Both are implemented using the n principal components for the orthogonal regressors and using a common prior distribution G. The first empirical Bayes forecast uses the g-prior with mean zero, where g and σ^2_ε are estimated from the OLS estimators and residuals. The second empirical Bayes forecast uses a mixed normal prior, in which $\beta_j = 0$ with probability $1 - \pi$ and is normally distributed, according to a g-prior with mean zero, with probability π. In this case, the parameters g, π, and the scale σ^2 are estimated from the OLS coefficients estimates, which allows for heteroskedasticity and autocorrelation in the regression error (the autocorrelation is induced by the overlapping observations in the direct multiperiod-ahead forecasts).

7.2. Data and comparison methodology

Data The data set consists of 131 monthly U.S. economic time series (industrial production plus 130 predictor variables) observed from 1959:1–2003:12. The data set is an updated version of the data set used in Stock and Watson (1999). The predictors include series in 14 categories: real output and income; employment and hours; real retail, manufacturing and trade sales; consumption; housing starts and sales; real inventories; orders; stock prices; exchange rates; interest rates and spreads; money and credit quantity aggregates; price indexes; average hourly earnings; and miscellaneous. The series were all transformed to be stationary by taking first or second differences, logarithms, or first or second differences of logarithms, following standard practice. The list of series and transformations are given in Stock and Watson (2004a).

Method for forecast comparisons All forecasts are pseudo-out-of-sample and were computed recursively (demeaning, standardization, model selection, and all model estimation, including any hyperparameter estimation, was done recursively). The period for forecast comparison is 1974:7–(2003:12-h). All regressions start in 1961:1, with earlier observations used for initial conditions. Forecast risk is evaluated using the mean squared forecast errors (MSFEs) over the forecast period, relative to the AR(AIC) benchmark.

7.3. Empirical results

The results are summarized in Table 1. These results are taken from Stock and Watson (2004a), which reports results for other variations on these methods and for more variables to be forecasted. Because the entries are MSFEs, relative to the AR(AIC) benchmark, entries less than one indicate a MSFE improvement over the AR(AIC) forecast. As indicated in the first row, the use of AIC to select the benchmark model is not particularly important for these results: the performance of an AR(4) and the AR(AIC) are nearly identical. More generally, the results in Table 1 are robust to changes in the details of forecast construction, for example, using an information criterion to select lag lengths.

It would be inappropriate to treat this comparison, using a single sample period and a single target variable, as a horse race that can determine which of these methods is "best". Still, the results in Table 1 suggest some broad conclusions. Most importantly, the results confirm that it is possible to make substantial improvements over the univariate benchmark if one uses appropriate methods for handling this large data set. At forecast horizons of one through six months, these forecasts can reduce the AR(AIC) benchmark by 15% to 33%. Moreover, as expected theoretically, the OLS forecast with all 130 predictors much performs much worse than the univariate benchmark.

As found in the research discussed in Section 4, the DFM forecasts using only a few factors – in this case, three – improve substantially upon the benchmark. For the forecasts of industrial production, there seems to be some benefit from computing the factors

Table 1
Forecasts of U.S. industrial production growth using 130 monthly predictors: Relative mean square forecast errors for various forecasting methods

Method	1	3	6	12
Univariate benchmarks				
AR(AIC)	1.00	1.00	1.00	1.00
AR(4)	0.99	1.00	0.99	0.99
Multivariate forecasts				
(1) OLS	1.78	1.45	2.27	2.39
(2) Combination forecasts				
Mean	0.95	0.93	0.87	0.87
SSR-weighted average	0.85	0.95	0.96	1.16
(3) DFM				
PCA(3, 4)	0.83	0.70	0.74	0.87
Diagonal weighted PC(3, 4)	0.83	0.73	0.83	0.96
Weighted PC(3, 4)	0.82	0.70	0.66	0.76
(4) BMA				
X's, $g = 1/T$	0.83	0.79	1.18	1.50
Principal components, $g = 1$	0.85	0.75	0.83	0.92
Principal components, $g = 1/T$	0.85	0.78	1.04	1.50
(5) Empirical Bayes				
Parametric/g-prior	1.00	1.04	1.56	1.92
Parametric/mixed normal prior	0.93	0.75	0.81	0.89

Notes: Entries are relative MSFEs, relative to the AR(AIC) benchmark. All forecasts are recursive (pseudo-out-of-sample), and the MSFEs were computed over the period 1974:7–(2003:12-h). The various columns correspond to forecasts of 1, 3, 6, and 12-month growth, where all the multiperiod forecasts were computed by direct (not iterated) methods. The forecasting methods are described in the text.

using weighted PCA rather than PCA, with the most consistent improvements arising from using the nondiagonal weighting scheme. Interestingly, nothing is gained by trying to exploit the information in the additional factors beyond the third using either BMA, applied to the PCA factors, or empirical Bayes methods. In addition, applying BMA to the original X's does not yield substantial improvements. Although simple mean averaging of individual ADL forecasts improves upon the autoregressive benchmark, the simple combination forecasts do not achieve the performance of the more sophisticated methods. The more complete analysis in Stock and Watson (2004a) shows that this interesting finding holds for other horizons and for forecasts of other U.S. series: low-dimensional forecasts using the first few PCA or weighted PCA estimators of the factors forecast as well or better than the methods like BMA that use many more factors.

A question of interest is how similar these different forecasting methods are. All the forecasts use information in lagged Y_t, but they differ in the way they handle information in X_t. One way to compare the treatment of X_t by two forecasting methods is to compare the partial correlations of the in-sample predicted values from the two methods, after controlling for lagged values of Y_t. Table 2 reports these partial corre-

Table 2
Partial correlations between large-n forecasts, given four lags of Y_t

Method	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Combination: mean	1.00									
(2) Combination: SSR-wtd	0.63	1.00								
(3) PCA(3, 4)	0.71	0.48	1.00							
(4) Diagonal wtd PC(3, 4)	0.66	0.56	0.90	1.00						
(5) Weighted PC(3, 4)	0.78	0.57	0.82	0.86	1.00					
(6) BMA/X's, $g = 1/T$	0.73	0.77	0.67	0.71	0.71	1.00				
(7) BMA/PC's, $g = 1$	0.76	0.61	0.62	0.61	0.72	0.82	1.00			
(8) BMA/PC's, $g = 1/T$	0.77	0.62	0.68	0.68	0.77	0.80	0.95	1.00		
(9) PEB/g-prior	0.68	0.56	0.52	0.50	0.60	0.77	0.97	0.85	1.00	
(10) PEB/mixed	0.79	0.63	0.70	0.70	0.80	0.82	0.96	0.99	0.87	1.00

Notes: The forecasting methods are defined in the text. Entries are the partial correlations between the in-sample predicted values from the different forecasting models, all estimated using Y_{t+1} as the dependent variable and computed over the full forecast period, where the partial correlations are computed using the residuals from the projections of the in-sample predicted values of the two forecasting methods being correlated onto four lagged values of Y_t.

lations for the methods in Table 1, based on full-sample one-step ahead regressions. The interesting feature of Table 2 is that the partial correlations among some of these methods is quite low, even for methods that have very similar MSFEs. For example, the PCA(3, 4) forecast and the BMA/X forecast with $g = 1/T$ both have relative MSFE of 0.83, but the partial correlation of their in-sample predicted values is only 0.67. This suggests that the forecasting methods in Table 2 imply substantially different weights on the original X_t data, which suggests that there could remain room for improvement upon the forecasting methods in Table 2.

8. Discussion

The past few years have seen considerable progress towards the goal of exploiting the wealth of data that is available for economic forecasting in real time. As the application to forecasting industrial production in Section 7 illustrates, these methods can make substantial improvements upon benchmark univariate models. Moreover, the empirical work discussed in this review makes the case that these forecasts improve not just upon autoregressive benchmarks, but upon standard multivariate forecasting models.

Despite this progress, the methods surveyed in this chapter are limited in at least three important respects, and work remains to be done. First, these methods are those that have been studied most intensively for economic forecasting, but they are not the only methods available. For example, Inoue and Kilian (2003) examine forecasts of U.S. inflation with $n = 26$ using bagging, a weighting scheme in which the weights are produced by bootstrapping forecasts based on pretest model selection. They report improvements over PCA factor forecasts based on these 26 predictors. As mentioned in the Introduction, Bayesian VARs are now capable of handling a score or more of predictors, and a potential advantage of Bayesian VARs is that they can produce iterated multistep forecasts. Also, there are alternative model selection methods in the statistics literature that have not yet been explored in economic forecasting applications, e.g., the LARS method [Efron et al. (2004)] or procedures to control the false discovery rate [Benjamini and Hochberg (1995)].

Second, all these forecasts are linear. Although the economic forecasting literature contains instances in which forecasts are improved by allowing for specific types of nonlinearity, introducing nonlinearities has the effect of dramatically increasing the dimensionality of the forecasting models. To the best of our knowledge, nonlinear forecasting with many predictors remains unexplored in economic applications.

Third, changes in the macroeconomy and in economic policy in general produces linear forecasting relations that are unstable, and indeed there is considerable empirical evidence of this type of nonstationarity in low-dimensional economic forecasting models [e.g., Clements and Hendry (1999), Stock and Watson (1996, 2003)]. This survey has discussed some theoretical arguments and empirical evidence suggesting that some of this instability can be mitigated by making high-dimensional forecasts: in a sense, the instability in individual forecasting relations might, in some cases, average out. But whether this is the case generally, and if so which forecasting methods are best able to mitigate this instability, largely remains unexplored.

References

Aiolfi, M., Timmermann, A. (2004). "Persistence in forecasting performance and conditional combination strategies". Journal of Econometrics. In press.

Altissimo, F., Bassanetti, A., Cristadoro, R., Forni, M., Lippi, M., Reichlin, L., Veronese, G. (2001). "The CEPR – Bank of Italy indicator". Bank of Italy. Manuscript.

Anderson, T.W. (1984). An Introduction to Multivariate Statistical Analysis, second ed. Wiley, New York.

Artis, M., Banerjee, A., Marcellino, M. (2001). "Factor forecasts for the UK". Bocconi University – IGIER. Manuscript.

Avramov, D. (2002). "Stock return predictability and model uncertainty". Journal of Financial Economics 64, 423–458.

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

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

Bates, J.M., Granger, C.W.J. (1969). "The combination of forecasts". Operations Research Quarterly 20, 451–468.

Benjamini, Y., Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing". Journal of the Royal Statistical Society, Series B 57, 289–300.

Bernanke, B.S., Boivin, J. (2003). "Monetary policy in a data-rich environment". Journal of Monetary Economics 50, 525–546.

Bernanke, B.S., Boivin, J., Eliasz, P. (2005). "Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach". Quarterly Journal of Economics 120, 387–422.

Bickel, P., Klaassen, C.A.J., Ritov, Y., Wellner, J.A. (1993). "Efficient and Adaptive Estimation for Semiparametric Models". Johns Hopkins University Press, Baltimore, MD.

Boivin, J., Ng, S. (2003). “Are more data always better for factor analysis?” NBER. Working Paper No. 9829.

Boivin, J., Ng, S. (2005). “Understanding and comparing factor-based forecasts”. NBER. Working Paper No. 11285.

Brillinger, D.R. (1964). “A frequency approach to the techniques of principal components, factor analysis and canonical variates in the case of stationary time series”. Royal Statistical Society Conference, Cardiff Wales. Invited Paper. Available at http://stat-www.berkeley.edu/users/brill/papers.html.

Brillinger, D.R. (1981). “Time Series: Data Analysis and Theory”, expanded ed. Holden-Day, San Francisco.

Brisson, M., Campbell, B., Galbraith, J.W. (2002). “Forecasting some low-predictability time series using diffusion indices”. CIRANO. Manuscript.

Carlin, B., Louis, T.A. (1996). Bayes and Empirical Bayes Methods for Data Analysis. Monographs on Statistics and Probability, vol. 69. Chapman and Hall, Boca Raton.

Chamberlain, G., Rothschild, M. (1983). “Arbitrage factor structure, and mean-variance analysis of large asset markets”. Econometrica 51, 1281–1304.

Chan, L., Stock, J.H., Watson, M. (1999). “A dynamic factor model framework for forecast combination”. Spanish Economic Review 1, 91–121.

Chipman, H., George, E.I., McCulloch, R.E. (2001). The Practical Implementation of Bayesian Model Selection. IMS Lecture Notes Monograph Series, vol. 38. Institute of Mathematical Statistics.

Clayton-Matthews, A., Crone, T. (2003). “Consistent economic indexes for the 50 states”. Federal Reserve Bank of Philadelphia. Manuscript.

Clemen, R.T. (1989). “Combining forecasts: A review and annotated bibliography”. International Journal of Forecasting 5, 559–583.

Clements, M.P., Hendry, D.F. (1999). Forecasting Non-Stationary Economic Time Series. MIT Press, Cambridge, MA.

Clyde, M. (1999a). “Bayesian model averaging and model search strategies (with discussion)”. In: Bernardo, J.M., Dawid, A.P., Berger, J.O., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 6. Oxford University Press, Oxford.

Clyde, M. (1999b). “Comment on ‘Bayesian model averaging: A tutorial’”. Statistical Science 14, 401–404.

Clyde, M., Desimone, H., Parmigiani, G. (1996). “Prediction via orthogonalized model mixing”. Journal of the American Statistical Association 91, 1197–1208.

Clyde, M., George, E.I. (2000). “Flexible empirical Bayes estimation for wavelets”. Journal of the Royal Statistical Society, Series B 62 (3), 681–698.

Connor, G., Korajczyk, R.A. (1986). “Performance measurement with the arbitrage pricing theory”. Journal of Financial Economics 15, 373–394.

Connor, G., Korajczyk, R.A. (1988). “Risk and return in an equilibrium APT: Application of a new test methodology”. Journal of Financial Economics 21, 255–289.

Cremers, K.J.M. (2002). “Stock return predictability: A Bayesian model selection perspective”. The Review of Financial Studies 15, 1223–1249.

Diebold, F.X., Lopez, J.A. (1996). “Forecast evaluation and combination”. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics, vol. 14. North-Holland, Amsterdam.

Diebold, F.X., Pauly, P. (1987). “Structural change and the combination of forecasts”. Journal of Forecasting 6, 21–40.

Diebold, F.X., Pauly, P. (1990). “The use of prior information in forecast combination”. International Journal of Forecasting 6, 503–508.

Ding, A.A., Hwang, J.T.G. (1999). “Prediction intervals, factor analysis models, and high-dimensional empirical linear prediction”. Journal of the American Statistical Association 94, 446–455.

Doan, T., Litterman, R., Sims, C.A. (1984). “Forecasting and conditional projection using realistic prior distributions”. Econometric Reviews 3, 1–100.

Efron, B., Morris, C. (1973). “Stein’s estimation rule and its competitors – An empirical Bayes approach”. Journal of the American Statistical Association 68, 117–130.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). “Least angle regression”. Annals of Statistics 32, 407–499.

El Karoui, N. (2003). “On the largest eigenvalue of Wishart matrices with identity covariance when n, p and $p/n \to \infty$”. Stanford Statistics Department Technical Report 2003-25.

Engle, R.F., Watson, M.W. (1981). “A one-factor multivariate time series model of metropolitan wage rates”. Journal of the American Statistical Association 76 (376), 774–781.

Favero, C.A., Marcellino, M. (2001). “Large datasets, small models and monetary policy in Europe”. CEPR. Working Paper No. 3098.

Favero, C.A., Marcellino, M., Neglia, F. (2002). “Principal components at work: The empirical analysis of monetary policy with large datasets”. Bocconi University. IGIER Working Paper No. 223.

Federal Reserve Bank of Chicago. “CFNAI background release”. Available at http://www.chicagofed.org/economic_research_and_data/cfnai.cfm.

Fernandez, C., Ley, E., Steel, M.F.J. (2001a). “Benchmark priors for Bayesian model averaging”. Journal of Econometrics 100, 381–427.

Fernandez, C., Ley, E., Steel, M.F.J. (2001b). “Model uncertainty in cross-country growth regressions”. Journal of Applied Econometrics 16, 563–576.

Figlewski, S. (1983). “Optimal price forecasting using survey data”. Review of Economics and Statistics 65, 813–836.

Figlewski, S., Urich, T. (1983). “Optimal aggregation of money supply forecasts: Accuracy, profitability and market efficiency”. The Journal of Finance 28, 695–710.

Forni, M., Reichlin, L. (1998). “Let’s get real: A dynamic factor analytical approach to disaggregated business cycle”. Review of Economic Studies 65, 453–474.

Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2000). “The generalized factor model: Identification and estimation”. The Review of Economics and Statistics 82, 540–554.

Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2003a). “Do financial variables help forecasting inflation and real activity in the EURO area?” Journal of Monetary Economics 50, 1243–1255.

Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2003b). “The generalized dynamic factor model: One-sided estimation and forecasting”. Manuscript.

Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2004). “The generalized factor model: Consistency and rates”. Journal of Econometrics 119, 231–255.

Forni, M., Giannoni, D., Lippi, M., Reichlin, L. (2005). “Opening the black box: Structural factor models with large cross-sections”. Manuscript, University of Rome.

George, E.I. (1999). “Bayesian Model Selection”. Encyclopedia of the Statistical Sciences Update, vol. 3. Wiley, New York.

George, E.I., Foster, D.P. (2000). “Calibration and empirical Bayes variable selection”. Biometrika 87, 731–747.

George, E.I., McCulloch, R.E. (1993). “Variable selection via Gibbs sampling”. Journal of the American Statistical Association 88, 881–889.

George, E.I., McCulloch, R.E. (1997). “Approaches for Bayesian variable selection”. Statistica Sinica 7 (2), 339–373.

Geweke, J. (1977). “The dynamic factor analysis of economic time series”. In: Aigner, D.J., Goldberger, A.S. (Eds.), Latent Variables in Socio-Economic Models. North-Holland, Amsterdam.

Geweke, J.F. (1996). “Variable selection and model comparison in regression”. In: Berger, J.O., Bernardo, J.M., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 5. Oxford University Press, Oxford, pp. 609–620.

Giannoni, D., Reichlin, L., Sala, L. (2002). “Tracking Greenspan: Systematic and unsystematic monetary policy revisited”. ECARES. Manuscript.

Giannoni, D., Reichlin, L., Sala, L. (2004). “Monetary policy in real time”. NBER Macroeconomics Annual 2004, 161–200.

Granger, C.W.J., Ramanathan, R. (1984). “Improved methods of combining forecasting”. Journal of Forecasting 3, 197–204.

Hannan, E.J., Deistler, M. (1988). The Statistical Theory of Linear Systems. Wiley, New York.

Hendry, D.F., Clements, M.P. (2002). “Pooling of forecasts”. Econometrics Journal 5, 1–26.

Hendry, D.F., Krolzig, H.-M. (1999). "Improving on 'Data mining reconsidered' by K.D. Hoover and S.J. Perez". Econometrics Journal 2, 41–58.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T. (1999). "Bayesian model averaging: A tutorial". Statistical Science 14, 382–417.

Inoue, A., Kilian, L. (2003). "Bagging time series models". North Carolina State University. Manuscript.

James, A.T. (1964). "Distributions of matrix variates and latent roots derived from normal samples". Annals of Mathematical Statistics 35, 475–501.

James, W., Stein, C. (1960). "Estimation with quadratic loss". Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1, 361–379.

Johnstone, I.M. (2001). "On the distribution of the largest eigenvalue in principal component analysis". Annals of Statistics 29, 295–327.

Jones, C.S. (2001). "Extracting factors from heteroskedastic asset returns". Journal of Financial Economics 62, 293–325.

Kapetanios, G., Marcellino, M. (2002). "A comparison of estimation methods for dynamic factor models of large dimensions". Bocconi University – IGIER. Manuscript.

Kim, C.-J., Nelson, C.R. (1998). "Business cycle turning points, a new coincident index, and tests for duration dependence based on a dynamic factor model with regime switching". The Review of Economics and Statistics 80, 188–201.

Kitchen, J., Monaco, R. (2003). "The U.S. Treasury staff's real-time GDP forecast system". Business Economics, October.

Knox, T., Stock, J.H., Watson, M.W. (2001). "Empirical Bayes forecasts of one time series using many regressors". NBER. Technical Working Paper No. 269.

Koop, G., Potter, S. (2004). "Forecasting in dynamic factor models using Bayesian model averaging". Econometrics Journal 7, 550–565.

Kose, A., Otrok, C., Whiteman, C.H. (2003). "International business cycles: World, region, and country-specific factors". American Economic Review 93, 1216–1239.

Leamer, E.E. (1978). Specification Searches. Wiley, New York.

Leeper, E., Sims, C.A., Zha, T. (1996). "What does monetary policy do?" Brookings Papers on Economic Activity 2, 1–63.

Lehmann, E.L., Casella, G. (1998). Theory of Point Estimation, second ed. Springer-Verlag, New York.

LeSage, J.P., Magura, M. (1992). "A mixture-model approach to combining forecasts". Journal of Business and Economic Statistics 3, 445–452.

Madigan, D.M., York, J. (1995). "Bayesian graphical models for discrete data". International Statistical Review 63, 215–232.

Maritz, J.S., Lwin, T. (1989). Empirical Bayes Methods, second ed. Chapman and Hall, London.

Miller, C.M., Clemen, R.T., Winkler, R.L. (1992). "The effect of nonstationarity on combined forecasts". International Journal of Forecasting 7, 515–529.

Min, C., Zellner, A. (1993). "Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates". Journal of Econometrics 56, 89–118.

Newbold, P., Harvey, D.I. (2002). "Forecast combination and encompassing". In: Clements, M.P., Hendry, D.F. (Eds.), A Companion to Economic Forecasting. Blackwell Press, Oxford, pp. 268–283.

Otrok, C., Silos, P., Whiteman, C.H. (2003). "Bayesian dynamic factor models for large datasets: Measuring and forecasting macroeconomic data". University of Iowa. Manuscript.

Otrok, C., Whiteman, C.H. (1998). "Bayesian leading indicators: Measuring and predicting economic conditions in Iowa". International Economic Review 39, 997–1014.

Peña, D., Poncela, P. (2004). "Forecasting with nonstationary dynamic factor models". Journal of Econometrics 119, 291–321.

Quah, D., Sargent, T.J. (1993). "A dynamic index model for large cross sections". In: Stock, J.H., Watson, M.W. (Eds.), Business Cycles, Indicators, and Forecasting. University of Chicago Press for the NBER, Chicago. Chapter 7.

Raftery, A.E., Madigan, D., Hoeting, J.A. (1997). "Bayesian model averaging for linear regression models". Journal of the American Statistical Association 92, 179–191.

Robbins, H. (1955). “An empirical Bayes approach to statistics”. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1, 157–164.

Robbins, H. (1964). “The empirical Bayes approach to statistical problems”. Annals of Mathematical Statistics 35, 1–20.

Sargent, T.J. (1989). “Two models of measurements and the investment accelerator”. The Journal of Political Economy 97, 251–287.

Sargent, T.J., Sims, C.A. (1977). “Business cycle modeling without pretending to have too much a priori economic theory”. In: Sims, C., et al. (Eds.), New Methods in Business Cycle Research. Federal Reserve Bank of Minneapolis, Minneapolis.

Sessions, D.N., Chatterjee, S. (1989). “The combining of forecasts using recursive techniques with nonstationary weights”. Journal of Forecasting 8, 239–251.

Stein, C. (1955). “Inadmissibility of the usual estimator for the mean of multivariate normal distribution”. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1, 197–206.

Stock, J.H., Watson, M.W. (1989). “New indexes of coincident and leading economic indicators”. NBER Macroeconomics Annual, 351–393.

Stock, J.H., Watson, M.W. (1991). “A probability model of the coincident economic indicators”. In: Moore, G., Lahiri, K. (Eds.), The Leading Economic Indicators: New Approaches and Forecasting Records. Cambridge University Press, Cambridge, pp. 63–90.

Stock, J.H., Watson, M.W. (1996). “Evidence on structural instability in macroeconomic time series relations”. Journal of Business and Economic Statistics 14, 11–30.

Stock, J.H., Watson, M.W. (1998). “Median unbiased estimation of coefficient variance in a time varying parameter model”. Journal of the American Statistical Association 93, 349–358.

Stock, J.H., Watson, M.W. (1999). “Forecasting inflation”. Journal of Monetary Economics 44, 293–335.

Stock, J.H., Watson, M.W. (2002a). “Macroeconomic forecasting using diffusion indexes”. Journal of Business and Economic Statistics 20, 147–162.

Stock, J.H., Watson, M.W. (2002b). “Forecasting using principal components from a large number of predictors”. Journal of the American Statistical Association 97, 1167–1179.

Stock, J.H., Watson, M.W. (2003). “Forecasting output and inflation: The role of asset prices”. Journal of Economic Literature 41, 788–829.

Stock, J.H., Watson, M.W. (2004a). “An empirical comparison of methods for forecasting using many predictors”. Manuscript.

Stock, J.H., Watson, M.W. (2004b). “Combination forecasts of output growth in a seven-country data set”. Journal of Forecasting. In press.

Stock, J.H., Watson, M.W. (2005). “Implications of dynamic factor models for VAR analysis”. Manuscript.

Wright, J.H. (2003). “Bayesian model averaging and exchange rate forecasts”. Board of Governors of the Federal Reserve System. International Finance Discussion Paper No. 779.

Wright, J.H. (2004). “Forecasting inflation by Bayesian model averaging”. Board of Governors of the Federal Reserve System. Manuscript.

Zellner, A. (1986). “On assessing prior distributions and Bayesian regression analysis with g-prior distributions”. In: Goel, P.K., Zellner, A. (Eds.), Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finietti. North-Holland, Amsterdam, pp. 233–243.

Zhang, C.-H. (2003). “Compound decision theory and empirical Bayes methods”. Annals of Statistics 31, 379–390.

Zhang, C.-H. (2005). “General empirical Bayes wavelet methods and exactly adaptive minimax estimation”. Annals of Statistics 33, 54–100.