DISCUSSION PAPER SERIES

No. 6158

ECONOMIC FORECASTING

Graham Elliott and Allan Timmermann

FINANCIAL ECONOMICS

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP6158.asp**

ISSN 0265-8003

ECONOMIC FORECASTING

Graham Elliott, University of California, San Diego
Allan Timmermann, University of California, San Diego and CEPR

Discussion Paper No. 6158
March 2007

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 6158

March 2007

ABSTRACT

Economic Forecasting*

Forecasts guide decisions in all areas of economics and finance and their value can only be understood in relation to, and in the context of, such decisions. We discuss the central role of the loss function in helping determine the forecaster's objectives and use this to present a unified framework for both the construction and evaluation of forecasts. Challenges arise from the explosion in the sheer volume of predictor variables under consideration and the forecaster's ability to entertain an endless array of functional forms and time-varying specifications, none of which may coincide with the `true' model. Methods for comparing the forecasting performance of pairs of models or evaluating the ability of the best of many models to beat a benchmark specification are also reviewed.

JEL Classification: C53
Keywords: economic forecasting, forecast evaluation and loss function

Graham Elliott
University of California San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
USA
Email: gelliott@weber.ucsd.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=135787

Allan Timmermann
University of California San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
USA
Email: atimmerm@ucsd.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=116464

* We thank Gray Calhoun for excellent research assistance.

Submitted 09 February 2007

Economic Forecasting*

Graham Elliott
UC San Diego

Allan Timmermann
UC San Diego

February 7, 2007

Abstract

Forecasts guide decisions in all areas of economics and finance and their value can only be understood in relation to, and in the context of, such decisions. We discuss the central role of the loss function in helping determine the forecaster's objectives and use this to present a unified framework for both the construction and evaluation of forecasts. Challenges arise from the explosion in the sheer volume of predictor variables under consideration and the forecaster's ability to entertain an endless array of functional forms and time-varying specifications, none of which may coincide with the 'true' model. Methods for comparing the forecasting performance of pairs of models or evaluating the ability of the best of many models to beat a benchmark specification are also reviewed.

1 Introduction

Forecasting problems are ubiquitous in all areas of economics and finance where agents' decisions depend on the uncertain future value of one or more variables of interest. When a household decides how much labor to supply or how much to save for a rainy day, this presumes an ability to forecast a stream of future wages and returns on savings. Similarly, firms' choice of when to invest, how much to invest and how to finance it (the capital structure decision) depends on their forecasts of future cash flows from potential investments, future stock prices and interest rates. Indeed, all present value calculations, and hence the vast majority of questions in asset pricing, have embedded in them forecasts of future cash flows generated by uncertain payoff streams. In public finance, decisions on whether to go ahead with large infrastructure projects such as the construction of a new bridge or a tunnel require projecting traffic flows and income streams over the project's lifetime which may well be several decades.

Recent research has seen a virtual revolution in how economists compute, apply and evaluate forecasts. This research has occurred as a result of extensive developments in information technology that have opened access to thousands of new potential predictor variables (including tick-by-tick trading data, disaggregate survey forecasts and real-time macroeconomic data) and a wealth of new techniques that facilitate search over and estimation of the parameters of increasingly complicated forecasting models. Questions such as which particular predictor variables to

*We thank Gray Calhoun for excellent research assistance.

include, which functional form to use and how to weight old versus more recent data have become an essential part of forecast construction and evaluation.

Economic forecasting is unique in that forecasters are forced to 'show their hand' in real time as they generate their forecasts. Future outcomes of most predicted variables are observed within a reasonable period of time, so a direct sense of how well a forecasting model performed can be gained. If forecasting performance is poor as a result of overfitting or using a bad model, this will become clear to the forecaster once data on realizations of the predicted variable is revealed. This real time feedback may in turn lead to a change in the forecasting model itself, thus posing unique challenges to the process of evaluating how fast the forecaster is learning over time. This is in stark contrast to many econometric problems. For example, an verification of an estimate of the effect of schooling on wages may take generations before evaluation. In many such problems we do not obtain an objective confirmation of how good the original estimate is since we do not have reference data for evaluating the economic prediction.

Often the result of the feedback from forecasts has been disheartening, both to econometricians trying to utilize data as efficiently as possible and to economists whose theories result in predictions that appear unable to explain as much of the variation in the data as they had hoped. To take one example, a seemingly simple task such as estimating the weights on different models in least squares forecast combination regressions is commonly outperformed on real data by using a simple equal-weighted average of forecasts (Clemen (1989)). An infamous result of Meese and Rogoff (1983) shows that despite a great deal of theoretical work on exchange rates−and even with the benefit of using future data suggested by theory as relevant−the random walk 'no change' prediction cannot be beaten. This result has to a great extent held up for exchange rate forecasts of the not too distant future.

While the performance of a forecasting model often can be observed fairly quickly, only limited conclusions can be drawn from the model's historical track record. Forecasting models are best viewed as greatly simplified approximations of a far more complicated reality and need not reflect causal relations between economic variables. Indeed, simple mechanical forecasting schemes−such as the random walk−are often found to perform well empirically although they do not provide new economic insights into the underlying variable (Clements and Hendry (2002)).

In a unified framework this paper provides an understanding of the properties, construction and evaluation of economic forecasts. Our objective is to help explain differences among the many approaches used by various researchers and understand the breadth of results reported in the empirical forecasting literature.

Our coverage emphasizes the importance of integrating economic forecasts (including model specification, variable selection and parameter estimation) in a decision theoretical framework. This is, in our view, the defining characteristic of economic forecasts. From this perspective, forecasts do not have any intrinsic value and are only useful in so far as they help improve economic decisions. What constitutes a good forecast depends on how costly various prediction errors are to the forecaster and hence reflects both the forecaster's preferences and the manner in which forecasts are mapped into economic decisions. Economic forecasting is not an exercise in modelling the data disjoint from the purpose of the forecast provision. Section 2 illustrates these points initially through two examples from economics and finance.

We next provide a formal statement of the forecasting problem. Section 3 reviews both classical

and Bayesian approaches to forecasting and introduces the individual components of the economic forecasting problem such as the forecaster's prediction model, the underlying information set as well as the forecaster's loss function. Although often only treated implicitly, the loss function is essential to all forecasting problems and so we devote Section 4 to a deeper discussion of various types of loss functions and the restrictions and assumptions they embody.

Using the decision theoretic framework set out in Section 3, Section 5-7 review several issues that arise in the practical construction of economic forecasts. Each of these topics has been active areas of research in recent years. An over-arching problem in economic forecasting is the myriad of data that a forecaster could potentially employ. Estimating models with a large number of parameters relative to the sample size undermines one of the central methods of econometrics—OLS justified through properties such as asymptotic efficiency of the parameter estimates—and opens the possibility that other estimation techniques are better suited to the task of constructing forecasts. In concert with differences over loss functions this provides a partial explanation of the myriad of estimation methods seen in practice.

Section 5 addresses the choice of functional form of the forecasting model. Lack of guidance from economic theory is often an issue and so the functional form is commonly chosen on grounds such as empirical 'fit' or an ability to capture certain episodes in the historical data sample. We review several methods aimed at approximating unknown functional forms in a parsimonious, yet flexible manner−a task made essential by the short samples available in most forecasting applications.

A final problem is related to how the forecasting model and the underlying data evolve over time. When forecasting models are viewed as simple approximations to a complex and evolving reality that changes because of shifts in legislation, institutions and technology−or even wars and natural catastrophes−it is to be expected that the 'true' but unknown data generating process changes over time. In the forecasting literature this has been captured through various approaches that deal with model- and parameter instability. Since all estimation techniques essentially average over past data to obtain a forecasting model, this raises the problem of exactly how to choose the data sample and how to weight 'old' versus 'new' data. Other approaches attempt to directly model breaks in the parameters or model in order to increase the effective data sample. These are covered in Section 6.

An important part of the analysis of economic forecasts is to assess just how good they are. Until recently, forecasts were largely evaluated without the use of standard errors that account for parameter estimation error and model specification search. It is well understood that data mining programs that search over many models tend to overfit and hence inflate in-sample estimates of forecasting performance. However, little or no account is typically made for such model search that precede the analysis. Standard practice for dealing with data mining has been to hold back some data and check whether the forecasting model still performed well in future out-of-sample periods. Again, often average losses are compared without any regard to pre-testing biases. Recent work has resulted in methods that account for sampling error in various forecasting situations. These are reviewed in section 7.

Economic theory rarely identifies a single forecasting model that works well in practice and leaves open many degrees of freedom in forecast construction. The resulting plethora of economic forecasting models has given rise to procedures for forecast comparisons that can handle even situations with a very large set of models. An alternative to evaluating particular models and

attempting to select a single dominant model is to average over various forecasting methods. Both forecast comparison and forecast combination are reviewed in Section 8. Section 9 provides an empirical analysis of forecasts of inflation and stock returns. Finally Section 10 concludes.

2 Forecasts and Economic Decisions: Two Examples

We start with an illustration of how economic forecasts are embedded in the economic decision process using two examples from macroeconomics and finance.

2.1 Central Bank Forecasts

First consider the forecasting problem encountered by a central bank whose main role is to set interest rates and whose objectives are defined over inflation and economic activity as measured, e.g., by output growth and the unemployment rate. Because future values of output growth, unemployment and inflation are uncertain, the bank's interest rate decisions must in practice depend on its forecast of these variables as well as its understanding of how they will be affected by current and future interest rates. In the analysis by Svensson (1997), the central bank's inflation targeting implies targeting inflation forecasts and so the forecast acts as an intermediate target.

As part of formulating a forecasting model, the central bank must decide which variables are helpful in predicting future economic activity and inflation. Forecasts of output growth and inflation could be linked via the Phillips curve. Monetary theories of inflation may suggest one set of variables, while the theory of the term structure of interest rates would suggest others. Variables such as new housing starts, automobile sales, new credit lines, monetary growth, personal bankruptcies, capacity utilization, unemployment rates etc. must also be assessed. Even after determining which predictor variables to include in the forecasting model, questions such as how to measure a particular variable or which dynamic lag structure to use must also be addressed.

When more than one forecasting model is available, which model to use—or whether to use a combination of forecasts from separate models—also become an issue. Many central banks make use of what Pagan (2003) refers to as a diverse 'suite of models'. Indeed, according to Pagan, at some stage the Bank of England made use of 32 different models (although not all of these were used in forecasting), ranging from VARs, time-varying component models to factor models. Similar evidence on the use of multiple models by other central banks is reported by Sims (2002).

Because central banks' quantitative models serve the dual purposes of being used in policy analysis and forecasting, a trade-off is likely to exist between the models' theoretical and empirical coherence (Pagan (2003)). For example, theory may impose constraints on the behavior of equilibrium error correction mechanisms such as the gradual disappearance of the output gap. The existence of such a trade-off means that the central bank may choose not to maximize the pure statistical 'fit' of the forecasting model when this is deemed to compromise the model's theoretical coherence.

Central bankers come with certain subjective views about how the economy operates which they may wish to impose on their forecasting model—a theme that naturally leads to Bayesian forecasting methods. Should the central bank use a simple vector autoregression (VAR) fitted to historical data and thus use a model tailored to fit historical features of the data? Should it use a

more theoretically coherent dynamic stochastic general equilibrium (DSGE) model? Or, should it use some combination of the two? If the central bank adjusts forecasts from a formal model using judgemental information, an additional issue arises, namely how much weight to assign to the data versus the judgemental forecast. Implicitly or explicitly, such weights will reflect the bank's prior beliefs.

Quite frequently forecasters find themselves in situations that differ in important regards from the historical sample used to estimate their forecasting models. Pagan (2003) refers to the difficulties and uncertainties the Bank of England faced in their forecasts following the events of 11 September, 2001. Indeed, an important part of maintaining a good forecasting model is to monitor and evaluate its performance both historically and in real time. Because past forecast errors have often been found to have predictive power over future errors, monitoring for serial correlation in forecast errors potentially offers a simple way to improve upon a forecast. More generally, if the process generating the predicted variable is subject to change, it is conceivable that a forecasting model that performed well historically will fail to do so in the more recent past.

2.2 Portfolio Allocation Decisions

As a second example, consider an investor's portfolio allocation decisions. Under mean-variance preferences these will depend on the investor's forecasts of a set of assets' mean returns as well as their variances and covariances. Under more general preferences, higher order moments such as skew and kurtosis and possibly the full return distribution may also matter to the investor. In either case, the investor must be able to produce quantitative forecasts and trade off portfolios with different probability distributions through a loss function. The investor must also decide how to incorporate predictability into his actions. How predictability maps into portfolio allocations will depend on the form of the prediction signal—i.e., is the sign of asset returns predictable or only their magnitude.

Even if means and variances of asset returns are believed to be constant and hence essentially unpredictable, their estimates can still be surrounded by considerable uncertainty. As a consequence, how the moments are estimated in practice can have a large effect on the portfolio weights. Due to estimation error, often the raw estimates are shrunk towards their values implied by a simple benchmark model such as the CAPM (Ledoit and Wolf (2003)). Alternatively, the investor's choice variables—the portfolio weights—can be restricted through short sales restrictions and maximum holding limits (Jagannathan and Ma (2003)).

If the mean and variance of returns are allowed to depend on time-varying state variables, the question immediately arises which state variables to select among interest rates (levels and spreads), macroeconomic activity variables, technical variables such as price momentum or reversals, valuation measures such as the price-earnings or book-to-market ratios or the dividend yield etc. Asset pricing theory provides little guidance to the exact identity of the relevant state variables; this raises several questions such as how to avoid over-fitting the forecasting model—a risk always encountered when multiple prediction models are considered—and how to assess the forecasting models' performance against a benchmark strategy such as simply holding the market portfolio.

Another problem that is more unique to forecasting models for financial returns is that any predictability patterns that do not capture time-varying risk premia must, if markets are efficient,

be non-stationary because their discovery should lead to their self-destruction once investors act to take advantage of such predictability. For example, there is evidence suggesting that popular models for predicting stock returns based on the dividend yield ceased to be successful at some point during the nineties, perhaps because of changes in firms' dividend payout and share repurchase practices or perhaps because investors incorporated earlier evidence of predictability. Only if a model's forecasting performance is tracked carefully through time can this sort of evidence be uncovered.

3 A formal statement of the forecasting problem

Forecasting can broadly be viewed as the process involved in providing information on future values of one or more variables of interest. Towards this end, the variables of interest must be defined and the information set containing known data that will be considered to construct the forecast must also be determined. The latter can be problematic in practice since we often have very large amounts of information that could be used as inputs to the forecasting model.

Other elements are important to the process of deriving a forecast, some of which are often ignored to some extent even though implicitly they still play a role.

The first element is the loss function. No forecast is going to be correct always, so a specification of how costly different mistakes are is needed to guide the procedure. This helps to avoid−or at least lower the probability of−worst case scenarios. We examine the loss function in the next section.

The second element is the family of forecasting models to be considered, which guides the selection of possible methods used for forecast construction. Models may be parametric, semi-parametric or non-parametric. A parametric model is a model which is fully specified up to a finite dimensional unknown vector. A non-parametric model can be considered as a model with an infinite dimensional set of unknown parameters. Semi-parametric models fit in the middle. Since little is often known about the form of the 'true' forecasting model, ideally one would specify the forecasting model nonparametrically. However, this ignores the short data samples and the large dimension of the set of potential predictor variables in most empirical forecasting problems. In practice a flexible parametric forecasting model is often the best one can hope to achieve.

A third element concerns which type of information to report for the outcome of interest. We could report a single number (point estimate), a range estimate or perhaps an estimate of the full probability distribution of all possible values. Most of the theory of frequentist forecasting has been directed towards point forecasting. A more recent literature has examined interval forecasts or forecasts of the conditional distribution of the variable of interest rather than a summary statistic. From a Bayesian perspective similar issues arise, although it is natural in this approach to provide the full predictive distribution.

3.1 Notation

Throughout we let Y be the random variable that generates the value to be forecast. To begin with, we restrict attention to point forecasts, f, which are functions of the available data at the time the forecast is made. Hence, if we collect all relevant information at the time of the forecast into the outcome z of the random variable Z, then the forecast is $f(z)$. Discovering which variables

are informative from a forecasting standpoint is practically important. We examine this in greater depth later on but for now simply think of this information as being incorporated into some random variable that generates our data. Exactly how z maps into the forecast $f(z)$ depends on a set of unknown parameters, θ, that typically have to be estimated from data.

The loss function is a function $\mathcal{L}(f, Y, Z)$ which maps the data, outcome and forecast to the real number line, i.e. for any set of values for these random variables the loss function returns a single number. The loss function describes in relative terms how bad any forecast might be given the outcome and possibly other observed data that accounts for any state dependence in the loss. The loss function and its properties are examined in greater detail in Section 4.

3.2 Optimal Point Forecasts

The forecaster's objective is to use data−outcomes of the random variable Z−to predict the value of the random variable Y. Let T be the date where the forecast is computed and let h be the forecast horizon. Then Z is defined on the information set $\mathcal{F}_T$, while the outcome is defined on $\mathcal{F}_{T+h}$. Z comprises a sequence $\{Z_t\}_{t=1}^T$ that typically includes past values of the variable to be forecast, as well as other variables, so often $\{Z\}_{t=1}^T = \{Y_t, X_t\}_{t=1}^T$. The outcome, Y, may be a vector or could be univariate.

The forecaster's objective can be reduced to finding a decision rule $f(z)$ that will be used to 'choose' a value for the outcome of Y. The forecast is the decision rule. For any decision rule, there is an attached 'risk',

$$R(\theta, f) = E_{Y,Z}\left[\mathcal{L}(f(Z), Y, Z)\right]. \tag{1}$$

Here the expectation is over the data Z holding the forecast rule f and the unknown parameters, θ, fixed (which is why the risk is a function of θ and the particular rule chosen, f). That this is a function of θ will become clear below. A sensible rule has low risk and minimizes expected loss or equivalently maximizes expected utility.[1]

Assuming the existence of a density for both Y given Z and for Z (denoted $p_Y(y|z)$ and $p_Z(z)$), we can write the risk as

$$R(\theta, f) = \int\int \mathcal{L}(f(z), y, z) p_Y(y|z, \theta) p_Z(z|\theta) dy dz. \tag{2}$$

It is this risk that forecasting methods — methods for choosing $f(z)$ — attempt to control and minimize.

Forecasts are generally viewed as 'poor' if they are far from the observed realization of the outcome variable. However, as is clear from (2), point forecasts aim to estimate not the realization of the outcome but rather a function of its distribution. The inner integral in the risk function (2) is $E_Y\left[\mathcal{L}(f(z), Y, Z)|Z\right]$ which transforms the loss function from a function relating the outcome variable (Y) to a function ($\mathcal{L}$) in a function space to a problem relating the parameters of the model (θ) to the forecast ($f(z)$). Because we are estimating a function of the parameters, we are concerned with a quality of the distribution and not with the outcome itself.

[1] This representation of the problem limits further choices of the loss function and requires assumptions on the underlying random variables to ensure that the risk exists.

3.3 Classical Approach

Rather than focussing on the full risk function in (2), the classical approach to minimizing risk only considers the inner integral $\int \mathcal{L}(f(z), y, z)p_Y(y|z, \theta)dy$. This expectation is taken with respect to the outcome variable holding both the data used to construct the forecast (z) and the parameters θ fixed. For a given conditional density for Y (i.e. a model for Y conditional on $Z = z$, $p_Y(y|z, \theta)$) and a given loss function $\mathcal{L}$ we can minimize this directly for a rule $f(z)$.

If we are able to differentiate under the integral, we get the forecaster's first order condition

$$\frac{d}{df(z)} \int \mathcal{L}(f(z), y, z)p_Y(y|z, \theta)dy = E[\mathcal{L}'(f(Z), Y, Z)|Z = z] = 0. \tag{3}$$

Assuming squared loss in the difference between the forecast and outcome, we have

$$\int \mathcal{L}(f(z), y, z)p_y(y|z, \theta)dy = E[Y - f(Z)|Z = z]^2. \tag{4}$$

This is minimized by choosing $f(z) = E[Y|Z = z]$, i.e. the conditional mean as a function of the data, the forecasting model and its parameters, θ.

In practice the parameters θ are almost always unknown and so the second step in the classical approach involves selecting a 'plug-in' estimator for θ. The resulting estimator $\hat{\theta}(z)$ is a function of the data z and hence the forecasting rule $f(z)$ is only a function of the observable data.

For example under squared loss and $z = \{y_t, x_t\}_{t=1}^T$, when the conditional mean of Y_{T+1} is $\theta' x_T$ we might use OLS estimates from a regression of y_{t+1} on x_t over the available sample as the plug-in estimator for θ. The forecast is then $f(z) = (\sum_{t=2}^T x'_{t-1}y_t)'(\sum_{t=2}^T x'_{t-1}x_{t-1})^{-1}x_T$. Alternative plug-in estimators are discussed in detail below.

In choosing between plug-in estimators, one approach is to examine the risk functions for the various methods, $R(\theta, f)$. These are functions of both the method f and the parameters θ. Typically no risk function dominates uniformly over all θ, i.e. some are better for values of θ but work less well for other values. The classical forecaster could then choose a method that minimizes worst case risk or could alternatively consider a weighting function over θ, choosing the best method for that particular weighting. Denoting the weighting function by $\pi(\theta)$, one would choose the method that minimizes $\int R(\theta, f)\pi(\theta)d\theta$, i.e. the risk averaged over all models that are thought to be important.

3.4 Bayesian Approach

The Bayesian approach starts with the idea of risk averaged over all possible models by defining Bayes risk as

$$r(\pi, f) = \int R(\theta, f)\pi(\theta)d\theta. \tag{5}$$

If the forecast $f(z)$ minimizes Bayes risk, it is a Bayes decision rule. Notice the similarity to choosing the plug-in method that minimizes average risk over relevant models in the classical approach.

To construct a Bayes decision rule (i.e. a forecast) we require a weighting or prior $\pi(\theta)$ over the parameters of the model that tells us which parameter values are likely and which are not. We also require models for the random variables underlying the data, namely $p_Y(y_{T+1}|z_T)$ and $p_Z(z_T|\theta)$, which allow calculation of the posterior $\pi(\theta|z) = p_Z(z|\theta)\pi(\theta)/m(z)$ where $m(z) = \int p_Z(z|\theta)\pi(\theta)d\theta$.

Bayesian forecasts are then chosen conditional on observing $Z = z$, using the posterior. To see intuitively why this works in the sense of delivering a forecast that minimizes Bayes risk, consider expected loss conditional on $Z = z$. A Bayesian approach would be to minimize this for any $Z = z$, yielding a rule $f(z)$. Since this rule minimizes the conditional expected loss, it also minimizes the unconditional expected loss, i.e. it minimizes Bayes risk and is hence a Bayesian decision rule.

3.5 Relating the Methods

The complete class theorem tells us that if the classical method does not correspond to a Bayesian procedure for some prior, then it is inadmissible. Under the same problem setting (i.e. for identical loss function and densities), one could therefore find a Bayesian method with equal or smaller risk than the classical procedure for all possible values of θ. Conversely, if there is equivalence between the two methods, then the classical approach cannot be beaten.

The first problem that can arise in the classical setting is that the 'plug in' method of constructing $f(z)$ using the estimate $\hat{\theta}(z)$ is ad hoc. Often forecasters choose estimators that yield nice properties of the parameters themselves. For example estimators that are consistent, asymptotically normal and asymptotically efficient for θ may be employed. However, because the goal of forecasting is not to estimate θ, but to construct the density, $f(z)$, such methods may not yield good forecasting rules even for reasonable weighting functions $\pi(\theta)$.

Some practical considerations cloud the picture. In practice, differences between the risks of optimal and ad hoc methods need not be large. Moreover, such comparisons require that the model be correctly specified which is against both the spirit and practice of modern forecasting. Still, the Bayesian approach offers a construction method that is guaranteed to be admissible for the specified model. Even if the true model is not necessarily the one used to construct the forecasts, provided that the forecasting model is close to the true model we are assured that we use a method that works well for this possible true model.

It follows from this discussion that it is difficult to find optimal solutions even for very simple forecasting problems. Furthermore, often simple methods such as OLS are not optimal. Hence for various combinations of distributions of the data and values of the parameters of the model there is leeway for alternative methods to dominate. This lack of a single dominant approach explains much of the interest in different forecasting approaches seen in the last two decades.

3.6 Density Forecasts

Finding a solution to minimizing $E_Y\left[\mathcal{L}(f(z), Y, Z)|Z\right]$ over decision rules involves choosing a forecast, i.e. some function of the forecast density, that minimizes risk

$$f^*(z) = \arg\min_{f(z)} \int \mathcal{L}(f(z), y, z) p_Y(y|z) dy.$$

An alternative to the provision of a point forecast is to provide the predictive density $p_Y(y|z)$. If the loss function is of the squared loss variety then all that the forecaster does on receipt of the forecast density, $p_Y(y|z, \theta)$ is to take the mean of the distribution and use this as the forecast. More generally, however, numerical integration over the density forecasts is required to evaluate the risk.

Forecasters with different loss functions will generally construct different optimal forecasts even though the density for the data is the same for each of them. Under some loss functions (e.g., piece-wise linear loss), the optimal forecast will be a quantile of the outcome variable where the quantile depends on the degree of loss asymmetry. Two forecasters with different loss functions in this family will want different quantiles of the distribution, so an agency that merely reports a single number could never give them both the optimal forecast. It would be sufficient to provide the entire distribution (density forecast) because this has all the quantile information and hence works for any piece-wise linear loss function.

The Bayesian equivalent to this approach is to provide the predictive density via the posterior distribution for θ. The Bayesian chooses $f(z)$ to minimize

$$\begin{aligned} r(\pi, a) &= \int \left(\int \left\{ \int \mathcal{L}(f(z), y, z) p_Y(y|z, \theta) dy \right\} p_Z(z|\theta) \pi(\theta) d\theta \right) dz \\ &= \int m(z) \left(\int \left\{ \int \mathcal{L}(f(z), y, z) p_Y(y|z, \theta) dy \right\} \pi(\theta|z) d\theta \right) dz \\ &= \int m(z) \left(\int \left\{ \mathcal{L}(f(z), y, z) \int p_Y(y|z, \theta) \pi(\theta|z) d\theta \right\} dy \right) dz. \end{aligned} \tag{6}$$

Here we allowed the integral over θ to pass by the loss function. The loss function is not explicitly a function of the parameters. Now $\int p_Y(y|z, \theta)\pi(\theta|z)d\theta = p_Y(y|z)$ is the predictive density obtained by integrating over θ using $\pi(\theta|z)$ as weights.

4 Loss Functions

Short of the special (and ultimately uninteresting) case with perfect foresight it will not be possible to find a method that always sets $f(z)$ equal to the outcome y. A formal method of trading off potential forecast errors of different signs and magnitudes is therefore required. The loss function $\mathcal{L}(f(Z), Y, Z)$ maps the data, outcome and forecast to the real number line, i.e. for any set of values of these random variables the loss function returns a single number. The loss function describes in relative terms how costly any forecast is given the outcome and possibly other observed data.

It is conventional practice to assume that the economic loss only depends on the forecast error, $Y - f$, but this is likely to be far too restrictive an assumption in some situations. Patton and Timmermann (2006) find that it is difficult to understand the Fed's so-called Green Book forecasts of output growth if the loss is restricted to only depend on the forecast error. Rationalizing the Fed's forecasts requires not only that over-predictions of output growth are costlier than under-predictions, but that overpredictions are particularly costly during periods of low economic growth. This finding makes sense if the cost of an overly tight monetary policy is particularly high during periods with low economic growth where it may cause or extend a recession.[2]

Forecasters thus must pay attention to how errors will affect their results, which means constructing a mathematical representation of potential losses. A natural foundation for a loss function is a utility function that involves both the outcome and the forecast. For a given utility function

[2] Central banks commonly state a desire to keep inflation within a band of 0 to 2% per annum. Inflation within this band might be regarded as successful outcomes, whereas deflation or inflation above 2% are viewed as failures. Again this is indicative of a non-standard loss function.

$U(f(Z), Y, Z)$, we can set the loss $\mathcal{L}(f, Y, Z) = -U(f, Y, Z)$ in order to see how one might elicit the loss function (Granger and Machina (2006), Skouras (2001)).

It is possible that some loss functions cannot be tied to any utility function for the decision problem. For example Elliott and Lieli (2006) derive the loss function from first principles for binary decision and outcome variables. This setup does not admit commonly applied loss functions for any possible utility function.

For money managers, asymmetric loss may be linked to loss aversion or concerns related to liquidity, bankruptcy or regulatory constraints. Under the Basel II accord, banks are required to forecast their Value at Risk which is a measure of how much they expect to lose with a certain probability such as 1%. Capital provisions are affected by this forecast: Overpredicting the Value at Risk ties up more capital than necessary, while underpredicting it could lead to regulatory penalties and the need for increased future capital provisions.

4.1 Commonly used Loss Functions

Loss functions satisfy a set of common properties. Since the perfect forecast, $f(z) = y$, is the best possible outcome, loss functions achieve a minimum at this point. Thus loss is typically bounded from below at the point where the forecast equals the outcome. In practice loss functions are usually normalized so $\mathcal{L}(y, y, z) = 0$ for all y and z. For this to be a unique minimum we have $\mathcal{L}(f(z), y, z) > 0$ for all $f(z) \neq y$.

Restrictions on the form of the loss function are also needed to make sense of the ideas of minimizing risk, in particular we require that the expected loss exists. The existence of expected loss depends both on the loss function and on the conditional distribution of the outcome variable. Recall that expected loss is

$$E_Y[\mathcal{L}(f(z), Y, z] = \int \mathcal{L}(f(z), y, z) p_Y(y|z) dy. \tag{7}$$

Hence issues with the existence of expected loss revolve around how large the loss becomes for tail behavior of the predicted variable.

Symmetry of the loss function is the constraint that, for all d,

$$\mathcal{L}(y - d, y, z) = \mathcal{L}(y + d, y, z). \tag{8}$$

Most popular loss functions in economic applications are symmetric.

By far the most commonly employed loss function is mean squared error (MSE) loss,

$$\mathcal{L}(f(Z), Y, Z) = (Y - f(Z))^2. \tag{9}$$

This is a particularly tractable loss function since there are no unknown parameters and the optimal forecast is simply the conditional mean of Y : $f^*(Z) = E[Y|Z]$. Hence under MSE loss the classical 'plug in' approach to forecasting simply involves estimating the conditional mean of Y. This relates naturally to regression analysis and the greater part of econometric theory. Under the Bayesian approach the optimal forecast is the mean of the predictive density, $f(z) = \int Y P_Y(Y|z) dy$.

Mean absolute (MAE) loss is also very common:

$$\mathcal{L}(f(Z), Y, Z) = |Y - f(Z)|. \tag{10}$$

For all continuous distributions the optimal forecast is the conditional median of Y.

These two loss functions are nested in the family of loss functions considered by Elliott, Komunjer and Timmermann (2005)

$$\mathcal{L}(f(Z), Y, Z; p, \alpha) \equiv [\alpha + (1 - 2\alpha) \cdot 1(Y - f(Z) < 0)] \cdot |Y - f(Z)|^p. \tag{11}$$

For $p = 1$ this gives the lin-lin (piece-wise linear) loss function which nests MAE loss when $\alpha = 0.5$. For $p = 2$ the asymmetric quadratic loss function, which nests MSE loss when $\alpha = 0.5$, is obtained. Optimal forecasts from the lin-lin loss function are conditional quantiles, while those from the asymmetric quadratic loss function are expectiles (Newey and Powell (1987)).

Varian (1974) studied linex loss,

$$\mathcal{L}(f(Z), Y, Z) = \exp(b(Y - f(Z))) - b(Y - f(Z)) - 1, \tag{12}$$

where $b > 0$ is a parameter that controls the degree of asymmetry. If $b > 0$, large underpredictions ($f < y$) are costlier than overpredictions of the same magnitude, with the relative cost increasing as the magnitude of the forecast error rises. Conversely, for $b < 0$, large overpredictions are costlier than equally large underpredictions.

4.2 Asymmetric Loss

Most empirical work in forecasting assumes MSE loss. Apart from the fact that using MSE loss represents 'conventional practice', this choice is likely to reflect difficulties in putting numbers on the relative cost of over- and underpredictions. Construction of a loss function requires a deep understanding of the forecaster's objectives and this may not always be easily accomplished.

Still, the implicit choice of MSE loss by the majority of studies in the forecasting literature seems difficult to justify on economic grounds. As noted by Granger and Newbold (1986, p. 125), ".. an assumption of symmetry about the conditional mean ... is likely to be an easy one to accept ... an assumption of symmetry for the cost function is much less acceptable."

Papers that explicitly consider the forecasting problem under asymmetric loss include Granger (1969, 1999), Varian (1974), Zellner (1986), Ito (1990), West, Edison and Cho (1993), Weiss (1996), Christoffersen and Diebold (1997), Batchelor and Peel (1998), Granger and Pesaran (2000), Pesaran and Skouras (2002), Artis and Marcellino (2001), Elliott, Komunjer and Timmermann (2005, 2006), Capistran (2005) and Patton and Timmermann (2005, 2006).

Many economic considerations give rise to asymmetric loss. Consider a firm involved in forecasting the sales of a new product. Overpredicting sales leads to inventory and insurance costs and ties up capital. It may also give rise to discounts needed to sell the remaining surplus. Such costs are mostly known or can at least be estimated with a fair degree of precision. Contrast this with the cost of underpredicting sales which leads to stock-out costs, loss of goodwill and reputation and lost current and future sales. Such costs are less tangible and can be difficult to quantify.

Another reason for asymmetric loss arises when the forecast is itself best viewed as a signal in a strategic game that explicitly accounts for the forecast provider's incentives. The papers by Ehrbeck and Waldmann (1996), Scharfstein and Stein (1990) and Truman (1994) suggest more complicated loss functions grounded on game theoretical models. Forecasters are assumed to differ by their ability to forecast. The chief objective of the forecaster is to influence clients' assessment

of their ability. Such objectives are common for business analysts or analysts employed by financial services firms such as investment banks or brokerages whose fees are directly linked to clients' assessment of analysts' forecasting ability.

An interesting example comes from financial analysts' earnings forecasts which are commonly found to be upward biased (e.g., Hong and Kubik (2003) and Lim (2001)). A reason given for this bias is that by reporting a rosier picture of a firm's earnings prospects, the analyst gets favored by the firm's management and gets access to more precise and timely information. Too strong a bias will compromise the precision of the analysts' forecast and will be detrimental to the position of the analysts in the regular rankings that are important to their career prospects, particularly for "buy-side" analysts. Forecasts must trade off bias against precision. In general we would not expect the cost of over- and under-predicting earnings to be identical.

4.3 Backing Out the Loss Function

In situations where it is difficult to specify *a priori* the exact form of the loss function, it is attractive to attempt to 'reverse engineer' the loss function. The idea is to approximate the unknown loss using a flexible family of loss functions such as (11). While flexibility is important, parsimony is an equally important concern since it is rare to encounter cases where the number of forecasts amounts to more than a few hundred observations, at least if attention is restricted to the behavior of individual forecasters. Often the situation is even more limited than this. For data sources such as the Survey of Professional Forecasters or the Livingston surveys, individual forecasters with more than a few dozen predictions are fairly uncommon.

Within the context of a given family of loss functions, the unknown parameters of the loss function can be estimated from the forecaster's first order condition (3). For example, for a given value of p, α is the single parameter that controls the degree of asymmetry among the loss functions in (11). When $p = 2$, $\alpha/(1-\alpha)$ measures the relative cost of positive and negative forecast errors of the same magnitude. For example, a value of $\alpha = 0.4$ suggests that positive errors are two-thirds as costly as negative errors of the same magnitude.

The estimate of α thus provides economic information about the degree of asymmetry required to justify the observed sequence of forecasts. Sometimes such estimates can be rejected on economic grounds. Suppose, for example, that an estimate $\alpha = 0.1$ is required to justify rationality of the observed forecasts. This suggests that it is almost ten times costlier to underpredict than to overpredict the variable of interest. This may be deemed implausible on economic grounds and so asymmetric loss is unlikely to be the explanation for the observed behavior of the forecast error.

How the forecaster maps predictions into actions may also be helpful in explaining properties of the observed forecasts. Leitch and Tanner (1991) studied forecasts of T-bill futures contracts and found that professional forecasters reported predictions with higher MSE-values than those from simple time-series models. This is puzzling since the time-series models presumably incorporate far less information than the professional forecasts. When measured by their ability to correctly forecast the direction of future interest rate movements−a metric related to the forecasters' ability to make money−the professional forecasts did better than the time-series models. A natural conclusion to draw from this is that the professional forecasters' objectives are poorly approximated by the MSE loss function and are closer to a directional or 'sign' loss function. This would make sense if the

investor's decision rule is to go long if an asset's payoff is predicted to be positive and otherwise go short.

5 Estimation of Forecasting Models

Constructing a forecast for a particular problem requires (i) choosing the variables in z that we intend to employ as inputs in the forecasting model; (ii) taking a stand on the model or set of models for the conditional distribution $p_Y(y|z)$; (iii) specifying how the forecasting model tracks the predicted variable through time, accounting for possible instabilities. The Bayesian approach further requires eliciting priors on the models and their parameters.

While economic theory is useful at suggesting candidate variables for inclusion in z, rarely is theory so precise as to pinpoint directly which variables should be included or how they should be measured. Economic theory is often better at excluding variables from consideration and limiting the search over which variables to include in the forecasting model.

Moreover, economic theory rarely specifies the distributional form or model relating y to z. In most forecasting situations there is therefore considerable uncertainty over the functional form of the model. Finally, the stability through time of the functional form may be questionable. Economies continually evolve, regulations and technology change, suggesting a need to allow the functional form or the parameters of the forecasting model to change over time.

Each of these issues highlights a theme in recent work on economic forecasting which we review below. First, however, we review the workhorse in the forecasting literature, namely the linear forecasting model and its extension to vector autoregressions (VARs). Challenges faced in using VARs for forecasting foreshadow issues that arise for the general problem.

5.1 The Linear Forecasting Model

Vector autoregressions (VARs), originally proposed by Sims (1980), constitute a prominent class of forecasting models. Prior to their introduction, larger structural models were the norm in macroeconomic forecasting. While these forecasting models are still employed by central banks and other institutions, the theoretical points of Sims (1980) along with the empirical success of VARs (Litterman (1986)) has led many forecasters to employ this method.

Optimal forecasts are synonymous with linear regression models under MSE loss and a linear specification for the conditional mean. VARs are multivariate extensions of the autoregressive model so commonly used in forecasting and take the form

$$z_t = \beta_0 + \sum_{j=1}^{k} \beta_j' z_{t-j} + u_t \tag{13}$$

(so $\theta = (\beta_0', \beta_1', ..., \beta_k')'$). In as far as possible, the autoregressive order, k, is chosen so that u_t is serially uncorrelated. When $z_t = (y_t, x_t')'$ the first equation of the system is the forecasting equation.

Least squares is the standard plug-in method for estimating the conditional mean from a linear model. OLS estimates are consistent and asymptotically efficient when the number of regressors is

fixed or grows slowly enough as the sample size increases. There are limitations to using this result to justify the use of OLS, however. First, because we can write MSE loss as the variance of the forecast error plus the squared bias, the additional loss from using a biased estimator could well be offset by reductions in the variance term. Because the focus of providing good forecasts is not on the individual estimates for each parameter, other estimation approaches could lead to better forecasting performance. Second, we might not want to rely on asymptotic optimality properties for the OLS estimates of the parameters. In practice the small sample distributions may differ greatly from their asymptotic counterparts, making comparisons of estimators based on these asymptotic distributions misleading. Taken together, these points have led to a number of other estimation approaches.

Bayesian methods have been suggested as alternatives to OLS in the construction of forecasts from VAR models. Under the additional assumption that u_t is normally distributed, the likelihood is fully specified and has a well known form. Combined with a set of priors, one can then construct the posteriors and the desired forecasts. For example, the normal prior results in a closed form solution for the posterior distribution for $\beta = \{\beta_0, \beta_1, ..., \beta_k)$. In the linear regression model with independently and identically distributed residuals $\varepsilon_t \sim N(0, \sigma^2 I)$ with σ^2 known and prior $\beta \sim N(\beta^0, \Omega)$, the posterior distribution for the regression parameters is normal with mean $(\sigma^{-2}X'X + \Omega^{-1})^{-1}(\sigma^{-2}X'y + \Omega^{-1}\beta^0)$ and variance $(\sigma^{-2}X'X + \Omega^{-1})^{-1}$. The plug-in forecast simply uses the mean of the posterior, which takes the form of a shrinkage estimator. Setting $\Omega = \sigma^2 k^{-1} I$ we have $\tilde{\beta} = (X'X + kI)^{-1}X'y$ which is in the form of a Ridge estimator. Under MSE loss, normally distributed $\hat{\beta}$ and any general prior distribution $\pi(\beta)$, it can be shown that the Bayes rule forecast takes the form of a correction to the OLS estimator.

Litterman (1980, 1986) and Doan, Litterman and Sims (1984) suggested "Minnesota" priors on the parameters of a VAR which more heavily weight the parameter configuration towards a model where variables follow individual random walks. More distant lags are shrunk towards zero more heavily. This type of prior can be helpful in obtaining better forecasts of macroeconomic outcomes. Robertson and Tallman (1999) examine the forecasting performance of flat prior VARs (i.e. the usual OLS estimates) versus Bayesian methods based on more informative priors. They find that extensions to the Litterman priors revolving around long run properties provides forecasting gains for a number of macroeconomic variables (GDP growth, unemployment, Fed funds rate and CPI inflation). Kadiyala and Karlsson (1993,1997) examine more extensive priors than the Litterman approach (allowing for dependence between the equations) and find examples of improvements over the Minnesota prior.

Unconstrained VARs are not grounded in economic theory other than in so far that theory has been used to select the underlying variables. A promising recent literature takes the approach of using dynamic stochastic general equilibrium (DSGE) models to constrain VARs. In this vein, Del Negro, Schorfheide, Smets and Wouters (2006) cast SDGE models as (reduced-form) VARs that include an error correction term. Theoretical restrictions from firms' and households' optimizing behavior, along with assumptions about government expenditures and agents' intertemporal budget constraints, imply a set of cross-equation restrictions on the parameters of the VAR. Deviations of the VAR parameters from these cross-equation restrictions can be viewed as indications of model misspecification. Using a Bayesian approach, Del Negro et al (2006) relate these constraints to the priors on the model parameters; Ignoring the theory corresponds to diffuse priors while informative

priors pull the parameters towards the theoretical constraints. In a simulation study these authors find evidence that, for a range of macroeconomic variables, the DSGE-based VAR produces better out-of-sample forecasting performance than the standard unconstrained VAR.

5.2 Estimation with Many Variables

The inability of economic theory to precisely define which variables should be included in a forecasting model has become an important issue since thousands of variables are readily available from governments and other organizations. The virtual explosion in the number of potential predictor variables is exacerbated by the fact that the dynamic structure of the forecasting model is typically unknown. The addition of an extra economic variable therefore increases the model dimension not only by a single parameter but by many parameters to account for the dynamic effects of this variable on the outcome variable. At the same time, the length of the available time series is often relatively short because the frequency of time series observations and the period over which data have been constructed combine to limit the sample size. Model instability (which will be discussed below) may further limit the useful length of the data.

Short samples, large sets of predictor variables combined with the understanding that better estimates than least squares are available when prior information can be exploited, suggest that least squares might not be the most appropriate method for estimating forecast models. Indeed, many other methods for estimating forecasting models might be useful depending on the exact circumstances of the model. This appears to be true in practice and has led to the proposal of a wide range of estimation techniques in the forecasting literature which we next turn to.

5.2.1 Shrinkage Methods

A variety of estimation and variable selection methods have been suggested to attain a better trade-off between the bias and variance components of the forecast MSE. Most of the alternative plug-in estimators are modifications of the OLS estimator, $\hat{\beta}_i^{OLS}$, and fall in the general set of shrinkage estimators, $\hat{\beta}_i^s$, of the form

$$\hat{\beta}_i^s = \gamma_i \hat{\beta}_i^{OLS} + \tilde{\beta}_i \qquad 0 < \gamma_i < 1, \quad i = 1, ..., k. \tag{14}$$

where $\tilde{\beta}_i$ is the shrinkage target. γ_i generally depends on the data (e.g. James and Stein (1961)). It is common practice to set $\tilde{\beta}_i = 0$, in which case the OLS estimators are shrunk towards zero.

Estimators differ in how they specify the shrinkage weights γ_i and shrinkage target $\tilde{\beta}_i$. These include bagging and subset selection methods, as well as Stein regression and many Bayes and empirical Bayes methods. Other examples of shrinkage methods that have been less employed for forecasting include ridge regression, the lasso (Tibshirani (1996)) and the Garrote (Breiman (1995)).

A natural approach is to only use a subset of the available predictor variables for forecasting. Inclusion of additional variables in the regression model increases the variance of the forecast, although if their true coefficients are nonzero then this also reduces the bias. Removing variables with coefficients that are small enough (so the resulting bias is small) to avoid this additional increase in variance may yield a better performing forecasting method. Of course if the coefficient

on the omitted variable were truly zero, then there is no bias and the variable should always be excluded.

Subset regression methods set $\tilde{\beta}_i = 0$ while γ_i becomes an indicator function based on the adopted rule for variable inclusion or exclusion. Various methods for choosing the regressor subset are in use. When there is a natural ordering to the regressors, such as in vector autoregressions, penalized likelihood methods such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) have been employed. On average the BIC method includes a smaller subset of regressors than AIC. Such methods are less common without a natural ordering because the set of models to search over easily becomes very large without this ordering. For example, with 30 variables, there are 2^{30} or more than one billion different models that have to be considered.

An alternative approach is to evaluate each variable on its own or in smaller groups — a method advocated by, e.g., Hendry and Krolzig (2004). This step-wise procedure employs t-tests to remove individually variables with statistically insignificant coefficient estimates. While computationally highly attractive, this approach gives rise to problems of its own: Insignificant estimates can arise not only because the true parameters are small but also because of large sampling error. Moreover, if one re-estimates the model after removal of some parameters and again examines statistical significance, the method becomes path dependent. This matters because classical pre-tests need not result in consistent estimates of the model. Finally, the 'all or nothing' approach of either using the OLS estimate or omitting the regressor may be too restrictive (i.e. restricting γ_i equal to zero or one).

Methods that attempt to exploit the bias-variance trade-off without the 'all or nothing' approach have thus been suggested. The estimation method closest to pre-testing is bagging (Bootstrap averaging) proposed by Breiman (1996). In this method the forecast model is bootstrapped, re-estimated using tests for significance to omit variables, and the forecast from the bootstrapped model is computed. A final forecast is then computed as the average of the forecast over the bootstrapped models (see Inoue and Kilian (2005) for a more complete description of the method). Since variables are unlikely to be omitted for all bootstrap replications, the estimator can be viewed as a smoothed version of the 'all or nothing' approach. Bagging still sets $\tilde{\beta}_i = 0$ but γ_i is now equal to the average over this indicator function across the bootstrap replications. In this way bagging changes from a hard threshold (one or zero) to a soft one (a number between these values).

5.2.2 Dynamic Factor Models

An alternative to excluding variables from a large dimensional set of predictors is to extract a set of common features from the data set and then use these as the basis for the forecasting model. Indeed, the dominant classical approach for dealing with large dimensional data is to extract a set of common factors of much lower dimensionality than the original variables to summarize an otherwise overwhelming amount of information. Suppose X contains N economic variables whose common dynamics can be represented through the factors

$$Z_t = \Lambda(L) f_t + e_t, \tag{15}$$

where e_t is a vector of idiosyncratic shocks, f_t is a vector of common dynamic factors and $\Lambda(L)$ is a matrix of lag polynomials representing dynamic effects. For low dimensional systems (small

N), dynamic factor models can be estimated through the Kalman filter. When N is large, Stock and Watson (2002) propose a principal components approach to obtain the common factors as the solution to a simple least squares problem. An alternative approach proposed by Forni et al (2000, 2003) is to extract principal components from the frequency domain using spectral methods.

While the construction of a set of common factors resolves the question as to how to aggregate an otherwise far too large dimensional state vector, use of these techniques in forecasting also raises many new issues. First, there is the risk that the factor extraction serves as a 'black box' approach void of any economic interpretation. This risk arises in situations where the factors are not clearly identifiable with underlying blocks of economic variables. In practice, however, often the first few factors can be interpreted in a way that links them to a particular subset of variables.

Effectively different forecasting methods amount to different weighting functions on the underlying regressors (Stock and Watson (2005)). Methods that put full weight on all regressors tend to suffer from imprecisely estimated parameters. At the opposite extreme are purely autoregressive methods that ignore information in other variables and forecast based solely on a variable's own history. In the middle of these extremes lie methods that include the first few (and most significant) principal components in addition to own-variable autoregressive dynamics. In empirical analysis, Stock and Watson find that the latter methods generally work best and that a few principal components do most of the job for most macroeconomic variables.

5.2.3 Non-Standard Loss Functions

For loss functions other than MSE, a variety of methods have been considered to estimate the parameters of the forecasting model. In the context of forecast combination, Elliott and Timmermann (2004) propose methods of moments estimators for a variety of loss functions including lin-lin and asymmetric quadratic loss. Estimation methods under asymmetric quadratic loss are further examined by Weiss (1996).

The message that has emerged from this literature is that although it is not uncommon practice to use one loss function to estimate the parameters of the forecasting model and another loss function to evaluate the forecasts, forecasting performance can often be significantly improved by using the same loss function in the estimation and evaluation stages. This has been found both in simulation experiments (Elliott and Timmermann (2004)) and in empirical studies (Christoffersen and Jacobs (2004)).

5.3 Nonlinear Models

In general, there is little reason to expect that the economy yields linear relationships between the data and the predicted variable. Empirical tests for various forms of nonlinearity tend to reject (Terasvirta (2006)). For example, important nonlinearities have been identified in the behavior of asset returns, particularly in the large literature on volatility modeling and forecasting. Hence it makes sense to employ nonlinear models as a way to improve the performance of linear forecasts. Typically the literature on nonlinear forecasting assumes MSE loss so the focus remains on estimating or approximating the conditional mean of the predicted variable. Towards this end most

nonlinear models take the form

$$y_{t+1} = \theta_1' z_t + g(z_t, \theta_2) + \varepsilon_{t+1}. \tag{16}$$

Assuming that $E[\varepsilon_{t+1}|z_t] = 0$, this nests the linear model when $g(.) = 0$.

Extending the set of forecasting models under consideration to nonlinear specifications substantially expands the model set. Define a set of models $\mathcal{M}$ as the potential combinations of parameters and functional forms under consideration. When the true model $M_0 \in \mathcal{M}$, the model is said to be correctly specified, if not it is misspecified. A misspecified model may yield forecasts that are difficult to beat, even if the coefficients are not meaningful (Clements and Hendry (1998)). For example a linear forecasting model estimated by OLS results in the linear model that minimizes Kullback Leibler distance between the estimated 'approximate' model and the unknown nonlinear model. That said, however, when forecasters have to 'sell' their forecasts to decision makers, the misspecified model may be difficult to put a story to.

The forecaster's problem is to choose the best available model $M \in \mathcal{M}$. Because this involves a search over functional forms, to be able to actually perform this search the forecaster needs to restrict the problem further. As with the specification of the variables to be included in the regression model, economic theory is often not particularly precise on the exact form of the nonlinearity to be expected.

The literature has pursued two broad themes: (i) the formalization of nonlinear models that generalize the linear model in an intuitive way; or (ii) the use of approximation procedures that seek to approximate the unknown nonlinear function. We next examine each in turn.

5.4 Local Approximations

The first literature considers nonlinearities that are essentially ad hoc models designed to be more flexible than linear models which arise as special cases. The set $\mathcal{M}$ is typically relatively small, nesting a linear specification, and fully defined up to a set of unknown parameters. Much of this work has been to extend autoregressive models and has its roots in the historical domination of autoregressive integrated moving average (ARIMA) forecasting models (Box and Jenkins (1970)) of the form

$$\phi(L)\Delta^p y_t = \eta(L)\varepsilon_t. \tag{17}$$

Here $\phi(L)$ and $\eta(L)$ are lag polynomials while $\Delta y_t = y_t - y_{t-1}$ takes first differences and ε_t is serially uncorrelated. Examples include the family of smooth transition autoregressive (STAR) models, which typically set $z_t = (y_{t,}, y_{t-1}, ..., y_{t-p})$ and differ in the exact form for $g(.)$, e.g. the exponential STAR model sets elements of $g(.)$ equal to $\theta_{21}' z_t(1 - \exp(-\theta_{22}' z_t^2))$ whereas the logistic STAR (LSTAR) model sets $g(.)$ equal to $\theta_{21}' z_t(1 + \exp(-\theta_{22}' z_t))^{-1}$, commonly with restrictions on $\theta_2 = (\theta_{21}', \theta_{22}')'$. Similar nonlinear models can be used with specifications for z_t other than lagged dependent terms in situations where a candidate variable explaining the nature of the nonlinearity (e.g. financial crises as measured by default premia) is available.

Threshold and switching regressions, which set $g(.)$ equal to $\theta_{21}' z_t I(s_t < \theta_{22})$ for some state s_t, where $I(.)$ is the indicator function are also included in this family. When the state is unobserved, it is common to model it through a regime switching process (Hamilton (1989)). The literature

predominantly uses fully parameterized Gaussian mixture model with mixing weights determined from the updated state probabilities.

The danger of using a misspecified forecasting model is a real problem given the lack of theoretical underpinning of the choice of $\mathcal{M}$ as well as this set being a very small subset of the possible model specifications. Furthermore, for many of the tests of the null of linearity for these models, rejection does not necessarily imply that the particular nonlinear model chosen is implied. Hence it would be of prime concern to understand the estimation and forecasting properties of these models under misspecification. Such misspecification and the approximation properties of these models may explain why these models sometimes generate extreme forecasts.

Smooth threshold models have been employed with some success empirically to forecast real exchange rates (Taylor and Sarno (2002)), industrial production and a host of other macroeconomic series (Terasvirta, van Dijk and Medeiros (2005)). Garcia and Perron (1996) and Ang and Bekaert (2002) use regime-switching models to capture the dynamics in US interest rates, while Perez-Quiros and Timmermann (2000) use these models to predict stock returns. Some papers find evidence that letting the state transition probabilities depend on forward-looking variables such as the leading indicator helps improve forecasting performance.

The large literature on autoregressive conditional heteroskedasticity (ARCH) in asset returns (reviewed in the context of forecasting by Andersen et al (2006)) is another example of nonlinear dynamics that could be important to portfolio managers who are concerned with predicting returns and managing the risk of their assets. These models imply that the conditional variance of asset returns is persistent and hence partially predictable, particularly at short horizons.

Tests for the functional form $g(.)$ are complicated because θ_2 or a subset of this vector is not identified under the null hypothesis of linearity. As a result, standard methods such as the generalized likelihood ratio test lose optimality properties and are no longer approximated asymptotically by a chi-squared distribution. In the application to switching models, standard practice seems to be not to conduct much testing to identify the number of regimes, and many papers simply assume the presence of two states. A large literature has arisen to deal with these issues, however, see Terasvirta (2006).

In some cases nonlinear forecasting models can adapt more quickly to changes in the underlying time-series dynamics and avoid smoothing the data as much as linear models. However, this same feature means that nonlinear models can be highly sensitive to the sort of outliers found in many economic and financial time series. Furthermore, the parameters capturing nonlinear dynamics are often associated with a few episodes such as the change in the dynamics of US interest rates during the 'monetarist experiment' from 1979-1982. As a consequence, these parameters can be very imprecisely estimated for the typical sample sizes available to macroeconomic forecasters and so these models often produce quite poor out-of-sample MSE performance.[3] On the other hand, nonlinear forecasting models may perform quite well in certain states (e.g. recessions or periods of financial crises) and so can be used either in conjunction with other models that generate more smooth forecast (see the section on combination) or for non-convex loss functions such as the sign function that put smaller weight on outliers.

[3]Indeed, some simulation studies find that even when the nonlinear model is correctly specified, it often produces less precise forecasts than a simple misspecified linear approximation due to the greater uncertainty about the nonlinear model's parameters (Psaradakis and Spagnolo (2005)).

5.5 Flexible Approximations

The second approach to forecasting with nonlinear models has been to acknowledge the uncertainty over the form of $g(.)$ and attempt to construct a flexible approximation by considering a very large set of models to include in $\mathcal{M}$. To make the search over models operational, rather than search over all of $\mathcal{M}$ this approach instead searches over an approximating set $\tilde{\mathcal{M}}$ of the form

$$y_{t+1} = \theta_1' z_t + \sum_{j=2}^{J} g_j(z_t, \theta_j) + \varepsilon_{t+1}. \tag{18}$$

Often the simplification $g_j(z_t, \theta_j) = \tilde{g}_j(z_t)\theta_j$ is employed to make the forecasting model linear in the parameters (or at least more so, since additional parameters can be hidden inside $\tilde{g}(.)$). The idea is to choose the functions $g(.)$ carefully enough and the number of them, J, large enough to approximate a wide variety of possible nonlinear functions, see e.g. Swanson and White (1995).

There are a large number of theoretically well motivated choices for the basis functions $g_j(z_t, \theta_j)$. Most popular in the economic forecasting literature are artificial neural network models, where $g_j(z_t, \theta_j) = \theta_{j1}(1 + \exp(-z_t'\theta_{j2}))^{-1}$ and various methods are employed for choosing or estimating θ_{j2}. Other basis functions include Fourier series, polynomials, piecewise polynomials and splines. Methods such as wavelets, ridgelets, and the Gallant (1981) flexible fourier form also fit this set. Both theoretically and in practice, different methods work well against different classes of functions.

Practical problems arise for these methods both in terms of estimation and forecasting. First, unless $\tilde{g}(.)$ is fully specified, estimation requires nonlinear optimization. Variations have arisen in attempts to find simple methods to specify the functional form, so as to leave the remaining estimation linear in the parameters and hence estimable by OLS. This is standard for example in the application of neural net models.

Second, the order J must be chosen. Since an infinite number of possible terms could be included and the fit is improved by choosing J as large as possible, overfitting is likely unless the number of included terms is somehow restricted. Overfitted models tend to produce very good results for the data used to train and estimate the models, but very poor forecasts on fresh (out-of-sample) data. To deal with these issues, methods such as information criteria and cross validation are used to select among this class of models. Overfitting remains the Achilles heel of these methods, however.

Finally, as with parametric nonlinear models, the risk of generating relatively extreme forecasts remains when forecasting from sample points where the data is reasonably sparse. Many practitioners use 'insanity' filters, replacing these forecasts with a smoothed value when the forecast is too far from the outcome or its mean. Even so, the track record of these models in forecasting has been mixed. Nonlinearities do seem to be present in many macroeconomic series, but the data samples for these variables tend to be relatively short, thus hampering the precise estimation of nonlinear forecasting models.

For financial returns the signal-to-noise ratio−i.e. the fraction of predictable variation in asset returns−tends to be very low. Often this means that the observed nonlinearities are poorly identified, imprecisely estimated and so the risk of overfitting is very high. As a result, the deterioration in out-of-sample forecasting performance is likely to be very high when compared against the predictive performance during the training sample used to fit the model (see, e.g. Racine

(2001)). Consequently there is little evidence that such forecasting models dominate simpler linear specifications, at least under MSE loss.

Overall, the difficulties that arise when forecasting with nonlinear models revolve around the question whether a fitted nonlinear model provides a good approximation to the true nonlinear model. In the case of the ad-hoc specifications, rejection of a linear model in favor of a particular nonlinear specification does not necessarily indicate that the latter will produce good forecasts. Rejections of typical tests for nonlinearity tend to indicate a range of possible models rather than a particular model. In the case of model approximation, model specification search tends to result in models that overfit the data, again causing problems for forecasting.

Economic forecasts are often only one piece of information used in conjunction with a decision maker's prior beliefs and other sources of information. Moreover, such forecasts are often used as a way to assign different weights on various possible scenarios. Purely statistical approaches based on complicated 'black box' approaches have with few exceptions so far failed to generate much attention among economists and are not used to the extent one might otherwise have expected.

5.6 Multi-Period Forecasts

In many forecasting situations the forecast horizon, h, is longer than the frequency used in collecting data and estimating the forecasting model. For example, a central bank may be interested not only in forecasting short-run inflation but also in inflation over the medium and long run. In these situations the forecaster encounters a multi-period forecasting problem.

Computing multi-period forecasts is conceptually trivial, but can become complicated in practice. To illustrate this point, when forecasting from a univariate first-order autoregressive model $y_t = \phi y_{t-1} + \varepsilon_t$, forward iteration gives

$$y_{t+h} = \phi^h y_t + \sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i}. \tag{19}$$

Assuming that $E[\varepsilon_{t+j}|y_t] = 0$ for $j > 1$ and that both the model and its parameters are known, the optimal forecast under MSE loss is simply $\phi^h y_t$.

When the parameters are unknown, however, the problem becomes far more complicated. A simple solution would be to plug in the OLS estimate of the parameter, $\hat{\phi}^h y_t$. However, this is clearly only a solution of convenience: $\hat{\phi}$ is generally not unbiased, and thus, $\hat{\phi}^h$ will not be unbiased for ϕ^h either. Even if $\hat{\phi}$ were unbiased, in general $\hat{\phi}^h$ would not inherit this property.

An obvious alternative to iterating forward on a single-period model is to tailor the forecasting model directly to the forecast horizon. This is more in spirit with viewing forecasting models as misspecified simplifications of the underlying data generating process and entails a model of the form

$$y_{t+h} = g(z_t, \theta) + \varepsilon_{t+h}. \tag{20}$$

The chief problem is now the overlap in the forecast errors that will generally exhibit behavior similar to that of a moving average process of order $h-1$. However, this is easily handled through a number of procedures that account for autocorrelation in the forecast errors.

Which approach is best−the direct or the iterated−is an empirical matter since it involves trading off estimation efficiency against robustness to model misspecification. It is also not clear

how iterating multiple periods ahead on a misspecified model will affect the quality of the forecast. For example, the initial value of the conditioning information (z_t) could well matter in this situation. Even when the models are correctly specified, there is a trade-off between the cumulative effect on the forecast of using plug-in parameter estimates (which is avoided in the direct approach), versus the greater efficiency of the iterated approach that comes from estimating the forecasting model on data measured at a higher frequency.

Marcellino, Stock and Watson (2006) address these points empirically using a data set of 170 US monthly macroeconomic time series. They find that the iterated approach generates the lowest MSE-values, particularly if long lags of the variables are included in the forecasting models and if the forecast horizon is long. This suggests that reducing parameter estimation error can be more important than concerns related to model misspecification.

When several forecasts at multiple horizons are simultaneously available (as in the case with many survey forecasts), this offers significant advantages in terms of constructing tests for forecast efficiency that do not depend on knowing which information was available to the forecaster. Assuming that the forecaster makes efficient use of all historical information, under MSE loss and a stationary data generating process we have that $MSE_{h_L} > MSE_{h_S}$, where $h_L > h_S$ are long and short forecast horizons respectively.[4] To see this, suppose the bias is zero at all horizons and that the variance of the two-period forecast is smaller than that of the optimal one-period forecast. In this case the variance of the previous period's two-period forecast must be smaller than the variance of the current one-period forecast, contradicting the assumption that the current one-period forecast was optimal in the first place. Hence, under appropriate stationarity assumptions, expected loss must be non-decreasing in the length of the forecast horizon.

Special problems may arise when forecasting multiple steps ahead with nonlinear models, where numerical methods are typically required due to the nonlinear form. This problem stems directly from taking the ad hoc model to be the true model, which is a little perverse. To illustrate this, suppose that

$$y_{t+1} = g(y_t; \theta) + \varepsilon_{t+1}. \tag{21}$$

Iterating forward to the two-period horizon, we have

$$y_{t+2} = g(g(y_t; \theta) + \varepsilon_{t+1}) + \varepsilon_{t+2}. \tag{22}$$

Hence the function g (presumed to be nonlinear) needs to be invoked as many times as the length of the forecast horizon. Moreover, the entire distribution of ε becomes crucial even under MSE loss where interest is limited to forecasting the conditional mean. For example, if g is quadratic, the variance of ε matters to forecasting the conditional mean two or more periods ahead.

6 Model Instability

Economic institutions, tax rules and political regimes change over time and the economy evolves in response to technological and macroeconomic shocks such as the oil price changes in the seventies. Events such as these make it plausible that the underlying data generating process changes over

[4] For other loss functions, Patton and Timmermann (2005) prove that the expected loss at the longer horizons must be greater than or equal to the expected loss at short horizons.

time. Typically economic theory is silent on the exact form of these instabilities which is more up to empirical research to determine.

In the construction of the forecasting problem in Section 3, the specification of the likelihood for the data $p_Y(y|z)$ did not require that the relationship between the data remains stable over time, or that the underlying data itself is stable through time, although a model for the process that generates instability is required. A difficulty that arises in the presence of changes in the data generating process is the existence of a multitude of models that can capture potential instabilities. Unit root models are popular, although the root could be near one rather than exactly equal to one. Fractionally integrated models allow similar behavior at low frequencies. Breaks in regression parameters can also mimic this type of behavior. Beyond this we could allow the root to be stochastic and near one.

Model instability introduces at least three problems for forecasters. First, it complicates specification of the likelihood for the data. From a Bayesian perspective this can make it more difficult (at least analytically) to determine a closed form forecasting rule, depending on the form of the nonstationarity. Second, since the parameterization of the nonstationarity results in a larger dimension of θ, estimation is also affected. Finally, nonstationary data makes averaging over the past to obtain plug-in estimates more difficult. In classical estimation this can be a large problem. Further complications arise through nonstandard properties of the estimators that frequently arise in these models.

The two most common types of instability found in macroeconomic and financial data are breaks in the model parameters and unit root or long memory behavior of the data. We next discuss each of these.

6.1 Breaks in Coefficients

The existence of nonconstant parameters in many of the data series that economists have an interest in forecasting is well documented. Stock and Watson (1996) considered the stability of a large set of macroeconomic and financial variables and found evidence of model instability for the majority of these. One of the stylized facts of empirical macroeconomics is the 'Great Moderation', i.e. the lower volatility of many macroeconomic series after the mid-eighties.

Breaks are likely to be empirically relevant for central bankers trying to forecast inflation. In fact, inflation appears to be among the least stable macroeconomic variables because it depends on monetary policy regimes, macroeconomic shocks and other factors. Stock and Watson (1999) report evidence of instability in the parameters of the Phillips curve.

Most work on forecasting models with unstable parameters has considered linear specifications of the form

$$Y_{t+1} = \left(\beta_t - \bar{\beta}\right) X_t + \gamma W_t + u_{t+1}, \tag{23}$$

where the coefficients β_t on X_t are changing over time while the remaining coefficients are constant.

The first problem that arises in the construction of a forecasting model of this type is that there are many ways in which β_t can be nonconstant. We could parameterize β_t as a stochastic process (either mean reverting or not) or as a step function that changes at random times by random amounts. Examples of such models include the popular unobserved components model

(where $Y_{t+1} = \beta_t + u_{t+1}$ and β_t follows a random walk process) and extensions of this to the entire vector β_t (as in the models of West and Harrison (1989)).[5]

If the variation in β_t is not permanent in the sense that it can be characterized by a mean reverting process the linear specification that omits the breaks in β_t is essentially a heteroskedastic model and least squares estimation of the parameters will not be too misleading (White (2001)).

When the breaks are permanent, the coefficients of the linear model lose meaning and become similar to sample averages of a random walk, changing with time and not related directly to any parameter of the model. In either case, knowing the true model will enable better forecasts. Forecasters must decide whether or not to (a) use only part of the data available, assuming that the retained data is sufficiently stationary that it will provide a good approximation to a model with constant coefficients, or (b) attempt to model the breaking process.

To illustrate these approaches, suppose β_t is constant apart from a single break of unknown size δ at an unknown date τ,

$$Y_{t+1} = \begin{cases} \beta X_t + \gamma W_t + u_{t+1} & t < \tau \\ (\beta + \delta) X_t + \gamma W_t + u_{t+1} & t \geq \tau \end{cases} . \tag{24}$$

An example of the first approach would be to try and estimate $\hat{\tau}$ and base the estimates of the forecasting model solely on data after the break. Alternatively, a forecaster might consider constructing estimates for both the break date $\hat{\tau}$ and the size of the break $\hat{\delta}$ in order to construct a forecast from the full data incorporating the break into the model. This would be an example of the second approach.

Unfortunately, while tests for nonconstant parameters are quite good at detecting breaking behavior of this nature, they are not capable of distinguishing the particular type of nonstationarity beyond the distinction of 'permanent' deviations versus the mean reverting deviations mentioned above. Nearly all popular tests have no power against temporary deviations of β_t from its mean. Conversely, nearly all tests have similar power against a host of possible processes for β_t when it does depart permanently from any value. This is true for models with few breaks, many breaks, or breaks every period.[6]

The implication for forecasting is that once one has found evidence of breaks in the parameters of some variables, there is still a great deal of uncertainty as to the nature of the breaking process. Because it will be difficult to pin down the appropriate model, parameterizing and estimating the breaking process will generally be quite difficult.

Even if it were known that the forecasting model has a single break point, estimates of the break date are often not particularly useful in practice.[7] Tests for a break will often reject even though the break size is too small to permit precise estimation of exactly when the break occurred. One can still proceed with the first approach and try to estimate the window of data to use for the forecast. However, some account for the uncertainty surrounding the timing of the break is likely to be an important part of a successful forecasting strategy in the presence of breaks.

[5] Markov switching models (Hamilton (1989)) can also be considered a special case.

[6] Stock and Watson (1998) show that tests for a single break have power against random walk breaks. Elliott and Mueller (2006) consider a wide class of breaking processes and show that optimal tests for each of the breaking processes have equivalent asymptotic power against all of the other breaking processes in a wide class.

[7] Bai (1997) showed that least squares estimates of the break date in the single break model are consistent provided they are larger in magnitude that those detectable by break tests.

For the alternative of estimating the full model including both the break date and the size of the break, Elliott (2006) shows that estimation of the break size and its location results in very poor forecasts relative to knowing these parameters. Instead a method of averaging over all possible break dates with weights that depend on sample estimates of the probability that each date is the true break date is suggested, with substantial gains over least squares estimates of these two parameters.

Similarly, Pesaran and Timmermann (2005, 2006) find that the estimation window matters significantly to the out-of-sample forecasting performance of simple time-series models in the presence of breaks. Given the considerable uncertainty surrounding the time and the size of the break, they consider approaches that average forecasts generated under different estimation windows. They also derive analytical results for the normal model and MSE loss which show that the gains in forecast accuracy from using pre-break data increases when breaks are small and occur late in the sample.

When there is more than one break, these issues become even more difficult. Bai and Perron (1998) suggest an approach to determine the number of breaks through repeated tests on the data. This approach has been applied to forecast stock returns by Paye and Timmermann (2006) and Rapach and Wohar (2006). Their results suggest the presence of multiple breaks in standard forecasting models for stock returns and reveal wide variation in the extent of predictability in stock returns across break segments. They also find that the break dates are difficult to pin down, vary greatly across different model specifications and do not seem to be common across international markets. This makes the task of forecasting stock returns particularly difficult since the question of 'how much historical data to use' and how to weight new versus old data is both very important in practice and difficult to come up with a satisfactory solution to.

6.2 Modeling the Break Process

The presence of historical breaks in a time-series model requires that the possibility of future breaks be considered. This means that the process generating breaks must itself be modeled. In this regard, the forecasting problem is unique compared with the problem of detecting and dating past breaks. Approaches that do not model the break process itself (such as Bai and Perron (1998)) are not directly applicable to forecasting.

This is not a problem for the time-varying parameter specifications which directly posit a model for how the parameters evolve in future periods. The most popular approach is to parameterize β_t as a random walk and use the Kalman Filter to estimate the path for β_t and produce a forecast (Harvey (2006) covers the classical approach while West and Harrison (1997) cover the Bayesian approach).

The simplest example arises when $X_t = 1$, β_t is a random walk and both the innovations to β_t and Y_t are normally distributed, so the forecast of Y_{T+1} is $\hat{\beta}_T$. Then

$$\hat{\beta}_t = \hat{\beta}_{t-1} + \phi_t(Y_t - \hat{\beta}_{t-1}),$$

where ϕ_t depends on the variances of the two error terms. For a given choice of initial value, this recursion can be used to generate forecasts in real time. In the limit ϕ_t can be approximated by a

constant, which for a given value of ϕ yields the exponentially weighted average model

$$\hat{\beta}_t = \frac{(1-\phi)}{(1-\phi^t)} \sum_{s=0}^{t-1} \phi^s Y_{t-s}. \tag{25}$$

This approach is equivalent to the discounted least squares model which puts a decreasing weight on data further back in time. These methods can readily be extended to the general model with time-varying predictor variables.

Examples of these models are plentiful in the empirical literature and have been used to track the skills of mutual fund managers (Mamaysky, Spiegel and Zhang (2006)) and to predict variables such as GDP growth, inflation and electricity demand (Harvey and Koopman (1993)).[8]

Because of the time-variation, the least squares plug-in approach (25) leads to risk that may depend on the initial parameters of the model, even asymptotically. In contrast, when the model is stationary risk is not generally dependent on the initial values when one averages over the data.

>From the Bayesian perspective, the first of these points is important while the second is not. By accounting for model uncertainty, standard Bayesian methods are directly applicable for estimating the parameters even when these are time-varying. Furthermore, since the procedure is conditional on Z the fact that risk averages across sample information is incorporated through the prior i.e. by the weighting of the relevant parameters.

As an example of the Bayesian approach, Pesaran, Pettenuzzo and Timmermann (2006) propose a hidden Markov chain approach to forecast time series subject to multiple structural breaks. They assume a hierarchical prior setting in which breaks are viewed as shifts in an underlying Bernoulli process. The parameters of each regime are realizations of draws from a stationary meta distribution. Information about this distribution gets updated recursively through time as new breaks occur. As a result this approach provides a way to forecast the frequency and size of future breaks. Their empirical findings for US interest rates suggest that accounting for breaks in out-of-sample forecasts can be important, particularly at long forecast horizons.

6.3 Unit Roots

Granger (1966) found that many macroeconomic data have a spectral peak near frequency zero, and Nelson and Plosser (1982) followed this up by showing that it was difficult to reject the presence of unit roots in many macroeconomic and financial data.

From the classical perspective, most of the literature has revolved around whether or not to impose unit roots. In the univariate model, this amounts to choosing between a model in levels or differences. In multivariate models, the problem of choosing levels or differences also raises the possibility that error correction terms can be included. When each individual series has a unit root but some of them are common, we know from the Granger representation theorem (Engle and Granger (1987)) that the full system can be represented as an error correction model.

Imposing unit roots reduces risk when the parameter is imposed sufficiently close to its true value but increases it when it is further away, with the effect dependent on the forecast horizon. Sample size is also important since estimates become more precise with more data swinging the balance in favor of less constrained estimation methods.

[8] Risk Metrics use this method to track conditional volatility in financial markets and typically sets ϕ close to one.

The most prominent alternative to OLS estimation is the pretest estimator which sets the estimate equal to one if the pretest fails to reject a unit root and otherwise selects the OLS estimate. Diebold and Kilian (2000) examine this method, which increases the gain from always imposing a unit root at the cost of doing worse on average than the OLS estimator when the coefficient is further away from a unit root.

Trade-offs are more complicated in multivariate models because of the higher dimensionality of the problem which results in a much broader set of trade-offs in the effect of parameter estimation on risk.

Finally, with the addition of different specifications of the transitory dynamics, many of the calculations that lie behind the results mentioned in the previous two paragraphs are affected, both in magnitude and in some cases in sign. Hence general results — whether analytical or through simulation — are difficult to arrive at.

The issue of unit roots or near nonstationarity also arises for the common linear forecasting model $\hat{y}_t = \beta_0 + \beta_1 x_{t-1}$ when the regressor, x_{t-1}, has a trend of unknown form. There are many examples of such models being applied. Forecasts of stock returns by highly persistent variables such as the dividend-price ratio or earnings-price ratio fit this situation, as does inflation forecasting using interest rate levels, or forecasting the change in the exchange rate with the forward premium. Whilst methods have been proposed for hypothesis testing in these models, there is not much evaluation of the effect on forecasting.

As in the unit root case, when the innovations to the regressor are correlated with the residuals of the forecasting equation, risk becomes a nonconstant function of the nuisance parameters describing the form of the persistence in the data such as the degree of persistence of x and the covariance between innovations to y and x. While most theoretical work has focussed on testing $\beta_1 = 0$, little attention has been paid to designing good forecasting procedures or examining the trade-offs between possible forecasting methods.

7 Forecast Evaluation

As noted in the introduction, one of the major differences between standard econometric problems and the forecasting problem is that the researcher receives feedback on how well their forecast actually performed. Thus, when the IMF forecasts next-year output growth or inflation, the following year they are able to see how far off their forecasts were. Evaluating forecasting procedures in light of this new information generates a dynamic process through which a number of important issues arise.

Forecast evaluation usually comprises two separate, but related, tasks, namely (i) providing summary statistics on the precision of past forecasts and (ii) testing optimality properties of the forecasts by means of a variety of diagnostics. The latter involves checking whether the conditions implied by an optimal forecast hold in a particular sample. If the loss function is known up to a finite set of unknown parameters, this is a straightforward process. From the forecaster's first order condition (3) the generalized forecast errors should themselves be unpredictable, i.e. follow a martingale difference sequence.

The nature of this orthogonality condition will depend both on the shape of the forecaster's

loss function and on the presumed data generating process underlying future values of Y used to calculate the conditional expectation $E_Y[L'|Z]$. For example, under MSE loss the optimal forecast is, as we have seen, the conditional expectation of Y given all current information, Z, and the generalized forecast error is simply proportional to the forecast error. Forecast errors should therefore have zero mean, be serially uncorrelated and be unpredictable given all current information.

7.1 Forecast Precision

A variety of performance measures can be reported. It is common practice to use holdout (out-of-sample) observations to obtain a measure of risk. The idea is to split the available sample into two pieces, a regression set of R observations and a subsequent prediction set of P observations. For each of the P observations in the hold-out set we can employ the forecast procedure as if we were actually in the position of forecasting out of sample, constructing a sample of forecasts $f(z_t)$ for $t = R+1, ..., R+P$.

Three different updating schemes are commonly used: Recursive forecasts where all data up to time t are used in the construction of each forecast and the data expands as t increases; rolling forecasts which use only the most recent fixed interval of the past data so that the data window remains the same as t increases; and fixed forecasts where only data up to R is used for the entire future.[9]

The sample analog of the risk for either procedure is simply the average loss, $P^{-1}\sum_{t=R+1}^{R+P} L(f(z_t), y_t, z_{t-1})$. Such measures are routinely computed in Monte Carlo studies and in studies using real data. Real data sets the densities of both Y and Z to their empirically observed densities, and hence is generally viewed as more interesting.

Given the arbitrariness of the scale in most loss functions, the raw number is difficult to interpret. However under MSE and MAE loss, the number can be directly interpreted. For MAE the loss function is in the scale of the units of the outcome variable, and hence a clear picture of the loss is immediate. For MSE, as with variances more generally, the square root of the outcome is reported so that it is in units of the outcome variable (root MSE or RMSE).

These measures are estimates of the expected loss and hence are surrounded by sampling variability. West (1996) derives asymptotic representations for the sampling distribution of the average loss under quite general assumptions on the data, loss function, and forecast method. He also provides asymptotic normal limiting results when forecasts are constructed recursively, using one of the aforementioned estimation windows. Under a number of technical conditions,[10] the average risk functions are consistent and asymptotically normally distributed with a covariance matrix that depends on the randomness of the out-of-sample observations and has additional terms reflecting the variation that arises through the forecasts' dependence on estimated parameters.

[9] Fixed forecasts enable the theoretical simplification that the parameter estimates are based solely on data outside the period over which the data are averaged, though it would seem unlikely we would use this method in practice. The other two methods have the additional complication of the parameters being functions of the data in the forecasting period.

[10] The loss functions must be twice differentiable, estimators for $\hat{\theta}$ must be asymptotically linear, as well as mixing and moment assumptions on various functions of the data.

Only in special cases is it appropriate to use the standard variance covariance matrix that ignores randomness in the estimated parameters of the forecasting model. This happens when the same loss function is employed for estimating the parameters θ and evaluating the forecast. The most interesting case is the linear forecasting model used to minimize MSE loss for which standard errors can be computed as usual from the sequence of realized losses. Alternatively, if the estimation sample is large relative to the sample over which the forecasts are evaluated, then the additional variation due to estimating θ will be small and hence negligible.

Some issues limit direct application of these results, however. When the hold-out sample either remains a fixed or a negligible proportion of the full sample, the coefficients of the forecasting models converge to their pseudo-true values. If two or more of the models are asymptotically equivalent (for example if one nests another) then asymptotically we have perfect multicollinearity and the asymptotic approximation to the variance covariance matrix of the hold out sample risks is singular.

7.2 Efficiency Tests

The hold-out sample can also be employed to approximate the first order condition (3) through its sample equivalent

$$P^{-1} \sum_{t=R+1}^{R+P} L'(f(z_t), y_{t+1}, z_t) = 0. \tag{26}$$

Moreover, realizations of $L'(f(z_t), y_{t+1}, z_t)$ should also be uncorrelated with any information available at time t. Hence it is common to test the condition that

$$P^{-1} \sum_{t=R+1}^{R+P} L'(f(z_t), y_{t+1}, z_t) v_t = 0, \tag{27}$$

where v_t is any function of $\{z_s\}_{s=1}^{t}$. Such a test can be conducted by regressing $L'(f(z_t), y_{t+1}, z_t)$ on v_t and testing that the OLS coefficients are zero. A particular function of v_t that is often employed is the forecast itself, which is a function of z_t hence a possible choice for v_t.

Under MSE loss, $L'(f(z_t), y_{t+1}, z_t) = y_{t+1} - f(z_t) = e_{t+1}$, i.e. the forecast error. Hence (26) simply tests the forecast errors for mean zero, and (27) tests that forecast errors are not correlated with any information available at the time that the forecast is made. For these reasons (26) is known as an unbiasedness test and (27) is known as an orthogonality test. The most popular form of these tests is the Mincer-Zarnowitz (1969) regression

$$y_{t+1} = \beta_c + \beta f(z_t) + u_{t+1}, \tag{28}$$

where u_{t+1} is an error satisfying $E[u_{t+1}|z_t] = 0$. Unbiasedness can now be tested through the joint constraint that $\beta_c = 0$ and $\beta = 1$.

Tests such as (26) and (27) examine whether or not the information in z_t has been used efficiently in the construction of the forecast. This is an important issue because a rejection of the test would suggest that improved forecasts are possible given the available data. It is also important from the perspective of testing rationality when the forecasts $f(z_t)$ are constructed by agents that are

expected to be acting rationally, and z_t is data that would have been available to those agents when they constructed their forecasts.

To examine these tests from an econometric perspective, recall that $f(z_t)$−and possibly also the instrument v_t−is constructed using parameter estimates based on data up to time t. When evaluating the sampling distribution for the regression estimates in the unbiasedness or orthogonality tests (26) and (27), we must therefore consider the sampling variability that arises through the fact that the variables in the regression are constructed. West and McCracken (1998) provide results for this regression covering a number of methods for constructing the forecasts and v_t. Under assumptions similar to those in West (1996) they show that the coefficients in the regression tests are asymptotically normal, although the variance covariance matrix may need to be adjusted to allow for the additional variation arising from sampling variation in the underlying parameter estimates.

An additional practical concern involves the specification of v_t as a function of z_t. Often there are numerous candidate variables in z_t. This, combined with the possibility that we could use any functional form of z_t as the instrument, means that the list of candidates is practically unlimited. Nevertheless, any test of orthogonality has power only in the direction of the included instrument, v_t. For example, in forecasting inflation with v_t set to past interest rates, the test would be capable of picking up any additional explanatory power in interest rates but not for other variables. The same is true for getting the functional form correct. Avoiding the first problem — picking the wrong z_t to include — is difficult. For the second problem, Corradi and Swanson (2002) suggest a nonparametric method for estimating a general function of the included elements of z_t.

For loss functions other than the mean squared loss function, $L'(.)$ is no longer equivalent to the forecast errors. Hence it is possible that forecast errors are not mean zero and that past information may well be correlated with forecast errors even when the forecast is constructed optimally. Timmermann and Patton (2005) give examples. Indeed, it is clear from (26) and (27) that the tests rely on the use of the correct loss function. Keane and Runkle (1990, page 719) write "*If forecasters have differential costs of over- and underprediction, it could be rational for them to produce biased forecasts. If we were to find that forecasts are biased, it could still be claimed that forecasters were rational if it could be shown that they had such differential costs.*"

Rationality tests may thus reject, not because the forecaster is using information inefficiently but because the loss function has not been correctly specified. This is an important issue since the loss function is generally unknown even though it is invariably assumed to be of the MSE type. Elliott, Komunjer and Timmermann (2005) examine a class of asymmetric quadratic loss functions

$$L(e_{t+1};\alpha) \equiv [\alpha + (1-2\alpha)\mathbb{1}(e_{t+1}<0)]\,|e_{t+1}|^2, \tag{29}$$

where α $(0<\alpha<1)$ is the asymmetry parameter. This loss function reduces to MSE when $\alpha = 0.5$. Regressing forecast errors on v_t (as would be appropriate for MSE loss) results in coefficients on v_t that converge to the true coefficient plus an extra term $(1-2\alpha)E[v_t v_t']^{-1}E[v_t|e_{t+1}|]$. If v_t contains a constant term (which is usually the case) then $E[v_t|e_{t+1}|]$ is always nonzero and orthogonality tests based on MSE loss will reject with probability one asymptotically as a result of using a misspecified loss function.

In general, future values of $L'(.)$ should not themselves be predictable given any variables in the forecaster's current information set. A joint test of forecast efficiency (rationality) can thus readily be conducted within the context of a given family of loss functions which yields $L'(.)$ as a

function of a finite set of unknown parameters. If the test is rejected, either the forecaster did not use information efficiently or the family of loss functions was incorrectly specified.

In situations where the loss function is not known up to a small set of shape parameters, it is possible to use tests that trade off assumptions about the underlying data generating process against much weaker assumptions on the loss function (such as homogeneity properties). Patton and Timmermann (2006) show that when loss is only required to be a homogenous function of the forecast error, while the data generating process can have dynamics in the first- and second conditional moments (thus covering a large range of nonlinear-in mean specifications, ARCH models etc.), a quantile regression test can be used to test forecast optimality.

7.3 Survey and Real Time Forecasts

Survey data provide an ideal way to test whether economic forecasters use information efficiently. In this regard, the empirical evidence has been mixed. Brown and Maital (1981) studied average forecasts of US GNP and rejected unbiasedness and efficiency in six-month growth predictions. Zarnowitz (1985) found only weak evidence against efficiency for the average forecast of US growth, but stronger evidence against efficiency for individual forecasters. Batchelor and Dua (1991) report little evidence that forecast errors were correlated with their own past values. In contrast, Davies and Lahiri (1995) found evidence that forecast efficiency was rejected for up to half of the survey participants in their panel analysis.

Survey data on inflation expectations is another area where efficiency tests have been conducted. Figlewski and Wachtel (1981) analyze expectations of individual respondents and reject forecast rationality under squared loss. Mishkin (1981) also rejects rationality of survey forecasts of inflation. Zarnowitz (1985) finds evidence of systematic forecast errors for US inflation and rejects unbiasedness for more than half of the survey participants. Pesaran and Weale (2006) provide an extensive review of the literature on survey data in forecasting.

The real-time nature of economic forecasting is what distinguishes it most from other areas of economic analysis. This affects all stages of the forecasting process: Models must be formulated, selected and estimated in real time. Evidence of model break-down or misspecification must also be examined in real time. It is not clear, for example, what one can conclude from full-sample evidence of forecast inefficiency. Unless the inefficiency was detectable at an earlier stage of the sample, using information that was available historically, it cannot be established that the forecaster acted irrationally. For example, under MSE loss, the forecast errors should be mean-zero conditional only on the available information (including the forecasting model) at the point the forecast was formulated and not conditional on full-sample information.[11]

Real-time considerations even pertain to the data "vintage" that was available at a given point in time and could have been used to formulate and evaluate a forecasting model. Croushore (2006) and Croushore and Stark (2003) make it clear that key macroeconomic data such as GDP growth is subject to important revisions, partly due to regular updates from preliminary to secondary and later data releases, partly due to changes in the methodology used to measure a particular variable. These revisions can lead not only to changes in the estimated parameters but can also affect the dynamic lag structure or functional form of the forecasting model and hence change conclusions

[11] See further discussion in Pesaran and Timmermann (2005).

regarding predictive relationships (Amato and Swanson (2001)). Data revisions are even more important for composite series such as the index of leading indicators whose composition may change due to past failures in forecasting (Diebold and Rudebusch (1991)).

These points emphasize that it is important to use the original data vintages when simulating the real-time out-of-sample forecasting process and evaluating the precision of the resulting forecasts.

7.4 Evaluation of Density Forecasts

Forecasting is one case where 'one size fits all' does not hold. Forecast users have different loss functions and therefore require different optimal point forecasts. It may therefore be better to provide forecast densities instead of point forecasts. Many agencies now provide such information. For example, the Bank of England reports the 'river of blood' forecast that shows by various shades of red their forecast of likely inflation outcomes. Similarly, the European Forecasting Network reports density forecasts for a range of macroeconomic variables. With such a density forecast in hand, decision makers with different loss functions will be able to separately solve for their optimal decision.

Two potential problems arise in comparing density forecasts to point forecasts. First, it is more difficult to estimate the whole density than to provide a single point forecast. Second, different estimation methods will be better for certain features of the density, and the loss function has information that is useful in suggesting which features of the density are important and which are not.

For the first of these, consider that for any lin-lin loss function the best estimator estimates the quantile of interest and not the entire density. Any density estimator (see, e.g. Tay and Wallis (2000)) may well trade off precision at the required quantile against precision over the entire density. Hence sample information in the density estimator may not be as good as that of the quantile estimator. Not all estimators are created equal.

For the second case, consider a bivariate outcome. The density for a binary outcome is equivalent to an estimate of a probability of the positive outcome, which in turn is simply a parameter estimate. Hence for this special case parameter estimation and density estimation are equivalent. In the case of a misspecified parametric density, Elliott and Lieli (2006) show that estimation that takes into account the loss function can provide a better estimate of the probability of a positive outcome for those loss functions. The estimator depends on the loss function. However each case yields a different estimate of the density. This is generally true for parametric estimation. Because parametric density estimation is the estimation of the parameters of the density, and different loss functions suggest different estimation techniques for the parameters, estimating the density without paying attention to the loss function and ultimate use of the forecast density involves estimation trade-offs (either implicit or explicit) that favor some users at the expense of others.

Although density forecasts are still not commonly reported, a literature has emerged on how such forecasts should be evaluated. A basic tool used to this end is the so-called probability integral transform. This is simply the inverse of the cumulative density function, F^{-1}, implied by a particular parametric forecasting model. When applied to the actual realization of the predicted variable (y), $F^{-1}(y)$ should be drawn from a uniform distribution and be independently and identically distributed over time. This argument ignores the effect of using estimated parameters of

course, but this type of test has regained popularity following the study by Diebold, Gunther and Tay (1998). Corradi and Swanson (2006) provide a comprehensive summary of current tests in this area.

Bayesian methods provide the predictive density $f_Y(y) = \int f(y|z,\theta)\pi(\theta|z)d\theta$. When the outcome is realized, it can be compared to the density the model suggests it should be a draw from. A natural statistic to compute is the p-value of this outcome, y, i.e. $P(y < Y^p)$ where Y^p is the random variable with density $f_Y(y)$. If this p-value is extreme it might bring the quality of the forecasting model into question. This evaluation method is used by, for example, Pesaran, Pettenuzzi and Timmermann (2006) to assess the quality of forecasts of interest rates from various models.

Despite issues with estimation, there is one major advantage of the provision of a density forecast, especially when the decision maker and the forecaster are different. Density forecasts convey the uncertainty in the decision making environment, in perhaps a better way than expressions such as MSE do to decision makers. Whiteman (1996) recounts his experience with providing density forecasts to Iowa state officials.

8 Comparing and Combining Forecasts

Decision makers often have access to more than one forecast. When faced with multiple forecasts, two very different strategies are possible: to seek out the best single forecasting model or to attempt to combine forecasts generated across all or a subset of models. The first approach requires being able to formally compare the forecasting performance across several models, while the latter requires a method for estimating the weights on the models used in the combination. We cover both issues in this section.

8.1 Forecast Comparisons

Since there is typically considerable uncertainty over the forecasting model, often we observe a wide array of forecasts attempting to predict the same sequence of outcomes. This has led to a literature on comparing the performance of different forecasting approaches. The idea is to compare the risk of two or more forecasts in order to choose the one that is best. For forecasting procedures $f^i(z)$, $i = 1, ..., n$, this means trying to determine which of $R(f^i(z), \theta)$ is smallest.

As with the evaluation of a single forecast, we can examine in-sample and out-of-sample performance. A hold-out sample can be used to construct an estimate of average risk, but with n different forecasting procedures this now becomes an $n \times 1$ vector of averages. These are then compared. Often the ordering of estimates of risk is examined, as well as the forecasting ability of general classes of models compared to some benchmark model.

A large number of papers aim to show that one particular forecasting model (e.g. a nonlinear specification) outperforms another benchmark model. However, it is difficult to extract any general rules from empirical studies in this literature since the best approach generally depends on the type of variable under consideration (i.e. nominal versus real data, data with a small or large persistent component), the data frequency and even the sample period.

More interesting from a general perspective are attempts to rank forecasting procedures over a wider range of data sets and see which ones perform well on average. Such an exercise is presented in

Stock and Watson (1999), who examine linear autoregressive models (with different subset selection methods such as AIC and BIC) along with commonly employed nonlinear models such as neural networks and LSTAR models across a large number of US macroeconomic data series. In practice Stock and Watson found that LSTAR models on average were outperformed by neural network models, which in turn were outperformed (except at the one month horizon) by autoregressions.

Most comparisons rank forecasts by estimated risk which is affected by sampling error. Questions such as which forecasting procedure is best is usually posed by testing the null hypothesis that all forecasting procedures have the same average loss:

$$H_0 : E[\bar{R}(f^1(z), \hat{\theta}_1)] = ... = E[\bar{R}(f^n, \hat{\theta}_n)],$$

where bars indicate the sample mean of the risk and hats indicate estimated parameters.

For pairwise model comparisons ($n = 2$) Diebold and Mariano (1995) suggest using the standard t statistic for testing equivalence in forecasting performance by taking the difference of the estimated losses and test if the resulting time-series has zero mean. For scaling they suggest a robust estimator of the variance, and suggest comparing this t-statistic to the standard normal distribution. Special cases of the West (1996) results are able to justify use of standard variance estimators for the Diebold and Mariano test.[12]

Clark and West (2004) provide an interesting illustration of how important parameter estimation can be in the comparison of a benchmark model with few or none parameters (e.g. the prevailing mean) versus a more heavily parameterized alternative model that may include time-varying predictor variables. Even when the larger model is true, because it involves estimation of more parameters and hence is more subject to parameter estimation error, we would expect this model to perform worse in finite samples than the simpler (biased) model, unless the predictive power of the extra regressor(s) is sufficiently large. Clark and West propose a test that accounts for this problem by correcting for parameter estimation error.

Giacomini and White (2006) consider the comparison of forecasting methods which comprise not only the prediction model but also choices such as estimation method and length of the estimation sample. Their analysis shifts the focus away from comparisons based on average performance towards the conditional expectation of differences in performance across forecasting methods. One advantage of their approach is that it directly accounts for the effect of parameter uncertainty by expressing the null in terms of estimated parameters and estimation windows. This means that estimation uncertainty does not vanish even asymptotically so that nested models can be compared under their approach.

More generally, horse races between competing forecasting models abound in the empirical literature. Because of the presence of strong common components in many forecasts (often representing autoregressive dynamics in the predicted variable) and short, overlapping samples, forecasts produced by different models are often sufficiently close that it is not possible to distinguish between the models with much statistical precision (see, e.g., Timmermann (2007) for a comparison of the IMF's forecasts to private sector consensus forecasts).

[12] The most prominent special case is when the expected value of the derivative of the loss function with respect to θ is zero evaluated at the true θ.

8.2 Comparing Large Sets of Models

When a large set of models needs to be compared ($n > 2$), the 'data snooping' method of White (2000) can be employed. This method compares a set of risk estimates generated by a range of individual forecasts to the risk of a benchmark model. The null hypothesis is that the best of the forecasting methods is no better than the benchmark model

$$H_0 : \min_{i=1,\ldots,n} (R(f^i(z), \hat{\theta}_i) - R(f^b(z), \hat{\theta}_b)) = 0,$$

where $R(f^b(z), \hat{\theta}_b)$ is the benchmark performance. The alternative hypothesis is that the best of the forecast methods outperforms the benchmark, i.e. that it has lower risk:

$$H_0 : \min_{i=1,\ldots,n} (R(f^i(z), \hat{\theta}_i) - R(f^b(z), \hat{\theta}_b)) < 0$$

Since the distribution of the risk differentials is asymptotically normal under the assumptions of the method (based on the results of West (1996)) this amounts to constructing a test for the maximum of a set of joint normals with unknown covariance matrix. White solves this problem by employing a bootstrap procedure to the estimates of risk. This bypasses the need to compute the unknown variance covariance matrix and directly estimates the p-value for the test.[13]

Controlling for data snooping can be important empirically. In the context of forecasting models for daily stock market returns based on technical trading rules, Sullivan, Timmermann and White (1999) find that data snooping can account for what otherwise appears to be strong evidence of return predictability.

Chong and Hendry (1986) introduced the idea of forecast encompassing, which can be applied to the concept of choosing between forecasting models. Under MSE loss the idea is similar to the orthogonality regressions (27) although the additional information v_t is no longer a subset of z_t but instead consists of forecasts or forecast errors from other forecasting methods. The idea is simple: If other forecasts have information relevant for the predicted variable that is not contained in the original forecast, then such forecasts will enter the orthogonality regression with a nonzero weight. This would mean that the original forecast did not include all relevant information. Conversely, if orthogonality holds, then the first forecast is said to encompass the other forecasts because it incorporates all the relevant information that the other forecasts have.

Romer and Romer (2000) provide an interesting comparison of the Federal Reserve Green Book inflation forecasts with private sector forecasts using encompassing regressions. Assuming MSE loss, they find evidence that the Fed inflation forecasts encompass the private sector forecasts. This conclusion is questioned by Capistran (2005) who finds evidence of significant biases of opposite sign in the Fed's forecasts during the pre- and post-Volker periods. Averaged over the full sample the bias is small, but this conceals evidence of a tendency to underpredict inflation in the pre-Volker sample followed by subsequent overpredictions.

To test if a particular forecast (null model) encompasses a set of alternative forecasts, a regression of the forecast error from the null model on the difference between the other forecast errors

[13] Hansen (2005) shows that when poor models are added to the set of candidate models, such that the benchmark is better than other models, the asymptotic normal result fails. He suggests a procedure where underperforming models are first removed in an initial step.

and that of the null model (e_t^*) can be undertaken using a $t-$ or an $F-$test (see, e.g., Clements and Hendry (1998, p. 265).)

$$e_t^* = \beta_0 + \beta_1(e_t^1 - e_t^*) + ... + \beta_n(e_t^n - e_t^*) + \varepsilon_t. \tag{30}$$

Alternative forms of this test have been suggested by Harvey et. al. (1998) and Clark and McCracken (2001) when two forecasts are being compared. To handle the problem that forecast errors depend on the estimated parameter, θ, the results of West and McCracken (1998) can be used. When the models are nested, the singularity of the joint distribution of the forecast errors is again a problem. Clark and McCracken (2001) show that in these cases the asymptotic distribution of a rescaled statistic can be approximated with a function of Brownian motions and hence the distribution is nonstandard in this case.

The emphasis in the forecast comparison literature has been on out-of-sample comparisons and the goal has been to select the best forecasting model for a given loss function. However this problem can be recast to ask 'which model is better?', a problem that has been closely examined in econometrics. For at least some model comparisons, tests conducted on the entire sample rather than on an artificially extracted hold-out sample might therefore be more appropriate and powerful (Inoue and Kilian (2004)). Under MSE loss, the problem of comparing forecast models reduces to the question of which forecast procedure is closest to the conditional expectation. This is easily tested in the full sample without problems of nesting of the models so long as the data are sufficiently stationary and not too dependent.[14]

In part the desire to test out-of-sample forecasting performance is related closely to the uncertainty about the underlying data generating process and also a concern for the effect of any pre-testing that might have occurred in constructing the forecasting models.

8.3 Forecast Combinations

Despite the many attempts to choose a single forecasting model, empirically it seems that combining forecasts from multiple models often outperforms forecasts from a single model. Clemen (1989) reviews the literature and finds that combinations outperform individual models in a wide range of forecasting problems. Makridakis and Hibon (2000) find similar results involving the forecasting of 3003 data series. For US macroeconomic series, Stock and Watson (1999) find that combining the forecasts from several methods on average performed better than simply relying on forecasts from individual models such as neural networks, LSTAR or autoregressions. Marcellino (2004) reports similar results for European data.

An argument often used to justify forecast combinations is that they diversify against model uncertainty. Forecasting models are best viewed as simple approximations to a more complicated, and constantly evolving, reality. We would therefore expect them to be misspecified in many regards−for example, they may exclude important information that is not easily modeled or they may not adjust sufficiently fast to evidence of model break-down. Some models may adapt very quickly to a change in the behavior of the predicted variable, while others adapt more slowly.

[14] Such assumptions are of course also required in the tests proposed by West (1996) and for results built on this paper.

To some extent forecast combination therefore provides insurance against 'breaks' or other non-stationarities that may occur in the future.

Because it is not known a priori how and whether the world will change in the future, a sensible strategy is to combine the forecasts from two or more approaches. A key issue is to what extent different forecasts diversify against modeling risk−which depends on the correlation in forecast errors across models−and how much weight to assign to the various forecasts.

A direct answer to the question of how to obtain a set of combination weights is provided by Bates and Granger (1969) who suggest simply regressing the predicted variable y on the individual forecasts $f_i(z)$ along with a constant

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i f_i(z) + \varepsilon. \tag{31}$$

When the individual forecasts are believed to be unbiased, it is common to omit the intercept term and restrict the slope coefficients to sum to one in which case they can be interpreted as forecast combination weights. This approach assumes MSE loss but has been generalized to other loss functions and method of moment type estimators (Elliott and Timmermann (2004)).

A comparison of the combination approach to encompassing explains why combining may be expected to be a more reasonable approach than selecting a single forecast, unless of course the true model is known to be included in the set of models under consideration and can be identified in practice. A single forecast only gets selected when the combination puts full weight on one of the forecasting methods while the rest are given zero weights. This is precisely the case where the forecast with a weight of one encompasses the other forecasts. However, this is a special case of the general concept of forecast combination, and so might be expected to be less commonly supported empirically than more evenly distributed weights.

In practice, although empirical evidence suggests that forecast combinations tend to outperform forecasts from a single model, strategies designed to obtain optimal combination weights are often outperformed by simple measures such as averaging the raw forecasts (i.e. giving all forecasts equal weights) or a trimmed set of these. If the models use roughly the same data sources and empirical techniques so differences in the performance across forecasting models are too small to be easily rejected by the data, they will tend to have similar error variances and covariances. In this situation, giving each forecast identical weights can be relatively efficient. Palm and Zellner (1982) suggest other reasons— e.g. instability of the covariance between forecast errors or estimation error in the combination weights.

Surveys are another important source of information for up-to-date forecasts of current and future values of variables that only become available at a later point in time. This is particularly useful for macroeconomic variables such as GDP which are published with a considerable delay and are subject to future revisions. Surveys provide a direct source of forwardlooking information and are thus a natural candidate to be combined with forecasts from more traditional time-series models which attempt to extract historical patterns from the data.

Which methodology is best may well depend on the state of the economy because the speed with which different forecasts incorporate shifts in the economy could vary. For example, when the economy is running at a normal pace, time-series models may provide the most accurate forecast because they make efficient use of historical information. However, these models may be slower at

capturing or predicting turning points−such as the emergence of a recession−compared with seasoned professional forecasters with access to a much larger information set. This idea is consistent with findings reported in Elliott and Timmermann (2005) who use a regime switching approach to track variations in the forecasting performance of time-series and survey forecasts of six key macroeconomic variables and form a combined forecast.

Bayesian approaches to forecast combination are becoming increasingly popular in empirical studies. Bayesian Model Averaging has been proposed by, inter alia, Leamer (1978) and Raftery et al (1997). Under this approach, the predictive density can be computed by averaging over a set of models, M_i, $i = 1, ..., n$, each characterized by parameters θ_i :

$$p_Y(y|z) = \sum_{i=1}^{n} p(M_i|z)\, p_Y(y, \theta_i|M_i, z). \tag{32}$$

Here $p(M_i|z)$ is the posterior probability of model M_i obtained from the model priors $\pi(M_i)$, the priors for the unknown parameters, $\pi(\theta_i|M_i)$, and the likelihood of the models under consideration. $p_Y(y, \theta_i|M_i, z)$ is the density of y and θ_i under the ith model, M_i, given Z. Unlike the weights used in the classical least-squares combination literature, these weights do not account for correlations between forecasts and the weights are always confined to the zero-unity interval. More details are provided in Timmermann (2006) and Geweke and Whiteman (2006).

9 Empirical Application

To illustrate the issues discussed above we consider the predictability of US inflation and stock returns. For inflation we use log first differences of the CPI while stock returns are captured by the value-weighted portfolio of US stocks traced by the Center for Research in Security Prices (CRSP). Both series are measured at the monthly frequency and the sample period is 1959:1-2003:12. To initialize our parameter estimates we use data from 1959:1 - 1969:12. We then generate out-of-sample forecasts from 1970:01 to 2003:12. Parameter estimates are either updated recursively, expanding the estimation window by one observation each month, or by means of a 10-year rolling window. Only data up to the previous month is therefore used to estimate the model parameters and generate forecasts for the current month. This is commonly referred to as a pseudo out-of-sample forecasting exercise.

We consider twelve forecasting approaches. The first is an autoregressive (AR) model

$$y_{t+1} = \beta_0 + \sum_{j=1}^{k} \beta_j y_{t+1-j} + \varepsilon_{t+1}, \tag{33}$$

where k is selected to minimize the BIC with a maximum of 18 lags and ε_{t+1} here and in subsequent models is regarded as white noise. The second model is a factor augmented AR model, using up to five common factors:

$$y_{t+1} = \beta_0 + \sum_{j=1}^{k} \alpha_j y_{t+1-j} + \sum_{j=1}^{q} \beta_j f_{j,t} + \varepsilon_{t+1}, \tag{34}$$

where $f_{j,t}$ is the jth factor and k and q are again selected to minimize the BIC (with $k \leq 18$ and $q \leq 5$). Factors are obtained using the principal components approach of Stock and Watson (2002) to a cross-section of 131 macroeconomic time series which begin in 1960. The factors are extracted in (simulated) real time using either a recursive or a rolling 10-year estimation window.

The third and fourth models are Bayesian VARs (BVARs) fitted to the variable of interest (inflation or stock returns) and the five factors:

$$z_{t+1} = \beta_0 + \sum_{j=1}^{k} \beta_j z_{t+1-j} + \varepsilon_{t+1}. \tag{35}$$

Here $z_t = (y_t, f_{1t}, \ldots, f_{5t})'$ and we include the most recent six months lags, i.e. $k = 6$. Following Litterman, own-lag terms at lag j have a prior variance of $0.04/j^2$, while off-diagonal lags have a prior variance of $0.0004/j^2$. Both a random walk prior and a white noise prior are considered. Under the random walk prior, the autoregressive parameters are shrunk towards unity, while under the white noise prior they are shrunk towards zero. Clearly the random walk prior is reasonable for the inflation example while the white noise prior is more reasonable for stock prices. We report both for each example to show the effect of the differences in prior choice.

Turning to the non-linear specifications, we consider two logistic STAR models of the form

$$y_{t+1} = \theta_1' \eta_t + d_t \theta_{21}' \eta_t + \varepsilon_{t+1} \tag{36}$$

where

$$\begin{aligned} \eta_t &= (1, y_t)' \\ d_t &= \begin{cases} 1/(1 + exp(\gamma_0 + \gamma_1 y_{t-3})) \\ 1/(1 + exp(\gamma_0 + \gamma_1 (y_t - y_{t-6}))) \end{cases} . \end{aligned}$$

We refer to these as STAR1 and STAR2, respectively.

As more flexible nonlinear alternatives, we also consider a single layer neural net model

$$y_{t+1} = \theta_0' \eta_t + \sum_{i=1}^{n} \theta_i g(\beta_i' \eta_t) + \varepsilon_{t+1} \tag{37}$$

with two hidden units ($n = 2$) as well as a two-layer neural net model

$$y_{t+1} = \theta_1' \eta_t + \sum_{i=1}^{n_2} \theta_i g\left(\sum_{j=1}^{n_1} \beta_j g(\alpha_j' \eta_t) \right) + \varepsilon_{t+1} \tag{38}$$

with two hidden units in the first layer ($n_1 = 2$) and one hidden layer in the second layer ($n_2 = 1$). For both neural net models, g is the logistic function and $\eta_t = (1, y_t, y_{t-1}, y_{t-2})$. Estimation uses search methods since α_j enters nonlinearly.

We also consider more traditional time-series forecasting methods such as exponential smoothing where the forecast f_t is generated by the recursion

$$f_{t+1} = \alpha f_t + (1 - \alpha) y_t, \tag{39}$$

subject to the initial condition that $f_1 = y_1$, and double exponential smoothing:

$$\begin{aligned} f_{t+1} &= \alpha(f_t + \lambda_{t-1}) + (1-\alpha)y_t \\ \lambda_t &= \beta(f_{t+1} - f_t) + (1-\beta)\lambda_{t-1}, \end{aligned} \tag{40}$$

where $f_1 = 0, f_2 = y_2$ and $\lambda_2 = (y_2 - y_1)$. Here α and α and β, respectively, are determined so as to minimize the sum of squared forecast errors in real time.

We finally consider a forecast combination approach that simply uses the equal-weighted average in addition to a very different approach that, at each point in time, selects the forecasting model with the best track record up to the present time and then uses this to generate a forecast for the following period.

In all cases, we apply the following 'insanity filter' which constrains outlier forecasts: If the predicted change in the underlying variable is greater than any of the historical changes up to a given point in time, the forecast is replaced with a 'no change' forecast.

Results in the form of out-of-sample, annualized root mean squared forecast errors (computed by multiplying the monthly RMSE values by the square root of 12) are presented in Table 1. First consider the results under recursive parameter estimation. For inflation, the best model is the average forecast followed by the exponential smoothing, the previous best model, the simple and factor-augmented AR models and the two-layer neural net model. Slightly worse forecasts are generated by the one-layer neural net and the BVARs, while the STAR models generate somewhat worse performances.

Overall, these results indicate that there is not much to differentiate between a cluster of the best forecasting models. This point is reinforced by the plots of predicted values from three of the models shown in Figure 1. Clearly the inflation forecasts from seemingly very different approaches are quite similar and dominated by a persistent common component.

Turning to the stock returns and focusing again on the results under recursive estimation, Table 1 shows that the best overall performance is delivered by the combined forecast and the simple and factor-augmented AR models. Once again the BVAR models perform rather poorly as do the STAR models and single layer neural nets. Overall, however, while a few approaches perform quite poorly it is difficult to distinguish between the forecasting performance among a cluster of reasonable forecasting models. In the case of stock returns which are not dominated by a strongly persistent component, there is more to differentiate between the time-series of forecasts, however, as shown in Figure 2.

Forecast precision tends to deteriorate significantly for the BVAR and double exponential smoothing forecasts under the 10-year rolling estimation window. This happens both for inflation and stock returns. This deterioration is likely due to the larger estimation error associated with using a shorter estimation window, although one should not forget that there is a trade-off in the form of faster adaptability as witnessed by the improved forecasting performance observed for the first STAR model's inflation forecasts.

Evaluation of the forecasts from these models is complicated because some of them are pair-wise nested (for example, the STAR and neural net models nest the AR models), while others are not (e.g. the factor-augmented AR models are not nested by the neural net models). As a diagnostic test, we simply compare the MSE performance of pairs of models using the Giacomini-White (2006)

approach. Our results assume a rolling estimation window corresponding to 10 years of monthly observations, i.e. 120 data points and therefore reflects the RMSE values reported in columns two and four in Table 1.

Results from these pairwise comparisons are reported in Table 2. While the BVAR and double exponential smoothing inflation forecasts are soundly rejected against those produced by the better models, for most of the other comparisons these tests do not have sufficient power to choose one model over another. The results are somewhat different in the case of the stock returns which, unlike the inflation series, do not contain a large persistent component and hence are more difficult to predict. There is little evidence to distinguish between the simple and factor-augmented AR models, the exponential smoothing, average and previous best forecasts of stock returns. Conversely, the BVARs, double exponential smoothing, STAR and neural net forecasts are generally rejected against the first group of forecasts. Parsimony thus seems to be key to successfully predict stock returns, particularly when a relatively short rolling estimation window of 120 observations is used.

Once again it is clear that although a few approaches perform very poorly and can be rejected out of hand, it is difficult to systematically differentiate between many of the other approaches.

The previous results assumed MSE loss. Theory suggests that the form of the loss function alters the optimal functional form of the model. To illustrate this, we next generated forecasts under lin-lin loss, setting $p = 1$ and α equal to 0.35, 0.50 or 0.65 in equation (11), and considering either the AR model or the factor-augmented AR model.[15] All forecasts were generated using a recursive estimation window. Results from this analysis are presented in Table 3. For inflation the average value of the lin-lin loss function under the simple AR model is generally significantly below the values produced under the factor-augmented AR specification. This holds irrespective of which quantile is being considered. In contrast, for stock returns the two models produce almost identical out-of-sample forecasting performance.

These results support many of the themes of our analysis. First, forecasts from seemingly very different approaches (e.g. linear versus nonlinear models) often produce very similar results - witness the similar RMSE performance of the neural nets and the autoregressive models. In part these similarities arise because we truncate the forecasts from the nonlinear models when these are too far away from the historical sample data.[16] In other cases nonlinear models can generate poor forecasts due to their sensitivity to outliers and their imprecisely estimated parameters. This last point is illustrated through the performance of the STAR models which generally was quite poor.

Secondly, it is difficult to outperform simple approaches such as a parsimonious autoregressive model. Simple forecasting approaches tend to generate relatively smooth and stable forecasts without being subject to too much parameter estimation error.

Third, and as an extension of the previous point, it appears that in many cases there are only marginal gains (in terms of out-of-sample RMSE performance) over and above projection on past values of the series themselves from considering the additional information that can be extracted

[15] These forecasts were generated using quantile regression, see Koenker and Bassett (1978). This already presents a nonlinear optimization problem so we only consider linear quantile specifications in our analysis.

[16] When extreme forecasts are not truncated, the RMSE for the stock return forecasts rises to 50 and 32 under the one- and two-layer neural net models, respectively. These values are three times and two times as large as the values reported in Table 1.

from large data sets. For persistent variables such as inflation, a linear autoregressive component is clearly the single most important predictive component, while for stock returns it is difficult to come up with predictor variables with significant predictive value.

Fourth, our results support the finding that forecast combination offers an attractive approach for many economic and financial variables. The average forecast produced the best or second best performance among all approaches for both inflation and stock returns. Thus, while forecast combination does not always generate the single best performance, it usually beats most alternatives unless some extremely poor models have been left in the mix of models that get combined.

Fifth, the loss function clearly matters in practice. We saw that under MSE loss, the purely autoregressive and factor-augmented autoregressive models produced essentially indistinguishable forecasting performance. In contrast, under lin-lin loss, the simple autoregressive forecasts were better for the inflation series, although they were nearly identical in the case of the stock returns.

Finally, model instability and/or sensitivity of forecasting performance to the sample period is clearly an issue. Table 1 compares the MSE performance under a recursive estimation approach which uses an expanding estimation window against that of a rolling 10-year window which can better accommodate shifts in the underlying data generating process. In many cases, the choice of estimation window makes a sizeable difference. If estimation error was the predominant effect, we would expect the ten-year rolling forecasts uniformly to be worse than the forecasts based on the expanding estimation window. This is exactly what we find for stock returns where there is no evidence that shortening the estimation window leads to improvements in any of the models. For inflation, however, we see that for half of the models the out-of-sample forecasting performance either improves or stays the same as a result of going from the expanding to the rolling estimation window. Indeed, in the case of the first STAR model, the latter approach produced substantially better forecasts.

10 Conclusion

The menu of forecasting methodologies has vastly expanded over the last few decades. No single approach is currently dominant and choice of forecasting method is often dictated by the situation at hand such as the forecast user's particular needs, data availability and expertise in experimenting with different classes of models. Although the situation is still evolving, recent research in the forecasting literature has supported some broad conclusions:

- Careful attention to the forecaster's objectives is important not only in the forecast evaluation stage but also in the estimation and model selection stages. For example, if the forecaster's loss function suggests that a particular quantile of the forecast distribution best summarizes his objectives, then quantile rather than least squares estimation should be used.

- Models of economic and financial time series are often unstable and so forecasting models are best viewed as "approximations" or tracking devices. As a consequence one should not expect that the same forecasting model will continue to dominate in different historical samples;

- Choice of the sample period used to estimate the parameters of the forecasting model is important. Using the longest possible data sample or a simple rolling window is not necessarily

the best approach if more precise information about the nature of the source of the model instability is available (e.g. institutional shifts, changes in tax policy or legislation, large technology or supply shocks). Since the nature and form of model instability may often not be very clear, more research is required to design robust forecasting approaches that deal with model instability in a variety of situations.

- Forecast combination offers an attractive alternative to the approach of seeking to identify a single best forecasting model. In part this stems from the fact that combination allows forecasters to hedge against model uncertainty and shifts in models' (relative) forecasting performance;

- Overfitting is an important concern in forecasting because of the short time series often encountered and the difficulty in getting independent data samples that can be used to cross-validate the forecasting models. This problem is exacerbated for financial time series where the signal-to-noise ratio tends to be very low. Parameter estimation error also is the likely reason why including additional economic variables in a forecasting model, which seem reasonable ex-ante, often fails to lead to the expected improvement in terms of out-of-sample forecasting performance;

- It is often difficult to distinguish with much statistical precision between the forecasts generated by seemingly very different forecasting methods. When large differences in forecasting performance occur, this often has to do with the tendency of nonlinear forecasting models to generate outliers in the forecast error distribution due to their sensitivity to the particular sample used for parameter estimation. How such outliers are dealt with then becomes important in practice;

- Guidance from economic theory is important at several stages of the forecasting process. Besides assisting in the choice of the forecaster's objective function, economic theory can be helpful in selecting categories of variables to be considered as potential predictors and imposing long-run restrictions which may reduce parameter estimation error. Econometric methods can then be used for variable selection among the predictors deemed potentially relevant from a theoretical perspective (often a large set), for specification of the short-run dynamics and for determination of the functional form of the forecasting model.

References

[1] Amato, J.D. and N.R. Swanson, 2001, The Real Time Predictive Content of Money for Output. Journal of Monetary Economics 48, 3-24.

[2] Andersen, T.G., T. Bollerslev, P.F. Christoffersen and F.X. Diebold, 2006, Volatility and Correlation Forecasting. Pages 777-877 in G. Elliott, C.W.J. Granger and A. Timmermann (eds.) Handbook of Economic Forecasting. Amsterdam: North Holland.

[3] Ang, A. and G. Bekaert, 2002, Regime Switches in Interest Rates, Journal of Business and Economic Statistics, 20, 163-182.

[4] Artis, M. and M. Marcellino, 2001, Fiscal Forecasting: The Track Record of the IMF, OECD and EC. Econometrics Journal 4, S20-S36.

[5] Bai, J., 1997, Estimation of a Change Point in Multiple Regression Models. Review of Economics and Statistics 79, 551-563.

[6] Bai, J. and P. Perron, 1998, Estimating and Testing Linear Models with Multiple Structural Changes. Econometrica 66, 47-78.

[7] Bates, J.M. and C.W.J. Granger, 1969, The Combination of Forecasts. Operations Research Quarterly 20, 451-468.

[8] Batchelor R. and P. Dua, 1991, Blue Chip Rationality Tests. Journal of Money, Credit and Banking 23, 692-705.

[9] Batchelor, R. and D.A. Peel, 1998, Rationality Testing under Asymmetric Loss. Economics Letters 61, 49-54.

[10] Box, G. and G. Jenkins, 1970, Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

[11] Breiman, 1995, Better Subset Regression Using the Nonnegative Garrote. Technometrics 37, 373-384.

[12] Breiman, 1996, Bagging Predictors, Machine Learning, 24, 123-140.

[13] Brown, B.Y. and S. Maital, 1981, What do Economists Know? An Empirical Study of Experts' Expectations. Econometrica 49, 491-504.

[14] Campbell, B. and E. Ghysels, 1995, Federal Budget Projections: A Nonparametric Assessment of Bias and Efficiency. Review of Economics and Statistics, 17-31.

[15] Capistran, C., 2005, Bias in Federal Reserve Inflation Forecasts: Is the Federal Reserve Irrational or Just Cautious. Mimeo, Banco de Mexico.

[16] Chong, Y.Y. and D.F. Hendry, 1986, Econometric evaluation of linear macro-economic models, Review of Economic Studies 53:671-690.

[17] Christoffersen, P. and K. Jacobs, 2004, The Importance of the Loss Function in Option Valuation. Journal of Financial Economics, 72, 291-318.

[18] Christoffersen, P.F. and F.X. Diebold, 1997, Optimal Prediction under Asymmetric Loss. Econometric Theory 13, 808-817.

[19] Clark, T.E. and M.W. McCracken, 2001, Tests of Equal Forecast Accuracy and Encompassing for Nested Models. Journal of Econometrics 105, 85-110.

[20] Clark, T.E. and K.D. West. 2004. Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis. Working Paper 04-03, Kansas City Federal Reserve, Kansas City, USA.

[21] Clemen, R.T., 1989, Combining Forecasts: A Review and Annotated Bibliography. International Journal of Forecasting 5, 559-581.

[22] Clements, M.P. and D.F. Hendry, 1998, Forecasting Economic Time Series, Cambridge University Press.

[23] Clements, M.P. and D.F. Hendry, 2006, Forecasting with Breaks in Data Processes, in C.W.J. Granger, G. Elliott and A. Timmermann (eds.) Handbook of Economic Forecasting, 605-657, Amsterdam, North-Holland.

[24] Clements, M.P. and D.F. Hendry, 2002, Modelling Methodology and Forecast Failure. Econometrics Journal 5, 319-344.

[25] Corradi, V. and N.R. Swanson, 2002, A Consistent Test for Out of Sample Nonlinear Predictive Ability. Journal of Econometrics 110, 353-381.

[26] Corradi, V. and N.R. Swanson, 2006, Predictive Density Evaluation, in C.W.J. Granger, G. Elliott and A. Timmermann (eds.) Handbook of Economic Forecasting, 197-286, Amsterdam, North-Holland.

[27] Croushore, D., 2006, Forecasting with Real-Time Macroeconomic Data. Pages 961-982 in G. Elliott, C. Granger and A. Timmermann (eds.) Handbook of Economic Forecasting. North-Holland: Amsterdam.

[28] Croushore, D. and T. Stark, 2003, A Real-time Data Set for Macroeconomists: Does the Data Vintage Matter? Review of Economics and Statistics 85, 605-617.

[29] Davies, A. and K. Lahiri, 1995, A New Framework for Analyzing three-dimensional Panel Data. Journal of Econometrics 68, 205-227.

[30] Del Negro, M., F. Schorfheide, F. Smets and R. Wouters, 2006, On the Fit of New-Keynesian Models. Forthcoming in Econometric Reviews.

[31] Diebold, F., Gunther, T., and A., Tay, 1998, Evaluating Density Forecasts, International Economic Review 39, 863-883.

[32] Diebold, F. X. and L. Kilian, 2000, Unit-Root Tests are Useful for Selecting Forecasting Models, Journal of Business and Statistics, 18, 265-273.

[33] Diebold, F.X. and R. Mariano, 1995, Comparing Predictive Accuracy. Journal of Business and Economic Statistics 13, 253-65.

[34] Diebold, F.X. and P. Pauly, 1990, The use of prior information in forecast combination, International Journal of Forecasting 6:503-508.

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

[36] Doan, T., R. Litterman and C. Sims, 1984, Forecasting and Conditional Projection using Realistic Prior Distributions", Econometric Reviews, 3, 1-144 (with discussion).

[37] Ehrbeck, T. and Waldmann, R., 1996, Why are Professional Forecasts Biased? Agency versus Behavioral Explanations, Quarterly Journal of Economics 111, 21-40.

[38] Elliott, G., 2005, Forecasting in the presence of a break. Mimeo, UCSD.

[39] Elliott, G., I. Komunjer and A. Timmermann, 2005, Estimating Loss Function Parameters. Review of Economic Studies 72, 1107-1125.

[40] Elliott, G., I. Komunjer and A. Timmermann, 2006, Biases In Macroeconomic Forecasts: Irrationality or Asymmetric Loss? Mimeo UCSD.

[41] Elliott, G. and R. Lieli, 2006, .Predicting Binary Outcomes, manuscript, UCSD.

[42] Elliott, G. and U. Mueller, 2006, "Efficient Tests for General Persistent Time Variation in Regression Coefficients", Review of Economic Studies, 73, 907-940.

[43] Elliott, G. and A. Timmermann, 2004, Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions. Journal of Econometrics 122, 47-79.

[44] Elliott, G. and A. Timmermann. 2005. Optimal Forecast Combination Weights Under Regime Switching. International Economic Review 46, 1081-1102.

[45] Engle, R.F. and C.W.J. Granger, 1987, Co-integration and Error Correction: Representation, Estimation and Testing. Econometrica 55, 251-276.

[46] Fair, R.C. and R. Shiller, 1990, Comparing Information in Forecasts from Econometric Models, American Economic Review, 80, 375-89.

[47] Figlewski, S. and P. Wachtel, 1981, The Formation of Inflationary Expectations. Review of Economics and Statistics 63, 1-10.

[48] Forni, M., M. Hallin, M. Lippi, and L. Reichlin, 2000. The Generalized Factor Model: Identification and Estimation. Review of Economics and Statistics 82, 540-554.

[49] Forni, M., M. Hallin, M. Lippi, and L. Reichlin, 2003, The Generalized Dynamic Factor Model: Forecasting and One Sided Estimation, CEPR working paper 3432.

[50] Gallant, R. 1981, On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form. Journal of Econometrics 15, 211-245.

[51] Garcia, R. and P. Perron, 1996, An Analysis of the Real Interest Rate under Regime Shifts. Review of Economics and Statistics 78(1), 111-125.

[52] Geweke, J. and C. Whiteman, 2006, Bayesian Forecasting. Pages 3-80 in G. Elliott, C.W.J. Granger and A. Timmermann (eds.) Handbook of Economic Forecasting. North-Holland: Amsterdam.

[53] Giacomini, R., and H., White, 2006, Tests of Conditional Predictive Ability. Econometrica 74, 6, 1545-1578.

[54] Goyal, A. and I. Welch, 2003, Predicting the Equity Premium with Dividend Ratios. Management Science 49, 639-654.

[55] Granger, C.W.J., 1966, The Typical Spectral Shape of an Economic Variable. Econometrica 34, 179-192.

[56] Granger, C.W.J., 1969, Prediction with a generalized cost function, OR, 20, 199-207.

[57] Granger, C.W.J., 1999, Outline of Forecast Theory Using Generalized Cost Functions. Spanish Economic Review 1, 161-173.

[58] Granger, C.W.J. and M. Machina, 2006, Forecasting and Decision Theory. Pages 81-98 in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), Handbook of Economic Forecasting.

[59] Granger, C.W.J. and P. Newbold, 1986, Forecasting Economic Time Series, 2nd Edition. Academic Press, New York.

[60] Granger, C.W.J. and M.H. Pesaran, 2000, Economic and Statistical Measures of Forecast Accuracy. Journal of Forecasting 19, 537-560.

[61] Hamilton, J.D., 1989, A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. Econometrica 57, 357-384.

[62] Hansen, P.R., 2005, A Test for Superior Predictive Ability. Journal of Business and Economic Statistics 23, 365-380.

[63] Harvey, A.C. and S.J. Koopman, 1993, Forecasting Hourly Electricity Demand Using Time-Varying Splines. Journal of American Statistical Association 88, 1228-1236.

[64] Harvey, A.C., 2006, Forecasting with Unobserved Components Time Series Models. Pages 327-412 in G. Elliott, C.W.J. Granger and A. Timmermann.

[65] Hendry, D.F. and H-M. Krolzig, 2004, Automatic Model Selection: A New Instrument for Social Science. Electoral Studies, 23, 525–544.

[66] Hong, H. and J. D. Kubik, 2003, Analyzing the Analysts: Career Concerns and Biased Earnings Forecasts. Journal of Finance 58,1, 313-351.

[67] Inoue, A. and L. Kilian. 2004, In-sample or out-of-sample Tests of Predictability: Which one should we use? Econometric Reviews 23(4), 371-402.

[68] Inoue, A. and L. Kilian. 2005, How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation, manuscript, North Carolina State University.

[69] Ito, T., 1990, Foreign Exchange Rate Expectations: Micro Survey Data, American Economic Review, 80, 434-449.

[70] Jagannathan, R. and T. Ma, 2003, Risk Reduction in Large Portfolios: Why imposing the wrong constraints helps. Journal of Finance 58, 1651-1684

[71] James, W. and Stein, C., 1961, Estimation with Quadratic Loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1 Berkeley, CA: University of California Press, 361-379.

[72] Kadiyala, K.R. and S. Karlsson, 1993, Forecasting with Generalized Bayesian Vector Autoregressions. Journal of Forecasting 12, 365-378.

[73] Kadiyala, K.R. and S. Karlsson, 1997, Numerical Methods for Estimation and Inference in Bayesian VAR-Models", Journal of Applied Econometrics, 12, 99-132.

[74] Keane, M.P. and D.E. Runkle, 1990, Testing the Rationality of Price Forecasts: New Evidence from Panel Data. American Economic Review 80, 714-735.

[75] Koenker, R.W. and G.W. Bassett, 1978, Regression Quantiles, Econometrica 46, 33-50.

[76] Leamer, E., 1978, Specification Searches. Wiley, Oxford.

[77] Ledoit, O. and M. Wolf, 2003, Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection. Journal of Empirical Finance 10, 603-621.

[78] Leitch, G. and J.E. Tanner, 1991, Economic Forecast Evaluation: Profits Versus the Conventional Error Measures, American Economic Review 81, 580-90.

[79] Lettau, M. and S. Ludvigsson, 2001, Consumption, aggregate wealth, and expected stock returns. Journal of Finance 56, 815-850.

[80] Lim, T., 2001, Rationality and Analysts' Forecast Bias. Journal of Finance 56-1, 369-385.

[81] Litterman, R.B., 1980, A Bayesian Procedure for Forecasting with Vector Autoregressions. Working Paper, Massachusetts Institute of Technology.

[82] Litterman, R.B., 1986, Forecasting with Bayesian Autoregressions — Five Years of Experience, Journal of Business and Economic Statistics, 4, 25-38.

[83] Marcellino, M., 2004, Forecast pooling for short time series of macroeconomic variables, Oxford Bulletin of Economic and Statistics 66:91-112.

[84] Marcellino, M., J.H. Stock and M.W. Watson, 2006, A comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series. Journal of Econometrics 135, 499-526.

[85] Makridakis, S. and M. Hibon, 2000, The M3-Competition: Results, Conclusions and Implications. International Journal of Forecasting 16 451-476.

[86] Mamaysky, H., M. Spiegel and H. Zhang, 2006, Improved Forecasting of Mutual Fund Alphas and Betas. Mimeo, Yale University.

[87] Meese, R.A. and K. Rogoff, 1983, Empirical exchange rate models of the seventies : Do they fit out of sample? Journal of International Economics 14, 3-24.

[88] Mincer, J. and V. Zarnowitz, 1969, The Evaluation of Economic Forecasts. In J. Mincer, ed., Economic Forecasts and Expectations. National Bureau of Economic Research, New York.

[89] Mishkin, F.S., 1981, Are Markets Forecasts Rational? American Economic Review 71, 295-306.

[90] Nelson, C. and C. Plosser, 1982, Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications. Journal of Monetary Economics 10, 139-162.

[91] Newey, W. and J. Powell, 1987, Asymmetric Least Squares Estimation and Testing. Econometrica 55, 819-847.

[92] Pagan, A., 2003, Report on Modelling and Forecasting at the Bank of England. Bank of England.

[93] Palm, F. C. and A. Zellner, 1992, To combine or not to combine? Issues of combining forecasts, Journal of Forecasting 11:687-701.

[94] Patton, A. and A. Timmermann, 2005, Properties of Optimal Forecasts under Asymmetric Loss and Nonlinearity. Forthcoming in Journal of Econometrics.

[95] Patton, A. and A. Timmermann, 2006, Testing Forecast Optimality Under Unknown Loss. Forthcoming in Journal of American Statistical Association.

[96] Paye, B. and A. Timmermann, 2006, Instability of Return Prediction Models. Journal of Empirical Finance 13 (3), 274-315.

[97] Perez-Quiros, G. and A. Timmermann, 2000, Firm Size and Cyclical Variations in Stock Returns. Journal of Finance, 1229-1262.

[98] Pesando, J.E., 1975, A Note on the Rationality of the Livingston Price Expectations. Journal of Political Economy 83, 849-858.

[99] Pesaran, M.H., D. Pettenuzzo and A. Timmermann, 2006, Forecasting Time Series Subject to Multiple Structural Breaks. Review of Economic Studies 73, 1057-1084.

[100] Pesaran, M.H. and S. Skouras, 2002, Decision-based Methods for Forecast Evaluation. In Clements, M.P. and D. F. Hendry (Eds.), A Companion to Economic Forecasting. Blackwell, Oxford.

[101] Pesaran, M.H. and A. Timmermann, 2005a, Small Sample Properties of Forecasts from Autoregressive Models under Structural Breaks. Journal of Econometrics 129, 183-217.

[102] Pesaran, M.H. and A. Timmermann, 2005b, Real time Econometrics. Econometric Theory 11, 212-231

[103] Pesaran, M.H. and A. Timmermann, 2006, Selection of Estimation Window in the Presence of Breaks. Forthcoming in Journal of Econometrics.

[104] Pesaran, M.H. and M. Weale, 2006, Survey Expectations. Pages 715-776 in the Handbook of Economic Forecasting, G. Elliott, C.W.J. Granger, and A.Timmermann (eds.), North-Holland.

[105] Psaradakis, Z. and F. Spagnolo, 2005, Forecast Performance of Nonlinear Error-Correction Models with Multiple Regimes. Journal of Forecasting 24, 119-138.

[106] Racine, J., 2001, On the Nonlinear Predictability of Stock Returns using Financial and Economic Variables. Journal of Business and Economic Statistics 19, 380-382.

[107] Raftery, A.E., D. Madigan and J.A. Hoeting, 1997, Bayesian model averaging for linear regression models, Journal of the American Statistical Association 92, 179-191.

[108] Rapach, D. and M. Wohar, 2006, Structural Breaks and Predictive Regression Models of Aggregate US Stock Returns. Journal of Financial Econometrics 4(2), 238-274.

[109] Robertson, J. and E. Tallman, 1999, Vector Autoregressions: Forecasting and Reality, Federal Reserve Bank of Atlanta Economic Review, First Quarter.

[110] Romer, C.D. and D.H. Romer, 2000, Federal Reserve Information and the Behavior of Interest Rates. American Economic Review 90(3), 429-457.

[111] Scharfstein, D. and J. Stein, 1990, Herd Behavior and Investment. American Economic Review 80, 464-479.

[112] Sims, C.A., 1980, Macroeconomics and Reality. Econometrica 48, 1-48.

[113] Sims, C.A., 2002, The Role of Models and Probabilities in the Monetary Policy Process. Mimeo, Princeton University.

[114] Skouras, S., 2001, Decisionmetrics: A Decision-based Approach to Econometric Modeling. Mimeo, Santa Fe Institute.

[115] Stock, J.H. and M.W. Watson, 1996, Evidence on structural instability in macroeconomic time series relations. Journal of Business and Economic Statistics 14, 11-30.

[116] Stock, J.H. and M.W. Watson, 1998, Median Unbiased Estimation of Coefficient Variance in a Time Varying Parameter Model, Journal of the American Statistical Association, 93, 349-358.

[117] Stock, J.H. and M.W. Watson, 1999. A Comparison of Linear and Nonlinear Models for Forecasting Macroeconomic Time Series. In R. Engle and H. White (eds.), Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W.J. Granger. Oxford University Press.

[118] Stock, J.H. and M.W. Watson, 1999, Forecasting Inflation. Journal of Monetary Economics 44, 293-335.

[119] Stock, James H., and Mark W. Watson, 2002, Macroeconomic Forecasting Using Diffusion Indexes. Journal of Business and Economic Statistics 20:147-162.

[120] Stock, J.H. and M.W. Watson. 2005, An Empirical Comparison of Methods for Forecasting Using Many Predictors. Mimeo, Harvard and Princeton University.

[121] Sullivan, R., A. Timmermann and H. White, 1999, Data-Snooping, Technical Trading Rules and the Bootstrap. Journal of Finance 54, 1647-1692.

[122] Svensson, L.E.O., 1997, Inflation Forecast Targeting: Implementing and Monitoring Inflation Targets. European Economic Review 41, 1111-1146.

[123] Swanson, N. and H. White, 1995, A Model Selection Approach to Assessing the Information in the Term Structure using Linear Models and Artificial Neural Networks. Journal of Business and Economic Statistics 13, 265-276.

[124] Tay, A.S. and K.F. Wallis, 2000, Density Forecasting: A Survey. Journal of Forecasting 19, 235-254.

[125] Taylor, M.P. and Sarno, L., 2002, Purchasing Power Parity and the Real Exchange Rate. International Monetary Fund Staff Papers 49, 65-105.

[126] Terasvirta, T., 2006. Forecasting Economic Variables with Nonlinear Models. Pages 423-458 in G. Elliott, C.W.J. Granger, A. Timmermann, eds. Handbook of Economic Forecasting. North-Holland: Amsterdam.

[127] Terasvirta, T., van Dijk, D., Medeiros, M.C., 2005, Smooth Transition Autoregressions, Neural Networks, and Linear Models in Forecasting Macroeconomic Time Series: A Re-examination. International Journal of Forecasting 21, 755-774.

[128] Tibshirani, R., 1996, Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society B 58, 267-288.

[129] Timmermann, A. 2006. Forecast Combinations. Pages 135-196 in G. Elliott, C.W.J. Granger, A. Timmermann, eds. Handbook of Economic Forecasting. North-Holland: Amsterdam.

[130] Timmermann, A., 2007, An Evaluation of the World Economic Outlook Forecasts. Forthcoming in IMF Staff Papers.

[131] Truman, B., 1994, Analyst Forecasts and Herding Behavior, Review of Financial Studies 7, 97-124.

[132] Turner, J, 2004, Local to Unity, Long Horizon Forecasting Thresholds for Model Selection in the AR(1), Journal of Forecasting, 23, 513-539.

[133] Varian, H. R., 1974, A Bayesian Approach to Real Estate Assessment. In Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage, eds. S.E. Fienberg and A. Zellner, Amsterdam: North Holland, 195-208.

[134] Weiss, A.A., 1996, Estimating Time Series Models Using the Relevant Cost Function. Journal of Applied Econometrics 11, 539-560.

[135] West, M. and J. Harrison, 1997, Bayesian Forecasting and Dynamic Models, second edition, Springer Series in Statistics, Springer Verlag: New York.

[136] West, K.D., 1996, Asymptotic Inference about Predictive Ability. Econometrica 64, 1067-84.

[137] West, K.D., H.J. Edison and D. Cho, 1993, A Utility-based Comparison of Some Models of Exchange Rate Volatility. Journal of International Economics 35, 23-46.

[138] West, K.D. and M.W. McCracken, 1998, Regression-Based Tests of Predictive Ability, International Economic Review 39, 817-840.

[139] White, H., 2000, A Reality Check for Data Snooping. Econometrica 68, 1097-1127.

[140] White, H. 2001, Asymptotic Theory for Econometricians, 2nd Edition, Academic Press: New York.

[141] Whiteman, C.H, 1996, Bayesian Prediction under Asymmetric Linear Loss: Forecasting State Tax Revenues in Iowa. In W.O. Johnson, J.C. Lee and A. Zellner (eds.) Forecasting, Prediction and Modeling in Statistics and Econometrics: Bayesian and non-Bayesian Approaches. New York: Springer-Verlag.

[142] Zarnowitz, V., 1985, Rational Expectations and Macroeconomic Forecasts. Journal of Business and Economic Statistics 3, 293-311.

[143] Zellner, A., 1986, Bayesian Estimation and Prediction Using Asymmetric Loss Functions. Journal of the American Statistical Association, 81, 446-451.

Table 1: Out-of-sample Forecasting performance (annualized root mean squared error) for various forecasting models, 1970 - 2003

	Inflation		SP500 Return	
ModelName	**Expanding window**	**10-year rolling window**	**Expanding window**	**10-year rolling window**
Autoregressive (AR)	0.78	0.77	15.9	15.9
Factor-augmented AR	0.78	0.80	15.9	15.9
BVAR - random walk prior	0.81	0.92	17.7	20.0
BVAR - white noise prior	0.81	0.94	17.4	19.1
Exponential smoothing	0.76	0.76	16.0	16.3
Double exp. smoothing	0.78	0.83	16.2	18.6
STAR 1	0.88	0.80	16.8	17.5
STAR 2	0.83	0.81	17.0	17.4
One Layer neural net	0.80	0.82	17.1	17.4
Two Layer neural net	0.78	0.77	16.0	17.5
Combined forecast (average)	0.75	0.75	15.9	16.3
Previous best	0.77	0.78	16.1	16.5

Table 2. p-values for pairwise test of identical out-of-sample mean squared errors under the Giacomini-White (2006) test.

Inflation model two

model one	Factor AR	BVAR - RW	BVAR - WN	Exp. Smoothing	Double exp Smoothing	STAR_1	STAR_2	One Layer NN	Two Layer NN	Average	Best
Autoregressive (AR)	0.028	0.000	0.000	0.813	0.012	0.155	0.075	0.142	0.943	0.162	0.542
Factor-augmented AR		0.005	0.001	0.124	0.324	0.806	0.674	0.569	0.133	0.002	0.328
BVAR - random walk prior			0.226	0.000	0.027	0.003	0.008	0.047	0.001	0.000	0.001
BVAR - white noise prior				0.000	0.013	0.001	0.003	0.021	0.000	0.000	0.000
Exponential smoothing					0.000	0.125	0.084	0.128	0.766	0.347	0.364
Double exp. smoothing						0.224	0.659	0.913	0.015	0.000	0.020
STAR 1							0.507	0.496	0.115	0.003	0.427
STAR 2								0.807	0.079	0.010	0.220
One Layer neural net									0.132	0.049	0.242
Two Layer neural net										0.167	0.588
Combined forecast (average)											0.076

Stock Returns

Autoregressive (AR)	0.937	0.000	0.000	0.073	0.000	0.010	0.001	0.009	0.060	0.150	0.177
Factor-augmented AR		0.000	0.000	0.050	0.000	0.004	0.000	0.004	0.047	0.054	0.155
BVAR - random walk prior			0.009	0.000	0.065	0.001	0.000	0.000	0.007	0.000	0.000
BVAR - white noise prior				0.000	0.528	0.048	0.022	0.034	0.100	0.000	0.002
Exponential smoothing					0.000	0.015	0.005	0.024	0.120	0.063	0.509
Double exp. smoothing						0.059	0.072	0.038	0.185	0.000	0.000
STAR 1							0.928	0.901	0.968	0.009	0.008
STAR 2								0.982	0.922	0.005	0.038
One Layer neural net									0.904	0.007	0.017
Two Layer neural net										0.098	0.176
Combined forecast (average)											0.528

Table 3. Forecasting performance under lin-lin loss, forecast models estimated by quantile regression.

Model	Quantile	Inflation	stock return
Autoregressive	0.35	0.27	5.72
Factor-augmented AR	0.35	0.34	5.71
Autoregressive	0.5	0.30	6.09
Factor-augmented AR	0.5	0.40	6.07
Autoregressive	0.65	0.29	5.64
Factor-augmented AR	0.65	0.42	5.68

Figure 1: Inflation Forecasts

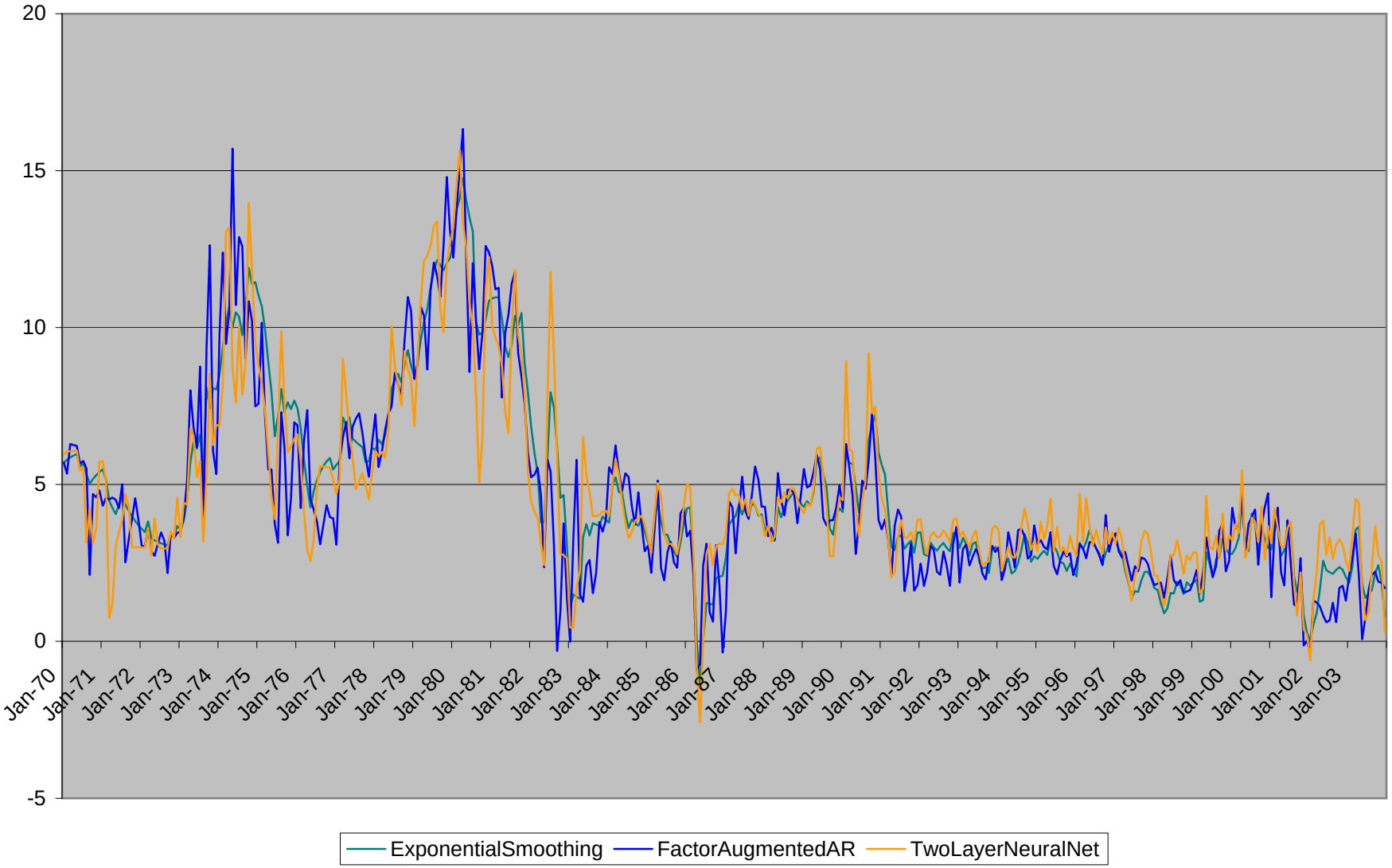

Figure 2: Forecasts of Stock Returns