1

Introduction

Our aim with this book is to present an overview of the theory and methods underlying forecasting as currently practiced in economics and finance, but more widely applicable to a great range of forecasting problems. We hope to provide an overview that is useful to practitioners in places such as central banks and financial institutions, academic researchers as well as graduate students seeking a point of entry into the field. The assumed econometric level of the reader is that of someone who has taken a graduate or advanced undergraduate course in econometrics.

Whenever a forecast is being constructed or evaluated, an overriding concern revolves around the practical problem that the best forecasting model is not only unknown but also unlikely to be known well enough to even correctly specify forecasting equations up to a set of unknown parameters. We view this as the only reasonable description of the forecaster's problem. Some methods do claim to find the correct model (oracle methods) as the sample gets very large. However, in any problem with a finite sample there is always a set of models—as opposed to a single model—that are consistent with the data. Moreover, in many situations the data-generating process changes over time, further emphasizing the difficulty in obtaining very large samples of observations on which to base a model. These foundations—using misspecified models to forecast outcomes generated by a process that may be evolving over time—generate many of the complications encountered in forecasting. If the true models were fully known apart from the values of the parameters, Bayesian methods could be used to construct density and point forecasts that, for a given loss function, would be difficult or impossible to beat in practice.

Without knowing the true data-generating process, the problem of constructing a good forecasting method becomes much more difficult. Oftentimes very simple (and clearly misspecified) methods provide forecasts that outperform more complicated methods that seek to exploit the data in ways we would expect to be important and advantageous. As a case in point, simple averages of forecasts from many models, even ones that on their own do not seem to be very good, are often found empirically to outperform carefully chosen model averages or the best individual models.

1.1 OUTLINE OF THE BOOK

The approach of this book is for the most part based on forecasting as a decision-theoretic problem. By this we mean that the forecaster has a specific objective in

mind (i.e., wishes to make a decision) and wants to base this decision on some data. Setting up this approach comprises most of the first part of the book. This part details the basic elements of the decision problem, with chapters on the decision maker's loss function, forecasting as a decision-theoretic problem, and an overview of general approaches to forecasting employing either classical or Bayesian methods. This part of the book provides foundations for understanding how different methods fit together. We also provide details of methods that are subsequently applied to many of the issues examined in the next part of the book, e.g., model selection and forecast combination.

The second part of the book reviews various approaches to constructing forecasting models. Methods employed differ for many reasons: lack of relevant data or the existence of a great deal of potentially relevant data, as well as assumptions made on functional forms for the models. In these chapters we attempt, as far as possible, to present the methods in enough detail that they can be employed without reference to other sources.

The third part of the book examines the evaluation of forecasts, while the fourth part covers forecasting models that deal with special complications such as model instability (breaks) and highly persistent (trending) data. This part also discusses data structures of special interest to forecasters, including real-time data (revised data) and data collected at different frequencies.

Finally, the fourth part of the book presents various extensions and refinements to the forecasting methods covered in the earlier parts of the book, including forecasting under model instability, long-run forecasting, and forecasting with data that either take a non-standard form (count data and durations) or are measured at irregular intervals and are subject to revisions.

1.1.1 Part I

The first part of the book motivates that point forecasting should be thought of simply as an application of decision theory. Since much is known about decision theory, much is also known about forecasting. This perspective makes point forecasting a special case of estimation, a field where excellent texts already exist. What makes economic forecasting interesting as a separate topic is the particular details of how decision theory is applied to the problem at hand. To apply this approach, we require a clear statement about the costs of forecast errors $e = y - f$, where y is the outcome being predicted and f is the forecast. The trade-off between different forecasting mistakes is embodied in a loss function, $L(f, y)$ which is discussed in chapter 2, with additional material on the binary case available in chapter 12. We regard loss functions as realistic expositions of the forecaster's objectives, and consider the specification of the loss function as an integral part of the forecaster's decision problem.[1] Different forecasters approaching the same outcome may well have different loss functions which could result in different choices of forecasting models for the same outcome.

The specification of loss functions is often disregarded in economic forecasting, and instead "standard" loss functions such as mean squared error loss tend to

[1] An alternative literature considers features of loss functions and attempts to suggest a good loss function for all forecasting problems. We do not consider this approach and view loss functions as primitive to the forecaster's problem.

be employed. This can prove costly in real forecasting situations as it overlooks directions in which forecast errors are particularly costly. Nonetheless, much of the academic literature is based on these standard loss functions and so we focus much of our survey of methods throughout the second part of the book on these standard loss functions.

Chapter 3 provides a general description of the forecaster's problem as a decision problem. It may strike some readers, more used to the "art" of forecasting, as unusual to cast point forecasting as a decision-theoretic problem. However, even readers who do not explicitly follow this approach are indeed operating within the decision-theoretic framework. For example, most forecasting methods are motivated in one of two ways: either the methods are demonstrated to provide better performance given a loss function (or set of loss functions) through Monte Carlo simulations for reasonable data-generating processes, or alternatively, the forecasting methods are shown to work well for some loss function for a particular set of empirical data $z = (y, x)$, where x represents the set of predictor variables used to forecast the outcome y. Both ways of measuring performance place the forecasting problem within the decision-theoretic approach.

To illustrate this point, consider Monte Carlo simulations of a data-generating process (joint density for the data) regarded as a reasonable representation of some data of interest. The simulation method suggests constructing N independent pseudo samples from this density, constructing N forecasts and evaluating $N^{-1} \sum_{n=1}^{N} L(f^{(n)}, y^{(n)})$, where $y^{(n)}$ is the outcome we wish to forecast, $f^{(n)}$ is the forecast generated by a prediction model, and $L(f, y)$ is the loss function which measures the costs of forecast inaccuracies. Superscripts refer to the individual simulations, $n = 1, \ldots, N$. The simulated average loss is usually thought of as a measure of the performance of the forecasting method or model for this data-generating process. This is reasonable since as N gets large, the sample average is, by standard laws of large numbers, a consistent estimate of the risk at the point of the parameter space for the data-generating process chosen for the Monte Carlo, i.e., as long as $E\,[L(f, y)]$ exists,

$$N^{-1} \sum_{n=1}^{N} L(f^{(n)}, y^{(n)}) \rightarrow^{p} E\,[L(f, y)]\,, \tag{1.1}$$

where the Monte Carlo estimates a point on the risk function and $\rightarrow^{p}$ means convergence in probability. Finding a forecast that minimizes the risk is precisely the setup of a decision-theoretic problem.

The third part of the book discusses methods for evaluation of sequential out-of-sample predictions. In each case, one obtains from the data set T observations of the "realized" loss from the data. One then evaluates the time-series average $T^{-1} \sum_{t=1}^{T} L(f_t, y_t)$, where the t subscript refers to time, as a measure of the expected loss. In this case the assumptions that underlie results such as (1.1) are much more stringent because the sequence of expected losses generated from data are not independently and identically distributed (i.i.d.) as in the Monte Carlo simulations. However, under suitable assumptions again this method estimates risk. When using real (as opposed to simulated) data, we do not know the true parameter values of the data-generating process. Analyzing a variety of economic variables, we get a sense of how well different forecasting methods work for different types of data.

The general setup in chapter 3 is common to forecasters basing their estimation strategies either on frequentist or on Bayesian approaches. Chapters 4 and 5 build on this setup separately for these two approaches. Chapter 4 examines the typical frequentist approaches, explaining general pitfalls that can occur as well as highlighting special cases arising later in the book. Chapter 5 does the same for the Bayesian approach.

Viewing forecasting as a decision-theoretic problem sometimes means that the best forecasting model, despite working well in practice, may actually be a model that is very difficult to interpret economically. This becomes a problem when the forecasting exercise is a step in a decision process, and the forecaster must "explain" the forecast to decision makers or forecast users. In these cases an inferior point forecast that tends to be further away from the outcome may be preferred because it is easier to explain and may be seen to be more credible. Of course in situations where we suspect a lot of overfitting or instability in the relationships between the variables, we might prefer forecasting models that conform to economic theory since they are expected to be more robust. Practically, economically motivated restrictions on forecasting models can just be seen as following the decision-theoretic approach for a restricted set of models.

The final chapter of the first part of the book, chapter 6, examines issues related to model selection. By now the econometrics literature has a very good understanding of the merits and limitations of model selection, which we discuss for general models. From the perspective of forecasting, however, we regard model selection as simply part of the model estimation process. Of interest to the forecaster is the risk of the final forecasting model computed in a way that accounts for the full estimation process. Given the complexity of the distributions of estimators obtained from models whose selection is driven by the data, this issue is difficult to address analytically although it is still of direct relevance to the forecaster.

1.1.2 Part II

Part II of the book provides an overview of the various approaches to forecasting that have become standard in many areas, including the economic and finance forecasting literature. Chapters are based around either the amount of information available—from only the past history of the predicted variable through very large panels of variables—or the general estimation approach, principally parametric or nonparametric methods. To the extent possible, we provide details of how to go about constructing forecasts from the various methods, or alternatively direct readers to explanations available in the literature. We also discuss the trade-offs between different methods. In this sense we endeavor to provide a "first stop" for practitioners wishing to apply the methods covered in this section.

An important insight that arises from the decision-theoretic approach is that there is no single best or dominant approach to constructing a forecast for all possible forecasting situations. We discuss the types of forecasting situations where each individual method is likely to be a reasonable approach and also highlight situations where other approaches should be considered.

Throughout the book, we use a variety of empirical applications to illustrate how different approaches work. In most applications we use so-called pseudo out-of-sample forecasts which simulate the forecast as it could have been generated using data only up to the date of the prediction. This method restricts both

model selection and parameter estimation to rely on data available at the point of the forecast. As time progresses and more data become available, the forecasting method, including the parameter estimates, are updated recursively. Such methods are commonly used to evaluate the usefulness of forecasts; a critical discussion of such out-of-sample forecasting methods versus in-sample methods is provided in part three of the book.

When building a forecasting model for an economic variable, the simplest specification of the conditioning information set is the variable's own past history. This leads to univariate autoregressive moving average, or ARMA, models. Since Box and Jenkins (1970) these models have been extensively used and often provide benchmarks that are difficult to beat using more complicated forecasting methods. Linear ARMA models are also easy to estimate and a large literature has evolved on how best to cover issues in implementation such as lag length selection, generation of multiperiod forecasts, and parameter estimation. We discuss these issues in chapter 7. The chapter also covers exponential smoothing, unobserved components models, and other ways to account for trends when forecasting economic variables.

Chapter 8 continues under the assumption that the information set is limited to the predicted variable's own past, but focuses on nonlinear parametric models. Examples include threshold autoregressions, smooth threshold autoregressions, and Markov switching models. These models have been used to capture evidence of nonlinear dynamics in many macroeconomic and financial time series. Unlike nonparametric models they do not, however, have the ability to provide a global approximation to general data-generating processes of unknown form.

Chapter 9 expands the information set to include multivariate information by considering a natural extension to univariate autoregressive models, namely vector autoregressions, or VARs. VARs provide a framework for producing internally consistent multiperiod forecasts of all the included variables. As used in macroeconomic forecasting VARs typically include a relatively small set of variables, often less than 10, but they still require a large number of parameters to be estimated if the number of included lags is high. To deal with the resulting negative effects of estimation errors on forecasting performance, a large literature has developed Bayesian methods for estimating and forecasting with VARs. Both classical and Bayesian estimation of VARs is covered in the chapter which also deals with forecasting when the future paths of some variables are specified, a common practice in scenario analysis or contingent forecasting.

The emergence of very large data sets has given rise to a wealth of information becoming readily available to forecasters. This poses both a unique opportunity—the potential for identifying new informative predictor variables—but also some real challenges given the limitations to most economic data. Suppose that N potential predictor variables are available, and that N is a large number, i.e., in the hundreds or thousands. Including all variables in the forecasting model—the so-called kitchen sink approach—is generally not feasible or desirable even for linear models since parameter estimation error becomes too large, unless the length of the estimation sample, T, is very large relative to N. Standard forecasting methods that conduct comprehensive model selection searches are also not feasible in this situation. If the true model is sparse, i.e., includes only few variables, one possibility is to use algorithms such as the Lasso, covered in chapter 6, to identify a few key predictors. Another strategy is to develop a few key summary measures that aggregate information from a large cross section of variables. This is the approach

used by common factor models. Chapter 10 describes how these methods can be used in forecasting, including in factor-augmented VAR models that include both univariate autoregressive terms along with information in the factors. Finally, we discuss the possibility of using methods from panel data estimation to generate forecasts.

While chapters 7–10 focus on parametric estimation methods and so assume that a certain amount of structure can be imposed on the forecasting model, chapter 11 considers nonparametric forecasting strategies. These include kernel regressions and sieve estimators such as polynomials and spline expansions, artificial neural networks, along with more recent techniques from the machine-learning literature such as boosted regression trees. Although these methods have powerful abilities to approximate many data-generating processes as the number of terms included by the approach gets large, in practice any given estimated nonparametric model is itself an approximation to this approximation. Notably, the number of terms that can be successfully included in empirical applications will often be severely restricted by the available data sample. These approximate models thus do not have the same approximation ability as the models and thus themselves are approximations. Once again, the algorithm used to fit these forecasting models—along with the loss function used to guide the estimation—become key to their forecasting performance and to avoiding issues related to overfitting.

Forecasts of binary variables, i.e., variables that are restricted to take only two possible values, play a special role in decisions such as households' choice on whether or not to buy a car, the decision on whether to pursue a particular education, or banks' decisions on whether to change interest rates for short-term deposits. Restricting the outcome to only two possible values has the advantage that it crystallizes the costs of making wrong forecasts, i.e., false positives or false negatives. Chapter 12 takes advantage of these simplifications to cover point and probability forecasts of binary outcomes and discusses both statistical and utility-based estimators for such data.

The decision-theoretic approach embodies a loss function that is appropriate for the decision to be made and not, as is so often the case, chosen for convenience. It results in a decision, i.e., a choice of an action to be made. This directs itself to basing estimation on an objective of providing the best decision. Alternatively, we might consider provision of a predictive distribution (density forecast) for an outcome as the objective of the forecasting problem. In chapter 13 we see that this perspective is useful for a wide range of decisions.[2] Distribution forecasts also serve the important role of quantifying the degree of uncertainty surrounding point forecasts.

Distributional forecasting fills an important place in any forecaster's toolbox but it does not replace point forecasting. First, although density forecasts can be used to construct point forecasts, typically it is the point forecast or decision that is required. Second, distributional forecasts rely on the distribution being estimated from data. This brings the loss function or scoring rule—the loss function used to estimate the

[2] Dawid (1984) introduced what he termed the "prequential" approach to statistics, where prequential is a fusing of the words "probability" and "sequential." This approach argued that rather than parameters being the object of statistical inference, the proper approach was to provide a sequence of probability forecasts for an outcome of interest. Hence, the provision of a density is important not just for forecasting, but for statistics in general.

density—back into the problem. Often ad hoc loss functions are employed to estimate the distributional forecast, leading to problems when the distributional forecast is subsequently used to construct the point forecast.

Given the plethora of different modeling approaches for construction of forecasts throughout chapters 7–13, it is not surprising that forecasters frequently have access to multiple predictions of the same outcome. Instead of aiming to identify a single best forecast, another strategy is to combine the information in the individual forecasts. This is the topic of forecast combinations covered in chapter 14. If the information used to generate the underlying forecasts is not available, forecast combination reduces to a simple estimation problem that basically treats the individual forecasts as predictors that could be part of a larger conditioning information set. Special restrictions on the forecast combination weights are sometimes imposed if it can be assumed that the individual forecasts are unbiased. If more information is available on the models underlying the individual forecasts, model combination methods can be used. These weight the individual forecasts based on their marginal likelihood or some such performance measure. Bayesian model averaging is a key example of such methods and is also covered in this chapter.

1.1.3 Part III

The third part of the book deals with forecast evaluation methods. Evaluation of forecast methods is central to the forecasting problem and the difficulties involved in this step explain both the plethora of methods suggested for forecasting any particular outcome and the need for careful evaluation of forecasting methods.

To see the central issue, consider the simple problem of forecasting the next outcome, y_{T+1}, in a sequence of independently and identically distributed data y_t, $t = 1, \ldots, T$ with mean μ, variance σ^2, and no explanatory variables. It is well known that under mean squared error (MSE) loss the best forecast is an estimate of the mean, μ, such as the sample mean $\bar{y}_T = T^{-1} \sum_{t=1}^{T} y_t$. Since the outcome y_{T+1} is a random variable whose distribution is centered on μ, the forecast is typically different from the outcome even if we had a perfect estimate of μ, i.e., if we knew μ, as long as $\sigma^2 > 0$. Observing a single outcome far away from the forecast is therefore not necessarily indicative of a poor forecast. More generally, methods for forecast evaluation have to deal with the fact that (in expectation) the average in-sample loss and the average out-of-sample loss differ. To see this, suppose we use the sample mean as our forecast. For any in-sample observation, $t = 1, \ldots, T$, the MSE of the forecast (or fitted value) is

$$\begin{aligned} E[y_t - \bar{y}_T]^2 &= E\left[(y_t - \mu) - T^{-1} \sum_{t=1}^{T} (y_t - \mu)\right]^2 \\ &= \sigma^2 \left(1 + \frac{T}{T^2} - 2\frac{1}{T}\right) \\ &= \sigma^2 (1 - T^{-1}). \end{aligned}$$

Here the third term in the second line comes from the cross product when we compute the squared terms in the first line.

In contrast, the MSE of out-of-sample forecasts of y_{T+1} is

$$\begin{aligned} E[y_{T+1} - \bar{y}_T]^2 &= E\left[(y_{T+1} - \mu) - T^{-1}\sum_{t=1}^{T}(y_t - \mu)\right]^2 \\ &= \sigma^2\left(1 + \frac{T}{T^2}\right) \\ &= \sigma^2(1 + T^{-1}). \end{aligned}$$

Here there is no cross-product term. Comparing these two expressions, we see that estimation error reduces the in-sample MSE but increases the out-of-sample MSE. In both cases the terms are of order T^{-1} and so the difference disappears asymptotically. However, in many forecasting problems this smaller-order term is important both statistically and economically. When we consider many different models of the outcome, differences in the MSE across models are of the same order as the effects on estimation error. This makes it difficult to distinguish between models and is one reason why model selection is so difficult. The insight that the in-sample fit improves by using overparameterized models, whereas out-of-sample predictive accuracy can be reduced by using such models, strongly motivates the use of out-of-sample evaluation methods, although caveats apply as we discuss in part III of the book.

In the past 20 years many new forecast evaluation methods have been developed. Prior to this development, most academic work on evaluation and ranking of forecasting performance paid very little attention to the consideration that forecasts were obtained from recursively estimated models. Thus, often studies used the sample mean squared forecast error, computed for a particular empirical data set, to give an estimate of a model's performance without accompanying standard errors. An obvious limitation of this approach is that such averages often are averages over very complicated functions of the data. Through their dependence on estimated parameters these averages are also typically correlated across time in ways that give rise to quite complicated distributions for standard test statistics. For some of the simpler ways that forecasts could have been generated recursively, recent papers derive the resulting standard errors, although much more work remains to be done to extend results to many of the popular forecasting methods used in practice.

Chapter 15 first establishes the properties that a good forecast should have in the context of the underlying loss function and discusses how these properties can be tested in practice. The chapter goes from the case where very little structure can be imposed on the loss function to cases where the loss function is known up to a small set of parameters. In the latter case it can be tested that the derivative of the loss with respect to the forecast, the so-called generalized forecast error, is unpredictable given current information. The chapter also shows how assumptions about the loss function can be traded off against testable assumptions on the underlying data-generating process.

Chapter 16 gives an overview of basic issues in evaluating forecasts, along with a description of informal methods. This chapter examines the evaluation of a sequence of forecasts from a single model. Critical values for the tests of forecast efficiency depend on how the forecast was constructed, specifically whether a fixed, rolling, or expanding estimation window was used.

Chapter 17 extends the assessment of the predictive performance of a single model to the situation with more than one forecast to examine and so addresses the issue of which, if any, forecasting method is best. We review ways to compare the forecasting methods and strategies for testing hypotheses useful to identifying methods that work well in practice. Special attention is paid to the case with nested forecasting models, i.e., cases where one model includes all the terms of another benchmark model plus some additional information. We distinguish between tests of equal predictive accuracy and tests of forecast encompassing, the latter case referring to situations where one forecast dominates another. We also discuss how to test whether the best among many (possibly thousands) of forecasts is genuinely better than some benchmark.

Chapter 18 examines the evaluation of distributional forecasts. A complication that arises is that we never observe the density of the outcome; only a single draw from the distribution gets observed. Various approaches have been suggested to deal with this issue, including logarithmic scores and probability integral transforms. We discuss these as well as ways to evaluate whether the basic features of a density forecast match the data.

1.1.4 Part IV

The fourth part of the book covers a variety of topics that are specific to forecasting. Chapter 19 discusses predictions under model instability. This chapter builds on the earlier observation that all forecasting models are simplified representations of a much more complex and evolving data-generating process. A key source of model misspecification is the constant-parameter assumption made by many prediction models. Empirical evidence suggests that simple ARMA models are in fact misspecified for many macroeconomic variables. The chapter first discusses how model instability can be monitored before moving over to discuss prediction approaches that specifically incorporate time-varying parameters, including random walk or mean-reverting parameters and regime switching parameters.

The previous chapters deal with cases where the forecast horizon is relatively short. Chapter 20 directly attacks the case where the forecast horizon can be long. Oftentimes a policy maker or budget office is interested in 5 or 10-year forecasts of revenue or expenditures. Interest may also lie in forecasts of the average growth rate over some period. From an estimation perspective, whether the forecast horizon is short or long is measured relative to the length of the data sample. We discuss these issues in chapter 20.

Real-time forecasting methods emphasize the need to ensure that all information and all methods used to construct a forecast would have been available in real time. This consideration becomes particularly relevant in so-called pseudo out-of-sample forecasts that simulate a sequence of historical forecasts. Many macroeconomic time series are subject to revisions that become available only after the date of the forecast. Since the selection of a forecasting model and estimation of its parameters may depend on the conditioning information set, which vintage of data is used can sometimes make a material difference. Similar issues related to data availability are addressed by a relatively new field known as nowcasting which uses filtering and updating algorithms to account for the jagged-edge nature of data, i.e., the fact that data are released at different frequencies and on different dates. These issues are covered in chapter 21.

This chapter also covers models for predicting data that take the format of either counts, and so are restricted to being an integer number, or durations, i.e., the length of the time intervals between certain events. The nature of the dependent variable gives rise to specific forecasting models, such as Poisson models, that are different from the models covered in the previous chapters of the book. Count models have gained widespread popularity in the context of analysis of credit events such as bankruptcies or credit card default, while duration analysis is used to predict unemployment spells and times between trades in financial markets.

1.2 TECHNICAL NOTES

Throughout the book we follow standard statistical methods which view the data as realizations of underlying random variables. Objective functions and other functions of interest are then also functions of random variables. Further, we assume that all functions are measurable, including functions that arise from maximizations of functions over parameters. We are rarely explicit about these assumptions, though this is seldom an issue for the functions examined in the book.

The decision-theoretic approach relies on the existence of risk or expected loss. For loss functions that are bounded, this is usually not problematic, but many popular loss functions are not bounded. For example, mean squared error loss and mean absolute error loss are the most popular loss functions in practice, and neither is bounded. It is fairly standard in the forecasting literature to simply assume that the expected loss exists, and further assume that the asymptotic limit of expected loss is the expected value of the limiting random variable that measures the loss. Throughout the book we follow this practice without giving conditions. Forecasting practice in some instances does seem to enforce "boundedness" of a sort on forecast losses; for example, in evaluating nonlinear models with mean squared error loss, often extreme forecasts that could lead to very large losses are removed and so the loss is in effect bounded.

Throughout the book we tend not to present results as fully worked theorems but instead give the main conditions under which the results hold. Original papers with the full set of conditions are cited. The reasons for this approach are twofold. First, often there are many overlapping sets of conditions that would result in lengthy expositions on often very straightforward methods if we were to include all the details of a result. Second, many of the conditions are highly technical in nature and often difficult or impossible to verify.