Forecasting in Economics and Finance

Graham Elliott[1] and Allan Timmermann[1,2]

[1]Department of Economics, University of California, San Diego, La Jolla, California 92093; email: atimmermann@ucsd.edu

[2]Center for Research in Econometric Analysis of Time Series, Aarhus University, DK-8210 Aarhus, Denmark

Annu. Rev. Econ. 2016. 8:81–110

The *Annual Review of Economics* is online at economics.annualreviews.org

This article's doi: 10.1146/annurev-economics-080315-015346

JEL codes: C53, E17, G17

Keywords

risk, forecast models, big data, parameter estimation, model misspecification, model instability, forecast evaluation

Abstract

Practices used to address economic forecasting problems have undergone substantial changes over recent years. We review how such changes have influenced the ways in which a range of forecasting questions are being addressed. We also discuss the promises and challenges arising from access to big data. Finally, we review empirical evidence and experience accumulated from the use of forecasting methods to a range of economic and financial variables.

1. INTRODUCTION

Methods for selecting a forecasting model, estimating its parameters, and evaluating the precision of the model's forecasts have improved in fundamental ways over the past 20 years. This review provides a critical survey of how these developments have helped address practical questions of critical importance to economic forecasters and points to some of the remaining issues.

Our review of the state of economic forecasting emphasizes a few key points. First, economic forecasting is fundamentally a decision problem, and thus the economics underlying the forecasting problem deserves to play a prominent role. A good forecast is one that generates low expected loss when used in economic decisions. The costs of different mistakes—typically different magnitudes of over- and underpredictions of the outcome—must therefore be considered in selecting a forecasting model, estimating its parameters, and generating forecasts. Policy makers, individual households, or firms can only trade off between alternative forecasting methods if they understand their underlying loss functions. This point may seem obvious, but in practice the vast majority of empirical studies on economic forecasting resorts to assuming squared error loss without dedicating much time to addressing whether this loss function is sensible for the decision problem at hand.

Viewing economic forecasting as a decision theoretic problem that can be informed by observed data implies that forecasting becomes an estimation problem. For example, the point forecasting problem becomes equivalent to the statistical problem of estimating a parameter of the conditional probability distribution of the outcome. This insight means that we can draw on a large body of literature on how to select among forecasting models, how to estimate their parameters, and how to evaluate the resulting forecasts.

Forecasting models that are simple enough to lend themselves to empirical estimation must be strongly condensed representations of a far more complex—and possibly changing—data-generating process (DGP). The correct perspective is therefore to regard all forecasting models as being misspecified. This means that estimation of forecasting models by methods such as maximum likelihood may not be a good approach, and other estimation methods should reasonably be considered.

Evaluation of a particular forecasting method is undertaken by considering its expected loss, or risk.[1] The notion of "risk" used in this context is different from its meaning in finance. A forecasting method's risk depends on the unknown parameters of the DGP and will of course also depend on the forecaster's loss function. In most forecasting problems, the underlying DGP is a function of a large number of parameters, making a full examination of risk difficult in practice. However, as shown below, the risk perspective offers important insights. When models and parameters are unknown, individual forecasting approaches may have attractive risk functions for certain values of the parameters of the DGP, but offer a less attractive risk profile when evaluated at other values.

A key point in the review is therefore that there is almost never a single forecasting approach that uniformly dominates all other alternatives to forecasting. Indeed, suppose we knew the true process that generated a particular data series, but did not know the values of the parameters. Then there would be no model uncertainty. However, the need to estimate model parameters means that simpler, misspecified models might actually produce better forecasts than the true model with estimated parameters. This is an important insight and helps explain why, in practice, no single method dominates economic forecasting and why some methods seem to work better for certain types of variables (e.g., persistent variables such as price inflation or wages) than for other

[1] Early work highlighting the importance of the forecast user's loss function to the evaluation of economic forecasts includes Granger & Pesaran (2000), Pesaran & Skouras (2002), and Skouras (2007).

variables (e.g., real economic growth) that correspond to very different values for the parameters of the DGP.

Accounting for model misspecification and parameter estimation error leads to a second key point, namely that different forecasting methods often can be combined to produce improved forecasts. Model or forecast combination has been used to deal with a variety of issues, ranging from model instability to the effect of uncertainty about the best forecasting model, the presence of parameter estimation error, or the pooling of information from different surveys or data sources. In practice, we often cannot distinguish statistically between a group of models with similar forecasting performance, so it makes sense to combine forecasts from these models, rather than arbitrarily attempt to identify a single best model.

A third key point is that forecast evaluation and model comparison are an important part of the forecasting process and can be conducted at a higher level of rigor today than in earlier years. Historically, common practice was to report estimates of different methods' risk—typically sample averages—and, possibly, use an informal ranking to compare the risk of different forecasting methods, in all cases without accompanying standard errors. In the past 20 years, a large literature has developed test statistics and limit theory that allow us to evaluate and compare different models' forecasting performance. Forecast comparisons can be undertaken even for very large sets of models and can be used to control for the effects of data mining arising from the search for a superior prediction model across multiple specifications.

We hope that our review addresses the types of questions that economic forecasters are likely to ask. We further hope that it will stimulate readers to ask new critical questions. We have therefore structured our review around a set of key practical issues, each of which is raised and then addressed. Invariably, these questions are related, so we cross-reference some of the important issues that arise across different areas.

First, however, we provide a framework for understanding the economic forecasting problem that is broad enough to explain many of the central issues that are outstanding in the economic forecasting literature.

2. FOUNDATIONS FOR ECONOMIC FORECASTING

This section provides the theoretical foundation for the forecasting problem, offering a unified perspective on many of the most important issues that arise. We refer back to this section extensively in our subsequent analysis.

2.1. Economic Forecasting as a Decision Problem

To represent the economic forecasting problem, let $f_{t+h|t}$ be the h-period-ahead forecast at time t of some outcome variable, y_{t+h}. Though $f_{t+h|t}$ and y_{t+h} are often scalars, they can also be vectors, typically of the same dimension as we have a forecast for each outcome. At a given point in time, t, the forecast $f_{t+h|t} = f(z_t)$ is a function of data observed at time t, $z_t = \{x_\tau, y_\tau\}_{\tau=1}^{t}$, so the data available for construction of the forecast given a sample of T observations are $z_T = \{x_t, y_t\}_{t=1}^{T}$.

A key object in generating as well as evaluating forecasts is the loss function, $L(f_{T+h|T}, y_{T+h})$. This is a mapping from $f \in \mathcal{F}$ and $y \in \mathcal{Y}$ to a subset of $\mathbb{R}$, typically $\mathbb{R}^+$. Alternatively, if $f_{t+h|t}$ is a distribution forecast (as opposed to a point forecast), the loss function is typically referred to as a scoring rule (see Gneiting & Raftery 2007 for further discussion of scoring rules).

The decision theoretic approach to constructing an economic forecast typically results in searching for a model that makes the sample average loss as small as possible, given the observed data. Given a data sample $z_T = \{x_\tau, y_\tau\}_{\tau=1}^{T}$, this can be accomplished by choosing the

forecasting model $f(z_t)$ that minimizes the sample average loss:

$$(T-b)^{-1}\sum_{t=1}^{T-b} L(f(z_t), y_{t+b}). \tag{1}$$

Alternatively, we might instead employ a structural economic model to generate a forecast $f(z_t)$ at any time t.

The hope, either way, is to obtain a forecasting model with the property that the forecast that arises from Equation 1 approximates $f^*(z_T) = \arg\min_{f(\cdot)\in\mathcal{F}} E\,[L(f(z_T), y_{T+b})]$. This is the forecast that makes the expected loss, calculated at the present point in time, T, as small as possible among all forecasting methods in the set $\mathcal{F}$. Because we are interested in forecasting the future outcome, y_{T+b} given information up to time T, z_T, in calculating this expected loss, we would prefer the expectation to be taken over the random variables generated by y_{T+b} given z_T. To characterize the optimal forecast, denote the joint density of (y_{T+b}, z_T) as $p_{y_{T+b},z_T}(y, z|\theta)$, where $\theta \in \Theta$ are unknown parameters of the DGP, and note that $p_{y_{T+b},z_T}(y, z|\theta) = p_{y_{T+b}|z_T}(y|z_T, \theta)p_{z_T}(z|\theta)$.[2] The conditional expectation of the loss, given the data z_T, can then be written

$$E_{y_{T+b}|z_T}[L] = \int L(f(z_T), y)p_{y_{T+b}|z_T}(y|z_T, \theta)\mathrm{d}y. \tag{2}$$

The optimal forecast is the function $f(z_T) \in \mathcal{F}$ that minimizes Equation 2.

For some loss functions, a solution can be found for the optimal forecast, $f^*(z_T)$. As an example, consider the popular mean squared error (MSE) loss function

$$L(f(z_T), y_{T+b}) = (y_{T+b} - f(z_T))^2. \tag{3}$$

Under MSE loss, the optimal forecast is the conditional mean of y_{T+b}, i.e., $f^*(z_T) = E_{y_{T+b}|z_T}[y_{t+b}]$, which depends on both z_T and θ. Such "optimal" forecasts, though unique for any known DGP, are functions of the unknown parameters, θ, and thus will typically be estimated by functions of the observed data.

2.2. Risk for Forecasting Methods

More generally, because the estimates of the forecasting models are functions of the full data set, we might consider the unconditional expectation—or risk—of the estimated forecasting method computed as of the time the forecast is constructed, i.e.,

$$R(f, \theta) = \int L(f(z_T), y_{T+b})p_{y_{T+b}|z_T}(y|z_T, \theta)p_{z_T}(z|\theta)\mathrm{d}y\mathrm{d}z. \tag{4}$$

As defined in Equation 4, the risk is a function of the forecasting method, $f(\cdot)$, the parameters of the joint density for the data, θ, and the loss function, L. The best forecasting model can then be defined as the function $f(\cdot) \in \mathcal{F}$ that minimizes risk as defined in Equation 4.

Minimizing Equation 4 requires searching over a function space, $\mathcal{F}$. Less parametric approaches, such as those typically used for data mining such as sieves or kernel estimation, can be viewed as attempts to reduce risk by searching over a wide space of models, $\mathcal{F}$. Assuming that the underlying basis in sieve models has been judiciously chosen (e.g., the logistic function in a neural net), theoretical results show that by including sufficiently many terms, these models can approximate very general sets of functions arbitrarily well (see Hornik et al. 1989). In practice,

[2] We refer to this as the DGP; candidates for $f(z)$ are referred to as models.

of course, these models must themselves be approximated with a finite number of terms and estimated parameters, so there will be a trade-off between the costs to increasing risk through the estimation of many terms and the decline in risk from being better able to capture nonlinearities in the forecast model. These methods have not gained widespread traction in the economics profession, perhaps due to their somewhat black box approach to generating a forecast and their reputation for overfitting. However, economic constraints (e.g., monotonicity) can be used to constrain the search associated with these methods and rule out economically implausible estimates. Undoubtedly, these methods will receive further consideration in future work. Hastie et al. (2009) provide a terrific introduction to statistical learning methods, and Bai & Ng (2009) and Rossi & Timmermann (2015) provide applications to economic and financial forecasting problems.

At the other end of the spectrum are more parametric methods, which specify the set of models $\mathcal{F}$ to be known up to a finite-dimensional parameter β so that $\mathcal{F}$ is $f(z_T, \beta)$ for $\beta \in \mathcal{B}$. The parametric approach results in a much simpler search problem. Both nonparametric and parametric methods, as well as intermediate methods, are often used to construct forecasting models.

Regardless of the estimation approach, no single optimal forecasting method will be uniformly dominant even in the simplest of forecasting situations. To illustrate this, consider again the case with MSE loss, Equation 3, for which the optimal forecast is the conditional mean. When the DGP is known, the optimal forecast becomes a parameter of the conditional distribution of y_{T+b} given z_T (namely, the conditional mean). We know from the extensive literature on estimating conditional means that different estimation procedures yield risk functions whose rankings cross for different values of θ. Some estimation methods are therefore better for certain parts of the parameter space, Θ, whereas other estimation methods work better in other regions of Θ.

Several aspects of the construction of a forecast model beyond estimating the vector of unknown parameters of the forecasting model, $f(\cdot)$, such as model selection and model averaging can be considered as complicated parameter estimation methods. To see this, consider a linear model with two predictors, one of which is always included, whereas a model selection method is used to choose whether to include the second variable. The estimator for the parameter associated with the second predictor is simply $\hat{\theta}_2 = 1(g(z) \in G)\tilde{\theta}_2$, where $\tilde{\theta}_2$ is the least squares estimator when both variables are included and $1(g(z) \in G)$ is an indicator variable that equals 1 if the variable is selected and zero if it is excluded.

2.3. Risk for the Linear Regression Model

To allow us to be more specific about the broad issues discussed so far, consider a simple linear regression model with normally distributed innovations. Specifically, assume that the joint DGP for the predictors and the dependent variable takes the form

$$\begin{pmatrix} y_{t+b} \\ x_t \end{pmatrix} \sim \text{ind}N\left[\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \Sigma_{yx} \\ \Sigma_{yx}' & \Sigma_x \end{pmatrix}\right]. \tag{5}$$

The parameters of the model are $\theta = (\mu_y, \mu_x, \sigma_y^2, \Sigma_{yx}, \Sigma_x)$. Using the matrix

$$\begin{pmatrix} 1 & -\Sigma_{yx}\Sigma_x^{-1} \\ 0 & I_k \end{pmatrix}$$

to rotate the data in Equation 5, we can write

$$\begin{pmatrix} y_{t+b} - \Sigma_{yx}\Sigma_x^{-1}x_t \\ x_t \end{pmatrix} \sim \text{ind}N\left[\begin{pmatrix} \mu_y - \Sigma_{yx}\Sigma_x^{-1}\mu_x \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{yx}' & 0 \\ 0 & \Sigma_x \end{pmatrix}\right]. \tag{6}$$

Thus, it follows that $y_{t+b}|x_t \sim \text{ind} N(\mu_y - \Sigma_{yx}\Sigma_x^{-1}(x_t - \mu_x), \sigma_y^2 - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{yx}')$, and $x_t \sim \text{ind}$ $N(\mu_x, \Sigma_x)$. Moreover, the rotated random variables in Equation 6 are orthogonal. Note also that the conditional distribution of y_{t+b} given z_t is a function of only a subset of θ.

The expected loss conditional on the observed data, z_t, is given by

$$E_{y_{T+b}|z_T}[L(f(z_T), y_{T+b})] = \int L(f(z_T), y) p_{y_{T+b}|z_T}(y|z_t, \theta)\mathrm{d}y$$
$$= \int L(f(z_T), y) p_{y_{T+b}|x_T}(y|x_t, \theta)\mathrm{d}y. \quad (7)$$

Under MSE loss, this is given by

$$E_{y_{T+b}|z_T}[L(f(z_T), y_{T+b})] = \int (y - f(z_T))^2 p_{y_{T+b}|z_T}(y|z_t, \theta)\mathrm{d}y, \quad (8)$$

which is minimized at

$$f(z_T) = E[y_{T+b}|z_T] = E[y_{T+b}|x_T] = \beta'\tilde{x}_T, \quad (9)$$

where $\tilde{x}_T = [1, x_T]$ and $\beta = [\mu_y - \Sigma_{yx}\Sigma_x^{-1}\mu_x, -\Sigma_{yx}\Sigma_x^{-1}] = \beta(\theta)$. This expression clearly shows that the parameters of the optimal forecasting model, β, are a function of the parameters of the DGP, θ, as we would expect.

The forecast in Equation 9 is the unique optimal forecast. However, this forecast is also infeasible as it depends on the unknown population parameters, β. In practice, β is unknown, and we have to rely on an estimate of β to construct the forecast. There are multiple candidates for such an estimator. The most obvious one is the ordinary least squares (OLS) estimator for β, which depends on all of $z_T = \{y_t, x_t\}_{t=0}^T$, not just x_T as in the optimal forecast. Setting $h = 1$ and using $\hat{\beta}_{\text{OLS}} = (\sum_{t=0}^{T-1} x_t x_t')^{-1}(\sum_{t=0}^{T-1} x_t y_{t+1})$, we can construct the forecast

$$f(z_T) = \hat{\beta}_{\text{OLS}}' x_T = x_T' \left(\sum x_t x_t'\right)^{-1} \left(\sum x_t y_{t+1}\right), \quad (10)$$

where all summations in this section are $t = 0$ to $T - 1$. This is a function of all of z_T.

With the use of in-sample evaluation, the expected loss is

$$E\left[T^{-1}\sum\left(y_{t+1} - \hat{\beta}_{\text{OLS}}' x_t\right)^2\right] = E\left[T^{-1}\sum\left(\varepsilon_{t+1} - (\hat{\beta}_{\text{OLS}} - \beta)' x_t\right)^2\right] \quad (11)$$
$$= T^{-1}\sum E\varepsilon_{t+1}^2 - E\left[(\hat{\beta}_{\text{OLS}} - \beta)'\left(T^{-1}\sum x_t x_t'\right)(\hat{\beta}_T - \beta)\right]$$
$$= \sigma^2 - T^{-1}E\left[\left(\sum x_t\varepsilon_{t+1}\right)'\left(\sum x_t x_t'\right)^{-1}\left(\sum x_t\varepsilon_{t+1}\right)\right]$$
$$= \sigma^2(1 - T^{-1}k).$$

Note that adding more parameters (increasing k) improves the in-sample fit and thus reduces the expected loss.

This result can be contrasted with what happens for the case with out-of-sample evaluation. To this end, note that the risk associated with this linear OLS forecast is given by

$$E_{y_{T+b},z_T}[L(f(z_T), y_{T+b})] = \int L(\hat{\beta}_{\text{OLS}}' x, y) p_{y_{T+b}|z}(y|z, \theta) p_{z_T}(z|\theta)\mathrm{d}y\mathrm{d}z. \quad (12)$$

Under MSE loss, this simplifies to

$$E_{y_{T+b},z_t}[L(f(z_T), y_{T+b})] = \int (y - \hat{\beta}_{\text{OLS}}' x)^2 p_{y_{T+b}|z}(y|z, \theta) p_{z_T}(z|\theta)\mathrm{d}y\mathrm{d}z. \quad (13)$$

For the DGP in Equation 5, the MSE loss in Equation 13 simplifies to

$$E\left[(y_{T+b} - \hat{\beta}'_{\text{OLS}} x_T)^2\right] = E\left[\varepsilon^2_{T+b} - (\hat{\beta}_{\text{OLS}} - \beta)' x_T x'_T (\hat{\beta}_{\text{OLS}} - \beta)\right]. \tag{14}$$

Conditional on the observed data, $x = \{x_0, \ldots, x_T\}$, the out-of-sample MSE loss is

$$\begin{aligned} E\left[\varepsilon^2_{T+1} - (\hat{\beta}_{\text{OLS}} - \beta)' x_T x'_T (\hat{\beta}_{\text{OLS}} - \beta) | x\right] &= \sigma^2 + E(\hat{\beta}_{\text{OLS}} - \beta)' x_T x'_T (\hat{\beta}_{\text{OLS}} - \beta) | x \\ &= \sigma^2 + Etr\left[(\hat{\beta}_{\text{OLS}} - \beta)(\hat{\beta}_{\text{OLS}} - \beta)' x_T x'_T | x\right] \\ &= \sigma^2 + tr\left[\sigma^2 \left(\sum x_t x'_t\right)^{-1} x_T x'_T\right] \\ &= \sigma^2 \left(1 + x'_T \left(\sum x_t x'_t\right)^{-1} x_T\right). \end{aligned} \tag{15}$$

Using this expression, the unconditional out-of-sample MSE loss associated with the OLS forecast becomes

$$\begin{aligned} E\left[\sigma^2 \left(1 + x'_T \left(\sum x_t x'_t\right)^{-1} x_T\right)\right] &= \sigma^2 \left(1 + E\left[x'_T \left(\sum x_t x'_t\right)^{-1} x_T\right]\right) \\ &= \sigma^2 \left(1 + T^{-1} Etr\left[\left(T^{-1} \sum x_t x'_t - \Sigma + \Sigma\right)^{-1} x_T x'_T\right]\right) \\ &= \sigma^2 \left(1 + T^{-1} Etr\left[\left(\Sigma + O_p(T^{-1/2})\iota\iota'\right)^{-1} x_T x'_T\right]\right) \\ &= \sigma^2 \left(1 + T^{-1} Etr\left[\Sigma^{-1} x_T x'_T\right]\right. \\ &\quad \left. + T^{-1} Etr\left[0_p(T^{-1/2})\Sigma^{-1}\iota\iota'\Sigma^{-1-1} x_T x'_T\right]\right) \\ &\approx \sigma^2 \left(1 + T^{-1} tr\left[\Sigma_X^{-1} \Sigma_X\right]\right) \\ &= \sigma^2 \left(1 + T^{-1} tr(I_k)\right) \\ &= \sigma^2 \left(1 + T^{-1} k\right). \end{aligned} \tag{16}$$

In sharp contrast to the case with in-sample evaluation (Equation 11), the out-of-sample risk is seen to increase in k, the number of unknown parameters that have to be estimated. The second term, $\sigma^2 k/T$, is due to estimation error and will vanish in large samples. However, estimation error can be important in finite samples as we next demonstrate through a set of Monte Carlo simulations.

2.4. Monte Carlo Simulations

We illustrate the methods discussed up to this point through a simple Monte Carlo experiment that uses the linear forecasting model and MSE loss setting from Section 2.3. The experiment, further described in the Appendix, assumes that there are 10 potential predictors of the outcome variable. The predictors are mutually correlated, joint normally distributed random variables, and independent over time. The outcome is also assumed to be normally and independently distributed over time, and it depends on either $k = 3$ or $k = 6$ nonzero coefficients for the 10 predictors, so that $10 - k$ predictors enter with a zero coefficient and thus are irrelevant.

A single parameter, α, varies the strength of the predictors over the outcome; this is best understood in terms of a hypothesis test of whether all 10 predictors, as a group, have predictive power over the outcome, i.e., a test that all coefficients, apart from the constant, are zero. When $k = 3$ for $\alpha < 2.5$, the power of this test is less than 80%. For values of $\alpha > 3$, the power is nearly one. In this case, the parameters are nonlocal, and it is statistically much easier to distinguish nonzero parameters from those that are equal to zero.

We consider forecasts generated by the following methods: (*a*) OLS using all possible predictors, (*b*) the complete subset regression (CSR) approach to forecast combination with $k = 3$ predictors included,[3] (*c*) a weighted average of forecasts computed over all possible estimators using weights based on the Akaike information criterion (AIC), (*d*) the single best model chosen by the AIC, and (*e*) the least absolute shrinkage and selection operator (LASSO) and post-LASSO methods.[4]

Figure 1 plots the risk functions for these methods, i.e., the MSE values as a function of α, which tracks the predictive power of the predictors with nonzero coefficients. **Figure 1*a*,*b*** assumes $k = 3$ predictors with nonzero coefficients, whereas **Figure 1*c*,*d*** assumes $k = 6$ nonzero coefficients. Thus, more predictors are relevant in the bottom panels than in the top panels, where, conversely, predictive ability is more sparse. The strength of the nonzero predictors is varied across the left and right columns of the figure; the left column represents the case with weak (local) predictors, whereas the right column assumes strong predictors.[5] For each case, we use 100 observations to estimate the forecasting models.

Figure 1 clearly demonstrates that there is no uniformly dominant forecasting method. Methods such as the LASSO and CSR, which shrink the estimated coefficients, are preferred for small values of the coefficients (i.e., for small values of α). However, for larger values of the coefficients, the biases in these methods kick in, and their MSE rises. Still, these methods work well for most of the range of values for which the predictors are sufficiently weak that it is not statistically obvious that they should be included (**Figure 1*a*,*c***).

In situations in which some predictors are clearly useful, methods that attempt to determine which variables to include perform better than shrinkage methods such as the LASSO and CSR.[6] However, for intermediate values of α for which pretests for inclusion of the relevant predictors are neither quite weak nor have power near one to identify the relevant predictors, there is a hump in the MSE of such pretest methods as well as in the MSE of the model selection method. When $k = 3$, the pretest or model selection methods are strictly preferred to the OLS kitchen sink forecasts based on inclusion of all 10 predictors despite OLS having nice theoretical properties (i.e., being the minimum variance unbiased, minimax method). This result does not hold uniformly when $k = 6$, and a larger fraction of the predictors is relevant from a forecasting perspective.

The difference between in- and out-of-sample risk is the reason why model selection is not a straightforward problem to solve. Many methods attempt to deal with this issue: The AIC adjusts the in-sample loss estimate to get closer to out-of-sample loss, whereas cross-validation attempts to estimate the out-of-sample loss directly. Other methods such as the BIC (Bayes information criterion) and LASSO take different approaches based on the belief that a more parsimonious model is likely to be preferred.

3. IMPORTANCE OF THE CHOICE OF LOSS FUNCTION

Ideally, the loss function used to construct a forecasting model should be tailored to the economic costs of over- or underpredicting the outcome and thus requires a good understanding of the

[3]This approach is further described in Section 6.

[4]The post-LASSO method uses the LASSO method to choose which variables to include in the regression, and then re-estimates the model using OLS with the selected variables.

[5]For the experiments with nonlocal parameters (**Figure 1*b*,*d***), we remove the CSR and LASSO methods, which are biased and hence have larger risk.

[6]We exclude these methods from **Figure 1*b*,*d*** because their MSE becomes very large.

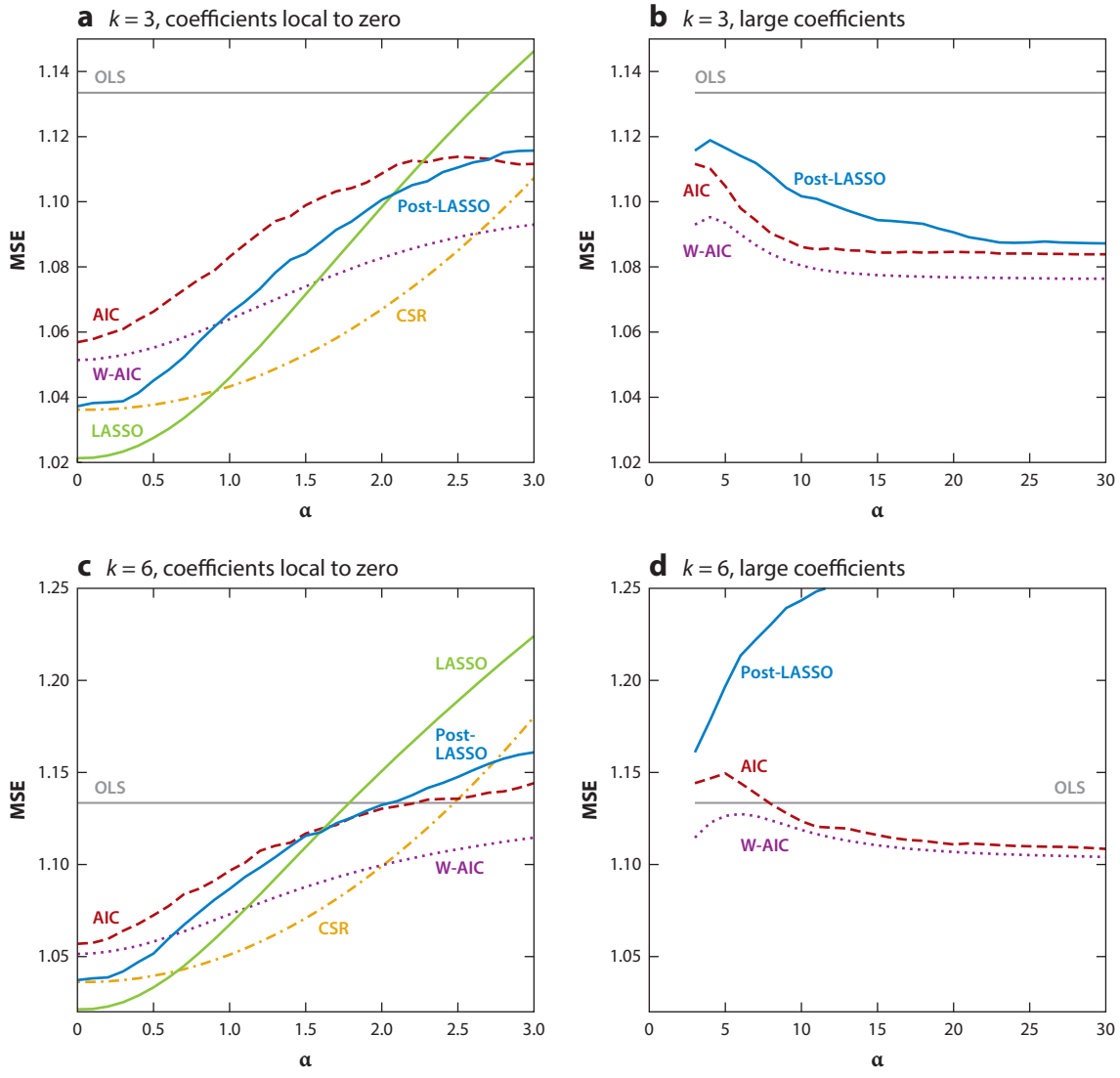

Figure 1

Risk functions for different forecast methods: ordinary least squares (OLS) using all possible predictors (*horizontal gray line*), the complete subset regression (CSR) approach to forecast combination with $k = 3$ predictors included (*yellow dashed-dotted line*), a weighted average of forecasts computed over all possible estimators using weights based on the Akaike information criterion (W-AIC; *purple dotted line*), the single best model chosen by the Akaike information criterion (AIC; *red dashed line*), and the least absolute shrinkage and selection operator (LASSO) (*green solid line*) and post-LASSO (*blue solid line*) methods. Other abbreviation: MSE, mean squared error.

economics underlying the forecasting situation. We next provide two examples of issues that may be relevant for the loss function used by the forecaster.

For a monetary policy maker, the loss function could involve not only the magnitude of the forecast error (i.e., the difference between the outcome and the forecast), but also the level of the outcome itself, or other state variables, as the cost of erring varies with the underlying economic

state. In their study of the Federal Reserve's Greenbook forecasts, Patton & Timmermann (2007) find that the Federal Reserve tends to provide conservative forecasts of economic growth (which is consistent with a loss function that penalizes underpredictions less than it penalizes overpredictions). However, they also find that this matters particularly in states with low or moderate economic growth. Overpredictions of economic growth in states with low growth could be particularly costly because the Federal Reserve would fail to apply appropriately aggressive monetary policy measures in such circumstances.

As a second example, for a company engaged in sales forecasting, there may be important technological and economic constraints related to the possibility (and cost) of expanding production or managing inventories. The effect on customer behavior in case the firm produces too little and is unable to supply enough products is another concern. Engineering data along with data on customer behavior would therefore appear to be important for constructing the firm's loss function.

As suggested by these examples, the derivation of a loss function is in many regards analogous to the elicitation of priors in Bayesian analysis. For example, one can imagine controlling the relative costs of small and large forecast errors through a single parameter and controlling the degree of asymmetry through another parameter. This approach is followed in the construction of the EKT error loss function suggested by Elliott et al. (2005):

$$L(e_{t+1|t}) = \left[\alpha + (1-2\alpha)1(e_{t+1|t} < 0)\right]|e_{t+1|t}|^p. \tag{17}$$

Here $e_{t+1|t} = y_{t+1} - f_{t+1|t}$ is the one-step-ahead forecast error at time $t+1$. The parameter α controls the degree of asymmetry, with $\alpha = 1/2$ representing the symmetric case. The parameter p, which Elliott et al. (2005) set to 1 or 2, controls whether the cost of forecast errors grows linearly or quadratically in $|e|$. The EKT loss nests popular loss functions as special cases: $p = 2$ and $\alpha = 1/2$ yield MSE loss, whereas $p = 1$ and $\alpha = 1/2$ result in mean absolute error loss.

Financial forecasting is one area where loss functions have proven relatively straightforward to motivate and construct. A single investor is often assumed to have mean variance or power utility over final wealth. Provided that a mapping from forecasts to portfolio weights can be established, the loss function is fully specified.

Example 1 (risk-averse investor's choice of stock position). Consider the single-period portfolio decision of an investor who can invest in risk-free bonds, with a guaranteed payoff of zero, or in risky stocks, with a future payoff of y_{t+1}. Let the investor's portfolio allocation to stocks be ω_t and assume that the investor's initial wealth at time t is $W_t = 1$. The investor's wealth at time $t+1$ is then given by $W_{t+1} = \omega_t y_{t+1}$. Under mean-variance preferences, the investor's objective is to maximize expected utility given current information, Z_t:

$$E[U(W_{t+1})|Z_t] = E[W_{t+1}|Z_t] - \frac{a}{2}Var(W_{t+1}|Z_t). \tag{18}$$

Here a reflects the investor's risk aversion, and $E[W_{t+1}|Z_t]$, $Var(W_{t+1}|Z_t)$ is the conditional mean and conditional variance of W_{t+1}, given current information, Z_t. Both these moments will generally depend on ω_t, so the forecasting problem involves modeling both these conditional moments. For the simple forecasting model

$$y_{t+1} = \mu + z_t + \varepsilon_{t+1},\ \varepsilon_{t+1} \sim (0, \sigma_\varepsilon^2),$$

where z_t is a variable observed by the investor at time t, the investor's first-order condition yields the optimal stock holding as

$$\omega_t^*(z_t) = \frac{\mu + z_t}{a\sigma_\varepsilon^2}. \tag{19}$$

The loss function is important for the estimation of the parameters of the forecasting model. Most empirical forecasting studies rely on standard loss functions that are common to estimation, such as MSE or mean absolute error loss. This approach may be appropriate in some situations. For example, in cases with a linear forecasting model and data that are jointly normal in the outcome and predictor variables, under MSE loss the forecasting model minimizes the equivalent of the negative of the maximum likelihood estimator (MLE). Assuming that the model is known and the variables are joint normally distributed, the MLE is an efficient estimator of the model parameters, regardless of the loss function. One can then proceed by simply plugging the maximum likelihood parameter estimates into the optimal forecast for the problem that still involves the correct loss function.

The assumptions of a known DGP and joint normally distributed variables are unlikely to hold for many empirical applications. This has important implications for how we approach the forecasting problem. If we do not know the correct DGP, maximum likelihood estimation might not be efficient, and it may be better to use the loss function to estimate the parameters of the model. Access to longer data sets would tend to make this point more important because the increased efficiency of the parameter estimates in a longer sample means that differences between maximum likelihood and loss-based estimators can grow larger.

The loss function also matters for how we evaluate a sequence of forecasts.

Example 2. Under MSE loss, the one-step-ahead forecast error $e_{t+1|t} = y_{t+1} - f_{t+1|t}$ should be orthogonal to all information in the forecaster's information set, a condition that can be tested through the orthogonality regression

$$e_{t+1|t} = v_t\delta + u_{t+1}, \tag{20}$$

where $v_t \in Z_t$ is any variable known to the forecaster at time t, and u_{t+1} is unpredictable conditional on Z_t: $E[u_{t+1}|Z_t] = 0$. Under MSE loss, no unknown parameters of the loss function enter into the forecaster's first-order condition $E[e_{t+1|t}|v_t] = 0$, so forecast efficiency implies testing that $\delta = 0$, or, more generally, by testing for absence of predictability of the conditional mean of the forecast error, $e_{t+1|t}$. Of course, such tests remain tests of the joint null of MSE loss and efficient use of information by the forecaster. Rejections of the null could arise because either of these conditions fails.

When the loss function depends on unknown parameters that have to be estimated, the joint hypothesis testing problem becomes clearer because any test of forecast efficiency now depends not only on the loss function belonging to a certain family, but also on the parameter estimates of this loss function. To see this, consider the case with so-called quad-quad loss that arises as a special case of Equation 17 with $\alpha \in (0, 1)$ and $p = 2$. Elliott et al. (2008) show that the null of forecast efficiency can be tested through a modified orthogonality regression

$$e_{t+1|t} = (1 - 2\alpha)|e_{t+1|t}| + v_t\delta + u_{t+1}. \tag{21}$$

Comparing Equations 20 and 21, it follows that under asymmetric loss ($\alpha \neq 0.5$), the standard orthogonality regression that assumes MSE loss (Equation 20) will suffer from an omitted variable bias. This happens because of the absence of the term $(1 - 2\alpha)|e_{t+1|t}|$, which picks up the omitted variable bias in the forecast error resulting from asymmetric loss. In practice, α is unknown, and any test of $\delta = 0$ in Equation 21 will depend on the estimated value for the asymmetry parameter, α.

Finally, we provide a note of caution. Even if the forecaster's loss function is well understood, forecasts are often used as inputs in complex decisions involving multiple outcomes and nonlinear effects, examples of which include value-at-risk calculations and stress testing of banks. In such situations, it can be difficult to formalize the exact decision rule for mapping forecasts to actions, and this introduces a tension between inference about predictive accuracy and decision making.

4. USING ECONOMIC THEORY TO COMPUTE FORECASTS

Economic theory has the potential to play an important role in many steps in the economic forecasting problem. At its most primitive, setting up the loss function requires understanding the basic trade-offs between forecast errors of different signs and magnitudes and thus presupposes a clear understanding of the economics underlying the forecasting problem. The risk measure in Equation 4 depends directly on this understanding. Economic theory also can be used to limit the size of $\mathcal{F}$, the set of models to consider, in three ways. First, theory can provide guidance on which variables to consider (model selection). Second, economic theory might be suggestive of the functional form used by the forecasting model (e.g., linear versus nonlinear). Third, economic theory can be used to impose constraints on the parameters or moments of the forecasting model (parametric restrictions). Such considerations tend to reduce the space of models $\mathcal{F}$ considered in the search. Alternatively, if a structural model is employed to construct the forecasts, the model itself arises as a result of economic theory. We discuss these points further below.

Economic theory may be suggestive of which variables to include. In particular, equilibrium theory may suggest that factors that determine households' and firms' first-order conditions should influence economic dynamics. Theory may also help identify the nature of the shocks (e.g., technology or preference shocks or shocks to government activity) that will most affect economic growth. Although an understanding of the source of the economic shocks may not be helpful in producing point forecasts, this information can potentially be useful when generating probability forecasts and evaluating risks.

Economic theory is perhaps less likely to provide specific predictions about which functional forms to use in the prediction model. Although economic theory frequently assumes linearity or relies on linearized models, this is more a matter of tractability rather than a direct result of economic restrictions or properties of the assumed technology or preferences. However, monotonicity or sign constraints on the coefficients of the forecasting model may prove helpful. For example, one may want to impose that the coefficient on valuation ratios such as the dividend-price ratio or the book-to-market ratio is positive in a forecasting model for stock market returns (see Campbell & Thompson 2008).

Pettenuzzo et al. (2014) develop a Bayesian approach that constrains the forecasts of stock market returns through restrictions either on the conditional expected excess return on stocks (over T-bills)—which has to be positive to ensure that investors are willing to hold risky stocks—or on the conditional expected excess return per unit of risk, which can be constrained by the "price of risk" in the economy. Empirically, they find that imposing such restrictions on the forecasting models leads to improvements in forecasts of US stock market returns.

Structural economic models can be used directly either to construct forecasts or to provide restrictions on forecasts. Larger-scale structural models are employed for a variety of purposes. Examples include the FRB/US model used by the Federal Reserve and Moody's Analytics macroeconomic model. Such models have the ability to construct counterfactual predictions as well as forecasts.

Dynamic stochastic general equilibrium (DSGE) models form a class of models that have proven particularly popular among central banks and economic researchers seeking stronger economic foundations for their forecasting models. These models incorporate intertemporally optimizing household and firm decisions and combine them with monetary and fiscal policy decision rules along with assumptions about shocks to technology and monetary and fiscal policy.[7] By

[7] Smets & Wouters (2003) provide a particularly influential study in this literature (see also the recent survey in Del Negro & Schorfheide 2013).

leaning heavily on economic theory, DSGE models ensure that a set of internally consistent model forecasts are being generated. Moreover, these models are well suited for understanding the economic reasons for specific forecasts. They can also be used to analyze the effect on predictions of changes to economic policy; this helps explain their popularity among central banks.

DSGE models make simplifying assumptions in their representation of the economic dynamics, and inevitably this can adversely affect the resulting forecasts. Some economic constraints are also more likely to hold in the long run and may not be helpful in predicting short-run dynamics. Indeed, several empirical studies have found that DSGE forecasts frequently get beaten by survey forecasts or even by simpler time-series models. Del Negro & Schorfheide (2013, p. 86) summarize the literature as follows: "The empirical evidence in the literature suggests that DSGE model forecasts are comparable to standard autoregressive or vector autoregressive models but can be dominated by more sophisticated univariate or multivariate time series models."

5. OPPORTUNITIES, LIMITATIONS AND CHALLENGES POSED BY BIG DATA

Improved access to vast data structures—often called big data—has facilitated new lines of analysis in many areas of economics, such as applied microeconomics, high-frequency finance, and studies of the impact of news events such as monetary policy announcements. A key question is how access to big data is likely to affect practices in economic forecasting.

Most forecasting problems involve two dimensions, namely a time-series dimension (T) and a cross-sectional dimension (K), where the former refers to the number of observations available for estimation (i.e., the sample length), and the latter typically refers to the number of potential predictors.

5.1. Big(ger) T

From a forecasting perspective, a bigger T offers the hope for more precise model estimates and improvements in our ability to choose between competing forecasting models. In practice, however, two factors tend to limit the benefits from a bigger T in many economic applications. First, often larger samples mean sampling the data at a higher frequency. Second, longer data sets often mean that estimates are constructed from data far in the past and of less relevance for forecasting today. This problem of model instability is discussed in Section 7. We next discuss these issues in turn.

Access to more frequently observed data—sometimes observed tick by tick—has allowed us to estimate models for high-frequency movements in variables such as stock prices, currencies, and interest rates and has facilitated important progress in our understanding of high-frequency movements in financial markets. The literature on high-frequency estimation and prediction of financial market volatility has made important progress over the past 15 years. We note that market microstructure effects and limits to liquidity for individual assets mean that there are also limits on how far we can push the sampling frequency. For example, sampling much more frequently than every five minutes may, for many assets, introduce more noise in the observed prices than is desirable from an estimation point of view. Still, such high-frequency data have enabled us to compute real-time (daily) estimates and forecasts of the risk of individual assets as well as portfolios and, it is fair to say, make up one of the great success stories of recent years (for a survey of volatility forecasting, see Andersen et al. 2006).

It is less obvious that access to data measured at, e.g., the daily frequency will help us produce more accurate forecasts of macroeconomic aggregates such as inflation or the unemployment rate.

Much of the variation in economic variables is related to the economic cycle, and an increased frequency of observation does not increase the number of cycles we observe and hence could have limited impact on our ability to produce better forecasting models.

Although this point shows that there are limits to how useful high-frequency data can be to our ability to predict macroeconomic variables over fixed horizons such as one month or one quarter ahead in time, other arguments give rise to more optimism. First, many macroeconomic variables (e.g., GDP) are measured with considerable error and get revised over time, so it is possible that financial variables, established in fluid and transparent markets, contain information that complements the information already contained in the macroeconomic variables of interest.[8] Second, financial variables are forward looking and so, although driven also by factors other than forecasts of macroeconomic fundamentals, might contain information that is helpful for forecasting the latter.

Exploring such questions, a recent literature on estimation with mixed data sampling (MIDAS) addresses whether variables observed at a high frequency such as daily interest rates or daily stock price movements can be used to generate improved forecasts of less-often observed variables such as quarterly GDP or monthly inflation (see Andreou et al. 2011). Andreou et al. (2013) use a rich set of daily variables and find that it can be used to generate some improvements in quarterly GDP forecasts. Pettenuzzo et al. (2015) add variables observed at the daily frequency to a specification for dynamics in the volatility equation for inflation and industrial production growth. They find that the addition of interest rate variables, a daily business cycle proxy, and information from stock markets can be used to generate improved out-of-sample density forecasts, whereas the benefits from using such information to improve the accuracy of point forecasts tend to be weaker.

With more frequently observed data sets also come new challenges such as irregularly sampled variables observed at different frequencies (i.e., daily stock prices, weekly payroll figures, monthly unemployment rates, and quarterly GDP growth). The Kalman filter can be used to handle many of the practical issues. One particularly interesting application is the construction of a daily measure of the business cycle. Building on work by Aruoba et al. (2009) that uses a Kalman filter to estimate a common factor model, the Federal Reserve Bank of Philadelphia now provides daily updates to the estimated business cycle measure, which is treated as a latent process that is correlated with observable variables. A related literature on nowcasting uses the same notion of "jagged edge" data (Banbura et al. 2011) sampled at different frequencies to produce estimates of the current (unobserved) state, which itself follows some process that can be used to predict the next period's state and so may be useful for forecasting purposes.

The other way to get a bigger T is by getting a longer time series of data. However, problems of model instability become more apparent in longer samples. Model instability, or heterogeneity in the DGP, can cause problems for our ability to find a good forecasting model. Even for mild unmodeled heterogeneity, we do not have a limit result that suggests that Equation 1 converges to Equation 4, as we would like when making a forecast at time T. Instead, we would have a result that Equation 1 converges to the average risk over the sample period. Given sufficiently strong heterogeneity, this might be problematic. For extreme levels of variability in the DGP, such as large, discrete breaks in the DGP, past data can be of limited use in constructing useful

[8]Revisions to data observed at different points in time pose additional challenges to macroeconomic forecasting. Macroeconomic series such as GDP are subject to revisions over time as more accurate estimates become available or new methodologies (e.g., weighting schemes) get introduced. This means that care has to be exercised with regard to which "vintage" of a particular variable is being modeled and predicted and even with regard to modeling the joint process for different vintages (see Croushore & Stark 2001 for further discussion of this point).

forecasting models to be used at the end of the sample (time T) unless these breaking processes can be modeled.

5.2. Bigger K

A bigger K (i.e., more potential predictors) is the big data effect most consistent with what we observe in practice in much of economic forecasting. Larger sets of predictors offer both opportunities and challenges. On the one hand, it is possible that some of the new predictors have genuine predictive power over the outcome, in which case these variables might be included in the forecasting model. On the other hand, a larger K increases the dimension of the set of predictors and hence the dimension of the set of potential forecasting models, $\mathcal{F}$, making it more difficult to identify the best forecasting model.

Dealing with the increased dimension of predictors requires strategies to reduce the effect of estimation error that arises from attempting to incorporate more variables in the model. One approach is to use a few, judiciously chosen, linear combinations of the predictors in the forecasting model. The most prominent approach, which has generated a large literature over the past 20 years, is to use dynamic factor models. Dynamic factor models summarize the information from a potentially vast array of predictors through the first few principal components, F_t, of the individual predictors x_{it}, $i = 1, \ldots, K$:

$$x_{it} = \Lambda_i F_t, \quad i = 1, \ldots, K. \tag{22}$$

Here $F_t = (F_{1t}, \ldots, F_{qt})'$ is a set of q common factors, with q much smaller than the original value of K, which can run in the hundreds—typically q is of the order of five or ten—and Λ_i are the factor loadings, which are assumed to be constant over time. Estimates of the common factors, along with lagged values of the dependent variable, are then used as conditioning information in the forecasting model:

$$y_t = \alpha + \sum_{i=1}^{n_F} \beta_i' L^i F_t + \sum_{i=1}^{n_y} \gamma_i L^i y_t + \varepsilon_t, \tag{23}$$

where n_F and n_y are the number of lags of the common factors and the dependent variable. This approach has proven very popular in practice. Important progress has been made in terms of understanding the theoretical properties of forecasting models that include estimated common factors (see Stock & Watson 2006 for a summary) and their ability to deal with certain forms of instability (Stock & Watson 2002).[9]

Another strand in the forecasting literature assumes that only a small subset of the predictors $x_{it} \in X_t$ truly enter into the prediction model (this assumption is known as sparseness). Assuming MSE loss and certain sparseness conditions, estimation of the forecasting model proceeds by minimizing the penalized loss function

$$T^{-1} \sum_{t=0}^{T-1} (y_{t+1} - \beta' x_t)^2 + \lambda \sum_{i=1}^{n_k} |\beta_i|, \tag{24}$$

where $\lambda > 0$ determines the penalty from inclusion of additional variables in the forecasting model. The combination of a squared loss function with an absolute value penalty term yields a

[9]Other approaches for constructing aggregate summary measures that can be used as predictors have been proposed. D'Amuri & Marcucci (2012) use Google searches to construct an index of Internet job search intensity, which they use to predict monthly unemployment in the United States. They find that models that add this search index variable perform better in out-of-sample forecasts of future unemployment than models that exclude this variable.

solution with the property that many of the parameter estimates for β_i are set to zero, making this an estimation procedure and a model selection procedure in one. Efficient algorithms have been developed to find the important predictor variables (see Tibshirani 1996). These are known as LASSO regressions and can handle situations in which $K > T$. More recent methods such as the elastic net of Zou & Hastie (2005) adjust the metric for the penalty, and there is also work on how to choose the penalty coefficient λ (see Belloni & Chernozhukov 2011). Chudik et al. (2016b) discuss the choice of penalty function and propose a new approach—multiple testing with one covariate at a time?—for dealing with the multiple testing problem involved in model selection.

The objective in Equation 24 assumes squared error loss and penalizes the estimates in a way that is consistent with an expectation that many of the predictors have zero coefficients. One might consider using a similar approach for different loss functions, noting that the forecaster's loss function should dictate the choice of penalty function in such cases.

The jury is still out on which of these approaches—dynamic factor models versus LASSO sparseness regressions—works best, or if they should even be viewed as alternatives. Results from LASSO estimation are not invariant to adding linear combinations of the original K variables, so one approach is to simply add the principal components and see if they get included by the LASSO algorithm.[10] An alternative strategy is to not treat these methods as competing alternatives, but instead try to combine the individual forecasts. This leads to the question of model combination, which is covered in Section 6.

Other issues arise when interest lies in forecasting a possibly large-dimensional vector of dependent variables. Standard vector autoregression (VAR) methods are not well suited for this, but Bayesian methods have been developed to handle large-dimensional VARs. Another interesting approach that can be used to ensure coherence between individual forecasts is the global VAR model proposed by Pesaran et al. (2004). This approach has been used with success in empirical forecasting (see, e.g., Chudik et al. 2016a for a recent application to output prediction). Forecasting with dynamic panels that allow for cross-sectional dependencies is another promising way to go, although such methods are only in their infancy at the present time.

6. MODEL SELECTION VERSUS FORECAST COMBINATION

If the DGP were known, in most forecasting situations we could obtain efficient estimates of θ by maximum likelihood and use these to construct a forecasting model. However, a premise that is broadly accepted in forecasting analysis is that all forecasting models are misspecified. Such model misspecification arises for a number of reasons: (*a*) The underlying (joint) DGP for the outcome of interest and the predictor variables often undergoes change and so is difficult to track accurately through time; (*b*) the functional form of the mapping from predictors to the outcome is unknown and difficult to pin down; (*c*) the identity of the best predictors is unknown and subject to a complex search; and (*d*) the parameters of the forecasting model are estimated with error. Importantly, model misspecification is not easily sidestepped as the complexity of the problem of selecting and estimating a forecasting model can only be expected to grow over time with the arrival of new information.

Concerns such as these suggest that we might not attempt to identify and use a single "best" forecasting model. In fact, even if we knew the predictors and functional form of the true forecasting model, the presence of parameter estimation error means that simpler, smaller models with fewer

[10] In a recent application to inflation forecasting, Medeiros & Mendes (2015) find that the principal components do not necessarily get selected by the LASSO approach when added to a larger set of individual predictor variables.

parameters to estimate might produce better forecasts in finite samples. Similarly, the effect of individual predictor variables might be sufficiently small ("local to zero") that the additional signal value gained from their inclusion in a particular forecasting model is outweighed by the effect of parameter estimation error.

These are reasons why it is often unlikely that a single forecasting model always dominates all other alternative specifications. In such situations, a viable alternative is to use forecast combination or model combination. Forecast combination treats a set of underlying forecasts, $f_1, \ldots, f_n$, as any other data available for generating a consolidated forecast, $f^c(f_1, \ldots, f_n, z)$. Linear forecasting schemes of the form $f^c = \omega_0 + \omega_1 f_1 + \omega_2 f_2 + \cdots + \omega_n f_n$ are particularly popular. Often the weights $\omega = (\omega_0, \omega_1, \ldots, \omega_n)$ are estimated by OLS or obtained by some weighting scheme such as using the inverse of the MSE of the forecasts relative to the average MSE. Weighting schemes such as these introduce estimation error into the combined forecast.[11]

One alternative is to use equal weights on the forecasts (i.e., setting $\omega_0 = 0$ and $\omega_1 = \ldots = \omega_n = 1/n$) (for a survey of forecast combination, see Timmermann 2006). Empirically, it has proven surprisingly difficult to come up with forecast combination schemes that perform better than simple equal-weighted averages (see Genre et al. 2013 for a comparison of different combination schemes for survey forecasts of a range of macroeconomic variables). Undoubtedly, this reflects that clearly poor models are usually trimmed from the set of models being combined. However, it may also reflect that many forecasting methods lead to forecasts with similar error variances and similar covariances, perhaps due to the presence of a large, common unexplained component in the forecast error. In this case, there is little scope for improvements by deviating from the simple equal-weighted average of forecasts.

In settings with many potential predictors, each of which has a "small" (local to zero) effect on the future outcome, the trade-off between omitted variable bias—resulting from the exclusion of potential predictor variables—and estimation error—due to the inclusion of additional weak predictors—takes a particularly interesting form. Elliott et al. (2013) show that in such settings it can be favorable to combine over forecasts generated by models that include only a small, k, fixed set of predictors such as five or ten variables. In a setting with K possible predictors, there are $n_{k,K} = K!/(k!(K-k)!)$ different k-variate models. For example, there will be as many different models that include one predictor as models that exclude one predictor from the model, and thus include $K-1$ variables. However, when K is large relative to T, the (kitchen sink) model that includes all, or almost all, possible predictors tends to produce very poor out-of-sample forecasts due to the effect of estimation error.

In empirical work for stock returns (Elliott et al. 2013) and inflation rate, GDP growth, and unemployment rate forecasts (Elliott et al. 2015), Elliott et al. find that the CSR forecasts perform very well compared to univariate ARIMA (autoregressive integrated moving average) or common factor models. Equal-weighted combinations appear to work as well as alternative combination schemes that estimate the weights of the individual forecasting models and do not sample the models at random if the total number of models is too large to allow all possible models with k predictors to be included in the combination.

Forecast combination can be used to incorporate insights from very different forecasting methods. DSGE-based forecasts (described in Section 4), data-based machine learning methods, and forecasts from surveys represent very different approaches to modeling and incorporating very

[11] In situations with large sets of prediction models, attempting to estimate the combination weights becomes even more difficult although in principle one could use methods such as LASSO to select a smaller subset of forecasts and then estimate the weights on this subset.

different information sources. Given such differences, it would seem like a good idea to combine different approaches. This has been done. For example, Wright (2013) proposes a "democratic prior" approach that incorporates information from long-run Blue-Chip survey forecasts into the prior of a VAR. This anchoring of the priors seems to lead to improved forecasting performance over simply using a Bayesian VAR.

Model combination methods have been used to generate density forecasts, particularly in the Bayesian literature. Bayesian model averaging methods have been in use for some time. These utilize a linear combination scheme based on the individual models' marginal likelihood. Let $p(y|M_i, Z)$ be the predictive density for y_i given model M_i and the data, Z, with $p(M_i|Z)$ the posterior probability for model M_i given the data, Z. Then the Bayesian model average is a weighted average of the individual models' densities:

$$p^{\text{BMA}}(y|Z) = \sum_{i=1}^{m} p(y|M_i, Z)p(M_i|Z). \tag{25}$$

The combination weights in this equation sum to 1 but do not account for any correlations between models. An alternative weighting scheme is proposed by Geweke & Amisano (2011), who suggest choosing combination weights $\omega_t = (\omega_{1t}, \ldots, \omega_{mt})'$ to maximize the weighted log score (LS)

$$\omega_t = \arg\max \sum_{\tau=1}^{t-1} \log\left(\sum_{i=1}^{m} \omega_{it} \exp(\text{LS}_{\tau+1,i})\right) \text{ s.t.} \sum_{i=1}^{m} \omega_{it} = 1, \omega_{it} \geq 0 \text{ for } i = 1, \ldots, m.$$

Geweke & Amisano point out that this weighting scheme does not require that the true model is included in the set of models under consideration—an assumption that is of course unlikely to hold. In an empirical application to density forecasting for stock returns, they find that a variety of models get nonzero weights, suggesting that the approach genuinely diversifies across different model specifications.

Empirically, model and forecast combination has proven to be one of the few forecasting methods with a broad ability to improve forecasting performance across a large range of economic and financial variables. Its appeal derives from the existence of many forecasts with broadly similar forecasting performance—in which case a "portfolio" of these offers diversification gains. Moreover, in situations in which the relative performance of various forecasting models changes over time, combination approaches that use estimated weights can allow the weights to shift, better emphasizing models with improved performance at the cost of models whose performance is deteriorating, a point emphasized by Pettenuzzo & Timmermann (2016). Such model instability is discussed directly below.

7. DEALING WITH MODEL INSTABILITY

Model instability refers to the situation in which the DGP, $p_{y_{T+h},z_T}(y, z|\theta)$, varies over time. This could result from either the density changing or the parameters of the forecasting model changing over time, or both. Constant changes in the makeup of the economy, with new industries replacing old, and changes in regulations mean that it may not be reasonable to assume that the DGP is constant for very long periods of time.

The obvious problem that arises from model instability is that estimates of the forecasting model from past data are not necessarily useful for forecasting today. Although sample analogs of risk such as Equation 1 still may converge to the average risk (Equation 4), the latter object might not have any meaningful interpretation at a single value for θ when θ is varying over time.

Similarly, the estimate for β in a forecasting model might be good on average over the full data sample, but not at time T where we make our forecast.

Model instability appears to be pervasive. Stock & Watson (1996) find that instability was present for the majority of time-series forecasting models fitted to a range of macroeconomic variables. Clements & Hendry (1998) rate model instability as a key influence on the performance of macroeconomic forecasting models.

In the presence of such model instability, a number of practical issues arise. First, can we detect the presence—and, perhaps, timing and nature—of such model instability? Second, how should forecasts be constructed when there is time variation in the DGP?

To begin, consider our ability to detect parameter instability, which is the topic of a large literature. Although many of the tests are designed for particular models of instability (e.g., the standard SupF test for a single break, covered in Andrews 1993), typically these tests have power against a wide variety of models for the time variation in the parameters. For a wide variety of models for the break process, Elliott & Müller (2006) show that optimal tests for breaks have similar power against different forms of model instability. It follows that a test for breaks against a particular model is not indicative of that model being the correct representation of the DGP. On a more positive note, many break tests are useful in detecting whether there is model instability in the first place—they just cannot identify the exact form of this instability.

When model instability is suspected (or tests reject the null of stability), there are two main strategies for building a forecasting model. First, we could attempt to model the instability parametrically. The chief challenge to the parametric approach comes from choosing which model of instability is likely to be the correct one. The majority of the work in forecasting has operated under the assumption that the correct model has been chosen, rather than addressing how to choose the correct model. Second, we could use more robust procedures for generating a forecast, which do not require taking a stand on the form of the instability.

Let us consider parametric modeling approaches to instability. The difficulty here is that there exists a large set of candidate models for the instability. For example, the parameters might change values at some unknown point in time in a one-off change, or they might vary each period. Between these possibilities are models that shift less frequently but by larger (discrete) amounts.

Specifically, the parameters could undergo changes every period and follow a mean-reverting process,

$$\begin{aligned} y_{t+1} &= \beta_t x_t + \varepsilon_{t+1}, \\ \beta_t &= \bar{\beta} + \kappa(\bar{\beta} - \beta_{t-1}) + u_t, \end{aligned} \tag{26}$$

where $\kappa \geq 0$ is the speed of mean reversion, and $\kappa = 0$ corresponds to a random walk model (see Engle & Watson 1985 for an analysis of this type of model).

Alternatively, the parameters could change less frequently but in a more discrete manner as captured by a regime switching process,

$$\begin{aligned} \beta_t &= \beta_{s_t}, \\ \Pr(s_{t+1} = j | s_t = i) &= p_{ij}, \end{aligned} \tag{27}$$

$s_t \in \{1, \ldots, K\}$. This model assumes repeated regimes for the parameters. Alternatively, the parameters could simply be drawn from a nonrepeated change-point process, as assumed by Chib (1998). This amounts to assuming that $\Pr(s_{t+1} = k | s_t = k-1) = p_{k-1,k} > 0$, whereas $p_{j,k} = 0$ for $j \leq k-2$ and for $k > j$, which means that the process cannot go back from the current state to an earlier one. Rather, if it leaves the current state, it must exit to a new state. The number of states can therefore be expected to grow in proportion with the sample size under this specification.

Allowing for time variation in second moments has proven important for both economic and financial variables. A popular approach is to assume a stochastic volatility process of the form

$$y_{t+1} = \beta x_t + \sqrt{h_t}\varepsilon_{t+1},\ \varepsilon_{t+1} \sim N(0,1), \tag{28}$$
$$\log(h_t) = \gamma_0 + \gamma_1 \log(h_{t-1}) + u_{t+1},\ \ u_{t+1} \sim N(0, \sigma_u^2).$$

Models with stochastic volatility have proven popular in macroeconomic forecasting (see Clark 2011 and Clark & Ravazzolo 2015 in the context of density forecasting).

Alternatively, Markov switching models with regime switching in both the mean and variance parameters can be used:

$$y_{t+1} = \beta_{s_t} x_t + \sigma_{s_t}\varepsilon_{t+1}, \varepsilon_{t+1} \sim N(0,1). \tag{29}$$

Approaches such as Equations 26–29 are fully parametric in that they assume a fully specified process (including functional form and distribution) for changes to the parameters. Such approaches take into account the insight that if breaks affected the DGP in the past, then forecasts of future outcomes ought also to account for the possibility of future breaks, particularly at long forecast horizons (see Giacomini & Rossi 2009 for a related discussion and Giacomini & Rossi 2015 for a survey of forecasting under model instability).

In the context of a single break, one approach is to attempt to test for the presence and most likely location of the break. Estimates of the parameters of the forecasting model can then be based on only the data after the break (postbreak estimation). This approach is unlikely to work well if the location of the break is close to the end of the sample so that there are only few data points with which to train the postbreak estimation. Alternatively, one can use a procedure that attempts to trade off bias and variance effects by including prebreak data so as to minimize the expected loss of the forecast (Pesaran & Timmermann 2007).

When it is expected that there are multiple but rare breaks, one could consider attempting to estimate the most recent break and base forecasts on the recent stable period. Bai & Perron (1998) provide methods for estimating the break points. Pesaran & Timmermann (2002) suggest an alternative method based on reversing the time series. A third possibility is to use multiple estimation windows and average across the windows based on the inverse of the associated MSE values, assigning less weight to estimation methods that yield relatively poor forecasting performance (see Pesaran & Timmermann 2007).

Robust methods that do not rely on estimating the number of breaks include the use of a rolling window estimator, discounted least squares methods, and intercept corrections. The idea behind rolling windows is that "old" data get excluded in the estimation and so do not cause biases in the parameter estimates. Rolling window estimates tend to be more volatile than full-sample (recursive) estimates, which make more efficient use of all data in situations in which the parameters do not change much over time. Rolling window estimates can also critically depend on how the length of the estimation window is set, and there appear to be no procedures for addressing this important issue in practice.

Discounted least squares, in the context of a univariate autoregression, is the same as exponential smoothing, which is popular when there are few or no covariates. There is a wide variety of methods for exponential smoothing (see Hyndman et al. 2008 for a book-length examination). The idea of discounting is similar to that of rolling regressions; however, instead of having a zero-one cutoff for including past data, the weight on past data declines the further back it is in the past. Intercept corrections adjust the forecast using the past forecast error to "correct" the forecast. The main difficulty here is that the past forecast, although under some conditions being an unbiased estimator of the bias due to ignoring the parameter instability, is also a noisy estimate of this effect and can add greatly to the risk of the forecast.

One approach for handling model instability that seems to work well in practice is to consider a variety of model specifications and then use an equal-weighted average of the forecasts generated by the different approaches. For example, Clark & McCracken (2010) combine forecasts of output, inflation, and short-term interest rates across VAR specifications that allow for model instability in a variety of ways (e.g., by using recursive lag selection, rolling estimation windows, and intercept corrections). Equal-weighted averages of such forecasts are found to be consistently good. Rossi (2013) also finds that equal-weighted forecast combinations produce relatively good out-of-sample forecasts. Pettenuzzo & Timmermann (2016) confirm these findings but further find that a variety of forecast combination schemes—such as equal-weighted combination, Bayesian model averaging, and the optimal pool of Geweke & Amisano (2011)—produce accurate out-of-sample forecasts of the distribution of inflation and growth in industrial production when applied to density forecasts from models that allow for stochastic volatility, time-varying parameters, and regime shifts.

8. DENSITY VERSUS POINT FORECASTS

Density forecasts provide a full summary of forecast uncertainty, which is invaluable in many situations. In practice, public agencies have therefore moved toward providing density forecasts. For example, the Bank of England reports a "fan chart" forecast for inflation as does the IMF in their *World Economic Outlook* publication. Fan charts use different shades of colors to illustrate bands of quantiles starting from the median forecast and fanning out toward coverage of an increasingly likely range of outcomes for variables such as the inflation rate measured over increasing forecast horizons.[12]

A common suggestion is that density forecasting offers a superior approach to point forecasting because such forecasts are not tied to a particular loss function. In practice, however, the data must be employed to select the model and estimate the parameters, all of which involves the addition of a loss function that may not bear any relation to the point forecasting problem that the density is used for. In turn, given such a loss function and an estimate of the density forecast, $p_{y_{T+h}|z_T}(y|z_T, \theta)$, we can use Equation 2 to construct a point forecast.

In addition, the provision of density forecasts is important in situations with many forecast users whose loss functions are heterogeneous. For example, consider a public weather forecasting service tasked with predicting whether it will rain tomorrow. A density forecast is the conditional probability that it will rain tomorrow; point forecasts are either {rain, no rain}. Users will forecast rain when this probability is greater than the utility gain from forecasting a sunny day (utility from correctly forecasting a sunny day minus the utility from incorrectly forecasting it) as a proportion of the utility gain from forecasting sun plus the utility gain from forecasting rain. A conditional forecast of say a 40% chance of rain might lead those whose day would be ruined by rain (e.g., they were driving a long way to the beach or having an outdoor party) to make alternative arrangements; thus, their forecast is rain given that an incorrect forecast of sun is relatively costly. For people going to the (indoor) shopping mall who are indifferent between rain and sun, their forecast is a sunny day because the conditional probability of rain is below their relative utility (0.5) of a sunny day.

Reporting density forecasts in a way that allows users to construct their point forecasts also poses challenges. For binary outcomes such as many weather forecasts, a single number defines the conditional density. When the density forecast is a parametric density, the parameters of the

[12] An interesting development in the recent forecasting literature is the use of downside risk measures. For example, the IMF *World Economic Outlook* reports balance of risks (coefficients of skewness) for a selection of risk factors.

density can be presented. More often, though, features of the density are presented instead. For example the Bank of England's density forecast for inflation is a histogram with a number of other distribution features (mean, median, mode, skewness) reported. Obviously, a histogram suppresses information that could be useful in computing the optimal forecast.

Density forecasts depend on estimates of unknown parameters and hence require a loss function. This loss function may not be related to the forecast user's loss function. In the binary problem in which loss is over two outcomes and the density forecast is the conditional probability of one of the outcomes, there is a direct link between the scoring rule used to estimate the forecast density and the users' individual utility functions. Shuford et al. (1966) and Schervish (1989) show that proper scoring rules are weighted averages of the individual utility functions of users. For more general problems, the relationships between the scoring rules popular in the literature and the underlying loss functions for individual users is not clear. Thus, it could well be that the scoring rule used provides a poor estimate of the feature of the conditional distribution that some users wish to construct.

One area in which density forecasting has been particularly important is in the Bayesian forecasting literature. The posterior predictive density is naturally obtained as part of the Bayesian analysis and so is typically readily available to be used to form out-of-sample forecasts and for model evaluation (see Karlsson 2013 for a summary of the extensive Bayesian literature).

9. EVALUATING AND COMPARING FORECASTING PERFORMANCE

It is common to evaluate forecasts out of sample by splitting a data sample with T observations into an initial estimation sample of length R, used for parameter estimation and model selection, and an out-of-sample period of length $P = T - R$, used to evaluate the forecasts. The practice of evaluating forecast models using out-of-sample evaluation reflects the desire to obtain a better estimator for a forecast method's risk than that provided by an in-sample estimator. In-sample methods estimate the risk using the same data as that used to estimate the forecasting model. This provides an estimator whose in-sample loss is smaller on average than the true risk because the estimated model parameters have been obtained by minimizing the estimate of this measure.

Out-of-sample experiments generally assume that the parameters of the forecasting model, $\hat{\beta}_t$, are updated using a fixed, rolling, or expanding estimation window. Under the fixed estimation scheme, an initial data sample $\tau = 1, \ldots, R$ is used to estimate the parameters that do not get updated subsequently. Under the rolling estimation scheme, a fixed window of the data (of length w) up to the present point in time, $\tau = t-w+1 : t$, is used to estimate the parameters. As new data are observed, old observations are therefore dropped. Under the expanding estimation scheme, all data from the initial time up to the present $\tau = 1, \ldots, t$ are used to estimate the parameters.

Risk can be estimated from out-of-sample data by computing the sample average. Even in recent work in many fields, it is typical to report these sample estimates without construction of standard errors. West (1996) shows how to construct standard errors accounting for the estimation error in the construction of the forecasts for a wide range of possible loss functions and regression models. In many situations (e.g., under MSE error where the forecasts are constructed from linear regressions), estimation error plays no role and standard (robust) error measures are appropriate. This work has been generalized in a number of papers. However, asymptotic results remain to be developed for many practical forecasting problems.

9.1. Comparing Models Out of Sample

Pairwise comparisons of two models' out-of-sample forecasting performance are commonly conducted using the approach of Diebold & Mariano (1995). Specifically, let $L(f_{1t+1}(\hat{\beta}_{1t}), y_{t+1})$,

$t = R, \ldots, T-1$, be a sequence of one-step-ahead losses associated with a particular model (model 1) whose parameter estimates, $\hat{\beta}_{1t}$, only use information available at time t. In the case of MSE loss, $L(f_{1t+1|t}(\hat{\beta}_{1t}), y_{t+1}) = (y_{t+1} - f_{1t+1|t}(\hat{\beta}_{1t}))^2$. Similarly, let $L(f_{2t+1|t}(\hat{\beta}_{2t}), y_{t+1})$ be the sequence of losses generated by a competing forecasting model (model 2). Then the loss differential, d_{t+1}, is given by

$$d_{t+1} = L(f_{1t+1|t}(\hat{\beta}_{1t}), y_{t+1}) - L(f_{2t+1|t}(\hat{\beta}_{2t}), y_{t+1}). \tag{30}$$

To conduct a test of equal expected loss for models 1 and 2, $H_0 : E[d_{t+1}] = 0$, Diebold & Mariano (1995) propose using a robust t-test for the intercept in a regression of d_{t+1} on a constant. A sufficient condition ensuring that the Diebold-Mariano test is well behaved is that the loss differentials (d_{t+1}) are covariance stationary.

Under the out-of-sample estimation schemes, recursive updates to $\hat{\beta}_{jt}$ may result in important learning-induced nonstationarities in d_{t+1}. We would expect such effects to be quite small if the initial estimation scheme is long relative to the length of the evaluation sample. However, if this condition is not satisfied, alternative procedures for forecast evaluation can be used. Specifically, West (1996) and McCracken (2007) develop methods that can be used in the comparison of non-nested and nested forecasting models accounting for recursive updates to the parameter estimates of linear regression models.

To see the difference between the nested and non-nested case, consider two forecasting models:

$$\begin{aligned} y_{t+1} &= \beta_1' x_{1t} + \varepsilon_{1t+1}, \\ y_{t+1} &= \beta_2' x_{2t} + \varepsilon_{2t+1}. \end{aligned} \tag{31}$$

In situations in which both x_{1t} and x_{2t} contain elements that are not included in the other set, the two models are non-nested. Conversely, if $x_{1t} \subset x_{2t}$, the two models are nested with model 1 being the small model and model 2 being the large model.

The results of McCracken (2007) show that the distribution (and thus critical values) of the resulting test statistic for the null that the two models have equal expected loss depends on the method used to estimate the parameters of the models (fixed, rolling, or expanding estimation window) and the proportion of the overall sample, T, used for initial estimation (R) and forecast evaluation ($P = T - R$), respectively. For nested models, the critical values used in the model comparison also depend on the relative dimension of the small and large forecasting models, i.e., $\dim(x_{2t}) - \dim(x_{1t})$.

A somewhat surprising result that arises from these studies is that it is possible for a large forecasting model to perform worse (e.g., produce larger MSE values) in a given sample than the small forecasting model, yet at the same time be deemed to be the best forecasting model. This situation may arise because the hypothesis of equal expected loss is evaluated at the probability limits of the parameters estimates $\beta_i = p\lim(\hat{\beta}_{it})$, i.e.,

$$H_0 : E[L(f_{1t+1|t}(\beta_1), y_{t+1})] = E[L(f_{2t+1|t}(\beta_2), y_{t+1})]. \tag{32}$$

The null in Equation 32 implies that the additional regressors included in the large model have zero coefficients. If this holds, the large model can be expected to perform worse than the small model in a given sample simply because the large model has more parameters to estimate. The smaller model (model 1) gains efficiency by correctly imposing that the additional parameters equal zero. In other words, the large model is more adversely affected by estimation error than the small model. This means that the distribution for the relative forecasting performance of the small versus the large model gets shifted to the left. In cases with many additional parameters to estimate, this effect can be strong enough that the 95% right-tail critical values of the distribution for the test statistic are negative. The reason is that a large model found to underperform in a

given sample can be judged to be significantly better than the smaller model, when evaluated at the limit of the two sets of parameter estimates, because it underperformed by less than we would have expected if the additional parameters of the large model truly were zero in the population.

Giacomini & White (2006) propose a different approach to comparing equal expected loss. They replace the null hypothesis in Equation 32 that compares the two models' expected loss at the probability limits of the parameters with a null of equal expected loss evaluated at the current parameter estimates:

$$H_0' : E[L(f_{1t+1|t}(\hat{\beta}_{1t}), y_{t+1})] = E[L(f_{2t+1|t}(\hat{\beta}_{2t}), y_{t+1})]. \quad (33)$$

Again the object underlying their test is an observed sequence of forecasts from two models. Giacomini & White assume that these forecasts are generated using rolling window estimators.[13] This assumption preserves the effect of estimation error on the two forecasts and also enables them to establish the distribution of the sequence of losses generated by the two forecasts. The difference between the null hypotheses in Equations 32 and 33 is far from trivial. As an example, suppose that the finite-sample bias in the small model (model 1) arising from the omission of relevant predictor variables exactly cancels out against its smaller estimation error (relative to model 2). Then the null hypothesis in Equation 33 should not be rejected, whereas the null in Equation 32 should be rejected as the estimation sample expands and estimation error vanishes because the additional predictors included by the large model actually contain useful information. Stated differently, tests based on Equation 33 may set the bar higher for the large model than tests based on Equation 32 because the former requires that the large model outperforms the smaller forecasting model by a sufficiently large margin so as to make up for the larger model's greater estimation error.

The null in Equation 33 compares two models' forecasting performance using sequences of parameter estimates. This means that the null will change if the same models are maintained but their parameters are estimated differently. For example, model 1 may be preferred over model 2 if a rolling estimation window of 100 observations is being used, whereas the reverse may hold if the rolling estimation window is instead 500 observations. This highlights that Equation 33 really tests the equivalence of the performance of pairs of forecasting methods as opposed to comparing specific models. Not only the models but also how they are implemented makes a difference when testing the null. In fact, Equation 33 can be used to test the relative accuracy of the same model, estimated using different methods or using rolling estimation windows of different lengths (e.g., 200 versus 500 observations).

The ability to put standard errors on comparisons of different models' forecasting performance—sometimes called horse races—represents a major improvement in the literature on forecast evaluation. However, key challenges remain. First, the results established by West (1996) and Clark & McCracken (2001) are limited to a small set of estimators and forecasting models and exclude a variety of nonlinear, nonparametric, and Bayesian approaches, as well as forecasting methods that involve model selection. Similarly, although the approach advocated by Giacomini & White (2006) is conceptually elegant, in practice the reliance on a rolling window estimator can lead to substantial drops in statistical power. In situations with weak predictors, using an expanding estimation window rather than a rolling window can lead to somewhat better forecasting performance. However, models estimated in this manner cannot be evaluated using Equation 33.

[13] With rolling regressions, their result motivates the Diebold & Mariano (1995) method as asymptotically appropriate.

9.2. In-Sample Versus Out-of-Sample Forecast Evaluation

In-sample evaluation methods use the same data sample to estimate the parameters of a forecasting model and to evaluate it on. In contrast, out-of-sample evaluation models separate the samples used to estimate (and select) the model and to evaluate it with. In part, the answer depends on what the evaluation is used for.

In-sample forecast evaluation methods typically result in an estimated risk measure of the form of Equation 4. When forecasting with a linear regression model with k predictors, the in-sample MSE risk estimate is on average $\sigma^2(1 - T^{-1}k)$ as in Equation 11. Conversely, the out-of-sample expected risk for this model is roughly $\sigma^2(1 + T^{-1}k)$ as in Equation 16. Hence, it may be expected that the sample analogs to out-of-sample risk would be better estimators of the relevant risk for forecasting.

When comparing two models, in-sample tests have asymptotic optimality properties under standard assumptions on the DGP. However, given concerns related to data mining as well as a desire to mimic the actual forecast procedures used, out-of-sample forecast evaluation results are often examined. There are good reasons for this. Hansen & Timmermann (2015a) consider a simulation experiment in which a forecaster searches across multiple models, the set of which is obtained by considering all possible linear models formed by selecting from an increasing list of predictors. Assuming that the statistical tests used for evaluating forecasting performance do not correct for such model specification search—which in practice means that conventional critical values are used to evaluate the chosen model—they find that in-sample tests of forecasting performance tend to massively over-reject the null of no predictability even in situations where the null is used as the DGP. Although data mining also causes out-of-sample tests to over-reject, the effect is substantially weaker for such tests. These results suggest that, even though over-rejections of the test for no predictability due to data mining do not disappear in out-of-sample tests, it is considerably less likely that good out-of-sample forecasting performance is spurious compared with similarly good performance observed in sample.

This result would seem to suggest that out-of-sample tests are to be preferred over in-sample tests of forecasting performance. However, Inoue & Kilian (2005) and Hansen & Timmermann (2015b) show that out-of-sample tests are associated with considerable losses in power due to their use of a subset of the data for forecast evaluation as well as the greater estimation error resulting from recursive estimation of the parameters.

For the case with a recursively expanding estimation window, Hansen & Timmermann (2015b) derive the power function for comparison of two nested regression models analytically and show that the power—the ability of the out-of-sample test to correctly detect that the large model is "best" when this holds—gets weaker when the "hold-out" part of the data (i.e., the proportion of the data used for forecast evaluation) is a larger proportion of the sample size. Their results also suggest that findings of superior forecasting performance are more likely to be spurious when the forecast evaluation sample is short and, conversely, more likely to reflect genuine predictive ability for the larger forecasting model when the forecast evaluation sample is long.

Besides greater robustness against data mining, another benefit from inspecting a model's out-of-sample forecasting performance is that it offers insights into how the model's forecasting performance evolved over time. By plotting cumulative sums of squared forecast errors, perhaps compared for pairs of models, one can gain insights into the stability of a model's performance and any periods of unusually poor or good forecasting performance.

10. DEALING WITH DATA MINING

In situations with multiple competing forecasting models, there are limitations to conducting pairwise comparisons of forecasting performance. Most notably, even if individual comparisons

can be conducted using tests with a certain (fixed) size, the statistical properties of a sequence of joint pairwise tests are unclear and will depend on the joint distribution of the involved test statistics.

In a very influential paper that addresses the multiple comparison issue, White (2000) proposes a "reality check" procedure for testing whether at least one model, selected from a larger set of m models, is capable of beating some benchmark specification, labeled model 0. Let $d_{kt+1}(\hat{\beta}_t) = L(f_{0t+1|t}(\hat{\beta}_{0t}), y_{t+1}) - L(f_{kt+1|t}(\hat{\beta}_{kt}), y_{t+1})$ be the loss difference between the benchmark and the k-th model. The null that the benchmark model is not beaten by any of the competing models takes the form

$$H_0 : \max_{k=1,\ldots,m} E[d_{kt+1}(\beta_k)] \leq 0, \tag{34}$$

where $\beta_k = \text{plim}(\hat{\beta}_{kt})$. Three points arise from Equation 34. First, the null uses a $\leq$ inequality. If the benchmark produces a lower expected loss than the alternative models, the loss differential d_{kt+1} will be negative, so the null is that even the best of the m forecasts cannot reduce the benchmark model's expected loss. Second, the null in Equation 34 is evaluated at the probability limits of the parameter estimates, β_k. This is similar to the null used by West (1996) and Clark & McCracken (2001). Third, Equation 34 is a composite null that involves multiple inequalities and depends on the joint distribution of m loss differentials. White establishes high-level assumptions under which the joint distribution of the vector of (scaled) sample means $(\bar{d}_{1t+1}, \ldots, \bar{d}_{mt+1})'$ converges in distribution to $N(0, \Omega)$, where Ω is an $m \times m$ covariance matrix. Moreover, he develops a bootstrap procedure for sampling from this distribution under the null of equal predictive accuracy.[14] These bootstrap draws can be used to conduct inference to see whether the best model's performance is sufficiently far out in the right tail of the distribution of the test statistic to reject the null in Equation 34.[15]

The reality check approach developed by White (2000) can be used to conduct metastudies of the existing literature, i.e., to assess whether, across all models used in published studies, there is robust evidence of superior out-of-sample forecasting performance (relative to the benchmark) for at least one model. It can also be used in a constructive manner by a modeler who is interested in testing if the best model, selected from a larger set of candidate models, is genuinely able to beat some benchmark, after accounting for the multiple hypothesis testing problem in Equation 34. The more models are included in the search, typically the higher the bar is set for the best model to be deemed capable of outperforming the benchmark. In this way, pure data mining can be costly as it can confound our ability to identify a genuinely good forecasting model. The null in Equation 34 may be the relevant hypothesis if we are interested in testing market efficiency, i.e., in analyzing whether there exists an investment strategy that, net of transaction costs and on a risk-adjusted basis, beats holding the market portfolio.

We may also be interested in finding out how many models can beat the benchmark and in identifying such models. Romano & Wolf (2005) develop a stepwise procedure that iterates on White's bootstrap to recursively identify the set of superior models, while simultaneously controlling the probability of wrongly classifying at least one forecasting model as being superior.

In situations without an obvious benchmark to compare forecasting performance against, Hansen et al. (2011) develop ways to recursively trim the set of models deemed to be better

[14] To deal with the effect of recursive estimation error that arises for forecasts generated by models with estimated parameters, one can alternatively test a null hypothesis such as that of Giacomini & White (2006) or use the bootstrap approach proposed by Corradi & Swanson (2007).

[15] Hansen (2005) provides further refinements to this procedure through his test for superior predictive ability.

than others. Their approach uses an equivalence test for comparing models and an elimination rule for trimming inferior models. This approach can handle large-dimensional sets of competing forecasting models and so is useful when considering which models to drop.

11. CONCLUSION

Economic forecasting has seen many exciting new developments over the past couple of decades related to new developments in areas such as model selection, real-time forecasting, and estimation with large cross sections of potentially relevant predictors. Some of these developments have been theoretical in nature—such as our improved ability to put standard errors on estimates of out-of-sample measures of forecasting performance—whereas others have largely been driven by access to new empirical data and computer algorithms designed to search among a vast list of variables in the hope of identifying individual predictors that could make a difference under assumptions of sparseness.

A common theme emerging from our review is that no single model or forecasting method can be expected to be dominate over time and across different economic variables. Individual models are invariably coarse approximations to a far more complex and evolving reality with biases that shift over time. This helps explain the existence of a plethora of different approaches to forecasting, and it also explains the success of forecast combinations in many different arenas.

Improvements to our ability to produce more real-time forecasts, often on a daily basis, remain one of the most exciting areas of research. Access to high-frequency data sources such as scanner data from supermarkets, credit card transactions, and factory-level activity offers the hope for improvements in the accuracy of such forecasts along with the challenges that arise from pooling such data sources.

APPENDIX: DETAILS OF THE MONTE CARLO SIMULATIONS USED IN FIGURE 1

The Monte Carlo simulations used to construct **Figure 1** assume a linear forecasting model $y_{t+1} = \beta' x_t + \varepsilon_t$. The estimation sample has 100 observations, and the evaluation is for a one-step-ahead forecast. We use 30,000 replications. We draw ε_t as a standard normal random variable independent over time and independent of the $\{x_t\}_{t=1}^{T}$ random variables. We draw $x_t \sim N(0, \Sigma)$ independent over time with Σ being random across Monte Carlo draws. Specifically, we set $\Sigma = 0.1 Q' Q$, where each element of the $K \times K$ matrix Q is drawn from an independent normal distribution (which is also independent of all other draws). This method ensures that Σ is positive definite regardless of K and results in a spread of eigenvalues of Σ that appears reasonable.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Hashem Pesaran and an anonymous reviewer for many helpful comments on the article.

LITERATURE CITED

Andersen TG, Bollerslev T, Christoffersen PF, Diebold FX. 2006. Volatility and correlation forecasting. See Elliott et al. 2006, pp. 777–878

Andreou E, Ghysels E, Kourtellos A. 2011. Forecasting with mixed frequency data. In *Oxford Handbook of Economic Forecasting*, ed. MP Clements, DF Hendry, pp. 225–45. New York: Oxford Univ. Press

Andreou E, Ghysels E, Kourtellos A. 2013. Should macroeconomic forecasters use daily financial data and how? *J. Bus. Econ. Stat.* 31:240–51

Andrews DWK. 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* 821–56

Aruoba SB, Diebold FX, Scotti C. 2009. Real-time measurement of business conditions. *J. Bus. Econ. Stat.* 27:417–27

Bai J, Ng S. 2009. Boosting diffusion indices. *J. Appl. Econom.* 24:607–29

Bai J, Perron P. 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 66:47–78

Banbura M, Giannone D, Reichlin L. 2011. Nowcasting. In *The Oxford Handbook of Economic Forecasting*, ed. MP Clements, DF Hendry, pp. 193–224. New York: Oxford Univ. Press

Belloni A, Chernozhukov V. 2011. High dimensional sparse econometric models: an introduction. In *Inverse Problems and High-Dimensional Estimaton*, ed. P Alquier, E Gautier, G Stolz, pp. 121–56. New York: Springer

Campbell JY, Thompson SB. 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? *Rev. Financ. Stud.* 21:1509–31

Chib S. 1998. Estimation and comparison of multiple change-point models. *J. Econom.* 86:221–41

Chudik A, Grossman V, Pesaran MH. 2016a. A multi-country approach to forecasting output growth using PMIs. *J. Econom.* 192:349–65

Chudik A, Kapetanios G, Pesaran MH. 2016b. *Big data analytics: a new perspective*. Unpublished manuscript, Univ. South. Calif., Los Angeles

Clark TE. 2011. Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *J. Bus. Econ. Stat.* 29:327–41

Clark TE, McCracken MW. 2001. Tests of equal forecast accuracy and encompassing for nested models. *J. Econom.* 105:85–110

Clark TE, McCracken MW. 2010. Averaging forecasts from VARs with uncertain instabilities. *J. Appl. Econom.* 25:5–29

Clark TE, Ravazzolo F. 2015. Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *J. Appl. Econom.* 30:551–75

Clements MP, Hendry DF. 1998. Forecasting economic processes. *Int. J. Forecast.* 14:111–31

Corradi V, Swanson NR. 2007. Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. *Int. Econ. Rev.* 48:67–109

Croushore D, Stark T. 2001. A real-time data set for macroeconomists. *J. Econom.* 105:111–30

D'Amuri F, Marcucci J. 2012. *The predictive power of Google searches in forecasting unemployment*. Work. Pap., Bank of Italy

Del Negro M, Schorfheide F. 2013. DSGE model-based forecasting. See Elliott & Timmermann 2013, pp. 57–140

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *J. Bus. Econ. Stat.* 13:253–63

Elliott G, Gargano A, Timmermann A. 2013. Complete subset regressions. *J. Econom.* 177:357–73

Elliott G, Gargano A, Timmermann A. 2015. Complete subset regressions with large-dimensional sets of predictors. *J. Econ. Dyn. Control* 54:86–110

Elliott G, Granger CWJ, Timmermann A, eds. 2006. *Handbook of Economic Forecasting*, Vol. 1. Amsterdam: North-Holland

Elliott G, Komunjer I, Timmermann A. 2005. Estimation and testing of forecast rationality under flexible loss. *Rev. Econ. Stud.* 72:1107–25

Elliott G, Komunjer I, Timmermann A. 2008. Biases in macroeconomic forecasts: irrationality or asymmetric loss? *J. Eur. Econ. Assoc.* 6:122–57

Elliott G, Müller UK. 2006. Efficient tests for general persistent time variation in regression coefficients. *Rev. Econ. Stud.* 73:907–40
Elliott G, Timmermann A, eds. 2013. *Handbook of Economic Forecasting*, Vol. 2. Amsterdam: North-Holland
Engle RF, Watson MW. 1985. Testing for regression coefficient stability with a stationary AR(1) alternative. *Rev. Econ. Stat.* 67:341–46
Genre V, Kenny G, Meyler A, Timmermann A. 2013. Combining expert forecasts: Can anything beat the simple average? *Int. J. Forecast.* 29:108–21
Geweke J, Amisano G. 2011. Optimal prediction pools. *J. Econom.* 164:130–41
Giacomini R, Rossi B. 2009. Detecting and predicting forecast breakdowns. *Rev. Econ. Stud.* 76:669–705
Giacomini R, Rossi B. 2015. Forecasting in nonstationary environments: what works and what doesn't in reduced-form and structural models. *Annu. Rev. Econ.* 7:207–29
Giacomini R, White H. 2006. Tests of conditional predictive ability. *Econometrica* 74:1545–78
Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* 102:359–78
Granger CWJ, Pesaran MH. 2000. Economic and statistical measures of forecast accuracy. *J. Forecast.* 19:537–60
Hansen PR. 2005. A test for superior predictive ability. *J. Bus. Econ. Stat.* 23:365–80
Hansen PR, Lunde A, Nason JM. 2011. The model confidence set. *Econometrica* 79:453–97
Hansen PR, Timmermann A. 2015a. Comment on "Comparing predictive accuracy, twenty years later" by FX Diebold. *J. Bus. Econ. Stat.* 33:17–21
Hansen PR, Timmermann A. 2015b. Equivalence between out-of-sample forecast comparisons and Wald statistics. *Econometrica* 83:2485–505
Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning*. New York: Springer. 2nd ed.
Hornik K, Stinchcombe M, White H. 1989. Multilayer feedforward networks are universal approximators. *Neural Netw.* 2:359–66
Hyndman RJ, Koehler AB, Ord JK, Snyder RD. 2008. *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin: Springer Verlag
Inoue A, Kilian L. 2005. In-sample or out-of-sample tests of predictability: Which one should we use? *Econom. Rev.* 23:371–402
Karlsson S. 2013. Forecasting with Bayesian vector autoregression. See Elliott & Timmermann 2013, pp. 791–897
McCracken MW. 2007. Asymptotics for out of sample tests of Granger causality. *J. Econom.* 140:719–52
Medeiros M, Mendes EF. 2015. *Regularization of high-dimensional time-series models with flexible innovations*. Unpublished manuscript, Pontif. Cathol. Univ. Rio and Univ. New South Wales, Sydney
Patton AJ, Timmermann A. 2007. Testing forecast optimality under unknown loss. *J. Am. Stat. Assoc.* 102:1172–84
Pesaran MH, Schuermann T, Weiner SM. 2004. Modelling regional interdependencies using a global error-correcting macroeconometric model. *J. Bus. Econ. Stat.* 22:129–62
Pesaran MH, Skouras S. 2002. Decision-based methods for forecast evaluation. In *A Companion to Economic Forecasting*, ed. MP Clements, DF Hendry, pp. 241–67. Hoboken, NJ: Blackwell
Pesaran MH, Timmermann A. 2002. Market timing and return prediction under model instability. *J. Empir. Finance* 9:495–510
Pesaran MH, Timmermann A. 2007. Selection of estimation window in the presence of breaks. *J. Econom.* 137:134–61
Pettenuzzo D, Timmermann A. 2016. Forecasting macroeconomic variables under model instability. *J. Bus. Econ. Stat.* In press
Pettenuzzo D, Timmermann A, Valkanov R. 2014. Forecasting stock returns under economic constraints. *J. Financ. Econ.* 114:517–53
Pettenuzzo D, Timmermann A, Valkanov R. 2015. *A MIDAS approach to modeling first and second moment dynamics*. Unpublished manuscript, Brandeis Univ., Waltham, MA, and Univ. Calif., San Diego
Romano JP, Wolf M. 2005. Stepwise multiple testing as formalized data snooping. *Econometrica* 73:1237–82
Rossi B. 2013. Advances in forecasting under instability. See Elliott & Timmermann 2013, pp. 1203–324

Rossi AG, Timmermann A. 2015. Modeling covariance risk in Merton's ICAPM. *Rev. Financ. Stud.* 28:1428–61
Schervish MJ. 1989. A general method for comparing probability assessors. *Ann. Stat.* 17:1856–79
Shuford EH, Albert A, Massengill HE. 1966. Admissible probability measurement procedures. *Psychometrika* 31(2):125–45
Skouras S. 2007. Decisionmetrics: a decision-based approach to econometric modelling. *J. Econom.* 137:414–40
Smets F, Wouters R. 2003. An estimated dynamic stochastic general equilibrium model of the euro area. *J. Eur. Econ. Assoc.* 1:1123–75
Stock JH, Watson MW. 1996. Evidence on structural instability in macroeconomic time series relations. *J. Bus. Econ. Stat.* 14:11–30
Stock JH, Watson MW. 2002. Forecasting using principal components from a large number of predictors. *J. Am. Stat. Assoc.* 97:1167–79
Stock JH, Watson MW. 2006. Forecasting with many predictors. See Elliott et al. 2006, pp. 515–54
Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 73:267–88
Timmermann A. 2006. Forecast combinations. See Elliott et al. 2006, pp. 135–96
West KD. 1996. Asymptotic inference about predictive ability. *Econometrica* 64:1067–84
White H. 2000. A reality check for data snooping. *Econometrica* 68:1097–26
Wright JH. 2013. Evaluating real-time VAR forecasts with an informative democratic prior. *J. Appl. Econom.* 28:762–76
Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67:301–20

Annual Review of Economics

Volume 8, 2016

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Economics* articles may be found at http://www.annualreviews.org/errata/economics