2

Loss Functions

Short of the special and ultimately uninteresting case with perfect foresight, it is not possible to find a method that always sets the forecast equal to the outcome. A formal method for trading off potential forecast errors of different signs and magnitudes is therefore required. The loss function, $L(\cdot)$, describes in relative terms how costly it is to use an imperfect forecast, f, given the outcome, Y, and possibly other observed data, Z. This chapter examines the construction and properties of loss functions and introduces loss functions that are commonly used in forecasting.

A central point in the construction of loss functions is that the loss function should reflect the actual trade-offs between different forecast errors. In this sense the loss function is a primitive to the forecasting problem. From a decision-theoretic perspective the forecast is the action that must be constructed given the loss function and the predictive distribution, which we discuss in the next chapter. For example, the Congressional Budget Office must provide forecasts of future budget deficits. Their loss function in providing the forecasts should be based on the relative costs of over- and underpredicting public deficits. Weather forecasters face very different costs from underpredicting the strength of a storm compared to overpredicting it.

The choice of a loss function is important for every facet of the forecasting exercise. This choice affects which forecasting models are preferred as well as how their parameters are estimated and how the resulting forecasts are evaluated and compared against forecasts from competing models. Despite its pivotal role, it is common practice to simply choose off-the-shelf loss functions. In doing this it is important to choose a loss function that at least approximately reflects the types of trade-offs relevant for the forecast problem under study. For example, when forecasting hotel room bookings, it is hard to imagine that over- and underpredicting the number of hotel rooms booked on a particular day lead to identical losses because hotel rooms are a perishable good. Hence, using a symmetric loss function for this problem would make little sense. Asymmetric loss that reflects the larger loss from over- rather than underpredicting bookings would be more reasonable.

There are examples of carefully grounded loss functions in the economics literature. For example, sometimes a forecast can be viewed as a signal in a strategic game that is influenced by the forecast provider's incentives. Studies such as Ehrbeck and Waldmann (1996), Hong and Kubik (2003), Laster, Bennett, and Geoum (1999), Ottaviani and Sørensen (2006), Scharfstein and Stein (1990) and Trueman (1994) suggest loss functions grounded on game-theoretical models. Forecasters are

assumed to differ in their ability to predict future outcomes. The chief objective of the forecasters is to influence forecast users' assessment of their ability. Such objectives are common for business analysts or analysts employed by financial services firms such as investment banks or brokerages whose fees are directly linked to clients' assessment of their forecasting ability.

The chapter proceeds as follows. Section 2.1 examines general issues that arise in construction of loss functions. We discuss the mathematical setup of a loss function before relating it to the forecaster's decisions and examining some general properties that loss functions have. Section 2.2 reviews specific loss functions commonly used in economic forecasting problems, assuming there is only a single outcome to predict, before extending the analysis in section 2.3 to cover cases with multiple outcome variables. Section 2.4 considers loss functions (scoring rules) for distributional forecasts, while section 2.5 provides some concrete examples of loss functions and economic decision problems from macroeconomic and financial analysis. Section 2.6 concludes the chapter.

2.1 CONSTRUCTION AND SPECIFICATION OF THE LOSS FUNCTION

Let Y denote the random variable describing the outcome of interest and let $\mathcal{Y}$ denote the set of all possible outcomes. For outcomes that are either continuous or can take on a very large number of possible values, typically $\mathcal{Y}$ is the real line, $\mathbb{R}$. In some forecasting problems the set of possible outcomes, $\mathcal{Y}$, can be much smaller, such as for a binary random variable where $\mathcal{Y} = \{0, 1\}$. For multivariate outcomes typically $\mathcal{Y} = \mathbb{R}^k$ for some integer k, where k is the number of forecasts to be evaluated.

Point forecasts are denoted by f and are defined on the set $\mathcal{F}$. Typically we assume $\mathcal{F} = \mathcal{Y}$ since in most cases it does not make sense to have forecasts that cannot take on the same values as Y or, conversely, have forecasts that can take on values that the outcome Y cannot. There are exceptions to this rule, however. For example, a forecast of the number of children per family could be a fraction such as 1.9, indicating close to 2 children, even though Y cannot take this value. We assume that the predictors Z (as well as the outcome Y and hence the forecast f) are real valued. Formally, the loss function, $L(f, Y, Z)$, is then defined as a mapping $L : \mathcal{Y} \times \mathcal{Y} \times \mathcal{Z} \mapsto \mathcal{L}$, where $\mathcal{L}$ is in $\mathbb{R}^1$, and $\mathcal{Z}$ contains the set of possible values the conditioning variables, z, can take. Often $\mathcal{L} = \mathbb{R}^1_+$, the set of nonnegative real numbers. Alternatively, we could constrain the forecasts to lie in the convex hull of the set of all possible outcomes, i.e., $\mathcal{F} = \text{conv}(\mathcal{Y})$. We discuss this further below.

A common assumption for loss functions is that loss is minimized when the forecast is equal to the outcome—$\min_f L(f, y, z) = L(y, y, z)$. The idea is that if we are to find a forecast that minimizes loss, then nothing dominates a perfect forecast. In cases where the loss function does not depend on Z, so $L(f, Y, Z) = L(f, Y)$, it is natural to normalize the loss function so that it takes a minimum value at 0. This can be done without loss of generality by subtracting the loss associated with the perfect forecast $f = y$, i.e., $L(f, Y) = \tilde{L}(f, Y) - \tilde{L}(Y, Y)$ for any loss function, $\tilde{L}$. For $f = y$ to be a unique minimum we must have $L(f, y) > 0$ for all $f \neq y$.[1] More generally, when the loss function $L(f, Y, Z)$ varies with Z, it may not be possible to

[1] In binary forecasting this condition is often not imposed. This usually does not affect the analysis but only the interpretation of the calculated loss figures.

rescale the loss function in this manner. For example, a policy maker's loss function over inflation forecasts might depend on the unemployment rate so that losses from incorrect inflation forecasts depend on whether the unemployment rate is high or low. For simplicity, in what follows we will mostly drop the explicit dependence of the loss function on Z and focus on the simpler loss functions $L(f, Y)$.

2.1.1 Constructing a Loss Function

Construction of loss functions, much like construction of prior distributions in Bayesian analysis, requires a careful study of the forecasting problem at hand and should reflect the actual trade-offs between forecast errors of different signs and magnitudes. Laying out the trade-off can be straightforward if the decision environment is fully specified and naturally results in a measurable outcome that depends on the forecast. For example, for a profit-maximizing investor with a specific trading strategy that requires forecasts of future asset prices, the natural choice of loss is the function relating payoffs to the forecast and realized returns. Other problems may not lead so easily to a specific loss function. For example, when the IMF forecasts individual countries' budget deficits, both short-term considerations related to debt financing costs and long-term reputational concerns could matter.[2] In such cases one can again follow a Bayesian prior selection strategy of defining a function that approximates a reasonable shape of losses associated with decisions based on incorrect forecasts.

Loss functions, as used by forecasters to evaluate their performance, and utility functions, as used by economists to assess the economic value of different outcomes, are naturally related. Both are grounded in the same decision-theoretic setup which regards the forecast as the decision and the outcome as the true state and maps pairs of outcomes (states) and forecasts (Y, f) to the real line. In both cases we are interested in minimizing the expected loss or disutility that arises from the decision.[3]

The relationship between utility and loss is examined in Granger and Machina (2006), who show that the loss function can be viewed as the negative of a utility function, although a more general relation of the following form holds:

$$U(f, Y) = k(Y) - L(f, Y), \tag{2.1}$$

where $k(Y)$ plays no role in the derivation of the optimal forecast.[4]

Example 2.1.1 (Squared loss and utility)**.** *Granger and Machina (2006) show that a utility function $U(f, Y)$ generates squared error loss, $L(f, Y) = a(Y - f)^2$, for $a > 0$, if and only if it takes the form*

$$U(f, Y) = k(Y) - a(Y - f)^2. \tag{2.2}$$

It follows that utility functions associated with squared error loss are restricted to a very narrow set.

[2] Forecasts can even have feedback effects on outcomes as in the case of credit ratings companies whose credit scores can trigger debt payments for private companies that affect future ratings (Manso, 2013).

[3] The first section of chapter 3 examines this issue in more detail.

[4] Granger and Machina (2006) allow decisions to depend on forecasts without requiring that the two necessarily be identical. Instead they require that the function mapping forecasts to decisions is monotonic.

Academic studies often do not derive loss functions from first principles by referring to utility functions or fully specified decision-theoretic problems, though there are some exceptions. Loss functions that take the form of profit functions have been used to evaluate forecasts by Leitch and Tanner (1991) and Elliott and Ito (1999). West et al. (1993) compare utility-based and statistical measures of predictive accuracy for exchange rate models. Examples of loss functions derived from utility are provided in the final section of this chapter.

2.1.2 Common Properties of Loss Functions

Reasonable loss functions are grounded in economic decision problems. Under the utility-maximizing approach, loss functions inherit well-known properties from the utility function. Rather than deriving loss functions from first principles, however, it is common practice to instead use loss functions with a "reasonable shape." For the loss function to be "reasonable," a set of minimal properties should hold. Other properties such as symmetry or homogeneity may suggest broad families of loss functions with certain desirable characteristics. We cover both types of properties below.

Trade-offs between different forecast errors when $f \neq y$ are quantified by the loss function. To capture the notion that bigger errors imply bigger losses, often it is imposed that the loss is nondecreasing as the forecast moves further away from the outcome. Mathematically, this means that $L(f_2, y) \geq L(f_1, y)$ for either $f_2 > f_1 > y$ or $f_2 < f_1 < y$ for all real y. Nearly all loss functions used in practice have this feature.

For loss functions that depend only on the forecast error, $e = y - f$, and thus take the form $L(f, y) = L(e)$, Granger (1999) summarized these requirements:

$$L(0) = 0 \text{ (minimal loss of 0);} \tag{2.3a}$$

$$L(e) \geq 0 \text{ for all } e; \tag{2.3b}$$

$L(e)$ is nonincreasing in e for $e < 0$ and nondecreasing in e for $e > 0$:

$$L(e_1) \leq L(e_2) \text{ if } e_2 < e_1 < 0, \qquad L(e_1) \leq L(e_2) \text{ if } e_2 > e_1 > 0. \tag{2.3c}$$

As in the case with more general loss, $L(f, y)$, condition (2.3a) simply normalizes the loss associated with the perfect forecast ($y = f$) to be 0. The second condition states that imperfect forecasts ($y \neq f$) generate larger loss than perfect ones. Most common loss functions depend only on e; see section 2.2 for examples.

Other properties of loss functions such as homogeneity, symmetry, differentiability, and boundedness can be used to define broad classes of loss functions. We next review these.

Homogeneity can be used to define classes of loss functions that lead to the same decisions. Homogeneous loss functions factor in such a way that

$$L(af, ay) = h(a)L(f, y), \tag{2.4}$$

for some positive function $h(a)$, where the degree of homogeneity does not matter. For loss functions that depend only on the forecast error, homogeneity amounts to $L(ae) = h(a)L(e)$ for some positive function $h(a)$. Homogeneity is a useful property when solving for optimal forecasts since the optimal forecast will be invariant to different values of $h(a)$.

Symmetry of the loss function refers to symmetry of the forecast around y. It is the property that, for all f,

$$L(y - f, y) = L(y + f, y). \tag{2.5}$$

For loss functions that depend only on the forecast error, symmetry reduces to $L(-e) = L(e)$, so that over- and underpredictions of the same magnitude lead to identical loss.[5]

Most empirical work in economic forecasting assumes symmetric loss. This choice reflects the difficulties in putting numbers on the relative cost of over- and underpredictions. Construction of a loss function requires a deeper understanding of the forecaster's objectives and this may be difficult to accomplish. Still, the implicit choice of MSE loss by the majority of studies in the forecasting literature seems difficult to justify on economic grounds. As noted by Granger and Newbold (1986, page 125), "an assumption of symmetry about the conditional mean... is likely to be an easy one to accept... an assumption of symmetry for the cost function is much less acceptable."

Differentiability of the loss function with respect to the forecast is again a regularity condition that is useful and helps simplify numerically the search for optimal forecasts. However, this condition may not be desirable and is certainly not required for a loss function to be well defined. In general, a finite numbers of points where the loss function fails to be differentiable will not cause undue problems at the estimation stage. However, when the loss function is extremely irregular, different methods are required for understanding the statistical properties of the loss function (see the maximum utility estimator in chapter 12).

Finally, loss functions may be bounded or unbounded. As a practical matter, there is often no obvious reason to let the weight the loss function places on very large forecast errors increase without bound. For example, the squared error loss function examined below assigns very different losses to forecasts of, say, US inflation that result in errors of 100% versus 500% even though it is not obvious that the associated losses should really be very different since both forecasts would lead to very similar actions. Unbounded loss functions can create technical problems for the analysis of forecasts as the expected loss may not exist, so most results in decision theory are derived under the assumption of bounded loss. In practice, forecasts are usually bounded and extremely large forecasts typically get trimmed as they are deemed implausible.

2.1.3 Existence of Expected Loss

Restrictions must be imposed on the form of the loss function to make sense of the idea of minimizing the expected loss. Most basically, it is required that the expected loss exists. Suppose the forecast depends on data Z through a vector of parameters, β, which depends on the parameters of the data generating process, θ, so $f = f(z, \beta)$. From the definition of expected loss, we have

$$E_Y[L(f(z, \beta), Y)] = \int L(f(z, \beta), y) p_Y(y|z, \theta) dy, \tag{2.6}$$

[5] A related concept is the class of bowl-shaped loss functions. A loss function is bowl shaped if the level sets $\{e : L(e) \leq c\}$ are convex and symmetric about the origin.

where $p_Y(y|z,\theta)$ is the predictive density of y given z,θ. When the space of outcomes $\mathcal{Y}$ is finite, this expression is guaranteed to be finite. However, for outcomes that are continuously distributed, restrictions must sometimes be imposed on the loss function to ensure finite expected loss. The existence of expected loss depends, both, on the loss function and on the distribution of the predicted variable, given the data, $p_Y(y|z,\theta)$, where θ denotes the parameters of this conditional distribution. Existence of expected loss thus hinges on how large losses can get in relation to the tail behavior of the predicted variable, as captured by $p_Y(y|z,\theta)$.

A direct way to ensure that the expected loss exists is to bound the loss function from above.[6] From a practical perspective this would seem to be a sensible practice in constructing loss functions. Even so, many of the most popular loss functions are not bounded from above. In part this practice stems from not considering the loss related to the forecasting problem at hand, but instead borrowing "off-the-shelf" loss functions from estimation methods that lead to simple closed-form expressions for the optimal forecast.

It is useful to demonstrate the conditions needed to ensure that the expected loss exists. Following Elliott and Timmermann (2004), suppose that L depends only on the forecast error, $e = y - f$, and lends itself to a Taylor-series expansion around the mean error, $\mu_e = E_Y[Y - f]$:

$$L(e) = L(\mu_e) + L'_{\mu_e}(e-\mu_e) + \frac{1}{2}L''_{\mu_e}(e-\mu_e)^2 + \sum_{k=3}^{\infty}\left(\frac{1}{k!}\right)L^k_{\mu_e}(e-\mu_e)^k, \quad (2.7)$$

where $L^k_{\mu_e}$ denotes the kth derivative of L evaluated at μ_e. Suppose there are only a finite number of points where L is not analytic and that these can be ignored because they occur with probability 0. Taking expectations in (2.7), we then get

$$\begin{aligned} E[L(e)] &= L(\mu_e) + \frac{1}{2}L''_{\mu_e}E_Y[(e-\mu_e)^2] + \sum_{k=3}^{\infty}\left(\frac{1}{k!}\right)L^k_{\mu_e}E_Y[(e-\mu_e)^k] \\ &= L(\mu_e) + \frac{1}{2}L''_{\mu_e}E_Y[(e-\mu_e)^2] + \sum_{k=3}^{\infty}\left(\frac{1}{k!}\right)L^k_{\mu_e}\sum_{i=0}^{k}\binom{k}{i}E_Y[e^{k-i}\mu_e^i] \\ &= L(\mu_e) + \frac{1}{2}L''_{\mu_e}E_Y[(e-\mu_e)^2] + \sum_{k=3}^{\infty}L^k_{\mu_e}\sum_{i=0}^{k}\frac{1}{i!(k-i)!}E_Y[e^{k-i}\mu_e^i]. \end{aligned} \quad (2.8)$$

This expression is finite provided that all moments of the error distribution exist for which the corresponding derivative of the loss function with respect to the forecast error is nonzero. This is a strong requirement and rules out some interesting combinations of loss functions and forecast error distributions. For example, exponential loss (or the Linex loss function defined below) and a student-t distribution with a finite number of degrees of freedom would lead to infinite expected loss since all higher-order moments do not exist for this distribution. What is required to make

[6] This is sufficient since we have already bounded the loss function (typically at 0) from below.

the higher-order terms in (2.8) vanish is that the tail decay of the predicted variable is sufficiently fast relative to the weight on these terms implied by the loss function.

2.1.4 Loss Functions Not Based on Expected Loss

So far we have characterized the loss function $L(f, Y)$ for a univariate outcome, and defined its properties with reference to a "one-shot" problem. This makes sense when forecasting is placed in a decision-theoretic or utility-maximization context. This approach to forecasting is internally consistent, from initially setting up the problem to defining the expected loss and conducting model estimation and forecast evaluation.

Some loss functions that have been used in practice are based directly on sample statistics without relating the sample loss to a population loss function. In cases where such a population loss function exists and satisfies reasonable properties, this does not cause any problems. Basing the loss function directly on a sample of losses can, however, sometimes yield a loss function that does not make sense in population or for fully specified decision problems. Loss functions that do not map back to decision problems often have poor and unintended properties. We consider one such example below.

Example 2.1.2 (Kuipers score for binary outcome)**.** *Let $f = \{1, -1\}$ be a forecast of the binary variable $y = \{1, -1\}$ and let $n_{j,k}$, $j, k \in \{-1, 1\}$ be the number of observations for which the forecast equals j and the outcome equals k. The Kuipers score is given by*

$$\frac{n_{1,1}}{n_{1,1} + n_{-1,1}} - \frac{n_{1,-1}}{n_{1,-1} + n_{-1,-1}}. \tag{2.9}$$

This is the positive hit rate, i.e., the proportion of times where $y = 1$ is correctly predicted less the "false positive rate," i.e., the proportion of times where $y = 1$ is wrongly predicted. This can equivalently be thought of as

$$\text{KuS} = \frac{n_{1,1}}{n_{1,1} + n_{-1,1}} + \frac{n_{-1,-1}}{n_{1,-1} + n_{-1,-1}} - 1, \tag{2.10}$$

which is the hit rate for $y = 1$ plus the hit rate for $y = -1$ minus a centering constant of 1. The Kuipers score is positive if the sum of the positive and negative hit rates exceeds 1. For a sample with a single observation, this definition makes no sense, as one of the denominators in (2.10) is 0: either $n_{1,1} + n_{-1,1} = 0$ or $n_{1,-1} + n_{-1,-1} = 0$.

For a single observation, this sample statistic does not follow from any obvious loss function. The first term in (2.10) is the sample analog of $P[f = 1|Y = 1]$ and the second is the sample analog of $P[f = -1|Y = -1]$. However, they do not combine to a loss function with this sample analog. This failure to embed the loss function into the expected loss framework results in odd properties for the objective. For example, the definition of KuS *in (2.9) implies that the marginal value of an extra "hit," i.e., a correct call, depends on the sample proportion of hits. To see this, consider the improvement in* KuS *from adding a single successfully predicted observation $y = 1$, $f = 1$.*

The resulting improvement in the hit rate is

$$\begin{aligned}\Delta\text{KuS} &= \frac{n_{1,1}+1}{n_{1,1}+1+n_{-1,1}} - \frac{n_{1,1}}{n_{1,1}+n_{-1,1}} \\ &= \frac{n_{-1,1}}{(n_{1,1}+n_{-1,1})(n_{1,1}+1+n_{-1,1})}.\end{aligned}$$

Thus the marginal value of a correct call depends on the total number of observations and the proportion of missed hits prior to the new observation. The Kuipers score's poor properties arise from the lack of justification of its setup for a population problem.

2.2 SPECIFIC LOSS FUNCTIONS

We next review various families of loss functions that have been suggested in the forecasting literature. The vast majority of empirical work on forecasting assumes that the loss function depends only on the forecast error, $e = Y - f$, i.e., the difference between the outcome and the forecast. In this case we can write $L(f, Y, Z) = L(e)$. In general, loss functions can be more complicated functions of the outcome and forecast and take the form $L(f, Y)$ or $L(f, Y, Z)$.

2.2.1 Loss That Depends Only on Forecast Errors

The most commonly used loss functions, including squared error loss and absolute error loss, depend only on the forecast error. For such loss functions, $L(f, Y, Z) = L(e)$, so the loss function takes a particularly simple form.

2.2.1.1 Squared Error Loss

By far the most popular loss function in empirical studies is squared error loss, also known as quadratic or mean squared error (MSE) loss:

$$L(e) = ae^2, \quad a > 0. \tag{2.11}$$

This loss function clearly satisfies the three Granger properties listed in (2.3). When viewed as a family of loss functions—corresponding to different values of the scalar a—squared error loss forms a homogeneous class.[7] It is symmetric, bowl shaped, and differentiable everywhere and penalizes large forecast errors at an increasing rate due to its convexity in $|e|$. The loss function is not bounded from above. Large forecast errors or "outliers" are thus very costly under this loss function.

2.2.1.2 Absolute Error Loss

Rather than using squared error loss, which results in increasingly large losses for large forecast errors, the absolute error is preferred in some cases. Under mean absolute error (MAE) loss,

$$L(e) = a\,|e|\,, \quad a > 0. \tag{2.12}$$

[7] While the scaling factor, a, does not matter to the properties of the optimal forecast, it is common to set $a = 0.5$, which removes the "2" that arises from taking first derivatives.

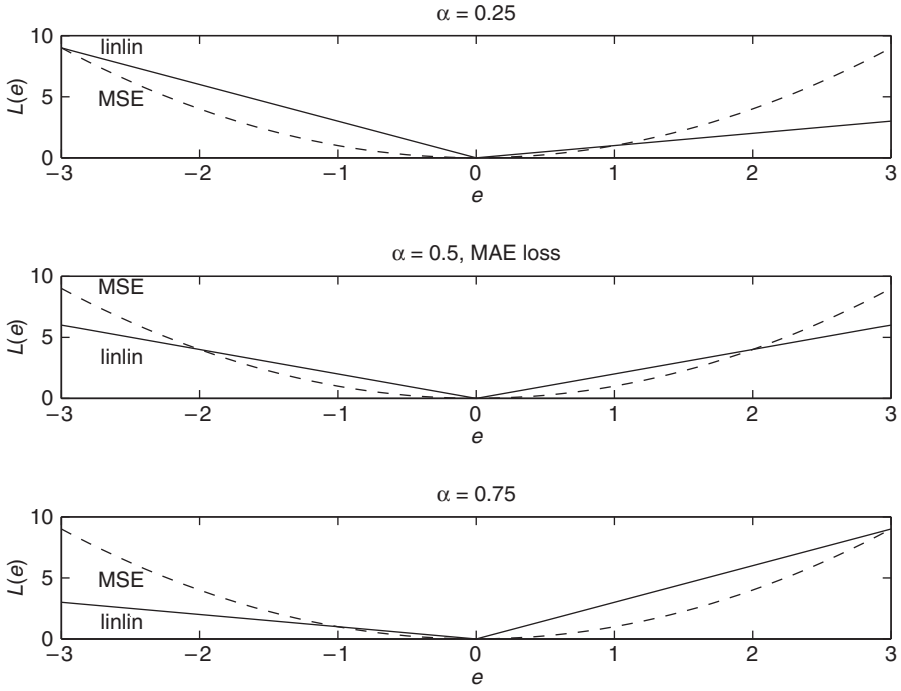

Figure 2.1: MSE loss versus lin-lin loss for different values of the lin-lin asymmetry parameter, α.

Like MSE loss, this loss function satisfies the three Granger properties listed in (2.3). The loss function is symmetric, bowl shaped, and differentiable everywhere except at 0. It is again unbounded. However, the penalty to large forecast errors increases linearly rather than quadratically as for MSE loss.

2.2.1.3 Piecewise Linear Loss

Piecewise linear, or so-called lin-lin loss, takes the form

$$L(e) = \begin{cases} -a(1-\alpha)e & \text{if } e \leq 0, \\ a\alpha e & \text{if } e > 0, \end{cases} \qquad a > 0, \tag{2.13}$$

for $0 < \alpha < 1$. Positive forecast errors are assigned a (relative) weight of α, while negative errors get a weight of $1 - \alpha$. The greater is α, the bigger the loss from positive forecast errors, and the smaller the loss from negative errors. Again, this loss function forms a homogeneous class for all positive values of a. It is common to set $a = 1$, so that the weights are normalized to sum to 1.

Lin-lin loss clearly satisfies the three Granger properties. Moreover, it is differentiable everywhere, except at 0. Compared to MSE loss, this loss function does not penalize large errors as much. MAE loss arises as a special case of lin-lin loss if $\alpha = 0.5$, in which case (2.13) simplifies to (2.12).

Figure 2.1 plots lin-lin loss against squared error loss. The middle window shows the symmetric case with $\alpha = 0.5$, and so corresponds to MAE loss. Small forecast errors ($|e| < 1$) are costlier under MAE loss than under MSE loss, while conversely

large errors are costlier under MSE loss. The top window assumes that $\alpha = 0.25$, so negative forecast errors are three times as costly as positive errors, reflected in the steeper slope of the loss curve for $e < 0$. In the bottom window, $\alpha = 0.75$ and so positive forecast errors are three times costlier than negative errors.

2.2.1.4 Linex Loss

Linear-exponential, or Linex, loss takes the form

$$L(e) = a_1(\exp(a_2 e) - a_2 e - 1), \quad a_2 \neq 0, a_1 > 0. \tag{2.14}$$

Linex loss is differentiable everywhere, but is not symmetric. Varian (1975) used this loss function to analyze real estate assessments, while Zellner (1986a) used it in the context of Bayesian prediction problems.

The parameter a_2 controls both the degree and direction of asymmetry. When $a_2 > 0$, Linex loss is approximately linear for negative forecast errors and approximately exponential for positive forecast errors. In this case, large underpredictions ($f < y$, so $e = y - f > 0$) are costlier than overpredictions of the same magnitude, with the relative cost increasing as the magnitude of the forecast error rises. Conversely, for $a_2 < 0$, large overpredictions are costlier than equally large underpredictions.

Although Linex loss is not defined for $a_1 = 0$, setting $a_1 = 2/a_2^2$ and taking the limit as $a_2 \to 0$, by L'Hôpital's rule the Linex loss function approaches squared error loss:

$$\lim_{a_2 \to 0} L(e) = \lim_{a_2 \to 0} \frac{\exp(a_2 e) - e}{2a_2} = \lim_{a_2 \to 0} \frac{e^2 \exp(a_2 e)}{2} = \frac{e^2}{2}.$$

Figure 2.2 plots MSE loss against Linex loss for $a_2 = 1$ (top) and $a_2 = -1$ (bottom). Measured relative to the benchmark MSE loss, large positive (top) or large negative (bottom) forecast errors are very costly in these respective cases. This loss function has been used in many empirical studies on variables such as budget forecasts (Artis and Marcellino, 2001) and survey forecasts of inflation (Capistrán and Timmermann, 2009). Christoffersen and Diebold (1997) examine this loss function in more detail.

2.2.1.5 Piecewise Asymmetric Loss

A general class of asymmetric loss functions can be constructed by letting the loss function shift at a discrete set of points, $\{\bar{e}_1, \ldots, \bar{e}_{n-1}\}$:

$$L(e) = \begin{cases} L_1(e) & \text{if } e \leq \bar{e}_1, \\ L_2(e) & \text{if } \bar{e}_1 < e \leq \bar{e}_2, \\ \quad \vdots & \quad \vdots \\ L_n(e) & \text{if } e > \bar{e}_{n-1}. \end{cases} \tag{2.15}$$

Here $\bar{e}_{i-1} < \bar{e}_i$ for $i = 2, \ldots, n-1$. It is common to set $n = 2$, choose $\bar{e}_1 = 0$ and assume that both pieces of the loss function satisfy the usual loss properties so that the loss is piecewise asymmetric around 0 and continuous (but not necessarily

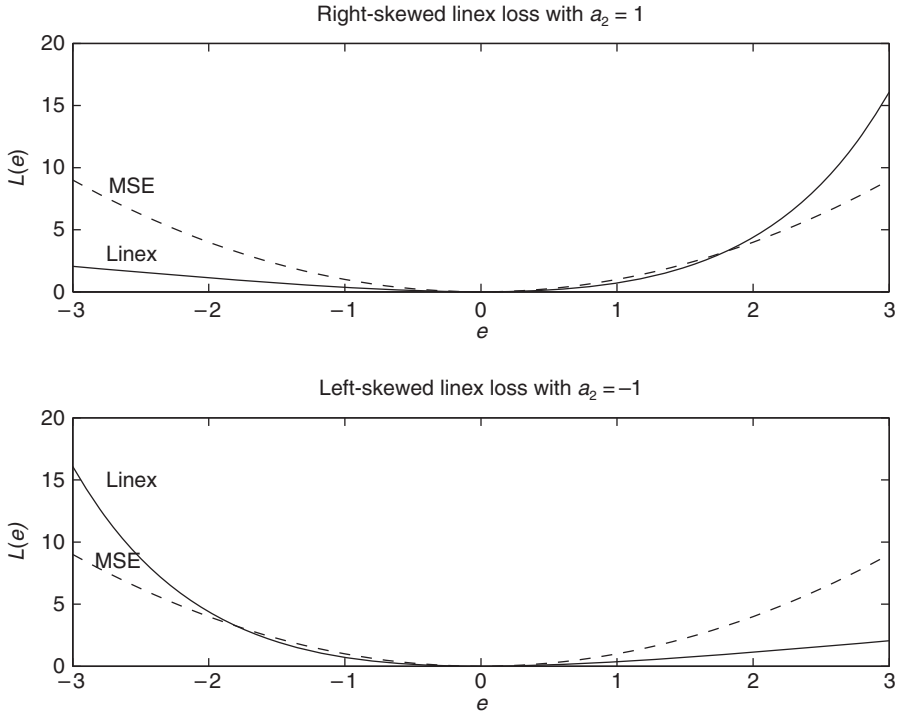

Figure 2.2: MSE loss versus Linex loss for different values of the Linex parameter, a_2.

differentiable) at 0. Lin-lin loss in (2.13) is a special case of (2.15) as is the asymmetric quadratic loss function

$$L(e) = \begin{cases} (1-\alpha)e^2 & \text{if } e \leq 0, \\ \alpha e^2 & \text{if } e > 0, \end{cases} \tag{2.16}$$

considered by Artis and Marcellino (2001), Newey and Powell (1987) and Weiss (1996).

A flexible class of loss functions proposed by Elliott, Komunjer, and Timmermann (2005) sets $n = 2$ and $\bar{e}_1 = 0$ in (2.15), while $L_1(e) = (1-\alpha)|e|^p$ and $L_2(e) = \alpha|e|^p$, where p is a positive integer, and $\alpha \in (0, 1)$. This gives the EKT loss function,

$$L(e) \equiv [\alpha + (1-2\alpha)\mathbb{1}(e < 0)]|e|^p, \tag{2.17}$$

where $\mathbb{1}(e < 0)$ is an indicator function that equals 1 if $e < 0$, otherwise equals 0. Letting α deviate from 0.5 produces asymmetric loss, with larger values of α indicating greater aversion to positive forecast errors. Imposing $p = 1$ and $\alpha = 0.5$, MAE loss is obtained. More generally, setting $p = 1$, (2.17) reduces to lin-lin loss since the loss is linear on both sides of 0, but with different slopes. Setting $p = 2$ and $\alpha = 0.5$ gives the MSE loss function which is therefore also nested as a special case, as is the asymmetric quadratic loss function (2.16) for $p = 2, \alpha \in (0, 1)$. Hence, the EKT family of loss functions nests the loss functions in (2.11), (2.12), (2.13), and (2.16) as special cases and generalizes many of the commonly employed loss functions.

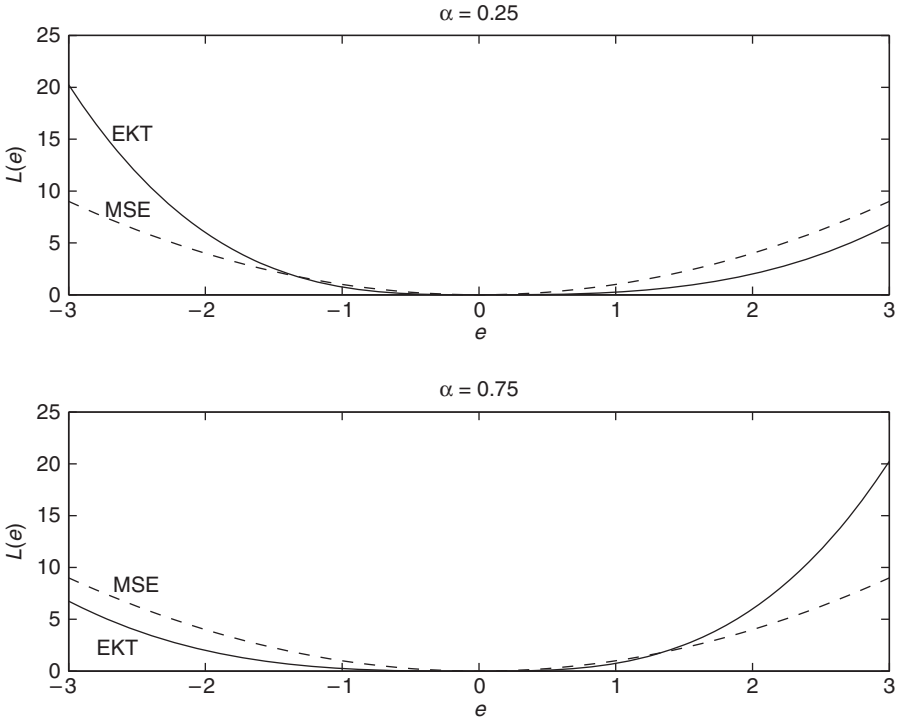

Figure 2.3: MSE loss versus EKT loss with p=3 for different values of the asymmetry parameter, α.

Figure 2.3 plots the EKT loss function for $p = 3$, $\alpha = 0.25$ (top) and $\alpha = 0.75$ (bottom). Compared with MSE loss, substantial asymmetries can be generated by this loss function.

Empirically, the EKT loss function has been used to analyze forecasts of government budget deficits produced by the IMF and OECD (Elliott, Komunjer, and Timmermann, 2005), the Federal Reserve Board's inflation forecasts (Capistrán, 2008), as well as output and inflation forecasts from the Survey of Professional Forecasters (Elliott, Komunjer, and Timmermann, 2008).

2.2.1.6 Binary Loss

When the space of outcomes $\mathcal{Y}$ is discrete, the forecast errors typically take on only a small number of possible values. Hence in constructing a loss function for such problems, all that is required is to evaluate each of a small number of possibilities. The simplest case arises when forecasting a binary outcome so that $\mathcal{Y} = \{-1, 1\}$ or $\mathcal{Y} = \{0, 1\}$. In this case there are only four possible pairings of the point forecast and outcome: two where the forecast gives the correct outcome and two errors. If we restrict the loss function to not depend on Z (this case is examined below) and also restrict the problem so that a correct forecast has the same value regardless of the

value for Y, then the binary loss function can be written as[8]

$$L(f, y) = \begin{cases} 0 & \text{if } f = y = 0, \\ (1-c) & \text{if } f = 0,\ y = 1, \\ c & \text{if } f = 1,\ y = 0, \\ 0 & \text{if } f = y = 1. \end{cases} \tag{2.18}$$

Here we have set the loss from a correct prediction to 0 and normalized the losses from an incorrect forecast to sum to 1 by dividing by their sum; see Schervish (1989), Boyes, Hoffman, and Low (1989), Granger and Pesaran (2000), and Elliott and Lieli (2013).

For (2.18) to be a valid loss function, we require that $0 < c < 1$. This ensures that the properties of the loss function listed in (2.3) hold. Notice that the binary loss function can be written as $L(e)$, since the loss is equal to

$$L(f, y) = c\mathbb{1}(e < 0) + (1-c)\mathbb{1}(e > 0).$$

2.2.2 Level- and Forecast-Dependent Loss Functions

Economic loss is mostly assumed to depend on only the forecast error, $e = Y - f$. This is too restrictive an assumption for situations in which the forecaster's objective function depends on state variables such as the level of the outcome variable Y. More generally, we can consider loss functions of the form $L(f, y) \neq L(e)$. The most common level-dependent loss function is the mean absolute percentage error (MAPE), given by

$$L(e, y) = a\left|\frac{e}{y}\right|. \tag{2.19}$$

Since the forecast and forecast error have the same units as the outcome, the MAPE is a unitless loss function. This is considered to be an advantage when constructing the sample analog of this loss function and employing it to evaluate forecast methods across outcomes measured in different units. If the loss function is well grounded in terms of the actual costs arising from the forecasting problem, dependence on units does not seem to be an important issue—comparisons across different forecasts with different units should be related not through some arbitrary adjustment but instead in a way that trades off the costs associated with the forecast errors for each of the outcomes. This is achieved by the multivariate loss functions examined in the next section.

Scaling the forecast error by the outcome in (2.19) has the effect of weighting forecast errors more heavily when y is near 0 than when y is far from 0. This is difficult to justify in many applications. Moreover, if the predictive density for Y has nontrivial mass at 0, then the expected loss is unlikely to exist, hence invalidating many of the results from decision theory for this case. Nonetheless, MAPE loss remains popular in many practical forecast evaluation experiments.

More generally, level- and forecast-dependent loss functions can be written as $L(f, y)$ but do not reduce to $L(e)$ or $L(e, y)$. Although loss functions in this class

[8] See chapter 12 for a comprehensive treatment of forecast analysis under this loss function.

are not particularly common, there are examples of their use. For example, Bregman (1967) suggested loss functions of the form

$$L(f, y) = \phi(y) - \phi(f) - \phi'(f)(y - f), \tag{2.20}$$

where ϕ is a strictly convex function, so $\phi'' > 0$. Squared error loss is nested as a special case of 2.20.

Differentiating (2.20) with respect to the forecast, f, we get

$$\begin{aligned}\frac{\partial L(f, y)}{\partial f} &= -\phi'(f) - \phi''(f)(y - f) + \phi'(f) \\ &= -\phi''(f)(y - f),\end{aligned}$$

which generally depends on both y and f. This, along with the assumption that $\phi'' > 0$, ensures that the conditional mean is the optimal forecast. Bregman loss is further discussed in Patton (2015).

In an empirical application of level-dependent loss, Patton and Timmermann (2007b) find that the Federal Reserve's forecasts of output growth fail to be optimal if their loss is restricted to depend only on the forecast error. Rationalizing the Federal Reserve's forecasts requires not only that overpredictions of output growth are costlier than underpredictions, but also that overpredictions of output are particularly costly during periods of low economic growth. This finding can be justified if the cost of an overly tight monetary policy is particularly high during periods with low economic growth when such a policy may cause or extend a recession.[9]

2.2.3 Loss Functions That Depend on Other State Variables

Under some simplifying assumptions we saw earlier that the binary loss function takes a particularly simple form. More generally, if the loss function depends on Z and the loss associated with a perfect forecast depends on the outcome Y, then the loss function for the binary problem becomes

$$L(f, y, z) = \begin{cases} -u_{1,1}(z) & \text{if } f = 1 \text{ and } y = 1, \\ -u_{1,0}(z) & \text{if } f = 1 \text{ and } y = 0, \\ -u_{0,1}(z) & \text{if } f = 0 \text{ and } y = 1, \\ -u_{0,0}(z) & \text{if } f = 0 \text{ and } y = 0, \end{cases} \tag{2.21}$$

where $u_{i,j}(z)$ are the utilities gained when $f = i$, $y = j$, and $Z = z$.

In this general form, the loss function cannot be simplified to depend only on the forecast error. Again restrictions need to be imposed on the losses in (2.21). First, we require that $u_{0,0}(z) > u_{1,0}(z)$ and $u_{1,1}(z) > u_{0,1}(z)$ so that losses associated with correct forecasts are not higher than those associated with incorrect forecasts. We might also impose that $\min\{u_{0,0}(z), u_{1,1}(z)\} > \max\{u_{1,0}(z), u_{1,0}(z)\}$ so that correct forecasts result in a lower loss (higher utility) than incorrect forecasts. Finally, it is

[9] Some central banks desire to keep inflation within a band of 0 to 2% per annum. Inflation within this band might be regarded as a successful outcome, whereas deflation or inflation above 2% is viewed as failure. Again this is indicative of a nonstandard loss function; see Kilian and Manganelli (2008).

quite reasonable to assume that correct forecasts are associated with different losses $u_{0,0}(z) \neq u_{1,1}(z)$, in which case normalizing the loss associated with a perfect forecast to 0 will not be possible for both outcomes. This is an example of level-dependent loss being built directly into the loss function.

2.2.4 Consistent Ranking of Forecasts with Measurement Errors in the Outcome

Hansen and Lunde (2006) and Patton (2011) consider the problem of comparing and consistently ranking volatility forecasts from different models when the observed outcome is measured with noise. This situation is common in volatility forecasting or in macro forecasting where the outcome may subsequently be revised. The volatility of asset returns is never actually observed although a proxy for it can be constructed. Volatility forecast comparisons typically use realized volatility, squared returns, or range-based proxies, $\hat{\sigma}^2$, in place of the true variance, σ^2.

Hansen and Lunde establish sufficient conditions under which noisy proxies can be used in the forecast evaluation without giving rise to rankings that are inconsistent with the (infeasible) ranking based on the true outcome.

Patton defines a loss function as being robust to measurement errors in the outcome if it gives the same expected-loss ranking of two forecasts whether based on the true (but unobserved) outcome or some unbiased proxy thereof. Specifically, a loss function is robust to such measurement errors if, for two forecasts f_1 and f_2, the ranking based on the true outcome, y,

$$E[L(f_1, y)] \gtreqless E[L(f_2, y)]$$

is the same as the ranking based on the proxied outcome, $\hat{y}$:

$$E[L(f_1, \hat{y})] \gtreqless E[L(f_2, \hat{y})],$$

for unbiased proxies $\hat{y}$ satisfying $E[\hat{y}|Z] = y$, where Z is again the information set used to generate the forecasts.

Patton (2011, Proposition 1) establishes conditions under which robust loss functions must belong to the following family:

$$L(f, \hat{y}) = \tilde{C}(f) + B(\hat{y}) + C(f)(\hat{y} - f), \tag{2.22}$$

where B and C are twice continuously differentiable functions, C is strictly decreasing, and $\tilde{C}$ is the antiderivative of C, i.e., $\tilde{C}' = C$.[10] In Patton's analysis f is a volatility forecast and $\hat{y}$ is a proxy for the realized volatility. Examples of loss functions in the family (2.22) include MSE and QLIKE loss:

$$\text{MSE} : L(f, \hat{y}) = (\hat{y} - f)^2,$$
$$\text{QLIKE} : L(f, \hat{y}) = \log(f) + \frac{\hat{y}}{f}.$$

[10] If $B = -\tilde{C}$, this family of loss functions yields the Bregman family in equation (2.20).

2.3 MULTIVARIATE LOSS FUNCTIONS

When a decision maker's objectives depend on multiple variables, the loss function needs to be extended from being defined over scalar outcomes to depend on a vector of outcomes. This situation arises, for example, for a central bank concerned with both inflation and employment prospects.

Conceptually it is easy to generalize univariate loss functions to the multivariate case, although difficulties may arise in determining how costly different combinations of forecast errors are. How individual forecast errors or their cross products are weighted becomes particularly important.

The most common multivariate loss function is multivariate quadratic error loss, also known as multivariate MSE loss; see Clements and Hendry (1993). This loss function maps a vector of forecast errors $e = (e_1, \ldots, e_n)'$ to the real number line and so is simply a weighted average of the individual squared forecast errors and their cross products:[11]

$$\text{MSE}(A) = e' Ae. \tag{2.23}$$

Here the $(n \times n)$ matrix A is required to be nonnegative and positive definite. This is the matrix equivalent of the univariate assumption for MSE loss that $a > 0$ in (2.11).

As noted in the discussion of MAPE loss, the loss function in (2.23) may be difficult to interpret when the predicted variables are measured in different units. This concern is related to obtaining a reasonable specification of the loss function whose role it is to compare and trade off losses of different sizes across different variables. Hence this is not really a limitation of the loss function itself but of applications of the loss function.

The loss function in (2.23) is "bowl shaped" in the sense that the level sets are convex and symmetric around 0. It is easily verified that (2.23) satisfies the basic assumptions for a loss function in (2.3). If the entire vector of forecast errors is 0, then the loss is 0. A positive-definite and nonnegative weighting matrix A ensures that losses rise as forecast errors get larger, so assumption (2.3c) holds.[12]

A special case arises when $A = I_n$, the $(n \times n)$ identity matrix. In this case covariances can be ignored and the loss function simplifies to $\text{MSE}(I_n) = E[e'e] = \text{tr}\, E[(ee')]$, i.e., the sum of the individual mean squared errors. Thus, a loss function based on the trace of the covariance matrix of forecast errors is simply a special case of the general form in (2.23). In general, however, covariances between forecast errors come into play, reflecting the cross products corresponding to the off-diagonal terms in A.

As a second example of a multivariate loss function, Komunjer and Owyang (2012) provides an interesting generalization of the Elliott, Komunjer, and Timmermann (2005) loss function in (2.17) to the case where $e = (e_1, \ldots, e_n)'$.

[11] While the vector of forecast errors could represent different variables, it could also comprise forecast errors for the same variable measured at different horizons, corresponding to short and long-horizon forecasts.

[12] Positive-definiteness alone is not sufficient to guarantee that the multivariate equivalent to (2.3) holds. Suppose $n = 2$ and let A be a symmetric matrix with 2 on the diagonals and -1 in the off-diagonal cells. A is positive definite but the marginal effect of making a bigger error on the second forecast is $4e_2 - 2e_1$, where $e = (e_1, e_2)'$. Hence if $e_2 < e_1/2$, increasing the error associated with the second forecast would reduce loss, thus violating (2.3).

Let $||e||_p = (|e_1|^p + \cdots + |e_n|^p)^{1/p}$ be the l_p norm of e and assume that the n-vector of asymmetry parameters, α, satisfies $||\alpha||_q < 1$. Further, let $1 \leq p \leq \infty$ and, for a given value of p, set q so that $1/p + 1/q = 1$. The multivariate loss function proposed by Komunjer and Owyang takes the form

$$L(e) = (||e||_p + \alpha' e)||e||_p^{p-1}. \tag{2.24}$$

As in the univariate case, the extent to which large forecast errors are penalized relative to small ones is determined by the exponent, p. However, now the full vector $\alpha = (\alpha_1, \ldots, \alpha_n)$ characterizes the asymmetry in the loss function, with $\alpha = 0$ representing the symmetric case. Since α is a vector, this loss function offers great flexibility in both the magnitude and direction of asymmetry for multivariate loss functions.

Other multivariate loss functions have been used empirically. Laurent, Rombouts, and Violante (2013) consider a multivariate version of the family of loss functions introduced by Patton (2011), and apply it to volatility forecasting.

2.4 SCORING RULES FOR DISTRIBUTION FORECASTS

So far we have focused our discussion on point forecasts, but forecasts of the full distribution of outcomes are increasingly reported. Just as point forecasting requires a loss-based measure of the distance between the forecast f and the outcome Y, distribution forecasts also require a loss function. These are known as scoring rules and reward forecasters for making more accurate predictions, i.e., predictions that are "closer" to the observed outcome get a higher score, where closeness depends on the shape of the scoring rule. Gneiting and Raftery (2007) provide a survey of scoring rules and discuss their properties.

Scoring rules, $S(p, y)$, are mappings of predictive probability distributions, p, and outcomes, y, to the real line. Suppose a forecaster uses the predictive probability distribution, p, while the probability distribution used to evaluate the "goodness of fit" of p is denoted p_0. Then the expected value of $S(p, y)$ under p_0 is denoted $S(p, p_0)$. A scoring rule is called strictly proper if the forecaster's best probability distribution is p_0, i.e., $S(p_0, p_0) \geq S(p, p_0)$ with equality holding only if $p = p_0$. In this situation there will be no incentive for the forecaster to use a probability distribution $p \neq p_0$ since this would reduce the score. The performance of a given candidate probability distribution, p, relative to the optimal rule, can be measured through the so-called divergence function

$$d(p, p_0) = S(p_0, p_0) - S(p, p_0). \tag{2.25}$$

Notice the similarity to the normalization in equation (2.3a) for loss functions based on point forecasts in (2.3): the divergence function obtains its minimum value of 0 only if $p = p_0$, and otherwise takes a positive value. The forecaster's objective of maximizing the scoring rule thus translates into minimizing the divergence function.

Several scoring rules have been used in the literature. Many of these have been considered for categorical data limited to discrete outcomes $y = (y_1, \ldots, y_m)$ with associated probabilities $\{p_1, \ldots, p_m\}$. Denote by p_i the predicted probability that

corresponds to the range that includes y_i. The logarithmic score,

$$S(p, y_i) = \log(p_i), \tag{2.26}$$

gives rise to the well-known Kullback–Leibler divergence measure,

$$d(p, p_0) = \sum_{j=1}^{m} p_{0j} \log(p_{0j}/p_j). \tag{2.27}$$

Similarly, the quadratic or Brier score,

$$S(p, y_i) = 2p_i - \sum_{j=1}^{m} p_j^2 - 1, \tag{2.28}$$

generates the squared divergence

$$d(p, p_0) = \sum_{j=1}^{m} (p_j - p_{0j})^2. \tag{2.29}$$

For density forecasts defined over continuous outcomes the logarithmic and quadratic scores take the form

$$\log S(p, y) = \log p(y),$$
$$S(p, y) = 2p(y) - \left(\int p(y)^2 \mu(dy) \right)^{1/2},$$

where $\mu(\cdot)$ is the probability measure associated with the outcome, y. Both are proper scoring rules. By contrast, the linear score, $S(p, y) = p(y)$, can be shown not to be a proper scoring rule; see Gneiting and Raftery (2007).

Which scoring rule to use in a given situation depends, of course, on the underlying objectives for the problem at hand and the choice should most closely resemble the costs involved in the decision problem. To illustrate this point, we next provide an example from the semiconductor supply chain.

Example 2.4.1 (Loss function for semiconductors)**.** *Cohen et al. (2003) construct an economically motivated loss or cost function for a semiconductor equipment supply chain. Supply firms are assumed to hold soft orders from clients which may either be canceled (with probability π) or get finalized (with probability $1 - \pi$) at some later date, y_N, when the final information arrives. Given such orders, firms attempt to optimally determine the timing of the production start, y_π, where $y_N > y_\pi$ due to a production lead-time delay. If an order is canceled, the supplier incurs a cancelation cost, c, per unit of time. Let y denote the final delivery date in excess of the production lead time. If this exceeds the production date, the supplier will incur holding (inventory) costs, h, per unit of time. Conversely, if the production start date, y_π, exceeds y, the company will not be able to meet the requested delivery date and so incurs a delay cost of g per unit of time. Cohen et al. (2003) assume that suppliers choose the production*

date, y_π, so as to minimize the expected total cost

$$E[L(y_\pi, y, y_N)] = \pi \times c \int_{y_\pi}^{\infty} (y_N - y_\pi) d P_N(y_N) + (1-\pi) \left[h \int_{y_\pi}^{\infty} (y - y_\pi) d P_y(y) + g \int_{-\infty}^{y_\pi} (y_\pi - y) d P_y(y) \right],$$

where $P_y(y)$ and $P_N(y_N)$ are the cumulative distribution functions of y and y_N, respectively. Provided that this expression is convex in y_π, the cost-minimizing production time, y_π^*, can be shown to solve the first-order condition

$$\pi \times c \times P_N(y_\pi^*) + (1-\pi)(g+h) P_y(y_\pi^*) = \pi \times c + (1-\pi)h, \tag{2.30}$$

and so implicitly depends on the cancelation probability, cancelation costs, inventory and delay costs, in addition to the predictive distributions for the finalization and final delivery dates. Cohen et al. (2003) use an exponential distribution to model the arrival time of the final order, P_N, and a Weibull distribution to model the distribution of the final delivery date, P_Y. To estimate the model parameters and predict the lead time, the authors use data on soft orders, final orders, and order lead time. Empirical estimates suggest that $\hat{g} = 1.0$, $\hat{h} = 3.0$, $\hat{c} = 2.1$, indicating that holding costs are three times greater than delay costs, while cancelation costs are twice as high as the delay costs. This in turn helps the manufacturer decide on the optimal start date for production, y_π^*.

2.5 EXAMPLES OF APPLICATIONS OF FORECASTS IN MACROECONOMICS AND FINANCE

Forecasts are of interest to economic agents only in so far as they can help improve their decisions, so it is useful to illustrate the importance of forecasts in the context of some simple economic decision problems. This section provides three such examples from economics and finance.

2.5.1 Central Bank's Decision Problem

Consider a central bank with an objective of targeting inflation by means of a single policy instrument, y_t, which could be an interest rate such as the repo rate, i.e., the rate charged on collateralized loans. Svensson (1997) sets out a simple model in which the central bank's loss function depends on the difference between the inflation rate (y_t) and a target inflation rate (y^*). Svensson shows that, conditional on having chosen a value for its instrument (the repo rate), the central bank's decision problem reduces to that of choosing a forecast that minimizes the deviation from the target. Although the forecast does not enter directly into the central bank's loss function, it does so indirectly because the actual rate of inflation (which is what the central bank really cares about) is affected by the bank's choice of interest rate which in turn reflects the inflation forecast.

Specifically, the central bank is assumed to choose a sequence of interest rates $\{i_\tau\}_{\tau=t}^{\infty}$ to minimize a weighted sum of expected future losses,

$$E_t \sum_{\tau=t}^{\infty} \lambda^{\tau-t} L(y_\tau - y^*), \tag{2.31}$$

where $\lambda \in (0, 1)$ is a discount rate and $E_t[\ ]$ denotes the conditional expectation given information available at time t. Both current and future deviations from target inflation affect the central bank's loss.

Following Svensson's analysis, suppose the central bank has quadratic loss

$$L(y_\tau - y^*) = \tfrac{1}{2}(y_\tau - y^*)^2. \tag{2.32}$$

Future inflation rates depend on the sequence of interest rates which are chosen to minimize expected future loss and hence satisfy the condition

$$\left\{i_\tau^*\right\}_t^{\infty} = \arg\min_{\{i_\tau\}_t^{\infty}} \sum_{\tau=t}^{\infty} \lambda^{\tau-t} E_t \left[(y_\tau - y^*)^2\right]. \tag{2.33}$$

Complicating matters, inflation is not exogenous but is affected by the central bank's actions. Solving (2.33) is therefore quite difficult since current and future interest rates can be expected to affect future inflation rates. Because inflation forecasts matter only in so far as they affect the central bank's interest rate policy and hence future inflation, a model for the data-generating process for inflation is needed. Svensson proposes a tractable approach in which inflation and output are generated according to the equations[13]

$$y_{t+1} = y_t + \alpha_1 z_t + \epsilon_{t+1}, \tag{2.34}$$

$$z_{t+1} = \beta_1 z_t - \beta_2(i_t - y_t) + \eta_{t+1}, \tag{2.35}$$

where z_t is current output relative to its potential level, and all parameters are positive, i.e., $\alpha_1, \beta_1, \beta_2 > 0$. The quantities ϵ_{t+1} and η_{t+1} are unpredictable shocks to inflation and output, respectively. The first equation expresses the change in inflation as a function of the lagged output, while the second equation shows that the real interest rate $(i_t - y_t)$ impacts output with a lag and also allows for autoregressive dynamics assuming $\beta_1 < 1$. Using these equations to solve for inflation two periods ahead, we obtain the following equation:

$$y_{t+2} = (1 + \alpha_1\beta_2)y_t + \alpha_1(1 + \beta_1)z_t - \alpha_1\beta_2 i_t + \epsilon_{t+1} + \alpha_1\eta_{t+1} + \epsilon_{t+2}. \tag{2.36}$$

Notice that the policy instrument (i) impacts the target variable (y) with a two-period delay. Moreover, each interest rate affects one future inflation rate and so a solution to the infinite sum in (2.33) reduces to choosing i_t to target y_{t+2}, choosing i_{t+1} to target y_{t+3}, etc. Hence, the central bank's objective in setting the current interest

[13] We have simplified Svensson's model by omitting an additional exogenous variable.

rate, i_t, simplifies to

$$\min_{i_t} E_t \left[\lambda^2 (y_{t+2} - y^*)^2\right].$$

Using the quadratic loss function in (2.32), the first-order condition becomes

$$E_t \left[\frac{\partial L(y_{t+2} - y^*)}{\partial i_t}\right] = E_t \left[(y_{t+2} - y^*)\frac{\partial y_{t+2}}{\partial i_t}\right] = 0. \tag{2.37}$$

From (2.36) this means choosing i_t so that $E_t[y_{t+2}] = y^*$, which can be accomplished by setting

$$i_t^* = y_t + \frac{(y_t - y^*) + \alpha_1(1 + \beta_1)z_t}{\alpha_1 \beta_2}. \tag{2.38}$$

It follows that the optimal current interest rate, i_t^*, should be higher, the higher the current inflation rate as well as the higher the output relative to its potential, i.e., the lower the output gap.

Under this choice of interest rate level, the argument in the loss function reduces to

$$y_{t+2} - y^* = (\epsilon_{t+1} + \alpha_1 \eta_{t+1} + \epsilon_{t+2}).$$

This is just an example of certainty equivalence, which relies heavily on the chosen squared error loss function in (2.32). If the original loss function did not have a first-order condition (2.37) that is linear in inflation, then the solution would not be so simple and the expected loss would not be a straightforward function of the expected inflation rate.

2.5.2 Portfolio Choice under Mean–Variance Utility

As an illustration of the relationship between economic utility and predictability, consider the single-period portfolio choice problem for an investor who can either hold T-bills which, for simplicity we assume pay a zero risk-free rate, or stocks which pay an excess return over the T-bill rate of y_{t+1}. Assuming that the investor has initial wealth $W_t = 1$, and letting ω_t be the portion of the investor's portfolio held in stocks at time t, future wealth at time $t + 1$, W_{t+1}, is given by

$$W_{t+1} = \omega_t y_{t+1}. \tag{2.39}$$

The portion of wealth held in stocks, ω_t, is the investor's choice variable. To analyze this decision we need to specify the investor's utility as well as how accurately y_{t+1} can be predicted.

Suppose the investor has mean–variance utility over future wealth and maximizes expected utility:

$$E[U(W_{t+1})|Z_t] = E[W_{t+1}|Z_t] - \frac{a}{2}\,\mathrm{Var}(W_{t+1}|Z_t), \tag{2.40}$$

where a captures the investor's risk aversion. The quantities $E[W_{t+1}|Z_t]$ and $\text{Var}(W_{t+1}|Z_t)$ are the conditional mean and variance of W_{t+1} given information at time t, $Z_t = \{z_1, \ldots, z_t\}$. Under this utility function, the investor's expected utility increases in expected returns but decreases in the amount of risk, as measured by the conditional variance of wealth. We can think of expected loss as the negative of (2.40).

Following Campbell and Thompson (2008), consider the following data-generating process for excess returns on stocks:

$$y_{t+1} = \mu + z_t + \varepsilon_{t+1}, \tag{2.41}$$

where $z_t \sim (0, \sigma_z^2)$, $\varepsilon_{t+1} \sim (0, \sigma_\varepsilon^2)$ and $\text{Cov}(z_t, \varepsilon_{t+1}) = 0$. Here z_t represents a potentially predictable return component which may be known at time t, while ε_{t+1} is an unpredictable shock to returns. For an investor without information on z_t, the expected value of y_{t+1} is μ, while the variance of y_{t+1} is $(\sigma_z^2 + \sigma_\varepsilon^2)$, and so

$$\omega_t^* = \underset{\omega_t}{\arg\max} \left\{ \omega_t \mu - \frac{a}{2} \omega_t^2 (\sigma_z^2 + \sigma_\varepsilon^2) \right\},$$

which implies the following optimal holdings of stocks:

$$\omega_t^* = \frac{\mu}{a(\sigma_z^2 + \sigma_\varepsilon^2)}. \tag{2.42}$$

Given this weight on stocks, the average (unconditional expectation) of the excess return on the uninformed investor's stock holdings becomes

$$E[\omega_t^* y_{t+1}] = E\left[\frac{\mu(\mu + z_t + \varepsilon_{t+1})}{a(\sigma_z^2 + \sigma_\varepsilon^2)} \right] = \frac{\mu^2}{a(\sigma_z^2 + \sigma_\varepsilon^2)} = \frac{S^2}{a}, \tag{2.43}$$

where $S = \mu / \sqrt{\sigma_z^2 + \sigma_\varepsilon^2}$ is the unconditional Sharpe ratio, i.e., the expected excess return per unit of risk (volatility). Similarly, $\text{Var}[\omega_t^* y_{t+1}] = \mu^2 / [a^2(\sigma_z^2 + \sigma_\varepsilon^2)]$, and so the expected utility, evaluated at the optimal stock holdings, ω_t^*, is

$$E[U(W_{t+1}(f_t^*))] = \frac{\mu^2}{2a(\sigma_z^2 + \sigma_\varepsilon^2)} = \frac{S^2}{2a}. \tag{2.44}$$

Turning to the informed case where the investor exploits the predictable component in stock returns, z_t, the conditional expectation and variance of future wealth are $E[W_{t+1}|Z_t] = \omega_t(\mu + z_t)$ and $\text{Var}(W_{t+1}|Z_t) = \omega_t^2 \sigma_\varepsilon^2$, respectively, and so this investor's optimal stock holding is

$$\tilde{\omega}_t^*(z_t) = \frac{\mu + z_t}{a\sigma_\varepsilon^2}. \tag{2.45}$$

This investor's expected average excess return becomes

$$E\left[\frac{(\mu + z_t)(\mu + z_t + \varepsilon_{t+1})}{a\sigma_\varepsilon^2} \right] = \frac{\mu^2 + \sigma_z^2}{a\sigma_\varepsilon^2}. \tag{2.46}$$

Noting that the predictive R^2 in (2.41) is given by $R^2 = \sigma_z^2/(\sigma_z^2 + \sigma_\varepsilon^2)$, the informed investor's average (unconditionally expected) excess return in (2.46) can be written as

$$E[\tilde{\omega}_t^*(z_t)y_{t+1}] = \frac{S^2 + R^2}{a(1 - R^2)}. \tag{2.47}$$

Comparing expected returns under the unconditional forecast in (2.43) to the expected return under the conditional forecast in (2.47), the proportional increase in the investor's expected excess returns is

$$\frac{E[\tilde{\omega}_t^*(z_t)y_{t+1}]}{E[\omega_t^* y_{t+1}]} = \frac{1 + (R^2/S^2)}{1 - R^2}, \tag{2.48}$$

whereas the simple return difference (2.47)–(2.43) amounts to $R^2(1 + S^2)/(a(1 - R^2))$.

Empirical work indicates that predictive return regressions have an R^2 close to 0, so the ratio in (2.48) is close to $(1 + R^2/S^2)$, suggesting that the magnitude of the predictive R^2 should be evaluated relative to the squared Sharpe ratio. Campbell and Thompson (2008) use historical data to estimate a squared monthly Sharpe ratio of 0.012 or 1.2%. Hence, even a monthly R^2 of "only" 0.5% would increase the average portfolio excess return by a factor 0.5/1.2, i.e., by roughly 40%. Given their historical data, this corresponds to an increase in the expected portfolio return of approximately 1.7% per annum assuming a risk aversion coefficient of $a = 3$. Even small R^2-values can thus make a considerable difference to portfolio performance in this case. Moreover, the predictive R^2 can be used as a measure of the expected return gains arising from predictability. [14]

Mean–variance investors are concerned with expected utility rather than expected returns. For uninformed investors their expected utility is given by (2.44). For informed investors, using the optimal stock holdings in (2.45), we have

$$E[\tilde{\omega}_t^*(z_t)y_{t+1}|Z_t] = E\left[\frac{(\mu + z_t)(\mu + z_t + \varepsilon_{t+1})}{a\sigma_\varepsilon^2}\middle| Z_t\right] = \frac{(\mu + z_t)^2}{a\sigma_\varepsilon^2},$$

$$\mathrm{Var}[\tilde{\omega}_t^*(z_t)y_{t+1}|Z_t] = \frac{(\mu + z_t)^2\sigma_\varepsilon^2}{a^2\sigma_\varepsilon^4} = \frac{(\mu + z_t)^2}{a^2\sigma_\varepsilon^2},$$

so that

$$E[U(W_{t+1}(\tilde{\omega}_t^*(z_t)))|Z_t] = \frac{(\mu + z_t)^2}{a\sigma_\varepsilon^2} - \frac{a}{2}\frac{(\mu + z_t)^2}{a^2\sigma_\varepsilon^2} = \frac{(\mu + z_t)^2}{2a\sigma_\varepsilon^2}.$$

The average (unconditional expectation) value of this expression is

$$E\left[E[U(W_{t+1}(\tilde{\omega}_t^*(z_t)))|Z_t]\right] = \frac{\mu^2 + \sigma_z^2}{2a\sigma_\varepsilon^2}. \tag{2.49}$$

[14] Of course, these calculations ignore transaction costs associated with portfolio turnover, as well as parameter estimation error which can be expected to be considerable.

Comparing (2.49) to (2.44), it is clear that the two are identical only when $\sigma_z^2 = 0$, otherwise (2.49)>(2.44), and the increase in expected utility due to using the predictor variable is given by

$$\text{CER} = \frac{\sigma_z^2}{2a\sigma_\varepsilon^2} = \frac{R^2}{2a(1-R^2)}.$$

This is the certainty equivalent return (CER), i.e., the additional guaranteed return which, if paid to uninformed investors, would equate their expected utility with that of investors with access to the predictor variable. Using the earlier empirical numbers, for $R^2 = 0.005$ and $a = 3$, this amounts to an annualized certainty equivalent return of about 1%.

2.5.3 Directional Trading System

Forecasters' objectives reflect their utility and action rules. As a third example, consider the decisions of a risk-neutral market timer whose utility is linear in the payoff, $U(\delta(f_t), y_{t+1}) = \delta_t y_{t+1}$, where y_{t+1} is the return on the market portfolio in excess of a risk-free rate at time $t+1$ and δ_t is the investor's holding of the market portfolio which depends on his forecast of stock returns as of time t,

$$U(\delta(f_t), y_{t+1}) = \delta_t y_{t+1}. \tag{2.50}$$

Again, we can think of the market timer's loss as the negative of (2.50). Moreover, assume that the investor follows the decision rule, $\delta(f_t)$, of going "long" one unit in the risky asset if a positive return is predicted ($f_t > 0$), otherwise going short one unit. In this case the investor's decision, $\delta(f_t)$, depends only on the predicted sign of y_{t+1}:

$$\delta(f_t) = \begin{cases} 1 & \text{if } f_t \geq 0, \\ -1 & \text{if } f_t < 0. \end{cases} \tag{2.51}$$

Trading profits depend on the sign of y_{t+1} and f_t as well as on the magnitude of y_{t+1}. To see this, let $\mathbb{1}(f_t > 0)$ be an indicator function that equals 1 if $f_t > 0$ and otherwise equals 0. Then the return from the trading strategy in (2.51) becomes

$$U(\delta(f_t), y_{t+1}) = (2\mathbb{1}(f_t > 0) - 1)y_{t+1}. \tag{2.52}$$

As one would expect from (2.51), both the sign (in relation to that of the forecast, f_t) and magnitude of excess returns, y_{t+1}, matters to the trader's utility, while only the sign of the forecast enters into the utility function. Note that large forecast errors for forecasts with the correct sign lead to smaller loss than small forecast errors for forecasts with the wrong sign.[15] This example also raises the issue of which forecast approach would be best suited given the directional trading rule. Since the trader ignores information about the magnitude of the forecast, an approach that focuses on predicting only the sign of the excess return could make sense.

[15] In related work, Elliott and Lieli (2013) derive the loss function from first principles for binary decision and outcome variables. This setup does not admit commonly applied loss functions for any possible utility function.

How the forecaster maps predictions into actions may thus be helpful in explaining properties of the observed forecasts. Leitch and Tanner (1991) studied forecasts of Treasury bill futures contracts and found that professional forecasters reported predictions with higher MSE than those from simple time-series models. At first, this seems puzzling since the time-series models presumably incorporate far less information than the professional forecasts. When measured either by their ability to generate profits or to correctly forecast the direction of future interest rate movements the professional forecasters did better than the time-series models, however. A natural conclusion to draw from this is that the professional forecasters' objectives are poorly approximated by the MSE loss function and are closer to a directional or "sign" loss function. This would make sense if investors' decision rule is to go long if an asset's excess payoff is predicted to be positive, and otherwise go short, i.e., sell the asset.

2.6 CONCLUSION

In any nontrivial forecasting situation, any forecasting method is going to make errors as the forecast will not equal the outcome with probability 1. As a consequence, forecasters need to assess the impact these errors will have on decisions based on imperfect forecasts. The loss function quantifies how costly forecast errors are for the decision maker. Formally, loss functions map the space of outcomes and decisions (forecasts) to the real number line (which is typically normalized to the nonnegative part of the real number line, although this is just a convenience) and so allow us to directly measure the economic effects of forecast errors.

In many situations the step of formally constructing a loss function that is relevant to the particular forecasting problem is skipped, and instead an informal statistic—often an intuitive function of the outcomes and forecasts assessed over a number of forecast situations or time periods—is used to evaluate forecasting performance. We showed above (using the Kuiper's score as an example) that such approaches can be difficult to interpret and often the resulting measures of loss have poor properties. This approach is therefore to be avoided.

Instead, any approach to a real decision or forecasting problem should carefully consider the relevance of the loss function to the real costs of the errors that are sure to arise when the forecasting method is put into practice. In some cases this involves constructing a loss function that is specific to a particular problem. In situations with a financial outcome that can be directly measured, this can and should be employed as the loss function. In other situations care needs to be taken to ensure that the loss function employed approximates to a reasonable extent the actual costs associated with the forecast errors.

In the next few chapters, as well as later in the book when we examine forecast evaluation, we show that loss functions matter for every step of the classical forecasting process. This includes estimation of parameters, choice of models, and the eventual evaluation of the forecasting method. It follows that the best forecasting method plausibly will depend on the choice of the loss function and a forecasting model built for one loss function may be inferior when evaluated on a different loss function. This is highly suggestive of taking seriously the step of constructing the loss function when building a forecasting model.

There are situations where it might be relevant to simply choose an "off-the-shelf" method, many of which we discuss in this chapter. Often forecasts are "intermediate" inputs provided to higher-level decision makers (e.g., the Greenbook forecasts computed by the Federal Reserve) or to the public (e.g., public weather forecasts provided by the government). Sometimes the end use is either not known with sufficient precision or the end uses are diverse enough across different agents that it is simply not possible to construct the loss function of the end user. In these cases it is typical to use mean squared error or similarly simple loss functions. This is a useful approach, although (a) perhaps a density forecast in this situation makes more sense, and (b) there are costs from not matching the forecast with the loss function. Indeed, the density forecasting approach is becoming more prevalent in these situations as weather forecasts are given as probabilities and government agencies make increasing use of fan charts, etc.