Forecasting Conditional Probabilities of Binary Outcomes under Misspecification

Graham Elliott[*] UC San Diego

Dalia Ghanem[†] UC Davis

Fabian Krüger[‡] Heidelberg Institute for Theoretical Studies

March 6, 2015

Abstract

We examine the problem of constructing density forecasts from a parametric binary choice model under a large family of loss functions ("scoring rules"). This problem is of economic interest since the scoring rules can be viewed as weighted averages over the utilities that heterogeneous decision makers derive from a publicly announced forecast (Shuford, Albert, and Massengill, 1966; Schervish, 1989). Our contribution is twofold. First, we review the scattered theoretical literature on the subject, and provide conditions that ensure the consistency of estimators implied by a variety of scoring rules. Second, we use an analytical example to illustrate how the scoring rules yield (asymptotically) identical results, if the model is correctly specified. We also show that under misspecification the choice of scoring rule may be inconsequential under specific symmetry conditions on the data-generating process, which are quite restrictive. Numerical results illustrate that if these symmetry conditions are violated, the scoring rules may lead to different parameter estimates, forecasts, and consequently decisions. Thus, by choosing a particular scoring rule, the forecaster is implicitly favoring certain decision makers over others.

[*] Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508, grelliott@ucsd.edu.

[†] Department of Agricultural and Resource Economics, University of California, Davis, and Member of the Gianini Foundation, One Shields Ave, Davis, CA 95616, dghanem@ucsdavis.edu.

[‡] Computational Statistics group, Heidelberg Institute for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany, fabian.krueger@h-its.org. The hospitality of UCSD during a visit there, as well as financial support from the Deutsche Forschungsgemeinschaft (grant "PO 375/13-1") and the European Union Seventh Framework Programme (grant agreement no. 290976) is gratefully acknowledged.

1 Introduction

Consider the problem of forecasting an as yet unobserved outcome represented by the random variable Y which takes on values $\{0, 1\}$ with a vector of observables X. The conditional probability that $Y = 1$ conditional on $X = x$, denoted by $p[x]$, is equivalent to the conditional mean and distributional forecast for the binary outcome Y. By contrast, a point forecast in this situation equals either zero or one. Point forecasting thus corresponds to choosing a binary action, and is natural if the forecaster and the decision maker are one entity (Elliott and Lieli, 2013; Lieli and White, 2010). In situations where the forecaster and the decision maker are separate entities, probability forecasts are often provided since they allow decision makers to construct their own point forecasts using their respective loss functions. Examples of such "public forecasting" scenarios include recession probabilities (e.g. forecasts by Hamilton and Chinn, 2014), sovereign default probabilities (e.g. Deutsche Bank, 2014), or probabilities of binary weather outcomes, such as rain (e.g. Mass et al., 2009).

It is common in practice to estimate the conditional probability forecast using a parametric model, such as logit or probit. To estimate this model, the forecaster must choose a loss function. If the model is correctly specified, the consistency and efficiency of the maximum likelihood estimator (MLE) justifies its choice as an estimation strategy. In practice, the model is very likely to be misspecified, and the choice of loss function will typically matter for the forecast, even asymptotically. Loss functions for distributional forecasts, such as predicted probabilities, are called "scoring rules" (e.g. Gneiting and Raftery, 2007). The Log score and the Brier score, which give rise to the MLE and nonlinear least squares, respectively, are widely used in practice. However, there are many other scoring rules that may be of interest.

A first contribution of this paper is to review the scattered theoretical literature on scoring rules for the binary case. In the context of a "double binary" decision problem (two outcomes, two possible actions), scoring rules are weighted averages over the utility functions of heterogeneous decision makers (Shuford, Albert, and Massengill, 1966; Schervish, 1989). This heterogeneity stems from the different costs that individual decision makers face under false positives versus false negatives. For instance, in forecasting the probability of currency crises (Inoue and Rossi, 2008), currency traders' costs of false positives and false negatives will vary with their degree of exposure. Another example is forecasting default probabilities of federally insured student loans (Knapp and Seaks, 1992). Lenders will tend to prefer false positives to false negatives, where the degree to which they prefer the former over the latter depends on their exposure and other loans on their menu. Student borrowers will prefer false negatives to false positives at varying degrees depending on their neediness, the loan amount, and how much they value their future credit history.

Popular scoring rules are based on a certain symmetry in these weighted averages. This symmetry may or may not be appropriate for a given empirical problem. Asymmetric rules relate to situations in which false negatives are much more costly than false positives or vice versa. For example, Lieli and Springborn (2013) analyze the environmental policy decision of whether or not to admit possibly invasive biological imports. From the consumer's point of view, it may be devastating to mistakenly classify an import as safe; in contrast, classifying a safe import as unsafe is typically less harmful. From the importer's perspective, however, a false positive is much more costly than a false negative. A symmetric scoring rule, such as the Log Score, weighs the consumers' and importers' utility functions equally, which may not be appropriate if the safety of the general public is at stake. There are many settings where symmetry may not be justified, such as forecasting natural disasters (say, wildfires or earthquakes) and economic disasters (say, recessions).

Despite their empirical relevance, such asymmetries are not reflected in common rules such as the Log or Brier score. We hence show how to construct proper asymmetric scoring rules in the spirit of Buja, Stuetzle, and Shen (2005). This involves prioritizing some decision makers over others, and hence resembles the aggregation of utilities in a social planner's problem (c.f. Lieli and Nieto-Barthaburu, 2010). In order to actually use a wide variety of scoring rules for parameter estimation, convergence of the resulting estimators is a key concern. We therefore provide conditions for a weak law of large numbers, allowing for time series dependence in the data.

Under misspecification, different choices of scoring rules may lead to different estimates of the forecast model. Matching the scoring rule used for parameter estimation with the one used for forecast evaluation has been recommended in the literature for this reason. However, there has been less work on examining the magnitude of these effects. In (non-binary) MSE-based forecasting, Weiss and Andersen (1984) make this point with respect to using autoregressions as forecast models. Granger (1993) makes this suggestion without elaboration, whilst Weiss (1996) makes the point more generally giving results. Hand and Vinciotti (2003) make the same suggestion in examining models for binary forecasting, while Gneiting (2011) and Patton (2014) consider several types of point forecasts.

We contribute to this literature by providing novel analytical and numerical evidence for the binary case. With the aid of an analytical example, we illustrate how the choice of scoring rule is inconsequential in the case of correct specification. In the case of misspecification, we characterize the conditions under which the choice is inconsequential. These conditions consist of specific symmetry conditions on the data-generating process that are likely to be violated in practice. A Monte Carlo study illustrates that if a subset of the conditions is violated, the choice of scoring rule affects parameter estimates, forecasts and – most importantly – decisions. While these effects are qualitatively robust, their magnitude is necessarily case specific and depends on factors such as the true DGP, the set of scoring rules being compared, and the preferences of the decision maker. These preferences determine whether or not the differences between two predicted probabilities (say, obtained under scoring rules A and B) are relevant in the sense of leading to different decisions.

The plan of this paper is as follows. The following section presents some theoretical results. Section 2.1 reviews results that link the scoring rule and the decision maker loss functions. We present results for the consistency of estimators using a wide variety of scoring rules for conditional probability forecasting in Section 2.2. Section 2.3 illustrates why matching loss functions for estimation and evaluation may be important in the case of misspecification with the aid of an analytical example. Section 3 provides a Monte Carlo demonstration of the theoretical points. Section 4 concludes.

2 Scoring Rules, Estimation and Evaluation

2.1 Characterization of scoring rules

Consider the problem of forecasting a binary random variable Y with outcomes $y \in \{0,1\}$ given some predictors denoted by the random variable(s) X with outcomes $x \in \mathbb{X}$, where $\mathbb{X}$ denotes the support of X.[1] We will denote by $p[X] \in P$ models of the conditional probability that $Y = 1$, where $p_0[X]$ is the correctly specified model. We use square brackets $p[x]$ to distinguish from the notation $p(x)$. The latter notation implies that p is a function defined on the support of X. For p defined on the support of a linear index $x'\theta$, $p[x] = p(x'\theta)$. The square brackets hence allow us to subsume θ. We do not assume that $p_0[X] \in P$. For notational brevity, we will often refer to p and p_0 instead of $p[\cdot]$ and $p_0[\cdot]$. A decision maker chooses a function $f(X)$ from the space of X to $\{0,1\}$. The optimal choice of this function depends on both the conditional probability that $Y = 1$ and the utility function of the decision maker. The decision maker's utility function has the form

$$U(y,f,c) = \left\{ \begin{array}{cc} 0 & \text{if } f=1 \text{ and } y=1 \\ -c & \text{if } f=1 \text{ and } y=0 \\ -(1-c) & \text{if } f=0 \text{ and } y=1 \\ 0 & \text{if } f=0 \text{ and } y=0 \end{array} \right\} \tag{1}$$

where $0 < c < 1$. Now the utility function can be re-written as follows

$$\begin{aligned} U(y,f,c) &= -c1(y-f=-1) - (1-c)1(y-f=1) \\ &= -c(1-y)f - (1-c)y(1-f), \end{aligned} \tag{2}$$

where $1(A)$ denotes the indicator function of the event A. Note that the utility function is normalized such that a correct decision yields zero utility. The utilities for incorrect decisions are normalized to sum to one in absolute value. This is without loss of generality when $U(y,f)$ depends only on these two outcomes.[2] Thus, $c = 0.5$ indicates a decision maker's indifference between false positives ($f = 1$ and $y = 0$) and false negatives ($f = 0$ and $y = 1$).

To motivate the problem, we will use the example of forecasting a storm at a coastal location. We

[1] We assume that X is observable and does not include lagged values of Y.

[2] Elliott and Lieli (2013) consider the more general case in which utility depends not only on the realization y and point forecast f, but also on other state variables such as measured covariates x. In this case, the normalizations we use here are not available, and the utility function takes a more complicated form.

will consider two types of decision makers, the local restaurant owners and fishermen. Let y denote whether a storm takes place or not. For a restaurant owner, if $f = 1$, the restaurant owners will allocate fewer staff members, since it expects to be serving fewer customers. If $f = 0$, the restaurant will hire its full staff. A fishermen will only go fishing if $f = 0$. We expect restaurant owners to prefer false negatives to false positives, i.e. $c \geq 0.5$. In the case of a false negative, tourists will be visiting the coastal location expecting good weather. Since a storm occurs, they will be spending more time at restaurants. Restaurant owners will have hired their full staff, and hence would be prepared to serve a lot of customers. In the case of a false positive, the restaurant does not hire additional staff and fewer tourists will be visiting the location. Thus, the restaurant owners' profits will be smaller. The fishermen, on the other hand, are likely to prefer false positives to false negatives, i.e. $c \leq 0.5$. In the case of a false positive (staying at home, but no storm), even though they lose the catch, they save on fuel. In the case of a false negative (going fishing when there is a storm), however, they may lose their equipment or even put their own life in danger. In reality, we have a continuum of heterogeneous fishermen and restaurant owners with different values of c. The exact value $c \in [0.5, 1]$ of a fisherman's utility takes will depend on the value of his/her equipment and the number of staff on his/her crew. Similarly, a restaurant owner's $c \in [0, 0.5]$ will depend on the restaurant size, menu and how much additional staff he/she hires.

Optimal forecasts for this problem are to set $f(x) = 1(p_0[x] > c)$, see Schervish (1989), Boyes, Hoffman, and Low (1989) and Granger and Pesaran (2000). This result assumes the knowledge of the true conditional probability. In practice, the unknown true probability $p_0[x]$ is replaced by an estimate $p[x]$ which can be frequentist (as in our analysis below) or Bayesian (c.f. Lieli and Springborn, 2013, Section 2). The utility function can now be written as

$$U(y, p, c) = -y(1 - c)1(p \leq c) - (1 - y)c1(p > c). \tag{3}$$

It is important to note that the parameter c plays a dual role in the above equation. In addition to determining the decision maker's preference over false positives and negatives, c is part of the optimal forecasting rule and determines how the decision maker interprets a probability forecast. If $p \leq c$, then the decision maker will interpret it as $f = 0$, otherwise the decision maker will interpret it as $f = 1$. Coming back to our example, consider a fisherman with $c = 0.25$, and a restaurant owner with $c = 0.75$. In this case, the optimal forecasting rule has the following implications. If $p < 0.25$, then neither the fisherman nor the restaurant owner will interpret the probability forecast to indicate that a storm will occur. If $0.25 \leq p < 0.75$, then only the fisherman will interpret the probability forecast to indicate that $f = 1$ and will not go fishing. If $p \geq 0.75$, then both the restaurant owner and the fisherman will interpret the probability forecast to indicate that $f = 1$. This shows how the preference over false positives and negatives informs the interpretation of the conditional probability forecast under the optimal forecasting rule.

To construct a forecast, we require an estimate of the true conditional probability of $Y = 1$, i.e. an estimate of $p_0[x]$ or a procedure that directly estimates $1(p_0[x] > c)$. Manski and Thompson (1989) and Elliott and Lieli (2013) examine the latter approach and show how direct estimation lessens the need for an exact understanding of the true conditional probability. Essentially, the function $1(p_0[x] > c)$ is easier to specify correctly than $p_0[x]$ since the former is a step function. Our example shows why we might estimate the conditional probability instead, since it gives the individual decision makers, i.e. the restaurant owners and fishermen, the opportunity to interpret the forecast in a manner that is optimal based on their own preferences. In general, when there are users with a range of utility functions, i.e. values for c, then provision of an estimate for $p_0[x]$ enables all users to construct their own forecast rules, see e.g. Lieli and Nieto-Barthaburu (2010).

When constructing a model p for the conditional probability, we require a scoring rule for estimation. By definition, a proper scoring rule $S(y, p)$ is a function for which $E[S(y, p)]$ is finite and maximized at $p = p_0$. It is considered to be a strictly proper scoring rule if this maximum is unique, i.e. the rule is maximized only at the true value for the probability, see e.g. Gneiting and Raftery (2007). From an econometric perspective, this means that the conditional probability is identified by the scoring rule. For binary outcomes, all proper scoring rules have the form

$$S(y, p) = yf_1(p) \ + (1 - y)f_2(p). \tag{4}$$

Schervish (1989, Theorem 4.2) shows that proper scoring rules can be seen as weighted averages of many utility functions, where the weights are over different cutoff values c. Denote by $\nu(c)$ a nonnegative weighting function over c defined on $[0, 1]$. By integrating the utility for a single decision maker in (3), we obtain the weighted average utility function

$$S(y, p) = \ -y \int_0^1 (1 - c)1(p \leq c)\nu(c)dc - (1 - y) \int_0^1 c1(p > c)\nu(c)dc. \tag{5}$$

$\nu(c)$ may be viewed as the density of the preference parameter c in a population of decision makers that the forecaster seeks to inform. Hence, it has an intuitive economic interpretation.

Equating (4) and (5), we see that $f_1(p) = -\int_0^1 (1 - c)1(p \leq c)\nu(c)dc$ and $f_2(p) = -\int_0^1 c1(p > c)\nu(c)dc$. As shown by Schervish, scoring rules with this form are strictly proper if $\nu(c)$ gives a nonzero weighting over all $c \in [0, 1]$. These results are useful in a number of ways. First, through specification of $\nu(c)$, they provide a constructive approach to designing scoring rules. Second, for existing scoring rules, they show the implicit weighting over decision makers' utility functions that underlie the construction of that particular scoring rule. Table 1, which extends Table 1 of Gneiting and Raftery (2007), gives several scoring rules and their implicit weights, $\nu(c)$. This includes popular approaches. Notice that the Log scoring rule is simply (pseudo) maximum likelihood for a parametrized model of $p[x]$. This is the most common approach to parametrically estimating models of the conditional probability, where the models are typically either logit or probit. See Lieli and Nieto-Barthaburu (2010) for an economic interpretation of the weighting that underlies maximum likelihood.

It is also possible through defining $f_1(p)$ to provide a positive approach to constructing proper scoring rules. If $f_1(p)$ and $f_2(p)$ are once differentiable such that $\partial f_1(p)/\partial p > 0$ for $p \in (0,1)$ and

$$\frac{\partial f_2(p)}{\partial p} = -\frac{\partial f_1(p)}{\partial p}\left(\frac{p}{1-p}\right), \tag{6}$$

then $S(y,p)$ in (4) is a proper scoring rule with

$$\nu(c) = \frac{\partial f_1(c)}{\partial c}\left(\frac{1}{1-c}\right).$$

This result was obtained by Shuford, Albert, and Massengill (1966), restated in Theorem 4.1 of Schervish (1989). Notice that this relates $f_1(p)$ to $f_2(p)$ through their slopes at any p. Hence, one can construct a proper scoring rule by defining $f_1(p)$ and constructing $f_2(p)$ using this restriction. For example, set $f_1(p) = p$, so $\partial f_1(p)/\partial p = 1 > 0$ for all p. Now $\nu(c) = (1-c)^{-1} > 0$ for $c \in (0,1)$. Using this $\nu(c)$ results in a proper scoring rule. To obtain $f_2(p)$, we solve

$$\begin{aligned} f_2(p) &= -\int_0^p c\nu(c)dc \\ &= -\int_0^p \frac{c}{1-c}dc. \end{aligned}$$

Integrating we obtain $f_2(p) = p + \ln(1-p)$ and hence

$$S(y,p) = yp + (1-y)[p + \ln(1-p)].$$

This is a strictly proper scoring rule. It is also worth mentioning that convex combinations of strictly proper scoring rules are also strictly proper.[3]

For either of these directions, it is an outcome of the process that we obtain an understanding of $\nu(c)$, the weights over the individual decision makers. The weighting functions for various popular scoring rules given in Table 1 are pictured in Figure 1. We see that the Log score and Boosting loss correspond to U-shaped weighting functions, each placing very similar weights over c. The weighting functions of the Brier and Spherical score are flat and bell-shaped, respectively. All of the popular rules are symmetric around $c = 0.5$, the point of indifference between false negatives and false positives. There is no obvious reason why this might be appropriate in general for situations where distributional forecasts are to be provided. In our simplified example of forecasting storms, where we have the restaurant owners ($c \geq 0.5$) and the fishermen ($c \leq 0.5$), the forecaster may prefer to weigh the restaurant owners' and fishermen's utility functions, e.g. according to their proportion in the region's population or based on economic revenues. Thus, this weighting may have a social,

[3] To see this, consider the example of a convex combination of the Log score and As1 score, which we use as an example in Section 2.3. Each of these scores is maximized in expectation by the true probability. Hence, any convex combination of the two is also maximized in expectation by the true probability, and thus defines a strictly proper scoring rule itself. Alternatively, note that a convex combination of the Log and As1 scores again satisfies the relationship in (6), and thus inherits strict propriety.

political or economic motivation. This paper is not concerned with the justification of a particular weighting scheme over another. Our goal is to show that the choice of the weighting function and thereby the scoring rule may have consequences on conditional probability forecasts and individual decision making in practice.

Buja, Stuetzle, and Shen (2005) and Merkle and Steyvers (2013) use the beta distribution to parametrize the weighting function $\nu(c)$. This leads to a flexible two-parameter family of scoring rules. A somewhat simpler approach is to directly choose a given shape for $\nu(c)$. This is exemplified by the As1 and As2 scoring rules shown above. In the first of these rules we set $\nu(c)$ so that it heavily weights small values of c relative to large values. This would be a situation where forecasters that are extremely averse to losses from false negatives are heavily weighted. The specification of As2 does the reverse of this, heavily weighting forecasters who are heavily averse to false positives.[4] By specifying $\nu(c)$ directly according to a reasonable weighting function, the results presented above allow us to construct economically meaningful scoring rules that are strictly proper. This situation, which draws on the existence of the Schervish (1989) representation for scoring rules, helps to bridge the gap between economic and statistical forecast evaluation criteria.

Name	$f_1(p)$	$f_2(p)$	$\nu(c)$	Source
Log score	$\ln(p)$	$\ln(1-p)$	$[c\,(1-c)]^{-1}$	Good (1952)
Half Brier Score	$-\frac{1}{2}\,(1-p)^2$	$-\frac{1}{2}\,p^2$	1	Brier (1950)
Spherical Score	$\frac{p}{\sqrt{1-2p+2p^2}}$	$\frac{1-p}{\sqrt{1-2p+2p^2}}$	$(1-2c+2c^2)^{-3/2}$	Toda (1963)
Boosting	$-\sqrt{\frac{1-p}{p}}$	$-\sqrt{\frac{p}{1-p}}$	$\frac{1}{2}[c\,(1-c)]^{-\frac{3}{2}}$	Buja, Stuetzle, and Shen (2005)
As1	$\ln(p)-p+1$	$-p$	c^{-1}	
As2	$p-1$	$p+\ln(1-p)$	$(1-c)^{-1}$	Gneiting and Raftery (2007)

Table 1: Summary of the scoring rules we consider, see Equations (4) and (5) for details.

2.2 Scoring rules and parameter estimation

For all of the choices of proper scoring rules, one can consider estimating $p[x]$ using the scoring rule as a loss function. Consider linear index models, i.e. $p[x] = p(x'\theta)$. With data $\{y_t, x_t\}_{t=1}^{T}$, we can consider estimating the parameters of the model from the maximization

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \sum_{t=1}^{T} S(y_t, p(x_t'\theta)).$$

[4] Gneiting and Raftery (2007, Example 5) show that the As2 rule is also a member of the Buja, Stuetzle, and Shen (2005) family of scoring rules.

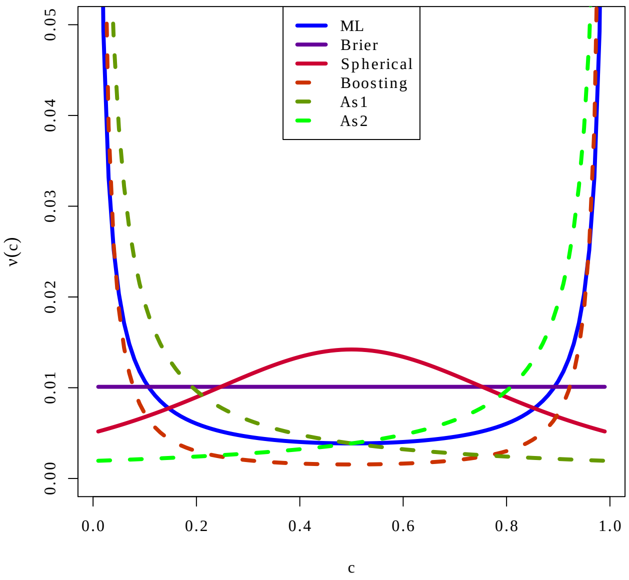

Figure 1: Weighting functions $\nu(c)$ for the scoring rules in Table 1. All functions have been multiplied by a scaling factor for comparability.

For example, for the Log scoring rule and $p(x_t'\theta) = e^{x_t'\theta}/(1+e^{x_t'\theta})$, this would be maximum likelihood of the logit model. For the same model with the half Brier score as the scoring rule, this would be nonlinear least squares estimation of the logit model. Various combinations of scoring rules and models could be employed to obtain a parameter estimate $\hat{\theta}$, and from this an estimate of the conditional probability that the outcome is one, $p(x'\hat{\theta})$ for any possible x. Under fairly general conditions, $\hat{\theta} \xrightarrow{p} \theta^*$, where

$$\theta^* = \arg\max_{\theta\in\Theta} E[S(y_t, p(x_t'\theta)].$$

The following theorem provides a set of conditions for achieving this consistency result. As detailed below, the theorem can be seen as a special case of M-estimation (e.g. Wooldridge, 1994, Section 4), adapted to the situation of using strictly proper scoring rules and binary models.

Theorem 1 *Assume*

(a) $S(.)$ *is a strictly proper scoring rule.*

(b) $\theta \in \Theta \subset \mathbb{R}^k$*, where* Θ *is compact.*

(c) For each $y_t \in \{0,1\}, x_t \in X,\ f_1(p), f_2(p)$ *are measurable and differentiable in* p *and* $p(x_t,\theta)$ *is measurable and differentiable in* θ.

(d) $E|f_i(p(x_t'\theta))|^{r+\delta} < \Delta < \infty$ *for* $i = 1, 2,\ \delta > 0,\ r \geq 1$ *and* $\sup_{x_t\in\mathbb{X}} \sup_{\theta\in\Theta} \left|(1-p(x_t'\theta))^{-1} f_1'(p(x_t'\theta))p'(x_t'\theta)\right| < K < \infty.$

(e) $\{y_t, x_t\}$ *are strictly stationary mixing processes with uniform mixing size* $-r/(2r-1)$ *with* $r \geq 1$ *or strong mixing of size* $-r/(r-1)$*,* $r > 1$.

(f) $E\left\|x_t\right\|^{r+\lambda} < \Delta_1 < \infty$ *for* $r \geq 1$ *and* $\lambda > 0$.

Then $\hat{\theta} \xrightarrow{p} \theta^*$ *where* $\theta^* = \max_{\theta\in\Theta} E[S(y_t, p(x_t'\theta))]$.

Proof The proof follows from using these conditions to show the conditions of Theorems 4.2 and 4.3 of Wooldridge (1994) hold. First, via Theorem 4.2, the conditions are sufficient so that $\max_{\theta\in\Theta}\left|T^{-1}\sum_{t=1}^{T} S(y_t, p(x_t'\theta)) - E[S(y_t, p(x_t'\theta))]\right|\,| \xrightarrow{p} 0$, so the objective function converges to its expected value uniformly in θ. Conditions (b) and (c) yield the theorem's (i) and (ii). For Theorem 4.2 part (iv), notice that for all $\theta_1, \theta_2 \in \Theta$, the mean value theorem and Equation (6) yield that

$$\begin{aligned}
\left|S(y_t, p(x_t'\theta_1)) - S(y_t, p(x_t'\theta_2))\right| &= \left| f_1'(p(x_t'\tilde{\theta}))p'(x_t'\tilde{\theta})\left(\frac{y_t - p(x_t'\tilde{\theta})}{1 - p(x_t'\tilde{\theta})}\right) x_t'(\theta_1 - \theta_2)\right|, \\
&\leq \left\| f_1'(p(x_t'\tilde{\theta}))p'(x_t'\tilde{\theta})\left(\frac{y_t - p(x_t'\tilde{\theta})}{1 - p(x_t'\tilde{\theta})}\right) x_t\right\| \left\|\theta_1 - \theta_2\right\|, \\
&= \left| f_1'(p(x_t'\tilde{\theta}))p'(x_t'\tilde{\theta})\left(\frac{y_t - p(x_t'\tilde{\theta})}{1 - p(x_t'\tilde{\theta})}\right)\right| \left\|x_t\right\| \left\|\theta_1 - \theta_2\right\|, \\
&\leq \left| f_1'(p(x_t'\tilde{\theta}))p'(x_t'\tilde{\theta})\left(\frac{1}{1 - p(x_t'\tilde{\theta}))}\right)\right| \sqrt{x_t'x_t} \left\|\theta_1 - \theta_2\right\|, \\
&\leq K\sqrt{x_t'x_t} \left\|\theta_1 - \theta_2\right\|,
\end{aligned}$$

where $\tilde{\theta}$ is an intermediate value for θ. Parts (iii) and (iv b) require that $S(y_t, p(x_t'\theta))$ and $\sqrt{x_t'x_t}$ satisfy a WLLN pointwise in θ. These results follow from assumptions (d) through (f), which are sufficient for a WLLN via Corollary 3.48 of White (2001). Notice that $E[S(y_t, p(x_t'\theta))]^q$ for any integer q is equal to $E[y_t f_1(p(x_t'\theta))^q + (1 - y_t)f_2(p(x_t'\theta))^q]$ which is finite given (d) for $q \leq r + \delta$. A WLLN for $\sqrt{x_t'x_t}$ follows directly from assumptions (e) and (f). Consistency of the estimate $\hat{\theta}$ follows from the conditions being sufficient for Theorem 4.3 of Wooldridge (1994). Condition M1 and M2 follow directly from (b) and (c) along with the results presented for uniform convergence of the average of the objective function. Condition M3 follows directly from assumption (a). □

Assumption (a) ensures that the scoring rule is strictly proper, and thus admits the decomposition in Equation (4). Furthermore, it ensures that the objective function attains a unique maximum at θ^*. If $S(.)$ is relaxed to be a proper scoring rule that is not necessarily strictly proper, then a result similar to Theorem 1 still holds. In this case, $E\left[S(y_t, p(x_t'\hat{\theta})\right] \xrightarrow{p} \max_{\theta\in\Theta} E\left[S(y_t, p(x_t'\theta)\right]$ which is sufficient to justify the procedure. Assumption (b) is the standard requirement to ensure a maximum. Assumption (c) is a standard regularity condition. The conditions in (d) relate to the scoring rule and model being employed, and are functions of both these choices. The first part of (d) ensures that expected loss exists. The second part is employed as part of the requirements for uniform consistency of the objective function. This assumption seems strong, however it holds widely since these objects are functions of $p(x_t'\theta)$ which is bounded between zero and one for all x_t and θ. For notational brevity, let $s_t = x_t'\theta$. For example, consider the half Brier score with a logit model for the conditional probability that $y_t = 1$. Then $f_1'(p(s_t)) = 1 - p(s_t)$ and

$p'(s_t) = p(s_t)(1-p(s_t))$ so $\left|(1-p(s_t))^{-1}f_1'(p(s_t))p'(s_t)\right| = |p(s_t)(1-p(s_t))| \leq 0.5$, hence the second part of (d) is satisfied. In this case $E|f_1(p(s_t))|^{r+\delta} = E|-0.5(1-p(s_t))^2|^{r+\delta} \leq 0.5^{r+\delta}$ and so is finite for all r, δ finite. For the Log score with a logit model for the conditional probability, we have that $f_1'(p(s_t))p'(s_t) = (1-p(s_t))$ and so $\left|(1-p(s_t))^{-1}f_1'(p(s_t))p'(s_t)\right| = 1$. For the Spherical scoring rule, $f_1'(p(s_t)) = (1-p(s_t))/((p(s_t))^2 + (1-p(s_t))^2)^{3/2}$. Hence, $\left|(1-p(s_t))^{-1}f_1'(p(s_t))p'(s_t)\right| = \left|(p(s_t)(1-p(s_t)))/((p(s_t))^2 + (1-p(s_t))^2)^{3/2}\right| \leq 2^{-1/2}$ and so this is also bounded.

The mixing assumptions in (e) impose a limit on the degree of time series dependence in the data. The requirement of strict stationarity gives meaning to the idea that we obtain the true conditional probability, at least asymptotically, when the model is correctly specified. If the data are not strictly stationary, then we can still obtain consistency results, however the interpretation of θ^* changes to being a limiting value that minimizes the average expected losses over time. It is worth noting that some strictly proper scoring rules, such as the Half Brier and Spherical scores considered in this paper, are bounded. For bounded scoring rules, the assumptions of Theorem 1 can be relaxed. First, the technical requirements of assumption (d) either become obsolete or are trivially satisfied. Second, assumption (f), which ensures a WLLN for $\sqrt{x_t'x_t}$, is not required.[5]

Conceptually, the main purpose of Theorem 1 is to illustrate that the structure of scoring rules makes them well suited for designing (consistent) parameter estimators in the tradition of M-estimators (e.g. Hayashi, 2000, Chapter 7), see also Gneiting and Raftery (2007, Section 9.1). There are many possible sets of assumptions (e.g. various ways of restricting time series dependence) that could lead to the statement of Theorem 1. Our chosen set of assumptions aims to strike a balance between generality and clarity of presentation, although results under alternative trade-offs between conditions are possible. Furthermore, results under more primitive conditions are available in more specialized settings. For example, de Jong and Woutersen (2011) analyze consistency in the important special case of a correctly specified probit model with lagged dependent variables, estimated via maximum likelihood. See their Theorems 1 and 2 for low-level conditions that guarantee limited dependence properties of the data-generating process, and their Theorem 3 on consistency.

In terms of understanding the results for forecasting binary outcomes, two results follow directly. First, if the model is correctly specified, i.e. $p_0[x_t] = p(x_t'\theta_0)$, then for all strictly proper scoring rules $\theta^* = \theta_0$. Hence, all strictly proper scoring rules will give the same true conditional probability asymptotically. This follows directly from the fact that strictly proper scoring rules are uniquely maximized by the true conditional probability. Thus, if the model is correctly specified, the choice of the best scoring rule to use depends not on the reasonableness of θ^* but instead on the efficiency of the estimator $\hat{\theta}$ obtained by maximizing a particular scoring rule. The popularity of the MLE derives

[5] To see this, note that if $\sup_{\theta\in\Theta} |S(y_t, p(x_t'\theta))| < C$, it follows that $|S(y_t, p(x_t'\theta_1)) - S(y_t, p(x_t'\theta_2))| < 2C$. This means that the stochastic upper bound in the proof of Theorem 1 can be replaced by a constant. Hence, a WLLN for the upper bound is trivially satisfied, and Assumption (f) (which ensures a WLLN for $\sqrt{x_t'x_t}$) becomes obsolete. Furthermore, the second part of Assumption (d), which is used in the general case to bound the term $|S(y_t, p(x_t'\theta_1)) - S(y_t, p(x_t'\theta_2))|$, is no longer required. Finally, the first part of Assumption (d) is automatically satisfied since the $f_i(p(x_t'\theta)),\ i = 1, 2,$ are bounded.

from the fact that it is an efficient parameter estimator *under correct specification*. However, it should be understood that the latter strong assumption is crucial in establishing the optimality of the MLE.

When the model is not correctly specified, then there is no reason that θ^* should be the same over different scoring rules. In practice, they will differ, and then so will the estimated conditional probability even asymptotically. Hence, decisions made for any particular loss function for the decision maker (value for c) will also differ across scoring rules. Scoring rules placing more weight on high values of c will provide probability forecasts which are most useful for decision makers with high values of c, and vice versa. In order to attain (asymptotic) optimality, the scoring rule chosen for estimating the parameters of the model should match the scoring rule used to evaluate the probability forecast. Under the conditions above, the magnitude of this effect depends on how θ^* varies with the choice of the scoring rule. Since both scoring rules and models tend to be very nonlinear, this relationship will generally be complex. The analytical example in Section 2.3 and the Monte Carlo results in Section 3 provide evidence on this issue.

In choosing between scoring rules, a forecaster needs to trade off the loss from using a scoring rule other than the Log score under correct specification with the gains this approach brings when the model is misspecified. The first consideration then is how plausible it would be to assume that the model is correctly specified. In most applications, especially those unmotivated by any underlying economic or scientific theory, this would be a difficult assumption to make. Nonetheless, it would generally be considered, and the answer is specific to the forecasting problem. The second consideration is how large the gains are from using the matching strategy under misspecification of the model.

2.3 Misspecification: An analytical example

In order to illustrate how the choice of scoring rule matters, we will give an analytical example, where we examine the effect of trading off between two specific scoring rules on θ^*. The scoring rules we consider are given in Table 1, the Log and As1 scores. The former is the Log-likelihood, the latter is an asymmetric scoring rule that emphasizes a better fit for smaller probabilities versus larger probabilities. We can write a composite scoring rule indexed by $\lambda \in [0, 1]$ that nests both of them,

$$\begin{aligned} S_\lambda(y, p(x'\theta)) &= y\ln(p(x'\theta)) + y\lambda(1 - p(x'\theta)) \\ &+ (1-y)(1-\lambda)\ln(1-p(x'\theta)) + (1-y)(-\lambda)p(x'\theta). \end{aligned} \tag{7}$$

For $\lambda = 0$, $S_0(y, p(x'\theta))$ gives the Log score. For $\lambda = 1$, $S_1(y, p(x'\theta))$ is the As1 score. For all $\lambda \in [0, 1]$, $S_\lambda(y, p(x'\theta))$ is a proper scoring rule, since propriety carries over to convex combinations of two proper scoring rules as mentioned above.

Let θ^* denote the maximizer of the objective function defined by the scoring rule, with

$$\begin{aligned}\theta^* &= \arg\max_{\theta\in\Theta} E_{X,Y}\left\{S_\lambda(Y, p(X'\theta))\right\} \\ &= \arg\max_{\theta\in\Theta} E_X\left\{E_{Y|X}\left(S_\lambda(Y, p(X'\theta))\right)\right\}.\end{aligned}$$

We can write the conditional expectation in the above objective function as follows,

$$E_{Y|X}\left(S_\lambda(Y, p(X'\theta))\right) = p_0[X]\ln(p(X'\theta)) + \lambda p_0[X] + (1-p_0[X])(1-\lambda)\ln(1-p(X'\theta)) - \lambda p(X'\theta).$$

For simplicity, we assume in the following that X is scalar. As demonstrated in Appendix A.2, extending the example to include an intercept is possible but appears to complicate the analysis without a compensating gain in insight.

The probit and logit link functions $p(.)$ are by far the most common choices in the literature.[6] Koenker and Yoon (2009) survey several other choices. Here, we do not make specific assumptions about p, except that it does not depend on estimands other than θ. Assuming sufficient regularity conditions to apply the dominated convergence theorem (DCT), the first-order condition for a maximum is

$$\int_{\mathbb{X}} \left.\frac{\partial E_{Y|X}\left(S_\lambda(Y, p(X'\theta))\right)}{\partial\theta}\right|_{X=x} f_X(x)\,dx \quad = \quad 0,$$

where $f_X(.)$ is the unconditional p.d.f. of X. Computing the expression explicitly and subsuming x, we obtain

$$\int_{\mathbb{X}} \left\{p_0\left(\frac{1}{p} - \lambda\right) - (1-p_0)\left(\frac{1-\lambda}{1-p} + \lambda\right)\right\} p'x\, f_X(x)dx \quad = \quad 0, \tag{8}$$

where $p' \equiv \left.\frac{\partial p(z)}{\partial z}\right|_{z=x\theta^*}$.

The first-order condition in Equation (8) gives a highly nonlinear implicit characterization of the limiting parameter estimate θ^*. However, our setting (involving a single parameter λ to characterize the employed scoring rule) allows us to nevertheless analyze how θ^* varies across scoring rules. Again assuming applicability of the DCT and using implicit differentiation, we obtain

$$\frac{\partial\theta^*}{\partial\lambda} \quad = \quad \frac{\int_{\mathbb{X}} \frac{(p_0-p)pp'}{p(1-p)} x\, f_X(x)dx}{\int_{\mathbb{X}} \left[\frac{(p_0-p)(1-\lambda p)p''}{p(1-p)} - \frac{(p_0(1-p)^2+(1-p_0)(1-\lambda)p^2)p'^2}{p^2(1-p)^2}\right] x^2\, f_X(x)dx}. \tag{9}$$

Equation (9) makes explicit how a change in the scoring rule, which is expressed here by differentiating with respect to λ, affects the probability limit θ^* of the parameter estimator. The denominator of

[6] In the probit case, $p(\cdot)$ is the c.d.f. of a standard normal variable. For the logit, $p(z) = [1 + exp(-z)]^{-1}$ is the c.d.f. of a standard logistic variable.

the above expression is always negative because it is the second derivative of the objective function, $E_{X,Y}\left(S_\lambda(Y, p(X'\theta))\right)$, evaluated at θ^*. However, the sign of the numerator may change depending on the truth, the model, and the density of X.

Under correct specification ($p_0[X] = p(X'\theta_0)$), the numerator in (9) becomes zero, and we get $\frac{\partial\theta^*}{\partial\lambda} = 0$. This result holds irrespective of the link function p and the distribution $f_X(x)$. It mirrors the fact that the composite scoring rule in (7) is strictly proper for any $\lambda \in [0,1]$. This implies that $\theta^* = \theta_0$ regardless of the value of λ. Hence, the choice of scoring rule is irrelevant under correct specification, at least in terms of the probability limit of the parameter estimator.

Under misspecification, it still holds that the denominator of (9) is always negative. Hence, the numerator will determine the sign of $\partial\theta^*/\partial\lambda$. We first examine when the sign is zero, i.e. the choice of scoring rule does not matter, in the following theorem.

Theorem 2 *Assume*

(a) $p_0[x] = 1 - p_0[-x]$,

(b) $p[x] = 1 - p[-x]$, *p is differentiable in x,*

(c) $f_X(x) = f_X(-x)$, $f_X \geq 0$, $dim(X) = dim(\theta) = 1$, $E[X] = 0$.

In addition, assume conditions (a)-(f) in Theorem 1 hold, and define $\mathbb{X}^+ = \mathbb{X} \cap [0,\infty)$ *and* $\mathbb{X}^- = \mathbb{X} \cap (-\infty, 0]$.

Then

$$\frac{\partial\theta^*}{\partial\lambda} = 0$$

if and only if

$$\int_{\mathbb{X}^+} (p_0[x] - p[x])^+ \frac{p'[x]x}{p[x](1-p[x])} f_X(x)dx \quad = \quad \int_{\mathbb{X}^+} (p_0[x] - p[x])^- \frac{p'[x]x}{p[x](1-p[x])} f_X(x)dx, \tag{10}$$

By symmetry, (10) *also holds on* $\mathbb{X}^-$.

Proof We denote the integrand in the numerator of (9) by $g[x]$, where

$$g[x] = \frac{(p_0[x] - p[x])p[x]p'[x]}{p[x](1-p[x])}x.$$

Assumptions (a) and (b) from the statement of the theorem imply that

$$\begin{aligned} p_0[x] - p[x] &= -(p_0[-x] - p[-x]), \\ p[x](1 - p[x]) &= p[-x](1 - p[-x]), \\ p'[x] &= p'[-x], \end{aligned}$$

assuming that p is differentiable in x. Note that the first equality means that the approximation error due to the chosen model is point-symmetric about the origin. The latter equalities depend on the approximation model only and hold for commonly used specifications such as probit and logit. These relationships, together with calculations detailed in Appendix A.1, imply that

$$\begin{aligned} & \int_{\mathbb{X}} g[x] f_X(x) dx \\ = & \int_{\mathbb{X}^+} (p_0[x] - p[x]) \frac{p'[x]x}{p[x](1 - p[x])} f_X(x) dx, \end{aligned}$$

Note that all quantities on the right-hand side of the last equality are nonnegative, except for $p_0[x] - p[x]$. Thus, the result follows that it would equal zero iff

$$\int_{\mathbb{X}^+} (p_0[x] - p[x])^+ \frac{p'[x]x}{p[x](1 - p[x])} f_X(x) dx = \int_{\mathbb{X}^+} (p_0[x] - p[x])^- \frac{p'[x]x}{p[x](1 - p[x])} f_X(x) dx.$$

where $(h[x])^+ = |h[x]| 1\{\text{sign}(h[x]) = +1\}$ and $(h[x])^- = |h[x]| 1\{\text{sign}(h[x]) = -1\}$. By symmetry, the same holds for the integral of the same term over $\mathbb{X}^-$. □

Discussion of assumptions in Theorem 2: (a) and (b) are point-symmetry conditions on p_0 and p, respectively, (c) ensures that X is symmetric about the origin. Together, (a) and (b) imply the point-symmetry of $p_0 - p$ about the origin.

Theorem 2 gives precise conditions under which the choice of scoring rule, and thus the weighting over decision makers, has no impact on the limiting parameter estimate θ^*. The intuition here is that on each part of the support of X, the model over-estimates and under-estimates the true conditional probability. In addition, the upward and downward prediction error averaged as in the above equation are equal on $\mathbb{X}^+$ and $\mathbb{X}^-$, respectively. This excludes a situation where the model tends to over-estimate the true conditional probability on $\mathbb{X}^+$ and under-estimate it on $\mathbb{X}^-$, or vice versa. In practice, the symmetry requirements imposed on the model (Assumption (b) in Theorem 2) hold for the widely used logit and probit specifications. The symmetry requirements imposed on the true conditional probability as well as the density of X (Assumptions (a), (c) in Theorem 2) may be unrealistic for certain applications. We next provide Monte Carlo evidence on the role of scoring rules under various scenarios which either do or do not satisfy the conditions of Theorem 2.

3 Numerical Demonstration of the Results

This section illustrates the results in Theorem 2 both for the asymptotic problem and for finite samples. We consider four data-generating processes (DGPs) given in Table 2 as well as the scoring rules in Table 1. In DGP #1, the symmetry conditions on the conditional probability under the true and misspecified models as well as on the marginal distribution of X, (a)-(c) in Theorem 2, are fulfilled. DGP #2 and #3 are variants of DGP #1 where the symmetry condition on the true conditional probability (a) and on the distribution of X (c), respectively, are violated. DGP #4 presents an example where both conditions (a) and (c) are violated.[7]

DGP #	$p_0(X)$	f_X	Correctly Specified Model (p_0)	Misspecified Model (p)
1	$F(-0.5X + 0.2X^3)$	$\mathcal{U}(-2.5, 2.5)$	$F(\theta_1 X + \theta_2 X^3)$	$F(\theta X)$
$2^{\dagger}$	$F(-0.5X + 0.2X^2)$	$\mathcal{U}(-2.5, 2.5)$	$F(\theta_1 X + \theta_2 X^3)$	
3^{*}	$F(-0.5X + 0.2X^3)$	$\mathcal{U}(-1, 4)$	$F(\theta_1 X + \theta_2 X^3)$	
$4^{*\dagger}$	$F(\sqrt{X})$	$\mathcal{U}(0, 10)$	$F(\theta\sqrt{X})$	

Table 2: DGPs used in simulation. f_X denotes the p.d.f. of X, $F(s) = \frac{\exp(s)}{1+\exp(s)}$ denotes the c.d.f. of the logistic distribution, and $\mathcal{U}(a, b)$ denotes the uniform distribution with limits a and b. DGP #1 is taken from Elliott and Lieli (2013, pp. 16–19) and fulfills conditions (a)-(c). $*$ indicates that a DGP violates condition (c), † indicates a DGP violates condition (a).

3.1 Asymptotic Problem

The necessary and sufficient conditions in Equation (10) of Theorem 2 give us an insight into how certain symmetry conditions jointly determine whether the choice over scoring rules has an effect on the estimated parameter and thereby the conditional probability. In this section, we seek to illustrate this insight for the asymptotic problem. To do so, for each scoring rule j, we compute θ_j^* for the misspecified logit model given in Table 2 numerically, based on a sample of size $1,000,000$. The parameters are reported in Table 3. Since all other elements of the forecast are identical, the difference between the conditional probability under various scoring rules is due to the difference in θ_j^*. Figure 2 plots the conditional probability for all scoring rules under DGPs #1 to #4. The plot for DGP #1 gives a case where the choice of the scoring rule has no effect on the conditional probability. This is true not only for the choice between the Log and As1 scores, as shown in Theorem 2, but also holds for all other scoring rules we examine. The plot clearly shows other implications of Theorem 2: (1) the prediction error is point-symmetric about the origin, (2) the prediction error changes its sign on $\mathbb{X}^+$ and $\mathbb{X}^-$, (3) a weighted average of the positive and negative prediction error on $\mathbb{X}^+$ as well as $\mathbb{X}^-$ would be equal as indicated by Equation (10). For all the other DGPs, where either (a), (c) or both

[7] In fact, the true conditional probability $p_0(X)$ is not even defined for $X < 0$.

	DGP #1	DGP #2	DGP #3	DGP #4
		θ_j^*		
Log	0.22	-0.44	0.6	0.43
Brier	0.22	-0.44	0.51	0.57
Spherical	0.21	-0.44	0.45	0.65
Boosting	0.22	-0.43	0.68	0.39
As1	0.22	-0.33	0.46	0.61
As2	0.22	-0.64	0.66	0.41

Table 3: Asymptotic parameter estimates for various scoring rules j, based on $1,000,000$ observations.

are violated, our numerical results clearly show that the choice of scoring rule has an effect on the conditional probability approximation. For DGP #2, where only the symmetry condition on the true conditional probability is violated, the only scoring rules that result in different predicted conditional probabilities than the Log Score are the asymmetric scoring rules. For DGPs #3 and #4, we observe differences in the predicted probabilities for all pairs of scoring rules, even if both scoring rules under comparison are symmetric (such as the Log versus Brier score).

As discussed earlier, the binary action of an individual decision maker (such as a fisherman or restaurant owner in our example above) is determined by whether or not the predicted probability, $F(\theta_j^* X)$, exceeds the threshold c. For a given value of the regressor X, the chosen action may thus depend on the scoring rule j used for parameter estimation. Figure 3 illustrates this point for $x = 2$. It shows that the scoring rules generally yield different classification curves. Rather than looking at a single design point (such as $x = 2$ above), we next consider a broader summary measure of differences between scoring rules:

$$\mathbb{P}\big(\text{sign}\left\{F(\theta_j^* X) - c\right\} \neq \text{sign}\left\{F(\theta_{\text{Log score}}^* X) - c\right\}\big), \tag{11}$$

which is the probability (computed over the distribution of X) that scoring rule j implies a different binary action than the Log score. Figure 4 shows how the choice of scoring rules under DGP #1 is inconsequential, in the sense of always leading to identical decisions. For DGP #2, the choice between the Log and the two asymmetric rules is the only one that leads to different classifications. For $c = 0.6$ and $c = 0.7$, the probability of different classification is 0.05 and 0.12, respectively. For DGPs #3 and #4, the choice between the Log and any other scoring rule leads to different classifications. What is particularly interesting here is that even though choosing between the Log and other symmetrical rules, such as Brier, may be relatively inconsequential for $c = 0.6$, it can lead to an 0.15 and 0.175 probability of different classifications at greater values of c for DGP #3 and #4, respectively. Thus, whether or not the differences across scoring rules matter depends on a decision maker's preferences

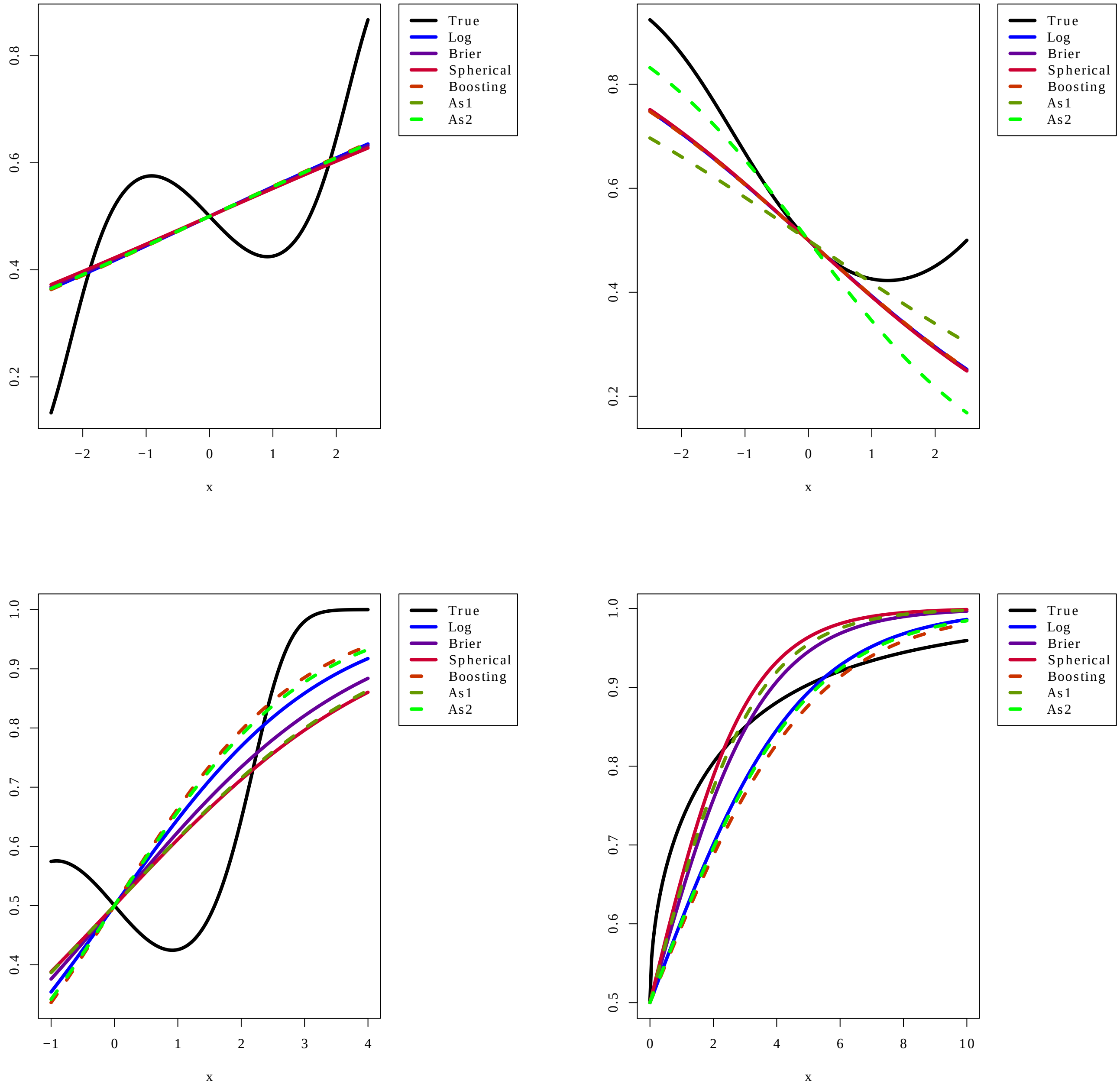

Figure 2: Asymptotic Problem under Misspecification for DGPs #1 (upper left), #2 (upper right), #3 (lower left), and #4 (lower right). The plots show the predicted conditional probability, whereby the parameter θ_j^* is computed using a sample of $1,000,000$ draws.

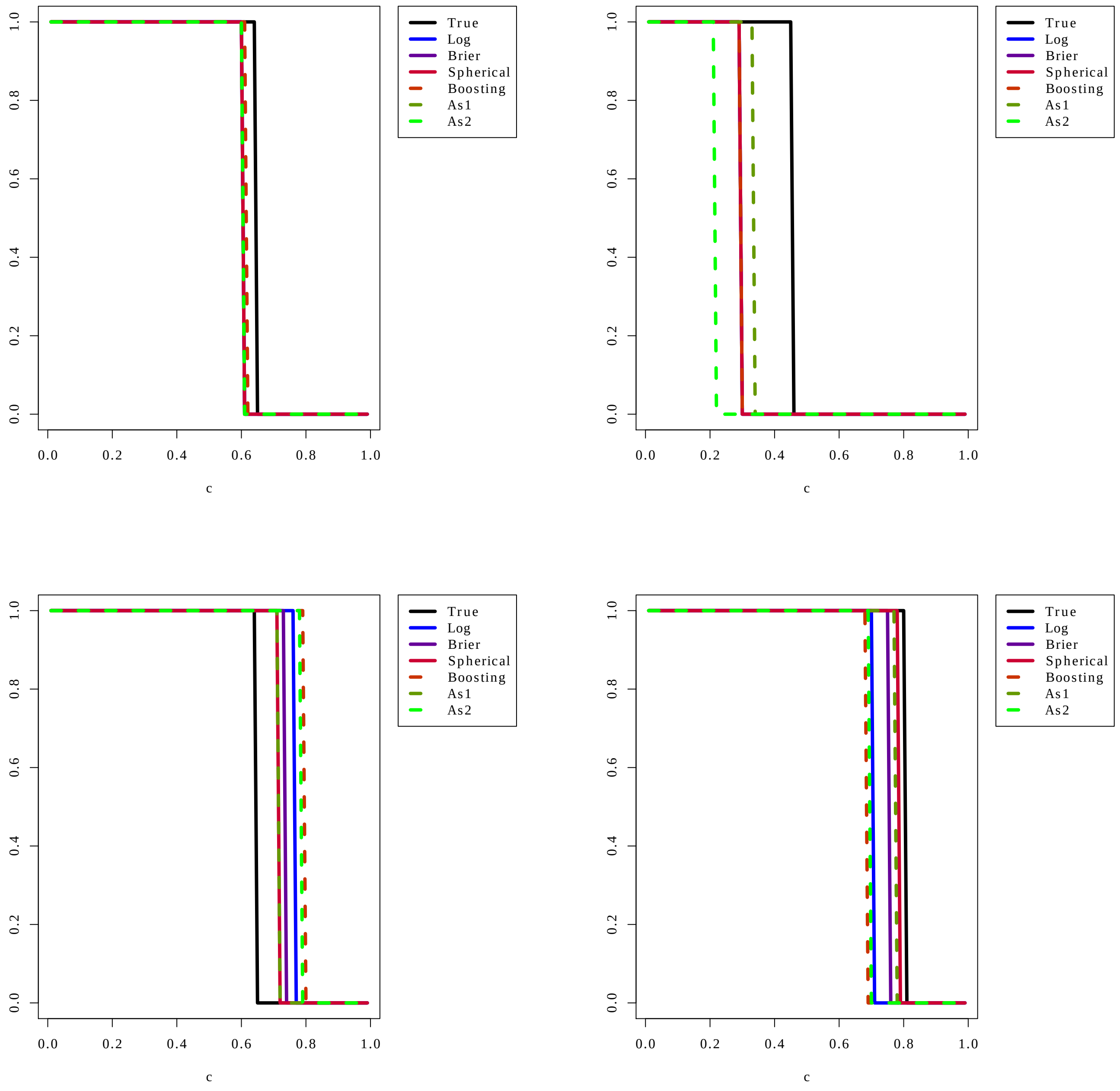

Figure 3: Asymptotic Problem under Misspecification for DGPs #1 (upper left), #2 (upper right), #3 (lower left), and #4 (lower right). The plots show the classification curves of the conditional probability given $x = 2$ (θ_j^* is computed using a sample of size $1,000,000$).

as embodied in c.

Put slightly differently, this number represents the probability that a decision maker with preference parameter c makes a correct decision when using a prediction model fitted via scoring rule j. Figure 5 shows how the ranking of the scoring rules differs across thresholds c. For example, consider the comparison of As1 and As2 in the figure for DGP #2 (upper right panel): While As1 performs better for values of c between 0.2 and 0.4, the reverse is true for c lying between 0.6 and 0.8. This result is closely in line with the fact that, when used as an estimation criterion, As1 places an emphasis on fitting small thresholds c correctly, whereas As2 focuses on high values of c (see Section 2.1). This analysis demonstrates how forecasters who are willing to favor a certain clientele (say, decision makers characterized by small values c, such as the fishermen in our example) can achieve this goal by issuing predictions based on an appropriate scoring rule (in this case, As1). As a further check, Figure 5 provides evidence on the probability that the estimator defined by scoring rule j delivers a correct classification. This probability is given by

$$\mathbb{P}\big(\text{sign}\left\{F(\theta_j^* X) - c\right\} = \text{sign}\left\{p_0(X) - c\right\}\big). \tag{12}$$

3.2 Finite-Sample Results

All of our results until now are for the case in which the limiting parameter values θ_j^* are known. We now briefly turn to the effects of sampling uncertainty. Specifically, we consider a rolling window estimation scheme for θ_j^* which is popular in practice (c.f. the discussion by Giacomini and White, 2006, p. 1548), using a window length of 120. Furthermore, we consider a forecast evaluation period of 100 periods.[8] In each Monte Carlo iteration, we thus simulate $120 + 100$ observations. For the first rolling window, we use observations 1 to 120 to estimate the parameter θ and make a forecast for observation 121. For the second rolling window, we use observations 2 to 121 for estimation and make a forecast for observation 122, and so forth.

The appendix reports variants of Figures 2 and 3 for the rolling window case, which we construct by averaging the probability and classification curves for each estimate of θ. The figures show that *on average*, the rolling window parameter estimates are very similar to their asymptotic counterparts.

Theorem 1 implies that, in the asymptotic case, it is generally optimal to use the same scoring rule for estimation and evaluation. We next analyze to which extent this statement carries over to the rolling window scenario. To this end, Table 4 summarizes the parameter estimates and predictive performance obtained under each scoring rule. The median estimates for each scoring rule (upper panel

[8] These sample sizes are typical in forecasting studies using quarterly macroeconomic data, for example when using an estimation sample from 1960 to 1989 ($30 \times 4 = 120$ observations), and an evaluation sample from 1990 to 2014 ($25 \times 4 = 100$ observations).

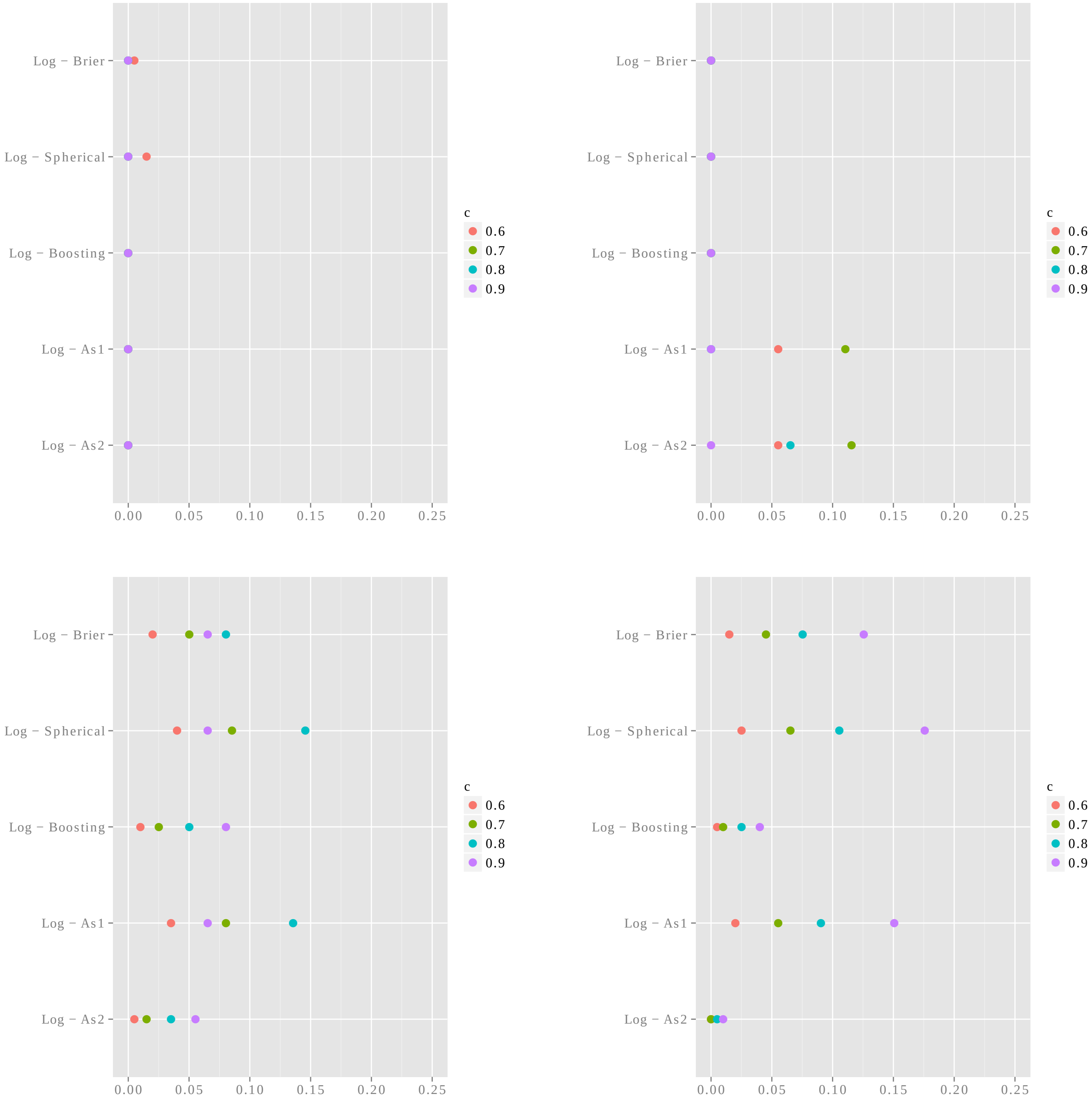

Figure 4: Asymptotic Problem under Misspecification for DGPs #1 (upper left), #2 (upper right), #3 (lower left), and #4 (lower right). The plots show the unconditional probability of different classifications using the Log score as opposed to other scoring rules, see Equation (11). θ_j^* is computed using a sample of size $1,000,000$.

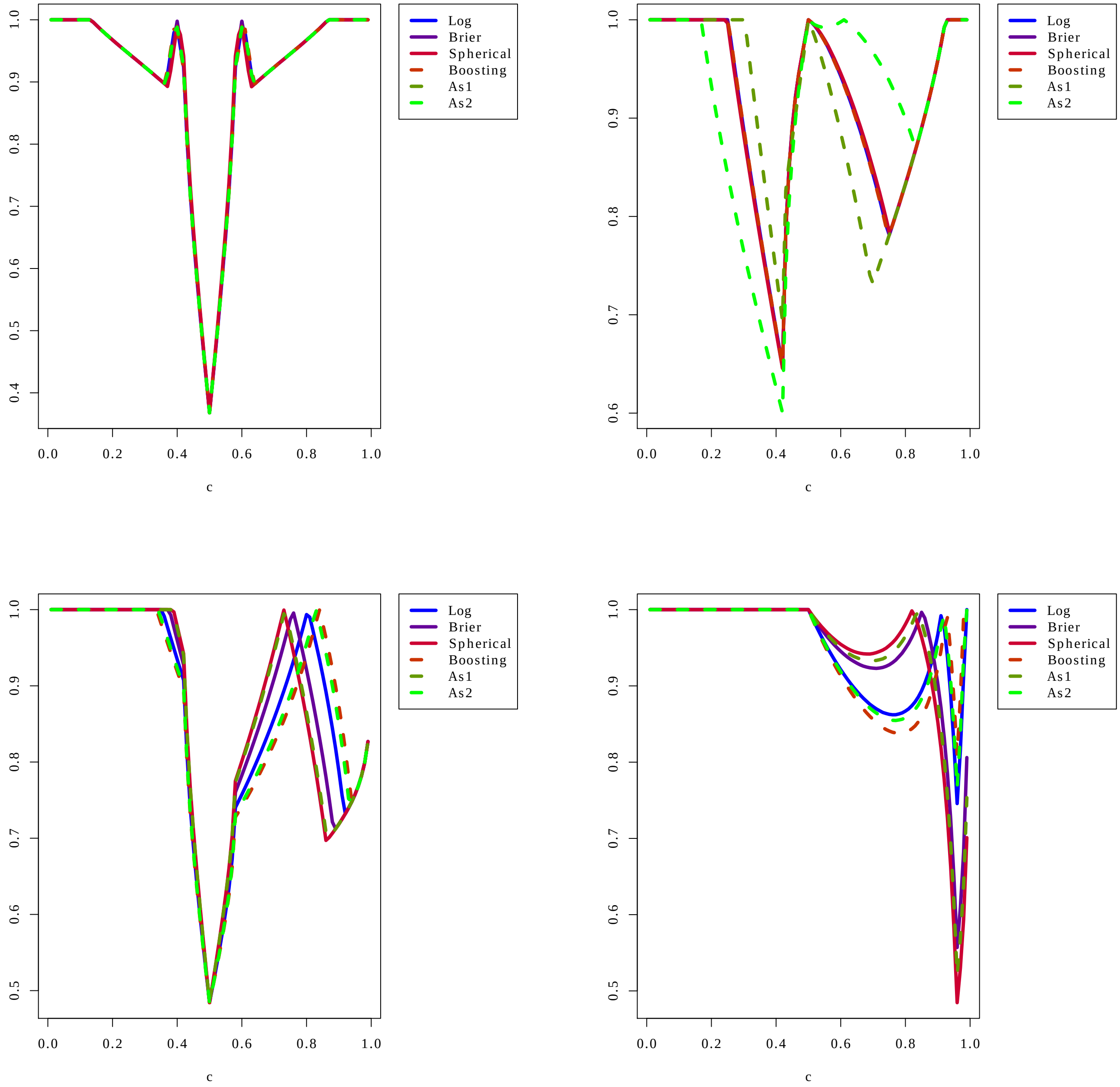

Figure 5: Asymptotic Problem under Misspecification for DGPs #1 (upper left), #2 (upper right), #3 (lower left), and #4 (lower right). The plots show the probability of correct classification (see Equation 12) for various thresholds c. θ_j^* is computed using a sample of size $1,000,000$.

of Table 4) are very close to their asymptotic limits in Table 3. This is in line with the similarity of the prediction and classification curves noted above. Estimators defined by alternative scoring rules clearly differ in their sampling variability, as measured by interdecile ranges (middle panel of Table 4), especially for DGPs #3 and 4.[9] That said, there is no simple relationship between the choice of scoring rule and the variability of the estimator it defines. For example, the spherical score defines the most precise estimator for DGP #1, whereas it defines the (by far) least precise estimator for DGP #4.

The bottom panel of Table 4 compares the forecast performance of two strategies: (1) using the same scoring rule for estimation and evaluation ("matching"), (2) simply using MLE for parameter estimation while using a different scoring rule for evaluation. In order to compare the performance of the two, we simply report the share of Monte Carlo iterations for which the first strategy performs better, whereby we average over an out-of-sample period of 100 observations in each Monte Carlo iteration. This share has a natural scaling between zero and one. It is therefore easily interpretable and comparable across scoring rules.[10]

In 13 of the 20 cases, shares in the bottom panel of Table 4 are strictly above 0.5, indicating better performance of matching compared to MLE. For some of these cases, we find that matching leads to a substantially different median estimate as well as lower variability as measured by the interdecile range. This holds true for As1 under DGP #2, as well as Brier, Spherical, and As1 under DGP#3. There are also some cases where the matching estimator is more variable than MLE, but nevertheless performs better out-of-sample. This happens for As2 under DGP #2, Boosting and As2 under DGP #3, as well as Spherical under DGP #4. In these cases, it seems that the relative gain from using an estimator that converges to the maximand of the scoring rule in question outweighs the relative loss in precision. For another subset of the cases where matching performs better, such as Brier and Spherical under DGP #1, Boosting under DGP #2, and As2 under DGP #4, the medians and interdecile ranges of the MLE and matching estimator are practically indistinguishable. We conjecture that in these cases, the matching strategy's improvement over MLE is marginal. Along the same lines, the six cases in which MLE does better than matching appear very close, with both strategies attaining similar medians and interdecile ranges.

To summarize, our results show that the "correct location" of the matching estimator puts it at an advantage over MLE under misspecification, which generally does not converge to the maximand of the scoring rule in question. To compensate for this, the MLE must be more precise (smaller interdecile range) in order to outperform matching in terms of out-of-sample scores.

[9] We use interdecile ranges, rather than variances, to eliminate the effect of outliers which are not surprising given the scale of our Monte Carlo experiment (for each scoring rule and DGP, we compute 100 times $10,000$ rolling window estimates). Measuring the estimators' variability by the median absolute deviation from the median (instead of the interdecile range) leads to the same qualitative interpretations.

[10] By contrast, the scoring rules themselves have no natural scaling, which makes it hard to judge whether differences of a given magnitude are practically relevant or not, and also impedes comparisons of effect sizes across scoring rules.

	DGP #1	DGP #2	DGP #3	DGP #4
	Median estimate $\hat{\theta}$			
Log	0.22	-0.44	0.6	0.43
Brier	0.21	-0.44	0.51	0.58
Spherical	0.21	-0.45	0.46	0.67
Boosting	0.22	-0.44	0.68	0.4
As1	0.22	-0.34	0.46	0.62
As2	0.22	-0.65	0.66	0.42
	Interdecile range of estimates $\hat{\theta}$			
Log	0.32	0.36	0.26	0.23
Brier	0.31	0.37	0.2	0.83
Spherical	0.29	0.38	0.19	1.58
Boosting	0.33	0.36	0.32	0.22
As1	0.34	0.3	0.2	1.1
As2	0.33	0.57	0.27	0.22
	Share of MC iterations for which matching scores better than MLE*			
Brier	0.76	0.44	0.75	0.47
Spherical	0.75	0.42	0.82	0.52
Boosting	0.23	0.52	0.53	0.53
As1	0.43	0.73	0.8	0.5
As2	0.45	0.57	0.59	0.52

Table 4: Summary results for rolling windows (window length 120). *In each MC iteration, we compute the average score over an out-of-sample period of 100 observations. “Matching” means to use the same scoring rule for estimation and evaluation.

4 Conclusion

This paper explores the nuances in forecasting conditional probabilities under misspecification. The natural choice under correct specification - regardless of the scoring rule used for out-of-sample evaluation - is indeed MLE. It is not only consistent for the maximand of the scoring rule in question, but also efficient. Under misspecification, however, there is no clear natural choice. The MLE is neither consistent for the maximand of the scoring rule in question, nor necessarily "efficient" in the sense of attaining lower sampling variability than other estimators. The paper shows in an analytical example that under certain symmetry conditions, the choice of scoring rule is inconsequential for parameter estimation. With the aid of numerical results for the asymptotic problem, we then illustrate how the violation of these conditions can lead to different probability limits of the parameter estimators and to different conditional probability forecasts. We also show how these different forecasts would lead to different interpretations by heterogeneous decision makers. In finite samples, we find an interesting relationship between the sampling distribution of the parameter estimators and the relative performance of the MLE (compared to the estimator that maximizes the scoring rule considered for evaluation).

Finally, our analysis has conceptual implications pertaining to the literature on distributional forecasting. It has been argued (e.g. Geweke and Amisano, 2011, p. 130) that the provision of distributional forecasts is superior to the provision of point forecasts because distributional forecasts can be employed to construct point forecasts for any loss function. While this argument seems valid in many situations (c.f. Section 1), it should not be misunderstood as saying that distributional forecasts were "loss function independent". Specifically, the present paper illustrates that probability forecasts – which are clearly distributional – are not loss function independent. A loss function is required for estimation and this choice makes explicit trade-offs regarding which aspects of the data to fit correctly, at the cost of neglecting other aspects.

References

BOYES, W. J., D. L. HOFFMAN, AND S. A. LOW (1989): "An Econometric Analysis of the Bank Credit Scoring Problem," *Journal of Econometrics*, 40(1), 3–14.

BRIER, G. W. (1950): "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78(1), 1–3.

BUJA, A., W. STUETZLE, AND Y. SHEN (2005): "Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications," Working Paper, Duke University.

DE JONG, R. M., AND T. WOUTERSEN (2011): "Dynamic Time Series Binary Choice," *Econometric Theory*, 27(4), 673–702.

DEUTSCHE BANK (2014): "Sovereign Default Probabilities Online," http://www.dbresearch.com/servlet/reweb2.ReWEB?rwnode=DBR_INTERNET_EN-PROD$NAVIGATION&rwobj=CDS.calias&rwsite=DBR_INTERNET_EN-PROD, Accessed: 2014-01-30.

ELLIOTT, G., AND R. P. LIELI (2013): "Predicting Binary Outcomes," *Journal of Econometrics*, 174(1), 15–26.

GEWEKE, J. W., AND G. AMISANO (2011): "Optimal Prediction Pools," *Journal of Econometrics*, 164(1), 130–141.

GIACOMINI, R., AND H. WHITE (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74(6), 1545–1578.

GNEITING, T. (2011): "Making and Evaluating Point Forecasts," *Journal of the American Statistical Association*, 106(494), 746–762.

GNEITING, T., AND A. E. RAFTERY (2007): "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102(477), 359–378.

GOOD, I. (1952): "Rational Decisions," *Journal of the Royal Statistical Society, Series B*, 14(1), 107–114.

GRANGER, C. W. J. (1993): "On the Limitations of Comparing Mean Square Forecast Errors: Comment," *Journal of Forecasting*, 12(8), 651–652.

GRANGER, C. W. J., AND M. H. PESARAN (2000): "Economic and Statistical Measures of Forecast Accuracy," *Journal of Forecasting*, 19(7), 537–560.

HAMILTON, J. D., AND M. CHINN (2014): "Econbrowser - Analysis of Current Economic Conditions and Policy," http://www.econbrowser.com, Accessed: 2014-01-30.

HAND, D. J., AND V. VINCIOTTI (2003): "Local versus Global Models for Classification Problems: Fitting Models Where it Matters," *The American Statistician*, 57(2), 124–131.

HAYASHI, F. (2000): *Econometrics*. Princeton University Press.

INOUE, A., AND B. ROSSI (2008): "Monitoring and Forecasting Currency Crises," *Journal of Money, Credit and Banking*, 40(2-3), 523–534.

KNAPP, L. G., AND T. G. SEAKS (1992): "An Analysis of the Probability of Default on Federally Guranteed Student Loans," *The Review of Economics and Statistics*, 74(3), pp. 404–411.

KOENKER, R., AND J. YOON (2009): "Parametric Links for Binary Choice Models: A Fisherian-Bayesian Colloquy," *Journal of Econometrics*, 152(2), 120–130.

LIELI, R. P., AND A. NIETO-BARTHABURU (2010): “Optimal Binary Prediction for Group Decision Making,” *Journal of Business & Economic Statistics*, 28(2), 308–319.

LIELI, R. P., AND M. SPRINGBORN (2013): “Closing the Gap between Risk Estimation and Decision Making: Efficient Management of Trade-Related Invasive Species Risk,” *Review of Economics and Statistics*, 95(2), 632–645.

LIELI, R. P., AND H. WHITE (2010): “The Construction of Empirical Credit Scoring Rules based on Maximization Principles,” *Journal of Econometrics*, 157(1), 110–119.

MANSKI, C. F., AND T. S. THOMPSON (1989): “Estimation of Best Predictors of Binary Response,” *Journal of Econometrics*, 40(1), 97–123.

MASS, C., J. BAARS, S. JOSLYN, J. PYLE, P. TEWSON, D. JONES, T. GNEITING, A. RAFTERY, J. SLOUGHTER, AND C. FRALEY (2009): “PROBCAST: A Web-Based Portal to Mesoscale Probabilistic Forecasts,” *Bulletin of the American Meteorological Society*, 90(7), 1009–1014.

MERKLE, E. C., AND M. STEYVERS (2013): “Choosing a Strictly Proper Scoring Rule,” *Decision Analysis*, 10(4), 292–304.

PATTON, A. J. (2014): “Comparing Possibly Misspecified Forecasts,” Working Paper, Duke University.

SCHERVISH, M. J. (1989): “A General Method for Comparing Probability Assessors,” *Annals of Statistics*, 17(4), 1856–1879.

SHUFORD, E. H., A. ALBERT, AND H. E. MASSENGILL (1966): “Admissible Probability Measurement Procedures,” *Psychometrika*, 31(2), 125–145.

TODA, M. (1963): “Measurement of Subjective Probability Distribution,” Report 3, State College, Pennsylvania, Institute for Research, Division of Mathematical Psychology.

WEISS, A. A. (1996): “Estimating Time Series Models Using the Relevant Cost Function,” *Journal of Applied Econometrics*, 11(5), 539–560.

WEISS, A. A., AND A. P. ANDERSEN (1984): “Estimating Time Series Models Using the Relevant Forecast Evaluation Criterion,” *Journal of the Royal Statistical Society, Series A*, 147(3), 484–487.

WHITE, H. (2001): *Asymptotic Theory for Econometricians*. Academic Press.

WOOLDRIDGE, J. M. (1994): “Estimation and Inference for Dependent Processes,” vol. 4 of *Handbook of Econometrics*, pp. 2639 – 2738. Elsevier.

Appendix

A Additional Material for Section 2.3

A.1 Detailed steps in the proof of Theorem 2

By Assumptions (a), (b) and (c) in the statement of the theorem and their direct implications mentioned in the proof,

$$\begin{aligned}
& \int_{\mathbb{X}} g[x] f_X(x)dx \\
= & \int_{\mathbb{X}} (p_0[x]-p[x])(p[x]-0.5)\frac{p'[x]}{p[x](1-p[x])} x f_X(x)dx + 0.5\int_{\mathbb{X}} (p_0[x]-p[x])\frac{p'[x]}{p[x](1-p[x])} x f_X(x)dx \\
= & \int_{\mathbb{X}+} (p_0[x]-p[x])(p[x]-0.5)\frac{p'[x]}{p[x](1-p[x])} x f_X(x)dx \\
+ & \int_{\mathbb{X}+} \{-(p_0[x]-p[x])\}\{-(p[x]-0.5)\}\frac{p'[x]}{p[x](1-p[x])} (-x) f_X(x)dx \\
+ & \ 0.5\int_{\mathbb{X}} (p_0[x]-p[x])\frac{p'[x]}{p[x](1-p[x])} x f_X(x)dx \\
= & \ 0.5\int_{\mathbb{X}} (p_0[x]-p[x])\frac{p'[x]x}{p[x](1-p[x])} f_X(x)dx.
\end{aligned}$$

Now the term on the right hand side of the last equality can be further simplified as follows:

$$\begin{aligned}
& \int_{\mathbb{X}} (p_0[x]-p[x])\frac{p'[x]x}{p[x](1-p[x])} f_X(x)dx \\
= & \int_{\mathbb{X}+} (p_0[x]-p[x])\frac{p'[x]x}{p[x](1-p[x])} f_X(x)dx + \int_{\mathbb{X}+} \{-(p_0[x]-p[x])\}\frac{p'[x](-x)}{p[x](1-p[x])} f_X(x)dx \\
= & \ 2\int_{\mathbb{X}+} (p_0[x]-p[x])\frac{p'[x]x}{p[x](1-p[x])} f_X(x)dx,
\end{aligned}$$

A.2 Including an intercept

Here we briefly consider extending the forecast model of the analytical example to include an intercept as well as a regressor. For this purpose, let $\mathbf{z} = \begin{bmatrix} 1 & x \end{bmatrix}'$ denote the 2×1 regressor vector. The matrix analogue of Equation (9) is given by

$$\begin{aligned}
\frac{\partial \theta^*}{\partial \lambda} = & \left[\int_{\mathbb{X}} \left[\frac{(p_0-p)(1-\lambda p)p''}{p(1-p)} - \frac{\left(p_0(1-p)^2 + (1-p_0)(1-\lambda)p^2\right)p'^2}{p^2(1-p)^2} \right] \mathbf{z}\, t(\mathbf{z})\ f_Z(\mathbf{z})d\mathbf{z} \right]^{-1} \times \\
& \left[\int_{\mathbb{X}} \frac{(p_0-p)pp'}{p(1-p)} \mathbf{z}\ f_Z(\mathbf{z})d\mathbf{z} \right], \\
\equiv & \begin{bmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{bmatrix} \times \begin{bmatrix} J_1 \\ J_2 \end{bmatrix} \qquad (13)
\end{aligned}$$

where $t(\mathbf{z})$ denotes the transpose of $\mathbf{z}$. Note that the p.d.f. $f_Z(\mathbf{z})$ is the "joint" distribution of the intercept and the regressor X, and is hence equivalent to the marginal distribution of X.

The derivative $\frac{\partial \theta^*}{\partial \lambda}$ is now a vector of dimension 2×1. Its first element expresses how the limiting intercept is affected by a change in the scoring rule. The second element indicates how the slope parameter is affected. Both elements are weighted sums of the terms J_1 and J_2 in (13), with weights given by the elements H^{ij} of the inverse matrix (which is well defined, since the inverted matrix is the Hessian matrix of the model at the optimum θ^*, and as such is negative definite). Three points are worth mentioning:

- A sufficient condition to ensure that the $\frac{\partial \theta^*}{\partial \lambda} = \begin{bmatrix} 0 & 0 \end{bmatrix}'$, such that the choice of λ is irrelevant for both the intercept and the slope, is that $J_1 = J_2 = 0$.
- This sufficient condition is satisfied under correct specification ($\exists\, \theta_0 \in \Theta$ s.t. $p_0 = p \,\forall\, x \in \mathbb{X}$), which is in line with the fact that the true model is optimal under any value of λ.
- Under misspecification, a sufficient condition for $J_1 = J_2 = 0$ can be constructed in close analogy to Theorem 2, by simpliying the expressions for J_1 and J_2 under additional symmetry assumptions as in the theorem. Note, however, that these conditions will not be necessary here, while they are necessary in the scalar setting of Theorem 2.

B Additional Figures for Section 3

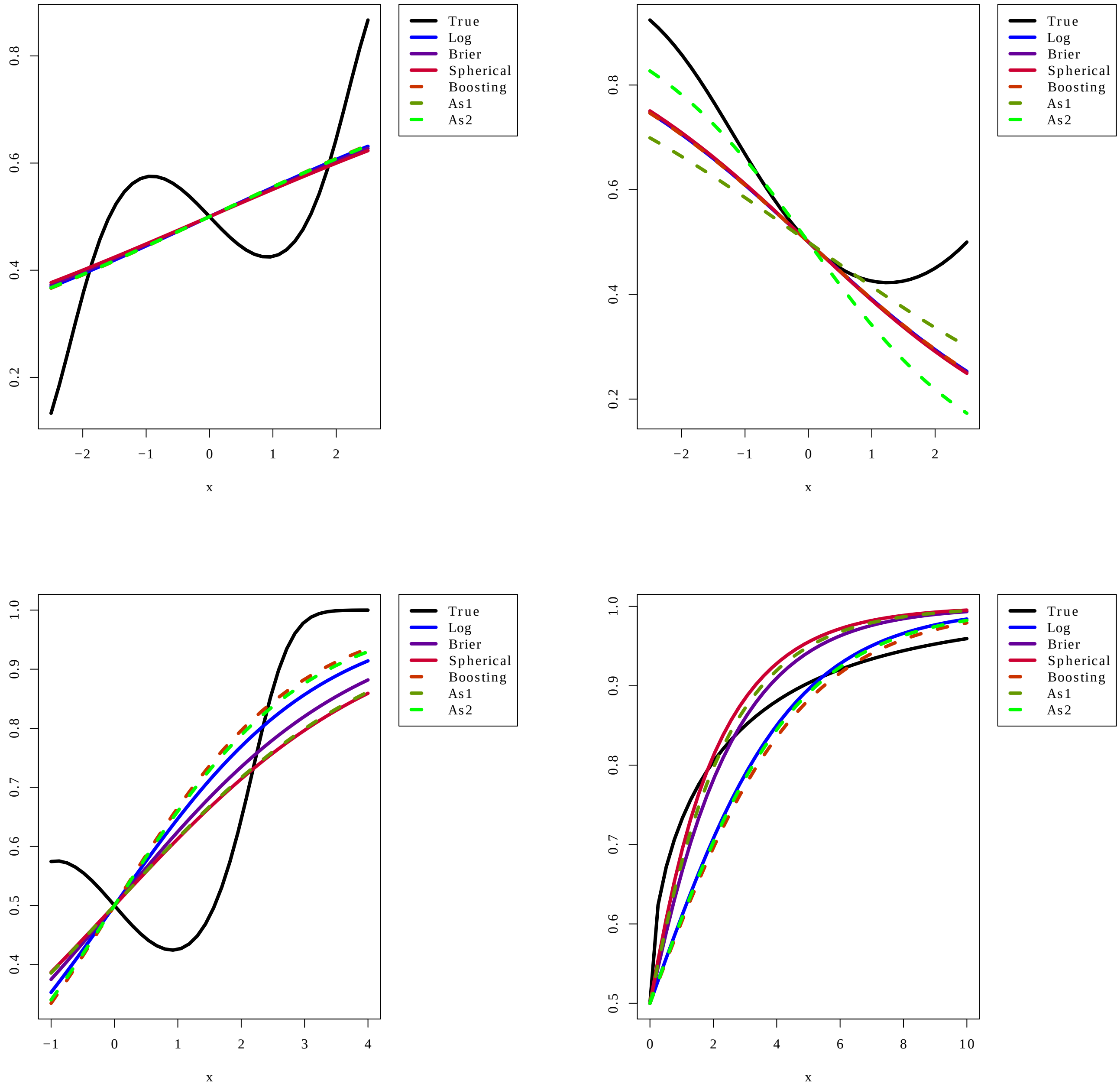

Figure 6: Predicted probability curves based on (misspecified) rolling window estimates for DGPs #1 (upper left), #2 (upper right), #3 (lower left), and #4 (lower right). The plots show the average conditional probability, $\frac{1}{L}\sum_{l=1}^{L} F(x\ \hat{\theta}_l)$, as a function of x. The $\{\hat{\theta}_l\}_{l=1}^{L}$ represent $L = 1,000,000$ parameter estimates per scoring rule ($10,000$ Monte Carlo iterations times 100 rolling windows per iteration).

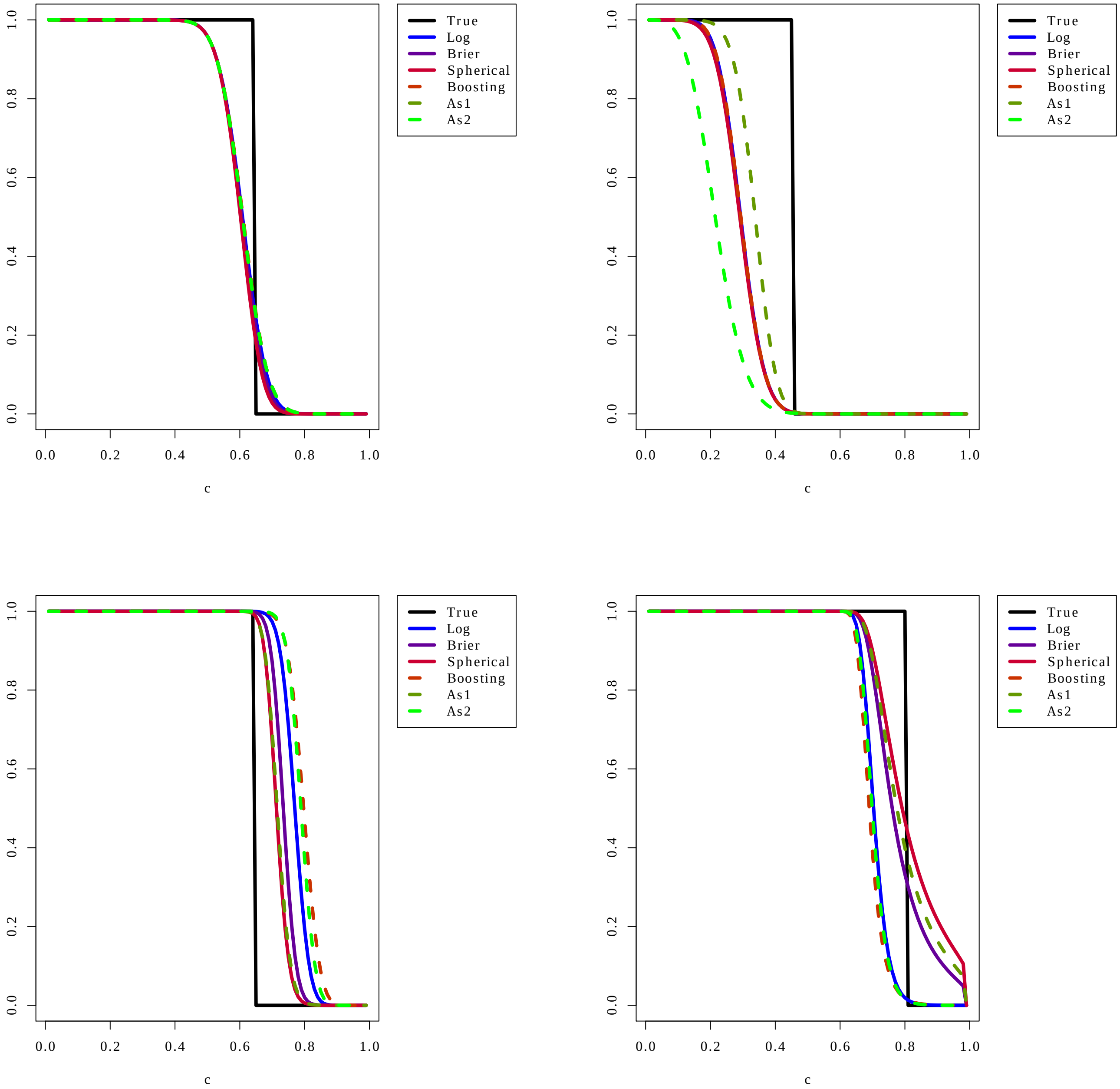

Figure 7: Classification curves based on (misspecified) rolling window estimates for DGPs #1 (upper left), #2 (upper right), #3 (lower left), and #4 (lower right). For a given scoring rule, the plots show the average classification curves at $x = 2$, $\frac{1}{L}\sum_{l=1}^{L} 1(F(2\,\hat{\theta}_l) > c)$, as a function of c. The $\{\hat{\theta}_l\}_{l=1}^{L}$ represent $L = 1,000,000$ parameter estimates per scoring rule ($10,000$ Monte Carlo iterations times 100 rolling windows per iteration).