Efficient probability sequences

Eva Regnier

August 18, 2014

Abstract

A *probability sequence* is an ordered set of probability forecasts for the same event. Although single-period probabilistic forecasts and methods for evaluating them have been extensively analyzed, we are not aware of any prior work on evaluating probability sequences. This paper proposes an efficiency condition for probability sequences and shows properties of efficient forecasting systems, including memorylessness and increasing discrimination. These results suggest tests for efficiency and remedial interventions for inefficient systems.

1 Background

A probability forecast is an estimate of the probability that a precisely defined event will occur. A *probability sequence* is an ordered set of probability forecasts for a single verifying event. An example currently appearing in the press is the forecast for the 2014 US Senate election at fivethirtyeight.com,[1] which in March of 2014 forecast a Republican Senate majority in the 114^{th} Congress with a probability of 0.508. The most recent earlier forecast, issued in July of 2013, was for a Republican majority with a probability of 0.504. The forecast will

[1] http://fivethirtyeight.com/features/fivethirtyeight-senate-forecast/, accessed May 2014

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
18 AUG 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Effi cient probability sequences	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Naval Postgraduate School,Defense Resources Management Institute,699 Dyer Road,Monterey,CA,93943	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A probability sequence is an ordered set of probability forecasts for the same event. Although single-period probabilistic forecasts and methods for evaluating them have been extensively analyzed, we are not aware of any prior work on evaluating probability sequences. This paper proposes an e ciency condition for probability sequences and shows properties of e cient forecasting systems including memorylessness and increasing discrimination. These results suggest tests for e ciency and remedial interventions for ine cient systems.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**26**	

		date forecast issued				
July 1	July 29	Sept. 2	Sept. 30	Oct. 31	Nov. 6	Outcome
0.425	0.550	0.475	0.198	0.081	0.047	**0**

Table 1: Probability forecasts issued in 2012 by the New York Times fivethirtyeight blog for the Republican party to win a Senate majority in the 113^{th} US Congress.

be updated more frequently as the election approaches, forming a probability sequence for the event that the Republicans win the Senate majority in the 114^{th} US Congress. A similar algorithm was used to forecast the 2012 US national elections, and issued probability forecasts at least weekly starting in July 2012 until the eve of the election.[2] For example, six of the forecasts for a Republican Senate majority in the 113^{th} Congress are given in Table 1. Wang and Campbell (2013) describe the fivethirtyeight forecasting system and other statistical modeling for election outcomes including their own.

Another familiar example of a probability sequence is the National Hurricane Center's (NHC's) probability forecast for winds exceeding a given threshold, generated every six hours over the course of the storm (DeMaria et al. 2009). The wind-speed probability product estimates the probability of one-minute sustained winds exceeding each of three thresholds within six-hour (or longer) periods for each half-degree cell in the region. For example, the NHC forecasts for tropical-storm force winds (winds exceeding 34 knots) for 2012's Hurricane Sandy affecting New York City for two 24-hour periods are given in Table 2. Two sequences are shown, for two distinct (though not independent) events.

In contrast to the outcome of an election, whose resolution is clearly tied to a specific time (election day), probability forecasting for time-series variables highlights the importance of the distinction between a rolling-horizon forecast and a fixed-event forecast. Every 24-hour period has a maximum sustained

[2] http://fivethirtyeight.blogs.nytimes.com/fivethirtyeights-2012-forecast/, accessed May 2014.

	date forecast issued					
event date	Oct. 25	Oct. 26	Oct. 27	Oct. 28	Oct. 29	Oct. 30
Oct. 28	0.06	0.06	0.11	0.04	**0**	
Oct. 29	0.14	0.31	0.42	0.80	0.92	**1**

Table 2: NHC's probability sequence issued at 11 am EDT each day for winds exceeding 34 knots in NYC between 8 am EDT on the event date indicated and 8 am the following day. Boldface indicates the outcome.

date forecast issued						
Oct. 24	Oct. 25	Oct. 26	Oct. 27	Oct. 28	Oct. 29	Oct. 30
<0.01	<0.01	0.06	0.42	<0.01	<0.01	<0.01

Table 3: NHC's probability forcasts issued at 11 am EDT for winds exceeding 34 knots in NYC between 8 am EDT two days later and 8 am three days later. This is an example of a rolling-horizon forecast and is not a probability sequence as defined in this paper.

wind in NYC. The event that the maximum sustained wind exceeds 34 knots between 8 am on October 28 and 8 am on October 29 is a distinct event from the event that the maximum sustained wind exceeds 34 knots between 8 am on October 29 and 8 am on October 30. In the context of forecast sequences, a sequence is a set of forecasts for an event whose definition does not change. We are interested in the relationships among multiple forecasts for the fixed event, not in relationships among, for example, the 48-hour lead forecasts for maximum sustained winds on different dates. Table 3 shows the rolling-horizon forecast for maximum sustained wind speeds exceeding 34 knots at NYC during the period approximately 45 to 69 hours after the 11 am issuance of the forecast. Note that the forecasts issued October 26 and 27 correspond to the fixed event forecast for the event dates October 28 and 29, respectively, shown in Table 2.

A *forecasting system* is a process for generating a probability forecast for many similar events, such as many elections or winds experienced in many storms, or many locations or verifying time periods. The forecasting process may be subjective, statistical, dynamical, or some combination of the above.

An example of subjective forecasts are the individual probabilistic forecasts for ranges of change in economic variables such as national gross domestic product and price levels collected in periodic Survey of Professional Forecasters, in the US by the Federal Reserve Bank, and in Europe by the European Central Bank. Another example of subjective forecasts are individual probabilistic forecasts for various world events elicited as part of the Intelligence Advanced Research Projects Activity (IARPA)'s Advanced Contingent Estimation Program,[3] one instance of which is described in Mellers et al. (2014).

Purely statistical forecasting systems, such as the fivethirtyeight election forecasting system, use historical and event-specific data to estimate the probability of events. Dynamical forecasting systems, such as numerical weather prediction models, may be used in simulation mode, as in ensemble-based forecasting systems, to produce probability forecasts (Sivillo, Ahlquist, and Toth 1997). Prediction market prices may also be interpreted as probabilistic forecasts reflecting a consensus of market participants, although they are not necessarily calibrated to the market participants' mean subjective probability (Manski 2006).

Many forecasting systems are combinations of subjective and statistical forecasts — for example averaging or otherwise post-processing subjective forecasts from many individuals, as in Baron, Mellers, Tetlock, Stone, and Ungar (2014), to produce a distinct forecasting system. The system consists of the method for eliciting individual subjective forecasts together with the aggregation process. Most meteorological ensemble-forecasting systems also include post-processing to produce a probabilistic forecast, for example adjusting for underdispersion in the ensemble.

We consider the evaluation of a forecasting system, theoretically capable of issuing forecasts for an infinite set of events. For a binary event, a single-period

[3] http://www.iarpa.gov/index.php/research-programs/ace, accessed June 2014.

probabilistic forecasting system may be represented as a joint probability distribution $f_{P,X}(p,x)$ of two random variables, a forecast P and an associated outcome X, with realizations $p \in [0,1]$, and $x \in \{0,1\}$. There is a large literature addressing the question of how to evaluate a single-period probabilistic forecasting system, or a sample from such a system. One criterion that is widely, perhaps universally, advocated is reliability. A forecasting system is perfectly *reliable* if $P(X = x|p) = p \ \forall p, x$. The degree of reliability can be measured in a number of different ways, and usually it is not measured separately from the second primary criterion, discrimination.

Discrimination, also called sharpness or resolution, is the ability of the forecasting system to differentiate among events by assigning high and low probabilities, relative to the base rate, $E[X]$. Discrimination is only valuable if paired with a high degree of reliability. At the extreme, a forecasting system that randomly issues forecasts of zero and one has high discrimination but the forecasts would be uncorrelated with the outcomes. The reverse is also true: a forecasting system that always forecasts $p = E[X]$ would be perfectly reliable but uninformative to anyone who knew the base rate.

There are many ways to measure imperfect reliability and discrimination. Many scoring functions, often called scoring rules, have been proposed for evaluating the overall performance of a forecasting system (or a sample). They combine the measure of reliability and discrimination into a scalar value, and take the form $s(p,x)$. When x has a finite number of values, the function can be separated, and for binary X, $s(p,x) = s_1(p) + s_0(p)$, where $s_1(p) = s(p,1)$ and $s_0(p) = s(p,0)$. A scoring function is positively (negatively) oriented if higher (lower) scores reflect better performance, and therefore $s_1'(p) > (<)0$ and $s_0'(p) < (>)0$.

A large literature discusses the advantages and disadvantages of various scor-

ing functions. Some of this work addresses the utility of the forecasting system to a user in a decision context (Murphy and Ehrendorfer 1987; Granger and Pesaran 2000; Jose, Nau, and Winkler 2008). Other work addresses desirable properties of a forecasting system based on axiomatic appeals (Selten 1998; Bröcker and Smith 2007; Winkler 1994; Jose, Nau, and Winkler 2009; Gneiting and Raftery 2007).

The best choice of a scoring function depends on the purpose of the score and the forecast. If the scoring function is used to evaluate human forecasters (Johnstone, Jose, and Winkler 2011) or alternative models whose designers are rewarded according to the function (Gneiting and Raftery 2007), and therefore creates an incentive scheme, then an important criterion is that the scoring rule is (strictly) proper meaning that the score is (uniquely) maximized by a probabilistic forecast equal to the forecaster's subjective probability, r, of the event. Mathematically, $s(p, x)$ is *strictly proper* if it satisfies(1). Even properness is not a universally accepted criterion for selecting a scoring function. Bickel (2007) points out that if the forecaster's utility function is not linear in the scoring function, a proper scoring function may not create the desired incentives. The meteorology community commonly uses improper skill scores in evaluating its forecasts, to capture the improvement in forecast relative to a baseline, as in **?**) which uses the Brier skill score to evaluate the NHC's wind-speed probability forecasting system. Research in evaluating single-period probability forecasts is active in the meteorology, decision science, and economics forecasting literature.

$$\text{definition of proper } s(\,):\ E_r\left[s(r,x)\right] > E_r\left[s(p,x)\right]\ \forall p \neq r,\ p, r \in [0,1] \qquad (1)$$

A few researchers have empirically compared the performance of the same forecasting system at different leads. For example, Mellers et al. (2014) allowed subjective forecasters to revise their probability forecasts over time before an

event was resolved, and found that subjective forecasts elicited in their experiments improve on average between the first and last week that the forecasters were able to forecast. Clements (2004) compared U.K. Monetary Policy Committee's forecasts of the probability of inflation exceeding 2.5% in the current quarter, the next quarter, and one year ahead, and found improvements in Brier score and log probability score as the event drew closer. We have not found any empirical examination of the relationships among the forecasts in a sequence, such as a search for inter-period correlation or trends, or any prescriptive analysis of probability sequences.

Nor are we aware of any research addressing appropriate scoring functions for probability sequences. It is interesting to note that Selten (1998) clearly thought of each single-period forecast as part of a sequence, but nevertheless limited his definition of a scoring rule to "measur[ing] the predictive success of a period for every period separately" (p. 44).

In the next section, we introduce notation and definitions for sequences. In Section 3, we propose an efficiency criterion for probability sequences, and use it to derive properties of an efficient probabilistic forecasting system, discussing some of their implications. Finally, we conclude with a discussion of the potential use of these results for diagnosing and remediating inefficiency in sequence-forecasting systems, both subjective and model-based.

2 Sequences

For a probability sequence, a forecasting system is a joint probability distribution over a finite ordered sequence of T probabilistic forecasts, indexed by t. Forecasts $p_t \in [0, 1]$ are realizations of the corresponding random variables P_t. For convenience, we will also write the vector $\mathbf{P}_t = (P_T, P_{T-1}, \ldots, P_t)$ and its realization $\mathbf{p}_t$.

Treating larger t as indicating a greater chronological distance from the outcome, t declines over a sequence, $t = T, \ldots, 1$ with forecast T being the first and $t = 0$ corresponding to the actual outcome of the event being forecast. The ordering of the sequence can also be interpreted as conditioning on successive subsets of the sample space, with no necessary relationship with timing. However, we rely on the assumption that any information available to the generation of forecast τ is also available for $t < \tau$.

We disallow perfect correlation between any pair P_t and P_τ, $t \neq \tau$ because perfect correlation effectively reduces the system to at most a $T - 1$-period forecasting system. This reasoning will be discussed in more detail in Section 3.1.

The forecasting system may be denoted $f_{P_T, P_T - 1, \ldots, P_1, X}(\,) = f_{\mathbf{P}_1, X}(\,)$, and may be used to describe functions of x and $\mathbf{p}_t$, such as $P(X = 1|\mathbf{p}_t)$ and $E\left[s\left(P_t, X\right)\right]$.

3 An efficiency condition for probability sequences

Criteria that apply to single-period probability forecasts, such as reliability, should also apply to probability sequences. However, sequences should satisfy additional criteria. For example, forecasts should improve over time with respect to a single-period scoring function, as information available to support forecasts in early periods is also available in later periods. An important question is how to determine whether they are improving *enough*.

As a basis for developing specific performance criteria for sequences, we use a minimal condition for a good forecasting system: it cannot be bootstrapped. For a *bootstrap-proof* sequence, in each period t, it is impossible to write a function with the earlier forecasts in the sequence itself as the only arguments that will score better in expectation than the last forecast in the sequence.

This is analogous to the criterion of weak efficiency for fixed event forecasts as defined by Nordhaus (1987) in the forecasting of non-probabilistic economic time series. Specifically, a forecast is *weakly efficient* if the forecast minimizes the expected error conditional on forecasts issued previously.

We adopt the term efficient to describe a bootstrap-proof system, and in the context of probability sequences, we operationalize this condition with the following axiom:

efficiency axiom

Given a strictly-proper positively-oriented single-period scoring function $s(p, x)$, a sequence forecasting system is *efficient* if $\forall t, p_t, \nexists g\left(\mathbf{p}_t\right)$ s.t.

$$E\left[s\left(g\left(\mathbf{p}_t\right), X\right)|\mathbf{p}_t\right] > E\left[s\left(p_t, X\right)|\mathbf{p}_t\right] \tag{2}$$

The efficiency axiom says that, given complete knowledge of the forecasting system, its performance cannot be improved by adjusting its forecasts based only on the earlier forecasts in the sequence itself. This is true not just in expectation over all possible values of $\mathbf{P}_t$, but conditional on any particular sequence $\mathbf{p}_t$.

Note that if $s(p, x) = -(p - x)^2$, a positively-oriented single-period function equal to the negative Brier score (Brier 1950), then the our efficiency condition is exactly equivalent to Nordhaus (1987)'s condition of weak efficiency for the forecast of a binary variable. Nordhaus (1987)'s standard of efficiency is that the forecast minimizes the squared error, whereas the efficiency condition above is defined with respect to any strictly proper scoring function.

Next, we derive properties of efficient sequence-forecasting systems.

3.1 Reliable and memoryless

We call a sequence-forecasting system is *each-period reliable* if it satisfies (3) and *memoryless* if it satisfies (4).

$$\text{each-period reliability: } P(X=1|\mathbf{p}_t) = p_t \quad \forall t, \mathbf{p}_t \tag{3}$$

$$\text{memorylessness: } P(X=1|\mathbf{p}_t) = P(X=1|p_t) \quad \forall t, \mathbf{p}_t. \tag{4}$$

Proposition 1 *An efficient sequence-forecasting system is each-period reliable and memoryless.*

Proof By definition of a strictly proper scoring function, if (3) is violated, then $g(\mathbf{p}_t) = P(X=1|\mathbf{p}_t)$ satisfies (2) and the system is inefficient.

Since (3) has been shown for an efficient system, the conditioning on $\mathbf{p}_{t+1}$ may be suppressed as in (4), and an efficient system is memoryless. ■

Each-period reliability is appealing as a performance criterion for sequences because reliability is a standard of optimality for single-period probability forecasts. An unreliable forecasting system is called poorly calibrated and if the properties of the system are known, it can be calibrated in post-processing: a new forecast $p^* = g(p) = P(X=1|p)$ can be issued that is reliable.

Memorylessness is perhaps less intuitive. An efficient system is fully updated at each period in that prior forecasts contain no additional information independent the most recent forecast p_t. If this were not true, the forecasting system could be bootstrapped.

Some readers may find the implication of (4) that $P(X=1|p_T=0.9) = P(X=1|p_1=0.9)$ counterintuitive because a forecast of $p=0.9$ (if far from the base rate) is highly informative, and if $T>1$, more information is anticipated to become available between time T and time 1. We would expect, however, that an extreme forecast (far from the base rate) is much less common at time

T than at time 1. If it occurs, it is realiable, but it is unlikely to occur, due to lower discrimination of early forecasts, as discussed further in Section 3.4.

We specified in the definition of a forecasting system that no two periods' forecasts may be perfectly correlated. Perfect correlation between any P_t and P_τ for $t \neq \tau$, together with a reliability constraint (3) on each period, implies that $p_t = p_\tau$ always, and therefore the forecasts are identical, and the system is at most a length-T sequence of distinct probability forecasts.

3.2 Inter-period reliable

A further implication of each-period reliability is inter-period reliability. We call a forecasting system *inter-period reliable* if it satisfies (5).

$$\text{inter-period reliability: } E\left[P_t|p_\tau\right] = p_\tau \ \forall t < \tau. \tag{5}$$

Proposition 2 *An efficient forecasting system is inter-period reliable.*

Proof

$$p_\tau \overset{(3),(4)}{=} P\left(X = 1|p_\tau\right) \tag{6a}$$

$$= \int_0^1 P\left(X = 1|p_\tau, p_t\right) f_{P_t|P_\tau,}\left(p_t|p_\tau\right) dp_t \tag{6b}$$

$$\overset{(4)}{=} \int_0^1 p_t f_{P_t|P_\tau}\left(p_t|p_\tau\right) dp_t = E\left[P_t|p_\tau\right]. \tag{6c}$$

where (6b) uses the law of total probability. ■

Each-period reliability is also an intuitive property of an efficient system. If a system is not inter-period reliable, then a forecast $g(\mathbf{p}_\tau) = E\left[P_t|p_\tau\right]$ will satisfy (2).

Both inter-period reliability and each-period reliability holds for a forecasting system that is efficient with respect to *any* strictly proper scoring function. They

do not depend on the scoring function.

3.3 Unpredictable revisions

A direct consequence of inter-period reliability is unpredictability in revisions. Equation (7) gives specific statements of unpredictability in period-t revision $P_{t-1} - p_t$.

$$\forall \mathbf{p}_t, t > 1 :$$

$$\text{zero expected revisions: } E[P_{t-1} - p_t|\mathbf{p}_t] = 0 \tag{7a}$$

$$\text{zero expected autocorrelation: } E\left[(P_{t-1} - p_t)\,(p_t - p_{t+1})\,|\mathbf{p}_t\right] = 0,\ t < T \tag{7b}$$

Proposition 3 *For an efficient sequence-forecasting system, expected revisions and autocorrelation in revisions are always zero.*

Although these properties follow directly from inter-period reliability, some readers may find them counter-intuitive, so we offer a formal proof.

Proof

$$E[P_{t-1} - p_t|\mathbf{p}_t] \overset{(4)}{=} E[P_{t-1}|p_t] - p_t \overset{(5)}{=} 0 \Rightarrow (7a)$$

$$E\left[(P_{t-1} - p_t)\,(p_t - p_{t+1})\,|\mathbf{p}_t\right] \overset{(4)}{=} E\left[(P_{t-1} - p_t)\,|p_t\right](p_t - p_{t+1}) \overset{(7a)}{=} 0 \Rightarrow (7b)$$

■

The properties defined in (7) also imply the weaker but more familiar properties:

$$E[P_{t-1} - P_t] = 0 \text{ and} \tag{8a}$$

$$E\left[(P_{t-1} - P_t)\,(P_t - P_{t+1})\right] = 0 \tag{8b}$$

Neither trends (positive autocorrelation) nor noise (negative autocorrelation) is expected in forecast from an efficient system. In an efficient system, there is no predictability in revisions, conditional on prior forecasts. If there were, it could be exploited to improve the forecast.

This result may be counterintuitive to some readers. (Nordhaus 1987) offers an intuition for this result:

> efficient forecasts appear jagged because they incorporate all news quickly. Inefficient forecasts appear smoother and more consistent, for they let the news seep in slowly. (p. 669)

The fact that these properties are not universally intuitive indicates that subjective probabilistic forecasts may violate this criterion and therefore be susceptible to improvement as discussed further in Section 4.

3.4 Strictly Improving

A forecasting system is *strictly improving* with respect to a positively-oriented scoring function $s(\,)$ if it satisfies (9).

$$\text{strict improvement: } t < \tau \Rightarrow E\left[s(P_t, X)|p_\tau\right] > E\left[s(p_\tau, X)\right], \forall p_\tau,\ \tau > 1 \quad (9)$$

Proposition 4 *An efficient sequence-forecasting system is strictly improving with respect to* any *strictly proper scoring function.*

Proof The proof of depends on the strict convexity of $E\left[s(p, X)|p\right]$ for proper $s(\,)$ and reliable P, which is shown in Appendix 1.1. For an efficient forecasting system, P_τ is reliable, and therefore for proper $s(\,)$, by Jensen's inequality,

$$E\left[s\left(P_t, X\right)|p_\tau\right] > E\left[s\left(E\left[P_t|p_\tau\right], X\right)\right] \overset{(5)}{=} E\left[s\left(p_\tau, X\right)\right] \Rightarrow (9)$$

■

Note that (9) is a stricter condition than $t < \tau \Rightarrow E\left[s(P_t, X)\right] > E\left[s(P_\tau, X)\right]$, which is also true for an efficient forecasting system. By (9), the expected score for later periods improves conditional on every value of p_τ (and therefore conditional on every $\mathbf{p}_\tau$).

The strictly-improving property of efficient forecast systems means that a system that is efficient with respect to any strictly proper scoring function is strictly improving with respect to all strictly proper scoring functions.

Since each period's forecasts are perfectly reliable, the improvement must come from better discrimination. Different scoring functions measure discrimination differently. Conditional on reliability, the expected negative Brier score (which is positively oriented) is (10a) while the expected log score is (10b). As discussed in Section 4.2, the Brier score discrimination component is equal to the sample standard deviation, a common measure of dispersion. For an efficient forecast all of these measures of dispersion will be increasing in expectation over the forecast sequence. Although large dispersion is usually associated with a less informative forecast, in this context, dispersion increases with the probability of extreme (far from the base rate) forecasts that are also reliable, and therefore informative.

$$\text{Brier score discrimination:} \quad -\int_0^1 f(p)p(1-p)dp \tag{10a}$$

$$\text{log score discrimination:} \quad \int_0^1 f(p)\left(p\,ln(p) + (1-p)ln(1-p)\right)dp \tag{10b}$$

		P_1	
		$p_1 = 0.8$	$p_1 = 0.2$
P_2	$p_2 = 0.6$	0.160 / 0.040	0.020 / 0.080
	$p_2 = 0.4$	0.187 / 0.047	0.093 / 0.373

(a) Forecasting system 1.

		P_1	
		$p_1 = 0.8$	$p_1 = 0.2$
P_2	$p_2 = 0.6$	0.120 / 0.000	0.060 / 0.120
	$p_2 = 0.4$	0.227 / 0.087	0.053 / 0.333

(b) Forecasting system 2.

Table 4: Two forecasting systems. The values shown are the joint probabilities $P(P_2 = p_2, P_1 = p_1, X = 1) \,/\, P(P_2 = p_2, P_1 = p_1, X = 0)$. Recall that forecast P_1 is issued after forecast P_2.

4 Discussion

4.1 Diagnosing inefficiency: an example

Each of the properties of efficient sequence-forecasting systems derived in Section 3 is testable and may be used to diagnose inefficiency. A simple example illustrates this for a forecasting system whose complete joint probability distribution is known. Table 4 shows two two-period forecasting systems for a binary event, each with a base rate, $E[X] = 0.460$. Table 4 gives the distribution of the forecasts and outcome.

Both systems are each-period reliable, i.e. $P(X = 1|p_t) = p_t$ for each value of t and of p_t. For example, for System 1, shown in Table 4a,

$$P(X = 1|p_2 = 0.6) = \frac{0.160 + 0.020}{0.160 + 0.040 + 0.020 + 0.080} = 0.6.$$

Both systems are also improving, with negative Brier scores -0.240 for P_2 and -0.160 for P_1 (the negative Brier score is positively oriented), so the later forecast, P_1 is better than the earlier forecast as measured by the Brier score.

Both systems are improving with respect to any single-period score. Since the two systems have identical marginal probabilities $P(P_t = p_t)$ and conditional probabilities $P(X = x|P_t = p_t)\ \forall t, p_t, x$, and therefore they have identical single-period expected scores $E\left[s(P_t, X)\right]$ regardless of the choice of $s(\,)$.

However, System 1 is efficient while System 2 is not. This can be diagnosed by testing for inter-period reliability. For System 1, $E[P_1|p_2 = 0.6] = 0.6$ and $E[P_1|p_2 = 0.4] = 0.4$, while for System 2, $E[P_1|p_2 = 0.6] = 0.440$ and $E[P_1|p_2 = 0.4] = 0.469$. This means System 2 is bootstrappable as illustrated in Section 4.3.

4.2 Diagnosing inefficiency based on a sample

The properties of an efficient forecasting system suggest statistical tests for efficiency in the more common situation in which a sample from the forecasting system is available, but not the complete distribution. One of the challenges in evaluating single-period probabilistic forecasts is the data requirements, a challenge that is compounded for probability sequences. A sample from a sequence-forecasting system consists of N observations where the n^{th} observation is $(\mathbf{p}_{t,n}, x_n)$, consisting of T forecasts $p_{t,n}$ plus the outcome x_n, for a total of $N \times (T + 1)$ observed values.

However, a sample from a sequence forecasting system may be tested for violations of the efficiency properties based on one or two periods' forecasts. The inter-period reliability property and its corollaries, zero expected revisions and zero autocorrelation, are especially interesting because they depend only on the inter-period covariance structure of the forecasts and do not depend on the relationship with the outcome X. Moreover, a forecasting system in which realizations of the outcome are costly, delayed, or otherwise difficult to assess, may be evaluated with respect to violations of these properties *without*

data about the actual outcomes. In the context of NHC wind-speed probability forecasts, for example, it is sometimes impossible to be sure whether winds exceeded a given threshold because the maximum sustained wind might not have occurred at a functional measuring station.

For example, a null hypothesis of zero autocorrelation in revisions may be tested (for a given t) simply by regressing the t-period revisions against the $t-1$-period revisions, and using a Student's[4] t-test for a linear relationship, interpreting the p-value as the probability of rejecting the null hypothesis (zero autocorrelation) if it holds.

In the remainder of this section, we propose statistics that may be used as the basis of tests of efficiency. Common approaches for evaluating single-period probability forecasts class forecasts into discrete bins. Following this convention, we classify $p_{t,n}$ into discrete bins $j = 1, \ldots, J$, defined identically $\forall t$ (although it is not necessary for them to be identical). We define $b(p)$ as the index of the bin that probability p falls into, $b(p) \in \{1, \ldots, J\}$, and

$$q_t(j) = \sum_{n:b(p_{t,n})=j} x_n,$$

equivalently the relative frequency of the event $X = 1$ conditional on the forecast p_t falling in bin j. $q_t(j)$ is also equal to the within-sample reliability-calibrated forecast for bin j.

Each-period reliability (3) may be tested for each t using the same methods used for testing single-period probabilistic forecasts. Statistical tests for the null hypothesis of forecast reliability (which could be applied for a given t to test for each-period reliability) have been suggested. Bröcker (2012) suggests a family of confidence intervals on $q(j)$ over J bins as a test for single-period reliability. A family test for each-period reliability over many periods, based on the null of

[4] We refer to the Student's t-test to avoid confusion with period-t.

each-period reliability could be developed.

The statistic in (11) is a single-period measure of inter-period unreliability. Under the null hypothesis of efficiency, its expected value is zero for each t, $1 < t \leq T$.

inter-period reliability statistic:

$$\frac{1}{N}\sum_{j=1}^{J}\left(\sum_{n:b(p_{t,n})=j}\left(p_{t,n}-q_{t-1}\left(b(p_{t-1,n})\right)\right)^2\right) \qquad (11)$$

where $q_{t-1}\left(b(p_{t-1,n})\right)$ is the reliability-calibrated $t-1$ forecast

Testing whether a forecast is improving, even for a single pair of periods t and $t-1$, is not entirely straightforward. However, as discussed in Section 3.4, if a forecasting system is reliable, improvement must come from discrimination. Moreover, given enough data, each-period reliability may be readily enforced by post-processing. Therefore, we might prefer a statistic that separates the measure of discrimination from the measure of reliability. Conveniently, per Proposition 4, an efficient forecast is improving with respect to any strictly proper scoring function. This suggests that a test for the difference in the discrimination component of any scoring function would be a useful test for the strictly improving property (sufficiency of later forecasts).

The Brier score may be decomposed into a reliability, discrimination, and variability components, as in Wilks (2011) (p. 333). The sample estimate of the discrimination components is shown in (12). Large values of (12) lead to better Brier scores. Using $q_t(j)$ instead of $p_{t,n}$ means that (12) estimates the standard deviation of the reliability-calibrated forecast. For an efficient forecasting sequence, therefore, (12) should be decreasing in t. A null hypothesis that (12) is not improving (the system is inefficient) would assume that (12)

evaluated at τ and at $t < \tau$ are equal.

$$\text{discrimination statistic: } \frac{1}{N}\sum_{j=1}^{J} m_t(j)\left(q_t(j) - \bar{x}\right)^2 \tag{12}$$

$$\text{where } m_t(j) = \sum_{n:b(p_{t,n})=j} 1 \quad \text{and} \quad \bar{x} = \sum_{n=1}^{N} x_n$$

At the limit as J increases, (12) is equal to the sample standard deviation of the reliability-calibrated forecast. For a reliable forecast, the sequence is improving in discrimination if its standard deviation is increasing, and statistical tests for differences of standard deviation may be applied for each t.

The statistics in (12) and (11) can be used as the basis for designing hypothesis tests for each t. For multiple single-period tests, an appropriate test of the family of inferences could be designed. Novel statistical tests for efficiency of sequence-forecasting systems may be a productive area for future research.

The diagnosis of inefficiency is essentially a search for unexploited structure in the forecasts. Any predictability that is not captured in the forecast can be used in post-processing, or point to a way to improve the forecasting system. Patterns such as conditional bias that might not show up in an average over j, may also be of interest, such as trends that are a function of p_t.

4.3 Remediation

Once diagnosed, inefficiency can be remedied in post-processing. For subjective forecasts, post-processing, such as debiasing to correct for over-precision, is already a common intervention for subjective single-period probability forecasts. Similarly, in ensemble-based meteorological forecasting, a correction for under-dispersion (insufficient spread) in the ensemble is common. In the context of probability sequences, empirical or parametric functions of the form

		P_1	
		$p_1 = 0.8$	$p_1 = 0.2$
P_2	$p_2 = 0.6$	1.000	0.333
	$p_2 = 0.4$	0.723	0.138

Table 5: Conditional probabilities for System 2 using both periods' forecasts. The values shown are $P(X = 1|p_2, p_1)$.

$p_t^*(\mathbf{p}_t) = P(X = 1|\mathbf{p}_t)$ estimated from the available sample from the forecasting system may be used to generate better forecasts, such that $E\left[s(p^*(\mathbf{p}_t), X)\right] > E\left[s(p_t, X)\right]$.

$$\text{post-processing function: } p_t^*(\mathbf{p}_t) = P(X = 1|\mathbf{p}_t) \tag{13}$$

For example, System 2 in Table 4b can be improved in period $t = 1$ by post-processing using the information from P_1 and P_2. Table 5 shows the conditional probabilities $P(X = 1|p_1, p_2)$ for System 2. A modified System 2 using the probabilities in Table 5 in place of p_1, i.e. $p_1^*(\mathbf{p}_1) = P(X = 1|p_2, p_1)$, has a period-1 negative Brier score of -0.149, which is better System 2's period-1 negative Brier score of -0.160.

Perhaps more interesting, a diagnosis of inefficiency implies that the forecasting system does not fully incorporate the available information and is susceptible to improvement and therefore may benefit from a re-examination of the forecasting process, motivated and informed by the results of efficiency tests.

For subjective forecasts, diagnostic test results could be provided to the forecasters as feedback to help them self-calibrate and motivate a search for more valid predictors or better synthesis of information.

For dynamical model-based forecasts, diagnostic tests showing inefficiency suggest a search for improving the underlying dynamical model or the addition of statistical post-processing to exploit any detectable structure in the system. For statistical models, the implications are the same—for example, it may sug-

gest the possibility of a prior distribution that the model does not capture, or predictors used in early forecasts that provide independent information that could be exploited in later forecasts.

4.4 Implications for scoring functions

Scoring functions that are separable functions of single-period scores do not diagnose violations of efficiency. For example, a separable scoring function would not detect inter-period structure such as regression to the base rate or other trends that could be removed in post-processing. In fact, the two systems shown in Table 4 have identical marginal distributions $P(P_t = p_t),\ \forall t, p_t$, therefore would have identical single-period scores regardless of the scoring function. Tests that depend on the correlation structure are necessary to diagnose efficiency.

This suggests that sequence-scoring functions of a form that depend on the inter-period structure of the system, which excludes functions of the form (14), may be desirable. The question of which functions are appropriate in a given situation is open.

$$\sum_{t=1}^{T} w_t \frac{1}{N} \sum_{n=1}^{N} s(p_{t,n}, x_n) \tag{14}$$

After more than 60 years of research on probabilistic forecasting, there is no consensus on the best single-period scoring function, and ongoing research explores the relative merits of various scoring functions in particular contexts. For example, an interesting difference arises from the consideration of probability forecasts used to make investment decisions in competitive markets, where the purchase price of an asset depends on the market (or a competitor's) probability forecast, and small differences in extreme (small or large) probabilities are very important. Hence Jose, Nau, and Winkler (2008) do not adopt Selten (1998)'s criterion that a scoring rule should not be hypersensitive (defined p. 50) to

small differences. The meteorology community commonly uses skill scores that are not proper in evaluating its forecasts (Gneiting and Raftery 2007). Skill scores adjust for a measure of difficulty in making the forecast, often a reference or baseline forecasting system. In that context, the benefits of this adjustment outweigh any incentive for untruthful forecasting (or biased models). We anticipate that a similar breadth of multi-period scoring functions would be useful and appropriate for different forecasting and decision contexts.

The decision context in which the forecasting system might be used could determine the relative importance of its performance over t, for example, informing the selection of a scoring function. Moreover, it is very possible that a decision maker could benefit more from using a particular inefficient forecasting system, instead of an alternative efficient forecasting system. However, the inefficient forecasting system is clearly improvable, and therefore although it may be superior in user value to other efficient forecasting systems, it should not satisfy a user (or, for that matter, a forecaster) because its inefficiency demonstrates that it is possible to create a system that will perform even better.

5 Conclusion

Probability sequences are becoming more common and relevant to decision makers in many domains, including finance and intelligence. To our knowledge, this is the first formal examination of the inter-period behavior of probability sequences. We showed several properties of an efficient sequence forecasting system, i.e. that it is reliable for each period, inter-period reliable and therefore memoryless, always has zero expected revision, and strictly improving with respect to any strictly proper scoring function. Some of these properties are intuitive, but others—in particular inter-period reliability and zero autocorrelations—are not. Analogous conditions of efficiency are commonly vi-

olated in forecasting of non-probability variables and therefore we should expect to find probability sequences that violate these properties and are susceptible to improvement.

References

Baron, J., B. A. Mellers, P. E. Tetlock, E. Stone, and L. H. Ungar (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis 11*(2), 133–145.

Bickel, J. E. (2007). Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis 4*(2), 49–65.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*(1), 1–3.

Bröcker, J. (2012). Probability forecasts.

Bröcker, J. and L. A. Smith (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting 22*(2), 382–388.

Clements, M. P. (2004). Evaluating the Bank of England density forecasts of inflation. *The Economic Journal 114*(498), 844–866.

DeMaria, M., J. A. Knaff, R. Knabb, C. Lauer, C. R. Sampson, and R. T. DeMaria (2009). A new method for estimating tropical cyclone wind speed probabilities. *Weather & Forecasting 24*(6), 1573–1591.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Granger, C. W. and M. H. Pesaran (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting 19*(7), 537–560.

Johnstone, D. J., V. R. R. Jose, and R. L. Winkler (2011). Tailored scoring rules for probabilities. *Decision Analysis 8*(4), 256–268.

Jose, V. R. R., R. F. Nau, and R. L. Winkler (2008). Scoring rules, generalized entropy, and utility maximization. *Operations Research 56*(5), 1146–1157.

Jose, V. R. R., R. F. Nau, and R. L. Winkler (2009). Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science 55*(4), 582–590.

Manski, C. F. (2006). Interpreting the predictions of prediction markets. *Economics Letters 91*(3), 425–429.

Mellers, B., L. Ungar, J. Baron, J. Ramos, B. Gurcay, K. Fincher, S. E. Scott, D. Moore, P. Atanasov, S. A. Swift, et al. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science 25*(5), 1106–1115.

Murphy, A. H. and M. Ehrendorfer (1987). On the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Weather and Forecasting 2*(3), 243–251.

Nordhaus, W. D. (1987). Forecasting efficiency: Concepts and applications. *The Review of Economics and Statistics 69*(4), 667–674.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics 1*(1), 43–62.

Sivillo, J. K., J. E. Ahlquist, and Z. Toth (1997). An ensemble forecasting primer. *Weather and Forecasting 12*(4), 809–818.

Wang, S. and B. C. Campbell (2013). Mr. Bayes goes to Washington. *Science 339*(6121), 758–759.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (3 ed.). Academic Press.

Winkler, R. L. (1994). Evaluating probabilities: Asymmetric scoring rules. *Management Science 40*(11), 1395–1405.

1 Appendix

1.1 Convexity of expected score

For differentiable $s(\,)$, properness (1) implies that $\frac{d}{dp}E_r\left[s(p,X)\right]=0$ when evaluated at $p=r$, and therefore, for propert $s(/,)$,

$$\frac{s_1'(p)}{s_0'(p)}=1-\frac{1}{p}\Rightarrow s_1'(p)=s_0'(p)-\frac{s_0'(p)}{p}. \tag{15}$$

For reliable P,

$$\begin{aligned}
E\left[s(p,X)|p\right]&=E_p\left[s(p,X)\right]=ps_1(p)+(1-p)s_0(p)\\
\frac{d}{dp}E\left[s(p,X)|p\right]&=ps_1'(p)+s_1(p)+(1-p)s_0'(p)-s_0(p)\\
&\overset{(15)}{=}p\left(s_0'(p)-\frac{s_0'(p)}{p}\right)+s_1(p)+(1-p)s_0'(p)-s_0(p)\\
&=s_1(p)-s_0(p)
\end{aligned}$$

For positively-oriented $s(\,)$, recall that $s_1'(p)>0$ and $s_0'(p)<0$, so

$$\frac{d^2}{dp^2}E\left[s(p,X)|p\right]=s_1'(p)-s_0'(p)>0$$

and therefore, for positively-oriented, proper $s()$ and reliable P, $E\left[s(p,X)|p\right]$ is convex in p.