## Prequential Analysis

Philip Dawid

University of Cambridge

NIPS 2008 Tutorial

Forecasting	2
	כ ⊿
Une-step Forecasts	4
	5
Some comments	6
Forecasting systems	7
Probability Forecasting Systems.	8
Statistical Forecasting Systems	9
Prequential consistency	10
Absolute assessment	11
Weak Prequential Principle.	12
Calibration	13
Example	14
Calibration plot	15
Computable calibration	16
Well-calibrated forecasts are essentially unique.	17
Significance test	18
Other tests	-0 10
	20
	20
Prequential frame of reference	21
Combining contingency tables	22
Sequential observation	23
Sequential observation — with drop-out	24
Censored survival data [1].	25
Censored survival data [2]	26

Censored survival data [3]	27
Game-theoretic probability	28
Sequential prediction of binary variables	29 30
Almost sure event	31
Prequential probability	32
Comparative assessment	33
Loss functions and scoring rules.	34
Examples:	35
Single distribution P	36
Likelihood	31
	20 20
	39
Prequential efficiency	40
Efficiency	41
Model testing	42
Model choice	43
Prequential consistency	44
Out-of-model performance	45
Conclusions	46
Conclusions	47
References	48
References	49

## Forecasting

#### **Context and purpose**

*Prequential* = [Probabilistic]/Predictive/Sequential

— a general framework for assessing and comparing the predictive performance of a FORECASTING SYSTEM.

□ We assume reasonably extensive data, that either arrive in a time-ordered stream, or can be can be arranged into such a form:

$$\mathbf{X} = (X_1, X_2, \ldots).$$

- $\hfill\square$  There may be patterns in the sequence of values.
- □ We try to identify these patterns, so as to use currently available data to form good forecasts of future values.

Basic idea: Assess our future predictive performance by means of our past predictive performance.

3 / 49

#### **One-step Forecasts**

 $\Box$  Introduce the data-points  $(x_1, \ldots, x_n)$  one by one.

- $\Box$  At time *i*, we have observed values  $\mathbf{x}^i$  of  $\mathbf{X}^i := (X_1, \ldots, X_i)$ .
- $\Box$  We now produce some sort of forecast,  $f_{i+1}$ , for  $X_{i+1}$ .
- $\Box$  Next, observe value  $x_{i+1}$  of  $X_{i+1}$ .
- $\Box$  Step up *i* by 1 and repeat.
- $\Box$  When done, form overall assessment of quality of forecast sequence  $\mathbf{f}^n = (f_1, \ldots, f_n)$  in the light of outcome sequence  $\mathbf{x}^n = (x_1, \ldots, x_n)$ .

We can assess forecast quality either in absolute terms, or by comparison of alternative sets of forecasts.





#### **Forecasting systems**

#### **Probability Forecasting Systems**

Very general idea, e.g.:

No system: e.g. day-by-day weather forecasts

**Probability model:** Fully specified joint distribution P for  $\mathbf{X}$  (allow arbitrary dependence)

 $\Box$  probability forecast  $f_{i+1} = P(X_{i+1} \mid \mathbf{X}^i = \mathbf{x}^i)$ 

**Statistical model:** Family  $\mathcal{P} = \{P_{\theta}\}$  of distributions for **X** 

 $\Box$  forecast  $f_{i+1} = P^*(X_{i+1} | \mathbf{X}^i = \mathbf{x}^i)$ , where  $P^*$  is formed from  $\mathcal{P}$  by somehow estimating/eliminating  $\theta$ , using the currently available data  $\mathbf{X}^i = \mathbf{x}^i$ 

**Collection of models** *e.g.* forecast  $X_{i+1}$  using model that has performed best up to time *i* 

8 / 49

#### **Statistical Forecasting Systems**

—based on a statistical model  $\mathcal{P} = \{P_{\theta}\}$  for **X**.

**Plug-in forecasting system** Given the past data  $\mathbf{x}^i$ , construct some estimate  $\hat{\theta}_i$  of  $\theta$  (*e.g.*, by maximum likelihood), and proceed as if this were the true value:

$$P_{i+1}^*(X_{i+1}) = P_{\hat{\theta}_i}(X_{i+1} \mid \mathbf{x}^i).$$

NB: This requires re-estimating  $\theta$  with each new observation!

**Bayesian forecasting system (BFS)** Let  $\pi(\theta)$  be a prior density for  $\theta$ , and  $\pi_i(\theta)$  the posterior based on the past data  $\mathbf{x}^i$ . Use this to mix the various  $\theta$ -specific forecasts:

$$P_{i+1}^*(X_{i+1}) = \int P_{\theta}(X_{i+1} \mid \mathbf{x}^i) \,\pi_i(\theta) \,d\theta.$$

#### **Prequential consistency**

Gaussian process:  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ ,  $\operatorname{corr}(X_i, X_j) = \rho$ 

MLEs:

$$\hat{\mu}_n = \overline{X}_n \qquad \stackrel{L}{\to} \quad \mathcal{N}(\mu, \rho\sigma^2) 
\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2 \stackrel{p}{\to} (1 - \rho)\sigma^2 
\hat{\rho}_n = 0$$

- not classically consistent.

But the estimated predictive distribution  $\hat{P}_{n+1} = \mathcal{N}(\hat{\mu}_n, \hat{\sigma}_n^2)$  does approximate the true predictive distribution  $P_{n+1}$ : normal with mean  $\overline{x}_n + (1-\rho)(\mu - \overline{x}_n)/\{n\rho + (1-\rho)\}$  and variance  $(1-\rho)\sigma^2 + \sigma^2/\{n\rho + (1-\rho)\}$ .

nean  $x_n + (1 - \rho)(\mu - x_n)/\{n\rho + (1 - \rho)\}$  and variance  $(1 - \rho)\sigma^2 + \sigma^2/\{n\rho + (1 - \rho)\}$ . 10 / 49

#### Absolute assessment

11 / 49

Weak Prequential Principle

The assessment of the quality of a forecasting system in the light of a sequence of observed outcomes should depend only on the forecasts it in fact delivered for that sequence

- and not, for example, on how it might have behaved for other sequences.

#### Calibration

- $\Box$  Binary variables  $(X_i)$
- $\square$  Realized values  $(x_i)$
- $\Box$  Emitted probability forecasts  $(p_i)$

Want (??) the  $(p_i)$  and  $(x_i)$  to be close "on average":

 $\overline{x}_n - \overline{p}_n \to 0$ 

where  $\overline{x}_n$  is the average of all the  $(x_i)$  up to time n, etc.

Probability calibration: Fix  $\pi \in [0,1]$ , average over only those times *i* when  $p_i$  is "close to"  $\pi$ :

$$\overline{x}_n' - \pi \to 0$$

Exa	mple														
	Probability	0.4	0.6	0.3	0.2	0.6	0.3	0.4	0.5	0.6	0.2	0.6	0.4	0.3	0.5
	Outcome	0	0	1	0	1	0	1	1	1	0	1	0	0	1
															_
	Probability p			0.2			0.3		0.4		0.5		0.6		
	Ins	Instances n			2			3		3		2		4	
	Su r	Successes r			0	)	1		1		2		3		
	Pro	Proportion			0	)	0.3	3	0.	33		1		0.75	Ì
	ρ														-
														14	/ 49



#### **Computable calibration**

Let  $\sigma$  be a computable strategy for selecting trials in the light of previous outcomes and forecasts

- e.g. third day following two successive rainy days, where forecast is below 0.5.

Then require asymptotic equality of averages,  $\overline{p}_{\sigma}$  and  $\overline{x}_{\sigma}$ , of the  $(p_i)$  and  $(x_i)$  over those trials selected by  $\sigma$ .

#### Why?

Can show following. Let P be a distribution for  $\mathbf{X}$ , and  $P_i := P(X_i = 1 | \mathbf{X}^{i-1})$ . Then

$$\overline{P}_{\sigma} - \overline{X}_{\sigma} \to 0$$

P-almost surely, for any distribution P.

#### Well-calibrated forecasts are essentially unique

Suppose  ${\bf p}$  and  ${\bf q}$  are computable forecast sequences, each computably calibrated for the same outcome sequence  ${\bf x}.$ 

Then  $p_i - q_i \rightarrow 0$ .

17 / 49

Significance testConsider e.g. $Z_n := \frac{\sum (X_i - P_i)}{\{\sum P_i(1 - P_i)\}^{\frac{1}{2}}}$ where  $P_i = P(X_i = 1 \mid X^{i-1})$ .Then $Z_n \xrightarrow{L} \mathcal{N}(0, 1)$ for (almost) any P.So can refer value of  $Z_n$  to standard normal tables to test departure from calibration, even without knowing generating distribution P— "Strong Prequential Principle"

#### **Other tests**

Suppose the  $X_i$  are continuous variables, and the forecast for  $X_i$  has the form of a continuous cumulative distribution function  $F_i(\cdot)$ .

If  $\mathbf{X} \sim P$ , and the forecasts are obtained from P:

 $F_i(x) := P(X_i \le x \mid \mathbf{X}^{i-1} = \mathbf{x}^{i-1})$ 

then, defining

 $U_i := F_i(X_i)$ 

we have

 $U_i \sim U[0,1]$ 

independently, for any P.

19 / 49

So we can apply various tests of uniformity and/or independence to the observed values

$$u_i := F_i(x_i)$$

to test the validity of the forecasts made

— again, without needing to know the generating distribution P.















## Game-theoretic probability

### 28 / 49

# Sequential prediction of binary variables At successive times t = 1, 2, ...: (Optional) Nature N chooses (and reveals) value $w_t$ of $W_t$ Forecaster F chooses probability $p_t \in [0, 1]$ Adversary A chooses stake $h_t \in \mathcal{R}$ Nature N chooses value $x_t \in \{0, 1\}$ of $X_t$ F pays A $h_t(x_t - p_t)$ ("fair bet") $K_n := A$ 's accumulated fortune at time n (starting with 1) $= 1 + \sum_{t=1}^n h_t(x_t - p_t)$

#### **Full event**

Let C be a "prequential event" (possible property of infinite sequence  $(w_1, p_1, x_1, w_2, p_2, x_2, ...)$  of plays of N and F)

#### - automatically respects WPP

Call C full if A has a *strategy* that ensures:

1. 
$$K_n \ge 0$$
, all  $n$ , AND

2. either

(a) C holds;

(b) 
$$K_n \to \infty$$

Example:  $C = "n^{-1} \sum_{t=1}^{n} (x_t - p_t) \rightarrow 0"$  (calibration).

Almost sure event Can show that any full event has probability 1 under any joint distribution P for  $(W_1, X_1, W_2, X_2, ...)$ - so long as ("compatibility")  $p_t = P(X_t = 1 \mid W_1 = w_1, X_1 = x_1, ..., W_t = w_t)$ - a strong prequential property - but criterion is meaningful (and sensible) even in the absence of any P (which we may not be willing to specify) - justifies prequential compatibility criteria 31 / 49

#### **Prequential probability**

C a prequential event. For  $\beta > 0$ , say  $C \in S(\beta)$  if A has a strategy ensuring that, whenever C occurs, the process  $(K_t)$  reaches or exceeds  $1/\beta$  before ever having been negative.

 $\Box \ C \in S(1)$ 

 $\Box \ \beta < \beta' \Rightarrow S(\beta) \subseteq S(\beta')$ 

The (upper) prequential probability of C is

$$PP(C) := \inf\{\beta > 0 : C \in S(\beta)\}$$

 $\Box PP(C) \in [0,1]$ 

 $\Box$  C is full  $\iff$  its complement has prequential probability 0

 $\Box$  For any compatible probability distribution *P*,  $P(C) \leq PP(C)$ 

- justifies limiting normal-based tests, etc.

#### Loss functions and scoring rules

Measure inadequacy of forecast f of outcome x by

loss function: 
$$L(x, f)$$

Then measure of overall inadequacy of forecast sequence  $f^n$  for outcome sequence  $x^n$  is cumulative loss:

$$L^n = \sum_{i=1}^n L(x_i, f_i)$$

We can use this to compare different forecasting systems.

34 / 49

#### **Examples:**

Squared error: f a point forecast of real-valued X $L(x, f) = (x - f)^2$ .

Scoring rule: f a probability forecast Q for XL(x, f) = S(x, Q).

Logarithmic score:  $S(x,Q) = -\log q(x)$ , where  $q(\cdot)$  is the density function of Q. The logarithmic score is proper: for given P,  $S(P,Q) = E_P\{S(X,Q)\}$  is minimised by taking Q = P. Consider only this from now on.

#### **Single distribution** *P*

At time *i*, having observed  $\mathbf{x}^i$ , probability forecast for  $X_{i+1}$  is its conditional distribution  $P_{i+1}(X_{i+1}) := P(X_{i+1} \mid \mathbf{X}^i = \mathbf{x}^i)$ . When we then observe  $X_{i+1} = x_{i+1}$ , the associated logarithmic score is

$$S(x_{i+1}, P_{i+1}) = -\log p(x_{i+1} \mid \mathbf{x}^i).$$

So the cumulative score is

$$L_n(P) = \sum_{i=0}^{n-1} -\log p(x_{i+1} \mid \mathbf{x}^i)$$
$$= -\log \prod_{i=1}^n p(x_i \mid \mathbf{x}^{i-1})$$
$$= -\log p(\mathbf{x}^n)$$

where  $p(\cdot)$  is the joint density of **X** under *P*.

36 / 49

#### Likelihood

 $L_n(P)$  is just the (negative) log-likelihood of the joint distribution P for the observed data-sequence  $\mathbf{x}^n$ .

If P and Q are alternative joint distributions, considered as forecasting systems, then the excess score of Q over P is just the log likelihood ratio for comparing P to Q for the full data  $\mathbf{x}^n$ .

This gives an interpretation to and use for likelihood that does not rely on the assuming the truth of any of the models considered.

#### Bayesian forecasting system

For a BFS:

$$P_{i+1}^*(X_{i+1}) = \int P_{\theta}(X_{i+1} | \mathbf{x}^i) \pi_i(\theta) d\theta$$
$$= P_B(X_{i+1} | \mathbf{x}^i)$$

where  $P_B := \int P_{\theta} \pi(\theta) d\theta$  is the Bayes mixture joint distribution.

This is equivalent to basing all forecasts on the single distribution  $P_B$ . The total logarithmic score is thus

$$L_n(\mathcal{P}) = L_n(P_B)$$
  
=  $-\log p_B(\mathbf{x}^n)$   
=  $-\log \int p_\theta(\mathbf{x}^n) \pi(\theta) \, d\theta$ 

38 / 49

## Plug-in SFS

For a plug-in system:  $L_n = -\log \prod_{i=0}^{n-1} p_{\hat{\theta}_i}(x_{i+1} \mid \mathbf{x}^i).$ 

 $\Box$  The outcome  $(x_{i+1})$  used to evaluate performance, and the data  $(\mathbf{x}^i)$  used to estimate  $\theta$ , do not overlap

- "unbiased" assessments (like cross-validation)

- $\Box$  If  $x_i$  is used to forecast  $x_j$ , then  $x_j$  is *not* used to forecast  $x_i$ 
  - "uncorrelated" assessments (unlike cross-validation)

Both under- and over-fitting automatically and appropriately penalized.

## **Prequential efficiency**

#### Efficiency

Let P be a SFS. P is prequentially efficient for  $\{P_{\theta}\}$  if, for any PFS Q:

 $L_n(P) - L_n(Q)$  remains bounded above as  $n \to \infty$ , with  $P_{\theta}$  probability 1, for almost all  $\theta$ .

[In particular, the losses of any two efficient SFS's differ by an amount that remains asymptotically bounded under almost all  $P_{\theta}$ .]

 $\Box$  A BFS with  $\pi(\theta) > 0$  is prequentially efficient.

□ A plug-in SFS based on a Fisher efficient estimator sequence is prequentially efficient.

41 / 49

## Model testing

Model:

$$\mathbf{X} \sim P_{\theta} \quad (\theta \in \Theta)$$

Let P be prequentially efficient for  $\mathcal{P} = \{P_{\theta}\}$ , and define:

$$\begin{array}{lll} \mu_{i} & = & \mathsf{E}_{P}(X_{i} \mid \mathbf{X}^{i-1}) \\ \sigma_{i}^{2} & = & \mathsf{var}_{P}(X_{i} \mid \mathbf{X}^{i-1}) \\ Z_{n} & = & \frac{\sum_{i=1}^{n} (X_{i} - \mu_{i})}{\left(\sum_{i=1}^{n} \sigma_{i}^{2}\right)^{\frac{1}{2}}} \end{array}$$

Then  $Z_n \xrightarrow{L} \mathcal{N}(0,1)$  under any  $P_{\theta} \in \mathcal{P}$ .

So refer  $Z_n$  to standard normal tables to test the model  $\mathcal{P}$ .

#### Model choice

#### **Prequential consistency**

**Probability models** Collection  $C = \{P_j : j = 1, 2, ...\}.$ 

□ Both BFS and (suitable) plug-in SFS are prequentially consistent: with probability 1 under any  $P_i \in C$ , their forecasts will come to agree with those made by  $P_i$ .

**Parametric models** Collection  $C = \{P_j : j = 1, 2, ...\}$ , where each  $P_j$  is itself a parametric model:  $P_j = \{P_{j,\theta_j}\}$ . Can have different dimensionalities.

- $\Box$  Replace each  $\mathcal{P}_j$  by a prequentially efficient single distribution  $P_j$  and proceed as above.
- $\Box$  For each j, for almost all  $\theta_j$ , with probability 1 under  $P_{j,\theta_j}$  the resulting forecasts will come to agree with those made by  $P_{j,\theta_j}$ .

44 / 49

#### **Out-of-model performance**

Suppose we use a model  $\mathcal{P} = \{P_{\theta}\}$  for  $\mathbf{X}$ , but the data are generated from a distribution  $Q \notin \mathcal{P}$ . For an observed data-sequence  $\mathbf{x}$ , we have sequences of probability forecasts  $P_{\theta,i} := P_{\theta}(X_i \mid \mathbf{x}^{i-1})$ , based on each  $P_{\theta} \in \mathcal{P}$ : and "true" predictive distributions  $Q_i := Q(X_i \mid \mathbf{x}^{i-1})$ . The "best" value of  $\theta$ , for predicting  $\mathbf{x}^n$ , might be defined as:

$$\theta_n^Q := \arg\min_{\theta} \sum_{i=1}^n K(Q_i, P_{\theta,i})$$

NB: This typically depends on the observed data With  $\hat{\theta}_n$  the maximum likelihood estimate based on  $\mathbf{x}^n$ , we can show that for any Q, with Q-probability 1:

$$\hat{\theta}_n - \theta_n^Q \to 0.$$

#### Conclusions

Prequential analysis:

- $\hfill\square$  is a natural approach to assessing and adjusting the empirical performance of a sequential forecasting system
- $\hfill\square$  can allow for essentially arbitrary dependence across time
- □ has close connexions with Bayesian inference, stochastic complexity, penalized likelihood, etc.
- $\hfill\square$  has many desirable theoretical properties, including automatic selection of the simplest model closest to that generating the data
- $\hfill\square$  raises new computational challenges.

#### References

- Dawid, A. P. (1982). The well-calibrated Bayesian (with Discussion). J. Amer. Statist. Ass. 77, 604–613. Reprinted in Probability Concepts, Dialogue and Beliefs, edited by O. F. Hamouda and J. C. R. Rowley. Edward Elgar Publishing Ltd. (1997), 165–173.
- [2] Dawid, A. P. (1984). Present position and potential developments: some personal views. Statistical theory. The prequential approach (with Discussion). J. Roy. Statist. Soc. A 147, 278–292.
- [3] Dawid, A. P. (1985). The impossibility of inductive inference. (Invited discussion of 'Self-calibrating priors do not exist', by D. Oakes.) *J. Amer. Statist. Ass.* **80**, 340–341.
- [4] Dawid, A. P. (1985). Calibration-based empirical probability (with Discussion). Ann. Statist. 13, 1251–1285. Reprinted in Probability Concepts, Dialogue and Beliefs, edited by O. F. Hamouda and J. C. R. Rowley. Edward Elgar Publishing Ltd. (1997), 174–208.
- [5] Dawid, A. P. (1986). Probability Forecasting. *Encyclopedia of Statistical Sciences* vol. 7, edited by S. Kotz, N. L. Johnson and C. B. Read. Wiley-Interscience, 210–218.
- [6] Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference (with Discussion). J. Roy. Statist. Soc. B 53, 79–109.
- [7] Dawid, A. P. (1992) Prequential data analysis. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, edited by M. Ghosh and P. K. Pathak. IMS Lecture Notes-Monograph Series 17, 113-126.
- [8] Dawid, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference (with Discussion). Bayesian Statistics 4, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith. Oxford University Press, 109–125.
- [9] Seillier-Moiseiwitsch, F., Sweeting, T. J. and Dawid, A. P. (1992). Prequential tests of model fit. Scand. J. Statist. 19, 45–60.
- [10] Seillier-Moiseiwitsch, F. and Dawid, A. P. (1993). On testing the validity of sequential probability forecasts. J. Amer. Statist. Ass. 88, 355–359.
- [11] Cowell, R. G., Dawid, A. P. and Spiegelhalter, D. J. (1993). Sequential model criticism in probabilistic expert systems. *IEEE Trans. Pattern Recognition and Machine Intelligence* **15**, 209–219.
- [12] Dawid, A. P. (1993). Invited discussion of 'The logic of probability', by V. G. Vovk. J. Roy. Statist. Soc. B 55, 341–343.
- [13] Dawid, A. P. (1997). Prequential analysis. *Encyclopedia of Statistical Sciences*, Update Volume 1, edited by S. Kotz, C. B. Read and D. L. Banks. Wiley-Interscience, 464–470.
- [14] Skouras, K. and Dawid, A. P. (1998). On efficient point prediction systems. J. Roy. Statist. Soc. B 60, 765–780.
- [15] Dawid, A. P. and Vovk, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli* 5, 125–162.
- [16] Skouras, K. and Dawid, A. P. (1999). On efficient probability forecasting systems. Biometrika 86, 765-784.
- [17] Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999). Probabilistic Networks and Expert Systems (Chapters 10 and 11). Springer, xii + 321 pp.
- [18] Skouras, K. and Dawid, A. P. (2000). Consistency in misspecified models. Research Report 218, Department of Statistical Science, University College London. Available from: http://www.ucl.ac.uk/Stats/research/Resrprts/abs00.html#218