Fundamentals of Prequential Analysis

Philip Dawid

Statistical Laboratory
University of Cambridge

Forecasting

Context and purpose

Prequential = [Probabilistic]/Predictive/Sequential
— a general framework for assessing and comparing the predictive performance of a FORECASTING SYSTEM.

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:

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

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:

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

- There may be patterns in the sequence of values.

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:

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

- 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.

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:

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

- 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.

One-step Forecasts

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

One-step Forecasts

- Introduce the data-points $(x_1, \ldots, x_n)$ one by one.
- At time i, we have observed values $\mathbf{x}^i$ of $\mathbf{X}^i := (X_1, \ldots, X_i)$.

One-step Forecasts

- Introduce the data-points $(x_1, \ldots, x_n)$ one by one.
- At time i, we have observed values $\mathbf{x}^i$ of $\mathbf{X}^i := (X_1, \ldots, X_i)$.
- We now produce some sort of forecast, f_{i+1}, for X_{i+1}.

One-step Forecasts

- Introduce the data-points $(x_1, \ldots, x_n)$ one by one.
- At time i, we have observed values $\mathbf{x}^i$ of $\mathbf{X}^i := (X_1, \ldots, X_i)$.
- We now produce some sort of forecast, f_{i+1}, for X_{i+1}.
- Next, observe value x_{i+1} of X_{i+1}.

One-step Forecasts

- Introduce the data-points $(x_1, \ldots, x_n)$ one by one.
- At time i, we have observed values $\mathbf{x}^i$ of $\mathbf{X}^i := (X_1, \ldots, X_i)$.
- We now produce some sort of forecast, f_{i+1}, for X_{i+1}.
- Next, observe value x_{i+1} of X_{i+1}.
- Step up i by 1 and repeat.

One-step Forecasts

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

- Introduce the data-points $(x_1, \ldots, x_n)$ one by one.
- At time i, we have observed values $\mathbf{x}^i$ of $\mathbf{X}^i := (X_1, \ldots, X_i)$.
- We now produce some sort of forecast, f_{i+1}, for X_{i+1}.
- Next, observe value x_{i+1} of X_{i+1}.
- Step up i by 1 and repeat.
- 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.

Time development

t	1	2	3	$\cdots$
f				
x				

Time development

t	1	2	3	$\cdots$
f	f_1			
x				

t	1	2	3	$\cdots$
f	f_1			
x	x_1			

Time development

t	1	2	3	$\cdots$
f	f_1	f_2		
x	x_1			

Time development

t	1	2	3	$\cdots$
f	f_1	f_2		
x	x_1	x_2		

Time development

t	1	2	3	$\cdots$
f	f_1	f_2	f_3	
x	x_1	x_2		

Time development

t	1	2	3	$\cdots$
f	f_1	f_2	f_3	
x	x_1	x_2	x_3	

Time development

t	1	2	3	$\cdots$
f	f_1	f_2	f_3	$\cdots$
x	x_1	x_2	x_3	

Time development

t	1	2	3	$\cdots$
f	f_1	f_2	f_3	$\cdots$
x	x_1	x_2	x_3	$\cdots$

Some comments

Forecast type: Pretty arbitrary: *e.g.*

- ☐ Point forecast
- ☐ Action
- ☐ Probability distribution

Some comments

Forecast type: Pretty arbitrary: *e.g.*

- Point forecast
- Action
- Probability distribution

Black-box: Not interested in the truth/beauty/. . . of any theory underlying our forecasts—only in their performance

Forecast type: Pretty arbitrary: *e.g.*

- Point forecast
- Action
- Probability distribution

Black-box: Not interested in the truth/beauty/. . . of any theory underlying our forecasts—only in their performance

Close to the data: Concerned only with realized data and forecasts — not with their provenance, what might have happened in other circumstances, hypothetical repetitions,. . .

Forecast type: Pretty arbitrary: *e.g.*

- ☐ Point forecast
- ☐ Action
- ☐ Probability distribution

Black-box: Not interested in the truth/beauty/. . . of any theory underlying our forecasts—only in their performance

Close to the data: Concerned only with realized data and forecasts — not with their provenance, what might have happened in other circumstances, hypothetical repetitions,. . .

No peeping: Forecast of X_{i+1} made before its value is observed — unbiased assessment

Forecasting systems

Very general idea, *e.g.*:

Probability Forecasting Systems

Very general idea, *e.g.*:

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

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)

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)

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

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)

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

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

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)

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

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

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

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)

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

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

- forecast $f_{i+1} = P^*(X_{i+1} \mid \mathbf{X}^i = \mathbf{x}^i)$, where P^* is formed from $\mathcal{P}$ by somehow estimating/eliminating θ, 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

Statistical Forecasting Systems

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

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

Plug-in forecasting system Given the past data $\mathbf{x}^i$, construct some estimate $\hat{\theta}_i$ of θ (*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 θ with each new observation!

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

Plug-in forecasting system Given the past data $\mathbf{x}^i$, construct some estimate $\hat{\theta}_i$ of θ (*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 θ with each new observation!

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

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

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

Plug-in forecasting system Given the past data $\mathbf{x}^i$, construct some estimate $\hat{\theta}_i$ of θ (*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 θ with each new observation!

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

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

Other *e.g.* fiducial predictive distribution, . . .

Prequential consistency

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

Prequential consistency

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

MLEs:

$$\begin{aligned}
\hat{\mu}_n &= \overline{X}_n & \xrightarrow{L} \quad & \mathcal{N}(0, \rho\sigma^2) \\
\hat{\sigma}^2_n &= n^{-1}\textstyle\sum_{i=1}^n (X_i - \overline{X}_n)^2 & \xrightarrow{p} \quad & (1-\rho)\sigma^2 \\
\hat{\rho}_n &= 0 & &
\end{aligned}$$

Prequential consistency

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

MLEs:

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

— not classically consistent.

Prequential consistency

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

MLEs:

$$\begin{aligned} \hat{\mu}_n &= \overline{X}_n & \xrightarrow{L} \quad & \mathcal{N}(0, \rho\sigma^2) \\ \hat{\sigma}_n^2 &= n^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2 & \xrightarrow{p} \quad & (1-\rho)\sigma^2 \\ \hat{\rho}_n &= 0 & & \end{aligned}$$

— 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 $(1-\rho)\sigma^2 + \sigma^2/\{n\rho + (1-\rho)\}$.

Absolute assessment

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

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

- Binary variables (X_i)
- Realized values (x_i)
- Emitted probability forecasts (p_i)

Calibration

- Binary variables (X_i)
- Realized values (x_i)
- 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.*

Calibration

- Binary variables (X_i)
- Realized values (x_i)
- 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" π:

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

Example

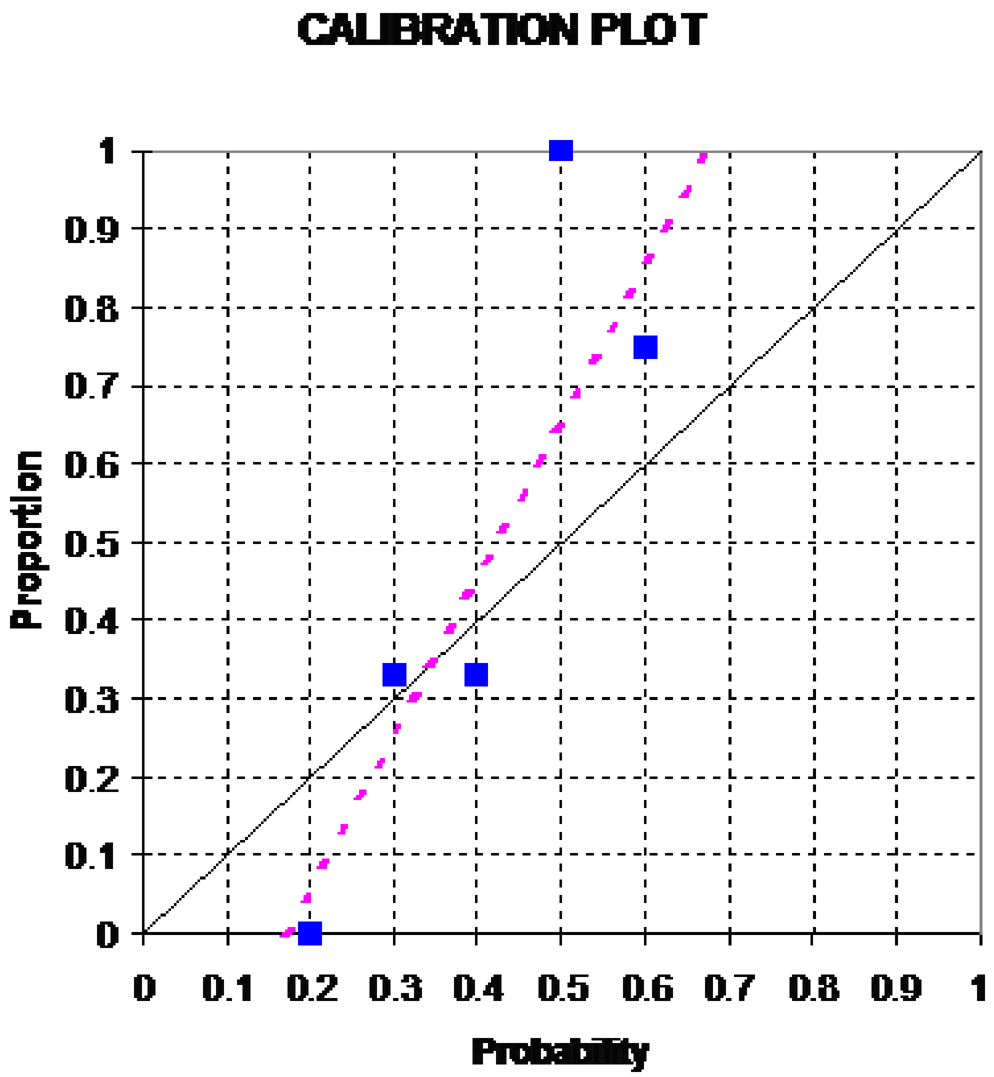
Probability	0.4	0.6	0.3	0.2	0.6	0.3	0.4	0.5	0.6	0.2	0.6	0.4	0.3	0.5
Outcome	0	0	1	0	1	0	1	1	1	0	1	0	0	1

Example

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
Instances n	2	3	3	2	4
Successes r	0	1	1	2	3
Proportion ρ	0	0.33	0.33	1	0.75

Calibration plot



Computable calibration

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

Computable calibration

Let σ 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.

Computable calibration

Let σ 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 σ.

Computable calibration

Let σ 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 σ.

Why?

Computable calibration

Let σ 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 σ.

Why?

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

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

P-almost surely, for any distribution P.

Well-calibrated forecasts are essentially unique

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

Well-calibrated forecasts are essentially unique

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

Then $p_i - q_i \to 0$.

Consider *e.g.*

$$Z_n := \frac{\sum (X_i - P_i)}{\sum P_i(1 - P_i)}$$

where $P_i = P(X_i = 1 \mid X^{i-1})$.

Consider *e.g.*

$$Z_n := \frac{\sum (X_i - P_i)}{\sum P_i(1 - P_i)}$$

where $P_i = P(X_i = 1 \mid X^{i-1})$.

Then

$$Z_n \xrightarrow{L} \mathcal{N}(0, 1)$$

for (almost) any P.

Significance test

Consider *e.g.*

$$Z_n := \frac{\sum (X_i - P_i)}{\sum P_i(1 - P_i)}$$

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

Significance test

Consider *e.g.*

$$Z_n := \frac{\sum (X_i - P_i)}{\sum P_i(1 - P_i)}$$

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"

Recursive residuals

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

Recursive residuals

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:

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

then, defining

$$U_i := F_i(X_i)$$

we have

$$U_i \sim U[0,1]$$

independently, for any P.

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

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.

Comparative assessment

Loss function

Measure inadequacy of forecast f of outcome x by

loss function: $L(x, f)$

Loss function

Measure inadequacy of forecast f of outcome x by

loss function: $L(x, f)$

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

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

We can use this to compare different forecasting systems.

Examples:

Squared error: f a point forecast of real-valued X

$$L(x, f) = (x - f)^2.$$

Squared error: f a point forecast of real-valued X

$$L(x, f) = (x - f)^2.$$

Logarithmic score: f a probability density $q(\cdot)$ for X

$$L(x, q) = -\log q(x).$$

Single distribution P

At time i, having observed $\mathbf{x}^i$, probability forecast for X_{i+1} is its conditional distribution $P_{i+1}(X_{i+1}) := P(X_{i+1} \mid \mathbf{X}^i = \mathbf{x}^i)$.

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

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

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

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

So the cumulative score is

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

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

Likelihood

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

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 $\mathbf{x}^n$.

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:

$$\begin{aligned} P^*_{i+1}(X_{i+1}) &= \int P_\theta(X_{i+1} \mid \mathbf{x}^i)\, \pi_i(\theta)\, d\theta \\ &= P_B(X_{i+1} \mid \mathbf{x}^i) \end{aligned}$$

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

Bayesian forecasting system

For a BFS:

$$\begin{aligned} P^*_{i+1}(X_{i+1}) &= \int P_\theta(X_{i+1} \mid \mathbf{x}^i)\, \pi_i(\theta)\, d\theta \\ &= P_B(X_{i+1} \mid \mathbf{x}^i) \end{aligned}$$

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

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

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

Plug-in SFS

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

- The data (x_{i+1}) used to evaluate performance, and the data $(\mathbf{x}^i)$ used to estimate θ, do not overlap
  - "unbiased" assessments (like cross-validation)

Plug-in SFS

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

- The data (x_{i+1}) used to evaluate performance, and the data $(\mathbf{x}^i)$ used to estimate θ, do not overlap
  - "unbiased" assessments (like cross-validation)
- If x_i is used to forecast x_j, then x_j is *not* used to forecast x_i
  - "uncorrelated" assessments (unlike cross-validation)

Plug-in SFS

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

- The data (x_{i+1}) used to evaluate performance, and the data ($\mathbf{x}^i$) used to estimate θ, do not overlap
  - "unbiased" assessments (like cross-validation)
- 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_θ probability 1, for almost all θ.

[In particular, the losses of any two efficient SFS's differ by an amount that remains asymptotically bounded under almost all P_θ.]

Efficiency

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

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

[In particular, the losses of any two efficient SFS's differ by an amount that remains asymptotically bounded under almost all P_θ.]

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

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_θ probability 1, for almost all θ.

[In particular, the losses of any two efficient SFS's differ by an amount that remains asymptotically bounded under almost all P_θ.]

- A BFS with $\pi(\theta) > 0$ is prequentially efficient.
- A plug-in SFS based on a Fisher efficient estimator sequence is prequentially efficient.

Model testing

Model:

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

Model testing

Model:

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

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

$$\begin{aligned} \mu_i &= \mathsf{E}_P(X_i \mid \mathbf{X}^{i-1}) \\ \sigma_i^2 &= \mathsf{var}_P(X_i \mid \mathbf{X}^{i-1}) \\ Z_n &= \frac{\sum_{i=1}^n (X_i - \mu_i)}{\left(\sum_{i=1}^n \sigma_i^2\right)^{\frac{1}{2}}} \end{aligned}$$

Model testing

Model:

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

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

$$\begin{aligned} \mu_i &= \mathsf{E}_P(X_i \mid \mathbf{X}^{i-1}) \\ \sigma_i^2 &= \mathsf{var}_P(X_i \mid \mathbf{X}^{i-1}) \\ Z_n &= \frac{\sum_{i=1}^n (X_i - \mu_i)}{\left(\sum_{i=1}^n \sigma_i^2\right)^{\frac{1}{2}}} \end{aligned}$$

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

Model:

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

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

$$\begin{aligned} \mu_i &= \mathsf{E}_P(X_i \mid \mathbf{X}^{i-1}) \\ \sigma_i^2 &= \mathsf{var}_P(X_i \mid \mathbf{X}^{i-1}) \\ Z_n &= \frac{\sum_{i=1}^n (X_i - \mu_i)}{\left(\sum_{i=1}^n \sigma_i^2\right)^{\frac{1}{2}}} \end{aligned}$$

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

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

Prequential consistency

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

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

Prequential consistency

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

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

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

Prequential consistency

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

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

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

- Replace each $\mathcal{P}_j$ by a prequentially efficient single distribution P_j and proceed as above.

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

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

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

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

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 θ, 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

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 θ, 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:

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

Happy Birthday George!