Part I
Statistical Prediction Theory

1

Statistical prediction

1.1 Filtering

Filtering is searching information provided by observed events on nonobserved events. These events are assumed to be associated with random experiments.

To describe such a problem, one may define a *probability space* $(\Omega, \mathcal{A}, P)$ and two sub-σ-algebras of $\mathcal{A}$, say $\mathcal{B}$ and $\mathcal{C}$, respectively observed and non-observed. '$\mathcal{B}$-observed' means that, for all $B \in \mathcal{B}$, the observer knows whether B occurs or not. Information provided by $\mathcal{B}$ on $\mathcal{C}$ may be quantified by P^B, the *conditional probability* with respect to $\mathcal{B}$. In general the statistician does not know P^B.

In the following, we are mainly interested in *prediction* (or *forecasting*). This signifies that $\mathcal{B}$ is associated with the past and $\mathcal{C}$ with the future. In practice one tries to predict a $\mathcal{C}$-measurable *random variable* Y from a $\mathcal{B}$-measurable random variable X.

If X is partly controllable, one can replace prediction by *foresight*, which consists in preparing for the future by constructing *scenarios* and then by selecting the most favourable option. It is then possible to make a *plan* for the future. We refer to Kerstin (2003) for a discussion about these notions.

Predictions can also sometimes modify the future. For example the publication of economic forecasts may change the behaviour of the economic agents involved and, therefore, influences future data. This phenomenon is discussed in Armstrong (2001), among others.

Inference and Prediction in Large Dimensions D. Bosq and D. Blanke

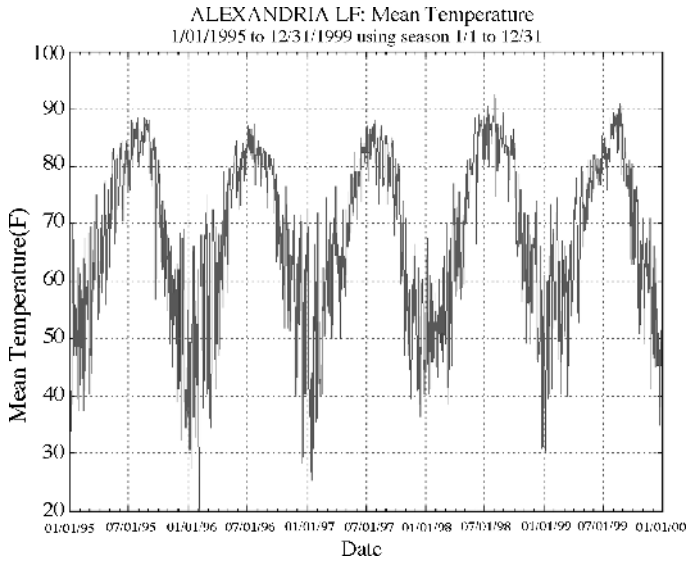

Figure 1.1 Example of a time series.

In this book we don't deal with this kind of problem, but we focus on *statistical prediction*, that is, prediction of future from a *time series* (i.e. a family of data indexed by time, see figure 1.1).[1]

1.2 Some examples

We now state some prediction problems. The following examples show that various situations can arise.

Example 1.1 *(The Blackwell's problem)*
The purpose is to predict occurrence or nonoccurrence of an event associated with an experiment when n repetitions of this experiment are available. Let $(x_n, n \geq 1)$ be the sequence of results ($x_n = 1$ if the nth event occurs, $= 0$ if not) and note by $p_n = p_n(x_1, \ldots, x_n)$ any predictor of x_{n+1}. The problem is to choose a 'good' p_n. Note that the sequence (x_n) may be deterministic or random. $\diamond$

Example 1.2 *(Forecasting a discrete time process)*
Let $(X_n,\ n \geq 1)$ be a real square integrable ⋆random process⋆. One observes $X = (X_1, \ldots, X_n)$ and intends to predict $Y = X_{n+h}$ $(h \geq 1)$; h is called the *forecasting horizon.*

[1]Provided by the NOAA-CIRES Climate Diagnostics Center: http://www.cdc.noaa.gov

This scheme corresponds to many practical situations: prediction of temperature, unemployment rate, foreign exchange rate, ... ◇

Example 1.3 *(Forecasting a continuous time process)*
Let $(X_t, t \in \mathbb{R}_+)$ be a real square integrable ⋆continuous time random process⋆. One wants to predict $Y = X_{t+h}$ $(h > 0)$ from the observed piece of sample path $X = (X_t, 0 \leq t \leq T)$ or $X = (X_{t_1}, \ldots, X_{t_n})$.

The difference with the above discrete time scheme is the possibility of considering small horizon forecasting (i.e. $h \to 0(+)$).

Classical applications are: prediction of electricity consumption, evolution of market prices during a day's trading, prediction of the ⋆counting process⋆ associated with a ⋆point process⋆, ... ◇

Example 1.4 *(Predicting curves)*
If the data are *curves* one may interpret them as realizations of random variables with values in a suitable function space. Example 1.3 can be rewritten in this new framework.

This kind of model appears to be useful when one wishes to predict the future evolution of a quantity during a full time interval. For example, electricity consumption for a whole day or variations of an electrocardiogram during one minute. ◇

Example 1.5 *(Prediction of functionals)*
In the previous examples one may predict functionals linked with the future of the observed process. For instance:

- The *conditional density* f_Y^X or the *conditional distribution* P_Y^X of Y given X.
- The *next crossing* at a given level, that is

$$Y_x = \min\{\tau : \tau > n, X_\tau = x\},$$

and, more generally, the *next visit* to a given Borel set,

$$Y_B = \min\{\tau : \tau > n, X_\tau \in B\}.$$

Notice that the forecasting horizon is not defined for these visit questions.

- Finally it is interesting to construct *prediction intervals* of the form $P(Y \in [a(X), b(X)])$. ◇

We now specify the prediction model, beginning with the real case.

1.3 The prediction model

Let $(\Omega, \mathcal{A}, P_\theta, \theta \in \Theta)$ be a *statistical model*, where $\mathcal{A}$ is the σ-algebra of events on Ω and (P_θ) a family of probability measures on $\mathcal{A}$ indexed by the unknown

parameter θ. $\mathcal{B} = \sigma(X)$ is the observed $^\star\sigma$-algebra* and $\mathcal{C}$ the nonobserved one. X takes its values in some *measurable space* $(E_0, \mathcal{B}_0)$.

A priori one intends to predict a $\mathcal{C}$-measurable real random variable Y. Actually, for reasons that are specified below, we extend the problem by considering prediction of $g(X, Y, \theta) \in \bigcap_{\theta\in\Theta} L^2(P_\theta)$, where g is known and Y nonobserved.

If g only depends on Y we will say that one deals with *pure prediction*, if it depends only on θ it is an *estimation problem*; finally, if g is only a function of X the topic is *approximation* (at least if g is difficult to compute!). The other cases are *mixed.* So, prediction theory appears as an extension of estimation theory.

Now a *statistical predictor* of $g(X, Y, \theta)$ is a known real measurable function of X, say $p(X)$. Note that the *conditional expectation* $\mathrm{E}^X_{\theta,g}$ is not in general a statistical predictor. In the following we assume that $p(X) \in \bigcap_{\theta\in\Theta} L^2(P_\theta)$, unless otherwise stated.

In order to evaluate the accuracy of p one may use the *quadratic prediction error* (QPE) defined as

$$R_\theta(p, g) = \mathrm{E}_\theta(p(X) - g(X, Y, \theta))^2, \theta \in \Theta.$$

This *risk function* induces the *preference relation*

$$p_1 \prec p_2 \Leftrightarrow R_\theta(p_1, g) \leq R_\theta(p_2, g), \theta \in \Theta. \tag{1.1}$$

If (1.1) is satisfied we will say that 'the predictor p_1 is preferable to the predictor p_2 for predicting g', and write $p_1 \prec p_2\ (g)$.

The QPE is popular and easy to handle, it has withstood the critics because it is difficult to find a good substitute. However, some other preference relations will be considered in Section 1.7 of this chapter.

Now, let $\mathcal{P}_G$ be the class of statistical predictors p such that $p(X) \in G = \bigcap_{\theta\in\Theta} G_\theta$ where G_θ is some closed linear space of $L^2(P_\theta)$, with *orthogonal projector* Π^θ, $\theta \in \Theta$. The following lemma shows that a prediction problem is, in general, mixed.

Lemma 1.1 *(Decomposition of the QPE)*
If $p \in \mathcal{P}_G$, its QPE has decomposition

$$\mathrm{E}_\theta(p - g)^2 = \mathrm{E}_\theta(p - \Pi^\theta g)^2 + \mathrm{E}_\theta(\Pi^\theta g - g)^2, \theta \in \Theta. \tag{1.2}$$

Hence, $p_1 \prec p_2$ for predicting g if and only if $p_1 \prec p_2$ for predicting $\Pi^\theta g$.

PROOF:
Decomposition (1.2) is a straightforward application of the Pythagoras theorem.

Therefore

$$\mathrm{E}_\theta(p_1 - g)^2 \leq \mathrm{E}_\theta(p_2 - g)^2 \Leftrightarrow \mathrm{E}_\theta(p_1 - \Pi^\theta g)^2 \leq \mathrm{E}_\theta(p_2 - \Pi^\theta g)^2, \theta \in \Theta.$$

■

Lemma 1.1 is simple but crucial: one must focus on the *statistical prediction error* $\mathrm{E}_\theta(p - \Pi^\theta g)^2$, since the *probabilistic prediction error* $\mathrm{E}_\theta(\Pi^\theta g - g)^2$ is not controllable by the statistician.

Thus, predicting $g(X, Y, \theta)$ or predicting $\Pi^\theta g(X, Y, \theta)$ is the same activity. In particular, to predict Y is equivalent to predicting $\Pi^\theta Y$; this shows that a nondegenerated prediction problem is mixed.

Example 1.6
Suppose that $X_1, \ldots, X_n$ are observed temperatures. One wants to know if X_{n+1} is going to exceed some threshold θ; then $g = \mathbb{1}_{X_{n+1} > \theta}$ and the problem is mixed. ◇

Let us now give some classical examples of spaces G.

Example 1.7
If $G = \bigcap_{\theta \in \Theta} L^2(\Omega, \mathcal{B}, P_\theta)$, then Π^θ is the *conditional expectation* given $\mathcal{B}$. ◇

Example 1.8
Suppose that $X = (X_1, \ldots, X_n)$ where $X_i \in L^2(P_\theta)$, $1 \leq i \leq n$ and $G = \mathrm{sp}\{1, X_1, \ldots, X_n\}$ does not depend on θ. Then $\Pi^\theta g$ is the *affine regression* of g on X. ◇

Example 1.9
If $X = (X_1, \ldots, X_n)$ with $X_i \in L^{2k}(P_\theta)$, $1 \leq i \leq n$, $(k \geq 1)$ and $G = \mathrm{sp}\{1, X_i, \ldots, X_i^k, 1 \leq i \leq n\}$ does not depend on θ. Then $\Pi^\theta g$ is a *polynomial regression*:

$$\Pi^\theta g(X_1, \ldots, X_n) = a_0 + \sum_{i=1}^{n} \sum_{j=1}^{k} b_{ji} X_i^j,$$

where a_0 and (b_{ji}) only depend on θ. ◇

1.4 P-sufficient statistics

As in estimation theory, sufficient statistics play a significant role in prediction. The definition is more restrictive.

Definition 1.1
A statistic $S(X)$ is said to be sufficient for predicting $g(X, Y, \theta)$ *(or* P-sufficient*) if*

(i) $S(X)$ is sufficient in the usual sense, i.e. the conditional distribution $P_X^{S(X)}$ of X with respect to $S(X)$ does not depend on θ.

(ii) For all θ, X and *$g(X, Y, \theta)$ are *conditionally independent* given $S(X)$.*

Condition (ii) means that

$$P_\theta^{S(X)}(X \in B, g \in C) = P_\theta^{S(X)}(X \in B) P_\theta^{S(X)}(g \in C),$$

$\theta \in \Theta, B \in \mathcal{B}_\mathbb{R}, C \in \mathcal{B}_\mathbb{R}$.

Note that, if $g(X, Y, \theta) = g(S(X), \theta)$, (ii) is automatically satisfied. Moreover one may show (see Ash and Gardner 1975, p. 188) that (ii) is equivalent to
(ii)$'$ $P_{\theta,g}^X = P_{\theta,g}^{S(X)}, \theta \in \Theta$.

We now give a statement which connects sufficiency and P-sufficiency.

Theorem 1.1
Suppose that (X, Y) has a strictly positive density $f_\theta(x, y)$ with respect to a $^\star\sigma$-finite* measure $\mu \otimes \nu$. Then

(a) If $S(X)$ is P-sufficient for predicting Y, $(S(X), Y)$ is sufficient in the statistical model associated with (X, Y).

(b) Conversely if $f_\theta(x, y) = h_\theta(S(x), y)c(x)d(y)$, then $S(X)$ is P-sufficient for predicting Y.

PROOF:

(a) Consider the decomposition

$$f_\theta(x, y) = f_\theta(x) f_\theta(y|x),$$

where $f_\theta(\cdot)$ is the density of X and $f_\theta(\cdot|\cdot)$ the conditional density of Y given $X = \cdot$. If $S(X)$ is P-sufficient, the factorization theorem (see Lehmann and Casella 1998, p. 35) yields $f_\theta = \varphi_\theta(S(x))\psi(x)$, where ψ does not depend on θ.

Now (ii)$'$ entails $f_\theta(y|x) = \gamma_\theta(y|S(x))$ where $\gamma_\theta(\cdot|\cdot)$ is the conditional density of Y given $S(X) = \cdot$.

Finally $f_\theta(x, y) = \varphi_\theta(S(x))\gamma_\theta(y|S(x))\psi(x)$ and the factorization theorem gives sufficiency of $(S(X), Y)$ in the model associated with (X, Y).

(b) Conversely, the relation

$$f_\theta(x, y) = h_\theta(S(x), y)c(x)d(y)$$

may be replaced by $f_\theta(x,y) = h_\theta(S(x),y)$ by substituting μ and ν for $c\cdot\mu$ and $d\cdot\nu$ respectively.

This implies

$$f_\theta(x) = \int h_\theta(S(x),y)\mathrm{d}\nu(y) := H_\theta(S(x)),$$

thus, $S(X)$ is sufficient, and since

$$f_\theta(y|x) = \frac{h_\theta(S(x),y)}{H_\theta(S(x))},$$

(ii)$'$ holds, hence $S(X)$ is P-sufficient. ■

We now give some examples and counterexamples concerning P-sufficiency.

Example 1.10
Let $X_1,\ldots,X_{n+1}$ be independent random variables with common density $\theta\mathrm{e}^{-\theta x}\mathbb{1}_{x>0}$ $(\theta>0)$. Set $X=(X_1,\ldots,X_n)$ and $Y=X_{n+1}-X_n$. Then $\sum_{i=1}^n X_i$ is sufficient for X but $(\sum_{i=1}^n X_i, Y)$ is not sufficient for (X,Y).

This shows that, in Theorem 1.1 (a), P-sufficiency of $S(X)$ cannot be replaced by sufficiency. ◇

Example 1.11
Let $X=(X_1,X_2)$ and Y be such that the density of (X,Y) is

$$f_\theta(x_1,x_2,y) = \theta^{3/2}\mathrm{e}^{-\theta(x_1+y)}\mathrm{e}^{-\frac{\pi^2}{4}x_2^2 y}\mathbb{1}_{x_1>0,\,x_2>0,y>0},\,(\theta>0).$$

Then (X_1,Y) is sufficient in the model (X,Y) but X_1 is not sufficient in the model (X_1,X_2).

This shows that, in Theorem 1.1 (b), the result is not valid if the special form $f_\theta(x,y)=h_\theta(S(x),y)c(x)d(y)$ does not hold. ◇

Example 1.12 *(discrete time Gaussian process)*
Let $(X_i, i\geq 1)$ be a *Gaussian process*, where $X_i \overset{d}{=} \mathcal{N}(\theta,1)$, $i\geq 1$ $(\theta\in\mathbb{R})$. Suppose that the covariance matrix C_n of $X=(X_1,\ldots,X_n)$ is known and invertible for each n, and set

$$C_{n+1} = \begin{bmatrix} C_n & \gamma_{n+1} \\ \gamma_{n+1} & 1 \end{bmatrix}.$$

Then, in order to predict $Y = X_{n+1}$, the statistic

$$(C_n^{-1}u_nX, C_n^{-1}\gamma_{n+1}X),$$

where $u_n = (1, \ldots, 1)'$, is P-sufficient. The statistic $C_n^{-1}u_nX$ is sufficient but not P-sufficient. $\diamond$

Example 1.13 *(Poisson process)*
Let $N = (N_t, t \geq 0)$ be a homogeneous ⋆Poisson process⋆ with intensity λ, and observed over $[0, T]$. Then N_T is sufficient. Now, since N is a ⋆Markov process⋆ the σ-algebras $\sigma(N_s, s \leq T)$ and $\sigma(N_s, s > T)$ are independent given N_T. Hence N_T is P-sufficient for predicting N_{T+h} $(h > 0)$. $\diamond$

Example 1.14 *(Ornstein–Uhlenbeck process)*
Consider an ⋆Ornstein–Uhlenbeck process⋆ (OU) defined as

$$X_t = \int_{-\infty}^{t} \mathrm{e}^{-\theta(t-s)}\mathrm{d}W(s), t \in \mathbb{R}, (\theta > 0),$$

where $W = (W_t, t \in \mathbb{R})$ is a ⋆standard bilateral Wiener process⋆ and the integral is taken in ⋆Ito⋆ sense. (X_t) is a zero-mean stationary Gaussian Markov process. Here the observed variable is $X_{(T)} = (X_t, 0 \leq t \leq T)$.

The likelihood of $X_{(T)}$ with respect to the distribution of $W_{(T)} = (W_t, 0 \leq t \leq T)$ in the space $\mathcal{C}([0, T])$ of continuous real functions defined on $[0, T]$ with uniform norm is

$$L(X_{(T)}, \theta) = \exp\left(-\frac{\theta}{2}(X_T^2 - X_0^2 - T) - \frac{\theta^2}{2}\int_0^T X_t^2\mathrm{d}t\right), \tag{1.3}$$

(See Liptser and Shiraev 2001). Then, the factorization theorem yields sufficiency of the statistics $(X_0^2, X_T^2, \int_0^T X_t^2\mathrm{d}t)$. Consequently $Z_T = (X_0, X_T, \int_0^T X_t^2\mathrm{d}t)$ is also sufficient.

Now, since (X_t) is Markovian, we have

$$\sigma(X_{(T)}) \perp\!\!\!\perp \sigma(X_{T+h})|\sigma(X_T),$$

then

$$P_{\theta, X_{T+h}}^{Z_T} = P_{\theta, X_{T+h}}^{X_T} = P_{\theta, X_{T+h}}^{X_{(T)}}$$

hence (ii)′ holds and Z_T is P-sufficient for predicting X_{T+h}. $\diamond$

The next statement shows that one may use P-sufficient statistics to improve a predictor. It is a Rao–Blackwell type theorem (See Lehmann 1991, p. 47).

Theorem 1.2 *(Rao–Blackwell theorem)*
Let $S(X)$ be a P-sufficient statistic for $g(X, Y, \theta)$ and $p(X)$ a statistical predictor of g. Then $\mathrm{E}^{S(X)}p(X)$ is preferable to $p(X)$ for predicting g.

PROOF:
$P_{\theta,X}^{S(X)}$ does not depend on θ, thus $q(X) = \mathrm{E}^{S(X)}p(X)$ is again a statistical predictor. Furthermore

$$\mathrm{E}_\theta(p-g)^2 = \mathrm{E}_\theta(p-q)^2 + \mathrm{E}_\theta(q-g)^2 + 2\mathrm{E}_\theta((p-q)(q-g)). \tag{1.4}$$

Now, by definition of conditional expectation $\mathrm{E}_\theta((p-q)q) = 0$ and, from condition (ii) in Definition 1.1,

$$\mathrm{E}_\theta[(p-q)g] = \mathrm{E}_\theta[\mathrm{E}_\theta^s(p-q)g] = \mathrm{E}_\theta^s(p-q)\mathrm{E}_\theta^s(g) = 0$$

since $\mathrm{E}_\theta^s(p) = q$; thus (1.4) gives $\mathrm{E}_\theta(p-g)^2 \geq \mathrm{E}_\theta(q-g)^2$. ■

Note that, if (ii) does not hold, the result remains valid provided $\mathrm{E}_\theta[(p-q)g] = 0$, $\theta \in \Theta$.

Finally it is noteworthy that, if $S(X)$ is P-sufficient to predict g it is also P-sufficient to predict $\mathrm{E}_\theta^{S(X)}(g)(= \mathrm{E}_\theta^X(g))$; actually this conditional expectation is a function of $S(X)$, hence $X \perp\!\!\!\perp \mathrm{E}_\theta^{S(X)}(g)|S(X)$.

1.5 Optimal predictors

A statistical predictor p_O is said to be *optimal* in the family $\mathcal{P}$ of predictors of g if

$$p_O \prec p, p \in \mathcal{P}$$

that is

$$\mathrm{E}_\theta(p_O - g)^2 \leq \mathrm{E}_\theta(p-g)^2; \quad \theta \in \Theta, p \in \mathcal{P}.$$

Notice that, in the family of all square integrable predictors, such a predictor does not exist as soon as $\mathrm{E}_\theta^X(g)$ is not constant in θ. Indeed, $p_1(X) = \mathrm{E}_{\theta_1}^X(g)$ is optimal at θ_1 when $p_2(X) = \mathrm{E}_{\theta_2}^X(g)$ is optimal at θ_2 which is impossible if $\mathrm{E}_{\theta_i}(\mathrm{E}_{\theta_1}^X(g) - \mathrm{E}_{\theta_2}^X(g))^2 \neq 0$; $i = 1, 2$.

Thus, in order ro construct an optimal predictor, it is necessary to restrict $\mathcal{P}$. For example one may consider only *unbiased* predictors.

Definition 1.2
A predictor $p(X)$ of $g(X, Y, \theta)$ is said to be unbiased if

$$\mathrm{E}_\theta(p(X)) = \mathrm{E}_\theta(g(X, Y, \theta)), \theta \in \Theta. \tag{1.5}$$

Condition (1.5) means that p is an unbiased estimator of the parameter $\mathrm{E}_\theta(g)$.

Example 1.15 *(Autoregressive process of order 1)*
Consider an autoregressive process of order 1 (AR(1)) defined as

$$X_n = \sum_{j=0}^{\infty} \theta^j \varepsilon_{n-j}, n \in \mathbb{Z} \tag{1.6}$$

where $(\varepsilon_n, n \in \mathbb{Z})$ is *strong white noise* (i.e. a sequence of i.i.d. random variables such that $0 < \mathrm{E}\varepsilon_n^2 = \sigma^2 < \infty$ and $\mathrm{E}\varepsilon_n = 0$) and $0 < |\theta| < 1$. The series converges in mean square and almost surely.

From (1.6) it follows that

$$X_n = \theta X_{n-1} + \varepsilon_n, n \in \mathbb{Z},$$

then

$$\mathrm{E}_{\theta,\sigma^2}^{\sigma(X_i, i \le n)}(X_{n+1}) = \theta X_n.$$

For convenience we suppose that $\sigma^2 = 1 - \theta^2$, hence $\mathrm{E}_{\theta,\sigma^2}(X_n^2) = 1$. Now a natural unbiased estimator of θ, based on $X_1, \ldots, X_n$ (n $>$ 1) is

$$\hat{\theta}_n = \frac{1}{n-1} \sum_{i=1}^{n-1} X_i X_{i+1},$$

hence a predictor of X_{n+1} is defined as

$$\hat{X}_{n+1} = \hat{\theta}_n X_n.$$

Now we have

$$\mathrm{E}_\theta(\hat{X}_{n+1}) = \theta^2 \frac{1 - \theta^{n-1}}{(1-\theta)(n-1)} \mathrm{E}_\theta(X_0^3),$$

thus $\hat{X}_{n+1}$ is unbiased if and only if (iff)

$$\mathrm{E}_\theta(X_0^3) = 0, 0 < |\theta| < 1. \qquad \diamond$$

We now give an extension of the classical Lehmann–Scheffé theorem (see Lehmann 1991, p. 88).

First recall that a statistic S is said to be *complete* if

$$\mathrm{E}_\theta(U) = 0, \theta \in \Theta \text{ and } U = \varphi(S) \Rightarrow U = 0, P_\theta \text{ a.s. for all } \theta.$$

Theorem 1.3 *(Lehmann–Scheffé theorem)*
If S is a complete P-sufficient statistic for g and p is an unbiased predictor of g, then $\mathrm{E}^S(p)$ is the unique optimal unbiased predictor of g (P_θ a.s. for all θ).

PROOF:
From Theorem 1.2 any optimal unbiased predictor of g is a function of S. Thus $\mathrm{E}^S(p)$ is a candidate since $\mathrm{E}_\theta[\mathrm{E}^S p] = \mathrm{E}_\theta(p) = g$. But completeness of S entails uniqueness of an unbiased predictor of g as a function of S, hence $\mathrm{E}^S(p)$ is optimal. ■

Example 1.13 *(continued)*
N_T is a complete P-sufficient statistic for N_{T+h}, the consequently unbiased predictor

$$p(N_T) = \frac{T+h}{T} N_T$$

is optimal for predicting N_{T+h}. Its quadratic error is

$$\mathrm{E}_\lambda (p(N_T) - N_{T+h})^2 = \lambda h \left(1 + \frac{h}{T}\right).$$

It is also optimal unbiased for predicting

$$\mathrm{E}_\lambda^{N_T}(N_{T+h}) = \lambda h + N_T$$

with quadratic error

$$\mathrm{E}_\lambda (p(N_T) - \mathrm{E}_\lambda^{N_T}(N_{T+h}))^2 = \frac{\lambda h^2}{T}. \qquad \diamond$$

The following statement gives a condition for optimality of an unbiased predictor.

Theorem 1.4
Set $\mathcal{U} = \{U(X) : \mathrm{E}_\theta U^2(X) < \infty, \mathrm{E}_\theta U(X) = 0; \theta \in \Theta\}$. *Then an unbiased predictor* p *of* g *is optimal iff*

$$\mathrm{E}_\theta[(p-g)U] = 0; U \in \mathcal{U}, \theta \in \Theta. \tag{1.7}$$

PROOF:
If p is optimal, we set

$$q = p + \alpha U, U \in \mathcal{U}, \alpha \in \mathbb{R},$$

then

$$\mathrm{E}_\theta (p + \alpha U - g)^2 \geq \mathrm{E}_\theta (p-g)^2,$$

therefore

$$\alpha^2 \mathrm{E}_\theta U^2 + 2\alpha \mathrm{E}_\theta ((p-g)U) \geq 0, U \in \mathcal{U}, \alpha \in \mathbb{R}$$

which is possible only if (1.7) holds.

Conversely, if p satisfies (1.7) and p' denotes another unbiased predictor, then $p' - p \in \mathcal{U}$, therefore

$$\mathrm{E}_\theta[(p-g)(p'-p)] = 0,$$

which implies

$$\begin{aligned}\mathrm{E}_\theta[(p'-g)^2 - (p-g)^2] &= \mathrm{E}_\theta(p'^2 - p^2) - 2\mathrm{E}_\theta[(p'-p)g] \\ &= \mathrm{E}_\theta(p'^2 - p^2) - 2\mathrm{E}_\theta[(p'-p)p] \\ &= \mathrm{E}_\theta(p'-p)^2 \geq 0,\end{aligned}$$

thus p is preferable to p'. ■

Note that such a predictor is unique. Actually, if p' is another optimal unbiased predictor, (1.7) yields

$$\mathrm{E}_\theta((p'-p)U) = 0, U \in \mathcal{U}, \theta \in \Theta$$

which shows that $p' - p$ is an optimal unbiased predictor of 0. But 0 is an optimal unbiased predictor of 0, with quadratic error 0, thus $\mathrm{E}_\theta(p'-p)^2 = 0$, $\theta \in \Theta$ hence $p' = p$, P_θ a.s. for all θ.

Now, since an unbiased predictor of g is an unbiased estimator of $\mathrm{E}_\theta g$, it is natural to ask whether the best unbiased estimator (BUE) of $\mathrm{E}_\theta g$ and the best unbiased predictor (BUP) of g coincide or not.

The next theorem gives an answer to this question.

Theorem 1.5
The BUE of $\mathrm{E}_\theta g$ *and the BUP of* g *coincide iff*

$$\mathrm{E}_\theta(gU) = 0, U \in \mathcal{U}, \theta \in \Theta. \tag{1.8}$$

PROOF:
First suppose that (1.8) holds. If p is the BUE of $\mathrm{E}_\theta g$, it is also the BUP of $\mathrm{E}_\theta g$, then Theorem 1.4 implies that, for all $U \in \mathcal{U}$ and all $\theta \in \Theta$, we have

$$\mathrm{E}_\theta((p - \mathrm{E}_\theta g)U) = 0, \tag{1.9}$$

therefore $\mathrm{E}_\theta(pU) = 0$, and from (1.8) it follows that

$$\mathrm{E}_\theta[(p-g)U] = 0$$

then (1.7) holds and p is the BUP of g.

Conversely a BUP of g satisfies (1.7), and, if it coincides with the BUE of $\mathrm{E}_\theta g$, (1.9) holds. Finally, (1.7) and (1.9) give (1.8). ■

Note that, if $\mathrm{E}_\theta g = 0, \theta \in \Theta$, the BUE of $\mathrm{E}_\theta g$ is $p = 0$. Thus 0 is the BUP of g if and only if (1.8) holds. For example, if g is zero-mean and independent of X, 0 is the BUP of g and its quadratic error is $\mathrm{E}_\theta g^2$.

Example 1.13 *(continued)*
If $X = N_T$ is observed we have $\mathcal{U} = \{0\}$, hence $p(N_T) = N_T(T+h)/T$ is the BUP of N_{T+h} and the BUE of $\mathrm{E}_\lambda(N_{T+h}) = \lambda(T+h)$. ◇

We now indicate a method that allows us to construct a BUP in some special cases.

Theorem 1.6
If g is such that

$$\mathrm{E}_\theta^X(g) = \phi(X) + \psi(\theta), \theta \in \Theta$$

where $\phi \in L^2(P_{\theta,x})$ and ψ are known, and if $s(X)$ is the BUE of $\psi(\theta)$; then

$$p(X) = \phi(X) + s(X)$$

is the BUP of g.

PROOF:
First note that $\mathrm{E}_\theta(p - \mathrm{E}_\theta g)^2 = \mathrm{E}_\theta(s - \psi(\theta))^2$. Now, let q be another unbiased predictor of g, then $q - \phi$ is an unbiased estimator of $\psi(\theta)$, hence

$$\mathrm{E}_\theta(q - \phi - \psi(\theta))^2 \geq \mathrm{E}_\theta(s - \psi(\theta))^2$$

that is

$$\mathrm{E}_\theta(q - \mathrm{E}_\theta^X(g))^2 \geq \mathrm{E}_\theta(p - \mathrm{E}_\theta^X(g))^2$$

thus, $p \prec q$, for predicting $\mathrm{E}_\theta^X(g)$ and, from Lemma 1.1, it follows that $p \prec q$ for predicting g. ■

Example 1.10 *(continued)*
Here we have $\mathrm{E}_\theta^X(Y) = \theta - X_n$ and, by the Lehmann–Scheffé theorem, $\bar{X}_n$ is the BUE of θ. Thus Theorem 1.6 shows that $p(X) = \bar{X}_n - X_n$ is the BUP of Y. ◇

Example 1.16 *(Semi-martingales)*
Let $(X_t, t \in \mathbb{R}_+)$ be a real square integrable process and $m(\theta, t)$ a deterministic function, such that

$$Y_t = X_t + m(\theta, t), t \in \mathbb{R}_+ (\theta \in \Theta \subset \mathbb{R})$$

is a *martingale* with respect to $\mathcal{F}_t = \sigma(X_s, 0 \leq s \leq t)$, $t \in \mathbb{R}_+$, i.e.

$$\mathrm{E}_\theta^{\mathcal{F}_s}(Y_t) = Y_s, 0 \leq s \leq t, \theta \in \Theta.$$

In order to predict X_{T+h} ($h > 0$) given the data $X_{(T)} = (X_t, 0 \leq t \leq T)$ we write

$$\begin{aligned}\mathrm{E}_\theta^{X_{(T)}}(X_{T+h}) &= \mathrm{E}_\theta^{\mathcal{F}_T}(Y_{T+h} - m(\theta, T+h)) \\ &= X_T + [m(\theta, T) - m(\theta, T+h)],\end{aligned}$$

and it is possible to apply Theorem 1.6 if $\psi(\theta) = m(\theta, T) - m(\theta, T+h)$ possesses a BUE.

In particular if (X_t) has **independent increments** then $(X_t - \mathrm{E}_\theta(X_t))$ becomes a martingale with $m(\theta, t) = -\mathrm{E}_\theta(X_t)$.

A typical example is again the Poisson process: (N_t) has independent increments, then $(N_t - \lambda t)$ is a martingale and $\mathrm{E}_\lambda^{N_T} = N_T + \lambda h$; applying Theorem 1.6 one again obtains that

$$N_T + \frac{N_T}{T} h = \frac{T+h}{T} N_T$$

is the BUP of N_{T+h}. ◇

The next lemma shows that existence of unbiased predictors does not imply existence of a BUP.

Lemma 1.2
Let $(X_t, t \in I)(I = \mathbb{Z}$ or $\mathbb{R})$ be a square integrable, zero-mean real Markov process with

$$\mathrm{E}_\theta^{\mathcal{F}_T}(X_{T+h}) = \varphi_T(\theta, X_T), \theta \in \Theta,$$

where $\mathcal{F}_T = \sigma(X_s, s \leq T)$.
Suppose that

(i) $\mathrm{E}_{\theta'}[\varphi_T(\theta, X_T)] = 0, \theta \in \Theta, \theta' \in \Theta$,

*(ii) there exist θ_1 and θ_2 in Θ such that P_{θ_1, X_T} and P_{θ_2, X_T} are *equivalent* and*

$$P_{\theta_1}[\varphi_T(\theta_1, X_T) \neq \varphi_T(\theta_2, X_T)] > 0.$$

Then, the class of unbiased predictors of X_{T+h}, given $X = (X_t, 0 \leq t \leq T)$ does not contain a BUP.

PROOF:
Consider the statistical predictor $p_1(X) = \varphi_T(\theta_1, X_T)$; from (i) it follows that it is unbiased, and a BUP p_O must satisfy

$$p_O(X) = p_1(X)\, P_{\theta_1} \quad \text{a.s.} \tag{1.10}$$

Similarly, if $p_2(X) = \varphi_T(\theta_2, X_T)$, we have

$$p_O(X) = p_2(X)\, P_{\theta_2} \quad \text{a.s.} \tag{1.11}$$

and (ii) shows that (1.10) and (1.11) are incompatible. ■

Example 1.14 *(continued)*
For the OU process, $X_T \overset{d}{=} \mathcal{N}(0, 1/2\theta)$ and $\mathrm{E}_\theta^{\mathcal{F}_T}(X_{T+h}) = \mathrm{e}^{-\theta h} X_T$, then (i) and (ii) hold and there is no BUP. ◇

Example 1.15 *(continued)*
If (ε_n) is Gaussian, $X_n \overset{d}{=} \mathcal{N}(0,1)$, and, since $\mathrm{E}_\theta^{\mathcal{F}_T}(X_{T+1}) = \theta X_T$, no BUP may exist. In particular $\hat{X}_{T+1}$ is not BUP. ◇

1.6 Efficient predictors

Under regularity conditions it is easy to obtain a Cramér–Rao type inequality for unbiased predictors. More precisely we consider the following assumptions:

Assumptions 1.1 (A1.1)
$\Theta \subset \mathbb{R}$ is an open set, the model associated with X is dominated by a σ-finite measure μ, the density $f(x,\theta)$ of X is such that $\{x : f(x,\theta) > 0\}$ does not depend on θ, $\partial f(x,\theta)/\partial\theta$ does exist. Finally the Fisher information

$$I_X(\theta) = \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\ln f(X,\theta)\right)^2$$

satisfies $0 < I_X(\theta) < \infty$, $\theta \in \Theta$.

Theorem 1.7 *(Cramér–Rao inequality)*
If A1.1 holds, p is an unbiased predictor, and the equality

$$\int p(x) f(x,\theta)\mathrm{d}\mu(x) = \mathrm{E}_\theta(g(X,Y,\theta))$$

can be differentiated under the integral sign, then

$$\mathrm{E}_\theta(p-g)^2 \geq \mathrm{E}_\theta(g - \mathrm{E}_\theta^X g)^2 + \frac{[\gamma'(\theta) - \mathrm{E}_\theta(\mathrm{E}_\theta^X(g)\frac{\partial}{\partial\theta}\ln f(X,\theta))]^2}{I_X(\theta)}, \tag{1.12}$$

where $\gamma(\theta) = \mathrm{E}_\theta g(X,Y,\theta)$.

PROOF:
Clearly it suffices to show that $\mathrm{E}_\theta(p - \mathrm{E}_\theta^X g)^2$ is greater than or equal to the second term in the right hand side of (1.12).

Now the ⋆Cauchy–Schwarz⋆ inequality yields

$$\mathrm{Cov}\left(p - \mathrm{E}_\theta^X g, \frac{\partial}{\partial\theta}\ln f\right) \leq [\mathrm{E}_\theta(p - \mathrm{E}_\theta^X g)^2]^{1/2}[I_X(\theta)]^{1/2}$$

moreover

$$\begin{aligned}\mathrm{Cov}\left(p - \mathrm{E}_\theta^X g, \frac{\partial}{\partial\theta}\ln f\right) &= \mathrm{E}_\theta\left(p\frac{\partial}{\partial\theta}\ln f\right) - \mathrm{E}_\theta\left(\mathrm{E}_\theta^X(g)\frac{\partial}{\partial\theta}\ln f\right)\\ &= \gamma'(\theta) - \mathrm{E}_\theta\left(\mathrm{E}_\theta^X(g)\frac{\partial}{\partial\theta}\ln f\right),\end{aligned}$$

hence (1.12). ■

The next statement gives an inequality very similar to the classical Cramér–Rao inequality.

Corollary 1.1
If, in addition, the equality

$$\gamma(\theta) = \int \mathrm{E}_\theta^{X=x}(g) f(x,\theta)\mathrm{d}\mu(x) \tag{1.13}$$

is differentiable under the integral sign, then

$$\mathrm{E}_\theta(p(X) - \mathrm{E}_\theta^X(g))^2 \geq \frac{\left[\mathrm{E}_\theta\left(\dfrac{\partial \mathrm{E}_\theta^X(g)}{\partial\theta}\right)\right]^2}{I_X(\theta)}, \theta \in \Theta. \tag{1.14}$$

PROOF:
Differentiating (1.13) one obtains

$$\gamma'(\theta) = \int \frac{\partial}{\partial\theta}(\mathrm{E}_\theta^{X=x}(g)) f(x,\theta)\mathrm{d}\mu(x) + \int \mathrm{E}_\theta^{X=x}(g)\frac{\partial \ln f(x,\theta)}{\partial\theta} f(x,\theta)\mathrm{d}\mu(x)$$

hence (1.14) from (1.12) where g is replaced by $\mathrm{E}_\theta^X g$. ■

Of course, (1.12) and (1.14) reduce to the classical Cramér–Rao inequality (see Lehmann 1991, p. 120) if g only depends on θ. So we will say that p is *efficient* if (1.12) is an equality. Note that p is efficient for predicting g if and only if it is efficient for predicting $\mathrm{E}_\theta^X(g)$.

We now give a result similar to Theorem 1.6.

Theorem 1.8
Under A1.1 and conditions in Theorem 1.6, if s(X) is an efficient unbiased estimator (EUE) of $\psi(\theta)$, then $p(X) = \phi(X) + s(X)$ is an efficient unbiased predictor (EUP) of g.

PROOF:
Efficiency of $s(X)$ means that ψ is differentiable and

$$\mathrm{E}_\theta(s(x) - \psi(\theta))^2 = \frac{[\psi'(\theta)]^2}{I_X(\theta)}, \tag{1.15}$$

now we have

$$\mathrm{E}_\theta(p(X)) = \mathrm{E}_\theta(\mathrm{E}_\theta^X g) = \gamma(\theta) \tag{1.16}$$

and

$$\mathrm{E}_\theta(p(X) - \mathrm{E}_\theta^X(g))^2 = \mathrm{E}_\theta(s(X) - \psi(\theta))^2.$$

Noting that

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathrm{E}_\theta(\phi(X)) &= \frac{\partial}{\partial\theta}\int \phi(x) f(x,\theta)\mathrm{d}\mu(x) \\ &= \mathrm{E}_\theta\left[\phi(X)\frac{\partial \ln f(X,\theta)}{\partial\theta}\right]\end{aligned}$$

and

$$\mathrm{E}_\theta\left(\psi(\theta)\frac{\partial \ln f(X,\theta)}{\partial\theta}\right) = 0$$

we obtain

$$\gamma'(\theta) - \mathrm{E}_\theta\left(\mathrm{E}_\theta^X(g)\frac{\partial \ln f(X,\theta)}{\partial\theta}\right) = \psi'(\theta)$$

then, the lower bound in (1.12) is $[\psi'(\theta)]^2/I_X(\theta)$ and efficiency of $p(X)$ follows from (1.15) and (1.16). ■

Example 1.10 *(continued)*
It is easy to verify that $\bar{X}_n - X_n$ is an efficient predictor of $X_{n+1} - X_n$. ◇

Example 1.17 *(Signal with Gaussian noise)*
Consider the process

$$X_t = \theta \int_0^t f(s)\mathrm{d}s + W_t, t \geq 0$$

where $\theta \in \mathbb{R}$, f is a locally square integrable function such that $\int_0^t f^2(s)\mathrm{d}s > 0$ for $t > 0$ and (W_t) is a standard Wiener process.

One can show (see Ibragimov and Hasminskii (1981)) that the maximum likelihood estimator of θ based on $X = (X_t, 0 \leq t \leq T)$ is

$$\hat{\theta}_T = \frac{\int_0^T f(t)\mathrm{d}X_t}{\int_0^T f^2(t)\mathrm{d}t} = \theta + \frac{\int_0^T f(t)\mathrm{d}W_t}{\int_0^T f^2(t)\mathrm{d}t}.$$

$\hat{\theta}_T$ is an EUE with variance $\left(\int_0^T f^2(t)\mathrm{d}t\right)^{-1}$.

Now, for $h > 0$, we have

$$\mathrm{E}_\theta^X(X_{T+h}) = X_T + \theta \int_T^{T+h} f(t)\mathrm{d}t := \phi(X) + \psi(\theta).$$

Applying Theorem 1.8 we obtain the EUP

$$\hat{X}_{T+h} = X_T + \frac{\int_0^T f(t)\mathrm{d}X_t}{\int_0^T f^2(t)\mathrm{d}t} \int_T^{T+h} f(t)\mathrm{d}t$$

with quadratic error

$$\mathrm{E}_\theta(\hat{X}_{T+h} - X_{T+h})^2 = \frac{\left(\int_T^{T+h} f(t)\mathrm{d}t\right)^2}{\int_0^T f^2(t)\mathrm{d}t} + h. \qquad \diamond$$

The next example shows that A1.1 does not imply existence of an EUP.

Example 1.18
If $X = (X_1, \ldots, X_n) \stackrel{d}{=} \mathcal{N}(\theta, 1)^{\otimes n}$ $(\theta \in \mathbb{R})$ and $g = \theta X_n$, the Cramér–Rao bound is θ^2/n. Then an EUP $p(X)$ must satisfy $\mathrm{E}_0(p(X) - 0)^2 = 0$ hence $p(X) = 0$ P_θ a.s. for all θ, which is contradictory. $\diamond$

We now study conditions for existence of an EUP, beginning with a necessary condition.

Theorem 1.9 *(Extended exponential model)*
If A1.1 holds, $\partial \ln f/\partial\theta$ and $\mathrm{E}_\theta((\partial \ln f/\partial\theta)(p - \mathrm{E}_\theta^X g))/I_X(\theta)$ are continuous in θ and if p is EUP for g with

$$\mathrm{E}_\theta(p(X) - \mathrm{E}_\theta^X g)^2 > 0, \theta \in \Theta$$

then

$$f(x,\theta) = \exp(A(\theta)p(x) - B(x,\theta) + C(x)) \tag{1.17}$$

where

$$\frac{\partial B(x,\theta)}{\partial\theta} = A'(\theta)\mathrm{E}_\theta^X(g). \tag{1.18}$$

PROOF:
If p is efficient for g, it is efficient for $\mathrm{E}_\theta^X(g)$ and the Cauchy–Schwarz inequality becomes an equality. Hence $\partial \ln f(X,\theta)/\partial\theta$ and $p(X) - \mathrm{E}_\theta^X(g)$ are collinear, and since they are not degenerate, this implies that

$$\frac{\partial \ln f(X,\theta)}{\partial\theta} = a(\theta)(p(X) - \mathrm{E}_\theta^X(g)) \tag{1.19}$$

where

$$a(\theta) = \frac{I_X(\theta)}{\mathrm{E}_\theta\left(\dfrac{\partial \ln f}{\partial\theta}(p - \mathrm{E}_\theta^X g)\right)}.$$

Using the continuity assumptions one may integrate (1.19) to obtain (1.17) and (1.18). ■

Note that decomposition (1.17) is not unique. Actually one may rewrite it under the form

$$f(x,\theta) = \exp(A(\theta)(p(x) + h(x)) - [B(x,\theta) + A(\theta)h(x)] + C(x))$$

but the prediction problem is then different: $p(X) + h(X)$ is EUP for $\mathrm{E}_\theta^X(g) + h(X)$.

The next statement gives a converse of Theorem 1.9 in a canonical case.

Theorem 1.10
Consider the model

$$f(x,\theta) = \exp(\theta p(x) - B(x,\theta) + C(x)), \theta \in \Theta.$$

If A1.1 holds and the equality $\int f(x,\theta)\mathrm{d}\mu(x) = 1$ *is twice differentiable under the integral sign, then* $p(X)$ *is an EUP of* $\partial B(X,\theta)/\partial\theta$.

PROOF:
By differentiating $\int f(x,\theta)\mathrm{d}\mu(x) = 1$ one obtains

$$\int [p(x) - \frac{\partial B}{\partial \theta}(x,\theta)] f(x,\theta)\mathrm{d}\mu(x) = 0,$$

therefore

$$\mathrm{E}_\theta p(X) = \mathrm{E}_\theta\left(\frac{\partial B}{\partial \theta}(X,\theta)\right).$$

Differentiating again leads to

$$\int \left[p(x) - \frac{\partial B}{\partial \theta}(x,\theta)\right]^2 f(x,\theta)\mathrm{d}\mu(x) = \int \frac{\partial^2 B}{\partial \theta^2}(x,\theta) f(x,\theta)\partial\mu(x),$$

that is

$$\mathrm{E}_\theta\left(p(X) - \frac{\partial B}{\partial \theta}(X,\theta)\right)^2 = \mathrm{E}\left(\frac{\partial^2 B(X,\theta)}{\partial \theta^2}\right),$$

it is then easy to verify that $\mathrm{E}_\theta(\partial^2 B(X,\theta)/\partial\theta^2)$ is the Cramér–Rao bound: p is EUP. ■

Example 1.13 *(continued)*
Here the likelihood takes the form

$$f_1(N_T,\lambda) = \exp(-\lambda T + N_T \ln\lambda + c(N_T)),$$

putting $\theta = \ln\lambda$ yields

$$f(N_T,\theta) = \exp(\theta N_T - T\mathrm{e}^\theta + c(N_T))$$

hence N_T is EUP for $\partial(T\mathrm{e}^\theta)/\partial\theta = \lambda T$, thus $(N_T/T)h$ is EUP for λh. From Theorem 1.8 it follows that $(N_T/T)h + N_T$ is EUP for $\lambda h + N_T = \mathrm{E}_\lambda^{N_T}(N_{T+h})$ and for N_{T+h}.◇

Example 1.14 *(continued)*
Taking into account the form of the likelihood we set

$$p(X) = \frac{T - X_T^2 + X_0^2}{2} \quad \text{and} \quad B(X,\theta) = \frac{\theta^2}{2}\int_0^T X_t^2 \mathrm{d}t,$$

then, Theorem 1.10 gives efficiency of $p(X)$ for predicting $\theta\int_0^T X_t^2\mathrm{d}t$.

Taking $\theta' = \theta^2$ as a new parameter, one obtains

$$f_1(X, \theta') = \exp\left(-\frac{\theta'}{2}\int_0^T X_t^2 \mathrm{d}t - \sqrt{\theta'}\,\frac{X_T^2 - X_0^2 - T}{2}\right),$$

hence

$$-\frac{1}{2}\int_0^T X_t^2 \,\mathrm{d}t$$

is efficient for predicting

$$\frac{1}{2\sqrt{\theta'}}\,\frac{X_T^2 - X_0^2 - T}{2} = \frac{1}{4\theta}(X_T^2 - X_0^2 - T).$$

This means that the empirical moment of order 2,

$$\frac{1}{T}\int_0^T X_t^2 \mathrm{d}t$$

is efficient for predicting

$$\frac{1}{2\theta}\left(1 - \frac{X_T^2}{T} + \frac{X_0^2}{T}\right).$$

It can be shown that it is not efficient for estimating $\mathrm{E}_\theta(X_0^2) = 1/2\theta$. In fact, its variance is

$$\frac{1}{2\theta^3 T} + \frac{1}{4\theta^2 T^2}(1 - \mathrm{e}^{-2\theta T})$$

when the Cramér–Rao bound is $1/(2\theta^3 T)$.

Note that the above predicted variables are not natural conditional expectations. Of course the genuine problem is to predict $\mathrm{E}_\theta^{\mathcal{F}_T}(X_{T+h}) = \mathrm{e}^{-\theta h} X_T$, but we have seen in Section 1.5 that there is no BUP in that case. We will consider this question from an asymptotic point of view in the next chapter. ◇

Example 1.19 *(Noncentered Ornstein–Uhlenbeck process)*
Consider the process

$$X_t = m + \int_{-\infty}^t \mathrm{e}^{-\theta(t-s)} \mathrm{d}W(s), t \in \mathbb{R}\ (\theta > 0, m \in \mathbb{R})$$

where W is a standard bilateral Wiener process.

Suppose that θ is known and m is unknown, and consider the process

$$X_{0,t} = \int_{-\infty}^{t} \mathrm{e}^{-\theta(t-s)} \mathrm{d}W(s), t \in \mathbb{R}.$$

On the space $C[0,T]$ the likelihood of $(X_t, 0 \leq t \leq T)$ with respect to $(X_{0,t}, 0 \leq t \leq T)$ has the form

$$f(X,m) = \exp\left[-\frac{\theta m^2}{2}(2+\theta T) + \theta m\left(X_0 + X_T + \theta \int_0^T X_s \mathrm{d}s\right)\right],$$

(see Grenander 1981), therefore $\theta\left(X_0 + X_T + \theta \int_0^T X_s \mathrm{d}s\right)$ is EUE for $\theta m(2+\theta T)$ and

$$m_T = \frac{1}{(2+\theta T)}\left[X_0 + X_T + \int_0^T X_s \mathrm{d}s\right]$$

is EUE for m.

Now $\mathrm{E}_\theta^{\mathcal{F}_T}(X_{T+h}) = \mathrm{e}^{-\theta h}(X_T - m) + m = m(1 - \mathrm{e}^{-\theta h}) + \mathrm{e}^{-\theta h} X_T n$, and, from Theorem 1.8 it follows that $m_T(1 - \mathrm{e}^{-\theta h}) + \mathrm{e}^{-\theta h} X_T$ is EUP for $\mathrm{E}_\theta^{\mathcal{F}_T}(X_{T+h})$ and for X_{T+h}.

Finally, since θ is known, the efficient predictor is

$$\hat{X}_{T+h} = \mathrm{e}^{-\theta h} X_T + \frac{1 - \mathrm{e}^{-\theta h}}{2 + \theta T}\left[X_0 + X_T + \int_0^T X_s \mathrm{d}s\right].$$

◇

Example 1.13 *(continued)*
Suppose that one wants to predict $\mathbb{1}_{\{N_{T+h}=0\}}$. Then $p = (-h/T)^{N_T}$ is the unique unbiased predictor function of N_T. It is optimal but not efficient. Moreover the naive predictor $\mathbb{1}_{\{N_T=0\}}$ is not unbiased but preferable to p.

Thus an optimal predictor is not always efficient and an unbiased optimal predictor is not always a good predictor. ◇

1.7 Loss functions and empirical predictors

The quadratic prediction error is not the single interesting risk function. Other preference relations are also used in practice.

Another matter is optimality; it is sometimes convenient to use predictors that are suboptimal but easy to compute and (or) robust.

In the current section we glance at various preference relations and some empirical predictors.

1.7.1 Loss function

A *loss function* $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a positive measurable function such that $L(a, a) = 0$, $a \in \mathbb{R}$. It generates a *risk function* defined as

$$R_\theta(g, p) = \mathrm{E}_\theta[L(g(X, Y, \theta), p(X))], \theta \in \Theta$$

which measures the accuracy of p when predicting g.

The resulting preference relation is

$$p_1 \prec p_2 \Leftrightarrow R_\theta(g, p_1) \leq R_\theta(g, p_2), \theta \in \Theta.$$

The following extension of the Rao–Blackwell theorem holds.

Theorem 1.11
Let $\prec$ be a preference relation defined by a loss function $L(x, y)$ which is convex with respect to y. Then, if S is P-sufficient for g and p is an integrable predictor, we have

$$\mathrm{E}_p^S \prec p. \tag{1.20}$$

PROOF:
We only give a sketch of the proof; for details we refer to Adke and Ramanathan (1997).

First we have the following preliminary result:

Let $(\Omega, \mathcal{A}, P)$ be a Probability space; $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ *three sub-σ-algebras of* $\mathcal{A}$, *then*

$$\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | \mathcal{F}_3 \Rightarrow \sigma(\mathcal{F}_1 \cup \mathcal{F}_3) \perp\!\!\!\perp \sigma(\mathcal{F}_2 \cup \mathcal{F}_3 | \mathcal{F}_3).$$

A consequence is:

if U_1, U_2 are real random variables such that $U_1 \in L^1(\Omega, \sigma(\mathcal{F}_1 \cup \mathcal{F}_3), P)$, $U_2 \in L^0(\sigma(\mathcal{F}_2 \cup \mathcal{F}_3))$, the space of real and $\sigma(\mathcal{F}_2 \cup \mathcal{F}_3)$-measurable applications, and $U_1 U_2 \in L^1(\Omega, \mathcal{A}, P)$, then

$$\mathrm{E}^{\mathcal{F}_3}(U_1) = 0 \Rightarrow \mathrm{E}(U_1 U_2) = 0. \tag{1.21}$$

Now, by using convexity of L, one obtains

$$L(g, p) \geq L(g, \mathrm{E}^S(p)) + L'(g, \mathrm{E}^S(p))(p - \mathrm{E}^S(p)) \tag{1.22}$$

where L' is the right derivative of $L(g, \cdot)$.

Choosing $\mathcal{F}_1 = \sigma(X)$, $\mathcal{F}_2 = \sigma(g)$, $\mathcal{F}_3 = \sigma(S)$ and applying (1.21) one obtains (1.20) by taking expectations in (1.22). ■

Note that an immediate consequence of Theorem 1.11 is an extension of the Lehmann–Scheffé theorem (Theorem 1.3). We let the reader verify that various optimality results given above remain valid if the quadratic error is replaced by a convex loss function.

1.7.2 Location parameters

Suppose that L is associated with a location parameter μ defined by

$$\mathrm{E}L(Z,\mu) = \min_{a\in\mathbb{R}} \mathrm{E}L(Z,a)$$

with $P_Z \in \mathcal{P}_L$, a family of probability measures on $\mathbb{R}$. Then, since

$$\mathrm{E}_\theta[L(g,p)] = \mathrm{E}_\theta[\mathrm{E}_\theta^X L(g,p)], \tag{1.23}$$

if $P_{\theta,g}^X \in \mathcal{P}_L$, the right side of (1.23) is minimum for $p_0(X) = \mu_\theta(X)$, where $\mu_\theta(X)$ is the location parameter associated with $P_{\theta,g}^X$. If $x \mapsto \mu_\theta(x)$ is measurable, one obtains

$$\mathrm{E}_\theta L(g,p_0(X)) = \min_{p\in L^0(\mathcal{B})} \mathrm{E}_\theta L(g,p(X)), \theta \in \Theta.$$

Three particular cases are classical:

- If $L(u,v) = (v-u)^2$, $\mu_\theta(X) = \mathrm{E}_\theta^X(g)$,
- if $L(u,v) = |v-u|$, $\mu_\theta(X)$ is a *median* of $P_{\theta,g}^X$,
- if $L(u,v) = \mathbb{1}_{|v-u|\geq\varepsilon}$ $(\varepsilon > 0)$, $\mu_\theta(X)$ is a *mode* of $P_{\theta,g}^X$. Note that this last loss function is not convex with respect to v.

In order to construct a statistical predictor based on these location parameters, one may use a *plug-in method*: this consists of replacing θ by an estimator $\hat{\theta}(X)$ to obtain the predictor

$$p(X) = \mu_{\hat{\theta}(X)}(X).$$

Such a predictor is not optimal in general but it may have sharp asymptotic properties if $\hat{\theta}$ is a suitable estimator of θ. We give some details in Chapter 2.

In a *nonparametric framework* the approach is somewhat different: it uses direct estimators of the conditional location parameter. For example the regression kernel estimator allows us to construct a predictor associated with conditional expectation (see

Chapter 6). For the conditional mode (respectively median) it may be estimated by taking the mode (respectively median) of a nonparametric conditional density estimator.

1.7.3 Bayesian predictors

In the Bayesian scheme one interprets θ as a random variable with a prior distribution τ.

Then if (X, Y) has density $f(x, y, \theta)$ with respect to $\mu \otimes \nu$ σ-finite, (X, Y, θ) has density $f(x, y, \theta)$ with respect to $\mu \otimes \nu \otimes \tau$. Thus, the marginal density of (X, Y) is given by

$$\varphi(x, y) = \int f(x, y, \theta) \mathrm{d}\tau(\theta).$$

Now, if L is a loss function with risk R_θ, the associated *Bayesian risk* for predicting Y is defined as

$$\begin{aligned} r(Y, p(X)) &= \int R_\theta(Y, p(X)) \mathrm{d}\tau(\theta) \\ &= \int L(y, p(x)) f(x, y, \theta) \mathrm{d}\mu(x) \mathrm{d}\nu(y) \mathrm{d}\tau(\theta) \\ &= \int L(y, p(x)) \varphi(x, y) \mathrm{d}\mu(x) \mathrm{d}\nu(y). \end{aligned}$$

As before a solution of $\min_p r(Y, p(X))$ is a location parameter associated with P_Y^X. If L is quadratic error, one obtains the *Bayesian predictor*

$$\tilde{Y} = \mathrm{E}^X(Y) = \int y \varphi(y|X) \mathrm{d}\nu(y) \tag{1.24}$$

where $\varphi(\cdot|X)$ is the marginal conditional density of Y given X.

Of course the recurrent problem is choice of the prior distribution. We refer to Lehmann and Casella (1998) for a comprehensive discussion.

Now the Bayesian approach turns out to be very useful if X does not provide enough information about Y. We give an example.

Example 1.20
Consider the model

$$Y = \theta X + \varepsilon, (|\theta| < 1)$$

where X and Y have distribution $\mathcal{N}(0, 1)$ and $\varepsilon \perp\!\!\!\perp X$. Then (1.24) gives

$$\tilde{Y} = \mathrm{E}(\theta) X = \int_{-1}^{1} \theta \mathrm{d}\tau(\theta) \cdot X,$$

while no reasonable nonbayesian predictor is available if θ is unknown. ◇

1.7.4 Linear predictors

Let $(X_t, t \geq 1)$ be a real square integrable stochastic process; assume that $X = (X_1, \ldots, X_n)$ is observed and $Y = X_{n+1}$ has to be predicted.

A commonly used empirical predictor is the linear predictor

$$p(X) = \sum_{i=1}^{n} a_i X_i. \tag{1.25}$$

If $\sum_{i=1}^{n} a_i = 1$ and $X_1, \ldots, X_{n+1}$ have the same expectation, $p(X)$ is unbiased. A typical predictor of this form is the *simple exponential smoothing* given by

$$p_\beta(X) = \frac{X_n + \beta X_{n-1} + \cdots + \beta^{n-1} X_1}{1 + \beta + \cdots \lambda + \beta^{n-1}},$$

with $0 \leq \beta \leq 1$.

If $\beta = 0$, it reduces to the *naive predictor*

$$p_0(X) = X_n,$$

if $\beta = 1$, one obtains the *empirical mean*

$$p_1(X) = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

if $0 < \beta < 1$ and n is large enough, practitioners use the classical form

$$\tilde{p}_\beta(X) = (1 - \beta)(X_n + \beta X_{n-1} + \cdots + \beta^{n-1} X_1).$$

The naive predictor is interesting if the X_t's are highly locally correlated. In particular, if (X_t) is a martingale, we have

$$X_n = \mathrm{E}_\theta^{\sigma(X_t, t \leq n)}(X_{n+1}), \theta \in \Theta$$

and X_n is the BUP with null statistical prediction error.

In contrast, the empirical mean is BUP when the X_t's are i.i.d. and $(\sum_{i=1}^{n} X_i)/n$ is the BUE of $\mathrm{E}_\theta X_1$ (cf. Theorem 1.5).

Finally, if $0 < \beta < 1$, one has

$$\tilde{p}_\beta(X) = \mathrm{E}_\beta^{\mathcal{F}_n}(X_{n+1})$$

provided (X_t) is an ARIMA (0,1,1) process defined by

$$\begin{cases} X_t - X_{t-1} = \varepsilon_t - \beta\varepsilon_{t-1}; \quad t \geq 2 \\ \qquad X_1 = \varepsilon_1. \end{cases}$$

and (ε_t) is a strong white noise. In fact this property remains valid for the wider range $-1 < \beta < 1$, see Chatfield (2000).

Concerning the choice of β, a general empirical method is *validation.* Consider the empirical prediction error

$$\delta_n(\beta) = \sum_{k=k_n}^{n} |X_k - p_{\beta,k-1}(X_1, \ldots, X_{k-1})|^2$$

where $p_{\beta,k-1}$ is the simple exponential smoothing constructed with the data $X_1, \ldots, X_{k-1}$ and $k_n < n$ is large enough. Then an estimator of β is

$$\hat{\beta} = \underset{0 \leq \beta \leq 1}{\arg\min}\, \delta_n(\beta).$$

If the model is known to be an ARIMA (0,1,1) one may also estimate β in this framework. If in addition, (X_t) is Gaussian, the maximum likelihood estimator (MLE) of β is asymptotically efficient as $n \to \infty$, see Brockwell and Davis (1991).

Finally, concerning the general predictor defined by (1.25) one may use *linear regression* techniques for estimating $(a_1, \ldots, a_n)$. Similarly as above, specific methods are available if (X_t) is an ARIMA (p,d,q) process, see Brockwell and Davis (1991).

1.8 Multidimensional prediction

We now consider the case where θ and (or) g take their values in a multidimensional space, or, more generally, in an infinite-dimensional space (recall that X takes its value in an arbitrary measurable space $(E_0, \mathcal{B}_0)$).

For example $X = (X_1, \ldots, X_n)$ where the X_i's are $\mathbb{R}^d$-valued with common density θ, and g is the conditional density $f^X_{\theta,Y}(\cdot)$.

A general enough framework is the case where $\theta \in \Theta \subset \Theta_0$ and g is B-valued where Θ_0 and B are separable Banach spaces with respective norms $\| \cdot \|_0$ and $\| \cdot \|$. Now, assume that $\mathrm{E}_\theta \| g \|^2 < \infty$, $\theta \in \Theta$ and denote by B^* the topological dual space of B. Then a natural preference relation between predictors is defined by

$$p_1 \prec p_2 \Leftrightarrow \mathrm{E}_\theta(x^*(p_1 - g))^2 \leq \mathrm{E}_\theta(x^*(p_2 - g))^2, \theta \in \Theta, x^* \in B^*.$$

This means that p_1 is preferable to p_2 for predicting g, if and only if $x^*(p_1)$ is preferable to $x^*(p_2)$ for predicting $x^*(g)$ for all x^* in B^*, with respect to the preference relation (1.1).

Now let $\mathcal{P}_{\mathcal{G}}(B)$ be the class of B-valued predictors such that $x^*(p) \in \mathcal{P}_{\mathcal{G}}$ for each $x^* \in B^*$, where $\mathcal{P}_{\mathcal{G}}$ is defined in Section 1.3. Then we have the following extension of Lemma 1.1:

Lemma 1.3
If $p \in \mathcal{P}_{\mathcal{G}}(B)$, then

$$\mathrm{E}_\theta(x^*(p) - x^*(g))^2 = \mathrm{E}_\theta(x^*(p) - \Pi^\theta x^*(g))^2 + \mathrm{E}_\theta(\Pi^\theta x^*(g) - x^*(g))^2,$$
$$\theta \in \Theta, x^* \in B^*. \tag{1.26}$$

In particular, if Π^θ is conditional expectation, it follows that $p_1 \prec p_2$ for predicting g if and only if $p_1 \prec p_2$ for predicting $\mathrm{E}_\theta^X(g)$.

PROOF:
It suffices to apply Lemma 1.1 to $x^*(p)$ and $x^*(g)$ and then to use the property

$$x^*(\mathrm{E}_\theta^X(g)) = \mathrm{E}_\theta^X(x^*(g))(\text{a.s.}).$$

■

Now, if $B = H$ (a Hilbert space) we have the additional property

$$\mathrm{E}_\theta \parallel p - g \parallel^2 = \mathrm{E}_\theta \parallel p - \mathrm{E}_\theta^X(g) \parallel^2 \; + \mathrm{E}_\theta \parallel \mathrm{E}_\theta^X(g) - g \parallel^2, \theta \in \Theta,$$

which is obtained by applying (1.26) to $x^* = e_j, j \geq 1$ where (e_j) is an orthonormal basis of H, and by summing the obtained equalities.

In this context one may use the simpler but less precise preference relation:

$$p_1 \prec_1 p_2 \Leftrightarrow \mathrm{E}_\theta \parallel p_1 - g \parallel^2 \leq \mathrm{E}_\theta \parallel p_2 - g \parallel^2, \theta \in \Theta.$$

Clearly

$$p_1 \prec p_2 \Rightarrow p_1 \prec_1 p_2.$$

Now, the results concerning sufficient statistics remain valid in the multidimensional case. In particular, application of Theorem 1.2 to $x^*(p)$ and $x^*(g)$ shows that, if $S(X)$ is P-sufficient, one has $\mathrm{E}^{\mathcal{S}}(p) < p$. A similar method allows us to extend the results concerning BUP. Details are left to the reader.

We now turn to efficiency. The next theorem gives a multidimensional Cramér–Rao inequality for predictors. We consider the following set of assumptions.

Assumptions 1.2 (A1.2)
Θ is open in Θ_0, X has a strictly positive density $f(x, \theta)$ with respect to a σ-finite measure, and

(i) $(\forall\theta \in \Theta)$, $(\exists U_\theta \in \Theta_0)$, $(\forall u \in U_\theta)$, $\exists V_{\theta,u}$ a neighbourhood of 0 in $\mathbb{R}$: $(\forall\delta \in V_{\theta,u})$, $\theta + \delta u \in \Theta$ and $\partial f(x, \theta + \delta u)/\partial\delta$ does exist. Then, one sets

$$\dot{f}_u(x,\theta) = \frac{\partial}{\partial\delta} f(x, \theta + \delta u)|_{\delta=0}.$$

(ii) The relation

$$\int f(x, \theta + \delta u)\mathrm{d}\mu(x) = 1, \quad \theta \in \Theta, u \in U_\theta, \delta \in V_{\theta,u}, \tag{1.27}$$

is differentiable with respect to δ under the integral sign.

(iii) The B-valued predictor p is such that

$$\int x^*(p(x))f(x, \theta + \delta u)\mathrm{d}\mu(x) = x^*(\gamma(\theta + \delta u)), \tag{1.28}$$

$x^* \in B^*$, $u \in U_\theta$, $\delta \in V_{\theta,u}$, where $\gamma : \Theta_0 \mapsto B$ is linear. Moreover this equality is differentiable with respect to δ under the integral sign.

(iv) $(\forall\theta \in \Theta)$, $(\forall u \in U_\theta)$,

$$I_\theta(X, u) = \mathrm{E}_\theta\left(\frac{\dot{f}_u(X,\theta)}{f(X,\theta)}\right)^2 \in]0, \infty[.$$

Then:

Theorem 1.12
If A1.2 holds, we have the bound

$$\mathrm{E}_\theta(x^*(p-g))^2 \geq \mathrm{E}_\theta(x^*(g - \mathrm{E}_\theta^X g))^2 + \frac{x^*\left[\gamma(u) - \mathrm{E}_\theta\left(\mathrm{E}_\theta^X(g)\frac{\dot{f}_u(X,\theta)}{f(X,\theta)}\right)\right]^2}{I_\theta(X,u)},$$
$$\theta \in \Theta, u \in U_\theta, x^* \in B^*. \tag{1.29}$$

PROOF:
Differentiating (1.27) and taking $\delta = 0$, one obtains

$$\mathrm{E}_\theta\left(\frac{\dot{f}_u}{f}\right) = \int \dot{f}_u(x,\theta)\mathrm{d}\mu(x) = 0;$$

the same operations applied to (1.28), and linearity of γ give

$$\begin{aligned}\mathrm{E}_\theta\left(x^*(p)\frac{\dot{f}_u}{f}\right) &= \int x^*(p(x))\frac{\dot{f}_u(x,\theta)}{f(x,\theta)} f(x,\theta)\mathrm{d}\mu(x) \\ &= x^*(\gamma(u)).\end{aligned}$$

Now the Cauchy–Schwarz inequality entails

$$\left[\mathrm{E}_\theta\left(x^*(p-\mathrm{E}_\theta^X g)\cdot\frac{\dot f_u}{f}\right)\right]^2 \le \mathrm{E}_\theta(x^*(p-\mathrm{E}_\theta^X g))^2\cdot\mathrm{E}_\theta\left(\frac{\dot f_u}{f}\right)^2,$$

collecting the above results and using Lemma 1.3 one arrives at (1.29). ■

In a Hilbert space it is possible to obtain a global result.

Corollary 1.2
If $B = H$, a Hilbert space, then

$$\mathrm{E}_\theta \parallel p-g\parallel^2 \ge \mathrm{E}_\theta\parallel p-\mathrm{E}_\theta^X(g)\parallel^2 + \frac{\parallel \gamma(u)-\mathrm{E}_\theta\left(g\frac{\dot f_u}{f}\right)\parallel^2}{I_\theta(X,u)}, \tag{1.30}$$

$\theta\in\Theta, u\in U_\theta$.

PROOF:
Apply (1.29) to $x^* = e_j$, $j\ge 1$ where (e_j) is a complete orthonormal system in H, and take the sum. ■

In (1.29) and (1.30), that are slight extensions of the Grenander inequality (1981, p. 484), the choice of u is arbitrary. Of course it is natural to choose a u that maximizes the lower bound. If the lower bound is achieved p is said to be efficient, but, in general, this only happens for some specific values of (x^*, u), as shown in the next example.

Example 1.21 *(Sequence of Poisson processes)*
Let $(N_{t,j}, t\ge 0)$, $j\ge 0$ be a sequence of independent homogeneous Poisson processes with respective intensity λ_j, $j\ge 1$ such that $\sum_j \lambda_j < \infty$.

Since $\mathrm{E}N_{t,j}^2 = \lambda_j t(1+\lambda_j t)$ it follows that $\sum_j N_{t,j}^2 < \infty$ a.s., therefore $M_t = (N_{t,j}, j\ge 0)$ defines an ℓ^2-valued random variable, where ℓ^2 is the Hilbert space $\{(x_j)\in\mathbb{R}^{\mathbb{N}}, \sum_j x_j^2 < \infty\}$ with norm $\parallel (x_j)\parallel = \left(\sum_j x_j^2\right)^{1/2}$.

One observes $M_{(T)} = (M_t, 0\le t\le T)$ and wants to predict M_{T+h} $(h>0)$. It is easy to see that M_T is a P-sufficient statistic, then one only considers predictors of the form $p(M_T)$.

Now let $\mathcal{N}\subset\ell^2$ the family of sequences (x_j) such that $(x_j)\in\mathbb{N}^{\mathbb{N}}$ and $x_j = 0$ for j large enough. This family is countable, hence the counting measure μ on $\mathcal{N}$, extended by $\mu(\ell^2-\mathcal{N}) = 0$, is σ-finite.

Then, note that $\sum_j N_{T,j}^2 < \infty$ a.s. yields $N_{T,j} = 0$ almost surely for j large enough. Thus M_T is $\mathcal{N}$-valued (a.s.). This implies that M_T has a density with respect to μ and the corresponding likelihood is

$$f(M_T(\omega), (\lambda_j)) = \sum_{j=0}^{J(T,\omega)} \mathrm{e}^{-\lambda_j T} \frac{(\lambda_j T)^{N_{T,j}(\omega)}}{N_{T,j}(\omega)!} \mathrm{e}^{-\sum_{j=0}^{J(T,\omega)} \lambda_j T}, \omega \in \Omega$$

where $J(T,\omega)$ is such that $N_{T,j}(\omega) = 0$ for $j > J(T,\omega)$.

Hence the MLE of $\theta = (\lambda_j)$ is

$$\hat{\theta}_T(\omega) = \left(\frac{N_{T,j}}{T}, 0 \le j \le J(T,\omega)\right) = \left(\frac{N_{T,j}(\omega)}{T}, j \ge 0\right).$$

Then, an unbiased predictor of M_{T+h} should be

$$\hat{M}_{T+h} = \left(\frac{T+h}{T} N_{T,j}, j \ge 0\right).$$

In order to study its efficiency we consider the loglikelihood:

$$\ln f(M_T, \theta + \delta u) = \sum_{j=0}^{\infty} [-(\lambda_j + \delta u_j)T + N_{T,j} \ln((\lambda_j + \delta u_j)T) + \ln(N_{T,j}!)]$$

since if $u = (u_j) \in \ell^2$, then $\sum (\lambda_j + \delta u_j)^2 < \infty$. Therefore

$$\frac{\dot{f}_u(M_T, \theta)}{f(M_T, \theta)} = \sum_{j=0}^{\infty} u_j \left(\frac{N_{T,j}}{\lambda_j} - T\right),$$

and

$$I_\theta(X, \mu) = T \sum_{j=0}^{\infty} \frac{u_j^2}{\lambda_j},$$

which belongs to $]0, \infty[$ if $\sum_j u_j^2 > 0$ and $\sum_j u_j^2/\lambda_j < \infty$.
Now, on one hand we have

$$\mathrm{E}_{\theta+\delta u}(x^*(\hat{M}_{T+h})) = (T+h) \sum_{j=0}^{\infty} (\lambda_j + \delta u_j) x_j,$$

thus, in (1.28), $\gamma : \ell^2 \mapsto \ell^2$ may be defined by

$$\gamma(v) = (T+h)v, v \in \ell^2,$$

on the other hand

$$\mathrm{E}_\theta^X(g) = \mathrm{E}_\theta^{M_T}(M_{T+h}) = (N_{T,j} + \lambda_j h, j \geq 0),$$

hence

$$\mathrm{E}_\theta\left(\frac{\dot{f}_u}{f}\mathrm{E}_\theta^X(g)\right) = T(u_j)$$

and

$$x^*\left[\gamma(u) - \mathrm{E}_\theta\left(\frac{\dot{f}_u}{f}\mathrm{E}_\theta^X g\right)\right] = h\sum_j x_j u_j.$$

Finally, since

$$\mathrm{E}_\theta[x^*(p - \mathrm{E}_\theta^X g)]^2 = \frac{h^2}{T}\sum_j \lambda_j x_j^2,$$

(1.29) shows that $\hat{M}_{T+h}$ is efficient if and only if

$$\sum_j \lambda_j x_j^2 = \frac{\left(\sum_j x_j u_j\right)^2}{\sum_j \frac{u_j^2}{\lambda_j}}. \tag{1.31}$$

In particular, if $x^* = (0, \ldots, 0, x_{j_0}, 0, \ldots)$ efficiency holds, provided $0 < \sum_j u_j^2/\lambda_j < \infty$.

More generally, set $x_j = \alpha_j/\sqrt{\lambda_j}$ and $u_j = \sqrt{\lambda_j}, \beta_j, j \geq 0$. If (α_j) and (β_j) are in ℓ^2, (1.31) gives

$$(\sum \alpha_j^2)(\sum \beta_j^2) = (\sum \alpha_j \beta_j)^2,$$

thus (α_j) and (β_j) are collinear, i.e. $(\lambda_j x_j)$ and (u_j) are collinear. ◇

Notes

As far as we know a systematic exposition of the theory of statistical prediction is not available in the literature. In this Chapter we have tried to give some elements of this topic.

Presentation of the prediction model is inspired by Yatracos (1992). Lemma 1.1 belongs to folklore but it is fundamental since it shows that the statistician may only predict $\Pi^\theta g$, rather than g.

Definition of P-sufficient statistics appear in Takeuchi and Akahira (1975). Also see Bahadur (1954); Johansson (1990) and Torgersen (1977) among others.

Theorem 1.1 is simple but useful and probably new, while Theorem 1.2 is in Johansson. The study of optimal unbiased predictors comes from Yatracos (1992) and Adke and Ramanathan (1997).

Theorem 1.7 is also in Yatracos but the more compact Corollary 1.1 and results concerning efficiency seem to be new.

Theorem 1.11 is taken from Adke and Ramanathan (1997) and other results of Section 1.7 are classical.

The elements of multidimensional prediction theory stated in Section 1.8 are natural extensions of the one-dimensional theory. The bound in Theorem 1.12 is an extension of the Cramér–Rao type bound of Grenander (1981). The application to sequences of Poisson processes is new.