Proper Measures of Discrepancy, Uncertainty and Dependence, with Applications to Predictive Experimental Design

A. P. Dawid
Department of Statistical Science
University College London

November 16, 1994

SUMMARY

We show how, associated with any decision problem, we may derive related measures of discrepancy, uncertainty and dependence for distributions. Such “proper” measures have special properties, which we characterise, and each measure essentially determines the others. The theory is applied to the Bayesian formulation of the problem of choosing an experiment in order to make a subsequent prediction. It is shown that coherent choice criteria may be based on any of the proper measures, with related measures yielding identical solutions.

Some key words: Proper scoring rule; Kullback-Leibler distance; entropy; cross-entropy; expected value of sample information.
1 Introduction

Let $X$ be an uncertain quantity with values in $X$. We may regard $X$ as an observable outcome of an experiment, or as the parameter of a statistical model. If $X$ is assigned a distribution $P$, any decision problem, with loss function $L$ depending on the outcome of $X$, may be solved.

In this paper we show how such a decision problem may be used to quantify a variety of concepts relating to the space $P$ of distributions for $X$: a proper scoring rule for assessing the empirical adequacy of $P$ in $P$; the uncertainty in a distribution $P$ in $P$; the discrepancy between two distributions $P$ and $Q$ in $P$; and a measure of the dependence of $X$ on another variable. These measures inherit various special properties, which we here characterise, from their decision-theoretic origins. Moreover, each such measure determines any of the others, up to equivalence.

In Section 2, we introduce our general formulation of a decision problem, governed by the outcome of $X$, and in Section 3 we describe the associated proper scoring rule for assessing a probability distribution for $X$. Section 4 derives from this a measure of discrepancy between two distributions for $X$, and Section 5 a measure of the uncertainty in such a distribution. In Section 6 we introduce a related measure of the dependence of $X$ on another quantity. Section 7 collects together the various relationships established between the different measures discussed. Finally, Section 8 shows how any of the above measures may be used as a coherent criterion for the selection of an experiment to perform in order to make predictive inference about $X$.

2 Decision problem

Consider a decision problem with action space $A$, and loss $L(x,a)$ suffered if action $a \in A$ is taken and the value of $X$ turns out to be $x \in X$. Let $P$ be the set of distributions for $X$. We typically suppose that, for $P \in P$, $L(P,a) = EP \{L(X,a)\}$ exists, and $\inf_{a \in A} L(P,a)$ is achieved at a unique value, $a(P)$ say. Slightly more generally, it is enough to require that, for any $P \in P$ there exist a unique $a(P) \in A$ such that $EP \{L(X,a) - L(X,a(P))\}$ exists in $[0, \infty]$ for all $a \in A$. We can also weaken the uniqueness requirement, merely requiring that, for each $P$, a specific act $a(P)$ with the above property be nominated. Then $a(P)$ is the Bayes act when uncertainty about $X$ is represented by the distribution $P$.

Note that if

$$L^*(x,a) = cL(x,a) + k(x)$$  \hspace{1cm} (2.1)

where $c > 0$ and the function $k$ is arbitrary, then the optimal act $a(P)$ is the same, whether $L^*$ or $L$ is the loss function. We shall call two loss functions
equivalent if they are related as in (2.1), and strongly equivalent if moreover $c = 1$.

If $P \neq Q \Rightarrow a(P) \neq a(Q)$, we call $L$ strict. Evidently, this property can only hold when the action space $A$ is at least as rich as $\mathcal{P}$.

**Example 1** Let $y : \mathcal{X} \rightarrow \mathbb{R}^1$, $A = \mathbb{R}^1$ and take $L(x, a) = (y - a)^2$, i.e. $(y(x) - a)^2$, corresponding to the problem of estimating $Y := y(X)$, with quadratic loss. Then $a(P) = E_P(Y)$ when this exists. Unless $\#\{\mathcal{X}\} = 2$, $L$ is not strict.

**Example 2** Let $A = \mathcal{P}$, and take $L(x, P) = -\log p(x)$, where $p$ is the density of $P$ with respect to some base measure $\mu$ on $\mathcal{X}$. This is the negative log-likelihood function, also called the logarithmic score. Changing $\mu$ results in a strongly equivalent loss function. For arbitrary $P, Q \in \mathcal{P}$, we can interpret $L(x, P) - L(x, Q)$ as $-\log \{p(x)/q(x)\}$, where the densities are taken with respect to $P + Q$, thus avoiding the need to specify a fixed $\mu$.

By the inequality

$$\int p(x) \log \{p(x)/q(x)\} d\mu \geq 0, \text{ all } P, Q \in \mathcal{P}, \tag{2.2}$$

where equality holds only if $Q = P$, we have $a(P) = P$, and $L$ is strict. □

In the sequel we shall proceed as if $L(P, a)$ exists for all $a \in A$, merely noting here that this condition can be weakened.

Starting from any decision problem, we now develop quantitative measures for a number of concepts of independent interest.

## 3 Scoring rules

Suppose You have to quote a distribution $Q$ for $X$. If $X = x$, You will suffer a loss $S(x, Q)$. This is a particular form of decision problem in which the action space $A$ is the same as $\mathcal{P}$. Writing $S(P, Q) = E_P \{S(X, Q)\}$, such a “scoring rule” encourages honesty if

$$S(P, P) \leq S(P, Q), \text{ all } P, Q \in \mathcal{P}. \tag{3.1}$$

A scoring rule satisfying (3.1) is termed proper; it is strictly proper if inequality holds in (3.1) whenever $Q \neq P$. Proper scoring rules are an important tool of Bayesian decision theory and inference: see Savage (1971), de Finetti (1975).

Now, given a loss function as in §2, define

$$S(x, Q) = L\{x, a(Q)\}. \tag{3.2}$$
Then $S(P, P) = L\{P, a(P)\} \leq L\{P, a(Q)\} = S(P, Q)$, by definition of $a(P)$. Hence (3.2) always defines a proper scoring rule (Dawid, 1986) which is strictly proper if $L$ is strict. In this way, any decision problem may be interpreted as involving the assessment of a probability distribution. Note that, if (2.1) holds, then $S^*(x, P) = cS(x, P) + k(x)$, in which case we call $S^*$ and $S$ equivalent (strongly equivalent if $c = 1$). Then $S^*(P, Q) = cS(P, Q) + k(P)$, with $k(P) := E_P\{k(X)\}$.

For Example 1, we have

$$S(x, Q) = \{y - E_Q(Y)\}^2,$$

(3.3)

$$S(P, Q) = \text{var}_P(Y) + \{E_P(Y) - E_Q(Y)\}^2.$$  

(3.4)

Unless $X$ is binary, this scoring rule is not strict since $S(P, Q) = S(P, P)$ whenever $E_Q(Y) = E_P(Y)$. We note that the logarithmic score of Example 2 is strictly proper. Whenever our original decision problem has $A = P$ and $L$ is strictly proper, (3.2) reproduces $S \equiv L$. For Example 2, we have

$$S(P, Q) = -\int p(x) \log q(x) d\mu.$$  

(3.5)

Now let $\{S_\lambda : \lambda \in \Lambda\}$ be a family of proper scoring rules, and let $W$ be a measure on $\Lambda$. Define

$$S(x, Q) = \int S_\lambda(x, Q) dW(\lambda).$$  

(3.6)

Then $S$ is also proper, and $S(P, Q) = \int S_\lambda(P, Q) dW(\lambda)$.

**Example 3** Let $y$ be a complex function on $X$, $Y = y(X)$, and define

$$S(x, Q) = |y - E_Q(Y)|^2.$$  

(3.7)

Separating real and imaginary parts, this is seen to be a combination of two cases of Example 1, hence proper. We obtain

$$S(P, Q) = \text{var}_P(Y) + |E_P(Y) - E_Q(Y)|^2$$  

(3.8)

where $\text{var}_P(Y)$ is interpreted as $E_P(\{|Y|^2\} - |E_P(Y)|^2$.

In particular, with $X = \mathbb{R}^m$, take, for $t \in \mathbb{R}^m$, $y(x) = e^{it \cdot x}$. We obtain

$$S(x, Q) = |e^{it \cdot x} - \phi_Q(t)|^2,$$  

(3.9)

where $\phi_Q$ is the characteristic function of $Q$, and

$$S(P, Q) = 1 - |\phi_P(t)|^2 + |\phi_P(t) - \phi_Q(t)|^2.$$  

(3.10)

Note that $S$ in (3.9) is invariant under the translations $x \rightarrow x + a$, $Q \rightarrow Q * \delta_a$, $a \in \mathbb{R}^m$, where $\delta_a$ is the point mass at $a$ and $*$ denotes convolution. Likewise, (3.10) is invariant under $P \rightarrow P * \delta_a$, $Q \rightarrow Q * \delta_a$.
**Example 4** With $\mathcal{X} = \mathbb{R}^m$, it now follows from (3.6) and Example 3 that a proper scoring rule is given by

$$S(x, Q) = \int e^{ix^T \tau} - \phi_Q(t) \, dW(t),$$

yielding

$$S(P, Q) = \int \left\{ 1 - |\phi_P(t)|^2 \right\} dW(t) + \int |\phi_P(t) - \phi_Q(t)|^2 dW(t).$$

If the support of $W$ is $\mathbb{R}^m$, this will be strictly proper. □

**Example 5** (Eaton, 1982; Eaton et al., 1995). Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ be a non-negative definite kernel: thus $K$ is bounded, $K(u, v) = \overline{K(v, u)}$, and

$$\|\alpha\|_K^2 := \int \int K(u, v) \, d\alpha(u) d\alpha(v) \geq 0$$

for all bounded signed measures $\alpha$ on $\mathcal{X}$. Define

$$S(x, Q) = \|Q\|_K^2 - \int K(x, v) dQ(v) - \int K(u, x) dQ(u),$$

strongly equivalent to $\|Q - \delta_x\|_K^2$. Then

$$S(P, Q) = \int \int K(u, v) \left\{ dQ(u) dQ(v) - dP(u) dQ(v) - dQ(u) dP(v) \right\},$$

so that

$$S(P, P) = -\|P\|_K^2$$

and so

$$S(P, Q) - S(P, P) = \|P - Q\|_K^2.$$

Thus (3.13) defines a proper scoring rule, which is strictly proper if $K$ is positive definite.

When $K(u, v) = y(u) \overline{y(v)}$, (3.13) is equivalent to (3.7). In particular, for $K(u, v) = e^{it^T (u-v)}$, (3.13) is equivalent to (3.9), and thus to (3.11) for $K(u, v) = \int e^{i t^T (u-v)} dW(t)$. □

**Example 6** Let $\mathcal{E}$ be a regular full exponential family, having canonical statistic $T = t(X)$ with values in $\mathbb{R}^k$, and parameterised by $\tau = E(T)$. The density with respect to some underlying measure $\nu$ on $\mathcal{X}$ is thus of the form

$$f(x; \tau) = \exp \left\{ a(x) - b(\theta) + \theta^T t(x) \right\},$$

where $b'(\theta) = \tau$ defines a $(1, 1)$ relationship between $\tau$ and the natural parameter $\theta$. Then define

$$S(x, Q) = -\log f(x; \tau_Q)$$

(3.18)
where $\tau_Q = E_Q(T)$, so long as this yields an allowable value of $\tau$. Clearly, (3.18) coincides with the logarithmic score for $Q \in \mathcal{E}$, but not outside $\mathcal{E}$.

With $b'(\theta_Q) = \tau_Q$ we have

$$S(x, Q) = -a(x) + b(\theta_Q) - \theta_Q^T t(x),$$
$$S(P, Q) = -E_P \{ a(X) \} + b(\theta_Q) - \theta_Q^T \tau_P,$$

equivalent to $b(\theta_Q) - \theta_Q^T \tau_P$, so that

$$S(P, Q) - S(P, P) = \{ b(\theta_Q) - b(\theta_P) \} - (\theta_Q - \theta_P)^T \tau_P,$$ (3.19)

which, since $\tau_P = b'(\theta_P)$, is non-negative by convexity of $b$. Hence $S$ is proper. However, $S$ is not strictly proper, since $S(P, Q) = 0$ whenever $E_Q(T) = E_P(T)$.

When $T = Y$, $\mathcal{E} = \{ N(\tau, 1) \}$, we recover Example 1 (up to equivalence).

\[ \Box \]

**Example 7** As another instance of Example 6, let $\mathcal{E}$ be the family of $p$-variate normal distributions for $X$, with arbitrary mean vector $\mu$ and dispersion matrix $\Sigma$. We can take $T = (X, XX^T)$, $\tau = (\mu, \mu^T + \Sigma)$. Let $\mu_Q, \Sigma_Q$ be the mean and dispersion of $X$ under $Q \in \mathcal{P}$. We obtain

$$S(x, Q) = \frac{1}{2} \left\{ \log \det \Sigma_Q + (x - \mu_Q)^T \Sigma_Q^{-1} (x - \mu_Q) \right\},$$ (3.20)
$$S(P, Q) = \frac{1}{2} \left\{ \log \det \Sigma_Q + (\mu_P - \mu_Q)^T \Sigma_Q^{-1} (\mu_P - \mu_Q) + \text{tr}(\Sigma_Q^{-1} \Sigma_P) \right\}.$$ (3.21)

\[ \Box \]

### 4 Discrepancy measures

Let $P, Q \in \mathcal{P}$. We can attempt to measure the discrepancy between these distributions by means of some function $D(P, Q)$, not necessarily symmetric, such that $D(P, Q) \geq 0$, $D(P, P) = 0$. Such a discrepancy measure $D$ is termed strict if $D(P, Q) > 0$ for $P \neq Q$.

From (3.1), we see that, given any proper scoring rule $S$, we may define a discrepancy measure by

$$D(P, Q) = S(P, Q) - S(P, P),$$ (4.1)

and $D$ will be strict if $S$ is. In particular, this construction may be based on a loss function $L$, via (3.2), yielding

$$D(P, Q) = L\{ P, a(Q) \} - L\{ P, a(P) \}.$$ (4.2)
If (2.1) holds, we obtain $D^*(P, Q) = cD(P, Q)$, and we call $D$ and $D^*$ equivalent. In particular, all strongly equivalent decision problems yield the identical discrepancy measure.

Example 1 yields

$$D(P, Q) = \{E_P(Y) - E_Q(Y)\}^2,$$

(4.3)
typically not strict, while Example 2 gives

$$D(P, Q) = \int p(x) \log \{p(x)/q(x)\} d\mu,$$

(4.4)
the Kullback-Leibler distance between $P$ and $Q$. From (3.10), we obtain

$$D(P, Q) = |\phi_P(t) - \phi_Q(t)|^2.$$ 

(4.5)
Note that, if (3.6) holds, we have, in obvious notation,

$$D(P, Q) = \int D_\lambda(P, Q)dW(\lambda).$$

(4.6)
Thus for Example 4 we have

$$D(P, Q) = \int |\phi_P(t) - \phi_Q(t)|^2 dW(t),$$

(4.7)
which will be strict if $W$ has full support. Again, (4.7) is translation invariant. More generally, Example 5 yields the discrepancy measure

$$D(P, Q) = \|P - Q\|_K^2,$$

(4.8)
which is strict if $K$ is positive definite. This discrepancy function was introduced by Eaton et al. (1995).

Note that (4.8) is symmetric in $P$ and $Q$. An extension of the argument of Savage (1971, §5.2) suggests that only such quadratic discrepancies can be symmetric.

The discrepancy measure $D(P, Q)$ engendered by (3.18) is given by (3.19). In the case of Example 7, we obtain

$$D(P, Q) = \frac{1}{2} \left\{ (\mu_P - \mu_Q)^T \Sigma_Q^{-1} (\mu_P - \mu_Q) + \text{tr} \left( \Sigma_Q^{-1} \Sigma_P \right) \right. \
- \left. \log \det \left( \Sigma_Q^{-1} \Sigma_P \right) - p \right\},$$

(4.9)
which is not strict, vanishing if $\mu_Q = \mu_P$ and $\Sigma_Q = \Sigma_P$.

Discrepancy measures constructed as in (4.1) have some special properties. We note that, for $Q_0, Q_1, P \in \mathcal{P},$

$$D(P, Q_1) - D(P, Q_0) = \int \{S(x, Q_1) - S(x, Q_0)\} dP(x).$$
Thus we see that, for any $Q_0, Q_1 \in \mathcal{P}$,
\[ D(P, Q_1) - D(P, Q_0) \text{ is affine in } P. \tag{4.10} \]
That is to say, if $K(P) = D(P, Q_1) - D(P, Q_0)$, then when $\tilde{P}$ is a random distribution in $\mathcal{P}$, $E \{ K(\tilde{P}) \} = K \{ E(\tilde{P}) \}$. A discrepancy measure satisfying (4.10) will be termed proper.

Any proper discrepancy measure $D$ has a representation (4.1) in terms of a proper scoring rule. For define
\[ S(x, Q) = D(\delta_x, Q), \tag{4.11} \]
with $\delta_x$ the unit mass at $x \in \mathcal{X}$. Then $S(P, Q_1) - S(P, Q_0) = E \{ K(\delta_X) \}$, where $X \sim P$, which equals $K(P)$ by propriety. Thus, taking $Q_1 = Q, Q_0 = P, D(P, Q) = S(P, Q) - S(P, P)$, in which, since $D(P, Q) \geq 0$, $S$ is a proper scoring rule, strictly proper if $D$ is strict. If, moreover, $S^*$ also satisfies (4.1) then $S(x, Q) = D(\delta_x, Q) = S^*(x, Q) - S^*(x, \delta_x)$, so that $S^*$ is strongly equivalent to $S$.

Note that the above procedure does not work, without ad hoc modification, for the Kullback-Leibler discrepancy (4.4), for which $D(\delta_x, Q)$ is typically infinite. Such unbounded measures should be excluded if we wish to retain a simple theory; however, their usefulness and other desirable properties argue for their retention, with due care and cautions.

Aitchison (1990) introduced another property, equivalent to the following.

**Definition 4.1** The discrepancy measure $D$ is called coherent if, for $\tilde{P}$ a random distribution over $\mathcal{P}$, $E \{ D(\tilde{P}, Q) \}$ is minimised in $Q$ when $Q = E(\tilde{P})$.

**Theorem 4.1** If $D$ is proper, then $D$ is coherent.

**Proof** Fix $Q_0 \in \mathcal{P}$, and define $U(P, Q) = D(P, Q) - D(P, Q_0)$. By propriety of $D$, $E \{ U(\tilde{P}, Q) \} = U \{ E(\tilde{P}), Q \}$. Thus $E \{ D(\tilde{P}, Q) \} = D \{ E(\tilde{P}), Q \} + \left[ E \{ D(\tilde{P}, Q_0) \} - D \{ E(\tilde{P}), Q_0 \} \right]$, and this is minimised for $Q = E(\tilde{P})$. \[ \square \]

Since there is a wide variety of proper discrepancy functions, Theorem 4.1 refutes Aitchison’s conjecture that only the Kullback-Leibler distance is coherent. Eaton et al. (1995) have noted the coherence property of the proper discrepancy (4.7).

Under regularity conditions, the converse to Theorem 4.1 also holds. First we note the following.

**Lemma 4.1** Let $\mathcal{P}$ be the family of probability distributions over $\mathcal{X}$, and let $Q \in \mathcal{P}$. Let $h : \mathcal{P} \rightarrow \mathcal{V}, \lambda \in \mathcal{V}$, where $\mathcal{V}$ is an affine space, be such that, whenever $\tilde{P}$ is a random distribution in $\mathcal{P}$ with $E(\tilde{P}) = Q$, we have
$E \{ h(\hat{P}) \} = \lambda$. Then, for any $P$ for which $P \ll Q$ and $dP/dQ$ is bounded a.s. \([P]\), $h(P) = E_P \{ u(X) \}$, where $u(x) = h(\delta_x)$, $\delta_x$ being the unit mass at $x$.

**Proof** Take $\beta \in (0,1)$ such that $dP/dQ \leq \beta^{-1}$ (a.s. \([P]\)). Then $R = (Q - \beta P)/(1 - \beta) \in \mathcal{P}$. Consider the following distribution over $\mathcal{P}$: with probability $\beta$, $\hat{P} = P$; otherwise, with probability $1 - \beta$, $\hat{P} = \delta_X$, where $X \sim R$. Then $E(\hat{P}) = \beta P + (1 - \beta)R = Q$, and so $\lambda = E \{ h(\hat{P}) \} = \beta h(P) + (1 - \beta)E_R \{ u(X) \}$. But $(1 - \beta)E_R \{ u(X) \} = E_Q \{ u(X) \} - \beta E_P \{ u(X) \}$, and $E_Q \{ u(X) \} = \lambda$ since $E_Q(\delta_X) = Q$. Thus $h(P) = E_P \{ u(X) \}$.

The above states, essentially, that the only functions of $\hat{P}$ whose expectations depend only on $E(\hat{P})$ are affine. It can be considered as a generalisation of Theorem 2.1 of Genest and Schervish (1985). It seems plausible that the bounded density condition could be removed, and we shall proceed on this supposition.

We continue our investigation into the propriety of a coherent discrepancy measure under the simplifying assumption that $\mathcal{X}$ is finite, with say $\#(\mathcal{X}) = N$. Then we can, and shall, take $\mathcal{P} = \mathcal{S}$, the unit simplex in $\mathbb{R}^N$. Let $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ be a coherent discrepancy measure. We use $\nabla$ to denote differentiation with respect to $Q$ parallel to $S$, so that, for differentiable $f : \mathcal{S} \to \mathbb{R}$, $\nabla f(Q) \equiv 0$ and $f(Q + \varepsilon) - f(Q) = \{\nabla f(Q)\}^T \varepsilon + o(\|\varepsilon\|^2)$, for $\varepsilon \in \mathbb{R}^N$, $Q + \varepsilon \in \mathcal{S}$; in particular we must have $1^T \varepsilon = 0$. We assume that, for $P$ fixed, and $Q \in \mathcal{S}^0$ the interior of $\mathcal{S}$, $\nabla D(P, Q)$ exists.

If now $\hat{P}$ is random over $\mathcal{S}$, then, assuming we can interchange expectation over $\hat{P}$ and differentiation with respect to $Q$, $E \{ \nabla D(\hat{P}, Q) \} = \nabla E \{ D(\hat{P}, Q) \}$ vanishes at $Q = E(\hat{P})$, by coherence. Applying Lemma 4.1 with $h(P) = \nabla D(P, Q)$, we deduce

$$\nabla D(P, Q) = E_P \nabla D(\delta_x, Q) \quad (4.12)$$

where $\delta_x \in \mathcal{S}$ has entries 0, except for a 1 in the position corresponding to $x \in \mathcal{X}$.

Now let $S(x, Q) = D(\delta_x, Q)$, so that $S(P, Q) = E_P D(\delta_x, Q)$, and so $\nabla S(P, Q) \equiv \nabla D(P, Q)$. It follows that $D(P, Q)$ has the form $S(P, Q) - g(P)$, whence, since $D(P, P) = 0$, $D(P, Q) = S(P, Q) - S(P, P)$, and so $D$ is proper.

The above logic will also apply in more general sample spaces $\mathcal{X}$, under sufficient regularity conditions, with $\nabla$ now denoting the Gâteaux derivative with respect to $Q$.
5 Uncertainty measures

Given a decision problem as in \S 2, define

$$H(P) = L \{ P, a(P) \} ,$$  \hspace{1cm} (5.1)

the Bayes loss when \( X \sim P \).

When (2.1) holds, we obtain

$$H^*(P) = cH(P) + k(P) \hspace{1cm} (5.2)$$

where \( k(P) = E_P \{ k(X) \} \) is an affine function of \( P \). When \( H^* \) and \( H \) are related in this way they are termed \emph{equivalent} (\emph{strongly equivalent} if \( c = 1 \)).

If \( S \) is the proper scoring rule associated with \( L \), we have

$$H(P) = S(P, P) \hspace{1cm} (5.3)$$

In particular, from (3.6) we derive \( H(P) = \int H_3(P)dW(\lambda) \).

We find the following forms for \( H(P) \) in our examples:

\textbf{Examples 1 and 3:} \( \text{var}_P(Y) \)

\textbf{Example 2:} \( - \int p(x) \log p(x) d\mu \), the entropy.

\textbf{Example 4:} \( \int \left\{ 1 - |\phi_p(t)|^2 \right\} dW(t) \).

\textbf{Example 5:} \( -\|P\|^2_k \).

\textbf{Example 6:} equivalent to \( b(\theta_P) - \theta^T_P \tau_P \).

\textbf{Example 7:} equivalent to \( \log \det \Sigma_P \).

Now let \( \mathcal{M} \) denote the vector space of finite signed measures on \( \mathcal{X} \), and for \( Q \in \mathcal{P} \) define a linear function \( \psi_Q : \mathcal{M} \rightarrow \mathbb{R} \) by

$$\psi_Q(M) = \int S(x, Q)dM. \hspace{1cm} (5.4)$$

Then for \( P \in \mathcal{P} \)

$$\psi_Q(P - Q) = S(P, Q) - S(Q, Q) \geq S(P, P) - S(Q, Q).$$

Hence

$$H(P) \leq H(Q) + \psi_Q(P - Q). \hspace{1cm} (5.5)$$

When \( \mathcal{X} \) is finite, we can take \( \mathcal{P} = \mathcal{S} \), the unit simplex in \( \mathbb{R}^N \), \( \mathcal{M} = \{ u \in \mathbb{R}^N : 1^T u = 0 \} \), and \( \psi_Q(u) \) has the form \( \psi^T_Q u \) with \( \psi_Q \in \mathcal{M} \). Then (5.5) gives

$$H(P) \leq H(Q) + \psi^T_Q (P - Q). \hspace{1cm} (5.6)$$
For fixed $Q$, varying $P$, the right-hand side of (5.6) defines a “supporting hyper-plane” to $H$ at $Q$, and it follows that $H$ is concave; conversely, any concave $H$ has the property (5.6), for suitable $\psi_Q$, not necessarily unique. When $H$ is differentiable at $Q$, $\psi_Q = \nabla H(Q)$. In more general spaces, (5.5) implies concavity, as expressed in the property: if $P$ has a random distribution over $\mathcal{P}$, then

$$E\{H(P)\} \leq H\{E(P)\}. \quad (5.7)$$

Various authors (e.g. DeGroot, 1962; Rao, 1982) have used general concave functions on $\mathcal{P}$ as measures of uncertainty or diversity. Concavity need not imply (5.5): see Hendrickson and Buehler (1971), Example 4.1, for a counter-example. It will do so, however, under suitable continuity conditions (ibid., Theorem 4.1). Under differentiability, $\psi_Q$ will be the Gâteaux derivative of $H$ at $Q$.

When (5.5) holds, $\psi_Q$ is termed a super-gradient of $H$ at $Q$. A function $H$ possessing a super-gradient $\psi_Q$, not necessarily unique, at every $Q \in \mathcal{P}$ will be called a proper uncertainty measure. It will be called regular if each $\psi_Q$ is uniquely determined, and strict if, for all $Q$, $\psi_Q$ can be chosen so that strict inequality holds in (5.5) unless $P = Q$. A bounded proper uncertainty measure $H$ has an equivalent “canonical form” $H^*$, given by

$$H^*(P) = H(P) - E_P \{H(\delta_X)\}, \quad (5.8)$$

such that $H^*(P) \geq 0$, with equality when $P$ is concentrated on a point.

Now let $H$ be a given proper uncertainty measure. Then $H$ has a representation (5.3) in terms of a proper scoring rule $S$. For define

$$S(x, Q) = H(Q) + \psi_Q(\delta_x - Q). \quad (5.9)$$

Then $S(P, Q) = H(Q) + \psi_Q(P - Q)$, $S(P, P) = H(P)$, and (5.5) shows that $S$ is proper. If $H$ is strict, and $\psi_Q$ chosen accordingly, $S$ will be strictly proper. If $\psi_Q$ is not uniquely determined, nor is $S(x, Q)$. However, when $H$ is regular, so that $\psi_Q$ is unique, (5.4) determines $S(x, Q)$ up to an additive function of $Q$, which is then fixed by (5.3), so that $S$ is unique.

Given proper $H$, a corresponding proper discrepancy is given by

$$D(P, Q) = H(Q) - H(P) + \psi_Q(P - Q). \quad (5.10)$$

Conversely, from $D$ we can recover $S$, and hence $H$, up to equivalence, obtaining

$$H(P) = E_P \{D(\delta_X, P)\}. \quad (5.11)$$

In Example 1, we have

$$H(P) = H(Q) + \int y^2 d(P - Q)(y) - \left\{E_Q(Y) + \int y d(P - Q)(y)\right\}^2 + E_Q(Y^2),$$
whence $H(P) - H(Q) + \{E_P(Y) - E_Q(Y)\}^2$ is, for fixed $Q$, linear in $P - Q$, and thus, since the last term is of second order for small $P - Q$, must be $\psi_Q(P - Q)$. Hence $S(P, Q) = \text{var}_P(Y) + \{E_P(Y) - E_Q(Y)\}^2$, and we recover $S(x, Q) = S(\delta_x, Q) = \{y - E_Q(Y)\}^2$. For Example 2, $H(P) - H(Q) = -\int(1 + \log q)d(P - Q) + o(|P - Q|)$, whence $S(P, Q) = -\int \log q dP$, and we recover $S(x, Q) = -\log q(x)$. The other examples may be verified similarly.

The above shows, that, under broad conditions, any concave real functional $H$ on $\mathcal{P}$ may be regarded as the Bayes loss in a suitable decision problem.

6 Dependence measures

Let $L(x, a)$ be a loss function as in §2, with $x \in \mathcal{X}$. Let the space $\mathcal{U}$ be arbitrary, and consider a pair $(X, U)$ of random variables, with joint distribution $P^{X,U}$ over $\mathcal{X} \times \mathcal{U}$. Let $P^X$ denote the induced marginal distribution for $X$, and $P^X_u$ the conditional distribution for $X$, given $U = u$, etc. Then the Bayes loss if an action must be chosen immediately is $H(P^X) = L\{P^X, a(P^X)\}$, while the expected Bayes loss if $U$ is to be observed first is $E\{H(P^X_U)\}$, the expectation being over the marginal distribution $P^U$ of $U$. Since $H$ is concave and $E\{P^X_U\} = P^X$, we have $E\{H(P^X_U)\} \leq H(P^X)$, with equality if $X$ and $U$ are independent. The quantity

$$C(P^{X,U}) = H(P^X) - E\{H(P^X_U)\}$$

may be considered a measure of the dependence of $X$ on $U$ in $P^{X,U}$. It is non-negative and vanishes if $X \perp U$; when $H$ is strict, $C$ vanishes only in this case. Note that, even if $\mathcal{U} = \mathcal{X}$, $C(P^{X,U})$ need not be symmetric as between $X$ and $U$. In the original decision problem,

$$C(P^{X,U}) = E_{P^{X,U}}[L\{X, a(P^X)\} - L\{X, a(P^X_U)\}]$$

defines the “expected value of sample information” contained in $U$ (Raiffa and Schlaifer, 1961, §4.5.2). More generally, $C$ might be defined directly in terms of (6.1), for a given proper uncertainty function $H$. Note that $C$ is defined for all joint distributions over $\mathcal{X} \times \mathcal{U}$, for arbitrary sample space $\mathcal{U}$. If $V$ is a one-one function of $U$, $C(P^{X,V}) = C(P^{X,U})$. However, $C$ is typically not preserved under transformation of $X$.

Using the relations already established between $H, S$ and $D$, we can alternatively express (6.1) as:

$$C(P^{X,U}) = E\{S(X, P^X) - S(X, P^X_U)\},$$

or

$$C(P^{X,U}) = E\{D(\delta_X, P^X) - D(\delta_X, P^X_U)\}.$$
Strongly equivalent versions of $H$, or of $S$, lead to identical $C$.

We further have:

**Lemma 6.1** For any $Q \in \mathcal{P}$,

$$E \left\{ D(P_{U}^{X}, Q) \right\} - D(P^{X}, Q) = C(P^{X,U}). \tag{6.5}$$

**Proof** Follows from $S(P^{X}, Q) = E \left\{ S(P_{U}^{X}, Q) \right\}$, and $D(P, Q) = S(P, Q) - H(P)$. \qed

**Corollary 6.1** We have

$$C(P^{X,U}) = E \left\{ D(P_{U}^{X}, P^{X}) \right\}. \tag{6.6}$$

In Examples 1 and 3 we obtain $C = \text{var} \left\{ E(Y \mid U) \right\}$, while Example 2 yields

$$C = E \left\{ \log p(X, U)/p(X)p(U) \right\}, \tag{6.7}$$

the “mutual information” or “cross-entropy”. Formula (6.7) does treat $X$ and $Y$ symmetrically and it seems plausible that this is the only case in which this holds. For Example 4 we obtain

$$C = \int \left[ E \left\{ |\phi_{P^{X}}(t)|^2 - |\phi_{P^{X}}(t)|^2 \right\} dW(t) \right. \tag{6.8}$$

$$\left. = \int E \left\{ |\phi_{P^{X}}(t) - \phi_{P^{X}}(t)|^2 \right\} dW(t) \right. \tag{6.9}$$

and for Example 5, $C = E \left\{ \| P_{U}^{X} - P^{X} \|_{K}^2 \right\}$. In Example 6 we have

$$C = b(\theta_{P^{X}}) - \theta_{P^{X}}^{T} \tau_{P^{X}} - E \left\{ b(\theta_{P^{X}}) - \theta_{P^{X}}^{T} \tau_{P^{X}} \right\},$$

while the special case of Example 7 gives

$$C = \frac{1}{2} E \left[ \log \left\{ \det \Sigma_{P^{X}} / \det \Sigma_{P^{X}} \right\} \right].$$

**Lemma 6.2** Let $(X, U, V)$ have a joint distribution such that $X \perp \! \! \! \perp V \mid U$. Then $C(P^{X,U}) \geq C(P^{X,V})$.

**Proof** Since $E(P_{U}^{X} \mid V) = P_{V}^{X}$, and $H$ is concave, $E \left\{ H(P_{U}^{X}) \mid V \right\} \leq H(P_{V}^{X})$, so that $E \left\{ H(P_{U}^{X}) \right\} \leq E \left\{ H(P_{V}^{X}) \right\}$, and the result follows from (6.1). \qed

**Corollary 6.2** If $V$ is a function of $U$, $C(P^{X,U}) \geq C(P^{X,V})$. 

13
Corollary 6.3 If $H$ in (6.1) is strict, $X \perp V \mid U$ and $C(P^{X,U}) = C(P^{X,V})$, then $X \perp U \mid V$.

Proof In this case $E \left\{ H(P^{X,V}_U) \mid V \right\} = H(P^X_V)$ almost surely, so by strictness $P^X_{U,V} = P^X_V$. \qed

Corollary 6.4 For fixed $P^X$, $C(P^{X,U})$ is maximised for $U = X$.

In the case $U = X$ of Corollary 6.4, we obtain from (6.1)

$$H^*(P^X) = C(P^{X,X}),$$

(6.10)

where $H^*$ is the canonical version of $H$, given by (5.8). Hence, given $C$, $H$ may be recovered up to equivalence. A dependence measure $C$ which may be represented as in (7.1) for some proper uncertainty measure $H$ may be called proper, and strict if $H$ is strict. However, we do not currently have a simple characterisation of proper dependence measures, except by recovering $H$ through $H(P^X) = C(P^{X,X})$, and checking that $H$ is a proper uncertainty function and (6.1) is satisfied. Such a routine may not be available without modification when $C$ is unbounded, as is the case for the cross-entropy (6.7).

7 Summary of relationships

Table 1 summarises the major relationships between the various measures discussed in this paper. Each entry gives the relevant equation relating the quantity labelling its column to that labelling its row, such a relationship being determined up to strict equivalence. Given any of the other quantities, a suitable definition of a related loss function $L$ may be taken to be the same as $S$. Expressions for $S$ and $D$ in terms of $C$ are not given explicitly, but may be obtained by first expressing them in terms of $H$.

8 Predictive design

Suppose that we ultimately wish to predict, or otherwise take an action with value depending on, a quantity $X \in \mathcal{X}$, having a distribution $P^X_\theta$, conditional on the value $\theta$ for a parameter $\Theta$. There is a class $\mathcal{E}$ of possible experiments. If $e \in \mathcal{E}$ is performed, then observations will be made of a quantity $Y_e$, with distribution $P^Y_{\theta,e}$, over a sample space $\mathcal{Y}_e$ when $\Theta = \theta$. After observing $Y_e = y$, an action $a \in \mathcal{A}$ is to be chosen. Finally $X$ is observed and the loss suffered if $X = x$ is $L(x,a)$. It is supposed that, given $\Theta = \theta$, $X \sim P^X_\theta$ independently of the choice of $e$ and $a$ and the value $y$ of $Y_e$. To complete the specification,
Table 1: Equations specifying the pairwise relationships between a loss function $L(x, a)$, a proper scoring rule $S(x, P)$, a proper uncertainty measure $H(P)$, a proper discrepancy measure $D(P, Q)$, and a proper dependence measure $C(P^X, U)$.

<table>
<thead>
<tr>
<th></th>
<th>$S$</th>
<th>$H$</th>
<th>$D$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>(3.2)</td>
<td>(5.1)</td>
<td>(4.2)</td>
<td>(6.2)</td>
</tr>
<tr>
<td>$S$</td>
<td>(5.3)</td>
<td>(4.1)</td>
<td>(6.3)</td>
<td></td>
</tr>
<tr>
<td>$H$</td>
<td>(5.9)</td>
<td>(5.10)</td>
<td>(6.1)</td>
<td></td>
</tr>
<tr>
<td>$D$</td>
<td>(4.11)</td>
<td>(5.11)</td>
<td>(6.6)</td>
<td></td>
</tr>
<tr>
<td>$C$</td>
<td>(6.10)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

we assign to $\Theta$ a prior distribution $\Pi$, independent of the choice of $\epsilon \in \mathcal{E}$. The objective is, first to choose the optimal experiment $\epsilon$ from $\mathcal{E}$, and then, after observing $Y_\epsilon$, the optimal action $a$ from $\mathcal{A}$, in such a way as to minimise the overall expected loss. The emphasis here will be on the choice of $\epsilon$.

### 8.1 Direct analysis

A typical path through the decision tree for this problem looks like Figure 1:

\[ 
\begin{array}{c}
\square \rightarrow \bigcirc \rightarrow \square \rightarrow \bigcirc \rightarrow \square \rightarrow \bigcirc \rightarrow \bigcirc \\
\nu_0 \quad \nu_1 \quad \nu_2 \quad \nu_3 \quad \nu_4 \quad \nu_5 \\
\end{array}
\]

\[ L(x, a) \]

\textbf{Fig. 1}

We can solve this using the usual method of “backwards induction”. Thus:

(i) At $\nu_4$ we evaluate the expected loss as

\[ E\{L(X, a)|\epsilon, y, a, \theta\} = E\{L(X, a)|\theta\} = L(P_\theta^X, a). \]

(ii) At $\nu_3$ we then calculate $E\{L(P_\theta^X, a)|\epsilon, y, a\} = L(P_{\epsilon, y}^X, a)$, where $P_{\epsilon, y}^X$ denotes the predictive distribution for $X$ after observing $Y_\epsilon = y$ in $\epsilon$.

(iii) At $\nu_2$ we select $a$ to minimise $L(P_{\epsilon, y}^X, a)$, viz. take $a = a(P_{\epsilon, y}^X)$. The corresponding expected loss is then $H(P_{\epsilon, y}^X)$, where $H$ is the proper uncertainty measure corresponding to $L$.

(iv) At $\nu_1$ the expected loss is thus $E\{H(P_{\epsilon, Y_\epsilon}^X)|\epsilon\}$, the expectation being over $Y_\epsilon$. 

15
(v) Finally, at $\nu_0$, the experiment $\epsilon$ is chosen to minimise
\[
E\{H(P^X_{\epsilon,Y})|\epsilon\}.
\]

Thus we see that the criterion for choosing $\epsilon$ can be expressed as:

\[
\text{Minimise the expected posterior uncertainty about } X \text{ consequent on observing the outcome } Y_\epsilon \text{ of } \epsilon.
\]

Since any proper uncertainty measure can be regarded as arising from a suitable decision problem, the above criterion will thus yield a coherent approach to the choice of $\epsilon$ when any proper uncertainty measure over $\mathcal{P}^X$, the set of all distributions for $X$, is used. In particular, this establishes the coherence of basing the choice of experiment on $H(P) = -\int p \log pd\mu$, the entropy, as arises in Example 2; on $H(P) = -\int |\phi_P(t)|^2 dW(t)$, arising from Example 4, as done by Eaton et al. (1995); or on $\log \det \Sigma_P$, as in Example 7, thus extending the standard design criterion of $D$-optimality to distributions which need not be normal.

### 8.2 Scoring rules

Now consider replacing $\mathcal{A}$ above by $\mathcal{P}^X$, $a \in \mathcal{A}$ by $Q \in \mathcal{P}^X$, and $L(x,a)$ by the associated proper scoring rule $S(x,Q)$. The decision tree is thus:

```
\begin{array}{cccccc}
\epsilon & y & Q & \theta & x & \\
\square & \bigcirc & \square & \bigcirc & \bullet & S(x,Q) \\
\nu_0 & \nu_1 & \nu_2 & \nu_3 & \nu_4 & \nu_5
\end{array}
```

Fig. 2

Applying the analysis of §8.1 to this new problem, we have:

(i)' At $\nu_4$, find $S(P^X_\theta, Q)$.

(ii)' At $\nu_3$, find $S(P^X_{\epsilon,Y}, Q)$.

(iii)' At $\nu_2$, choose $Q = P^X_{\epsilon,Y}$, to obtain $H(P^X_{\epsilon,Y})$.

Steps (iv)' and (v)' then agree with (iv) and (v) of §8.1. Thus we may, without loss of generality, always assume that our problem is one of probability prediction with an arbitrary proper scoring rule $S$ for $X$, and its solution is then given by (8.11) for the associated uncertainty measure $H(P) = S(P, P)$.
8.3 Discrepancy functions

Next, suppose we start with a proper discrepancy measure $D$ on $\mathcal{P}^X$, and let $S$ and $H$ be the associated proper scoring rule and uncertainty measure, as given by (4.11) and (5.11). Step (i)' of §8.2 can then be reformulated as:

(i)" At $\nu_4$, find $H(P_{\Theta}^X) + D(P_{\Theta}^X, Q),$

whence follows:

(ii)" At $\nu_3$, find $E\{H(P_{\Theta}^X)|\epsilon, y\} + E\{D(P_{\Theta}^X, Q)|\epsilon, y\}.$

Then

(iii)" At $\nu_2$, choose, as at (iii)' above, $Q = P_{\epsilon, Y}^X,$ thus obtaining

$$E \left\{ H \left( P_{\Theta}^X \right) \mid \epsilon, y \right\} + E \left\{ D \left( P_{\Theta}^X, P_{\epsilon, Y}^X \right) \mid \epsilon, y \right\}.$$

(iv)" At $\nu_1$, find $E\{H(P_{\Theta}^X)\} + E\{D(P_{\Theta}^X, P_{\epsilon, Y}^X)|\epsilon\}$.

(v)" At $\nu_0$ choose $\epsilon$ to minimise

$$E\{D(P_{\Theta}^X, P_{\epsilon, Y}^X)|\epsilon\}, \quad (8.12)$$

the expected discrepancy between the sampling distribution of $X$ given $\Theta$, and the predictive distribution of $X$ given $Y_\epsilon$ in $\epsilon$, the expectation being over the joint distribution of $(\Theta, Y_\epsilon)$.

It follows that the criterion (8.12) for experimental design is coherent, for any proper discrepancy measure $D$ on $\mathcal{P}^X$. In particular, we can use the Kullback-Leibler discrepancy (4.4); the quadratic discrepancy (4.7) or more generally (4.8); or, corresponding to “$D$-optimality”, a discrepancy equivalent to (4.9).

8.4 Dependence measures

If $C$, satisfying (6.1), is the proper dependence measure associated with the decision problem, it follows from (8.11) that the optimal experiment $\epsilon$ minimises $C(P_{\Theta}^X, Y_\epsilon)$, the dependence of $X$ on $Y_\epsilon$. This criterion will be coherent for any proper dependence measure. However, the usefulness of this characterisation is limited, since we do not currently have a simple criterion for the propriety of a dependence measure.
8.5 Equivalent problems

Suppose that we have two equivalent decision problems with loss functions $L(x, a)$ and $L'(x, a)$ such that $L'(x, a) = cL(x, a) + k(x)$, for some function $k$ and constant $c > 0$. In this case we have, in a natural notation:

$$
L'(P, a) = cL(P, a) + k(P),\quad a'(P) = a(P),
$$
$$
S'(x, Q) = cS(x, Q) + k(x),\quad S'(P, Q) = cS(P, Q) + k(P),
$$
$$
H'(Q) = cH(Q) + k(P),\quad D'(P, Q) = cD(P, Q),
$$
$$
C'(P^X, V) = cC(P^X, V).
$$

In particular, we see from (8.12) that the optimal choice of $\epsilon$ at $\nu_0$ in Figure 1 will be the same for equivalent problems, as will be the optimal choice for $a$ at $\nu_2$.

8.6 Prediction and estimation

The problem of estimation may be regarded as a trivial special case of the above predictive framework, on taking $X \equiv \Theta$, and thus $P_\theta = \delta_\theta$, the point mass at $\theta$.

Conversely, any predictive decision problem may itself be recast as one of estimation. Define $L^*(\theta, a) = L(P^X_\theta, a)$, viz. $E\{L(X, a)\}$ for $X \sim P^X_\theta$. If we truncate Figure 1 at $\nu_4$, and attach there the loss $L^*(\theta, a)$, in accordance with step (i) of §8.1, this will not affect steps (ii)-(v), and thus we shall obtain same optimal $\epsilon$ and $a$. Only now the problem has been cast as one of estimation, with loss function $L^*(\theta, a)$.

If we now analyse this new estimation problem in its own right, we find, in an obvious notation, the following associated quantities, where, for a distribution for $\Theta$, $P^X_\Pi = \int P^X_\theta d\Pi(\theta)$ denotes the induced predictive distribution for $X$ in the original problem:

$$
L^*(\Pi, a) = L(P^X_\Pi, a),\quad a^*(\Pi) = a(P^X_\Pi),
$$
$$
S^*(\theta, \Gamma) = S(P^X_\theta, P^X_\Gamma),\quad S^*(\Pi, \Gamma) = S(P^X_\Pi, P^X_\Gamma),
$$
$$
H^*(\Gamma) = H(P^X_\Gamma),\quad D^*(\Pi, \Gamma) = D(P^X_\Pi, P^X_\Gamma),
$$

and, so long as $V \perp X \mid \Theta$,

$$
C^*(P^{\Theta, V}) = C(P^{X, V}).
$$
Hence the above relations define proper scoring rules, uncertainty functions, discrepancy measures and dependence measures on the parameter space, which could be applied directly in the problem of estimation. Note that their definitions depend on the form of the original distributions $P_{\theta}^X$, as well as on the original predictive loss function $L(x, a)$. An interesting problem is to characterise when, given (say) an estimative uncertainty function $H^*$, it can arise as above from some loss function in a predictive problem with given $\{P_{\theta}^X\}$.

Now consider again the analysis of §8.3. It is readily seen that the terms involving $H$ in $(i)'' - (v)''$ do not affect the minimising value for $Q$ in $(iii)''$ and for $\epsilon$ in $(v)''$. Consequently, the identical solutions are obtained from the decision problem of Fig. 3, where we have ignored $H$ and collapsed out the final stage using $(i)''$:

**Fig. 3**

\[
\begin{array}{cccccc}
  & & & \epsilon & y & Q \\
\square & \rightarrow & \bigcirc & \rightarrow & \square & \rightarrow & \bigcirc & \rightarrow & \bullet & D(P_{\theta}^X, Q) \\
  \nu_0 & \rightarrow & \nu_1 & \rightarrow & \nu_2 & \rightarrow & \nu_3 & \rightarrow & \nu_4 \\
\end{array}
\]

In this reformulation, the task is regarded as predicting the sampling distribution $P_{\theta}^X$, with loss function $L'(\theta, Q)$ given by the discrepancy $D(P_{\theta}^X, Q)$. This is the approach taken by Eaton et al. (1995) using (4.8) and, in particular, its special case (4.7). Note that $L'(\theta, Q)$ is an “estimative” loss function, with action space $\mathcal{P}^X$. The resulting $S'$, $H'$, $D'$ and $C'$ are strongly equivalent to $S^*$, $H^*$, $D^*$ and $C^*$ above, and of course the optimal $\epsilon$ and $a$ are the same.

### 8.7 Summary

We have seen above that the following equivalent criteria may be used for coherent choice of experiment $\epsilon$:

(a) Minimise $E\{\min_a L(P_{\epsilon Y_e}^X, a)\}$, where $P_{\epsilon Y_e}^X$ is the predictive distribution for $X$ given $Y_e = y$. The expectation is over the distribution of $Y_e$ in $\epsilon$. This is the standard Bayesian decision-theoretic extensive form for minimizing overall expected loss.

(b) Minimise $E[\min_Q \{S(P_{\epsilon Y_e}^X, Q)\}]$, where $S(\cdot)$ is a proper scoring rule. This is a special case of (a) with $L$ replaced by $S$.

(c) Minimise the expected predictive uncertainty $E\{H(P_{\epsilon Y_e}^X)\}$, where $H$ is a proper uncertainty measure.
(d) Minimise \( E \{ D(P^X_\Theta, P^X_{eY_e}) \} \), the expectation being over the joint distribution of \((\Theta, Y_e)\), with \( D \) a proper discrepancy measure. This is the expected discrepancy of the predictive distribution from the sampling distribution of \( X \) (a one point distribution in the case of estimation).

(e) Proceed as for (a), taking \( \mathcal{A} = \mathcal{P}^X \), and loss function \( L'(\theta, Q) = D(P^X_\theta, Q) \), where \( D \) is a proper discrepancy measure on \( \mathcal{P}^X \).

(f) Maximize \( C(P^{X,Y_e}) \), the dependence of \( X \) on \( Y_e \), where \( C \) is a proper dependence measure.

When the various measures used in the different criteria above are related to each other in the manner we have described, all the criteria will yield the same solution. In particular, since any proper measure can be regarded as decision-based, and thus as deriving from a loss function as used in (a), any of the above criteria may be considered coherent.

As particular cases of the above, we note the equivalence, and coherence, of the various criteria in the following examples:

Example 1 (with \( Y \equiv X \))
This loss function, together with additional cost terms, was applied to experimental design by Brooks (1972).

(a) Act optimally under the quadratic loss \( L(x, a) = (x - a)^2 \).

(b) Act optimally under the quadratic scoring rule
\[
S(x, Q) = \{ x - E_Q(X) \}^2.
\]

(c) Minimise
\[
E \{ \text{var}(X|Y_e) \}.
\]

(d) Minimise
\[
E \{ \{ E(X|\Theta) - E(X|Y_e) \}^2 \}.
\]

(e) Act optimally under the loss function
\[
L'(\theta, Q) = \{ E(X|\theta) - E_Q(X) \}^2,
\]
for \( Q \) a distribution for \( X \).

(f) Maximize
\[
C(P^{X,Y_e}) = \text{var} \{ E(X|Y_e) \}.
\]

Example 2
This approach to design was pioneered by Lindley (1956).
(a) and (b) Act optimally under the logarithmic penalty $S(x, P) = -\log p(x)$.

(c) Minimise the expected predictive entropy.

(d) Minimise the expected Kullback-Leibler distance of the predictive distribution from the sampling distribution.

(e) Solve the “estimation problem” in which the loss, when $\Theta = \theta$ and a distribution $Q$ for $X$ is quoted, is the Kullback-Leibler discrepancy between $P_\theta$ and $Q$.

(f) Maximise the cross-entropy between $X$ and $Y_e$.

**Example 4**

This is the basis of the approach of Eaton et al. (1995).

(a) and (b) Act optimally under the loss function

$$S(x, Q) = \int |e^{itx} - \phi_Q(t)|^2 dW(t).$$

(c) Maximise

$$\int E |\phi_{P_X}(t)|^2 dW(t).$$

(d) Minimise

$$\int E \left\{ |\phi_{P_{Y_e}}(t) - \phi_{P_X}(t)|^2 \right\} dW(t).$$

(e) Act optimally under the loss function

$$L'(\theta, Q) = \int |\phi_{P_{\hat{\theta}}}(t) - \phi_Q(t)|^2 dW(t).$$

(f) Maximise

$$\int E \left\{ |\phi_{P_{\hat{\theta}}}(t) - \phi_{P_X}(t)|^2 \right\} dW(t).$$

**Acknowledgments**

I have benefited greatly from discussions with P. Sebastiani and A. Giovagnoli.
References


