ORKUSTOFNUN

Summary of Main Concepts and Terminology of Preposteriory Analysis of the Worth of Hydrological Data

Tómas Jóhannesson

Greinargerð TJ-95-03

Summary of Main Concepts and Terminology of Preposteriory Analysis of the Worth of Hydrological Data

1. INTRODUCTION

This report defines the main concepts which are used in preposteriory analysis of the worth of hydrological data. The first section defines notation for the concepts and gives brief definitions. The second section provides definitions of the main concepts in terms of integrals over the relevant statistical distributions. The last section briefly describes MATLAB® routines which have been written to evaluate the integrals numerically for specific problems.

2. NOTATION

The following list defines a number of basic concepts frequently used in the literature on preposteriory analysis of the worth of hydrological data. The list is not meant to replace the papers from the literature. Rather, it defines a uniform notation and can serve as a reminder for people who are already familiar with these concepts from the literature.

θ	State of nature, *i.e.* value(s) of (statistical) parameter(s) which describe the system which is being analysed, *i.e.* mean discharge of a river or the parameters in the statistical distribution of floods.
a	Alternatives or actions, *i.e.* parameter(s) describing the variable aspects of a hypothetical hydrological project which is being evaluated, *e.g.* the height of bridge or a dam. A the purpose of the analysis is compute an optimal value of a.
I	Information, *i.e.* hydrological data which is used to reach a decision regarding a.
I^*	Perfect information, *i.e.* information which is sufficiently extensive to determine the state of nature θ exactly.
$p(\theta)$	Statistical distribution of state of nature θ. $p_\theta(\theta \mid i)$ is the distribution of θ given i.
$g(I)$	statistical distrubution of hydrological information I. $g_I(i \mid \theta)$ is the distribution of i given θ.
$d(I)$	Decision rule, which is used to reach a decision $a = d(I)$ regarding a on the basis of I.
$\hat{\theta}$	Parameter estimate, *i.e.* value(s) of θ estimated on the basis of the available information and used in the decision procedure.
$l(a, \theta)$	loss function, variable cost, goal function, *i.e.* the cost of a action a given that the state of nature is in fact equal to θ.
$E(x)$	Expected value of a statistical variable x.

$R(I,d)$ Risk function, $R(I,d) = E(l(d(I),\theta))$, *i.e.* the expected value of the loss function given the statistical distribution of the information I and of the state of nature θ and a decision rule $d(I)$.

$d^*(i)$ Bayesian decision rule. The decision a^*, which given the information i and the associated distribution of θ, $p_\theta(\theta \mid i)$, minimises the expected value of the the loss function $E(l(a,\theta))$.

$R^*(I)$ Bayesian risk, $R^*(I) = R(I,d^*(I))$ The expected value of the loss corresponding to a Bayesian decision rule, *i.e.* $E(l(d^*(I),\theta))$.

OL Opportunity loss. $OL = l(a,\theta) - l(d^*(I^*),\theta)$. The difference between the loss function corresponding to a certain decision and the loss corresponding to an optimal decision given a known state of nature.

XOL Expected Bayesian opportunity loss. $XOL = E(OL) = R^*(I) - R^*(I^*)$ $= R(I,d^*(I)) - R(I^*,d^*(I^*))$. The expected additional cost caused by imperfect information.

$XAOL$ Expected actual opportunity loss (possibly a non-optimal decision rule). $XAOL = E(OL) = R(I,d) - R^*(I^*)$. The expected additional cost caused by imperfect information *and* non-optimal decision rule.

$XXOL$ Expected expected opportunity loss! The expected value of the expected opportunity loss corresponding to added or improved information ΔI. ΔI has its own statistical distribution given certain information i which is already available and the statistical distribution of the state of nature.

$EV\Delta I$ Expected valued of added information!! $EV\Delta I = XOL - XXOL$ or $EV\Delta I = XAOL - XXOL$ depending on the decision rule. $EV\Delta I$ is an estimate of the marginal value of additional data for the decision which is being analysed.

p(A|B) The conditional probability $p(A \mid B)$ of an event A given that an event B occurred is by *Bayes rule* equal to

$$p(A \mid B) = \frac{p(A)p(B \mid A)}{p(B)} \quad .$$

Another useful form of Bayes rule for a set of disjoint events A_i, where $P(\bigcup_i A_i) = 1$, is

$$p(A_i \mid B) = \frac{P(A_k)p(B \mid A_k)}{\sum_i p(A_i)p(B \mid A_i)} \quad ,$$

and still another one for continuous probability distributions is

$$p_{X|Y}(x \mid y) = \frac{p_X(x)p_{Y|X}(y \mid x)}{\int p_X(x)p_{Y|X}(y \mid x)dx} \quad .$$

3. INTEGRALS

Several of the functions that are defined in the previous section are defined as expected values over a statistical distribution. The integrals which arise in these definitions are presented below.

Risk:

$$R(I,d) = \int l(d(i),\theta)g_I(i \mid \theta)p(\theta)did\theta = \int l(d(i),\theta)p_\theta(\theta \mid i)g(i)d\theta di \qquad (1)$$

Bayesian decision $a^* = d^*(i)$ satisfies the minimisation equation:

$$\int l(a^*,\theta)p_\theta(\theta \mid i)d\theta = \min_a \int l(a,\theta)p_\theta(\theta \mid i)d\theta \qquad (1)$$

Bayesian risk:

$$R^*(I) = R(I,d^*) = \int \left[\min_a \int l(a,\theta)p_\theta(\theta \mid i)d\theta\right] g(i)di \qquad (1)$$

4. MATLAB ROUTINES

The routines described below evaluate the statistical integrals and the minimisation which are formulated above. The relevant statistical distributions and loss functions must be formulated as MATLAB routines under specific names which can be freely chosen by the user (.m file in MATLAB). Given a family of such MATLAB functions corresponding to a hydrological problem, the routines can be called to evaluate the integrals or locate optimal Bayesian decisions. The MATLAB graphical environment can be used to plot or further analyse the results of the routines.

In general the routines are able to handle more than one parameter in the statistical distributions, *i.e.* the parameter θ can be a vector and the integrals involving θ can thus be many-dimensional. The actions or the decisions are, on the other, assumed to be described by one and only one parameter a.

risk — expected value of a loss function.

risk(lossfun,tpdffun,a,theta1,theta2 [,tol])

computes the expected value of the loss given the name of a loss function "lossfun(a,theta)", *i.e.* $l(a,\theta)$, state of the world probability density function "tpdffun(theta)", *i.e.* $p(\theta)$, a predefined action a, together with upper and lower limits for integration over the states of the world, *i.e.* θ.

risk(lossfun,tpdffun,apdffun,a1,a2,theta1,theta2 [,tol])

computes the expected value of the loss given a loss function "lossfun(a,theta)", *i.e.* $l(a,\theta)$, state of the world probability density function "tpdffun(theta,a)", *i.e.* $p(\theta \mid a)$, action probability density function "apdffun", *i.e.* $g(a)$, together with upper and lower limits for the numerical integration over actions, a, and states of the world, θ. This function assumes that the implicit probability density distribution of actions $d(i)$ which depends on the distribution $g(i)$ has been explicitly derived as the function $g(a)$ and the risk integral reformulated in terms of this function instead of $g(i)$.

An optional integration error tolerance may be given by the argument tol which has a default value of 0.001 if not specified.

The predefined action a may be an array of values. Actions are described by a single random variable. States of the world can be describe by more than one random variable.

bayesd	Bayesian decision. ...
brisk	Bayesian risk, i.e. [expected] value of a loss function given a Bayesian decision procedure. ...
xol	expected opportunity loss. ...

The MATLAB integrations are preformed using the public domains "Numerical Integration Toolbox" routines which implement composite Gauss integration methods. The minimisation is performed using the routines fmin/fmins which are part of MATLAB's Optimization Toolbox.