View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by RERO DOC Digital Library

Acta Mathematicae Applicatae Sinica, English Series
Vol. 21, No. 2 (2005) 247–256

Acta Mathematicae Applicatae
Sinica, English Series
© Springer-Verlag 2005

Conditionally Optimal Weights of Evidence

Stephan Morgenthaler[1], **Robert G. Staudte**[2]

[1]Ecole polytechnique federale de Lausanne (EPFL), SB IMA, Station 8, 1015 Lausanne, Switzerland
(E-mail: Stephan.Morgenthaler@epfl.ch)
[2]Statistics Department, LaTrobe University, Melbourne, 3086, Australia (E-mail: R.Staudte@latrobe.edu.au)

Abstract A weight of evidence is a calibrated statistic whose values in $[0,1]$ indicate the degree of agreement between the data and either of two hypothesis, one being treated as the null (H_0) and the other as the alternative (H_1). A value of zero means perfect agreement with the null, whereas a value of one means perfect agreement with the alternative. The optimality we consider is minimal mean squared error (MSE) under the alternative while keeping the MSE under the null below a fixed bound. This paper studies such statistics from a conditional point of view, in particular for location and scale models.

Keywords Structural models, invariance, confidence distributions, testing
2000 MR Subject Classification 62F25, 62F35

1 Introduction

Let Y be a random variable whose distribution is either F (under H_0) or G (under H_1). A (non randomized) statistical test is a binary $0-1$ random variable $T = T(Y)$. The event $\{T = 1\}$ is called the rejection region and optimality is defined by

$$E_{Y\sim F}[T(Y)] \le \alpha \text{ and } E_{Y\sim G}[1 - T(Y)] \text{ minimal.}$$

Tests are often defined via test statistics S as $\{T = 1\} = \{S > \text{cv}\}$, where cv is a so-called critical value. If this is the case, a test can be converted to a $[0,1]$-valued variable $p(Y)$, the p-value, in the following manner

$$p(Y) = P_{Z\sim F}\{S(Z) > S(Y)\}.$$

The p-value is informally interpreted as measuring the degree of concordance between the null hypothesis and the data, with larger values indicating better concordance. A weight of evidence formalizes this interpretation of the p-value mathematically. It takes values close to 0 if the null hypothesis is true and close to 1 if the alternative holds. Let $0 \le W(Y) \le 1$ be such a statistic. Its effectiveness can be measured for $q \ge 1$ by $E_{Y\sim F}[W(Y)^q]$ and $E_{Y\sim G}[(1 - W(Y))^q]$, which we will in the following call type-I risk and type-II risk. Admissible weights W minimize

$$E_{Y\sim G}\big[\big(1 - W(Y)\big)^q\big] + \lambda E_{Y\sim F}[W(Y)^q], \tag{1}$$

for some $\lambda \ge 0$. If F, G have densities f, g we can write (1) as

$$\int \big(\big(1 - W(y)\big)^q g(y) + \lambda W(y)^q f(y)\big)dy.$$

Manuscript received March 20, 2004.
Supported in part by a grant from the Swiss National Science Foundation.

The optimal W is such that for all possible variations $\delta(Y)$, we have

$$\int \big(- \delta(y)\big(1 - W(y)\big)^{q-1} g(y) + \lambda\delta(y)W(y)^{q-1} f(y)\big)dy \geq 0,$$

which for $q > 1$ implies $\lambda W(y)^{q-1} f(y) - (1 - W(y))^{q-1} g(y) \equiv 0$ and for $q = 1$, $\int \delta(y)\big(\lambda f(y) - g(y)\big)dy \geq 0$. Thus, for $q > 1$,

$$\frac{1 - W(y)}{W(y)} = \Big(\lambda \frac{f(y)}{g(y)}\Big)^{1/(q-1)}, \tag{2}$$

and for $q = 1$,

$$W(y) = \begin{cases} 0, & \text{if } \lambda f(y) - g(y) > 0; \\ 1, & \text{else.} \end{cases} \tag{3}$$

These admissible weights of evidence are functions of the likelihood ratio $f(y)/g(y)$. In the case of $q = 1$, we obtain the Neyman-Pearson tests with rejection ($W = 1$) if $g/f \geq \lambda$ and non-rejection ($W = 0$) if $g/f < \lambda$. This weight switches from choosing one hypothesis to a preference for the other one without any intermediate region. For big values of q the opposite behavior occurs. The corresponding weights of evidence have a large zone of indifference and choose clearly one of the two hypothesis only when the likelihood ratio is very small or very big. The choice $q = 2$ seems a good compromise and the corresponding optimal weights of evidence have been introduced by Blyth and Staudte[1,2]. The constant $\lambda > 0$ is chosen to bound the type-I risk and thus to calibrate the value of the weight of evidence. Since bigger values of λ imply smaller values of W and thus smaller type-I risk and bigger type-II risk, we have to choose its smallest possible value without violating the bound.

In this paper, we study weights of evidence in the context of structural models, that is parameters defined by groups of transformations applied to a random vector with a fixed distribution. We determine the optimal weight of evidence under the conditioning principle (Section 2). We then show how the optimal weight changes when uncertainty about the underlying distribution is introduced, in particular when allowing for outliers with the use of heavy-tailed distributions (Section 3).

2 Optimal Weights of Evidence for Transformation Models

The quality of a weight of evidence as described in the introduction is to be determined by its risks, that is a sample space average. Such a weight of evidence can reduce overall risks by allowing relatively large local risks in those regions of the sample space that are less probable either under the null or the alternative hypothesis. If a more uniform behavior of the loss is desired, the risk properties conditional on suitable subsets of the sample space are of interest. In the case of transformation models, a canonical ancillary division of the sample space into such subsets exists. In this section, we discuss the resulting weights of evidence.

Let $\boldsymbol{E} \in \mathbb{R}^n$ be a random vector with a known absolutely continuous n-dimensional distribution F and consider $\boldsymbol{Y} = \theta_0(\boldsymbol{E})$ where $\theta_0: \mathbb{R}^n \to \mathbb{R}^n$ is a measurable transformation of $\mathbb{R}^n$. About θ_0 it is only known that it belongs to a set, Θ, of transformations which under composition, $(\theta_1\theta_2)(\boldsymbol{y}) = \theta_1\big(\theta_2(\boldsymbol{y})\big)$, form a locally compact topological transformation group (see [5], first chapter) with the property that $\theta_1 \neq \theta_2 \to \theta_1(\boldsymbol{E}) \neq \theta_2(\boldsymbol{E})$. We can thus form products and inverses of these transformations and there is an identity transformation. The distribution of $\boldsymbol{Y}$ is equal to $F(\theta_0^{-1}(\boldsymbol{y}))$, from which it follows that the likelihood is equal to

$$L(\theta) = f(\theta^{-1}(\boldsymbol{y}))J(\theta^{-1}, \boldsymbol{y}), \tag{4}$$

where $J(\theta^{-1},\cdot)$ is the Jacobian of the transformation θ^{-1}. Within this structure we wish to draw inferences about θ_0 given an observed value $\boldsymbol{y}$ of $\boldsymbol{Y}$. This is done by acting conditionally on the set

$$\langle\boldsymbol{y}\rangle = \{\theta(\boldsymbol{y})\colon \theta\in\Theta\}\subset\mathbb{R}^n,$$

which is the the orbit of the transformation group containing the observation. In cases of interest to statisticians, the group Θ has a p-dimensional representation, each orbit is of dimension $p\le n$, and the set containing all orbits, the orbit space $\mathcal{A}$, is of dimension $n-p$. The orbit obtained from $\boldsymbol{y}=\theta_0(\boldsymbol{e})$ for any fixed $\theta_0\in\Theta$ is the same as that obtained from $\boldsymbol{e}$, that is $\{\theta(\boldsymbol{e})\colon\theta\in\Theta\}=\{\theta\big(\theta_0(\boldsymbol{e})\big)\colon\theta\in\Theta\}$. It follows that the distribution on $\mathcal{A}$ induced by the random vector $\boldsymbol{Y}$ is the same as the one induced by $\boldsymbol{E}$, which implies that the particular orbit picked by the data is an ancillary to the inference problem. This is fundamental to the conditioning argument. Special cases are translations (location problem) and scalings (scale problem). Inferences based on the conditional distribution given the orbit are straightforward because of the reduction in dimension. The above problem is, in more traditional terms, about a parametric family with a p-dimensional parameter and a p-dimensional statistic, namely the position of $\boldsymbol{Y}$ within the orbit $\langle\boldsymbol{y}\rangle$.

To compute conditional expectations such as $E_{\theta_0}(W(\boldsymbol{Y})^2|\langle\boldsymbol{y}\rangle)$ we need a coordinate system for each orbit. The appropriate tool for this are isomorphisms between the orbit and the group Θ. These are mappings

$$T\colon\mathbb{R}^n\to\Theta$$

that satisfy

$$T(\theta(\boldsymbol{Y}))=\theta T(\boldsymbol{Y})\text{ for all }\boldsymbol{Y}\in\mathbb{R}^n\text{ and for all }\theta\in\Theta \tag{5}$$

and are called equivariant statistics. Choosing such a statistic partitions the information in the data $\boldsymbol{Y}$ into two parts, $\boldsymbol{Y}\equiv\langle\boldsymbol{y}\rangle\oplus T(\boldsymbol{Y})$. If we act conditionally, the choice of T is irrelevant since within any orbit all equivariant maps have a very simple structure. If T and T' are two equivariant statistics, it follows from (5) that for an arbitrary $\boldsymbol{z}=\theta(\boldsymbol{y})\in\langle\boldsymbol{y}\rangle$

$$T(\boldsymbol{z})^{-1}T'(\boldsymbol{z})=T(\theta(\boldsymbol{y}))^{-1}T'\big(\theta(\boldsymbol{y})\big)=T(\boldsymbol{y})^{-1}T'(\boldsymbol{y}) \tag{6}$$

is constant. Note that $T(\boldsymbol{z})^{-1}$ refers to the inverse within the transformation group. The formulas are somewhat simplified if we use the particular equivariant map T^* that satisfies

$$T^*(\boldsymbol{y})=\text{identity}. \tag{7}$$

Let $I\subset\Theta$ and consider the probability measure on Θ describing the conditional sampling distribution of the estimator T^*

$$\nu_{\theta_0}(I|\langle\boldsymbol{y}\rangle)=P_{\theta_0}\{T^*(\boldsymbol{Y})\in I|\langle\boldsymbol{y}\rangle\}=P\{T^*(\boldsymbol{E})\in\theta_0^{-1}I|\langle\boldsymbol{y}\rangle\}. \tag{8}$$

This measure is absolutely continuous with respect to any left-invariant Haar measure μ on Θ, which by definition satisfies $\mu(I)=\mu(\theta I)$ for all μ-measurable subsets $I\subset\Theta$ and for all $\theta\in\Theta$, where θI denotes $\{\theta\eta\colon\eta\in I\}\subset\Theta$ (see [9]). When restricted to $\langle\boldsymbol{y}\rangle$, the mapping T^* is bijective and assigns to $t\in\Theta$ the inverse $(T^*)^{-1}(t)=t(\boldsymbol{y})$. From (8) it thus follows that the conditional density of $T^*(\boldsymbol{Y})$ given the orbit $\langle\boldsymbol{y}\rangle$ is equal to

$$d\nu_{\theta_0}(t|\langle\boldsymbol{y}\rangle)\propto f\big(\theta_0^{-1}t(\boldsymbol{y})\big)J(\theta_0^{-1}t,\boldsymbol{y})d\mu(t) \tag{9}$$

$$\propto L(t^{-1}\theta_0)d\mu(t), \tag{10}$$

where μ is a left-invariant Haar measure and where we made use of (4). The normalizing constant in (9) is

$$m(\langle\boldsymbol{y}\rangle)=\int_\Theta f(\theta_0^{-1}t(\boldsymbol{y}))J(\theta_0^{-1}t,\boldsymbol{y})d\mu(t), \tag{11}$$

which due to the left-invariance of μ does not depend on θ_0. We now have all the necessary tools to prove the following result.

Proposition 1. *Suppose* $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ *is equal to* $\theta(\boldsymbol{E})$ *with* $\theta \in \Theta$, *a compact topological group of transformations of* $\mathbb{R}^n$, *and* $\boldsymbol{E}$ *having an absolutely continuous distribution* F *with density* f. *The optimal weight of evidence* $W(\boldsymbol{y})$ *for distinguishing between* $H_0 : \theta = \theta_0$ *and* $H_1 : \theta = \theta_1$ *is a statistic such that* $E_{\theta_1}((1-W)^2)$ *is minimal subject to the constraints* $E_{\theta_0}(W^2|\langle\boldsymbol{y}\rangle) \leq \alpha$ *for all orbits. It is of the form*

$$\frac{1 - W(\boldsymbol{y})}{W(\boldsymbol{y})} = \lambda(\langle\boldsymbol{y}\rangle)\frac{L(\theta_0)}{L(\theta_1)}, \tag{12}$$

where $L(\theta)$ *is the likelihood function,* $L(\theta) = f\big(\theta^{-1}(\boldsymbol{y})\big)J(\theta^{-1},\boldsymbol{y})$, *and* $\lambda(\langle\boldsymbol{y}\rangle)$ *is the smallest positive real such that*

$$\int_\Theta W\big(t(\boldsymbol{y})\big)^2 L(t^{-1}\theta_0)d\mu(t) \leq \alpha m(\langle\boldsymbol{y}\rangle). \tag{13}$$

Proof. The conditionally optimal weight of evidence minimizes

$$E_{\langle\boldsymbol{Y}\rangle}\left\{E_{\theta_1}\big((1 - W(Y_1, \cdots, Y_n))^2|\langle\boldsymbol{Y}\rangle\big) + \lambda(\langle\boldsymbol{y}\rangle)E_{\theta_0}(W(Y_1, \cdots, Y_n)^2|\langle\boldsymbol{Y}\rangle)\right\}$$

and thus

$$\int_\Theta \left([1 - W(t(\boldsymbol{y}))]^2 d\nu_{\theta_1}(t|\langle\boldsymbol{y}\rangle) + \lambda(\langle\boldsymbol{y}\rangle)W\big(t(\boldsymbol{y})\big)^2 d\nu_{\theta_0}(t|\langle\boldsymbol{y}\rangle)\right)$$

for all orbits. The point-wise minimizer of the integrand, evaluated at $t =$ identity $\in \Theta$, thus satisfies

$$\frac{1 - W(y_1, \cdots, y_n)}{W(y_1, \cdots, y_n)} = \lambda\frac{d\nu_{\theta_0}(\text{identity}|\langle\boldsymbol{y}\rangle)}{d\nu_{\theta_1}(\text{identity}|\langle\boldsymbol{y}\rangle)}$$

and (9) gives the desired result (12). The value of λ has to be chosen in such a manner that the conditional type-I risk is bounded by α, that is

$$\int_\Theta W(t(\boldsymbol{y}))^2 L(t^{-1}\theta_0)d\mu(t)\Big/\int_\Theta L(t^{-1}\theta_0)d\mu(t) \leq \alpha,$$

which together with (11) proves (13).

Example 1. Let $y_1, \cdots, y_n$ be a sample from a uniform distribution on the interval $[\theta - 0.5, \theta + 0.5]$. We wish to compute the optimal weight for $\theta_0 = 0$ vs. $\theta_1 = 0.2$. The transformation group for this problem consists of the mappings

$$\theta_r(y_1, \cdots, y_n) = (y_1 + r, \cdots, y_n + r), \qquad r \in \mathbb{R}.$$

The transformation group Θ is the additive group on the reals and the left-invariant Haar measure is proportional to the Lebesgue measure.

The optimal unconditional weight satisfies

$$\frac{1 - W(y_1, \cdots, y_n)}{W(y_1, \cdots, y_n)} = \frac{\lambda L(\theta_0)}{L(\theta_1)} = \frac{\lambda \prod_{i=1}^{n} \{-0.5 \leq y_i \leq +0.5\}}{\prod_{i=1}^{n} \{-0.5 \leq y_i - 0.2 \leq +0.5\}},$$

where $\{a \leq y \leq b\}$ denotes the indicator function of the interval $[a, b]$ evaluated at y. If the true value of $\theta = \theta_0 = 0$, the numerator on the right hand side is equal to one, whereas the denominator is either equal to one or to zero; thus W is either equal to $1/(1+\lambda)$ or 0. It is non-zero, if $y_i \in [-0.3, 0.7]$ for all i, which happens with probability 0.8^n. The constant $\lambda > 0$ must be chosen to ensure $E_0[W^2] = 0.8^n/(1+\lambda)^2 = \alpha$. Clearly, for positive λ, only the values $\alpha \leq 0.8^n$ are possible. For $\alpha = 0.8^n$ the weight $W \equiv 1$, which has zero type-II risk, is optimal.

Conditionally on an observed configuration $\langle \boldsymbol{y} \rangle = \{(y_1 + r, \cdots, y_n + r) \colon r \in \mathbb{R}\}$, the optimal weight is the same as above, except that the proportionality constant depends on the orbit $\lambda = \lambda \langle \boldsymbol{y} \rangle$, which now has to be such that $E_0[W^2 | \langle \boldsymbol{y} \rangle] \leq \alpha$. It is also more convenient to reparametrize the weight and to write $W(\theta_r(\boldsymbol{y})) = W(r)$, which satisfies

$$\frac{1 - W(r)}{W(r)} = \frac{\lambda \langle \boldsymbol{y} \rangle \prod_{i=1}^{n} \{-0.5 \leq y_i + r \leq +0.5\}}{\prod_{i=1}^{n} \{-0.5 \leq y_i + r - 0.2 \leq +0.5\}}.$$

If the true value of $\theta = \theta_0 = 0$, the conditional distribution on the orbit $\langle \boldsymbol{y} \rangle$ has density (see 9)

$$d\nu_{\theta_0 = 0}(\theta_r | \langle \boldsymbol{y} \rangle) \propto \prod_{i=1}^{n} \{-0.5 \leq y_i + r \leq +0.5\},$$

that is r is uniformly distributed on the interval $[r_L = -\min(y_i) - 0.5,\ r_U = -\max(y_i) + 0.5]$. Two cases must be distinguished. First, if $r_U - r_L \leq 0.2$, then the intervals $[r_L, r_U]$ and $[r_L + 0.2, r_U + 0.2]$ do not overlap and $W(r) = \{r_L + 0.2 < r < r_U + 0.2\}$ and has zero conditional type-I and type-II risk. Second, if $r_U - r_L > 0.2$, then the numerator in $(1 - W(r))/W(r)$ is equal to one for all values of r in the interval $[r_L, r_U]$, whereas the denominator is equal to one for $r \in [r_L + 0.2, r_U + 0.2]$. This implies that the conditional weight is equal to

$$W(r) = \begin{cases} 0, & \text{if } r_L < r \leq r_L + 0.2, \\ 1/(1 + \lambda \langle \boldsymbol{y} \rangle), & \text{if } r_L + 0.2 < r \leq r_U, \\ 1, & \text{if } r_U < r \leq r_U + 0.2. \end{cases}$$

The conditional type-I risk is equal to $E_0[W^2 | \langle \boldsymbol{y} \rangle] = ((r_U - r_L - 0.2)/(r_U - r_L))/(1 + \lambda \langle \boldsymbol{y} \rangle)^2$, since the conditional probability for $r_L + 0.2 < r \leq r_U$ equals $(r_U - r_L - 0.2)/(r_U - r_L)$. The conditional type-I risk only take values between zero and $(r_U - r_L - 0.2)/(r_U - r_L)$. When $\alpha \geq (r_U - r_L - 0.2)/(r_U - r_L)$, the trivial weight $W \equiv 1$ is the optimal choice.

The conditional and unconditional solutions can be quite different. For $n = 3$, for example, $0.8^n = 0.512$, whereas $(r_U - r_L - 0.2)/(r_U - r_L)$ can take any value between zero to 0.8. If $(r_U - r_L - 0.2)/(r_U - r_L) \leq 0.512$, that is $\max(y_i) - \min(y_i) \geq 0.59 (= 1 - 0.2/0.488)$, then the unconditional solution uses too large a λ-value and thus has a decreased conditional type-I risk. If the opposite is true, the unconditional solution is dangerous, because in order to control the conditional type-I risk, a larger λ-value ought to be used.

2.1 Fiducial Probabilities

The more traditional inference for the parameter of a transformation model is by way of confidence sets and we are next exploring its link with weights of evidence. As before, let $\boldsymbol{E} \in \mathbb{R}^n$ be a random vector with a known absolutely continuous n-dimensional distribution F and and let $\boldsymbol{y}$ be an observed value of $\boldsymbol{Y} = \theta(\boldsymbol{E})$, with the help of which we want to make inferences about the parameter θ. For any equivariant statistic T we have $T(\boldsymbol{E}) = T(\theta^{-1}(\boldsymbol{Y})) = \theta^{-1}T(\boldsymbol{Y})$ (see 5), which reveals $T(\theta^{-1}(\boldsymbol{Y}))$ to be a pivot, that is a function of the parameter and the data with a constant distribution. Inversion of the pivot allows us to construct confidence sets. If one does not wish to appeal to the conditioning argument, one proceeds by choosing a subset $I \subset \Theta$ such that $P\{T(\boldsymbol{E}) \in I\} = 1 - \alpha$. It then follows that

$$P_\theta\{\theta^{-1}T(\boldsymbol{Y}) \in I\} = P_\theta\{\theta \in T(\boldsymbol{Y})I^{-1}\} = 1 - \alpha,$$

so that $\mathcal{C} = T(\boldsymbol{y})I^{-1} \subset \Theta$ is a $1 - \alpha$ confidence set. In these last equations the set I^{-1} consists of the inverses of the elements of I. With conditioning, the same applies except that the conditional coverage probability $P\{T(\boldsymbol{E}) \in I | \langle \boldsymbol{y} \rangle\}$ now determines our confidence coefficient.

The data-dependent probability measure assigning to subsets of $\mathcal{C} \subset \Theta$ their conditional confidence coefficient,

$$\phi_{\boldsymbol{y}}(\mathcal{C}) = P\{T(\boldsymbol{E}) \in \mathcal{C}^{-1}T(\boldsymbol{y}) | \langle \boldsymbol{y} \rangle\}, \tag{14}$$

was called fiducial measure by R. A. Fisher[3] (for example, Ch. III.3). In the form given here these probabilities were developed by Fraser[4] who later called them structural probabilities. If the parameter is one-dimensional, then the density $d\phi_{\boldsymbol{y}}$ is also called a confidence density. The conditional approach offers two important advantages. First, the confidence coefficient is valid for subsets $\langle \boldsymbol{y} \rangle$ of the sample space and not merely globally. Second, the method is unique, since the dependence of (14) on the choice of the equivariant statistic T is only seeming. Starting with another equivariant statistic T' instead of T, we have for all $\boldsymbol{z} \in \langle \boldsymbol{y} \rangle$ the equality $T(\boldsymbol{z}) = T'(\boldsymbol{z})T'(\boldsymbol{y})^{-1}T(\boldsymbol{y})$ (see 6). Thus,

$$\begin{aligned} P\{T'(\boldsymbol{E}) \in \mathcal{C}^{-1}T'(\boldsymbol{y}) | \langle \boldsymbol{y} \rangle\} =& P\{T(\boldsymbol{E}) \in \mathcal{C}^{-1}T'(\boldsymbol{y})T'(\boldsymbol{y})^{-1}T(\boldsymbol{y}) | \langle \boldsymbol{y} \rangle\} \\ =& P\{T(\boldsymbol{E}) \in \mathcal{C}^{-1}T(\boldsymbol{y}) | \langle \boldsymbol{y} \rangle\}, \end{aligned}$$

which shows that $\phi_{\boldsymbol{y}}$ defines a canonical, data-dependent distribution on the parameter space. It has the property that a set $\mathcal{C} \subset \Theta$ with $\phi_{\boldsymbol{y}}(\mathcal{C}) = 1 - \alpha$ is a confidence set for the unknown θ with exact confidence coefficient $1 - \alpha$, valid both conditionally and unconditionally. To compute $\phi_{\boldsymbol{y}}$ we will again use the particular equivariant map T^* (see 7). Now, by (8)

$$\phi_{\boldsymbol{y}}(\mathcal{C}) = P\{T^*(\boldsymbol{E}) \in \mathcal{C}^{-1} | \langle \boldsymbol{y} \rangle\} = \nu_{\text{identity}}(\mathcal{C}^{-1}),$$

which implies (see 9 and 10) that

$$d\phi_{\boldsymbol{y}}(\theta) \propto f\big(\theta^{-1}(\boldsymbol{y})\big) J(\theta^{-1}, \boldsymbol{y}) d\mu(\theta^{-1}) = L(\theta) d\mu(\theta^{-1}). \tag{15}$$

Note that $d\mu(\theta^{-1})$ defines a right-invariant measure on Θ (see [9] III.14). An intuitive weight of evidence for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ is based on the ratio of the fiducial densities, $(1 - W)/W = \lambda d\phi_{\boldsymbol{y}}(\theta_0)/d\phi_{\boldsymbol{y}}(\theta_1)$. Our formula shows that this is not optimal in the sense of the conditional risk, since we obtain the likelihood ratio modified by a ratio of Haar measures.

Example 2. In this example, we apply the general theory to the scale model. In this case, the transformation group can be represented by the positive reals $\mathbb{R}_+$, assigning to $s \in \mathbb{R}_+$ the transformation $\theta_s(\boldsymbol{E}) = (sE_1, \cdots, sE_n)$. The random variable $\boldsymbol{E} = (E_1, \cdots, E_n)$ has independent and identically distributed components with distribution F and thus $F(\boldsymbol{e}) = \prod_{i=1}^{n} F(e_i)$. The group structure is given by $\theta_s^{-1}(\boldsymbol{e}) = \boldsymbol{e}/s$ and $(\theta_s\theta_t)(\boldsymbol{e}) = \theta_{st}(\boldsymbol{e})$. The orbit of an observed $\boldsymbol{y}$ is equal to $\langle \boldsymbol{y} \rangle = \{s\boldsymbol{y}\colon s \in \mathbb{R}_+\}$, which is called the scale configuration. A left-invariant measure in this group is given by $d\mu(s) \propto ds/s$, where ds denotes the Lebesgue measure. The Jacobian of the transformation θ_s is $J(\theta_s, \boldsymbol{y}) = s^n$. An equivariant estimator is a map $T\colon \mathbb{R}^n \to \mathbb{R}_+$ such that for all positive s, $T(s\boldsymbol{y}) = sT(\boldsymbol{y})$. The conditional density of $T^*(\boldsymbol{Y})$ given the orbit $\langle \boldsymbol{y} \rangle$ is proportional to

$$d\nu_{\theta_0}(\theta | \langle \boldsymbol{y} \rangle) \propto \theta^{n-1} \prod_{i=1}^{n} f(\theta y_i/\theta_0)$$

and the confidence density is proportional to $sL(s)ds = s^{1-n} \prod_{i=1}^{n} f(y_i/s)ds$.

3 Robustness of Weights of Evidence

The optimality of the weights of evidence as defined up to now deals with properties when averaging over sets of samples generated with a known model. However, this model is itself uncertain and should not to be relied on to the extent of being the sole factor in judging the quality of an inference. There are several ideas for dealing with this difficulty.

3.1 Robustness Indicators

The effect of a single observation or of a small set of observations on the conclusions drawn from a weight of evidence ought not be overwhelming. If we conclude that there is a lot of evidence for θ_1 using all the data but very little evidence for θ_1 when setting one of the observations aside, then we should probably weigh the evidence more carefully. This comment shows that we ought to analyze the sensitivity of any weight of evidence, whether proposed as optimal under some model or obtained in some other way, to changes in the data. The sensitivity of a weight of evidence to an arbitrary additional value is most conveniently defined as a change in the value of $\log((1-W)/W)$, because on this logistic scale changes close to 0 and 1 are magnified and because the normalizing constant λ cancels out. Thus, the function

$$\mathrm{SC}(\Delta) = \log\Big(\frac{1-W}{W}(y_1+\Delta, y_2, \cdots, y_n)\Big) - \log\Big(\frac{1-W}{W}(y_1, \cdots, y_n)\Big)$$

is our measure of sensitivity (Note that in order to make the constants cancel, the sample size needs to remain constant under contamination).

Example 3. In the location case, a weight of evidence, optimal when averaging over the whole sample space, has sensitivity

$$\mathrm{SC}(\Delta) = \log\Big(\frac{f(y_1-\theta_0+\Delta)}{f(y_1-\theta_1+\Delta)}\Big) - \log\Big(\frac{f(y_1-\theta_0)}{f(y_1-\theta_1)}\Big),$$

whether or not a single observation can provide overwhelming evidence is determined by whether of not this is bounded, and this in turn is determined by the limiting behavior of $\log(f(x))$ as $|x| \to \infty$. Suppose $\log(f(x)) \sim -M|x|^k$ for large values of $|x|$ and for some $k > 0$. It then follows that

$$\log\big(f(y_1-\theta_0+\Delta)\big) - \log\big(f(y_1-\theta_1+\Delta)\big) \sim M\frac{|\Delta|^k}{\Delta}k(\theta_0-\theta_1).$$

The logarithm of the likelihood ratio under a large contamination is thus unbounded for $k > 1$, tends to a constant for $k = 1$ and tends to zero when $k < 1$. The Gaussian distribution for example has $k = 2$, exponential tails correspond to $k = 1$ and Weibull tails can have $k < 1$. In the Gaussian case, one has $\mathrm{SC}(\Delta) = \Delta(\theta_0 - \theta_1)$. Heavy-tailed distributions have $\log(f(x)) \sim -M\log(|x|)$ and the corresponding log likelihood ratio under large contamination tends to zero. The example shows that the weights of evidence derived from an assumed model are as a rule sensitive to outlying observations, unless the model satisfies certain tail conditions.

Example 4. For the scale model, one has

$$\mathrm{SC}(\Delta) = \log\Big(\frac{f((y_1+\Delta)/\theta_0)}{f((y_1+\Delta)/\theta_1)}\Big) - \log\Big(\frac{f(y_1-\theta_0)}{f(y_1-\theta_1)}\Big),$$

With $\log(f(x)) \sim -M|x|^k$ for large values of $|x|$, we find

$$\log\Big(f((y_1+\Delta)/\theta_0)\Big) - \log\Big(f\big((y_1+\Delta)/\theta_1\big)\Big) \sim M|\Delta|^k(\theta_1^{-1} - \theta_0^{-1}).$$

The situation is even worse than in the location case in that all such distributions have an unbounded sensitivity and one has to turn to heavy-tailed laws in order to obtain bounded sensitivities.

The conditionally optimal weight of evidence are more difficult to analyze because the constant of proportionality depends on the configuration and, of course, the configuration changes between $(y_1 + \Delta, y_2, \cdots, y_n)$ and $(y_1, \cdots, y_n)$.

The sensitivities defined above can be formalized to an influence function for weights of evidence, defined as the Gâteaux derivative of the functional corresponding to $\log((1-W)/W)$. For weights derived from the log likelihood, we have

$$\frac{1}{n}\log\Big(\frac{1-W}{W}(y_1,\cdots,y_n)\Big) = \mathrm{cte}(n) + \frac{1}{n}\sum_{i=1}^{n}\rho_F(y_i,\theta_0) - \frac{1}{n}\sum_{i=1}^{n}\rho_F(y_i,\theta_1), \tag{16}$$

where $\rho_F(x,\theta) = \log(f_\theta(x))$. The term of interest to us is, in functional form,

$$\mathrm{lw}(G) = \int_{-\infty}^{\infty}\rho_F(y,\theta_0)dG(y) - \int_{-\infty}^{\infty}\rho_F(y,\theta_1)dG(y),$$

where G denotes the distribution of the observations. This quantity could be called the mean information per observation from G in favor of F_{θ_0} and in disfavor of F_{θ_1}. In the particular case, where $G = F_{\theta_0}$, it coincides with Kullback's information number. The Gâteaux derivative of (18) in the direction of the Dirac measure Δ_x is defined as the derivative with respect to t, evaluated at $t = 0$, of $\mathrm{lw}((1-t)G + t\Delta_x) - \mathrm{lw}(G)$ and is equal to

$$\mathrm{IF}(x) = \rho_F(x,\theta_0) - \mathrm{lw}(G).$$

Optimal weights of evidence subject to the condition of a bounded influence function have been studied in Morgenthaler and Staudte[8].

Example 5. For $F_\theta(x) = \Phi(x-\theta)$, the Gaussian location model, one finds

$$\mathrm{IF}(x) = (x - \mu_G)(\theta_0 - \theta_1),$$

where μ_G denotes the mean of the distribution G.

3.2 Optimal Robust Weights of Evidence for Transformation Models

The study of influence and breakdown properties shows that methods derived from heavy-tailed distributions are automatically resistant to outliers, gross errors and other wild values. In the context of transformation models, it is therefore natural to consider families of possible models, including heavy-tailed ones, and to derive optimal methods in this context. Morgenthaler and Tukey[6] or Morgenthaler[7] give an introduction to related ideas. In the simplest such case we consider two distributions, F and G, for $\boldsymbol{E}$ and two hypotheses θ_0 and θ_1 in Θ. Since we have two distributions, we also have two likelihood functions, $L_F(\theta)$ and $L_G(\theta)$. Let $0 \le \pi_F$ and $0 \le \pi_G$ and consider the corresponding optimal weight $W_{F \text{ or } G}$ that solves the problem

$$\begin{aligned}&\text{Minimize}\quad \pi_F E_F\big((1 - W\{\theta_1(\boldsymbol{E})\})^2\big) + \pi_G E_G\big((1 - W\{\theta_1(\boldsymbol{E})\})^2\big),\\ &\text{subject to}\quad E_F(W^2\{\theta_0(\boldsymbol{E})\}) \le \alpha \text{ and } E_G(W^2\{\theta_0(\boldsymbol{E})\}) \le \alpha.\end{aligned} \tag{17}$$

Example 6. For the location model and sample size $n = 1$, this problem is equivalent to the problem of minimizing with respect to W the integral

$$\int_{-\infty}^{\infty}\Big(\big(\pi_F f(y-\theta_1) + \pi_G g(y-\theta_1)\big)\big(1 - W(y)\big)^2 + \big(\lambda_F f(y-\theta_0) + \lambda_G g(y-\theta_0)\big)W^2(y)\Big)dy,$$

subject to the constraints in (17). In this expression λ_F, λ_G are Lagrange multipliers. The solution is

$$\frac{1 - W_{F \text{ or } G}(y)}{W_{F \text{ or } G}(y)} = \frac{\lambda_F L_F(\theta_0) + \lambda_G L_G(\theta_0)}{\pi_F L_F(\theta_1) + \pi_G L_G(\theta_1)}, \tag{18}$$

which for $(\pi_F, \pi_G) = (1, 0)$ or $= (0, 1)$ leads us almost back to (3), were it not for the fact that both constraints in (17) must be satisfied. For $n > 1$, the same expression holds. Determining the correct values of the Lagrange multipliers requires in general the numerical computation of n-dimensional integrals.

This example shows that in a situation where, in addition to the unknown parameter value, uncertainty about the underlying distribution is mixed in, the weighing of the evidence still relies on a kind of likelihood ratio. The new "likelihood" is a mixture of the likelihoods included in the model, with the mixing weights depending on the orbit of the data $\langle \boldsymbol{y} \rangle$ and on user specified weights π_F, π_G. Instead of a single Lagrangian constant, used to calibrate the weight of evidence in order for it to have bounded type-I risk, there are now two such constants thus allowing the weight to have bounded type-I risk under both distributions. A conditional version of (17) is

$$\begin{aligned} &\text{Minimize} \quad \pi_F E_F((1 - W\{\theta_1(\boldsymbol{E})\})^2) + \pi_G E_G((1 - W\{\theta_1(\boldsymbol{E})\})^2), \\ &\text{subject to} \quad E_F(W^2\{\theta_0(\boldsymbol{E})\} | \langle \boldsymbol{y} \rangle) \leq \alpha \text{ and } E_G(W^2\{\theta_0(\boldsymbol{E})\} | \langle \boldsymbol{y} \rangle) \leq \alpha, \end{aligned} \tag{19}$$

which describes a weight of evidence with good conditional properties when sampling from F or from G.

Proposition 2. *Suppose $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ is equal to $\theta(\boldsymbol{E})$ with $\boldsymbol{E}$ having an absolutely continuous distribution, either equal to F or equal to G with densities f and g, respectively. The parameter $\theta \in \Theta$ is an element of a compact topological group of transformations of $\mathbb{R}^n$ with left-invariant Haar measure μ. Based on a realization $\boldsymbol{y}$ of $\boldsymbol{Y}$, we wish to weigh the merits of $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. The solution $W_{F \text{ or } G}$ of (19) is of the form*

$$\frac{1 - W_{F \text{ or } G}(\boldsymbol{y})}{W_{F \text{ or } G}(\boldsymbol{y})} = \frac{\lambda_F(\langle \boldsymbol{y} \rangle) L_F(\theta_0) + \lambda_G(\langle \boldsymbol{y} \rangle) L_G(\theta_0)}{\pi_F L_F(\theta_1) + \pi_G L_G(\theta_1)}, \tag{20}$$

where $\pi_F \geq 0$ and $\pi_G \geq 0$ are arbitrary, and $\lambda_F(\langle \boldsymbol{y} \rangle)$ and $\lambda_G(\langle \boldsymbol{y} \rangle)$ are the smallest positive reals such that both

$$\int_\Theta W_{F \text{ or } G}(t(\boldsymbol{y}))^2 L_F(t^{-1}\theta_0) d\mu(t) \leq m_F(\langle \boldsymbol{y} \rangle)\alpha$$

as well as the analogous inequality for G hold (see (11) for the definition of m_F and m_G).

Proof. We wish to minimize

$$\begin{aligned} &E^F_{\langle \boldsymbol{Y} \rangle}\{\pi_F E_{F,\theta_1}\big((1 - W)^2 | \langle \boldsymbol{Y} \rangle\big) + \lambda_F(\langle \boldsymbol{y} \rangle) E_{F,\theta_0}(W^2 | \langle \boldsymbol{Y} \rangle)\} \\ &\quad + E^G_{\langle \boldsymbol{Y} \rangle}\ \{\pi_G E_{G,\theta_1}\big((1 - W)^2 | \langle \boldsymbol{Y} \rangle\big) + \lambda_G(\langle \boldsymbol{y} \rangle) E_{G,\theta_0}(W^2 | \langle \boldsymbol{Y} \rangle)\}. \end{aligned}$$

This can be written as an integral over $\mathcal{A}$ of

$$\begin{aligned} &\pi_F m_F(\langle \boldsymbol{y} \rangle) E_{F,\theta_1}((1 - W)^2 | \langle \boldsymbol{y} \rangle) + \lambda_F(\langle \boldsymbol{y} \rangle) m_F(\langle \boldsymbol{y} \rangle) E_{F,\theta_0}(W^2 | \langle \boldsymbol{y} \rangle) \\ &\quad + \pi_G m_G(\langle \boldsymbol{y} \rangle) E_{G,\theta_1}((1 - W)^2 | \langle \boldsymbol{y} \rangle) + \lambda_G(\langle \boldsymbol{y} \rangle) m_G(\langle \boldsymbol{y} \rangle) E_{G,\theta_0}(W^2 | \langle \boldsymbol{y} \rangle). \end{aligned}$$

The optimal weight evaluated at $\langle \boldsymbol{y} \rangle$ thus satisfies

$$\frac{1 - W}{W} = \frac{\lambda_F m_F d\nu_{F,\theta_0}(\text{identity} | \langle \boldsymbol{y} \rangle) + \lambda_G m_G d\nu_{G,\theta_0}(\text{identity} | \langle \boldsymbol{y} \rangle)}{\pi_F m_F d\nu_{F,\theta_1}(\text{identity} | \langle \boldsymbol{y} \rangle) + \pi_G m_G d\nu_{G,\theta_1}(\text{identity} | \langle \boldsymbol{y} \rangle)}$$

and the proposition follows from (10) and (11).

Example 7. In this example we compare different weights of evidence for $H_0 : \theta = 1$ and $H_1 : \theta = 2$ with θ being a scale parameter. The observed data are $\boldsymbol{y} = (0.554, -0.166, 4.116, -0.213, -0.501)$. The data analyst who thinks that these observations were generated by a normal distribution will find $m_{\text{normal}} = 2.93 \times 10^{-5}$ and a value of the weight of evidence equal to $W_{\text{normal}} = 0.93$. He or she will conclude that the evidence in favor of H_1 is substantial.

Another data analyst, considering the data as being generated by a Cauchy distribution, will conclude the opposite, since in this case, $m_{\text{Cauchy}} = 4.00 \times 10^{-4} = 43 \times m_{\text{normal}}$ and a value of the weight of evidence equal to $W_{\text{Cauchy}} = 0.06$.

The weight (20) with $\pi_1 = \pi_2 = 1$ and using as possible generating distributions both the Cauchy and the normal provides an intermediate compromise. In this case the analyst will find $W_{\text{normal or Cauchy}} = 0.10$, that is, relatively weak evidence in favor of H_1. Since the weight of the observed orbit is much larger in the Cauchy case then in the normal case, the compromise favors the conclusion reached under the Cauchy scenario.

4 Conclusions

Weighing empirical evidence in favor or disfavor of a hypothesis is an archetypical statistical problem to which many different solutions have been proposed. We studied this problem in its simplest form, namely the comparison of two simple hypotheses, and by formulating it as an estimation problem. The restriction of the weight to the interval $[0, 1]$ is arbitrary, but important when using the L^q loss function as we did. Not surprisingly, the optimal solution turns out to be a function of the likelihood ratio. In the context of statistical models where the parameter is an element of a transformation group, the same result holds, except that the calibration of the weight of evidence is done conditionally on the ancillary statistic.

In the final section, we examined the robustness properties of such weights and showed that optimal weights based on heavy-tailed distributions are less sensitive to outliers. This suggests the use of models in which the shape of the error distribution is not completely determined, but rather given by a range of possibilities. The optimal weights for such models were derived in the paper.

References

[1] Blyth, C.R., Staudte, R.G. Hypothesis estimation and confidence profiles. *Proceedings of the 49th Session of the International Statistical Institute*, 135–136 (1993)

[2] Blyth, C.R., Staudte, R.G. Estimating Statistical Hypotheses. *Statistics and Probability Letters*, 23: 45–52 (1995)

[3] Fisher, R.A. Statistical Methods and Scientific Inference (3rd ed.). Hafner Press, New York, 1973

[4] Fraser, D.A.S. The fiducial method and invariance. *Biometrika*, 48: 261–280 (1961)

[5] Kawakubo, K. The theory of transformation groups. Oxford University Press, Oxford, New York, Tokyo, 1991

[6] Morgenthaler, S., Tukey, J.W. eds. Configural polysampling: a route to practical robustness. John Wiley, New York, 1991

[7] Morgenthaler, S. Configural polysampling. *Encyclopedia of Statistical Sciences, Update Vol. 1*, (Kotz, S., ed.), John Wiley, New York, 1997

[8] Morgenthaler, S., Staudte, R G. Optimal weights of evidence with bounded influence. *Metrika*, 55: 91–97 (2002)

[9] von Neumann, J. Invariant measures. American Mathematical Society, Providence (Rhode Island), 1999