Optimal Distributions

Stephan Morgenthaler
Ecole polytechnique fédérale de Lausanne (EPFL)
MA-Ecublens, 1015 Lausanne, Switzerland
stephan.morgenthaler@epfl.ch

1 The Problem

Given measurements $y_1, \ldots, y_n$ of some unknown quantity $\theta \in \mathbb{R}$, we wish to compute a confidence interval $[L(y_1, \ldots, y_n), U(y_1, \ldots, y_n)]$ that has robust behavior and is equivariant, that is $L(r + sy_1, \ldots, r + sy_n) = r + sL(y_1, \ldots, y_n)$ for any $r \in \mathbb{R}$ and any $s \in \mathbb{R}_+$. Robustness means two things, namely

$$P\left(L(Y_1, \ldots, Y_n) < \theta < U(Y_1, \ldots, Y_n)\right) \approx 1 - \alpha \tag{1}$$

$$U(Y_1, \ldots, Y_n) - L(Y_1, \ldots, Y_n) \qquad \text{as small as possible}, \tag{2}$$

where we assume that for some $\theta \in \mathbb{R}$ and some $\sigma \in \mathbb{R}_+$ $(Y_1 - \theta)/\sigma, \ldots, (Y_n - \theta)/\sigma$ is an i.i.d. sample from a continuous distribution $F \in \mathcal{F}$. The class $\mathcal{F}$ represents the uncertainty about the distribution of the measurement errors. Assuming $\mathcal{F} = \{F\}$, that is a single and known error distribution, is standard procedure for many statisticians, whereas choosing $\mathcal{F} = \{G\colon d(F, G) \leq \epsilon\}$ for a fixed F and some "distance" $d(\cdot, \cdot)$ is a standard choice in the theory of robust statistics.

In the first case, it was shown by Fisher (1934) that the fiducial density

$$co_F(t) \propto \int_0^\infty s^{n-1} \prod_{i=1}^n f\left(s(y_i - t)\right) ds,$$

determines L and U, for example by taking the $\alpha/2$ and $1 - \alpha/2$ quantiles of this fiducial distribution.

In the second case, when $\mathcal{F}$ contains several distributions, the situation is much more complicated. Fisher's fiducial distribution allows us to compute confidence points for any fixed distribution, but if we change the distribution different confidence statements occur. It is, however, evident that no data analyst will be comfortable reporting an upper confidence point, if another equally reasonable underlying distribution would lead to a markedly larger upper confidence point. In such a situation one would conclude that the assumption about the underlying law was such that an overly optimistic inference resulted. This thinking takes the concern (1) seriously, whereas (2) is left aside, and it suggests an optimization problem of interest to mathematical statisticians, namely the search for the error distribution $F \in \mathcal{F}$ that leads to the widest confidence intervals or the largest upper confidence point.

2 Numerical Solutions

Typically, $\mathcal{F}$ is a function space and direct numerical approaches are not available. If we can approximate $\mathcal{F}$ by a finite-dimensional space, then a numerical search over the elements of the approximating space becomes possible. The symmetric *HR*-distributions (Morgenthaler and Tukey, 2000) are such a low-dimensional class of error distributions. They are defined via a two-dimensional family of transformations of a unit Gaussian random variable, namely

$$Y = Z \exp(h\, Z^2/(2 + r\, Z^2)),$$

where $Z \sim \mathcal{N}(0,1)$ and $r > 0$, $h > -2r$ are reals. The distribution of Y is symmetric around 0 and is either unimodal or bimodal. The tails of the *HR*-distributions can be heavier than Gaussian, including Pareto tails of arbitrary tail index, lighter than Gaussian, or like Gaussian.

Example. (The example is taken from Morgenthaler and Tukey (2000).) If we consider samples of size $n = 5$, we can always scale and shift them in such a way that the minimal data point is at minus 1, the maximal is at plus 1 and the other three are distributed in between. Fisher called such a standardized arrangement of data a **configuration**. The extreme configurations, when pushing the intermediate points to their limits together with the most- and least-favorable distribution, that is those with the smallest and largest 95%-quantile of the fiducial distribution are shown in Table 1. The corresponding densities are shown in Figure 1.

Table 1: *The first column indicates the data configuration, the next two columns contain the approximate parameter values (h, r) corresponding to the least-favorable distributions, that is, the one maximizing the 95% confidence point. The last two columns give the most optimistic assessment, that is the model with smallest 95% confidence point.*

Configuration	Least-informative		Most-informative	
	h	r	h	r
$(-1,-1,0,1,1)$	2.0	0.7	-0.66	0.66
$(-1,0,0,0,1)$	-2.5	2.5	$h \geq 1$	0
$(-1,-1,1,1,1)$	0.5	0.5	-3	3
$(-1,1,1,1,1)$	0	$r \geq 0$	-2	2
$(-1,0,1,1,1)$	0.4	0.7	-2.2	2.2
$(-1,0,0,1,1)$	1.4	0.5	-0.7	0.7

Assuming a bimodal underlying distribution for the configurations in rows 1, 3 and 6 of Table 1 leads to a small upper confidence points for the location parameter. In these cases, symmetric Gaussian-tailed laws – sometimes more peaked, sometimes less peaked – give rise to the least-informative models. In the case of row 2, a configuration with an accumulation at the center and two "outliers", it is the heavy-tailed laws that are most-informative. Such a distribution "explains" the presence of the outliers and "rejects" them when computing the confidence points. The least-informative law within the *HR*-family, on the other hand, is bimodal.

3 Analytical Solutions

Finding the most unfavorable distribution in $\mathcal{F}$ by functional analytic means is not straightforward either. One can show (Fernholz and Morgenthaler, 2000) that for large sample sizes, the optimization of the width of the 95% confidence interval is equivalent to

$$I_H(F) = \inf_{G \in \mathcal{F}} I_H(G)\,, \tag{3}$$

where $I_H(F) = \int -(f'/f)'(x)\, dH(x)$ is the **generalized Fisher information** of the model F with respect to the actual error distribution H. This can be derived under the condition that the data $Y_1 - \theta, \ldots, Y_n - \theta$ is i.i.d. with distribution H and can be proved by using a special kind of Gâteaux derivative. In Fernholz and Morgenthaler (2000), the solution to this new optimization problem (3)

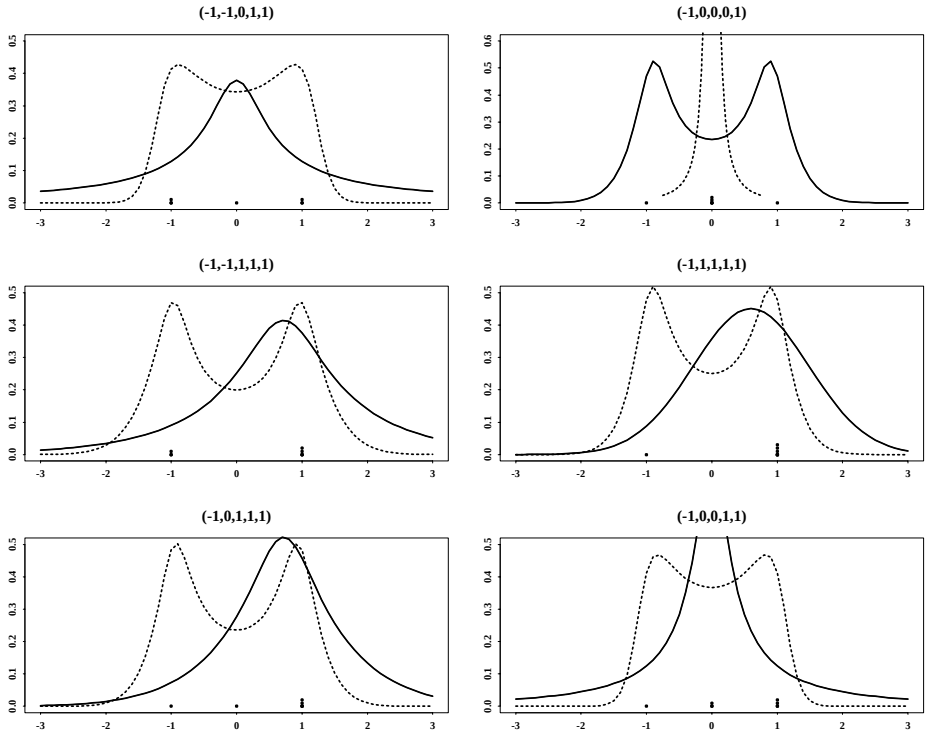

Figure 1: *For each of the six extreme data configurations the least-favorable HR-distribution is shown with its location and scale fixed at the maximum likelihood estimate. In each panel, the configuration is indicated by the points at the bottom. The solid lines show the least-favorable distribution, whereas the dotted lines correspond to the most-favorable one.*

for

$$\mathcal{F}_\epsilon = \{F \,|\, F = (1-\epsilon)H + \epsilon G \text{ such that } G' = g, g' \text{ and } g'' \text{ exist}\}$$

is discussed and the following result is proved.

Theorem. For a symmetric data distribution H with four times differentiable density h which satisfies $\{h'' < 0\} = (-a, a)$, $\{h'' > 0\} = (-\infty, -a) \cup (a, \infty)$, and h''/h increasing in (a, ∞), the asymptotically least informative model density in $\mathcal{F}_\epsilon$ for $0 < \epsilon < 1$ is

$$f_0(x) = \begin{cases} (1-\epsilon)\, h(x) & \text{if } |x| < b_\epsilon; \\ C_\epsilon h''(x) & \text{if } |x| \geq b_\epsilon; \end{cases}$$

where $C_\epsilon > 0$ and $b_\epsilon > a$ always exist and are such that $C_\epsilon h''(b_\epsilon) = (1-\epsilon)\, h(b_\epsilon)$ and $\int_{-\infty}^{\infty} f_0(x)\, dx = 1$.

Example. The second condition on the constants involved in the definition of f_0 in the Theorem is

$$2\, C_\epsilon\, h'(-b_\epsilon) + (1-\epsilon)(2\, H(b_\epsilon) - 1) = 1\,.$$

When $H = \Phi$, the unit Gaussian, the first condition is

$$b_\epsilon = (1 + (1-\epsilon)/C_\epsilon)^{1/2}\ .$$

The resulting locally least-favorable distributions are shown in Figure 2.

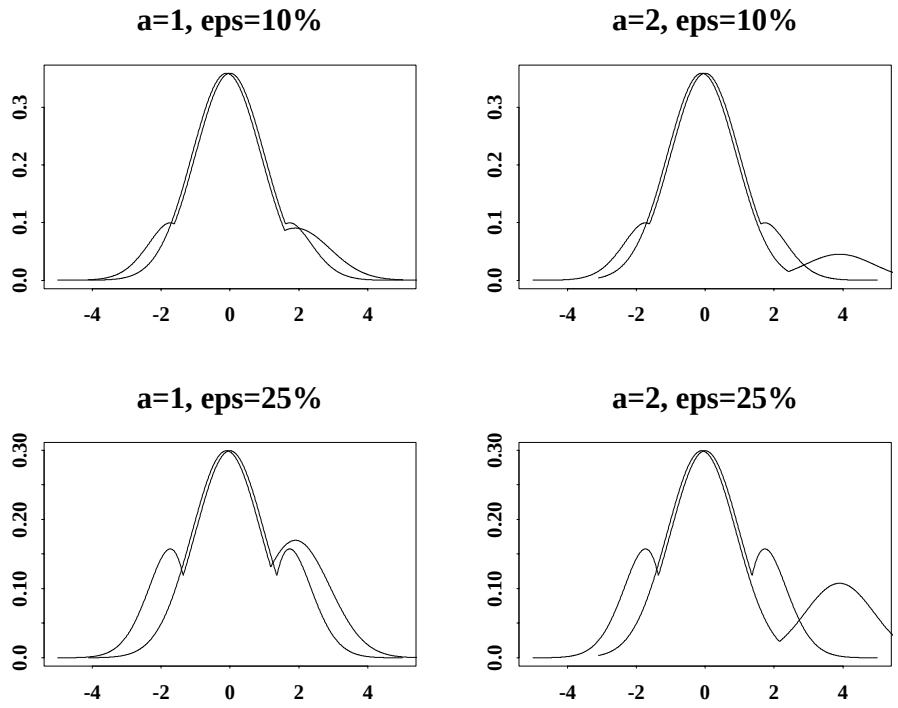

Figure 2: *The four panels show the asymptotically least-informative conditional model for normal data, $H = \Phi$, with $\epsilon = 10\%$ (top row) and $\epsilon = 25\%$ (bottom row). One of the pair of Huber's least informative distribution for interval estimation and tests (Huber, 1981, Chapter 10, Example 3.1) is also shown in each plot, the ones for an interval half width of $a = 1$ on the left and the ones for $a = 2$ on the right. In order to be able to distinguish the two densities, Huber's density is shifted slightly to the left, whereas ours is centered at zero.*

REFERENCES

Fernholz, L. and Morgenthaler, S. (2000). Least-informative distributions for robust confidence intervals. Technical report, Ecole polytechnique fédérale de Lausanne.

Fisher, R. A. F. (1934). Two new properties of maximum likelihood. *Proc. Royal Soc.*, A144:285–307.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Morgenthaler, S. and Tukey, J. W. (2000). Fitting quantiles: Doubling, HR, HQ, and HHH distributions. *Journal of Computational and Graphical Statistics*, 9:180–195.

RESUME

Bon nombre de problèmes statistiques font appel à une optimisation d'une fonctionnelle définie dans un ensemble de distributions. Ce papier présente des solutions possibles à un tel problème, à savoir l'identification de la distribution qui nous amène à la plus grand limite de confiance supérieure pour une quantité inconnue mesurée indépendamment n fois.