JOURNAL OF APPLIED STATISTICS
2019, VOL. 46, NO. 13, 2314–2337
https://doi.org/10.1080/02664763.2019.1586848

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Loss functions in restricted parameter spaces and their Bayesian applications

P. Mozgunov[a], T. Jaki[a] and M. Gasparini[b]

[a]Department of Mathematics and Statistics, Lancaster University, Lancaster, UK; [b]Dipartimento di Scienze Matematiche, Politecnico di Torino, Turin, Italy

ABSTRACT

Squared error loss remains the most commonly used loss function for constructing a Bayes estimator of the parameter of interest. However, it can lead to suboptimal solutions when a parameter is defined on a restricted space. It can also be an inappropriate choice in the context when an extreme overestimation and/or underestimation results in severe consequences and a more conservative estimator is preferred. We advocate a class of loss functions for parameters defined on restricted spaces which infinitely penalize boundary decisions like the squared error loss does on the real line. We also recall several properties of loss functions such as symmetry, convexity and invariance. We propose generalizations of the squared error loss function for parameters defined on the positive real line and on an interval. We provide explicit solutions for corresponding Bayes estimators and discuss multivariate extensions. Four well-known Bayesian estimation problems are used to demonstrate inferential benefits the novel Bayes estimators can provide in the context of restricted estimation.

ARTICLE HISTORY
Received 25 July 2018
Accepted 20 February 2019

KEYWORDS
Aitchison distance; Bayesian estimation; Scale parameter

1. Introduction

In many parameter estimation problems, the support of the parameter is either naturally restricted (e.g. probability, variance, exponential distribution parameter) or an investigator can restrict it based on previously obtained knowledge (e.g. treatment effects on children given the data for adults). The knowledge about restricted space can carry important information and improve estimation [19]. There are also application areas in which estimates on bounds of the restricted space are highly undesirable as they can lead to severe consequences. For instance, underestimating the potential of an event to have disastrous or life-threatening consequences may be worse than overestimating it [25]. An erroneously low estimated risk-level can lead to the absence of initiative to reduce it.

One of the ways to incorporate the information about a restricted space is to employ a Bayesian approach and to define a (uniform) prior distribution on the restricted space [10]. However, once a posterior is obtained, the squared error loss

$$L_q(\theta, d) = (\theta - d)^2 \tag{1}$$

CONTACT P. Mozgunov p.mozgunov@lancaster.ac.uk

Supplemental data for this article can be accessed here. https://doi.org/10.1080/02664763.2019.1586848

where d is a decision the statistician has to take in order to approximate an unknown estimand θ, called parameter, is often used to summarize a posterior distribution. The squared loss function (1) ignores the information about the restricted parameter space and is recognised to lead to suboptimal solutions [see e.g. 4,31, for alternative loss functions for a scale parameter]. Research on improving the Bayes estimator under squared loss function (a posterior mean) for a scale parameter has consequently attracted a great deal of attention [see e.g. 16, and references therein].

In addition, since loss function (1) does not penalize boundary values, it was found to be unacceptable in many application areas: see [25] for examples in reliability analysis, Karimnezhad et al. [13] in environmental sciences and [29] in drug development. To avoid boundary values of a scale parameter, [25] introduced the *precautionary loss function*

$$L_{sq}(\theta, d) = \frac{(d-\theta)^2}{d} \text{ where } \theta, d \in (0, +\infty) \tag{2}$$

which was used by many researchers [12,14].

The precautionary loss function covers the case of the scale parameter. There are, however, many applications in which the parameter of interest is restricted to an interval (a, b) and similar problems of severe consequences of boundary decisions can appear. We provide two motivating examples from the medical domain. Firstly, in the setting of outbreaks, the probability of response for a drug able to stop the outbreak should be high (say $> 90\%$). In this case, overestimation of the probability of response can lead to the approval of a drug which cannot stop the outbreak that can cost a lot of human lives. Secondly, in many paediatric trials, adult data responses can be used to define feasible values of responses (usually an interval) for children. At the same time, underestimation of the response effect for comparative treatments in paediatric clinical trials is highly undesirable as it might result in an underpowered and unethical study. In both settings, one can benefit from the application of specific loss function for parameters defined on an interval. We provide more details on the consequences in the later example in Section 6.

Despite the importance, the question of an appropriate loss function choice for a parameter θ defined on the interval (a, b) has been paid less attention in the statistical literature compared to a scale parameter. At the same time, its importance is acknowledged in many fields [see 2,22, for examples in compositional data analysis]. Specifically, Aitchison [2] proposed to use

$$L_{iB}(\theta, d) = \left(\text{logit}(d) - \text{logit}(\theta)\right)^2$$

as the measure of distance for $d, \theta \in (0, 1)$ where $\text{logit}(x) = \log \frac{x}{1-x}$ is the logit-transformation. This is the squared error loss after the logit-transformation of d and θ. While being intuitively clear, it is not convex and it has no explicit formula of the Bayes estimator, making its use challenging in applications. Furthermore, despite the variety of literature on families of loss functions for parameters restricted to an interval and on corresponding improved Bayes estimators (see e.g. [16,21]), they seem to be rarely applied in practice due to their complexity and to a lack of closed-form solutions. The choice of loss functions for parameters defined on the interval and on the positive real line is yet an under-represented area in the Bayesian literature and the usual mean still remains a common summary statistic.

The contribution of this work is twofold. Firstly, we provide a unified approach to define symmetry of a loss function when a parameter space is restricted to a particular open subset based on an appropriate definition of distance. We underline that our distances on corresponding parameter spaces share a common property – infinite penalization of the bounds which is also known as the *balance property* [19]. We also recall two other desirable properties of loss functions: convexity and invariance. Secondly, we propose several loss functions which are as simple as the squared loss function (1), have explicit solutions for the corresponding Bayes estimator and incorporate the information about the restricted parameter space in the corresponding Bayes estimator. In particular, we propose the scale invariant generalization of the the precautionary loss function for a scale parameter $\theta \in (0, +\infty)$ and the interval squared loss function

$$L_{iq}(\theta, d) = \frac{(d-\theta)^2}{(d-a)(b-d)}$$

for the parameter $\theta \in (a, b)$. We show that the Bayes estimator corresponding to the interval squared loss function includes the Bayes estimator of the squared loss function (1) and of the precautionary loss function as limiting cases. It is found that the interval squared and precautionary loss functions are both symmetric on the corresponding parameter spaces and can be useful in application areas where conservative estimates are preferred. We generalise the approach for the multivariate parameter space and demonstrate how Bayes estimators obtained using the proposed loss functions behave in four classic problems of Bayesian estimation compared to standard approaches.

The rest of the paper is organized as follows. A historical perspective for the scale parameter estimation and the case of symmetric loss function on the positive real line is given in Section 2. Section 3 introduces the novel loss function for an interval. The multivariate generalizations are given in Section 4. Four examples demonstrating novel loss functions and corresponding Bayes estimators are considered in Section 5. An application of a novel loss function to the problem of the sample size calculation in a clinical trial is given in Section 6.

2. Scale symmetry

2.1. A historical anecdote: galileo on scale symmetry

In the Spring of 1627, a *peculiar controversy*[1] arose in one of Florence intellectual circles, where *noble gentlemen* used to entertain *erudite talks*:

> Un cavallo, che vale veramente cento scudi, da uno è stimato mille scudi e da un altro dieci scudi: si domanda chi abbia di loro stimato meglio, e chi abbia fatto manco stravaganza nello stimare.

The problem translates into: ‘A horse, whose true worth is one hundred *scudi* [2], is estimated by someone to be one thousand *scudi* and by someone else to be ten *scudi*: the question is, who gave a better estimate, and who instead gave a more extravagant estimate?’. It is formulated in a letter from Andrea Gerini to Nozzolini, an *erudite priest*. Gerini wanted Nozzolini’s opinion on a sentence by Galilei [7], according to whom

> …li due stimatori abbiano egualmente esorbitato e commesse eguali stravaganze nello stimare l’uno mille e l’altro dieci quello che realmente val cento,

which translates to: 'The two estimators have been equally exorbitant and are responsible for an equal extravagance by estimating, one thousand the former and ten the latter, what is really worth one hundred'.

In the intense correspondence following the initial letters, Nozzolini argues that the estimates should be evaluated according to the *arithmetic proportion*, whereas Galileo insists that the correct method of judging is by *geometric proportion*. The crux of the problem is that the estimand is a positive quantity, for which the *geometric proportion* seems more appropriate, as wittingly argued by Galileo in another letter:

> Se uno stimasse alta dugento braccia una torre, che veramente fusse alta cento, con quale esorbitanza nel meno pareggerà il signor Nozzolini l'altra nel più ?

which translates as: 'If one were to overestimate a one-hundred arm high tower as two-hundred arm high, what underestimate would Nozzolini consider as equally deviating?'

2.2. Scale symmetry and scale invariance

Consider a toy example to illustrate Galileo's position in modern statistical terms. Two inferential procedures are based on two independent experiments:

(1) Estimate $\mu \in (-\infty, +\infty)$ given i.i.d. $X_i \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 is known;
(2) Estimate $\sigma \in (0, +\infty)$ given i.i.d. $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, μ is known.

Assume that the true parameters values are equal $\theta = \mu = \sigma$ and X_i and Y_k independent for all i,k. Using squared error loss (1), the decision $\mu = 0$ in the first experiment and $\sigma = 0$ in the second are equally penalized, while this should not be the case. The claim of $\sigma = 0$ implies that the Y's are degenerate random variables, an extremely strong statement which should be penalized similarly to the decision $\mu = +\infty$ or $\sigma = +\infty$. The squared error loss function imposes an infinite penalty to a boundary decision in the first experiment and does not in the second one. While the decision $\sigma = 0$ is usually prevented by a proper choice of the prior, the squared loss function does not imply that it should be avoided and associated with a severe penalty. In contrast, an appropriate loss function imposes such a penalty and can be also used to prevent boundary decisions. We define the properties of such loss function for a scale parameter in this section.

Let us start with the following definition for a parameter defined on the whole real line.

Definition 2.1: A loss function $L(\theta, d)$ is symmetric if, for every d_1, d_2 and $\theta \in \mathbb{R}^1$

$$(\theta - d_1)^2 = (d_2 - \theta)^2 \tag{3}$$

implies $L(\theta, d_1) = L(\theta, d_2)$.

The Definition 2.1 implies that two decisions defined on the real line should be equally penalized by a symmetric loss function $L(\cdot, \theta)$ if they stand on the same squared distance from θ. Note that for $d_1 < d_2$ Equation (3) can be rewritten as $\theta = (d_1 + d_2)/2$. It follows that if θ is the *arithmetic mean* of d_1 and d_2, then these decisions should be equally penalized. Clearly, the squared error loss of equation (1) is symmetric on the real line by definition.

Then, Galilei's claim of *eguali stravaganze* for a positive parameter θ can be expressed in modern terminology as the requirement of a *scale symmetric* loss function, as in the following definition.

Definition 2.2: A loss function $L(\theta, d)$ is scale symmetric if, for every d_1, d_2 and $\theta \in \mathbb{R}_+$

$$\frac{d_1}{\theta} = \frac{\theta}{d_2} \tag{4}$$

implies $L(\theta, d_1) = L(\theta, d_2)$.

Equation (4) can be rewritten as $\theta = \sqrt{d_1 d_2}$ or $\log(\theta) = (\log d_1 + \log d_2)/2$. In other words, if θ is the *geometric mean* of d_1 and d_2, then these decisions should be equally penalized by a scale symmetric loss function. As in Definition 2.1 of symmetry for the real line, two decisions are symmetric if the parameter θ is their appropriate mean – geometric in this case as opposed to arithmetic. This fact will be used for our proposal of the definition of symmetry on interval in Section 3.

The distance on the positive real line, $\mathbb{R}_+$, defined as [22]

$$\mathcal{D}_+(\theta, d) = (\log\theta - \log d)^2 \tag{5}$$

is known in Statistics as Brown's loss function [4]. Its motivation is to rescale the positive real line to the whole real line via the log transformation and to use the squared error loss function. Here, the logarithm function is a natural choice for a positive random variable. Note that $\mathcal{D}_+(\theta, d_1) = \mathcal{D}_+(\theta, d_2)$ implies either $d_1 = d_2$ or Equation (4). Therefore, we could also restate Definition 2.2 in terms of $\mathcal{D}_+(\cdot)$.

The Euclidean distance on the real line and $\mathcal{D}_+$ on the positive real line infinitely penalize boundary values on the corresponding parameter space. In case of $\theta \in \mathbb{R}$, the squared distance $L_q(\theta, d)$ takes an infinite value when $d = \pm\infty$. For similar reasons, we require that an appropriate loss function for a scale parameter should go to infinity as the decision approaches the natural boundaries of the parameter space, to reproduce the behaviour at $\pm\infty$ of the squared error loss function. A loss function with this property is also called *balanced* [19].

We recall one more property of loss functions for a parameter on the positive real line - the scale invariance.

Definition 2.3: Loss function $L(\theta, d)$ is scale invariant if for every $c > 0$ and every pair (θ, d),

$$L(\theta, d) = L(c\theta, cd).$$

Then, the following result can be obtained.

Lemma 2.4: *A loss function is scale invariant and scale symmetric if and only if it can be written as a scalar function g such that* $g(d/\theta) = g(\theta/d)$.

Proof: A loss function $L(\theta, d)$ is scale invariant if and only if it is ratio-based, i.e. if and only if there exists a scalar function $g(x), x > 0$ such that $L(\theta, d) = g(d/\theta)$. Scale invariance therefore implies $L(\theta, d) = g(d/\theta)$ for some g, whereas by Definition (2.2) scale symmetry implies $L(\theta, d) = L(\theta, \theta^2/d)$ and vice versa. ■

It follows that the squared loss function (1) is not scale symmetric, is not scale invariant and does not penalize all boundaries for a scale parameter.

2.3. Symmetric loss functions on the positive real line

In the modern statistical literature, the inadequacy of difference-based loss functions, like the squared error loss, for estimating certain positive quantities has often been recognized [11,31]. Several alternative loss functions have been proposed, the best-known being the normalized squared loss function proposed by Stein [31]

$$L_{nq}(\theta, d) = \left(\frac{d}{\theta} - 1\right)^2,$$

Stein's loss (or an entropy loss function)

$$L_S(\theta, d) = \frac{d}{\theta} - 1 - \log\left(\frac{d}{\theta}\right)$$

and Brown's loss function [4] itself, $\mathcal{D}_+(\theta, d)$. One can check that all of functions above are scale invariant, but only Brown's loss function is scale symmetric and infinitely penalizes the boundary decisions. Unfortunately, Brown's loss function is not *convex*, a feature of loss functions which is often required to represent risk aversion and for the sake of regularizing the associated minimization problems. Another unpleasant consequence of non-convexity is that the Bayes estimator associated to Brown's loss function is usually difficult to calculate. Below we propose simple alternative loss functions, which share the desirable properties of a loss function on the positive line and have explicit Bayes estimators.

We propose a family of loss functions defined for $k > 0$ as

$$L_k(\theta, d) = \left(\frac{d}{\theta}\right)^k + \left(\frac{\theta}{d}\right)^k - 2 \tag{6}$$

which are scale symmetric, scale invariant, convex, and which tend to infinity at the boundaries. Expression (6) is a function of the ratio $\frac{d}{\theta}$ to make it scale symmetric, and it satisfies Lemma 2.4 to make it scale invariant. The constant 2 is subtracted so the minimum value of the loss function $L_k = 0$ is attained at $d = \theta$. In this paper, we focus on the case $k = 1$

$$L_1(\theta, d) = \frac{(d - \theta)^2}{\theta d} \tag{7}$$

which can be considered as a modification of the squared error loss function. The numerator is again the squared distance, but the denominator guarantees the infinite penalization for $d = 0$. It is easy to see that the loss function (7) is a scale invariant version of the precautionary loss function (2).

2.4. Scale means (the minimizers) and scale variances

Within the Bayesian approach, θ is a random variable with a distribution which conveys the uncertainty the researcher has in a given state of information (whether prior, posterior, elicited, objective and so on). In such a scenario, a point summary of the distribution of θ minimizing the risk (i.e. the expected loss) associated with a given loss function is often required. Such a minimizer of an expected d is usually called a Bayes estimator. When a scale symmetric loss function is used, we propose to call such minimizers *scale means.* In case of convex loss functions, such as the novel ones listed in the previous section, minimization can be performed explicitly, as in Theorem 2.5.

Theorem 2.5: *Let θ be a positive random variable with a posterior density function f and such that $\mathbb{E}(\theta^k) < \infty$ and $\mathbb{E}(\theta^{-k}) < \infty$, where $\mathbb{E}$ denotes the posterior mean with respect to f and $k > 0$. Then,*

(a) *Expectation of the loss function $L_k(\theta, d)$ (6) with respect to f is minimized by the Bayes estimator (scale mean)*

$$\hat{d}_k = \left(\frac{\mathbb{E}(\theta^k)}{\mathbb{E}(\theta^{-k})}\right)^{1/2k}. \tag{8}$$

(b) *Expectation of the precautionary loss function L_{sq} (2) is minimized by the Bayes estimator (scale mean)*

$$\hat{d}_{sq} = \sqrt{\mathbb{E}(\theta^2)}, \tag{9}$$

for which the following bound holds: $\hat{d}_{sq} \geq \mathbb{E}(\theta)$.

Proof: **(a)** The expectation of the loss function (6) with respect to the posterior density function f takes the form

$$\mathbb{E}(L_k(\theta, d)) = \mathbb{E}\left(\frac{d}{\theta}\right)^k + \mathbb{E}\left(\frac{\theta}{d}\right)^k - 2 = d^k\mathbb{E}\left(\theta^{-k}\right) + d^{k-1}\mathbb{E}\left(\theta^k\right) - 2.$$

Then, the decision d minimizing the expected loss function is found solving

$$\frac{\partial\mathbb{E}(L_k(\theta, d))}{\partial d} = kd^{k-1}\mathbb{E}\left(\theta^{-k}\right) - kd^{-k-1}\mathbb{E}\left(\theta^k\right) = 0$$

This results in $\hat{d}_k = (\frac{\mathbb{E}(\theta^k)}{\mathbb{E}(\theta^{-k})})^{\frac{1}{2k}}$, and in the special case of $k = 1$, $\hat{d}_1 = \sqrt{\mathbb{E}(\theta)/\mathbb{E}(\theta^{-1})}$.

(b) Similarly to the previous point, the expectation of the precautionary loss function (2) with respect to the posterior density function f taken the form

$$\mathbb{E}(L_{sq}(\theta, d)) = \mathbb{E}\left(\frac{(d-\theta)^2}{d}\right) = \frac{d^2 - 2d\mathbb{E}(\theta) + \mathbb{E}(\theta^2)}{d}.$$

Then, the decision d minimizing the expected loss function is found by $\frac{\partial\mathbb{E}(L_{sq}(\theta,d))}{\partial d} = 0$. This results in $\hat{d}_{sq} = \sqrt{\mathbb{E}(\theta^2)}$. Using Jensen inequality for θ^2 one can obtain $\mathbb{E}(\theta^2) \geq$

$\mathbb{E}^2(\theta)$. Applying the squared root to both sides of the inequality the result immediately follows. ■

In a more fundamental Bayesian approach, a Bayes estimator is regarded only as a convenient summary of the posterior, and a loss function as a way to prescribe what kind of summary is appropriate. Typically, a posterior expectation is used as the Bayes estimator, implying that a squared loss function is being used. A second step is usually taken to accompany the Bayes estimator with a measure of uncertainty of the posterior. If a posterior mean is used, a posterior variance is usually presented. However, if a scale symmetric loss function is considered to be a reasonable criterion for choosing an estimator, i.e. a number which minimizes a posterior expected loss, then it is also reasonable to present the achieved minimum of the posterior expected loss as a second summary of the posterior. For the given loss functions (6) and (2), particularly simple expected posterior losses can be obtained. In particular, for the loss function (6) such *scale variance* of order k of the random variable θ in Theorem 2.5(a) can be written $\hat{\tau}_k(\theta) := 2\sqrt{\mathbb{E}(\theta^k)\mathbb{E}(\theta^{-k})} - 2$, whereas the scale variance for the precautionary loss function (2) in Theorem 2.5(b) is $\hat{\tau}(\theta) = 2(\sqrt{\mathbb{E}(\theta^2)} - \mathbb{E}(\theta))$.

3. Interval symmetry

The approach used above for a positive parameter can be generalized to the parameter defined on the interval (a, b). The issue of a restricted parameter space is not usually discussed in the choice of the loss function and corresponding Bayes estimator: bounds are taken into account through the prior specification only, then the squared loss function and posterior mean (the corresponding Bayes estimator) are used [21]. Such solutions can be suboptimal if boundary decisions are to be avoided. Below, we define the property of the symmetry on an interval and show that the novel definition generalizes the cases of parameters on the whole real line and on the positive real line. We provide the loss function with desirable properties which is, again, a generalization of the squared loss function and the precautionary loss function.

3.1. Symmetric loss functions on intervals

Let us consider an inferential problem for which the parameter of interest lies in a particular interval (a, b). Define the following transformation

$$\operatorname{logit}_{(a,b)}(x) = \log\frac{x-a}{b-x} \tag{10}$$

where $a < x < b$. Notice that, for $a = 0$ and $b = 1$, transformation (10) reduces to the common logit transformation widely used in Statistics and was used by Aitchison [2] to justify the definition of distance on the unit interval. Following the same lines of reasoning as in Section 2.2, we use this transformation to introduce the definition of *symmetric on the interval* (a, b) loss function (or simply *interval symmetric* loss function), and demonstrate why it is a convenient choice.

Definition 3.1: A loss function $L(\theta, d)$ is symmetric on the interval (a, b) if, for every choice of $d_1, d_2 \in (a, b)$ and $\theta \in (a, b)$

$$\operatorname{logit}_{(a,b)}(\theta) = \frac{\operatorname{logit}_{(a,b)}(d_1) + \operatorname{logit}_{(a,b)}(d_2)}{2}. \tag{11}$$

implies $L(\theta, d_1) = L(\theta, d_2)$.

In other words, two decisions d_1 and d_2 should be penalized equally if the mean of their logit transformation is equal to the logit transformation of θ. Lemma 3.2 justifies the use of the logit transformation (10).

Lemma 3.2: *Definition 3.1 is equivalent to Definition 2.1 when $a \to -\infty$ and $b \to +\infty$ and equivalent to Definition 2.2 when $a \to 0$ and $b \to +\infty$.*

Proof: Condition (11) for $a < d_1 < d_2 < b$ can be rewritten

$$\theta = f(a, b, d_1, d_2) \equiv \frac{ab - d_1 d_2 + \sqrt{(d_1 - a)(b - d_1)(d_2 - a)(b - d_2)}}{a + b - d_1 - d_2}.$$

Obviously, θ is a symmetric function of d_1 and d_2. Considering two limits

$$\lim_{a \to -\infty,\, b \to +\infty} f(a, b, d_1, d_2) = \frac{d_1 + d_2}{2}, \quad \lim_{a \to 0,\, b \to +\infty} f(a, b, d_1, d_2) = \sqrt{d_1 d_2},$$

it can be easily seen that the definitions are equivalent. ■

It follows from Lemma 3.2 that Definition 3.1 is a convenient generalization of the definition of symmetry and of scale symmetry.

3.2. An interval symmetric loss function and a Bayes estimator

As in the case of a positive parameter and scale symmetric loss functions, discussed in Section 2, the approach of [4] of specifying a squared loss function after rescaling the interval (a, b) to the real line via, for example, the logit transformation (10) provides the loss function

$$L_{iB}(\theta, d) = \left(\operatorname{logit}_{(a,b)} d - \operatorname{logit}_{(a,b)} \theta\right)^2. \tag{12}$$

On the unit interval, this loss function is equivalent to so-called Aitchison distance proposed by Aitchison [2] for parameters defined on a simplex. However, the loss function (12) is not convex and its minimization problem does not have an explicit solution. As an alternative, we propose the following loss function

$$L_{iq}(\theta, d) = \frac{(d - \theta)^2}{(d - a)(b - d)}. \tag{13}$$

which is interval symmetric and tends to infinity when the decision d tends to bounds a and b.

Note that loss function (13) for $a = 0$ and $b = 1$ looks similar to the well-known loss function $(d - \theta)^2/(\theta(1 - \theta))$ which, however, does not penalize boundary decisions.

The Bayes estimator corresponding to L_{iq} is given in Theorem 3.3.

Theorem 3.3: *Let $\theta \in (a, b)$ be a random variable with a posterior density function f and $\mathbb{E}(\theta^2) < \infty$ where $\mathbb{E}(\cdot)$ denotes the expectation with respect to f. Then,*

(a) *the expectation of the interval symmetric loss function L_{iq} (13) with respect to f is minimized by the Bayes estimator*

$$\hat{d}_{iq} = \frac{ab - \mathbb{E}(\theta^2) + \sqrt{(\mathbb{E}(\theta^2) - ab)^2 - (a + b - 2\mathbb{E}(\theta))(2ab\mathbb{E}(\theta) - (a + b)\mathbb{E}(\theta^2))}}{a + b - 2\mathbb{E}(\theta)} \tag{14}$$

(b) *In the limiting case $a \to -\infty$ and $b \to +\infty$ estimator* (14) *minimizes the expectation of squared loss function (1), and in the limiting case $a \to 0$ and $b \to +\infty$ estimator* (14) *minimizes the expectation of precautionary loss function (2).*

Proof: **(a)** The equality is proved by differentiating in d the expected losses L_{iq} (13).

(b) Denote the estimator (14) by $d \equiv g(a, b, \theta)$; then, taking the limits

$$\lim_{a \to -\infty,\ b \to +\infty} g(a, b, \theta) = \mathbb{E}(\theta), \quad \lim_{a \to 0,\ b \to +\infty} g(a, b, \theta) = \sqrt{\mathbb{E}(\theta^2)},$$

it is easy to see that the obtained estimators are equivalent to the minimizers of the squared loss function (1) and of the precautionary loss function (13), respectively. Note that $\hat{d}_{iq} \to \frac{a+b}{2}$ as $\mathbb{E}(\theta) \to \frac{a+b}{2}$. ■

It follows from Theorem 3.3 that the Bayes estimator $\hat{d}_{iq}$ includes the Bayes estimator under squared loss function (1) and precautionary loss function (2) as special cases.

4. Multivariate generalizations

The definition of symmetry can be generalized to the case of a parameter belonging to a subset of $\mathbb{R}^m$ by applying the same ideas to selected shapes of the parameter space as in the following definition.

Definition 4.1: Let $\boldsymbol{\theta} = (\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(m)})^{\mathrm{T}}$ be a parameter lying in one of the parameter spaces $\Theta \subset \mathbb{R}^m$ listed below. Let $\boldsymbol{d_i} = (d_i^{(1)}, d_i^{(2)}, \dots, d_i^{(m)})^{\mathrm{T}}$, $i = 1, 2$ be two vectors of decisions defined on the same parameter space. A loss function $L(\boldsymbol{\theta}, \mathbf{d})$ is a multivariate $\Theta-$symmetric if the equality

$$L(\boldsymbol{\theta}, \mathbf{d_1}) = L(\boldsymbol{\theta}, \mathbf{d_2})$$

is implied by each triple $\boldsymbol{\theta}, \mathbf{d_1}, \mathbf{d_2} \in \Theta$ satisfying the following respective definitions of distances:

(a) when $\Theta = \mathbb{R}^m$ (symmetry on $\mathbb{R}^m$ itself):

$$\sqrt{\sum_{j=1}^{m}\left(d_1^{(j)} - \theta^{(j)}\right)^2} = \sqrt{\sum_{j=1}^{m}\left(d_2^{(j)} - \theta^{(j)}\right)^2};$$

(b) when $\Theta = \mathbb{R}_+^m = \{\theta : \theta^{(i)} > 0, i = 1, \ldots, m\}$ (scale symmetry on $\mathbb{R}_+^m$):

$$\sqrt{\sum_{j=1}^{m}\log^2\left(\frac{d_1^{(j)}}{\theta^{(j)}}\right)} = \sqrt{\sum_{j=1}^{m}\log^2\left(\frac{d_2^{(j)}}{\theta_j}\right)};$$

(c) when $\Theta = \{\theta : (a_1 < \theta^{(1)} < b_1), \ldots, (a_m < \theta^{(m)} < b_m)\}$ (symmetry on an $\mathbb{R}^m$-rectangle):

$$\sqrt{\sum_{j=1}^{m}\left(\text{logit}_{(a_j,b_j)}d_1^{(j)} - \text{logit}_{(a_j,b_j)}\theta^{(j)}\right)^2} = \sqrt{\sum_{j=1}^{m}\left(\text{logit}_{(a_j,b_j)}d_2^{(j)} - \text{logit}_{(a_j,b_j)}\theta^{(j)}\right)^2};$$

(d) when $\Theta = \{\theta : \theta^{(1)} > 0,\ \theta^{(2)} > 0,\ \ldots, \theta^{(m)} > 0; \sum_{i=1}^{m}\theta^{(i)} = 1\}$ (symmetry on the unit simplex):

$$\sqrt{\frac{1}{m}\sum_{i<j}\left(\log\frac{d_1^{(i)}}{d_1^{(j)}} - \log\frac{\theta^{(i)}}{\theta^{(j)}}\right)^2} = \sqrt{\frac{1}{m}\sum_{i<j}\left(\log\frac{d_2^{(i)}}{d_2^{(j)}} - \log\frac{\theta^{(i)}}{\theta^{(j)}}\right)^2}.$$

The definition of the symmetric loss function in each case employs a distance corresponding to the particular restricted space. While the distances in (a)–(c) are natural extensions of the previously used, the definition in (d) is less straightforward. Definition 4.1(d) uses the Aitchison distance proposed by Aitchison [2] and employed in compositional data analysis. Regarding properties of the proposed definition, Lemma 4.2, similar to Lemma 3.2, holds.

Lemma 4.2: *Let $\boldsymbol{\theta} = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)})^{\mathrm{T}}$ be a vector of parameter of interest such that $\theta^{(1)} \in (a_1, b_1)$, $\theta^{(2)} \in (a_2, b_2), \ldots, \theta^{(m)} \in (a_m, b_m)$ and $\boldsymbol{d_i} = (d_i^{(1)}, d_i^{(2)}, \ldots, d_i^{(m)})^{\mathrm{T}}$ be a vector of corresponding decisions lying in corresponding intervals. Definition 4.1(c) is equivalent to Definition 4.1(a) when $a_i \to -\infty$ and $b_i \to \infty$ for all $i = 1, \ldots, m$ and to Definition 4.1(b) when $a_i = 0$ and $b_i \to \infty$ for all $i = 1, \ldots, m$.*

Following [4], all distances in Definition (4.1) could be taken as corresponding symmetric loss functions. For example, in case *(b)*, one could define

$$\mathcal{D}_+^{(m)}(\boldsymbol{\theta}, \boldsymbol{d}) = \sum_{j=1}^{m}\log^2\left(\frac{d^{(j)}}{\theta^{(j)}}\right). \tag{15}$$

At the same time, some convex alternatives could be considered when leading to simple solutions of minimization problems. However, the search of the symmetric multivariate

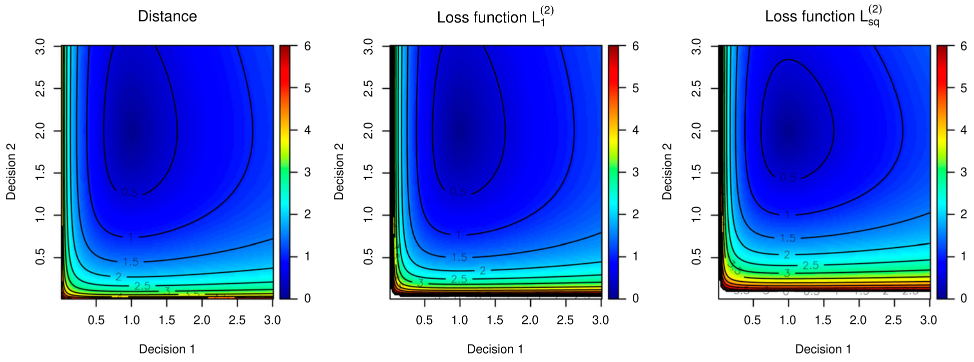

Figure 1. Contour plots of loss functions $\mathcal{D}_+^{(2)}, L_1^{(2)}, L_{sq}^{(2)}$ for the case $m = 2$ and $\boldsymbol{\theta} = (1, 2)^{\mathrm{T}}$.

generalization of our proposed loss functions L_k, L_{sq} and L_{iq} seems to be a non-trivial one. We propose the following loss functions for parameters with non-negative components.

Proposition 4.3: *Let $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(m)})^{\mathrm{T}} \in \mathbb{R}_+^m$ and $\boldsymbol{d} = (d^{(1)}, \ldots, d^{(m)})^{\mathrm{T}} \in \mathbb{R}_+^m$. The loss functions*

$$L_k^{(m)}(\boldsymbol{\theta}, \boldsymbol{d}) = \sum_{j=1}^{m} \left(\left(\frac{d^{(j)}}{\theta^{(j)}} \right)^k - \left(\frac{\theta^{(j)}}{d^{(j)}} \right)^k \right) - 2m \tag{16}$$

$$L_{sq}^{(m)}(\boldsymbol{\theta}, \boldsymbol{d}) = \sum_{j=1}^{m} \frac{(d^{(j)} - \theta^{(j)})^2}{d^{(j)}} \tag{17}$$

are additive multivariate generalizations of the loss function L_1 given in (6) and (2), respectively, which infinitely penalize each boundary decision $d^{(j)} = 0$ and $d^{(j)} = \infty, j = 1, \ldots, m$.

Clearly, loss functions $L_1^{(m)}$ and $L_{sq}^{(m)}$ penalize the boundaries as desired and some good performance of the corresponding estimators can be expected. However, the property of symmetry is not satisfied. One can find two decisions $\tilde{\boldsymbol{d}}_1$ and $\tilde{\boldsymbol{d}}_2$ for which $\mathcal{D}_+^{(m)}(\boldsymbol{\theta}, \tilde{\boldsymbol{d}}_1) = \mathcal{D}_+^{(m)}(\boldsymbol{\theta}, \tilde{\boldsymbol{d}}_2)$, but $L_1^{(m)}(\boldsymbol{\theta}, \tilde{\boldsymbol{d}}_1) \neq L_1^{(m)}(\boldsymbol{\theta}, \tilde{\boldsymbol{d}}_2)$. Even if the whole loss functions are not symmetric, they are 'component-wise' symmetric as shown above. The comparison of loss functions $L_1^{(2)}, L_{sq}^{(2)}$ and $\mathcal{D}_+^{(2)}$ for different values of decision $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ and fixed $\boldsymbol{\theta} = (1, 2)^{\mathrm{T}}$ is given in Figure 1.

The proposed loss functions perform similar to the distance $\mathcal{D}_+^{(2)}$, but have some more favourable properties, like convexity.

5. Examples

Loss functions penalising boundary decisions were found to be more beneficial in many applications areas. For instance, Saint-Hilary et al. [29] have shown that a loss function similar to the loss function (6) penalising the decision $d = 0$ can lead to a more reliable benefit-risk analysis of novel drugs. Similarly, Mozgunov and Jaki [24] have found that the loss function (13) penalising decisions $d=0,1$ incorporated into a model-based dose-finding design can improve a selection of the optimal doses without exposing patients to excessively toxic doses. Below, we investigate the performance of the proposed loss function and corresponding Bayes estimators in more general settings. We consider four classic examples of estimation to demonstrate the essential differences of estimators. We focus on the small sample size ($n=15$) to emphasize the difference in estimators. The results for moderate ($n = 100$) and large ($n = 1000$) sample sizes are given in Supplementary Materials. Software in the form of R code [27] is also provided in Supplementary Materials.

For all examples, the frequentist operating characteristic, Mean Squared Error (MSE), is chosen to compare the different estimators on common grounds. As advocated by Berger et al. [3,6], studying the frequentist properties of Bayes estimators is a way to study the properties independently of the prior distribution and to consider Bayesian point estimate simply as a function of the data. Furthermore, as we intend to compare several Bayes estimators, which minimize different loss functions, the frequentist characteristics are chosen to assess the performance of these estimators on an equal basis. Note that this choice is not favourable to our new proposals, since the MSE is derived from squared error loss.

5.1. Estimation of a probability

An important example of a parameter defined on the finite interval $[0, 1]$ is a probability. In the presence of a binary random sample with an unknown probability of success a uniform distribution, i.e. a Beta prior distribution $\mathcal{B}(1, 1)$ is often assumed, a proposal which dates back to Laplace. Having observed x successes out of n trials, the posterior distribution is a conjugate Beta distribution $\mathcal{B}(x + 1, n - x + 1)$. The estimator corresponding to the squared error loss function (posterior mean) has the form $\hat{p}_q = \frac{x+1}{n+2}$. Another widely used estimator is the so-called 'add two successes and two failures' Agresti-Coull estimator [1] $\hat{p}_{AC} = \frac{x+2}{n+4}$. Below we compare these approaches to the newly proposed estimator (14).

The symmetric optimal Bayes estimator (14) in the case $a=0$ and $b=1$ can be written

$$\hat{d}_{iq} = \frac{\mathbb{E}(\theta^2) - \sqrt{\mathbb{E}(\theta^2)(1 - 2\mathbb{E}(\theta) + \mathbb{E}(\theta^2))}}{2\mathbb{E}(\theta) - 1}$$

where θ is a probability of success lying in the interval (0,1), over which a posterior distribution is given. It is assumed that the extremes of the interval are not possible values for the parameter. The first and second moments of a Beta distribution can be computed explicitly and plugged in formula (14) to obtain the following interval symmetric optimal Bayes estimator

$$\hat{p}_{iq} = \left(1 + \sqrt{\frac{(n - x + 1)(n - x + 2)}{(x + 1)(x + 2)}}\right)^{-1}. \tag{18}$$

Simulated trials with sample size $n=15$ are considered. On a grid of values $\theta \in (0.01, 0.99)$, $N = 10^9$ trials were simulated. This means that for each value of θ on the grid, we simulate 10^9 trials with the total sample size $n=15$. This results in 10^9 point estimates found for

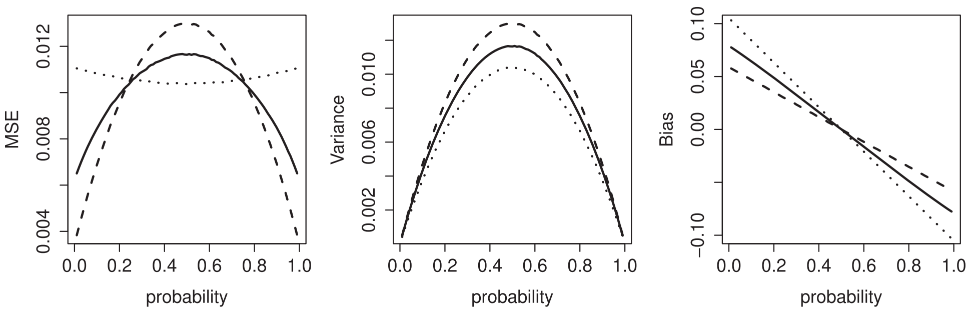

Figure 2. MSE, variance and bias for the restricted symmetric squared error loss function estimator $\hat{p}_{iq}$ (solid), the squared error loss function estimator $\hat{p}_q$ (dashed) and the Agresti-Coull estimator $\hat{p}_{AC}$ (dotted). Results are based on $n = 15$ observations and 10^9 simulations.

each method. Then, the MSE is computed as

$$MSE_k \equiv \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_k^{(i)} - \theta)^2, \tag{19}$$

where $\hat{p}_k^{(i)}$ is a corresponding value in the i^{th} simulation and $k = q, iq, AC$ corresponds to an estimation method. The results are given in Figure 2.

The proposed estimator $\hat{p}_{iq}$ outperforms (in terms of the MSE) the Bayes estimator obtained using the squared error loss function $\hat{p}_q$ in the interval $\theta \in (0.2, 0.8)$. The cost of this advantage is the worse performance on the intervals close to the bounds as the proposed form of the loss function penalizes boundary decisions and by that drives the final estimate away from them. However, the proposed estimator outperforms the Agresti-Coull estimator $\hat{p}_{AC}$ at the same intervals $\theta \in (0, 0.2)$ and $\theta \in (0.8, 1)$. Thus, the proposed estimator might be considered as a trade-off between currently used estimators $\hat{p}_q$ and $\hat{p}_{AC}$, that outperforms $\hat{p}_{AC}$ on bounds and $\hat{p}_q$ away from bounds.

In addition to the MSE, the associated confidence intervals and coverage probabilities are extensively studied in the literature [5]. In particular, coverage probabilities were shown to have an erratic behaviour and often to go below their nominal level. Corrections were proposed by Agresti and Coull [1]. Confidence intervals can also be constructed around our newly proposed point estimator $\hat{p}_{iq}$. The following confidence intervals are compared via simulated coverage probabilities in Figure 3:

(1) Normal approximation confidence interval centred around $\hat{p}_k, k = q, iq, AC$ as suggested by Brown et al. [5]

$$CI_N^{(k)} = \hat{p}_k \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_k(1-\hat{p}_k)}{n}},$$

where $1 - \alpha$ is the confidence level.

(2) Wilson confidence interval centred around $\hat{p}_{AC}$ [33]

$$CI_W^{(AC)} = \frac{x+2}{n+4} \pm 2\frac{\sqrt{n}}{n+2}\sqrt{\frac{x(n-x)}{n^2} + \frac{1}{n}}.$$

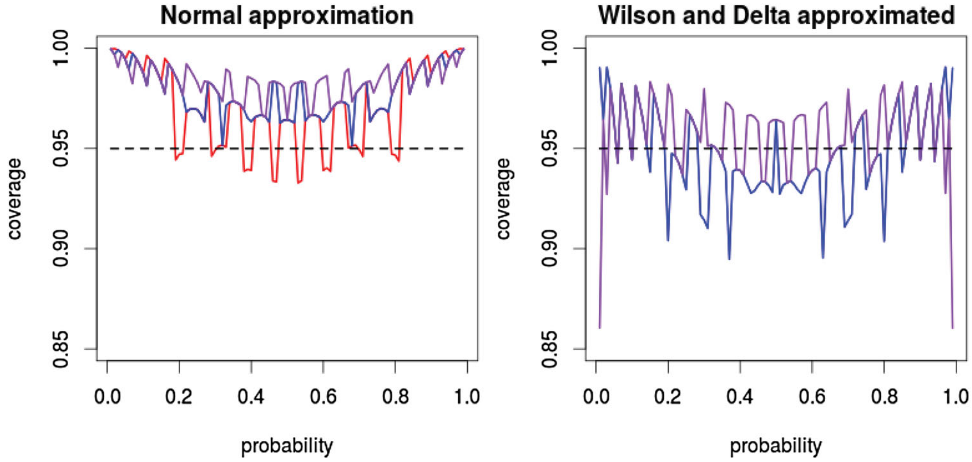

Figure 3. Left panel: Coverage probabilities of $CI_N^{(iq)}$ (middle line), $CI_N^{(q)}$ (lower line) and $CI_N^{(AC)}$ (upper line) using [1], the normal approximation interval. Right panel: Coverage probabilities of $CI_D^{(iq)}$ (lower line) and $CI_W^{(AC)}$ (upper line) using [2], the Wilson, and [3], the delta-method, confidence intervals respectively. Results are based on $n = 15$ observations and 10^9 simulations.

(3) Approximate confidence interval using the delta-method centred around the newly proposed $\hat{p}_{iq}$

$$CI_D^{(iq)} = \hat{p}_{iq} \pm 2\sqrt{\hat{V}_{iq}},\ \hat{V}_{iq} = \left(\frac{\partial f(x)}{\partial x}\right)^2 \Big|_{x=n\hat{p}_{iq}} n\hat{p}_{iq}(1-\hat{p}_{iq}),\ f(x) = \hat{p}_{iq}.$$

Using the normal approximation confidence interval, the coverage probability of $CI_N^{(q)}$ goes below the nominal value for several values of θ. The coverage probabilities of $CI_N^{(iq)}$ and $CI_N^{(AC)}$ also fluctuate but do not go below 0.95 for $N = 15$, which is a desirable property. While one can find combination of N and θ for which coverage probabilities $CI_N^{(iq)}$ and $CI_N^{(AC)}$ might be below 0.95 [5], it would be generally true that their coverage probabilities are greater than of $CI_N^{(q)}$ for larger intervals of θ and are more robust. In addition, a comparison between the normal approximation method (left panel of Figure 3) and the Wilson and delta method intervals (right panel of Figure 3) supports the suggestion by Brown et al. [5] that the normal approximation method gives a portmanteau way to construct simple confidence intervals with - on average - better coverage probabilities than more complicated methods.

5.2. Restricted estimation of a normal distribution mean

In the following example, it is demonstrated what benefits the proposed form of loss function (13) can provide in a Bayesian framework in the presence of the additional information that the true parameter lies in an interval (a, b).

The problem of restricted mean estimation has been known for a long time and has been extensively studied in the literature (see e.g. [17], and references therein). The previously proposed estimators were constructed using the squared error loss function and compared by the Bayesian risk. For an extensive overview of the problem, we refer the reader to [21] and for some recent generalizations to [20]. Interestingly, despite the variety of the literature on the problem and the fact that the Bayes estimator with respect to the uniform prior distribution on $(-a, a)$ outperforms uniformly the 'unrestricted' Bayes estimator under squared error loss function [10], the sample mean estimator is still widely used in practice. For this reason, we propose our simple alternative and compare it to the most commonly used estimators.

Consider the problem of estimating the mean μ of a normal sample of i.i.d. $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$ where σ^2 is known. Assume it is known that the parameter μ belongs to the interval $(-a, +a)$ where $a > 0$. A possible example is the estimation of the treatment effect in paediatric studies of a clinical treatment that was already tested on adults. An investigator might be sure that the same dosage of the drug as given to adults would cause a greater level of toxicity in children. It is of interest how much one can gain by incorporating this information in the estimator (14) compared to currently used approaches that use the squared error loss.

As stated above, the common way to incorporate this information in a Bayesian framework is restricting the prior distribution for the parameter of interest μ to the interval $(-a, +a)$ and then using the squared error loss to obtain a point estimate (as a summary of a posterior distribution). However, the information about the restricted space is used only on the prior and ignored when choosing summary statistics, while the proposed form of loss function (13) allows to incorporate it again. Moreover, in practice the prior information is often ignored and the sample mean estimator (corresponding to Jeffrey's prior) is used. Therefore, a comparison of the proposed loss function and the currently utilized approaches (with and without the incorporation the prior information) is of interest.

Consider samples of size $n = 15$ and alternative Bayes estimators as follows:

- Bayes estimator under Jeffrey's prior $g(\mu) \propto k\sqrt{\frac{n}{\sigma^2}}$ for μ and the squared error loss function. Denoted by J;
- Bayes estimator under the uniform prior $\mathbb{U} \sim (-a, +a)$ for μ and squared error loss function. Denoted by U_1.
- Bayes estimator (14) under the uniform prior $\mathbb{U}(-a, +a)$ for μ and the interval symmetric squared error loss function. Denoted by U_2.

Alternatively to U_2, the Bayes estimator U_2' defined by (14), but with a wider interval $(-1.25a, +1.25a)$ is also applied to investigate how less severe penalization of the bounds influences the estimation. This wider interval could be considered as a conservative way to incorporate information about the location of the parameter.

Two cases $a = 2$ and $a = 4$ are considered. The parameter $\sigma^2 = 4$ is assumed to be known in both cases. As before, the *MSE* is used for a fair comparison of methods. The results for 10^6 replications for each value of μ are given in Figure 4.

Incorporating the interval information in the Bayes estimator allows for improvement if the true value of the parameter μ does not lie close to the bound. In the case $a = 2$ estimator U_2 outperforms U_1 on the interval $\mu \in (-1, +1)$ and in the case $a = 4$ on the interval $\mu \in (-3, +3)$. The same holds for estimator U_2', however, its MSE never falls below the level of

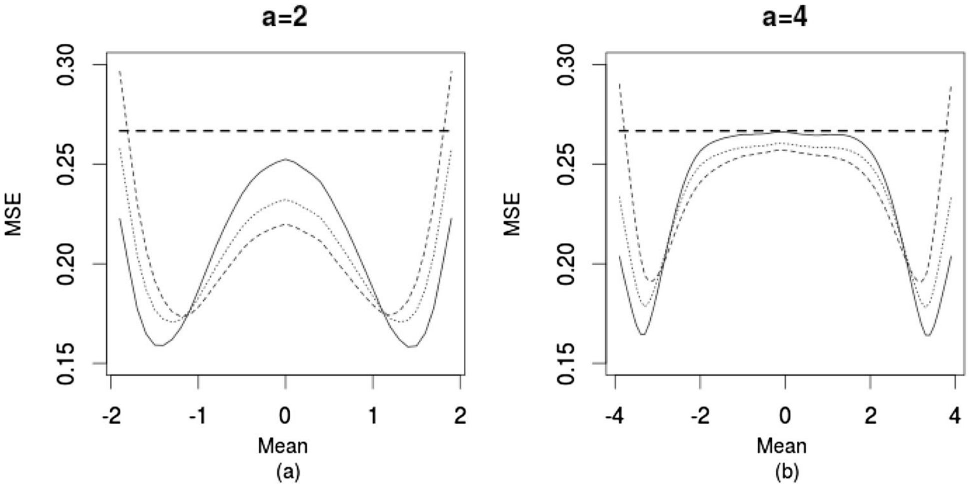

Figure 4. MSE corresponding to different values of the restricted mean parameter μ with (a) $a = 2$ and (b) $a = 4$ and the Bayes estimator U_1 (solid), U_2 (dashed) and U_2' (dotted) and simple mean estimator J (solid dashed). Results are based on 10^6 replications.

the MSE of estimator J. Clearly, the wider interval improves estimation on the bounds, at the cost of the higher MSE in the middle. Note that the wider interval $a = 4$ corresponds to weaker additional information and to a smaller benefit in the MSE comparing U_1, U_2 and U_2' against J. Overall, the Bayes estimator corresponding to the interval symmetric loss function avoids the boundary decisions, improves the estimation if the parameters lies away from bounds and can be recommended for the application if boundary decisions lead to severe consequences.

5.3. Bayesian estimation of the parameters of a gamma distribution

An important example of multidimensional restricted parameter estimation is a Gamma distribution with positive shape and scale parameters $\alpha_1, \alpha_2 > 0$. Bayesian inference for this problem has been studied, for example, in [23] and methods for approximate computation of Bayes estimators were proposed using the Lindley approximation [18].

We consider the function $L_{sq}^{(2)}$ given in (17) to obtain Bayes estimators. As the loss function (17) is the sum of the univariate precautionary loss functions (2), the following estimators are used $\hat{\alpha}_i = \sqrt{\mathbb{E}(\alpha_i^2)},\ i = 1, 2$ where the expectation $\mathbb{E}$ is taken with respect to the posterior distribution.

Let us consider an experiment with sample size $n = 15$. The parameters of the Gamma distribution are varied over a grid, $\alpha_1, \alpha_2 \in (0, 10)$ and the performance of the different approaches are compared by simulations. The two approaches compared use the same prior distribution for parameters α_1 and α_2 using the following estimators:

(1) Bayes estimator under the squared error loss function,
(2) Bayes estimator under the multivariate precautionary loss function.

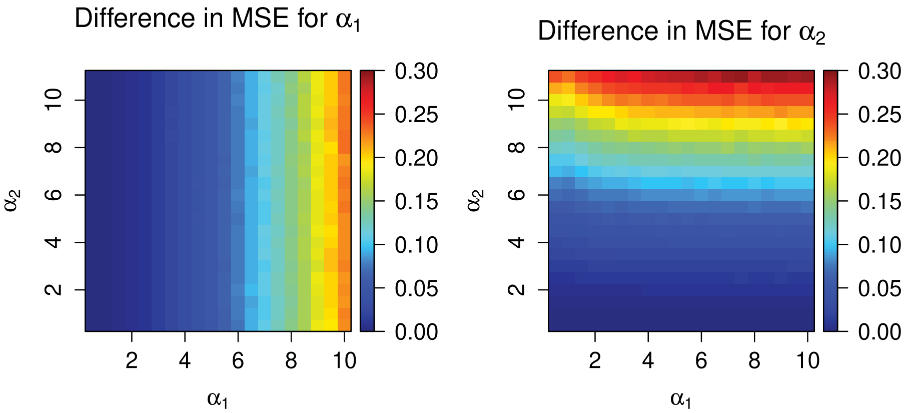

Figure 5. Difference in the MSEs for parameters α_1 and α_2 for their different true values and using Bayes estimator under the squared error loss function and Bayes estimator under $L_{sq}^{(2)}$. Results are based on 10^9 replications.

We use the same weakly informative Gamma prior distributions $\Gamma(10^{-4}, 10^{-4})$ of (α_1, α_2) with parameters for both estimation methods and an approximate method proposed by Lindley [18]. The weakly informative distribution are chosen to minimize the influence of the prior distribution on the comparison of estimators and corresponds to the situations of no prior knowledge about the parameters. We would like to emphasize that the choice of the prior distribution is out of the scope of this paper and the goal is to compare two Bayes estimators when all other parameters are equal. The difference between the MSE of estimators based on method (method 1 against method 2) for parameters α_1 and α_2 in 10^9 simulations is given in Figure 5.

The differences in the MSEs for both parameters are positive for all values of the true parameters α_1 and α_2. It means that the Bayes estimator from method 2) is associated with smaller MSE than the Bayes estimator from method 1). The difference in the MSE increases as the true value of the parameter increases. This result makes the proposed estimator and the associated loss function $L_{sq}^{(2)}$ good candidates for further investigation in multidimensional estimation problems.

5.4. Bayesian estimation of the parameters of a Weibull distribution

Another important example of multidimensional restricted parameters estimation is the Weibull distribution which positive scale and shape parameters $\lambda, \nu > 0$. The Weibull distribution is of great importance in applications as it is widely employed in, for example, reliability engineering, extreme value theory and survival analysis [26]. Bayesian inference for this problem has been studied in [9]. Importantly, despite both parameters being defined on the positive real line only, the squared error loss function (and associated posterior mean Bayes estimator) are used in these works. Below, we consider how the novel loss function (7) and the associated Bayes estimator behaves in this estimation problem.

We consider the function $L_k^{(m)}$ given in (16) to obtain Bayes estimators. As the loss function (16) is the sum of the univariate loss functions (6), the following Bayesian estimators are used for both scale parameters of Weibull distribution

$$\hat{\lambda}_k = \left(\frac{\mathbb{E}(\lambda^k)}{\mathbb{E}(\lambda^{-k})}\right)^{1/2k}, \quad \hat{\nu}_k = \left(\frac{\mathbb{E}(\nu^k)}{\mathbb{E}(\nu^{-k})}\right)^{1/2k} \tag{20}$$

where the expectations are taken with respect to the posterior density function. Note that the estimators depend on the parameter of the loss function k. We will investigate the influence of the parameter k on the estimation.

As above, we consider an experiments with a small sample size, $n = 15$. We consider several values of both scale and shape parameters, $\lambda \in \{1, 2, 3, 4, 5\}$, $\nu \in \{(0.5, 1, 5, 10, 15)\}$, and the performance of different approaches are compared by simulations. Again, the two approaches compared use the same weakly information Gamma prior distribution $\Gamma(10^{-4}, 10^{-4})$ for both positive parameters, as we would like to minimize the impact of the prior distribution on the comparison.

We start from the comparison of

(1) Bayes estimators under the squared error loss function,
(2) Bayes estimators (20) under the multivariate $L_1^{(m)}$ (16) for $k = 1$.

The difference between the MSE of estimators (method 1 against method 2) for positive parameters λ, ν in 10^4 simulations are given in Table 1. The MSE for ν are scaled by $\frac{1}{\nu^\lambda}$ to obtain the results on a similar scale for various parameters.

The differences in the MSEs for both parameters are positive for all considered true values of λ and ν. It follows that the Bayes estimator from method 2) leads to a smaller MSE than the estimator corresponding to the squared error loss function. For a fixed value of ν, the MSE corresponding to λ increases with the parameters. The scaled MSE corresponding to ν stays nearly the same for various values of λ. Overall, the proposed estimator and the associated loss function can be good candidates to be used for the parameters defined on the positive real line.

So far, only the loss function $L_k^{(m)}$ for $k = 1$ has been considered. To investigate the impact of the parameter k on the estimation characteristics, we fix the scale parameter $\nu = 1$ and vary the shape parameter over the grid $\lambda \in (1, 8)$. Specifically, we consider the following value of the parameter $k = \{1, 2, 3, 4\}$. The results are given in Table 2.

Table 1. Difference in the MSEs for parameters λ (upper lines) and ν (lower lines) for their different true values and using Bayes estimator under the squared error loss function and Bayes estimator under $L_1^{(m)}$.

	$\nu = 1$	$\nu = 2$	$\nu = 5$	$\nu = 10$	$\nu = 15$
$\lambda = 1$	0.0061	0.0065	0.0067	0.0069	0.0068
	0.0097	0.0081	0.0033	0.0067	0.0087
$\lambda = 2$	0.0268	0.0270	0.0274	0.0265	0.0269
	0.0090	0.0106	0.0103	0.0142	0.0142
$\lambda = 3$	0.0608	0.0616	0.0598	0.0622	0.0613
	0.0081	0.0125	0.0143	0.0136	0.0147
$\lambda = 4$	0.1075	0.1068	0.1065	0.1065	0.0987
	0.0104	0.0122	0.0146	0.0136	0.0151
$\lambda = 5$	0.1652	0.1649	0.1645	0.1652	0.1585
	0.0093	0.0131	0.0140	0.0115	0.0152

Note: Results are based on 10^4 replications.

Table 2. MSE, Bias and Variance for the Bayes estimator of $\lambda \in (1, 8)$ corresponding to the squared error loss function (posterior mean) and for the estimator $\hat{\lambda}_k$ given in Equation (20) for $k = 1,2,3,4$.

		$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$	$\lambda = 5$	$\lambda = 6$	$\lambda = 7$	$\lambda = 8$
Posterior Mean	MSE	0.072	0.286	0.653	1.150	1.781	2.582	3.515	4.571
	Bias	0.096	0.189	0.286	0.383	0.475	0.577	0.669	0.772
	Variance	0.063	0.250	0.571	1.003	1.555	2.249	3.067	3.975
Est. (20); $k = 1$	MSE	0.065	0.259	0.592	1.042	1.613	2.339	3.185	4.137
	Bias	0.072	0.142	0.215	0.287	0.356	0.434	0.502	0.581
	Variance	0.060	0.239	0.546	0.959	1.486	2.150	2.933	3.800
Est. (20); $k = 2$	MSE	0.065	0.258	0.589	1.036	1.604	2.325	3.166	4.112
	Bias	0.071	0.139	0.210	0.281	0.348	0.425	0.492	0.569
	Variance	0.060	0.238	0.545	0.957	1.482	2.144	2.924	3.789
Est. (20); $k = 3$	MSE	0.064	0.255	0.583	1.026	1.588	2.302	3.135	4.072
	Bias	0.068	0.134	0.203	0.271	0.336	0.410	0.474	0.549
	Variance	0.059	0.237	0.542	0.952	1.475	2.134	2.910	3.771
Est. (20); $k = 4$	MSE	0.063	0.252	0.575	1.012	1.567	2.271	3.092	4.016
	Bias	0.065	0.127	0.192	0.257	0.318	0.388	0.449	0.520
	Variance	0.059	0.236	0.538	0.946	1.465	2.120	2.891	3.746

Note: Results are based on 10^4 replications.

The posterior mean estimator (corresponding to the squared error loss function) corresponds to the highest MSE for all considered values of λ. As it was shown above, one can reduce the MSE by applying the estimator (20) for $k = 1$. Considering the bias $k = 1$, the estimator overestimates the true value of the parameter. When increasing the value of k, the value of the estimator decreases and approaches the true value of the parameter. This results in reduce bias of the estimator for all values of λ as k increases. Furthermore, the variance of the estimator decreases with parameter k. This results in decrease of the MSE as k increases. Overall, increasing values of k are found to lead lower estimates, which in the case of the overestimations leads to smaller MSE and more accurate estimation.

6. Application of the novel estimator to the paediatric clinical trial sample size calculation

6.1. Motivation

There is overwhelming medical evidence that children's response can noticeably differ from the adults' response in many diseases [15,32]. While there is generally a large amount of data available from adult studies, the knowledge about children remains quite limited. An example is drug-resistant partial epilepsy. Despite guidelines establishing the need to perform comprehensive paediatric drug development programmes, pivotal trials in children with epilepsy have been completed mostly in Phase IV as a postapproval replication of adult data [28]. To change this practice, more studies in children should be conducted. The planning of such studies, however, requires many assumptions, and the information from adults population can (and should) be efficiently used to justify them. Specifically, the planning requires the values of the expected responses (given previous trials) for an alternative treatment (or control) to which the comparison is to be made. The underestimation of this expected value can lead to a underpowered study and to unethical allocation of children in the study. In this section, we will demonstrate how the information about the relation between adults' and children's response can be incorporated in the proposed interval symmetric estimator and how it impacts the subsequent planning of the clinical trial.

6.2. Setting

Assume that one would like to conduct a randomized controlled clinical trial to study whether a novel intervention leads to a significantly different response in children with inadequately controlled partial seizures. A typical question that a clinician would ask a statistician before the trial is the sample size required to achieve a desirable level of statistical power. Suppose that equal sample sizes for the intervention and control (placebo) groups are to be used. Formally, the clinician would like to test the following hypothesis

$$H_0 : p \leq p_{placebo} \text{ versus } H_1 : p > p_{placebo}$$

where p is the unknown probability of response for the tested intervention and $p_{placebo}$ is the probability of response given the placebo. To achieve $1 - \beta$ in testing this hypothesis and given the type-I error α, one can obtain the following formula for the sample size

$$n \approx \frac{\left(z_{\alpha}\sqrt{2\bar{p}(1-\bar{p})} + z_{\beta}\sqrt{p_{target}(1-p_{target}) + p_{placebo}(1-p_{placebo})}\right)^2}{\left(p_{target} - p_{placebo}\right)^2}$$

where p_{target} is the clinically important response which the clinician would like to find in the trial, and z_{α} and z_{β} are $1 - \alpha$ and $1 - \beta$ quantiles of the normal distribution. The latter two values are to be defined by clinicians and statisticians. The clinically important response, given the severity of the diseases and alternative treatments [28], is assumed to be $p_{target} = 0.5$. However, the response corresponding to the placebo effect is much more challenging.

Given a vast amount of data from clinical trials in adults, it is known that the response to placebo is 0.10 [28]. However, one cannot use this value as there is an evidence that the placebo response can be lower for children. Moreover, the meta-analysis by Rheims et al. [28] suggested the placebo response is at 0.20 twice as large. Using this knowledge and the clinical data from the recent study [8], one can estimate the response for the placebo effect. We argue that the estimator corresponding to the interval symmetric loss function (13) is appropriate choice for at least two reasons:

(1) It is known that the clinically feasible values of the placebo response start at 0.1. Therefore, the probability of interest, $p_{placebo}$ lies in the interval $(0.1, 1)$.
(2) Underestimation of the placebo response is highly undesirable given that the trial is conducted in children. The underestimation leads to underpowered study, which might result in its failure and in the unethical 'waste' of children patients involved.

At the same time, one would like to limit a number of children enrolled in the study and the proposed sample size should be justified. The difference in the sample size calculations using the currently used 'naive' estimator for the placebo and the proposed interval symmetric estimator (14) is given below.

6.3. *Results*

Given the data obtained in the randomized clinical trial [8], there were 97 children patients assigned to the placebo group and 19 of them experienced a reduction of partial seizure frequency. Therefore, the 'naive' estimator that would be typically used to plan the clinical trial is

$$\hat{p}_{naive} = \frac{19}{97} \approx 0.196.$$

Alternatively, using the information that the estimator $p \in (0.1, 1)$ and assuming the uniform prior distribution for the probability and the Beta posterior $\mathcal{B}(x+1, n-x+1)$ as in Section 5.1, one can obtain the following formula for the interval symmetric estimator (14)

$$\hat{p}_{iq} = \frac{0.1 - \frac{(x+1)(x+2)}{(n+3)(n+2)} + \sqrt{\left(\frac{(x+1)(x+2)}{(n+3)(n+2)} - 0.1\right)^2 - \left(1.1 - 2\frac{x+1}{n+2}\right)\left(2.2\frac{x+1}{n+2} - 1.1\frac{(x+1)(x+2)}{(n+3)(n+2)}\right)}}{1.1 - 2\frac{x+1}{n+2}}.$$

Plugging-in the data from the study [8], $x = 19$ and $n = 97$, one can obtain

$$\hat{p}_{iq} \approx 0.209.$$

While the difference in the estimates can look quite marginal, it indeed leads to a difference in the required sample size per treatment group. Using $\alpha = 0.05$ type-I error, and the desirable power of $1 - \beta = 0.90$, one can obtain that the estimators lead to the following sample sizes: $n_{naive} = 41$ and $n_{iq} = 45$. Consequently, the proposed estimator suggests to enrol 8 more patients into the study. While this might seem to result in a minor change in the total sample size, this justified increase in the sample size can avoid a failure of the study and might lead to a new, better intervention available for children suffering from epilepsy.

7. Discussion

The concept of a symmetric loss function in a restricted parameter space is introduced in this paper. Scale symmetric and interval symmetric loss functions which share desirable properties are provided. On the basis of four examples, we show that the corresponding Bayes estimators perform well when compared to other available estimators based on squared error loss and improve the estimation if the parameter lies away from bounds. Following the real-life application example, it is found that the novel Bayes estimators allow avoiding boundary decisions that can be undesirable in paediatric clinical trials. Consequently, the estimator can be recommended in other applications where more conservative estimates are preferable.

Overall, when choosing a loss function for the particular application, we argue that a statistician should answer two questions: (i) is there credible information that the parameter of interest is restricted to particular space and (ii) are there any values of the parameters should be avoided as they might lead to undesirable consequences. Answering these questions will guide whether, for example, the squared error loss function, the scale symmetric loss function or the interval symmetric (with specified interval) loss function should be used. We would like to emphasize that we restrict our choice to some specific loss functions, mainly due to their simplicity and easy implementation. Alternative loss function sharing the stated properties can and should be considered.

The proposed definitions were generalized for a subset of $\mathbb{R}^m$, where distances on restricted spaces could themselves be used as loss functions, which usually are non-convex and do not result in explicit minimizers. While we have presented the modification of the squared error loss function for a restricted univariate parameter defined on an interval, the equivalent multidimensional extension seems to be non-trivial and requires further study.

Notes

1. The italic is a translation of a commentary to [7] appearing in the edition of Galilei's works mentioned in the bibliography, from which all of the quotes are taken, following [30].
2. A monetary unit, literally, a shield.

Acknowledgments

The authors acknowledge the insightful and constructive comments made by associate editor and two reviewers. These comments have greatly helped to sharpen the original submission.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567 and by the Medical Research Council (MR/M005755/1) and, in part, from Prof Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care (DHCS).

ORCID

P. Mozgunov http://orcid.org/0000-0001-6810-0284

References

[1] A. Agresti and B.A. Coull, *Approximate is better than "exact" for interval estimation of binomial proportions*, Am. Stat. 52 (1998), pp. 119–126.

[2] J. Aitchison, *On criteria for measures of compositional difference*, Math. Geol. 24 (1992), pp. 365–379.

[3] J. Berger, et al., *The case for objective bayesian analysis*, Bayesian Analy. 1 (2006), pp. 385–402.

[4] L. Brown, *Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters*, Anna. Math. Stat. 39 (1968), pp. 29–48.

[5] L.D. Brown, T.T. Cai and A. DasGupta, *Interval estimation for a binomial proportion*, Stat. Sc. 16 (2001), pp. 101–117.

[6] B. Efron, *Frequentist accuracy of bayesian estimates*, JRSS: B 77 (2015), pp. 617–646.

[7] G. Galilei, *Lettera (intorno la stima di un cavallo).*, Le opere di Galileo Galilei (1627), p. Prima edizione completa. Societa editrice fiorentina. Firenze.

[8] T. Glauser, R. Ayala, R. Elterman, W. Mitchell, C. Van Orman, L. Gauer, Z. Lu and N.S. Group, et al., *Double-blind placebo-controlled trial of adjunctive levetiracetam in pediatric partial seizures*, Neurology 66 (2006), pp. 1654–1660.

[9] P.K. Gupta and A.K. Singh, *Classical and bayesian estimation of Weibull distribution in presence of outliers*, Cogent Math. 4 (2017), pp. 1300975.

[10] J.A. Hartigan, *Uniform priors on convex sets improve risk*, Stat. Prob. Lett. 67 (2004), pp. 285–288.

[11] W. James and C. Stein, *Estimation with quadratic loss*, in *Pr. of the Fourth Berkeley Symposium on Math. Stat. and Prob.*, Vol. 1. 1961, pp. 361–379.

[12] A. Karimnezhad and F. Moradi, *Bayes, e-bayes and robust bayes prediction of a future observation under precautionary prediction loss functions with applications*, Appl. Math. Model. 40 (2016), pp. 7051–7061.

[13] A. Karimnezhad, S. Niazi and A. Parsian, *Bayes and robust bayes prediction with an application to a rainfall prediction problem*, J. Korean Stat. Soc. 43 (2014), pp. 275–291.

[14] A. Kiapour and N. Nematollahi, *Robust bayesian prediction and estimation under a squared log error loss function*, Stat. Prob. Let. 81 (2011), pp. 1717–1724.

[15] T.P. Klassen, L. Hartling, J.C. Craig and M. Offringa, *Children are not just small adults: the urgent need for high-quality trial evidence in children*, PLoS. Med. 5 (2008), pp. e172.

[16] T. Kubokawa, *A unified approach to improving equivariant estimators*, Ann. Stat. 22 (1994), pp. 290–299.

[17] S. Kumar and Y.M. Tripathi, *Estimating a restricted normal mean*, Metrika 68 (2008), pp. 271–288.

[18] D.V. Lindley, *Approximate bayesian methods*, Trabajos de estadística y de investigación operativa 31 (1980), pp. 223–245.

[19] E. Mahmoudi and H. Zakerzadeh, *An admissible estimator of a lower-bounded scale parameter under squared-log error loss function*, Kybernetika 47 (2011), pp. 595–611.

[20] É. Marchand, A. Najafabadi and T. Payandeh, *Bayesian improvements of a mre estimator of a bounded location parameter*, Elect. J. Stat. 5 (2011), pp. 1495–1502.

[21] E. Marchand and W.E. Strawderman, *Estimation in restricted parameter spaces: A review*, in *A Festschrift for Herman Rubin*, Inst. of Math. Stat., 2004, pp. 21–44.

[22] G. Mateu-Figueras, V. Pawlowsky-Glahn and J.J. Egozcue, *The normal distribution in some constrained sample spaces*, SORT 37 (2013), pp. 29–56.

[23] R.B. Miller, *Bayesian analysis of the two-parameter gamma distribution*, Technometrics 22 (1980), pp. 65–69.

[24] P. Mozgunov and T. Jaki, *Improving a safety of the CRM via a modified allocation rule*, Preprint, arXiv:1807.05781 (2018).

[25] J.G. Norstrom, *The use of precautionary loss functions in risk analysis*, IEEE Trans. Reliab. 45 (1996), pp. 400–403.

[26] F.N. Nwobi and C.A. Ugomma, *A comparison of methods for the estimation of Weibull distribution parameters*, Metodoloski zvezki 11 (2014), pp. 65.

[27] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2015). https://www.R-project.org/.

[28] S. Rheims, M. Cucherat, A. Arzimanoglou and P. Ryvlin, *Greater response to placebo in children than in adults: a systematic review and meta-analysis in drug-resistant partial epilepsy*, PLoS. Med. 5 (2008), pp. e166.

[29] G. Saint-Hilary, V. Robert, T. Jaki, M. Gasparini and P. Mozgunov, *A novel measure of drug benefit-risk assessment based on scale loss score (slos)*, Stat. Method. Med. Res. (2018). doi.org/10.1177/0962280218786526.

[30] I. Scàrdovi, *Appunti di statistica*, Vol. 1. Pàtron, Bologna, 1980.

[31] C. Stein, *Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean*, Ann. Inst. Stat. Math. 16 (1964), pp. 155–160.

[32] T. Stephenson, *How children's responses to drugs differ from adults*, Br. J. Clin. Pharmacol. 59 (2005), pp. 670–673.

[33] E.B. Wilson, *Probable inference, the law of succession, and statistical inference*, J. Am. Stat. Assoc 22 (1927), pp. 209–212.