Scale Symmetric Loss Functions for Bayesian Analysis

by

Mauro Gasparini*
Purdue University

Technical Report #95-16

Department of Statistics
Purdue University

April 1995

* Research partially supported by the National Science Foundation, Grant DMS-9303556.

Scale symmetric loss functions for Bayesian analysis

Mauro Gasparini[1]
Purdue University

Abstract

Scale symmetry is a natural property to be imposed on the loss function, when a positive parameter θ is being estimated. It can be justified with a four century old argument. Simple scale symmetric loss functions and related descriptive measures of a distribution are studied, for the purpose of summarizing a Bayesian posterior distribution.

[1]Research partially supported by the National Science Foundation, Grant DMS-9303556.
[0]*Key words and phrases.* Scale symmetry, scale invariance, scale mean, scale variance.
[0]*AMS 1980 subject classifications.*Primary 62G05;secondary 62P99.

1 Some history and a summary

In the Spring of 1627, a *peculiar controversy* [2]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* [a monetary unit, literaly, a shield], is estimated by someone to be one thousand *scudi* and by someone else to be ten *scudi* worth: 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 wants Nozzolini's opinion on a sentence by Galileo Galilei (1627), 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*

(" ... the two persons estimating [the horse worth] have been equally exorbitant and are responsible of 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 more 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 ?*

(" If one were to overestimate two-hundred-arm high a tower, which is really one-hundred-arm high, what underestimate would Nozzolini consider as equally deviating?").

It is apparent that, in modern statistical terms, the problem may be formulated as choosing between two kinds of loss functions in a decision-theoretic approach to estimation of an unknown, but intrinsically positive, parameter θ. The way the problem was discussed by the *noble gentlemen* is remarkably modern, in that the true value, its

[2]the italic is a translation of a commentary to Galilei (1627) appearing in the edition of Galilei's works mentioned in the bibliography, from which all of the quotes are taken.

estimates and the loss (or *extravagance*, using their colorful expression) are clearly and separately identified.

In section 2, it is seen how Galileo's requirement of a *geometric proportion* can be recast in modern terms as a simple and appealing property of the loss function, namely *scale symmetry*. Curiously, these four-century old scale symmetric loss functions have not received much attention, although they have been briefly discussed by Brown (1968, page 37). In section 3 and 4, the use of a particularly simple scale symmetric loss function for Bayesian analysis is analyzed. Section 5 illustrates how the use of other scale symmetric loss functions may lead to sensible computational problems.

2 Scale symmetric loss functions

In the modern statistical literature, the inadequacy of difference-based loss functions, like square error loss, for estimating certain positive quantities has always been recognized. Several alternative loss functions have been proposed, the best-known being the normalized quadratic loss function [Stein (1964)]

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

Stein's loss

$$L_S(\theta, d) = \frac{d}{\theta} - 1 - \log(\frac{d}{\theta}) \tag{2}$$

and Brown's loss function [Brown (1968)]

$$L_B(\theta, d) = (\log(\frac{d}{\theta}))^2, \tag{3}$$

where d is the estimate of the positive parameter θ and a constant is subtracted, when necessary, to obtain $L(\theta, \theta) = 0$. All loss functions above are *scale invariant*, in the sense that

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

for every $c > 0$ and every pair (θ, d). Equivalently, a loss function $L(\theta, d)$ is scale invariant if and only if there exist a scalar function $g(x), x > 0$ such that $L(\theta, d) = g(d/\theta)$.

Now, Galileo's claim of *eguali stravaganze* can be reexpressed in modern terminology as a requirement of a *scale symmetric* loss function, as in the following

Definition 1 *A loss function is called* scale symmetric *if, for every positive* d_1, d_2, θ,

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

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

It is clear that Galileo's expression *"geometric proportion"* comes from relationship (4).

Theorem 1 *A scale invariant loss function is scale symmetric if and only if it can be written as a scalar function g such that $g(d/\theta) = g(\theta/d)$.*

Proof. Scale invariance implies $L(\theta, d) = g(d/\theta)$ for some g, scale symmetry implies $L(\theta, d) = L(\theta, \theta^2/d)$ and viceversa. ∎

Loss function L_B is scale symmetric, L_Q and L_S are not. For any real function $w(\theta)$, and any scale symmetric and scale invariant loss function L, the product $w(\theta)L$ is scale symmetric but not, in general, scale invariant.

3 Scale means

We now focus on some particular scale invariant and scale symmetric loss functions. A natural subclass are the ones for which the function $g(x)$ of theorem (1) is written as an average of $h(x)$ and $h(1/x)$, for a nondecreasing function $h(x), x > 0$. Among the many possible choices, $h(x) = x^k$ seems to be a natural one, providing

$$L_k(\theta, d) = \frac{1}{2}[(\frac{d}{\theta})^k + (\frac{\theta}{d})^k] - 1, \tag{5}$$

where, as usual, a constant is subtracted for normalization purposes. We then have the following theorem:

Theorem 2 *Let Θ be a positive random variable and $k > 0$. If $E\Theta^k < \infty$ and $E\Theta^{-k} < \infty$, then the expectation of loss function (5)*

$$d = \nu_k(\Theta) := (\frac{E\Theta^k}{E\Theta^{-k}})^{\frac{1}{2k}}. \tag{6}$$

In particular, write $\nu(\Theta) := \nu_1(\Theta)$. Then

$$\frac{1}{E\Theta^{-1}} \leq \nu(\Theta) = \sqrt{\frac{E\Theta}{E\Theta^{-1}}} \leq E\Theta. \tag{7}$$

Proof. The first result is proved by differentiating in d, the inequalities are proved by Jensen's inequality applied to the function $1/\Theta$. ∎

Definition 2 *Let Θ be a positive random variable with finite $E\Theta^k$ and $E\Theta^{-k}$. Then $\nu_k(\Theta)$ is called the* scale mean *of order k of the random variable Θ.*

Scale means are particularly meaningful in the context of Bayes estimation, where Θ is an unknown parameter and its distribution a posterior distribution, given some sample information. If a point estimate has to be selected according to loss function (5), then $\nu_k(\Theta)$ is usally called a Bayes estimate. The scale mean of order 1, or simply the *scale mean* $\nu(\Theta)$, is particularly interesting. First, the fact that underestimation is penalized as much as overestimation does not imply that the resulting Bayes estimate $\nu(\Theta)$ is greater (further away from 0) than the Bayes estimate under squared error loss $E\Theta$. One would expect such a thing since squared error loss is finite at $d = 0$. While this may be true for other loss scale symmetric functions, as discussed for example in Brown (1968, page 37), it is not true for $L_1(\theta, d)$, as seen in (7). In a sense, this is a consequence of requiring $E\Theta^{-1} < \infty$, a condition which, at first, seems quite arbitrary, but is is indeed not any more special than requiring $E\Theta$ to exist finite. The scale group of transformations is a natural group for the problem of estimating a positive (scale) parameter, much in the same way the translation group is a natural group when estimating a location parameter. From this point of view, it is much too tempting looking at formula (7) as a very close analogue to the following redundant way of writing an expectation:

$$E\Theta = \frac{E\Theta - E(-\Theta)}{2}$$

The analogy between the scale mean and the usual mean then becomes clear if we substitute scale symmetry for location symmetry, ratios for differences and the geometric mean for the arithmetic mean. Much in the same way $E\Theta < \infty$ requires that both the expectation of the positive and the negative part of Θ be finite, so the finiteness of $\nu(\Theta)$ is a consequence of $E\Theta < \infty$ and $E\Theta^{-1} < \infty$.

In view of the above discussion, loss function $L_1(\theta, d)$, or more in general $L_k(\theta, d)$, may sometimes be preferable, for example, to $L_Q(\theta, d)$, which generates a Bayes estimator $E\Theta^{-1}/E\Theta^{-2}$, not sharing the same characteristics of symmetry.

It is convenient to use a power $k \neq 1$ in situations when the estimand is expressed in some units different from the data. For example, in estimating a variance, it may be more natural to use a scale mean of order $1/2$.

Example 1 Consider estimating the unknown common variance θ of a random sample $X_1, \ldots, X_n$ from a normal distribution with unknown mean μ, using loss function $L_{1/2}(\theta, d)$. The standard conjugate prior for (μ, θ) is an inverse gamma on θ, with density $\pi(\theta) = b^a e^{-b/\theta}/(\Gamma(a)\theta^{a+1})$, $a, b > 0$, and, conditionally on θ, a normal on μ with mean m and variance θ/p, $p > 0$. Standard computations show that the limit of the corresponding Bayes estimator of θ obtained using formula (6) with $k = 1/2$, as $a, b, p \to 0$,

is $\nu_{1/2}(\Theta) = \sum(X_i - \bar{X})^2/(n-1)$, which is the usual unbiased estimate of Θ and can also be obtained using loss function $L_S(\theta, d)$.

Example 2 As an example of a computation of a scale mean in the presence of nuisance parameters, consider estimating the scale parameter of a Weibull distribution. Given type II censored observations $X_1, \ldots, X_n$ from the density $(\beta/\theta)x^{\beta-1}e^{-x^\beta/\theta}$, let r be the number of observed order statistics and $\sum^* x_i^\beta := \sum_{i=1}^r x_i^\beta + (n-r)x_r^\beta$. Joint continuous conjugate priors on θ and β do not exist [Soland (1969)], but it is customary to choose a prior density on θ and β of the form $\pi(\beta)b^a e^{-b/\theta}/(\Gamma(a)\theta^{a+1})$. $\pi(\beta)$ is then the marginal prior on β and θ has, conditionally on β, an inverse gamma distribution, with parameters a and b which may depend on β. Standard calculations show that the posterior scale mean of Θ is

(8)
$$\left[\int \frac{\beta^r(\prod x_i)^{\beta-1}}{(b+\sum^* x_i^\beta)^{a+r-1}}\pi(\beta)d\beta\right]^{1/2}\left[(a+r)(a+r-1)\int \frac{\beta^r(\prod x_i)^{\beta-1}}{(b+\sum^* x_i^\beta)^{a+r+1}}\pi(\beta)d\beta\right]^{-1/2}$$

The computation of (8) does not seem to be any more difficult than the computation of the posterior expectation. For numerical and Gibbs sampling computational methods see for example Canavos and Tsokos (1973) and Berger and Sun (1993), respectively. If β is known and $a, b \to 0$, then expression (8) tends to $\sum^* x_i^\beta/\sqrt{r(r-1)}$, which is smaller then the usual Bayes estimator with an improper prior $\sum^* x_i^\beta/(r-1)$ and greater than the MLE estimator $\sum^* x_i^\beta/r$.

4 Scale variances

In the previous section, we have used symmetric loss functions to obtain "Bayes estimators", to be compared to standard frequentist estimators. This mixed approach to inference is sometimes convenient, since it allows for comparison between different procedures in terms of frequentist measures of performance.

In a more fundamental Bayesian approach to inference though, a Bayes estimator is regarded only as a convenient summarization of the posterior [for a discussion, see for example Box and Tiao (1973, Appendix A5.6)]. A loss function is then a way to prescribe what kind of summary is appropriate. Typically, a posterior expectation is used as the Bayes estimator, implying that a quadratic loss function is being used. A second step is usually taken to accompany the Bayes estimate with a measure of uncertainty of the posterior. If a posterior mean is used, a posterior variance is presented at this stage. But, if a specified 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 should also be reasonable to present the achieved minimum of the posterior expected loss as a second summary of the posterior. Perhaps the reason why a posterior variance, or even a frequentist MSE, are sometimes used instead of an expected posterior loss is that only few losses actually allow for simple expected posterior losses. For example, the use of loss function $L_S(\theta, d)$ involves the posterior expectation of the logarithm of Θ, which may not have an immediate intuition for a user.

In the previous section we have argued for the use of symmetric loss functions for certain kinds of problems. In particular, we tried to appreciate the usefulness of simple, albeit arbitrary, loss functions of the form (5). These loss functions give rise to particularly simple expected posterior losses, which may be of independent interest as measures of variability of a positive variate.

Definition 3 *Let Θ be a positive random variable with finite $E\Theta^k$ and $E\Theta^{-k}$. Then*

$$\tau_k(\Theta) := \sqrt{E\Theta^k E\Theta^{-k}} - 1$$

is called the scale variance *of order k of the random variable Θ. If either $E\Theta^k$ or $E\Theta^{-k}$ is infinite, then $\tau_k(\Theta) := \infty$.*

In a Bayesian framework, if Θ is assigned a posterior distribution given some data, then $\tau_k(\Theta)$ is the posterior expected loss, when loss function (5) is used.

Scale variances are descriptive measures of the distribution of a positive random variable. They are scale invariant and symmetric, in the sense that $\tau_k(c\Theta) = \tau_k(\Theta)$, for $c > 0$, and $\tau_k(\Theta^{-1}) = \tau_k(\Theta)$. In particular, let $\tau(\Theta) := \tau_1(\Theta)$ be called scale variance *tout court*. By Jensen's inequality, we have $0 \leq \tau(\Theta) \leq \infty$. Table 1 contains scale means and scale variances for a few commonly used distributions.

5 Scale centers

There are many other ways to define a g satifying the conditions of theorem 1 and obtain a scale symmetric loss function. Another simple one is the following: given a nondecreasing function $h(x), x > 0$, define

$$g(x) = \max\{h(x), h(1/x)\},$$

in analogy with a symmetric loss functions $\phi(|\theta - d|)$, for some nondecreasing ϕ, in a location parameter problem.

Example 3 Take $h(x) = (\log(x))^2$ and obtain Brown's loss function $L_B(\theta, d)$.

As in section 3, in connection with the choice $h(x) = x^k, k > 0$, we have

$$L_{max}(\theta, d) = (\max\{\frac{d}{\theta}, \frac{\theta}{d}\})^k - 1 \tag{9}$$

and the following theorem.

Theorem 3 *Let Θ be a positive random variable with density $\pi(\theta)$ and $k > 0$. If $E\Theta^k < \infty$ and $E\Theta^{-k} < \infty$, then the expectation of loss function (9) is minimized by the solution d_k to the following equation.*

$$d_k^{2k} = (\int_{d_k}^{\infty} \theta^k \pi(\theta) d\theta)(\int_0^{d_k} \theta^{-k} \pi(\theta) d\theta)^{-1}. \tag{10}$$

Proof. The minimization problem

$$\min_d \int (\max\{\frac{d}{\theta}, \frac{\theta}{d}\})^k \pi(\theta) d\theta,$$

is easily seen to be equivalent, by differentiating in d, to equation (10), which has a unique solution since the left hand side is an increasing and the right hand side a decreasing function of d, spanning all the positive reals. ∎

Definition 4 *Let d_k be called the* scale center *of order k of the random variable Θ.*

The center of order 1 is a sort of scale median for the parameter Θ, but see Lehmann (1983, page 176) for an alternative definition of *scale median.*

Unfortunately , finding a scale center usually requires numerical methods, for example Newton-Raphson, and it seems hopeless to obtain estimators and posterior expected losses of direct interpretability.

	density (for $x > 0$)		scale mean	scale variance
inverse gamma	$(\Gamma(\alpha)x^{\alpha+1})^{-1}\beta^{\alpha}e^{-\beta/x}$	$\alpha > 1$ $\beta > 0$	$\frac{\beta}{\sqrt{\alpha(\alpha-1)}}$	$\sqrt{\frac{\alpha}{\alpha-1}} - 1$
gamma	$\Gamma(\alpha)^{-1}\beta^{\alpha}x^{\alpha-1}e^{-\beta x}$	$\alpha > 1$ $\beta > 0$	$\frac{\sqrt{\alpha(\alpha-1)}}{\beta}$	$\sqrt{\frac{\alpha}{\alpha-1}} - 1$
Weibull	$(\beta/\lambda)(x/\lambda)^{\beta-1}e^{-(x/\lambda)^{\beta}}$	$\beta > 1$ $\lambda > 0$	$\lambda\sqrt{\frac{\Gamma(1+1/\beta)}{\Gamma(1-1/\beta)}}$	$\sqrt{\Gamma(1+\beta^{-1})}\times$ $\sqrt{\Gamma(1-\beta^{-1})} - 1$
lognormal	$(\sigma x\sqrt{2\pi})^{-1}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}$	$\mu > 0$ $\sigma > 0$	e^{μ}	$e^{\sigma^2/2} - 1$
inverse normal	$\sqrt{\frac{\lambda}{2\pi x^3}}e^{-\lambda\frac{(x-\mu)^2}{2\mu^2 x}}$	$\mu > 0$ $\lambda > 0$	$\sqrt{\frac{\mu^2\lambda}{\mu+\lambda}}$	$\sqrt{1+\frac{\mu}{\lambda}} - 1$

Table 1: Scale mean $\nu(X)$ and scale variance $\tau(X)$.

References

Berger, J.O. and Sun, D. (1993). Bayesian Analysis for the Poly-Weibull distribution. *J. Amer. Statist. Assoc.*, **88**, 1412-1418.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian inference in statistical analysis.* Wiley, New York.

Brown, L.D. (1968). Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. *Ann. Math. Statist.* **39**, 29-48.

Canavos, G.C. and Tsokos, C.P. (1973). Bayesian estimation of life parameters in the Weibull distribution. *Operations Research*, **21**, 755-763.

Galilei, Galileo (1627). Lettera (intorno la stima di un cavallo). In *Le opere di Galileo Galilei.* Prima edizione completa. Società editrice fiorentina. Firenze 1855.

Lehmann, E.L. (1983). *Theory of point estimation.* Wiley, New York.

Soland, R.M. (1969). Bayesian analysis of the Weibull process with unknown scale and shape parameters. *IEEE Trans. Reliab.*, **R-18**, 181-184.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.* **16**,155-160.