Median Loss Decision Theory

Chi Wai Yu[a,], Bertrand Clarke[b]

[a]*Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*
[b]*Department of Medicine, Department of Epidemiology and Public Health, and the Center for Computational Sciences, University of Miami, 1120 NW 14th Street, Miami, FL, 33136.*

Abstract

In this paper, we argue that replacing the expectation of the loss in statistical decision theory with the median of the loss leads to a viable and useful alternative to conventional risk minimization particularly because it can be used with heavy tailed distributions. We investigate three possible definitions for such *medloss* estimators and derive examples of them in several standard settings. We argue that the *medloss* definition based on the posterior is better than the other two definitions that do not permit optimization over large classes of estimators. We argue that median loss minimizing estimates often yield improved performance, have resistance to outliers as high as the usual robust estimates, and are resistant to the specific loss used to form them. In simulations with the posterior *medloss* formulation, we show how the estimates can be obtained numerically and that they can have better robustness properties than estimates derived from risk minimization.

Key words: Asymptotics, decision theory, loss function, median, model selection, posterior, prediction
2000 MSC: 62C10, 62F35

1. Introduction

Statistical decision theory in Wald [18, 19] is either Bayesian or Frequentist. Frequentist decision theory rests on expected utility as demonstrated in Von Neumann and Morgenstern [17]. Specifically, they identified two axioms, called Independence and Continuity, under which choices ranked by a preference relation could be represented by an expected utility. This meant that the optimal action under uncertainty could be found by maximizing expected utility. In the Bayesian context, Savage [16] used six axioms and again established an equivalence between ordering under a preference relation and an expected utility so that again optimal actions under uncertainty could be found by maximizing expected utility. Because of its appealing properties, maximizing expected utility became standard.

Email addresses: macwyu@ust.hk (Chi Wai Yu), bclarke2@med.miami.edu (Bertrand Clarke)

Despite this, expected utility as a criterion for decision making has been repeatedly criticized from several standpoints. First, the Allais paradox, in Allais [1], calls the Independence axiom, and hence standard Frequentist decision theory into question. Second, Ellsberg's paradox, in Ellsberg [2], shows that Savage's Sure-Thing principle, the counterpart of the Independence axiom, also contradicts real life decision making, thereby calling the standard Bayesian formulation into question. In addition, it is well known that, in both the Bayesian and Frequentist contexts, the decision rules depend delicately on the loss functions which are usually unknown, leading to poor robustness properties. These findings, among others, have motivated the development of several utility models that are not based on expectation.

One of the alternatives to expected utility models for decision making under uncertainty is due to Manski [9]. He suggested using quantiles of the distribution of the utility as the main criterion in the Frequentist context. One of the benefits of optimizing quantiles rather than means is that moments no longer need to be assumed. So, heavy tailed distributions are not a problem. In fact, quantile based criteria are well-established in many applications, especially in finance. For instance, one of the most popular risk measures in finance is the Value-at-risk, VaR, which is often taken to be a high quantile, say the 95-th or 99-th, of the negative of the utility which can be regarded as the loss function here. Despite this, no axiomatic justification for quantile optimization was offered by Manski [9]. This gap led Rostek [13] to develop axioms under which a preference ordering over actions would be equivalent to the ordering of those actions under a quantile representation. Her work made use of Machina and Schmeidler [8] which provided an axiomatization for non-expected-utility models in a Bayesian setting.

Here, we minimize the median of the loss rather than the expected loss and refer to this as *medloss* estimation. We choose the 0.5 quantile because our interest focuses on the central tendency of the response not its tail properties, which are necessarily less stable. This means that in an estimation setting both under- and over-estimation are regarded as equally bad and in a predictive setting under- and over-prediction are regarded as equally bad. Indeed, let $L(\delta(X), \theta)$ be used as a measure of adequacy of an action or estimator $\delta(X)$ for the parameter θ, taken as the true state of nature and suppose that that X is distributed according to P_θ. Now, for each θ, the loss $L(\delta(X), \theta)$ is a random variable that is typically (strongly) right skewed. If the skewness is not negligible, the median is more representative of the typical location of $L(\delta(X), \theta)$ than the mean is as well as being more stable in the sense of having a higher breakdown point.

There are no less than three distinct ways to define a *medloss* estimator in parametric settings. We present these three definitions and argue that two of them are very similar and only permit limited optimizations. That is, they are intractable unless the action space is small, having one of two real dimensions for instance. The third definition, which we prefer, is based on a posterior formulation and permits optimization over larger actions spaces. It may also have better robustness properties because the prior used to form the posterior often leads to inferences that are a little less sensitive to outliers than a real valued statistic.

So far, we have described axiomatic and conceptual arguments for looking at the median of a loss function or *medloss*. However, the applicability of median methods to heavy tailed distributions may be more important from a modeling and data analytic standpoint than is commonly realized. In particular, it is very easy to be misled into accepting a model that incorrectly treats a heavy tailed error as normal.

For clarity, let us see how this problem may occur in a simple example. Suppose that the correct data generating model for Y is

$$Y_i = 2X_{1i} + E_i, \text{ for } i = 1, \ldots, n, \tag{1}$$

where $\{E_i : i = 1, \ldots, n\}$ is IID from the standard Cauchy distribution and X_1 is generated IID $N(0,1)$. Let us define a new variable X_2 to be the result of setting $X_2 = \hat{e}$ + normal noise, where $\hat{e}$ is the residual from the least squares estimator in (1) using only Y and X_1. Then, it is straightforward to generate triples (Y_i, X_{1i}, X_{2i}) for $i = 1, \ldots, n$. These triples will fit the model $Y = \beta_1 X_1 + \beta_2 X_2 + E$, where E is IID normal error.

That is, if one has a multivariable regression problem and one of the variables roughly replicates the residuals of a simple model with heavy tailed error one may be misled into an incorrect model which has a normal error. In short, a conventional normal error model may merely be the construct of the variable selection and fitting rather than anything real.

Quantile optimization has been suggested from the standpoint of robustness against the loss function as well. An instance of this property is shown in Theorem 1 below. A related problem is the difficulty in reliably specifying a loss function in practice. The consequence of this is that it is desirable to have the right degree of insensitivity to the loss function. In the Bayesian setting, Rukhin [14] shows that a reasonable degree of robustness holds for even loss functions when the posterior density is symmetric and unimodal, under the assumption that the posterior expected loss exists and is finite. He also gives conditions under which Bayes estimators are independent of the loss functions, but this is a very special case. Other techniques for managing sensitivity of the loss include the Pitman measure of closeness, PMC, see Pitman [10], universal domination (UD) and stochastic domination (SD), see Hwang [4]. However, the PMC is not transitive and UD and SD are hard to implement in practice. Quantile optimization is not the only method for managing loss-robustness, however, other methods to date have limitations.

The rest of this paper is organized as follows. In Section 2 we recall Rostek's axiomatic foundation of the quantile utility model. This leads us to present three definitions for *medloss* estimators. In Section 3 we present examples under all three definitions. In Section 4 we relate the first two definitions to each other through the concepts of median admissibility and minimaxity and give further examples. In Section 5, we turn to our preferred definition based on the posterior. We give a simple technique for finding posterior *medloss* estimators computationally and show that in the case of a well-behaved symmetric posterior the posterior *medloss* estimator reduces to the median of the posterior. Then we look computationally at the robustness properties of the *medloss* estimator in the Gamma family. In a short concluding section we state the main implications of our work.

2. Quantile axiomatic foundation

In this section, we present the motivation for minimizing the median loss (hereafter *medloss*) from the standpoint of utility theory.

2.1. Quantile Utility

Advocating a quantile-utility approach to modeling choice under uncertainty, Manski [9] examined an agent's preferences by maximizing a quantile of the distribution of utility. Unfortunately, he did not provide any axioms under which necessary and sufficient conditions for preference ordering would admit a quantile representation. Machina and Schmeidler [8] did provide a foundation for non-expected-utility models by modifying Savage's axioms. However, this new axiomatic system still did not fit the quantile utility model well, because, as Rostek [13] points out, some of the axioms were too restrictive.

Consequently, Rostek [13] axiomatized quantile maximization for decision making in Savage's setting by combining the results of Manski [9] and Machina and Schmeidler [8]. More specifically, taking preferences over acts as a primitive, Rostek [13] finds conditions that are necessary and sufficient for those preferences to admit a quantile representation. Indeed, a detailed examination of the axioms for quantile and expected utility reveals that considerably weaker axioms are used in quantile utility representations for preference relations. In particular, quantile utility has properties such as robustness to the choice of utility functions in the sense that they are unique up to strictly increasing transformations and quantile utility models do not have moment restrictions. Thus, quantile utility optima have some inbuilt resistance to outliers, unlike expected utility optima.

Note that we are not saying that median loss models are always better than the expected loss models, or vice versa. Indeed, the relative usefulness of the two types of optimization depends on the application. We believe median loss approaches are better for most unimodal situations involving outliers, asymmetry, heavy tails or the possibility that the data generator is more dispersed than the model. By contrast, if the data comes from a unimodal distribution believed to be 'nice' – symmetric with light tails – then the expected utility representation for preference ordering should perform better.

Throughout this paper, we denote the set of real numbers by $\mathcal{R}$, and a sequence of n random variables by $X^n = (X_1, \ldots, X_n)$ with realizations denoted $x^n = (x_1, \ldots, x_n)$. Now, suppose a random variable X has a distribution F. We denote its median by $\operatorname{med} X$, $\operatorname{med}(X)$, or $\operatorname{med} F$. For the median of a function of X, say $g(X)$, we denote its median by $\operatorname{med}_X g(X)$ or $\operatorname{med}_F g(X)$.

2.2. Three medloss Criteria

Having justified quantile utilities and argued that UD, SD, and PMC are not satisfactory, we now introduce our three *medloss* criteria. All of them satisfy Savage's principle that no reasonable criterion can separate two estimators with the same marginal distribution, see Savage [16]. Also, all three of them are easier to satisfy than UD or SD.

2.2.1. A Frequentist medloss criterion

Our first *medloss* criterion is Frequentist and is the following.

Definition 1. *An estimator $\delta_1(X^n)$ is better than an estimator $\delta_2(X^n)$ in the sense of median loss (medloss) iff*

$$P_\theta\Big(L(\delta_1(X^n),\theta) \leq med_{X^n}\, L(\delta_2(X^n),\theta)\Big) \geq 0.5, \tag{2}$$

for all $\theta \in \Theta$, which is equivalent to

$$med_{X^n}\, L(\delta_1(X^n),\theta) \leq med_{X^n}\, L(\delta_2(X^n),\theta),\ \forall\, \theta \in \Theta,$$

where Θ is the parameter space and $med_{X^n}\, L(\delta(X^n),\theta)$ denotes the median of the loss $L(\delta(X^n),\theta)$.

In particular, for any strictly increasing loss function L of $\|\delta(X^n)-\theta\|_Q$, $\delta_1(X^n)$ is better than $\delta_2(X^n)$ in the sense of *medloss* iff

$$\text{med}_{X^n}\,(L\|\delta_1(X^n)-\theta\|_Q) \leq \text{med}_{X^n}\,(L\|\delta_2(X^n)-\theta\|_Q), \quad \forall\, \theta \in \Theta, \tag{3}$$

or equivalently,

$$\text{med}_{X^n}\,(\|\delta_1(X^n)-\theta\|_Q) \leq \text{med}_{X^n}\,(\|\delta_2(X^n)-\theta\|_Q), \quad \forall\, \theta \in \Theta.$$

In addition to giving solutions without requiring moment conditions, this *medloss* criterion has some of the other nice properties of PMC. For instance, Keating, Mason and Sen [6] identify robustness against the loss as an important desideratum. Expected-loss based criteria usually do not have this robustness, but the PMC does as does the *medloss*. In particular, Keating, Mason and Sen [6] show that the PMC is invariant under powers of the absolute error loss, and it is not hard to show this holds for the Frequentist *medloss* too. This invariance dramatizes that the *medloss* focuses only on the ranking of estimators not on evaluating the amount by which one estimator is preferred over another.

2.2.2. Bayes medloss Criterion

There are two notions of *medloss* in the Bayesian context. The first is analogous to the Bayes risk and so uses the joint distribution of (Θ, X^n).

Definition 2. *Let $\pi(\cdot)$ be the prior density of θ. The Bayes medloss of an estimator δ under π is*

$$m_\pi(\delta) = \underset{\pi}{med}[med_{X^n} L(\delta(X^n),\theta)],$$

where $\underset{\pi}{med}[med_{X^n} L(\delta(X^n),\theta)]$ means the median of the medloss $med_{X^n} L(\delta(X^n),\theta)$ with respect to the prior π. Thus, an estimator δ_1 is better than an estimator δ_2 in the sense of Bayes medloss iff

$$m_\pi(\delta_1) \leq m_\pi(\delta_2).$$

The estimate which minimizes the Bayes medloss is called Bayes medloss estimator.

It will be seen in the examples below that the Frequentist *medloss* criterion and the Bayes *medloss* criterion lead to the same estimators in several cases. Moreover, in Section 4 we will see that the Frequentist *medloss* and the Bayes *medloss* are related by concepts of median admissibility and median minimaxity.

2.2.3. Posterior medloss criterion

Our third *medloss* criterion is based on the posterior distribution and is analogous to the posterior risk.

Definition 3. *An estimator $\delta_1(x^n)$ is better than an estimator $\delta_2(x^n)$ in the sense of posterior medloss iff*

$$\underset{\Theta|X^n}{med} L(\Theta, \delta_1(X^n)) \leq \underset{\Theta|X^n}{med} L(\Theta, \delta_2(X^n)),$$

where $\underset{\Theta|X^n}{med} L(\Theta, \delta(X^n))$ means the median of the loss $L(\Theta, \delta(X^n))$ with respect to the posterior distribution of $\Theta|X^n = x^n$, or called the posterior medloss.

The estimator which minimizes the posterior medloss is called the posterior medloss estimator, which is indeed the mid-point of the smallest interval on which the posterior probability is 1/2.

We will see in the next section that the Bayes *medloss* and posterior *medloss* criteria are not equivalent. Despite the logically satisfying relationship between Frequentist and Bayes *medloss* criteria, we will argue that the posterior *medloss* is more useful for inference. This will provide small sample properties to support the asymptotic results in Yu and Clarke [21].

3. Examples under *medloss* criteria

Here we present several examples under the three *medloss* definitions. For the Frequentist and Bayes *medloss* we look at the normal, Cauchy and exponential distributions. For contrast, we describe the general result to be given in Section 5 for the posterior *medloss* and give an example to show it is not equivalent to the Bayes *medloss* in this section. Note that this differs from the expected utility based form of decision theory in which an estimator minimizing the posterior risk also minimizes the Bayes risk. It is the non-equivalence of the Bayes and posterior *medloss* estimators that forces us to advocate one – the posterior *medloss* –over the other – the Bayes *medloss*.

3.1. Examples of Frequentist and Bayes medloss estimators

Ideally, we seek an estimator δ which minimizes the Frequentist *medloss*, $\text{med}_{X^n} L(\delta(X^n) - \theta)$, at every value of θ. However, this does not seem to be possible in general. Indeed, in expected-utility based decision theory it is common to examine restricted families of estimators in specific parametric families. For instance, Puri, Ralescu and Ralescu [11] examine the minimaxity of estimators of the form $aX + b$ in the context of one-dimensional exponential families. Ralescu and Ralescu [12] examine the admissibility of estimators of the form

$(aX+b)/(cX+d)$ in the same families and Lehmann [7] devotes a section to the special case of translation-scale families and equivariant estimators. Indeed, Hoffman [3] examines the linear regression model for Bayes linear estimators that are admissible or minimax; Rukhin [15] looks at equivariant estimators for estimating exponential quantiles; and Kubokawa [5] examines equivariant estimators in location-scale families.

So, for the sake of giving some simple examples we begin by looking at the normal and Cauchy distribution under the translation class of estimators and then look at the exponential distribution under the scale class of estimators. We first derive the Frequentist *medloss* estimators and then look at the Bayes *medloss* estimators. In section 4, we will relate the Frequentist and Bayes *medloss* estimators by the concepts of median admissibility and minimaxity and give further examples.

3.1.1. Frequentist medoss, Normal distribution, translation class

Let $X \sim N(\theta, 1)$. For comparison purposes, we find the best Frequentist *medloss* estimator for θ and the best expected loss estimator for the unknown normal mean parameter θ in the translation class

$$\mathcal{C}_1 = \{\delta^1_c(X) = X + c, \forall c \in \mathcal{R}\}$$

under the squared error loss L^2. To start, note that we have

$$\text{med}_X\, L^2(X+c,\theta) = \text{med}_X\, (\theta - (X+c))^2, \tag{4}$$

so the minimum over $c \in \mathcal{R}$ can be found as follows. Since $(X+c) - \theta \sim N(c,1)$, we see that

$$((X+c)-\theta)^2 \sim \chi^2_{1,\alpha(c)}, \tag{5}$$

where $\chi^2_{1,\alpha(c)}$ denotes the non-central χ^2 distribution with 1 degree of freedom and non-centrality parameter $\alpha = \alpha(c) = |c|$. Letting $F_{1,\alpha}$ be the d.f. for $\chi^2_{1,\alpha}$, we have

$$\begin{aligned} \alpha_1 > \alpha_2 &\Rightarrow F_{1,\alpha_1} \leq F_{1,\alpha_2} \\ &\Rightarrow \text{med}\, F_{1,\alpha_1} \geq \text{med}\, F_{1,\alpha_2}. \end{aligned} \tag{6}$$

In other words, the best *medloss* estimator for θ minimizes the non-centrality parameter α. So, we have $\forall c \in \mathcal{R}$ and $\forall \theta \in \Theta$,

$$\text{med}_X\, (\theta - X)^2 \leq \text{med}_X\, (\theta - (X+c))^2, \tag{7}$$

which gives that X is the (unique) best estimate for the normal mean θ in $\mathcal{C}_1$ under median loss with L^2. The uniqueness follows because 0 is the unique minimizer of $\alpha(c) = |c| \geq 0$. (If a scale class of estimators, i.e., cX, is used to estimate a normal mean, the solution depends on c and is treated in Yu [20].)

In the expected utility setting we have

$$EL^2(X+c,\theta) = E(\theta - (X+c))^2 = 1 + c^2, \tag{8}$$

it suffices to find the best expected-loss estimator for θ by minimizing the function $1 + c^2$ over $c \in \mathcal{R}$. In other words, for all θ,

$$c_0^* = 0 = \arg\min_{c \in \mathcal{R}} EL^2(\theta, X + c).$$

Thus,

$$\forall c \in \mathcal{R} \text{ and } \forall \theta \in \Theta, \quad E(\theta - X)^2 \leq E(\theta - (X + c))^2,$$

which gives that X is also the unique best estimate for θ in the translation class $\mathcal{C}_1$ based on the expected loss with L^2.

3.1.2. Frequentist medloss, Cauchy distribution, translation class

The Cauchy example, like our numerical example in the introduction dramatizes that the *medloss* approach will often be better for heavy-tailed distributions. Consider a translation class of estimators. Finding an optimal estimator under the usual definition of risk is impossible because $E(X + c - \theta)^2$ does not exist. By contrast, the Frequentist *medloss* estimator does not have any moment conditions and includes essentially all heavy-tailed distributions, for instance the Cauchy.

To solve the problem for the Cauchy distribution, it is enough to find $\tilde{c}$ to minimize the median of the distribution of $(X + c - \theta)^2$. Recall that $W = X - \theta$ follows the standard Cauchy distribution with density $f(w) = 1/[\pi(1 + w^2)]$. It is symmetric around its mode of 0. Let $F_{(W+c)^2}(\cdot)$ denote the distribution function of $(W + c)^2$. We have

$$\forall y \geq 0, \quad F_{(W+c)^2}(y) = P((W + c)^2 \leq y) = P(-\sqrt{y} - c \leq W \leq \sqrt{y} - c).$$

Thus, $\tilde{c} = \arg\max_{c} P(-\sqrt{y} - c \leq W \leq \sqrt{y} - c), \forall y \geq 0$, i.e. $\tilde{c} = 0$. In other words, for the Cauchy distribution, X is the uniformly optimal *medloss* estimator in the translation class under the squared-error loss. This remains true for any L^p loss with $p > 0$.

3.1.3. Freqeuntist medloss, Exponential distribution, scale class

In the last two symmetric examples, the Frequentist *medloss* and expected-loss criteria gave the same solutions. However, in the case of an exponential distribution for X, the estimators from the Frequentist *medloss* and the expected-loss criteria are very different. Indeed, the best expected-loss estimator for the exponential mean λ is $0.5X$, while the best Frequentist *medloss* estimator is $\approx 0.85X$. It is well known that the best unbiased estimator for λ is X. Thus, the best *medloss* estimator $0.85X$ is a tradeoff between the efficient and the best expected-utility estimators.

Let $X \sim Exp(\lambda)$ have density $f(x|\lambda) = \lambda^{-1}e^{-x/\lambda}$, where $x \geq 0$ and $\lambda > 0$. Consider a scale invariant loss $L(\delta(x), \lambda) = \frac{(\lambda - \delta(x))^2}{\lambda^2}$ in the scale class $\{\delta_c(X) = cX, \forall c \in \mathcal{R}^+\}$. Here we fix the support of $\delta_c(X)$ to be the same as that of X, i.e. $\mathcal{R}^+ \cup \{0\}$.

Again, we find the best expected-loss estimator and the best Frequentist *medloss* estimator. The first is simpler, so we start with it. The expected loss $EL(\delta_c(X), \lambda)$ on $\{\delta_c(X) = cX, \forall c \in \mathcal{R}^+\}$ is $E\left[\frac{(\lambda - cX)^2}{\lambda^2}\right] = c^2 + (c - 1)^2$, which is a convex function of c with the unique minimizer at $c^* = 1/2$ for all λ. Thus, $X/2$ is the unique best estimate for the exponential mean λ under the scale invariant loss L in the scale class $\{\delta_c(X) = cX, \forall c \in \mathcal{R}^+\}$.

Now we obtain the Frequentist *medloss* estimator. Consider $\text{med}_X\, L(\delta_c(X), \lambda)$, let $Y_c = \frac{(\lambda - cX)^2}{\lambda^2}$ and let m_c be the median of Y_c. i.e. $m_c = \text{med}_X \frac{(\lambda-cX)^2}{\lambda^2}$. Again we want to find $\tilde{c}$ such that $m_{\tilde{c}} < m_c$ for any $c > 0$ and all $\lambda > 0$. Let $W = cX \sim Exp(c\lambda)$, we have that

$$\begin{aligned}\frac{1}{2} = P(Y_c \le m_c) &= P(\lambda - \lambda\sqrt{m_c} \le W \le \lambda + \lambda\sqrt{m_c}) \\ &= \begin{cases} e^{(\sqrt{m_c}-1)/c} - e^{(-\sqrt{m_c}-1)/c}, & for\, m_c < 1 \\ 1 - e^{(-\sqrt{m_c}-1)/c}, & for\, m_c \ge 1. \end{cases}\end{aligned} \tag{9}$$

The form of m_c can be found by examining two cases.

Case 1: For $m_c < 1$, let $\triangle = e^{\sqrt{m_c}/c} > 0$, so we have $0.5e^{1/c} = \triangle - \triangle^{-1}$, i.e. we solve for $\triangle$ from $\triangle^2 - 0.5e^{1/c} \triangle -1 = 0$. By some algebra, we get

$$\triangle = (e^{1/c} + \sqrt{e^{2/c} + 16})/4,$$

from which we can derive $m_c = A(c)$, where $A(c) = c^2\{\ln[(e^{1/c} + \sqrt{e^{2/c}+16})/4]\}^2$.

Case 2: For $m_c \ge 1$, we can derive $m_c = B(c)$, where $B(c) = (c \ln 2 - 1)^2$ by a technique similar to case 1.

From case 1 we get $0 < A(c) < 1$ and from case 2 we get $B(c) \ge 1$. It can be verified computationally that $A(c)$ and $B(c)$ cross each other at $c_0 = 2/\ln 2 \approx 2.8854$ and that $A(c_0) = 1 = B(c_0)$. Thus, we have the following expression for m_c.

$$m_c = A(c)I_{(0<c\le c_0)} + B(c)I_{(c>c_0)}. \tag{10}$$

The minimum of m_c in (10) is $\tilde{c} = \arg\min_c m_c = \arg\min_c \left[A(c)I_{(0<c\le c_0)} + B(c)I_{(c>c_0)}\right]$. Approximating $\tilde{c}$ numerically gives $\tilde{c} \approx 0.8498$. It is clear that $\tilde{c} \ne c^* = 1/2$.

To conclude, we observe that both $A(c)$ and $B(c)$ are independent of λ and hence so is $\tilde{c}$. Thus $\tilde{c}X$ is uniformly optimal for λ under L in the scale class $\{\delta_c(X) = cX, \forall c > 0\}$. Therefore, $\text{med}_X \frac{(\lambda-\tilde{c}X)^2}{\lambda^2} \le \text{med}_X \frac{(\lambda-cX)^2}{\lambda^2}$, for any $c > 0$ and all $\lambda > 0$.

3.1.4. Bayes medloss estimators

In the last subsections, we found that X was the Frequentist *medloss* estimator for the normal mean and Cauchy location. If we now look at the Bayes *medloss* criterion for those cases we find that again X is the Bayes *medloss* estimator. This follows because for the normal and Cauchy

$$\text{med}_{X^n} L(\delta(X^n), \theta) = \text{constant as a function of } \theta.$$

Thus, the inner median in Definition 2 is trivial—just the constant above. Similarly, in the exponential case, $\tilde{c}X$ is the Frequentist *medloss* estimator from Definition 1 as well as the optimal Bayes *medloss* estimator from Definition 2.

3.2. Posterior medloss examples

Posterior *medloss* estimators are easier to construct in general. Indeed, in section 5, we show that when the prior and parametric family are symmetric and the posterior is unimodel, the posterior *medloss* estimator is the median of the posterior, the same as would be found under absolute error loss. An important related point is that the Bayes *medloss* estimators are in general different from the posterior *medloss* estimators. This is reasonable because as already noted, Bayes *medloss* estimators can only be derived under narrow settings while posterior *medloss* estimators can be derived more generally.

This non-equivalence between Bayes and posterior *medloss* estimators differs from the usual expected-utility setting. Recall that the standard way to find a Bayes estimator under the usual risk criterion is to use the iterated expectation. That is, $E_{\theta,X^n}L(\delta,\theta) = E_m E_{\theta|X^n}L(\delta,\theta)$ and it is enough to minimize the inner conditional expectation pointwisely in X^n to find an optimal estimator. By contrast, it is not hard to give an example for which the Bayes and posterior *medloss* estimators differ. A particularly incisive one was provided by an anonymous referee and is as follows.

Let $X|\theta \sim$ Bernoulli(θ) and $\theta \in \{0.4, 0.5, 0.6\}$ and set the prior distribution of θ be

$$P(\theta = 0.4) = 0.275, \quad P(\theta = 0.5) = 0.45, \quad P(\theta = 0.6) = 0.275.$$

Now, the posterior distribution after observing $X = 0$ is

$$P(\theta = 0.4\,|X = 0) = 0.33, \quad P(\theta = 0.5\,|X = 0) = 0.45, \quad P(\theta = 0.6\,|X = 0) = 0.22.$$

and after observing $X = 1$ is

$$P(\theta = 0.4\,|X = 1) = 0.22, \quad P(\theta = 0.5\,|X = 1) = 0.45, \quad P(\theta = 0.6\,|X = 1) = 0.33.$$

Under absolute error loss, the posterior *medloss* is minimized by $\tilde{\delta}$ where $\tilde{\delta}(X = 0) = 0.45$ or $\tilde{\delta}(X = 1) = 0.55$, and 0.05 is the value of the posterior *medloss*. Thus, $\tilde{\delta}$ is the posterior *medloss* estimator.

On the other hand, since $P(X = 0\,|\theta = 0.4) = 0.6$ and $P(X = 1\,|\theta = 0.6) = 0.6$,

$$\underset{X\,|\theta=0.4}{\mathrm{med}}\, L(\theta,\, \delta(X)) \quad \text{and} \quad \underset{X\,|\theta=0.6}{\mathrm{med}}\, L(\theta,\, \delta(X)) \tag{11}$$

achieve their minimum (zero) at $\delta^*(X = 0) = 0.4$ and $\delta^*(X = 1) = 0.6$, respectively. It is easy to see that $\delta^*(X = 0)$ and $\delta^*(X = 1)$ also minimize the Bayes *medloss*, i.e. $\underset{\pi(\theta)}{\mathrm{med}}\Big(\underset{X\,|\theta}{\mathrm{med}} L(\theta,\, \delta(X))\Big)$, with a value of 0, respectively. Thus the Bayes and posterior *medloss* estimators are different.

4. Frequentist and Bayes *medloss* via median admissibility and minimaxity

In this section, we redefine admissibility and minimaxity with respect to *medloss* criteria. Several lemmas parallel to the conventional expected-loss based results can be readily given. Then, it is important to give a few examples of median admissibility and minimaxity before establishing the median-inadmissibility of the least squares estimator under a linear model.

4.1. Definitions and results.

First, we give two definitions under the Frequentist *medloss* criterion.

Definition 4. *An estimator $\delta(X^n)$ is median-inadmissible, or m-inadmissible, if there exists an estimator $\delta^*(X^n)$ such that $med_{X^n}L(\delta^*(X^n),\theta) \leq med_{X^n}L(\delta(X^n),\theta)$ for all θ, with strict inequality for some θ. By contrast, it is median-admissible if $\delta^*(X^n)$ does not exist.*

Definition 5. *An estimator $\delta^{**}(X^n)$ is a median-minimax, or m-minimax, decision rule if it minimizes $\sup\limits_{\theta} med_{X^n}L(\delta(X^n),\theta)$ among all estimators $\delta(X^n)$ in the decision space $\mathcal{D}$, i.e.*

$$\sup_{\theta} med_{X^n}L(\delta^{**}(X^n),\theta) = \inf_{\delta}\sup_{\theta} med_{X^n}L(\delta(X^n),\theta).$$

These parallel the conventional definitions under the expected utility. Now we have the following propositions; they are the natural extensions of results in Lehmann [7].

Proposition 1. 1. *If $\delta^*(X^n)$ is a unique m-minimax rule, then it is also m-admissible.*
2. *If $\delta^*(X^n)$ is a m-admissible rule with constant medloss, then $\delta^*(X^n)$ is also a m-minimax rule.*
3. *For any given prior $\pi(\theta)$, if the Bayes medloss estimator $\delta^{\pi}(X^n)$ is unique, then it is m-admissible.*
4. *Suppose that $\delta^{\pi^*}(X^n)$ is the Bayes medloss estimator with respect to a specific prior $\pi^* \in \Pi$. Then it is also m-minimax if it has a constant medloss, i.e. $med_{X^n}L(\delta^{\pi^*}(X^n),\theta) = m(\delta^{\pi^*},\theta) = \rho^*$, for all θ.*

Proposition 2. *Let $\{\pi_k : k \geq 1\}$ be a sequence of priors π_k on Θ. Denote the sequences of the corresponding Bayes medloss estimators and their Bayes medloss by $\{\delta_k(X^n) : k \geq 1\}$ and $\{m_{\pi_k}(\delta_k) : k \geq 1\}$, respectively. Suppose that $\delta^*(X^n)$ is an estimator for θ and its medloss satisfies*

$$\sup_{\theta\in\Theta} m(\delta^*(X^n),\theta) \leq \lim_{k\to\infty} m_{\pi_k}(\delta_k). \tag{12}$$

Then, $\delta^*(X^n)$ is m-minimax.

Proof. Assume that $\delta^*(X^n)$ is not *m-minimax*, i.e. $\exists\tilde{\delta}(X^n)$ such that

$$\sup_{\theta} m(\tilde{\delta}(X^n),\theta) < \sup_{\theta} m(\delta^*(X^n),\theta). \tag{13}$$

By the definitions of δ_k and $m_{\pi_k}(\delta_k)$ for $k \geq 1$, we have

$$m_{\pi_k}(\delta_k) \leq m_{\pi_k}(\tilde{\delta}) = \operatorname*{med}_{\pi_k} m(\tilde{\delta}(X^n),\theta) \leq \sup_{\theta} m(\tilde{\delta}(X^n),\theta).$$

Thus, by (13), we get $m_{\pi_k}(\delta_k) < \sup\limits_{\theta} m(\delta^*(X^n),\theta)$. Then taking limit on both sides, we have

$$\lim_{k\to\infty} m_{\pi_k}(\delta_k) < \sup_{\theta} m(\delta^*(X^n),\theta),$$

which contradicts the condition (12). So, $\delta^*(X^n)$ is m-minimax. □

Corollary 1. *If $\delta^*(X^n)$ is an estimator for θ with constant medloss, say $m(\delta^*(X^n), \theta) = \rho^*$, for all θ, and there exists a sequence of prior $\{\pi_k\}$ such that the corresponding Bayes medloss estimators $\delta_k(X^n)$ has Bayes medloss $m_{\pi_k}(\delta_k)$ satisfying*

$$\lim_{k\to\infty} m_{\pi_k}(\delta_k) = \rho^*. \tag{14}$$

Then, $\delta^*(X^n)$ is m-minimax.

4.2. Examples of median-admissible and median-minimax estimators

For the translation class of estimators $X + c$, where $X \sim N(\theta, 1)$ and $c \in \mathcal{R}$, we found that X is the Bayes *medloss* estimator in the class and it is unique. So, by Proposition 1, X is m-admissible. Note that the *medloss* of X, i.e. $med(X - \theta)^2$ is the median of a chi-squared distribution with 1 degree of freedom , which is approximately equal to 0.4549. Thus, by Proposition 1, X is also m-minimax.

More generally, if we consider a translation class of estimators $X + c$ of an unknown parameter θ and X is from a location family, then $X - \theta$ is distribution-free. So, the *medloss* of $X + c$, say $m_c = \text{med}_X\, |X + c - \theta|$, is also distribution-free. Thus, the Frequentist *medloss* estimator based on Definition 1 and the Bayes *medloss* estimator based on Definition 2 coincide in this setting, and the Bayes *medloss* estimator has a constant *medloss*, which implies that the estimator is also m-minimax, by Proposition 1. If the Bayes *medloss* estimator is unique, then it is also m-admissible. This also happens when the distribution of X is symmetric about the location parameter θ.

Referring to the example of a scale class of exponential distributed random variables, $\{cX : X \sim Exp(\lambda),\ c \geq 0\}$, with the scale invariant loss L, we showed that $\tilde{c}X$ is a Bayes *medloss* estimator with *medloss*

$$m_{\tilde{c}} = A(\tilde{c}),$$

where $\tilde{c} \approx 0.8498$. So, $\tilde{c}X$ is m-admissible because of its uniqueness, and is m-minimax because it has a constant *medloss*.

This result can be generalized to a scale class of estimators cX, where X is from a scale family with a scale parameter θ i.e. X/θ is distribution-free. Thus, the Bayes *medloss* estimator satisfies

$$\min_c \ \underset{\pi(\theta)}{\text{med}} \ \text{med}_X \, \frac{(cX - \theta)^2}{\theta^2} = \min_c \ \text{med}_X \, (cX/\theta - 1)^2.$$

Now, it can be seen that if there is only one value of c, say $\tilde{c}$, minimizing $\text{med}_X\,(cX/\theta - 1)^2$, then the corresponding Bayes estimator $\tilde{c}X$ is both m-admissible and m-minimax by Proposition 1. However, it is m-admissible but not m-minimax if we replace the scale invariant loss by the squared error loss, i.e. $(cX - \theta)^2$. This is so because we will have $\sup_\theta \theta^2 = \infty$, which implies that $\sup_\theta \text{med}_X\,(\tilde{c}X - \theta)^2 = \sup_\theta \theta^2\, \text{med}_X\,(\tilde{c}X/\theta - 1)^2$ does not exist.

4.3. Median-Inadmissibility for linear models

Since the Frequentist *medloss* criterion is weaker then the UD or SD criteria introduced in Hwang [4], the results in Hwang [4] for his U-admissibility defined under UD and SD continue to hold for median-admissibility. Following Hwang [4], we can verify the median-inadmissibility of the least squares estimator for linear models. Indeed, our first result is actually the direct consequence of Theorem 3.4 in Hwang [4]. Consider a p-dimensional random vector X with its median θ. Let $\delta_a(X)$ be the James-Stein positive part estimator, i.e.

$$\delta_a(X) = \Big(1 - \frac{a}{\|X\|^2}\Big)_+ X,$$

where $y_+ = \max\{y, 0\}$ and $\|z\|^2 = z^T z$ is the Euclidean norm. So, we have the following.

Proposition 3. *Assume that X has a density f in form of $f(\|x-\theta\|^2)$ and $f'(s)/f(s)$ is defined for all $s \in [\alpha_0, \alpha_1]$, where for some $c > 0$ and $a > 0$,*

$$\alpha_0 = (c - \sqrt{a})_+^2 \text{ and } \alpha_1 = c^2 + a.$$

If

$$\frac{-(p-2)a^{-1/2}}{2c} \ln\{[c + (c^2+a)^{1/2}]a^{-1/2}\} \leq \inf_{s\in(\alpha_0,\alpha_1)} f'(s)/f(s), \tag{15}$$

then for every θ,

$$P_\theta(\|\theta - \delta_a(X)\| \leq c) > P_\theta(\|\theta - X\| \leq c). \tag{16}$$

Clearly, if (16) is satisfied for $c = \text{med}_X\|\theta - X\|$, then $\delta_a(X)$ is better than (dominates) X under the *medloss* criterion under Euclidean error. Thus, we have the following.

Corollary 2. *With the assumptions and notations of Proposition 3 above, if there exists a real constant a such that (15) is satisfied for $c = med_X\|\theta - X\|$, then $\delta_a(X)$ dominates X under the medloss criterion.*

As an example, Hwang [4] considers the case that $X - \theta$ has a p-variate t distribution with N degrees of freedom, and find that for every N and $p \geq 3$, X is inadmissible under the Euclidean error in the sense of UD, and the James-Stein positive part estimator $\delta_a(X)$ universally dominates X under Euclidean error if $a > 0$ satisfies

$$\frac{p-2}{(N+a)^{1/2}a^{1/2}} \ln\{[(N+a)^{1/2} + (N+2a)^{1/2}]a^{-1/2}\} \geq (N+p)/N. \tag{17}$$

This result holds for all c, which implies that $\delta_a(X)$ dominates X under the *medloss* criterion. So, we have the following.

Corollary 3. *Assume that $X-\theta$ has a p-variate t distribution with N degree of freedom. If $a>0$ satisfies (17), then under the Euclidean error, X is median-inadmissible for every N and $p\geq 3$, and is dominated by $\delta_a(X)$ in the medloss sense.*

Using Corollary 3, we can show the median-inadmissibility of the least-squares estimator $\hat{\delta}^{LS}$ for θ in a linear model. Consider the linear model $X=A\theta+e$, where A is an $m\times p$ known design matrix with full rank p and e/σ has a t distribution with N degrees of freedom, where σ is known. For this model, we have the following.

Corollary 4. *If a satisfies (17), then for every N and $p\geq 3$, the least-squares estimator $\hat{\delta}^{LS}=(A^TA)^{-1}A^TX$ is median-inadmissible and is dominated by*

$$\delta(X)=\Big(1-\frac{a\sigma^2}{\|(A^TA)^{1/2}\hat{\delta}^{LS}\|^2}\Big)_+\hat{\delta}^{LS}, \tag{18}$$

in the medloss sense, under the generalized Euclidean error with the norm with respect to $\|\theta-\delta(X)\|_{A^TA}$.

5. Posterior medloss estimation

Here, we have proposed two definitions for *medloss* criteria in the Bayesian setting. In Section 3.2, we saw that the best estimators based on those criteria are not in general the same. While the Bayes *medloss* and the Frequentist *medloss* criteria are interesting variants on the expected-utility version of decision theory, we argue that the posterior *medloss* is more reasonable for general use in finite samples because it represents an optimization over a larger class of functions. Moreover, while all of the *medloss* estimators have the same type of loss-insensitivity, the posterior *medloss* also has the practical advantage of being easily computed in routine cases. Indeed, we will also show the claim made in Section 3.2, namely that the posterior *medloss* reduces to the median of the posterior in well-behaved cases. We conclude this section with a simulation verifying that the posterior *medloss* estimator can exhibit the kind of robustness to outliers that the usual Bayes estimator does not.

5.1. Robustness of the posterior medloss estimator to the loss function

In general, the usual risk-based estimator is invariant up to positive affine transformations of the loss. In addition, under the assumptions of unimodality and symmetry of the posterior density, Rukhin [14] showed that the posterior expected-loss estimator is only invariant up to the choice of even loss functions when the posterior risk is finite. Here we show that when the posterior density is continuous, the posterior *medloss* estimator is invariant up to any strictly increasing transformation of the absolute-error loss, which implies that the posterior *medloss* estimate has a higher loss robustness than the usual posterior expected-loss estimate does, under weaker assumptions on the posterior. This loss robustness also holds for median-admissible, median-minimax, and Bayes *medloss* estimators. Because of the similarity of these statements, it is enough to establish the result for the posterior *medloss* estimators.

Theorem 1. *Define*

$$\delta_1(x^n) = \arg\min_{d} \underset{\pi(\Theta|x^n)}{med} |\Theta - d(x^n)|,$$

where $\pi(\Theta|x^n)$ is any continuous posterior density of Θ given x^n. Assume that the median of the loss $\mathcal{L}$ is unique. Then for any strictly increasing functions $\mathcal{L}$ of $|\Theta - d(x^n)|$, we have

$$\delta_1(x^n) = \arg\min_{d} \underset{\pi(\Theta|x^n)}{med} \mathcal{L}(|\Theta - d(x^n)|).$$

i.e. all $\underset{\pi(\Theta|x^n)}{med} \mathcal{L}(|\Theta - d(x^n)|)$ attain their minimum at the same point $\delta_1(x^n)$.

Proof. let $\Pi^{-1}(\cdot|x^n)$ be the inverse of the continuous posterior distribution $\Pi(\cdot|x^n)$ of Θ given x^n, and consider the *medloss* for any strictly increasing loss function $\mathcal{L}$ of $|\Theta - d(x^n)|$,

$$m\mathcal{L} = \underset{\pi(\Theta|x^n)}{\text{med}} \mathcal{L}(|\Theta - d(x^n)|).$$

Here $m\mathcal{L} = m\mathcal{L}(d, x^n)$. By the property of median that $\text{med}(h(X)) = h(\text{med}(X))$ for strictly increasing functions h, we have

$$m\mathcal{L} = \mathcal{L}\Big(\underset{\pi(\Theta|x^n)}{\text{med}} |\Theta - d(x^n)|\Big) = \mathcal{L}(m_1),$$

so the median of the absolute error loss $m_1 = m_1(d, x^n) = \mathcal{L}^{-1}(m\mathcal{L})$, where $\mathcal{L}^{-1}$ is the inverse of $\mathcal{L}$. By the definition of $m\mathcal{L}$,

$$\frac{1}{2} = \Pi(\mathcal{L}(|\Theta - d(x^n)|) \leq m\mathcal{L} \quad |x^n) = \Pi(|\Theta - d(x^n)| \leq m_1 \quad |x^n).$$

□

5.2. General procedure for computing the posterior medloss estimator

Now we provide a general procedure for computing the posterior *medloss* estimators when the posterior density is continuous.

Let $\Xi = \{(a, b) \in \mathcal{R}^2 : \Pi(a \leq \Theta \leq b \quad |x^n) = 1/2\}$. Then for any $(a, b) \in \Xi$, let $a = d(x^n) - m_1$ and $b = d(x^n) + m_1$. After some algebra, we get

$$\begin{cases} d(x^n) = (b + a)/2 \\ m_1 = (b - a)/2. \end{cases}$$

These expressions suggest the following general procedure to find posterior *medloss* estimates.

1. Find the posterior distribution function $\Pi(\cdot|x^n)$ and its inverse $\Pi^{-1}(\cdot|x^n)$.
2. Define

$$a^* = \inf\{a : \Pi(a) > 0\} \text{ and } b^* = \sup\{b : \Pi(b) < 1\}.$$

Consider a sequence of $a_i \in (-\infty, m]$, where m is the posterior median, i.e. $m = \Pi^{-1}(1/2\,|x^n) \in (a^*, b^*)$. Thus, we can pick any finite number $r \in (a^*, m]$, and then define $a_i = r + si$ with step size s for $i = 1, \ldots, (m-r)/s$.

3. Define $b_i = \Pi^{-1}[1/2 + \Pi(a_i|x^n)|x^n]$.
4. Let $d_i(x^n) = (b_i + a_i)/2$. The corresponding value of the median loss is $m_1(i) = (b_i - a_i)/2$.
5. Find the minimum median $\tilde{m}_1$ in $\{m_1(i)\}_{i=1,\ldots,(m-r)/s}$. The corresponding $\tilde{d}(x^n)$ is the value of the posterior *medloss* estimate.

This procedure gives good performance in routine examples. Consider

$$\begin{cases} X \sim Poisson(\lambda) \\ \lambda \sim Ga(\alpha, \beta) \end{cases},$$

where the Poisson density is given by $P(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$ equipped with the conjugate Gamma prior for convenience. Thus, the posterior distribution for λ given $X = x$ is

$$\lambda|X = x \sim Ga((\alpha + x), \frac{\beta}{1+\beta}).$$

To compare the risk-based and *medloss*-based estimators, we must first find them. For the risk, under $L^2 = (\lambda - \delta(X))^2$, the optimal rule $\delta_2^*(x)$ is the posterior mean, i.e.

$$\delta_2^*(x) = E(\lambda|X = x) = \tfrac{\beta}{\beta+1}(x + \alpha).$$

For the *medloss*, we do not have a closed form for

$$\tilde{\delta}_2(x) = \arg\min_{\delta} \operatorname*{med}_{\pi(\lambda|x)} |\lambda - \delta(x)|^2,$$

however, our generic algorithm can be used to generate the curves in Figure 1.

For the fixed value $x = 2$ and a Gamma(2, 3) prior, we get $\delta_2^*(x) = 3$ and $\tilde{\delta}_2(x) \approx 2.3$. Figure 1 shows the curves for the risk and *medloss* as the value of δ varies.

5.3. Closed form of the posterior medloss estimator for unimodal, symmetric posteriors

As mentioned before, Rukhin [14] showed that the posterior risk estimator is invariant up to even loss functions if the posterior density is symmetric and strictly unimodal, and the posterior risk of this estimator is finite. Parallel to Rukhin's result, we show that for any strictly increasing function of the absolute error loss L^1 for the *medloss* the posterior *medloss* estimator reduces to the median of the posterior distribution.

Theorem 2. *If the posterior density is continuous, symmetric and decreases away from its median, then the posterior median is the posterior medloss estimator with respect to any strictly increasing function of the absolute error loss L^1.*

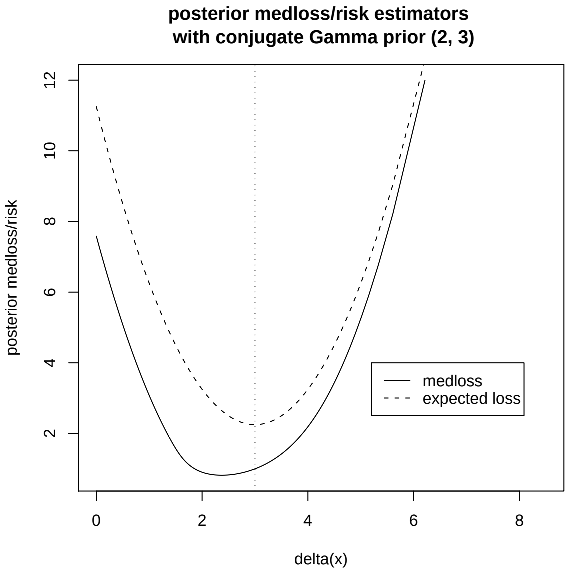

Figure 1: The posterior risk and *medloss* as a function of $\delta(x)$. The dashed curve represents the posterior risk for L^2, while the solid one is for the posterior *medloss* under L^2. The dotted vertical line shows the value of the posterior mean.

Theorem 2 is a direct consequence of the following two lemmas for any continuous and symmetric density that decreases away from its median.

Lemma 1. *If the density of Z is symmetric and decreases away from its median med(Z), then med(Z) maximizes $P(|Z-a| \le k)$ over a, for any $k \ge 0$. In other words,*

$$\textit{med(Z)} = \textit{arg} \sup_a P(|Z-a| \le k), \textit{ for any } k \ge 0.$$

Proof. Without loss of generality, assume that $\mathrm{med}(Z) = 0$ and $a > 0$. Thus,

$$\begin{aligned} P(|Z-a| \le k) = P(a-k \le Z \le a+k) &= \int_{-k}^{k} dF(z) + \int_{k}^{k+a} dF(z) - \int_{-k}^{a-k} dF(z) \\ &= \int_{-k}^{k} dF(z) + \int_{k}^{k+a} dF(z) - \int_{k-a}^{k} dF(z) \\ &\le \int_{-k}^{k} dF(z) = P(|Z-0| \le k). \end{aligned}$$

□

Lemma 2. *If Z has a symmetric distribution with unique median, then its median, med(Z), is the unique minimizer of the medloss for any strictly increasing function of $L^1 = |Z-m|$ loss for any $m \in R$.*

Proof. It suffices to show that $\text{med}(Z)$ is the unique minimizer of the *medloss* with L^1 loss. Then by Lemma 1, we have

$$P(|Z - \text{med}(Z)| \leq k) > P(|Z - a| \leq k), \text{ for any } k \geq 0 \text{ and any } a.$$

Now let $k = \text{med}(|Z - a|)$. Then the right hand side of the above inequality is 1/2, so $P(|Z - \text{med}(Z)| \leq \text{med}(|Z - a|)) > 1/2$, i.e.

$$\text{med}(|Z - \text{med}(Z)|) < \text{med}(|Z - a|), \text{ for all } a.$$

Thus, $\text{med}(Z)$ is the unique minimizer of the *medloss* with L^1, or any L^p loss function. □

5.4. Comparison of the usual Bayes estimator and posterior medloss estimator

To complete our development of the posterior *medloss* estimator we have done a simulation study to show that it performs well in a standard example. In particular, we argue that the posterior *medloss* estimators compared to posterior means or posterior medians have two sorts of desirable robsutness. The first follows from Rostek's axioms, see Rostek [13] that include conditions amounting to requiring the tail behavior of the likelihood not to matter very much. We verify that this leads to the anticipated robustness properties of the posterior *medloss* estimators by considering the case that data come from one model, the data generator (DG), but this is not known to the analyst who analyzes the data using a different model. The DG is chosen to have slightly heavier tails than the data analysis (DA) model. This is much more representative of the typical inference task than assuming the DG model is tighter than the DA model. The second sense of robustness is more direct: The posterior *medloss* estimator does not change as much as the data changes as the posterior mean or posterior median does.

To begin, we compare the optimal estimators based on the expectation and posterior *medloss* criteria under the absolute error loss L^1 and the squared error loss L^2. Our procedure has six steps. First, let $b = \{10, 11, \ldots, 20\}$ be a vector of values to be taken as true. Second, let the DG be $\text{Gamma}(a_0, b)$ so we generate n independent observations from $X \sim \text{Gamma}(a_0, b)$, where a_0 is a fixed shape parameter. Third, assume that the prior distribution on the scale parameter b is inverse gamma $\text{I.Ga}(\alpha, \beta)$ with mode $= [(\alpha + 1)\beta]^{-1} = 15$ with fixed α and β, so that the true values of b are around 15. Fourth, let the DA model be $X \sim \text{Gamma}(a_1, b)$, so that if $a_0 > a_1$, the DG has a heavier tail than DA model. Fifth, based on these assumptions on the priors and data distributions, we have the posterior distribution $b|X = x \sim \text{I.Ga}(\alpha^1, \beta^1)$, where $\alpha^1 = na_1 + \alpha$, and $\beta^1 = \beta[1 + \beta \sum_{i=1}^n x_i]^{-1}$. Sixth, we find the usual Bayes estimate using the posterior mean (L^2 case) or the posterior median (L^1 case) and the posterior *medloss* estimator under L^2 and L^1. Note that these two estimators are in fact the same, see Theorem 2.

For each b, we repeat steps 2 - 6 N times giving N Bayes and N posterior *medloss* estimates under L^1 loss and L^2 loss. Thus, there are effectively 4 sequences of estimators, $\hat{b}_{i,L,T}$ where $i = 1, \ldots, N$, L indicates the loss, L^2 or L^1 and T indicates the type of estimator, posterior Bayes or posterior *medloss*. For each sequence of N estimates, we calculate the sample expected value $\hat{b}^E_{L,T}$ and the sample median $\hat{b}^M_{L,T}$. This gives a total of 8 different estimates for each b and N. Finally, we find the 8 relative errors $|\hat{b}^E_{L,T} - b|/b$ and $|\hat{b}^M_{L,T} - b|/b$

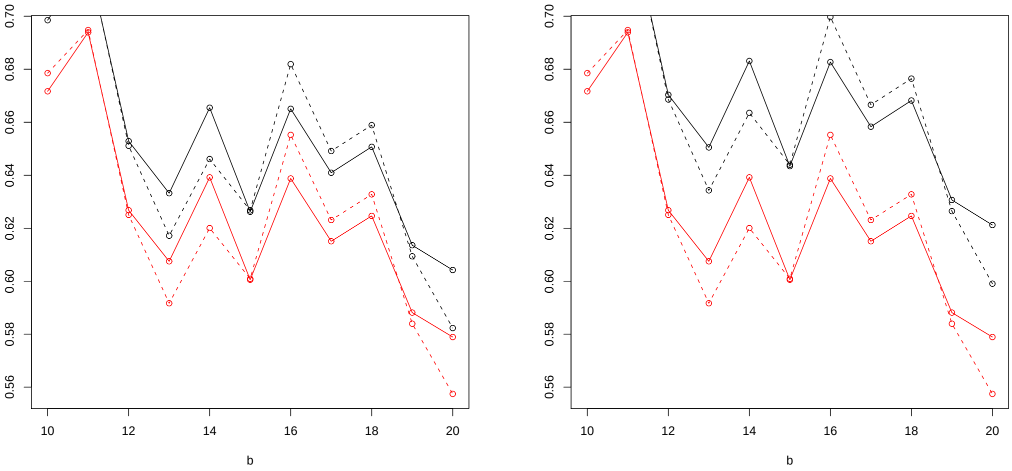

Figure 2: The dark lines are for the Bayes estimators and the light lines for the posterior *medloss* estimators. The solid lines represent the sample means of N realizations of the Bayes and posterior *medloss* estimators; while the dashed lines represent the sample medians of N realizations of the Bayes and posterior *medloss* estimators. The panel on the left use L^1 loss, and the panel on the right use L^2 loss. Since the DG has a heavier tail than the DA model, it is seen that the dark lines are always above the light lines indicating the posterior *medloss* estimators have smaller relative errors than the Bayes estimators.

for each choice of L and M, for each b. Using the relative error puts the comparisons on a common scale to ensure their validity.

In our simulation, we chose $\alpha = 4$, $\beta = 1/75$, $n = 20$, $N = 50$, and $(a_0, a_1) = (5, 3)$. When $a_0 = a_1$, the DG and DA have equally heavy tails and computations not shown here indicate that for $b < 15$ the light lines are below the dark lines indicating that the posterior medloss estimators have lower relative errors. Once $b > 15$, however, the two estimators appear roughly equally good. When $a_0 > a_1$, the DG has a heavier tail than the DA model and the results are seen in Figure 2. This shows the first sense of robustness: When the tails of the DG are heavier than the tails of the DA model that the posterior *medloss* estimators are much better. They are insensitive to tail behavior as suggested by the Rostek's axioms, and hence robust to outliers. The result is independent of the loss function as suggested by Theorem 2. Note that these statements are independent of whether we use the sample mean or sample median in the relative error.

To demonstrate a second sense of robustness, consider a single data point $X = x$ drawn from a Gamma(a, b) distribution. Since we are looking only at the estimator of b as a function of x, we can ignore the specific value of b. and plot $\hat{b}_1$, $\hat{b}_2$ and $\hat{b}_3$ defined to be the posterior mean (optimal under L^2 risk), the posterior median (optimal under L^1 risk) and the posterior *medloss* estimator (optimal under Def. 3). Let the hyperparameters again be $\alpha = 4$ and $\beta = 1/75$. The three estimators are shown for $a_1 = 1$ and $a_1 = 10$ in the panels on the left side of of Figure 3; other values are qualitatively similar. It is seen that one curve is above the other, but our interest is on the slope not the absolute value. Thus, in the right hand panels we have translated the top two lines so that the heads of the original bottom line and the two translated lines coincide. It is seen that the slope of the posterior *medloss* estimator is the smallest for for both cases and the slope of the posterior mean is the largest in both cases. The posterior median is between the other two. Overall this suggests the posterior *medloss* estimator is a little more robust against outliers or heavy tails without being genuinely insensitive.

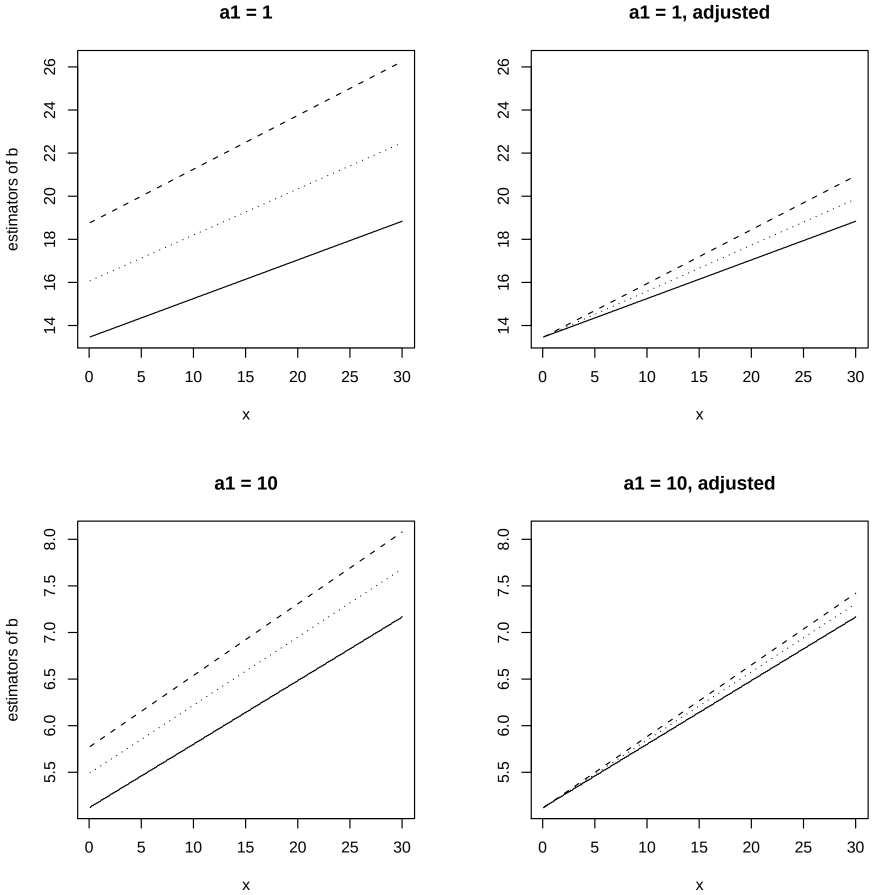

Figure 3: The left hand panels show the posterior mean (dashed line), the posterior median (dotted line) and the posterior *medloss* (solid line) for the indicated values of a_1. The right hand panels show the adjusted values of the same estimators. The posterior mean changes the most with a change in x, the posterior *medloss* changes the least and the posterior median is in between.

6. Discussion

In this paper we have developed the rudiments of a decision theory based on the median rather than the mean. While the Frequentist *medloss* criterion and the Bayes *medloss* criterion can be related via concepts such as median admissibility and median minimaxity, they seem, like their expected-utility counterparts to be applicable only in narrow settings. However, unlike the relationship between the Bayes risk and the posterior risk, the Bayes *medloss* and posterior *medloss* criteria are very different. From the development here, it can be seen that the posterior *medloss* can be optimized over a larger class of actions than either the Frequentist or Bayes *medloss* criteria. Indeed, the posterior *medloss* estimator is easier to compute and is asymptotically consistent, normal and efficient, see Yu and Clarke [21]. As seen in the example of Section 5, the posterior *medloss* is robust against heavy tails or outliers in two senses as well as having the usual loss-insensitive robustness of median based criteria.

We suggest that median based inference via the posterior *medloss* criterion is a fully viable alternative to the conventional expected-utility or risk based estimators commonly used. Moreover, median-based methods do not require moment conditions; they are applicable to any distribution. Our initial findings also suggest that median-based methods can, in some respects, outperform classical decision theory methods for parameter estimation in finite samples as well as being fully competitive asymptotically.

Acknowledgments

This research was supported by by NSERC operating grant number RGPIN 138122.

References

[1] Allais, M., 1953. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de lecole Americaine. Econometrica. 21, 503-546.

[2] Ellsberg, D., 1961. Risk, Ambiguity, and the Savage Axioms. Quarterly Journal of Economics. 75, 643-669.

[3] Hoffman, K. (1996) A subclass of Bayes linear estimators that are minimax. Acta Applicandae Mathematicae, 43, 87-95.

[4] Hwang, J.T., 1985. Universal Domination and Stochastic Domination: Estimation Simultaneously Under a Broad Class of Loss Functions. Ann. Statist. 13, 295-314.

[5] Kubokawa, T. 2004. Minimaxity in estimation of restricted parameters. J. Jap. Stat. Soc. Vol. 34-2, 1-19.

[6] Keating, J.P., Mason, R.L., Sen, P.K., 1993. Pitman's measure of closeness: a comparison of statistical estimators. Society for Industrial and Applied Mathematics. Philadelphia.

[7] Lehmann, H.L., 1983. Theory of Point Estimation, 1st ed. New York, Wiley.

[8] Machina, M., Schmeidler, D., 1992. A More Robust Definition of Subjective Probability. Econometrica. 60, 745-780.

[9] Manski, C.F., 1988. Ordinal Utility Models of Decision Making Under Uncertainty. Theory and Decision. 25, 79-104.

[10] Pitman, E.J.G., 1937. The "closest" estimates of statistical parameters. Proc. Cambridge Philos. Soc. 33, 212-222.

[11] Puri, M.L., Ralescu, D.A. and Ralescu, S.S., 1984. Linear minimax estimators for estimating a function of the parameter. Australian and New Zealand Journal of Statistics, 26, 277 - 283.

[12] Ralescu, D.A. and Ralescu, S.S., 1981. A Class of Nonlinear Admissible Estimators in the One-Parameter Exponential Family. Ann. Statist. 9, 177-183.

[13] Rostek, M.J., 2007. Quantile Maximization in Decision Theory. Unpublished Manuscript.

[14] Rukhin, A.L., 1978. Universal Bayes Estimators. Ann. Statist. 6, 1345-1351.

[15] Rukhin, A. 1986 Admissibility and minimaxity results in the estimation problem of exponential quantiles. Ann. Statist. 14, 220-237.

[16] Savage, L.J., 1954. The Foundations of Statistics. Wiley, New York.

[17] Von Neumann, J., Morgenstern, O., 1947. The Theory of Games and Economic Behaviour, 2nd ed. (1st edn 1944). Princeton, Princeton University Press.

[18] Wald, A., 1939. Contributions to the Theory of Statistical Estimation and Testing Hypotheses. The Annals of Mathematical Statistics. 10, 299-326.

[19] Wald, A., 1950. Statistical Decision Functions. John Wiley, New York.

[20] Yu, C. W. 2009 *Median Loss Analysis and It Application to Model Selection.* PhD Thesis, Department of Statistics, University of British Columbia.

[21] Yu, C.W., Clarke, B., 2010. Asymptotics of Bayesian Median Loss Estimation. To appear in Journal of Multivariate Analysis.