The New England Journal of Statistics in Data Science Volume 0, 1–10 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS106

Limiting the Shrinkage for the Exceptional by Objective Robust Bayesian Analysis: The *"Clemente Problem"*

Luis R. PERICCHI and María-Eglée PEREZ*

The authors want to dedicate this work to Roberto Clemente Walker, first Latin American to reach the Baseball Hall of Fame.

Abstract

Modern Statistics is made of the sensible combination of direct evidence (the data directly relevant or the "individual data") and indirect evidence (the data and knowledge indirectly relevant or the "group data"). The admissible procedures combine the two sources of information, and technology advances make indirect evidence more substantial and ubiquitous. It has been pointed out, however, that an important problem of Statistics when "borrowing strength" is to treat in a fundamentally different way exceptional cases that do not adapt to the central "aurea mediocritas". This is what has been coined as "the Clemente problem" in honor of R. Clemente, an exceptional batter [6]. In this article, we argue that the problem is caused by the simultaneous use of square loss function and conjugate (light-tailed) priors, which is the usual procedure. We propose in their place to use robust penalties, in the form of losses that penalize more severely huge errors, or (equivalently) priors of heavy tails which make the exceptional more probable. Using heavy-tailed priors, we can reproduce, in a Bayesian structured way, Efron and Morris' "limited translated estimators" (with Double Exponential Priors) and "discarding priors estimators" (with Cauchy-like priors), which discard the prior in the presence of prior-likelihood conflict. We show that both Empirical Bayes and Full Bayes approaches can alleviate the Clemente problem and beat the James-Stein estimator in terms of smaller square errors, for sensible Robust Bayes priors. We model in parallel Empirical Bayes and Fully Bayesian hierarchical models, illustrating that the differences among sensible versions of both are relatively small, as compared with the effect due to the robust assumptions. We follow [16] in using a heavy-tailed (scaled) Beta2 distribution for (squared) scales that arise naturally as an alternative to the usual Inverted-Gamma distribution. The combination of a Cauchy Prior for location and Scaled Beta2 for square scales yields a novel closed-form prior for location, extremely suitable for Objective Robust Bayesian Analysis. Finally, we calculate the predictive intervals and the Robust models covers the Clemente average at 80% of probability, which the conjugate do not. Our approach has connections with [25], which employs a completely different methodology. This is an instance of the convergence of the best frequentist and Bayesian analyses [see 3].

KEYWORDS AND PHRASES: Cauchy-Scale2 Beta2 prior, Clemente problem, Robust loss functions, Horseshoe priors, Indirect evidence, Objective robust Bayesian analysis, Scaled Beta2 priors.

1. SHRINKAGE ESTIMATORS THAT BORROW STRENGTH FROM INDIRECT EVIDENCE: TOO MUCH OF A GOOD THING?

1.1 An Historical Example

In [8] the authors obtained a sample of batting averages for 18 baseball players during the 1970 season. They used the average obtained during the first 45 at-bats to predict the batting average for the rest of the season for each player.

In Table 1, the relevant data are presented. Direct evidence is the observed individual average; thus, the temptation is to predict by the observed individual average, although it is known that this estimator is inadmissible. This is a terrible practical predictor in this case, which has been corroborated in other cases [see 4]. In contrast, indirect evidence comes in the form of the general mean $M = 0.265$ of all batters.

This "pure indirect evidence" estimator is surprisingly good in this case and far better than the "pure direct evidence", MLE, or naive estimator, as Brown calls it. However, both intuition and theory point to a sensible combination of the two sources of evidence to improve overall predictions. The problem is then: How much to weigh direct and indirect evidence in each individual case? Wouldn't it be reasonable to weigh the common indirect evidence less when there is reason to believe that the individual is exceptional? This

*Corresponding author.

Table 1. Original data: 1970 *batting averages for* 18 *MLB players.*

Player	Batting average for first 45 at bats	Batting average for remainder of season	At bats for remainder of season
Clemente (Pitts, NL)	0.400	0.346	367
F. Robinson (Balt, AL)	0.378	0.298	426
F. Howard (Wash, AL)	0.356	0.276	521
Johnstone (Cal, AL)	0.333	0.222	275
Berry (Chi, AL)	0.311	0.273	418
Spencer (Cal, AL)	0.311	0.270	466
Kessinger (Chi, NL)	0.289	0.263	586
Alvarado (Bos, AL)	0.267	0.210	138
Santo (Chi, NL)	0.244	0.269	510
Swoboda (NY, NL)	0.244	0.230	200
Unser (Wash, AL)	0.222	0.264	277
Williams (Chi, AL)	0.222	0.256	270
Scott (Bos, AL)	0.222	0.303	435
Petrocelli (Bos, AL)	0.222	0.264	538
E. Rodriguez (KC, AL)	0.222	0.226	186
Campaneris (Oak, AL)	0.200	0.285	558
Munson (NY, AL)	0.178	0.316	408
Alvis (Mil, NL)	0.156	0.200	70

was the original motivation of Efron and Morris for their clever (and "*ad-hoc*") "limited translation estimators" [7].

1.2 Efron and Morris Set Up

The initial assumption about the data in [7] is:

$$Y_i \sim \frac{1}{45}\text{Bin}(45, p_i)$$

where Y_i is the batting average for the first 45 at-bats, and p_i depends on each player's ability.

The batting average for the rest of the season, R_i, can be modeled as

$$R_i \sim \frac{1}{n_i}\text{Bin}(n_i, p_i)$$

where n_i is the number of at bats for player i throughout the remainder of the season.

They applied a variance stabilizing transformation to Y_i,

$$X_i = \sqrt{45}\arcsin\left(2Y_i - 1\right)$$

so that $X_i \sim N(\mu_i, 1)$, with μ_i approximately equal to the transformed value of p_i. The interest is to predict the final individual batting average for the rest of the season. In the sequel, we will use this transformed variable.

The analysis of these baseball data by [7] was widely cited and remains one of the clearest expositions in favor of combining "*indirect evidence*" with "*direct evidence*", a practice often termed "borrowing strength", "learning from the experience of others", and "shrinkage estimation".

1.3 The Clemente Problem

In [6] a fundamental problem is exposed: *"The Clemente Problem: How to protect atypical cases from too much indirect evidence?"*. Professor Efron refers to the Puerto Rican sportsman Roberto Clemente, an outstanding batter and human being, who had the highest batting average on the list of 18 players. After the first 45 turns, Clemente had a batting average of 0.400, or 40% of hits. Even though shrinking to a general mean improves the overall prediction of the 18 batters, Clemente's average was predicted as 0.290. The atypical Clemente was not protected from "*too much of a good thing*", and his personal prediction was very poor: he finished with a batting average of 0.346, much higher than predicted. The problem lies in the fact that the usual method shrinks a fixed proportion to all players; see Equation (3.3) below. It makes no exceptions for batters that are too good or too bad. This is a logical consequence of the assumptions made, since it corresponds to an optimal decision in Decision Theory and the coherence of Bayesian analysis. So, in *"What if?"* mode of thought, *"What if?"* if the logical consequences are not *"pleasant to the mind"*, *a fortiori* assumptions must be changed. Fixed proportion estimators are not robust in the sense that the amount of shrinkage is not limited, that is, the potential influence of indirect evidence is unbounded, or using a metaphoric expression, the procedure is "*myopic*" to the conflict between the bulk of the data and the individual [15]

1.4 Robust Penalties

Our starting point is re-analyzing the Loss (or minus Utility) function. In decision analysis, the Square Loss is by far

the most used (or over-used?) and the Clemente problem is (in part) a coherent consequence of its assumption. To see this, we recall the following result in Decision Theory.

Result 1. *Suppose that a function of the parameter* $\boldsymbol{\theta}$, $g(\boldsymbol{\theta})$, *is being estimated by* $\boldsymbol{\delta}(X_1, \ldots, X_m)$. *Assume the weighted square loss function:*

$$\begin{aligned} L(g(\boldsymbol{\theta}), \boldsymbol{\delta}) &= \sum_{i=1}^{m} L_i(g(\theta_i), \delta_i(\mathbf{X})) \\ &= \sum_{i=1}^{m} w(\theta_i) \cdot (\delta_i(\mathbf{X}) - g(\theta_i))^2, \ w(\theta_i) \geq 0. \end{aligned}$$

Then the optimal Bayes estimator is:

$$\delta_i(\mathbf{X}) = \frac{E[w(\theta_i) \cdot g(\theta_i)|data]}{E[w(\theta_i)|data]}. \tag{1.1}$$

Proof: Page 47 in [9].

Although this result may be termed as classical, its statistical consequences have not been fully appreciated. In fact (1.1) invites two strategies: the first is to weight the square loss and use (1.1) as the individual estimator, and the second is to change the prior in a way suggested by (1.1) while keeping the square loss function. These are two different viewpoints that shed different light and possibilities. In this article, we highlight assumptions to get robust solutions in both ways.

2. HEAVIER THAN QUADRATIC LOSSES

For simplicity, we will assume we are estimating $\boldsymbol{\theta}$, that is, $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$.

To diminish the shrinkage at the extremes, the loss function (centered around the overall group location) must penalize errors far from the overall location more heavily than the square loss. We also need an origin M to anchor the weighting function $w(\theta_i)$. This origin M might be chosen subjectively, using prior knowledge and experience, or empirically, and could be interpreted as the mean of the θ_i's.

Example 1 (Exponentially weighted loss)**.**

$$L_i^{exp}(\theta_i, \delta_i(\mathbf{X})) = \exp[r \cdot |\theta_i - M|] \cdot (\theta_i - \delta_i(\mathbf{X}))^2, \quad i = 1, \ldots, m, r > 0$$

Calculation yields the optimal estimator in (1.1) as

$$\delta_i^{exp}(\mathbf{X}) = \frac{a[b(\mu_{1i} - rv) - c] + a'[c' + (1 - b')(\mu_{1i} + rv)]}{ab + a'(1 - b')},$$

where $a = \exp(r(M - \mu_{1i}) + vr^2/2))$, $b = \Phi(\frac{(M-(\mu_{1i}-rv)}{\sqrt{v}})$, $c = \sqrt{v}\phi(\frac{M-(\mu_{1i}-rv)}{\sqrt{v}})$, $a' = \exp(r(\mu_{1i} - \tilde{\theta}) + vr^2/2))$,$b' = \Phi(\frac{\tilde{\theta}-(\mu_{1i}+vr)}{\sqrt{v}})$,$c' = \sqrt{v}\phi(\frac{\tilde{\theta}-(\mu_{1i}+rv)}{\sqrt{v}})$, and μ_{1i}, v are the posterior mean and variance, respectively, of a Normal likelihood with a Normal prior.

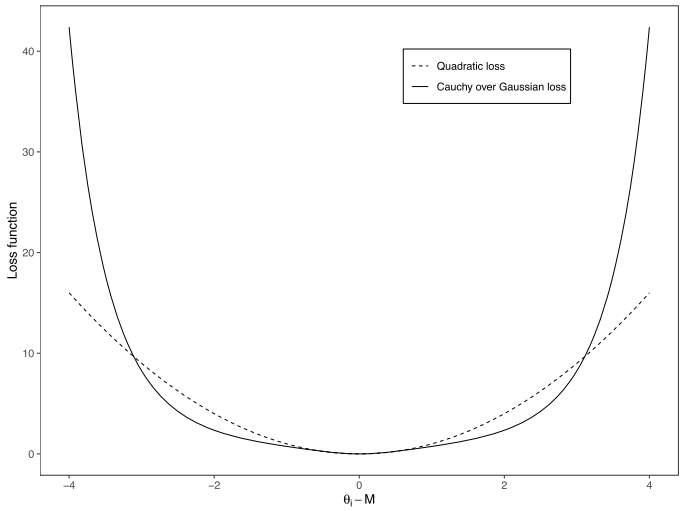


Figure 1: Quadratic loss and Cauchy over Gaussian loss ($\delta_i = M$).

This and other loss functions can lead to tractable results, but it is more convenient for the purposes of the present paper to work with losses that also have a direct interpretation in terms of heavy tailed (robust) priors, which penalize discrepancies more heavily and are naturally scaled.

Example 2 (Cauchy over Gaussian Loss:)**.** *Here we place a Cauchy density over a Gaussian, both centered at* M *with matched interquartile range.*

$$L_i^{CG}(\theta_i, \delta_i(\mathbf{X})) = \frac{Cauchy(\theta_i|M, 1)}{N(\theta_i|M, 2.19)} \cdot (\theta_i - \delta_i(\mathbf{X}))^2.$$

Figure 1 shows the Cauchy over Gaussian loss, being quite close to a square loss around zero but growing fast without bound for "exceptional" values.

Result 2. *In fact, the optimal estimator (1.1) under the* $L_i^{CG}(\theta_i, \delta_i(\mathbf{X}))$ *is the posterior expectation under a Cauchy prior, since using the Gaussian with mean* M *and variance* 2.19 *as a prior*

$$\begin{aligned} \delta(\mathbf{X}) = \frac{E[w(\theta_i) \cdot \theta_i|\mathbf{X}]}{E[w(\theta_i)|\mathbf{X}]} &= \frac{\int \frac{Cauchy(\theta_i|M,1)}{N(\theta_i|M,2.19)} \theta_i \cdot \pi(\theta_i|\mathbf{X}) d\theta_i}{\int \frac{Cauchy(\theta_i|M,1)}{N(\theta_i|M,2.19)} \pi(\theta_i|\mathbf{X}) d\theta_i)} \\ &= E[\theta_i|\mathbf{X}, \ Cauchy \ Prior] \end{aligned}$$

In words, the optimal estimator under the Robust Cauchy over Gaussian penalty (and Gaussian prior) is the posterior expectation under a Cauchy Prior, since with square loss, the optimal estimator is the posterior expectation.

So, a bridge has been established between Robust Losses and Robust Priors through equation (1.1). We call it comprehensive Robustness, the use of Robust Loss Functions, or the use of Robust (heavy tailed) Priors. For convenience, now we go to the "Robust Prior Route".

In what follows we try different models, in two methods: Empirical Bayes and Fully Robust Bayes, trying to solve, or at least alleviate, the Clemente problem, that is, the lack of robustness of the exponential family models with conjugate (light-tailed) priors and squared loss function.

3. ROBUST CONSEQUENCES OF ROBUST (HEAVILY TAILED) PRIORS

In this section, we motivate briefly the use of heavy tailed priors as a tool for robustness. Back in the situation of subsection 1.2, let us make the usual assumption of a Normal Likelihood and Prior:

$$X_i \sim \text{Normal}(\mu_i, 1), \quad i = 1, \ldots, 18 \tag{3.1}$$
$$\mu_i \sim \text{Normal}(\text{M}, \sigma_0^2), \tag{3.2}$$

where (M, σ_0^2) has been assigned, for instance via an Empirical Bayesian method as in [7] or using previous year batting averages (for 1969, the global batting average was 0.248, see for example http://www.baseball-reference.com/leagues/MLB/1969.shtml). These assumptions, coupled with square error loss, lead to the posterior conditional expectation as the optimal estimator, which can be written as:

$$E(\mu_i|x_i) = x_i + \frac{1}{1+\sigma_0^2}(M - x_i). \tag{3.3}$$

It is convenient here to define shrinkage as $|E(\mu_i|x_i) - x_i|$. There are several ways to analyze the lack of robustness of (3.3), but the one most relevant here is: *all batters, whether exceptional or average, are shifted to the mean of the means M by a **fixed** proportion* $c = 1/(1+\sigma_0^2)$, so the shrinkage is $c \cdot |M - x_i|$. That is the "Clemente Problem". Notice that increasing the prior variance σ_0^2 is **not** a fix to the problem: it would certainly reduce the shrinkage, but to all batters in equal proportion, even to those who are not exceptional. We are not proposing an indiscriminate reduction of the shrinkage, but rather a differential shrinkage. The posterior mean (3.3) is *myopic* to the exception. Now (3.3) is the logical consequence of the assumptions. Thus, the only logical way to change it is to change the assumptions. We could change the loss, but equivalently, we change the tail behavior of the priors: using flatter tails gives Bayes Theorem the input that the exceptional is more likely, and probabilistic (not ad-hoc impositions) coherence then acts as a "robustifier" of the estimation.

Our first alternative is a Double Exponential prior:

$$\mu \sim \text{DE}(\text{M}, \nu_0) = \frac{1}{\nu_0\sqrt{2}} \exp(-\frac{\sqrt{(2)}}{\nu_0}|\mu - \text{M}|), \tag{3.4}$$

where $\nu_0 = \frac{\sqrt{(2)}\sigma_0}{\log(2)}\Phi^{-1}(0.75)$, to match the quartiles of the Normal prior. Finally for even heavier tails we explore with a Cauchy prior.

$$\mu \sim \text{C}(\text{M}, \gamma_0) = \frac{1}{\pi\gamma_0}\frac{1}{1+(\mu - \text{M})^2/\gamma_0^2}, \tag{3.5}$$

where $\gamma_0 = \sigma_0\Phi^{-1}(0.75)$, again to match Normal quartiles.

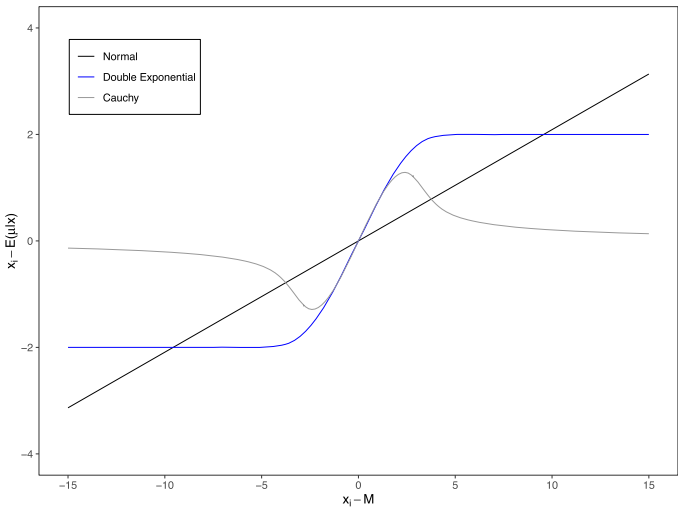


Figure 2: Observation fixed at zero prior location varying. Normal linear unbounded influence of prior location, monotone limited translation in Double Exponential and discarding influence in the Cauchy.

For exact and approximate results with these, and other priors, see for example [17], but Figures 2 and 3 tell the story. In Figure 2 the observation is kept fixed at zero and the prior location is moved to create a conflict between one data point and a prior. With a Normal prior the shrinkage grows linearly without bound. The other two priors yield robust estimators. For a Double Exponential prior, the posterior expectation becomes essentially a '*limited translation estimator*" in Efron and Morris' terminology. The influence of the overall mean M is bounded and monotonic. Finally, for the Cauchy prior, the posterior expectation is **not** monotonic in the conflict between the MLE and the General Mean. Furthermore, the prior is progressively discarded in favor of the MLE, as the general mean and MLE diverge. It is also quite interesting that the shrinkage of the two robust priors is almost the same close to the center, that is, around the general average M, and actually they give more shrinkage near the center than the Normal.

In Figure 3 we are showing the actual values of the transformed data X_i, $i = 1, \ldots, 18$ on the x axis. The prior location M is fixed at the sample overall average. Around the middle-ground the three estimators are very close together. But at some point, the robust estimators separate from the Normal, in the direction of the MLE, being the adjustment towards the MLE of the Cauchy somewhat stronger than that of the Double-Exponential.

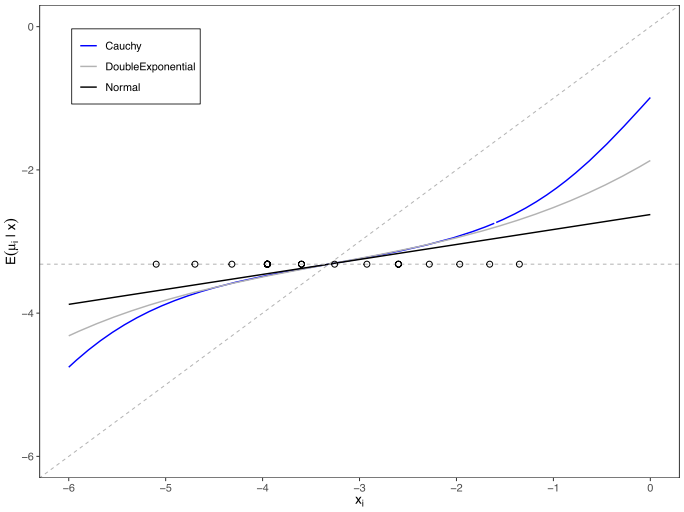


Figure 3: Prior location fixed at the general average. Observations displayed in the X-Axis. Different Shrinkage behavior displayed of Normal, Double Exponential and Cauchy priors.

4. MODELS

We will consider two types of strategies, Empirical Bayes and Full Bayesian strategies.

4.1 Empirical Bayes Strategy

Following [8], general location and scale parameters are calculated from the sample: $\mathrm{M} = \bar{X} = -3.3166$ and $\tilde{\sigma}^2$ such that $\frac{1}{(1+\tilde{\sigma}^2)} = \frac{k-3}{\sum_{i=1}^{k}(X_i-\bar{X})^2}$, so $\tau = (\sigma^2)^{-1} = 3.7853$.

Models 1, 2 and 3, are defined with the likelihood (3.1) and respectively priors (3.2), (3.4) and (3.5). The justification is as follows: the first prior corresponds to the original analysis by Efron and Morris, Double Exponential and Cauchy priors are two heavy tailed priors (as compared with Normal) that promote a qualitatively different behavior of the estimators. The three priors have the same origin given by an Empirical Bayes analysis, and their scales have been matched by equating interquartile ranges to facilitate comparison.

4.2 Full Bayesian Strategy

4.2.1 Model 4: Full Bayesian Conjugate Model

This model assigns vague conjugate priors to the common mean M and the common variance σ^2. This is the fully Bayes version of Efron and Morris Empirical Bayes model.

$$\begin{aligned} X_i &\sim \text{Normal}(\mu_i, 1), i = 1, \ldots, 18 \\ \mu_i &\sim \text{Normal}(\mathrm{M}, \sigma^2) \\ M &\sim N(0, 10^5), \quad \sigma^2 \sim \text{Inv-Gamma}(0.01, 0.01) \end{aligned}$$

To make Model 4 robust, it is not enough to change the Normal Prior assumption. It is also necessary to replace the Inverted-Gamma distribution assumption regarding the scales, see [12]. We first present the alternative models and delay until next section the discussion of the Scaled Beta2 model (SBeta2).

4.2.2 Model 5: Normal Likelihood, Double Exponential Prior for μ_i, Vague Double Exponential Prior for the General Mean M, Scaled Beta2$(1, 1, 1)$ Prior for the Scale Parameter $\sigma = \frac{\nu}{\sqrt{2}}$

$$\begin{aligned} X_i &\sim \text{Normal}(\mu_i, 1), i = 1, \ldots, 18 \\ \mu_i &\sim \text{DE}(\mathrm{M}, \sqrt{2}\sigma) \\ M &\sim \text{DE}(0, \sqrt{2} \times 10^3), \quad \sigma \sim \text{Beta2}(1, 1, 1) \end{aligned}$$

4.2.3 Model 6: Normal Likelihood, Cauchy Prior for μ_i, Vague Cauchy Prior for the General Mean M, Scaled Beta2$(1, 1, 4)$ Prior for the Scale Parameter

$$\begin{aligned} X_i &\sim \text{Normal}(\mu_i, 1), i = 1, \ldots, 18 \\ \mu_i &\sim \text{Cauchy}(\mathrm{M}, \sigma) \\ M &\sim \text{Cauchy}(0, 10^3), \quad \sigma \sim \text{SBeta2}(1, 1, 4) \end{aligned}$$

4.2.4 Model 7: Normal Likelihood, Cauchy Prior for μ_i, Vague Cauchy Prior for the General Mean M, Scaled Beta2$(1, 1, 4)$ Prior for the Squared Scale Parameter

$$\begin{aligned} X_i &\sim \text{Normal}(\mu_i, 1), i = 1, \ldots, 18 \\ \mu_i &\sim \text{Cauchy}(\mathrm{M}, \sigma) \\ M &\sim \text{Cauchy}(0, 10^3), \quad \sigma^2 \sim \text{SBeta2}(1, 1, 4) \end{aligned}$$

4.3 The Scaled Beta2 Distribution as a Prior for Scales

In these last three models, robust priors (Double Exponential and Cauchy) have been assigned for the locations. On the other hand, we propose the use of the **scaled Beta distribution of the second kind** with parameters p, q and b family (SBeta2(p, q, b)) as priors for the scale parameters or for squares of scales. Let Y be a random variable such that $Y \sim \text{Beta2}(p, q, b)$; its density function [13, 16, 10] is given by:

$$p(y|p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{1}{b} \cdot \frac{\left(\frac{y}{b}\right)^{p-1}}{\left(1+\frac{y}{b}\right)^{(p+q)}}, y > 0,$$

This family has a very natural justification as a prior for variances in hierarchical models, as it is obtained as a scale mixture of Gamma distributions, through a Gamma mixing distribution, in much the same way that the Student-t is obtained as a scale mixture of Normal distributions [16]. The

usual prior assumed for scales is the Inverted Gamma family with very large prior variance. This practice has come under criticism by [12], where the author, among other alternatives, proposes a half-Cauchy prior. The Scaled Beta2 prior has flexible tail behavior, which makes it particularly suitable for modeling. When p = q = 1, the SBeta2(1,1,b) prior is very close to the half-Cauchy, so we use it here. (If σ is Half-Cauchy scale b, then σ^2 is SBeta2(1/2,1/2, b^2)).

What is the assumed prior for the location parameter in Model 6 after integrating out the scale? This is the so-called Cauchy-Scaled Beta 2 [16], which deserves special mention on its own.

Definition 1 (Cauchy-Scaled Beta2 Prior)**.** *Assume that* $\theta|\sigma \sim Cauchy(0,\sigma)$*, and* $\sigma \sim Beta2(p,q,b)$*, then* θ *is distributed as a Cauchy - Scaled Beta2*$(0,p,q,b)$*.*

It is remarkable that when $p = q = 1$ (an assignment that makes the Beta2 distribution have Cauchy tails), the marginal density for θ has an explicit formula (after a long integration by simple fractions):

Result 3. *The marginal density for the location* θ *of a Cauchy-Scaled Beta2 prior is:*

$$\begin{aligned}\pi(\theta) &= \int_0^\infty \frac{b\tau}{\pi(b+\tau)^2\,(\theta^2+\tau^2)}\,d\tau \\ &= \frac{b}{\pi\,(b^2+\theta^2)^2}\left[-\left(b^2+\theta^2-\pi b|\theta|\right)+\right. \\ &\qquad \left.(b-\theta)(b+\theta)\,(\log(b)-\log(|\theta|))\right]\end{aligned}$$

[see 16].

In Figure 4, the Cauchy-Scaled Beta2 prior is displayed along with the Cauchy, Double-Exponential and Normal. This prior enjoys several features that explain why it is so efficient in predicting the batter's averages. This prior is unbounded as $\theta \to 0$ and has tails even heavier than Cauchy. In [5] [see also 18] finds such characteristics for a prior to be both robust and leading to efficient estimation (they propose particular prior which does not have an explicit form and that they call "horseshoe" prior). Furthermore, the Cauchy-Scaled Beta2 besides obeying such desiderata, has an explicit form (at least for p=q=1) which makes it amenable for mathematical analysis. We do not know of any other explicit "horseshoe" prior. (In [16] results for other values of hyper-parameters are obtained)

Using again a Cauchy prior for the location parameter, the Scaled Beta2 family can be also be used as a prior for the square of the scale, and a closed form for the marginal of the location parameter θ is also available [see 10].

Result 4. *Assume* $\theta|\tau^2$ *has the form*

$$\pi(\theta|\mu,\tau,b) = \frac{1}{\pi\sqrt{b}\tau}\cdot[1+\frac{(\theta-\mu)^2}{b\tau^2}]^{-1},$$

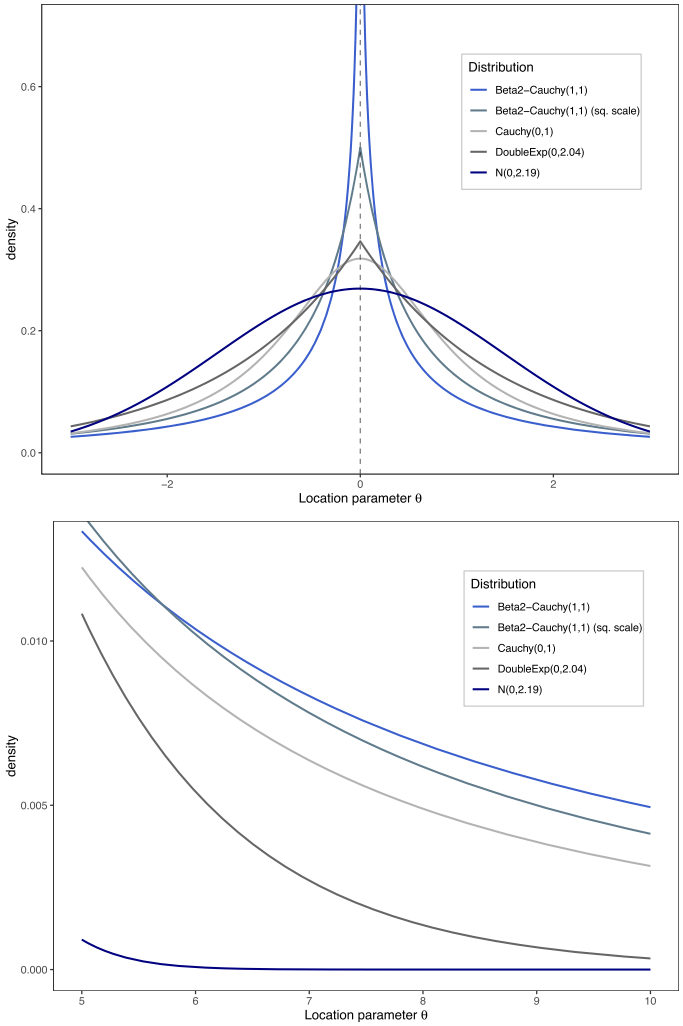


Figure 4: Comparison of Cauchy-Scaled Beta2, Normal, Double Exponential and Cauchy distributions, at the center (upper figure) and at the tail (lower figure).

and

$$\tau^2 \sim SBeta2(\tau^2|1,1,b).$$

Then the marginal density for the location θ *is*

$$\pi(\theta) = \frac{1}{2\sqrt{b}\cdot(1+\frac{|\theta-\mu|}{\sqrt{b}})^2}. \tag{4.1}$$

This is an interesting marginal on itself, close to a Cauchy, and it does not have a pole at zero, so it is not a Horseshoe prior. In [10] this prior is studied and applied in detail. In fact a general result for the marginal of the location, for any p and q is obtained in terms of the Hypergeometric Function.

Assessments of Hyper-parameters in the Scaled Beta 2 Distribution: We assessed $p = q = 1$ in model

Table 2. Estimators and mean square error of prediction for MLE, general mean, empirical Bayes models and full Bayesian models.

Player	Observed season	First 45 (MLE)	General mean	Model 1	Model 2	Model3	Model 4	Model 5	Model 6	Model 7
Clemente	0.346	0.400	0.265	0.290	0.304	0.314	0.282	0.298	0.291	0.309
Robinson	0.298	0.378	0.265	0.286	0.296	0.301	0.279	0.291	0.283	0.296
Howard	0.276	0.356	0.265	0.282	0.288	0.291	0.277	0.285	0.2762	0.287
Johnstone	0.222	0.333	0.265	0.277	0.281	0.282	0.273	0.279	0.272	0.279
Berry	0.273	0.311	0.265	0.273	0.275	0.275	0.270	0.273	0.269	0.273
Spencer	0.270	0.311	0.265	0.273	0.275	0.275	0.270	0.273	0.269	0.273
Kessinger	0.263	0.289	0.265	0.269	0.269	0.270	0.267	0.268	0.266	0.269
Alvarado	0.210	0.267	0.265	0.265	0.264	0.264	0.264	0.264	0.263	0.263
Santo	0.269	0.244	0.265	0.259	0.258	0.259	0.261	0.259	0.260	0.259
Swoboda	0.230	0.244	0.265	0.259	0.258	0.259	0.260	0.259	0.260	0.259
Unser	0.264	0.222	0.265	0.255	0.252	0.254	0.257	0.254	0.258	0.253
Williams	0.256	0.222	0.265	0.255	0.252	0.254	0.257	0.254	0.257	0.254
Scott	0.303	0.222	0.265	0.255	0.252	0.253	0.257	0.254	0.257	0.253
Petrocelli	0.264	0.222	0.265	0.255	0.252	0.253	0.257	0.254	0.257	0.253
Rodriguez	0.226	0.222	0.265	0.255	0.252	0.253	0.257	0.254	0.257	0.254
Campaneris	0.285	0.200	0.265	0.250	0.245	0.247	0.254	0.248	0.254	0.247
Munson	0.316	0.178	0.265	0.245	0.237	0.238	0.251	0.242	0.249	0.240
Alvis	0.200	0.156	0.265	0.240	0.228	0.226	0.247	0.234	0.242	0.230
MSE ($\times 10^3$)		**4.184**	**1.348**	**1.196**	**1.187**	**1.137**	**1.198**	**1.168**	**1.108**	**1.117**
$\frac{MSE(\text{Model})}{MSE(\text{Model 1})}$		**350%**	**113%**	**100%**	**99%**	**95%**	**100%**	**98%**	**93%**	**93%**

6 and 7, which is a sensible default assumption since then, both the value of the scale and its reciprocal are finite at zero, and both tails are very heavy. We also assumed $b = 4$, which is larger than three times the estimator of the between variance, and the results for larger values were found to be quite similar to those with $b = 3$.

5. MODEL PREDICTIONS

All proposed models were fitted using Stan (version 2.32) [19] through package `rstan` (version 2.37.2) [20]. Table 2 shows the batting averages predicted for each of the models, and in Figure 5 we display the milder shrinkage of the extremes using robust priors, particularly Cauchy priors, when compared to that of Model 1 and Model 4.

Robustifying the priors pays dividends twice: the relative shrinkage of the extremes is lower and, at the same time, the error of prediction is diminished up to 7% for Models 6 and 7, which use the Cauchy-Scaled Beta2 and the Cauchy-Scale2 Beta 2 priors respectively. Model 3, which incorporates the Cauchy prior has, for example, a square prediction error 5% lower than Model 1, and predicts for Clemente a more respectful 0.314 average, much higher than the 0.290 from Model 1. Something similar may be said for Model 6 and Model 7. On the other hand, the price paid seems less than modest: computational tools as approximations and MCMC algorithms that make the computations routine are now available.

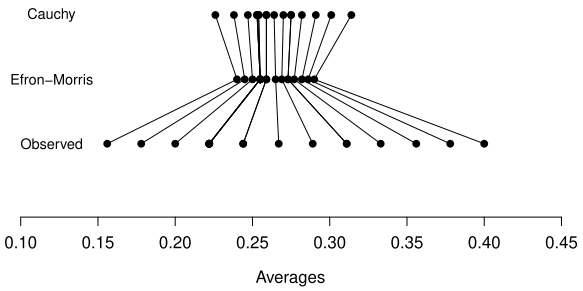
Thus by a very modest cost in computation, the "robustified" model has achieved both goals, decreasing the MSE and solving or at least alleviating the Clemente problem.

In terms of alternative approaches of Statistics that merge direct and indirect evidence, the difference between (sensible and objective versions) of Empirical and Fully Bayes Hierarchical Modeling is relatively small as compared with the difference between heavy and light tail priors.

In general, the Clemente problem is closely related to the implicit dogmatism inherent, not in Bayes in general, but in "Conjugate Bayes with Square Loss". The way out seems to be: either an Empirical or Fully Bayesian Hierarchical Model, but making emphasis on Robustness.

In many contexts, Bayesian models are compared using methods focused on the predictive performance in cross-validation using tools as Watanabe-Akaike Information Criterion [24] and Leave One Out cross validation (LOO-CV) criterion [22]. WAIC and LOO were calculated using R-package `loo` [23]. Table 3 shows the results

According to Table 3, these two popular methods have little discriminatory power between models but slightly prefer conjugate ones. This is in contrast to a prediction perspective. The 80% intervals of the Robust Models 3, 6 and 7, cover the Clemente Data but not the others (see Figure 6). It is striking that the best models in terms of WAIC and LOO are the poorest in prediction coverage. Coverage is a main objective of this article, as a consequence of structured Bayesian modeling. In summary, the best predictive models (in terms of coverage and in terms of Mean Squared Errors)



(a)MLE, Model 1, and Robust Empirical Bayes Model 3.

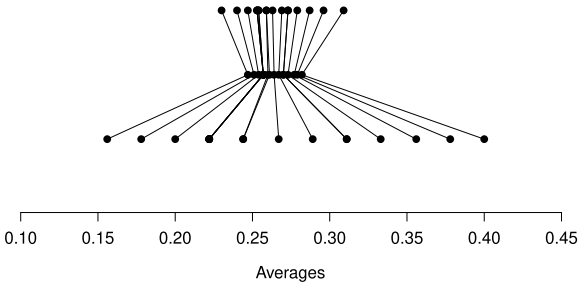


Figure 5: Comparison of shrinkages: upper figure shows observed data, Efron and Morris estimates and Robust Empirical Bayes estimates. The lower figure shows observed data, non-robust Full Bayes estimates and Robust Full Bayes estimates.

are robust models, based on the Cauchy and Scaled-Beta 2 distributions. This is the case whether from an Empirical Bayes or Full Bayesian methodology. On the other hand, methods for model selection like WAIC and LOO performed badly in prediction and coverage.

6. "OBJECTIVE ROBUST BAYESIAN ANALYSIS"

It can be argued about the great relevance of Objective Robust Bayesian Analysis (ORBA). To put this article

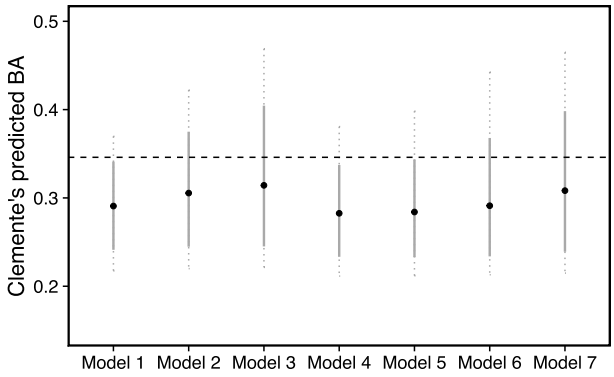


Figure 6: 80% (solid line) and 95% (dotted line) prediction intervals for Clemente's batting average by the end of the season. The actual value, 0.346, is shown by the dashed line.

in perspective, we should mention some contributions. The first, although its approach is subjective, is [1], which uses the theory of Regularly Varying (RV) functions to assess robustness for general location and scale parameters. The second is [11], where the Generalized Polynomial Tails Comparison (GPTC) Theorem is proved, and the properties of specific robust priors are analyzed. In [15], the relationship between RV and GPTC is established. On the other hand, [5] introduced the "Horseshoe Estimator," which obeys certain desiderata convenient for robust analysis in hierarchical models.

In the present article, we add the following insights: 1) We established a bridge between robust losses and robust priors, by weight ratios times the square loss, using the expression (1.1) to compute the optimal. Example 2 provides a specific example of how to assess a robust loss through the ratio of a flat-tailed prior over a Gaussian. The theoretical duality between losses and priors has been mentioned before, for example, by [2, p. 161], but we also search for the consequences of looking at robust procedures through the glass of robust penalties. Furthermore, note that the connection between priors and losses is done through an empirical Bayes component, since the loss is centered and scaled through an EB reasoning. 2) We employ systematically here a robust prior for scales, the Scaled Beta2, which

Table 3. WAIC and LOO values for proposed empirical Bayes models and full Bayesian models.

Model	WAIC	SE	p_{WAIC}	SE	LOO	SE	p_{LOO}	SE
Model 1	52.1	3.8	2.9	0.6	52.3	3.9	3.0	0.7
Model 2	52.4	3.4	3.4	0.7	53.0	3.5	3.8	0.8
Model 3	52.5	3.5	3.5	0.7	53.9	3.5	4.2	0.7
Model 4	54.2	4.5	3.0	0.7	54.6	4.5	3.2	0.7
Model 5	54.2	4.5	3.0	0.7	54.6	4.5	3.2	0.7
Model 6	54.2	4.5	3.0	0.7	54.6	4.5	3.2	0.7
Model 7	54.4	3.9	4.0	0.9	55.8	4.5	4.7	0.9

is a convenient alternative to the overused Inverted-Gamma prior. Furthermore, we show the Cauchy-Scaled Beta2 prior, which is an explicit objective prior that obeys the desiderata of a "Horseshoe" prior, as well as the Cauchy-Scaled Beta2 prior with general closed form marginal distributions for location parameters. 3) We illustrate, using a classical data set, that it is possible to alleviate the "Clemente problem" and at the same time reduce the mean square error of prediction as compared with non-robust conjugate approaches and with the James-Stein estimator. 4) We illustrate that the differences between sensible versions of Empirical Bayes and Objective Bayes are relatively small in practice. Much more important is the difference between robust Bayes and conjugate Bayes. This, coupled with the fact that Robust Bayes is, in general, less dogmatic, in the sense that in conflict prior information is discarded or at least with limited influence, makes Objective Robust Bayes a strong candidate for statistical synthesis.

The scope of applications of Robust Objective Procedures is wide-ranging and includes any problem calling for the use of hierarchical models. Just to mention a few, in the analysis of population dynamics for the orchid genus *Caladenia* presented in [21], hierarchical models based on vague conjugate priors were used, but although the model selection procedures clearly pointed towards a hierarchical model as the selected one, they gave results that were not biologically sound, probably due to excessive shrinkage toward the species with more data. Another example is the unfair "pulling down" of hitherto perfect-scoring hospitals in hospital profiling, simply because a couple of hospitals have poor performance [see 14]. *"Shrinkage is a good thing, but non-robust-conjugate methods yield too much of a good thing"*.

ACKNOWLEDGEMENTS

The authors would like to thank two anonymous referees, whose comments led to important improvements to this paper.

FUNDING

The first author acknowledges support from: NIH P20 GM148324, P20 GM103475, RO1 CA282427. Both authors acknowledge support from the Biostatistics and Bioinformatics Center (BBC) of the Department of Mathematics, University of Puerto Rico Rio Piedras Campus, and the grant: Demography PR 2026, of the University of Puerto Rico.

Accepted 22 May 2026

REFERENCES

[1] Andrade, J. A. A. and O'Hagan, A. (2006). Bayesian Robustness Modeling Using Regularly Varying Distribution. *Bayesian Analysis* **1** 169–188. https://doi.org/10.1214/06-BA106. MR2227369

[2] Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4757-4286-2. MR0804611

[3] Berger, J. O. (2023). Four Types of Frequentism and Their Interplay with Bayesianism. *New England Journal of Statistics in Data Science* **1(2)** 126–137. https://doi.org/10.51387/22-NEJSDS4.

[4] Brown, L. D. (2008). In-Season Prediction of Batting Averages: a Field Test of Empirical Bayes and Bayes Methodologies. *The Annals of Applied Statistics* **2** 113–214. https://doi.org/10.1214/07-AOAS138. MR2415597

[5] Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The Horseshoe Estimator for Sparse Signals. *Biometrika* **97(2)** 465–480. https://doi.org/10.1093/biomet/asq017. MR2650751

[6] Efron, B. (2010). The Future of Indirect Evidence. *Statistical Science* **25** 145–157. https://doi.org/10.1214/09-STS308. MR2789983

[7] Efron, B. and Morris, C. (1972). Limiting the Risk of Bayes and Empirical Bayes Estimators-Part II: The Empirical Bayes Case. *Journal of the American Statistical Association* **67** 130–139. MR0323015

[8] Efron, B. and Morris, C. (1975). Data Analysis using Stein's Estimators and Its Generalizations. *Journal of the American Statistical Association* **70** 311–319.

[9] Ferguson, T. (1967) *Mathematical Statistics. A Decision Theoretic Approach*. Academic Press. MR0215390

[10] Fúquene, J., Pérez, M. E. and Pericchi, L. (2014). An alternative to the Inverted Gamma for the variances to modelling outliers and structural breaks in dynamic models. *Brazilian Journal of Probability and Statistics* **28**(2). https://doi.org/10.1214/12-BJPS207. https://doi.org/10.1214/12-BJPS207. MR3189499

[11] Fúquene, J. A., Cook, J. D. and Pericchi, L. R. (2009). A Case for Robust Bayesian Priors with Applications to Clinical Trials. *Bayesian Analysis* **4** 817–846. https://doi.org/10.1214/09-BA431. MR2570090

[12] Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis* **1(3)** 515–533. https://doi.org/10.1214/06-BA117A. MR2221284

[13] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions* **2**. Wiley. MR1326603

[14] Normand, S. T. and Shahian, D. M. (2007). Statistical and Clinical Aspects of Hospital Outcomes Profiling. *Statistical Science* **22**(2) 206–226. https://doi.org/10.1214/088342307000000096. MR2408959

[15] O'Hagan, A. and Pericchi, L. (2012). Bayesian heavy-tailed models and conflict resolution: A review. *Brazilian Journal of Probability and Statistics* **26(4)** 372–401. https://doi.org/10.1214/11-BJPS164. https://doi.org/10.1214/11-BJPS164. MR2949085

[16] Perez, M. E., Pericchi, L. R. and Ramirez, I. C. (2017). The Scaled Beta2 Distribution as a Robust Prior for Scales. *Bayesian Analysis* **12**(3) 615–637. https://doi.org/10.1214/16-BA1015. MR3655869

[17] Pericchi, L. R. and Smith, A. F. M. (1992). Exact and approximate posterior moments for a Normal location parameter. *Journal of the Royal Statistical Society B* **54**(3) 793–804. MR1185223

[18] Polson, N. G. and Scott, J. G. (2011). Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction. In *Bayesian Statistics 9* 501–538 Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199694587.003.0017. MR3204017

[19] Stan Development Team (2023). *Stan Modeling Language Users Guide and Reference Manual, version 2.32*. https://mc-stan.org.

[20] Stan Development Team (2025). *RStan: the R interface to Stan*. R package version 2.32.7. https://mc-stan.org/.

[21] Tremblay, R. L., Pérez, M. E., Larcombe, M., Brown, D., Quarmby, J., Bickerton, D., French, G. and Bould, A. (2009). Population dynamics of *Caladenia*: Bayesian estimates of transition and extinction probabilities. *Australian Journal of Botany* **57** 351–360.

[22] Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27** 1413–1432. https://doi.org/10.1007/s11222-016-9696-4. https://doi.org/10.1007/s11222-016-9696-4. MR3647105

[23] Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P. -C., Paananen, T. and Gelman, A. (2025). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 2.9.0. https://mc-stan.org/loo/.

[24] Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *J. Mach. Learn. Res.* **11** 3571–3594. MR2756194

[25] Yu, C. and Hoff, P. D. (2018). Adaptive multigroup confidence intervals with constant coverage. *Biometrika* **105** (2) 319–335. https://doi.org/10.1093/biomet/asy009. https://doi.org/10.1093/biomet/asy009. MR3804405

Luis R. Pericchi. University of Puerto Rico, Rio Piedras Campus, USA. E-mail address: luis.pericchi@upr.edu

María-Eglée Perez. University of Puerto Rico, Rio Piedras Campus, USA. E-mail address: maria.perez34@upr.edu