Working Papers in Statistics
No 2015:6
Department of Statistics
School of Economics and Management
Lund University

A novel weighted likelihood estimation with empirical Bayes flavor

MD. MOBARAK HOSSAIN, UNIVERSITY OF NEVADA
TOMASZ J. KOZUBOWSKI, UNIVERSITY OF NEVADA
KRZYSZTOF PODGÓRSKI, LUND UNIVERSITY

A novel weighted likelihood estimation with empirical Bayes flavor

Md. Mobarak Hossain and Tomasz J. Kozubowski

Department of Mathematics & Statistics, University of Nevada, Reno NV 89557, USA

Krzysztof Podgórski

Department of Statistics, Lund University, Sweden

Abstract: We propose a novel approach to estimation, where each individual observation in a random sample is used to derive an estimator of an unknown parameter using the maximum likelihood principle. These individual estimators are then combined as a weighted average to produce the final estimator. The weights are chosen to be proportional to the likelihood function evaluated at the estimators based on each observation. The method can be related to a Bayesian approach, where the prior distribution is data driven. In case of estimating a location parameter of a unimodal density, the prior distribution is the empirical distribution of the sample, and converges to the true distribution that generated the data as the sample size increases.

We provide several examples illustrating the new method, argue for its consistency, and conduct simulation studies to assess the performance of the estimators. It turns out that this straightforward methodology produces consistent estimators, which seem to be comparable with those obtained by the maximum likelihood method.

AMS 2010 subject classifications: 62C12, 62F10, 62F12, 62F15

Key words and phrases: Consistency, data-dependent prior, empirical Bayes, exponentiated distribution, location parameter, maximum likelihood estimator, super-efficiency, unbounded likelihood

1. INTRODUCTION

The method of maximum likelihood (ML) is perhaps the most widely used estimation method in the parametric setting. However, the required optimization of the likelihood function is rarely possible explicitly, and finding the estimators may be computationally challenging. This is often the case when estimating a location parameter based on independent, identically distributed (IID)

observations from a probability density function (PDF)

$$f(x|\theta) = h(x-\theta),\ x \in \mathbb{R}, \tag{1.1}$$

where $h(\cdot)$ is a certain symmetric and unimodal function. There are many well-documented difficulties associated with the optimization of the likelihood function

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta) \tag{1.2}$$

in this case. A good illustration of which is the case of Cauchy distribution (see [5]),

$$f(x|\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2},\ x \in \mathbb{R}, \tag{1.3}$$

where the likelihood function is multimodal (see Figure 1) and the corresponding likelihood equation have multiple roots (see, e.g., [2]), making it computationally challenging to find the right one. An additional difficulty with estimating the location parameter via the ML approach arises

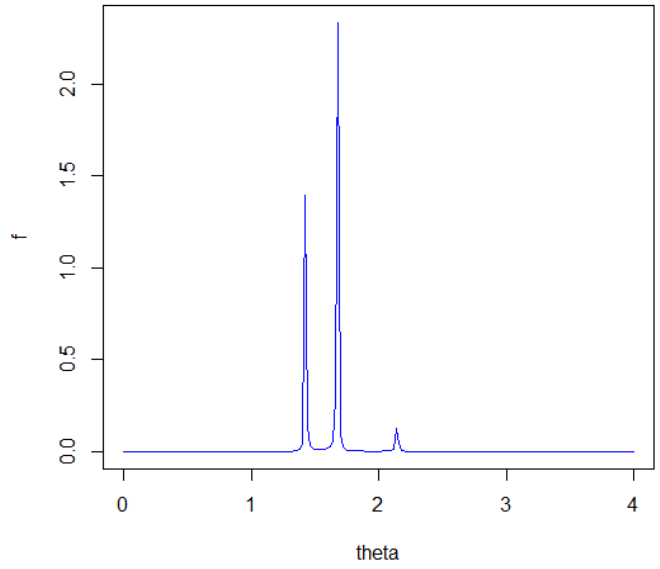

FIGURE 1. The likelihood function based on a random sample of size five from the Cauchy location model (1.3) with $\theta = 2$.

when the function $h(\cdot)$ is unbounded at the mode, as in this case the likelihood function (1.2) has

infinite values at the multiple modes that coincide with the data points (see [10]). A symmetric gamma location model with (known) shape parameter $\alpha \in (0,1)$, given by the PDF

$$f(x|\theta) = \frac{|x-\theta|^{\alpha-1}\exp(-|x-\theta|)}{2\Gamma(\alpha)},\ x \in \mathbb{R}, \tag{1.4}$$

provides a typical example of this problem. Here, the likelihood function attains a maximum value (of infinity) at each of the n points $\theta = x_i$, $i = 1,\ldots,n$, making it impossible to estimate θ by ML.

As seen above, the standard ML approach may be quite challenging. However, it is often rather straightforward when the sample size is $n = 1$. Indeed, if we only have one observation, the likelihood function in the Cauchy case (1.3) attains a maximum value with respect to θ at $\hat{\theta} = x$. The same is true in the sym-gamma case (1.4) when $n = 1$. With this in mind, we propose a novel estimation paradigm, where each individual observation in a random sample, x_i, is used to derive an estimator of an unknown parameter using the ML principle, leading to $\hat{\theta}_i = v(x_i)$. These individual estimators are subsequently combined as a weighted average to produce the final estimator,

$$\hat{\theta} = \sum_{i=1}^{n} w_i \hat{\theta}_i, \tag{1.5}$$

where the weights $\{w_i\}$ add up to one and are chosen to be proportional to the likelihood function evaluated at the estimators based on each observation, $L(\hat{\theta}_i|\mathbf{x})$, so that

$$w_i = \frac{L(\hat{\theta}_i|\mathbf{x})}{\sum_{j=1}^{n} L(\hat{\theta}_j|\mathbf{x})}. \tag{1.6}$$

We provide several examples illustrating the new method, and conduct simulation studies to assess the performance of the estimators. It turns out that this straightforward methodology produces consistent estimators, which seem to be comparable with those obtained by the ML method in large sample setting.

An important aspect of this estimation method is its relation to the Bayesian paradigm, and, in particular, to the empirical Bayes methodology. Note that $\hat{\theta}$ given by (1.5) is the mean value of a discrete random variable taking values $\{\hat{\theta}_i\}$ with probabilities $\{w_i\}$ given by (1.6). The distribution of this random variable is actually a posterior distribution in the Bayesian set-up, where the data-driven prior distribution is also discrete, supported on the same values $\{\hat{\theta}_i\}$ but with equal probabilities. In the case of estimating a location parameter of a unimodal density, the

prior distribution is the empirical distribution of the sample. This can be seen as a *non-parametric* empirical Bayes approach as described next.

Let recall that the hierarchical Bayes model can be written as

$$\begin{aligned} X_i|\theta &\sim f(x|\theta), \quad i=1,\ldots,n, \\ \Theta|\eta &\sim \lambda(\theta|\eta), \end{aligned} \tag{1.7}$$

where the parameter η is to be decided upon. For the location problem, both, the model PDF f and the prior PDF λ, arise from the location model driven by a function h:

$$f(x|\theta) = h(x-\theta),\ \lambda(\theta|\eta) = h(\theta-\eta). \tag{1.8}$$

According to the standard (parametric) empirical Bayes methodology, the unknown location parameter η of the prior is estimated from the marginal PDF of $\mathbf{X}$,

$$m(\mathbf{x}|\eta) = \int \prod_{i=1}^{n} f(x_i|\theta)\lambda(\theta|\eta)d\theta, \tag{1.9}$$

usually by maximizing the expression in (1.9) with respect to η.

In contrast, in the proposed method, the empirical distribution of the sample approximates the prior distribution $\lambda(\theta|\eta)$. The assumption that the prior in this setting coincides with the model is indeed a natural one. Since the model density is symmetric in x and θ, a particular sample value x "points" towards a likely value of the parameter θ that might have generated the x value in the same way as a particular value of θ leads to a likely x value that may show up in a sample. Collecting this information across the entire sample provides a clue about possible values of θ that might have generated the sample - which in a way is a reflection of the prior. Moreover, having $f(\theta|\theta_0)$ as the prior, where θ_0 is the true value of the parameter θ that generated the data, should be quite desirable, as in this case the prior would assign relatively more probability to a neighborhood of θ_0, in which case "it is intuitively plausible that a close approximation to the asymptotic result will tend to be achieved more quickly (i.e. for smaller n)" (cf., [8], p. 493).

We note that in our 'empirical' formulation there is no external input of any kind with regard to the prior distribution, and it is the random sample itself that essentially determines it. Moreover, our aim is to obtain a consistent estimator of a certain true generic parameter that we call θ_0, and we use the Bayesian setup primarily to show consistency of this construction in the frequentist meaning.

In this connection, it is important to distinguish two types of prior distributions for Θ: one that does not depend on data, denoted by $d\pi(\theta)$, and the one that is data dependent (which is our case), and denoted by $d\pi_n(\theta|\mathbf{x})$, where $\mathbf{x}$ represents the data and n is the data size. If these priors are re-centered at the true value θ_0, we write them as $\pi^0(u)$ and $\pi_n^0(u|\mathbf{x})$, respectively. The posterior distribution and its mean are expressed by means of the likelihood ratio process,

$$Z_n^0(u) = \frac{f_n(\mathbf{x}|\theta_0+u)}{f_n(\mathbf{x}|\theta_0)}, \tag{1.10}$$

where $f_n(\mathbf{x}|\theta_0+u)$ is the PDF of $\mathbf{X}$ given that the parameter is θ_0+u. Straightforward algebra shows that the posterior means, under these two types of priors, can be expressed using the likelihood ratio process (1.10) as follows:

$$\begin{aligned}
\hat{\theta}_b^{(n)} &= \frac{\int(\theta_0+u)Z_n^0(u)d\pi^0(u)}{\int Z_n^0(u)d\pi^0(u)} = \theta_0 + \frac{\int uZ_n^0(u)d\pi^0(u)}{\int Z_n^0(u)d\pi^0(u)}, \\
\hat{\theta}_{eb}^{(n)} &= \frac{\int(\theta_0+u)Z_n^0(u)d\pi_n^0(u|\mathbf{x})}{\int Z_n^0(u)d\pi_n^0(u|\mathbf{x})} = \theta_0 + \frac{\int uZ_n^0(u)d\pi_n^0(u|\mathbf{x})}{\int Z_n^0(u)d\pi_n^0(u|\mathbf{x})}.
\end{aligned} \tag{1.11}$$

There is a considerable body of literature regarding the asymptotics of $\hat{\theta}_b^{(n)}$ under variety of circumstances, and frequentist properties of such a 'Bayesian' estimator are well understood (see, e.g., [7]). In particular, certain regularity conditions (see Chapter 3 in [7]) for the IID case guarantee the asymptotic normality and efficiency of the estimator,

$$\lim_{n\to\infty}\sqrt{n}(\hat{\theta}_b^{(n)}-\theta_0) \stackrel{d}{=} N(0,\Sigma_0^2), \tag{1.12}$$

where $\Sigma_0^2 = I(\theta_0)^{-1}$ and $I(\theta_0)$ is the Fisher's information matrix. Although these results do not apply directly to our situation (since our empirical prior distribution is data dependent), in the important case where $\pi_n^0(u|\mathbf{x})$ converges to a certain distribution $\pi^0(u)$, one can argue that $\hat{\theta}_{eb}^{(n)}$ inherits some asymptotic properties of $\hat{\theta}_b^{(n)}$. Such a result is indeed presented in our paper. It is derived through an asymptotic closeness of $\hat{\theta}_{eb}^{(n)}$ and $\hat{\theta}_b^{(n)}$, showing that the asymptotic behavior of both estimators is similar. To the best of our knowledge, there are no readily available results on the asymptotics of Bayesian estimators derived from data-dependent priors to be utilized in our case. We believe that the asymptotic efficiency and normality in the context of data-driven prior could be established as well but we leave this topic for future studies.

Another interesting question concerns the asymptotics in the case when the empirical prior is based on the distribution of an estimator that is already consistent. The empirical prior distribution

can be based, for example, on the leave-one-out or bootstrap distribution of the estimator. As the sample size increases the prior should converge to a distribution having all mass concentrated at the true parameter. Consider, for example, the 'leave-one-out' prior that is residing on the n estimators $\hat{\theta}_i$ calculated using the sample *without* the observation x_i. In the Bayesian set-up, the discrete distribution supported on these values with equal probabilities would play the role of a *data-driven* prior distribution, whose mean

$$\hat{\theta}_{Jack} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_i \tag{1.13}$$

is recognized as the *jackknife estimator* of [11]. When this distribution is updated through the likelihood, the resulting mean of the posterior distribution will again be given by (1.5). One would expect that an estimator derived via Bayesian approach in this set-up will have attractive properties in finite sample setting.

Here is the organization of our paper. In Section 2, we describe the new estimation method, along with basic examples. Its Bayesian interpretation is discussed in Section 3, where we also address the issue of consistency. This is followed by Section 4, where we present further examples and numerical evidence in favor of the new method. Finally, more technical results are discussed in the Appendix.

2. Basic ideas and examples

Suppose that $\mathbf{x} = (x_1, \ldots, x_n)$ is a random sample from a PDF $f(x|\theta)$, where $\theta \in \Omega$ is an unknown parameter. The maximum likelihood estimator (MLE) of θ based on the ith data value is the quantity $\hat{\theta}_i = v(x_i)$ that maximizes the likelihood based on the individual x_i only,

$$\hat{\theta}_i = \text{argmax}\{f(x_i|\theta) : \ \theta \in \Omega\}. \tag{2.1}$$

Using the weights, which are given in (1.6), the estimator is defined by

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \hat{\theta}_i L(\hat{\theta}_i|\mathbf{x})}{\sum_{j=1}^{n} L(\hat{\theta}_j|\mathbf{x})}. \tag{2.2}$$

One advantage of this approach over the ordinary MLE is computational convenience. This is because it is often much easier to solve the optimization problem (2.1), in contrast with optimizing the likelihood function based on the entire sample. This is particularly true in the case of estimating

a location parameter, discussed in the introduction. To illustrate the proposed methodology, we present several simple examples

We start with the case of uniform distribution on $[0,\theta]$. The MLE of θ based on the entire sample is the largest order statistic, $\hat{\theta}_{MLE} = x_{(n)} = \max(x_1,...,x_n)$. Similarly, the MLE based on a single data point x_i is that point itself: $\hat{\theta}_i = x_i$, $i = 1,\ldots,n$. However, it is easy to see that the weight w_i assigned to $\hat{\theta}_i$ according to (1.6) is actually zero except for that corresponding to $\hat{\theta}_{(n)} = x_{(n)}$. As a result, the estimator (2.2) reduces to $\hat{\theta} = \sum_{i=1}^{n} w_i \hat{\theta}_i = x_{(n)}$ and thus coincides with the MLE.

Let us now consider estimating a scale parameter θ of an exponential distribution. The (full) likelihood function (1.2) takes on the form

$$L(\theta|\mathbf{x}) = \theta^n e^{-\theta n\bar{x}}, \tag{2.3}$$

with the MLE $\delta(\mathbf{x}) = 1/\bar{x}$. The maximum value of the likelihood based on a single data point x_i occurs at $\hat{\theta}_i = 1/x_i$, so that the estimator (2.2) takes on the form

$$\hat{\theta} = \frac{\sum_{k=1}^{n} x_k^{-n-1} e^{-n\bar{x}/x_k}}{\sum_{k=1}^{n} x_k^{-n} e^{-n\bar{x}/x_k}}. \tag{2.4}$$

We see that in the case of exponential distribution, the new estimator admits an explicit form, albeit much more complex than that of the ordinary MLE. To assess its utility, we compared the performance of the two methods in terms of their average estimated values as well as the mean squared errors (MSE) based on $k = 10,000$ simulations, with sample sizes ranging from $n = 2$ to $n = 100$. In each case the true value was chosen to be $\theta = 2$ and the MSEs were computed as $\sqrt{\frac{1}{k}\sum_{i=1}^{k}(\hat{\theta}-\theta)^2}$. As we see from the results reported in Table 1, the two approaches result in rather similar estimators.

These results indicate that the estimators do converge to the true value. However, the estimator derived via the new approach, although straightforward to compute, admits an explicit value that is a rather complex function of the data. This is not an exception: in general, the estimators obtained by this methodology are simple to compute, but their explicit forms are complicated functions of the data. Fortunately, their large sample properties can be approached using a Bayesian interpretation and utilizing the asymptotic theory for Bayesian estimators. This is discussed in the next section and summarized in Proposition 3.1.

TABLE 1. The averages and MSEs for estimators of the scale parameter θ of exponential distribution based on $10,000$ simulations: NEW – the new method and MLE – the maximum likelihood method.

		NEW	MLE
n	θ	$\hat{\theta}$(MSE)	$\hat{\theta}$(MSE)
2	2	3.63 (45.52)	3.89 (47.41)
5	2	2.4960 (2.62)	2.4964 (2.30)
10	2	2.232 (0.80)	2.230(0.70)
50	2	2.0486 (0.096)	2.0494 (0.091)
100	2	2.0195 (0.044)	2.0197 (0.042)

In conclusion of this section, let us discuss behavior of this method under transformations of either the parameter or the data. Consider a change of the parameter, $\tau = \tau(\theta)$, and transformed observations $\mathbf{Y} = (r(X_1), \ldots, r(X_n))$, where both τ and r are one-to-one functions. Let $\hat{\theta}^{(\mathbf{X})}$ be an estimator of θ obtained via the $\mathbf{X}$, given by (1.5). Similarly, let $\hat{\tau}^{(\mathbf{Y})}$ be an estimator of τ obtained through the $\mathbf{Y}$ by *the same methodology*.

Proposition 2.1. *In the above setting, suppose that $\hat{\theta}$ is an estimator of θ given by (1.5) with $\hat{\theta}_i = v(x_i)$, so that*

$$\hat{\theta}^{(\mathbf{X})} = \sum_{i=1}^{n} w_i v(x_i), \tag{2.5}$$

where the weights $\{w_i\}$ are given by (1.6) and the likelihood $L(\cdot|\mathbf{x})$ is given by (1.2). Then,

$$\hat{\tau}^{(\mathbf{Y})} = \sum_{i=1}^{n} w_i \tau(v(x_i)), \tag{2.6}$$

where the weights $\{w_i\}$ in (2.5) and (2.6) are the same.

This result follows from well-known invariance properties of the MLE on a change of parametrization. Namely, our estimate of τ based on a single data value x_i is given by $\hat{\tau}_i = \tau(\hat{\theta}_i)$, where $\theta_i = v(x_i)$ is the estimator of θ based on x_i. Moreover, it is clear that the likelihoods evaluated at θ_i and τ_i coincide. The data transformation part of the result follows from a routine change of variables argument.

3. BAYESIAN INTERPRETATION AND CONSISTENCY

As stated in the introduction, the new estimation methodology described in the previous section can be viewed as a *non-parametric empirical Bayes* method. This is because the estimator defined in (2.2) is the expectation of a discrete random variable taking on the values $\hat{\theta}_i = v(x_i)$ with probabilities given by the weights w_i specified by (1.6). Thus our estimator becomes a Bayes estimator under squared error loss function if the (data-driven) prior distribution is supported on the n values $\hat{\theta}_i = v(x_i)$, $i = 1, \ldots, n$, with equal probabilities $1/n$ (see the Appendix for more detailed arguments). When n increases, this empirical prior distribution converges to the distribution of $v(X)$, where X is a generic random variable following a model with the PDF $f(\cdot|\theta)$. The CDF of $v(X)$ is then given by $\Pi(\cdot|\theta) = F(v^{-1}(\cdot)|\theta)$, where F is the CDF associated with f and assuming the function v is one-to-one. Thus, this empirical distribution can be thought of as a non-parametric approximation to the prior distribution in the following Bayesian set-up:

$$\begin{aligned} &X_i|\theta \sim F(\cdot|\theta), \quad i = 1, \ldots, n, \\ &\Theta|\eta \sim \Pi(\cdot|\eta). \end{aligned} \tag{3.1}$$

In the above setup we assume that η is unknown. As noted in the introduction, in a location model driven by a function h, the corresponding PDFs will satisfy the relations (1.8). This should be contrasted with the standard empirical Bayes approach, where the unknown parameter η of the prior is estimated from the marginal PDF of $\mathbf{X}$, given by (1.9). That estimated value, $\hat{\eta}$, is subsequently plugged-in into a traditional Bayesian estimator derived using (3.1). Let us emphasize that the advantage of using such a non-parametric approach in the Bayesian set-up (3.1) with unknown η is its computational simplicity, as the marginal PDF given by (1.9) is usually difficult to evaluate explicitly, making the estimation of η computationally intensive.

To illustrate the above discussion, let us consider a standard example of a normal location model. For simplicity, we assume that the scale parameter is one, and we let $\mathbf{X} = X_1, \ldots, X_n$ be IID normal variables with an unknown mean $\theta \in \mathbb{R}$. The prior distribution is also assumed to be normal, with standard deviation one and mean $\eta \in \mathbb{R}$, so that

$$\pi(\theta|\eta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta-\eta)^2}, \ \theta \in \mathbb{R}. \tag{3.2}$$

It is well known that the posterior distribution of Θ given $\mathbf{X} = \mathbf{x}$ is also normal with mean

$$\mathbb{E}(\Theta|\mathbf{x}) = \frac{n}{n+1}\bar{x} + \frac{1}{n+1}\eta \tag{3.3}$$

and variance $1/(n+1)$ (cf. [8], p. 233). Consequently, under the square error loss function, the Bayesian estimate of θ is given by (3.3). Clearly, as $n \to \infty$, the estimator converges to the true value of θ, denoted by θ_0, regardless of the mean η of the prior distribution. However, in finite samples, the above Bayesian estimator may have a considerable bias, if the prior mean is not close enough to the true parameter θ_0.

On the other hand, the new method introduced in this paper does not utilize any external information regarding θ. Instead, the empirical distribution of the sample approximates the distribution of the prior, which is assumed to be *exactly the same* as the model itself (with $\theta = \theta_0$). This particular prior should be quite desirable, as *it is centered about the true value of* θ and consequently will minimize the Bias as seen above. Since such a prior in practice can not be used (as θ_0 is unknown), approximating it through data seems quite natural.

Assuming that the prior and the model are both normal distributions, brings us to the Bayesian set-up (3.1), where now the parameter $\eta \in \mathbb{R}$ is unknown and needs to be estimated from data. According to the classical empirical Bayes, the estimator of θ is still the posterior mean given by (3.3), where the unknown quantity η is replaced by its (parametric) estimate $\hat{\eta}$, such as the MLE obtained from the marginal distribution of the $\{X_i\}$. This case is rather exceptional, in that such an estimate is available in an explicit form. Indeed, since the marginal PDF of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sqrt{n+1}} e^{-\frac{1}{2}\left[\sum_{i=1}^n x_i^2 + \eta^2 - \frac{(n\bar{x}+\eta)^2}{n+1}\right]}, \ \mathbf{x} \in \mathbb{R}^n, \tag{3.4}$$

the MLE of η is the value of η that minimizes the quantity

$$\sum_{i=1}^n x_i^2 + \eta^2 - \frac{(n\bar{x}+\eta)^2}{n+1}$$

that appears in the exponent. Since this is a simple quadratic function of η, it is easy to see that the MLE is given by $\hat{\eta} = \bar{x}$. Plugging this into (3.3) shows that the empirical Bayes estimate of θ is actually the same, $\hat{\theta}_{EB} = \bar{x}$. Thus in this case the standard empirical Bayes approach coincides with the maximum likelihood method

Although our new methodology in the context of this example also starts with the same Bayesian set-up with an unknown θ, it differs from the above in the way we approximate the

posterior mean, which is now computed according to (2.2):

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j - x_i)^2}}{\sum_{i=1}^{n} e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j - x_i)^2}}. \tag{3.5}$$

As we can see, the new estimator of the normal mean is much more complex then the standard estimator provided by the sample mean.

To compare the performance of these two "empirical Bayes" estimators, we run a simulation experiment where we computed the averages and the MSEs connected with the two methods based on $k = 10,000$ simulations. Without loss of generality, we assumed the true value of the normal mean was $\theta_0 = 3$ and the standard deviation was taken as 1. The results appear in Table 2. As we can see, the performance of the estimators is rather close, with the new method having a slightly larger bias and MSE.

TABLE 2. A comparison of two empirical Bayes methods when estimating the mean $\theta = 3$ of normal distribution with standard deviation equal to 1.

n	μ (true)	$\hat{\mu}$(MSE) [NEW]	$\hat{\mu}$(MSE) [OLD]
5	3	2.995 (0.50)	2.999 (0.45)
11	3	2.997 (0.33)	2.999 (0.31)
25	3	2.998 (0.21)	2.998 (0.20)
51	3	3.000 (0.15)	3.000 (0.14)
101	3	3.003 (0.10)	3.003 (0.10)
501	3	3.000 (0.05)	3.000 (0.04)

In our approach, we utilize a Bayes estimator with the (empirical) prior distribution $\pi_n(\theta|\mathbf{x})$ obtained on the basis of a random sample $\hat{\theta}_i = v(X_i)$. This data driven distribution "points" towards a particular prior distribution. Namely, as the sample size increases, this *empirical prior* converges to that of $v(X)$, where X is a random variable with the PDF $f(x|\theta)$ centered about the true value $\theta = \theta_0$ and that can be denoted as $\pi(\theta)$. It is expected that the estimates obtained by this method will be consistent. The rationale behind this claim is provided by the well-established theory of Baysian estimators such as the mean $\hat{\theta}_b^{(n)}$ of the posterior distribution that is given by (1.11), see, e.g, [7]). It guarantees asymptotic normality of the Bayesian estimators for a large class of

non-data driven prior distributions. Frequently, one can assume some consistency between our empirical and non-empirical priors, π and $\pi_n(\cdot|\mathbf{x})$, such as

$$\lim_{n\to\infty} n^{1/2-\delta} \sup_{\theta} |\pi_n(\theta|\mathbf{x}) - \pi(\theta)| = 0 \tag{3.6}$$

for each $\delta > 0$. It is then natural to expect that $\hat{\theta}_{eb}^{(n)}$ given in (1.11), derived on the basis of a data-dependent prior, will behave similarly to $\hat{\theta}_b^{(n)}$.

We make these claims more precise for the special case of the location parameter. By virtue of the the empirical central limit theorem, see, e.g., [3], (3.6) holds for the empirical distribution of the observations that is taken as an empirical prior. Consequently, the empirical prior approximates the prior $\pi(\theta)$ distributed according to $f(\theta|\theta_0)$. This leads to the following consistency result for which a formal argument is sketched in the Appendix.

Proposition 3.1. *Let $\hat{\theta}_b^{(n)}$ be the posterior mean under the prior given by $\pi(\theta) = f(\theta|\theta_0)$, for the location problem with the true parameter θ_0. We assume asymptotic normality for this estimator as given in (1.12). Then, for each $\delta > 0$, we have*

$$\lim_{n\to\infty} n^{1/2-\delta} \left| \hat{\theta}_{eb}^{(n)} - \hat{\theta}_b^{(n)} \right| = 0.$$

Consequently, $\hat{\theta}_{eb}^{(n)}$ is also a consistent estimator of θ_0.

4. Further examples and simulations

Here we illustrate the new method in several common settings, including location and scale families, as well as shape parameters and mixtures. Rather than presenting comprehensive analyses, which are left for further studies, we want to illustrate the method with typical examples, and check its performance through (few) selected simulation experiments.

4.1. **Location families.** We start with a location model (1.1), defined through a unimodal density $h(\cdot)$, which is symmetric about the origin and supported on the entire real line. This generalizes the case of normal location model, presented in Section 2. It is clear that, similarly to the normal case, the MLE of the location parameter θ based on a single observation x_i, is given by $\hat{\theta}_i = x_i$. Moreover, the likelihood function evaluated at $\hat{\theta}_i$ is given by

$$L(\hat{\theta}_i|\mathbf{x}) = \prod_{j=1}^{n} h(x_j - x_i),\ i = 1,\ldots,n, \tag{4.1}$$

assuming the function h is bounded at the mode. Thus, according to (1.5), our estimator of θ takes on the form

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i \prod_{j=1}^{n} h(x_j - x_i)}{\sum_{i=1}^{n} \prod_{j=1}^{n} h(x_j - x_i)}. \tag{4.2}$$

In the normal case, this reduces to (3.5) derived in Section 2.

4.1.1. *Cauchy distribution.* Consider a Cauchy distribution, mentioned in the introduction, given by the PDF

$$f(x|\theta) = \frac{1}{\pi} \frac{1}{1 + (x-\theta)^2}, \ x \in \mathbb{R}. \tag{4.3}$$

This location model clearly satisfies the conditions stated above, so our estimator of θ is given by (4.2), leading to

$$\hat{\theta} = \frac{\sum_{i=1}^{n} y_i \prod_{j=1}^{n} \left(1 + |\, y_j - y_i \,|^2\right)^{-1}}{\sum_{i=1}^{n} \prod_{j=1}^{n} \left(1 + |\, y_j - y_i \,|^2\right)^{-1}}. \tag{4.4}$$

Thus, in contrast with the MLE of θ, which may require challenging numerical approximations, this estimator is straightforward to compute.

4.1.2. *Symmetric gamma distribution.* Consider a symmetric gamma location model, given in (1.4). This density is unimodal and symmetric, with the mode at θ, whenever $\alpha \in (0, 1]$. When $\alpha = 1$, we obtain a Laplace distribution, with the PDF bounded at the mode and our estimator simplifies to

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i e^{-\sum_{j=1}^{n} |x_j - x_i|}}{\sum_{i=1}^{n} e^{-\sum_{j=1}^{n} |x_j - x_i|}}. \tag{4.5}$$

However, when α is strictly less than 1, we encounter a problem: the function h is *unbounded* at the mode. Consequently, although we still have $\hat{\theta}_i = x_i$, the likelihood (4.1) evaluated at $\hat{\theta}_i$ is equal to infinity, and so the weights (1.6) are not well defined. There is a simple remedy to address this issue. Namely, the likelihood (4.1) is replaced by

$$\tilde{L}(\hat{\theta}_i|\mathbf{x}) = \prod_{j \neq i} h(x_j - x_i), \ i = 1, \ldots, n, \tag{4.6}$$

and the weights (1.6) are modified accordingly:

$$\tilde{w}_i = \frac{\tilde{L}(\hat{\theta}_i|\mathbf{x})}{\sum_{j=1}^{n} \tilde{L}(\hat{\theta}_j|\mathbf{x})}. \tag{4.7}$$

We then define the estimator as

$$\hat{\theta} = \sum_{i=1}^{n} \hat{\theta}_i \tilde{w}_i = \frac{\sum_{i=1}^{n} \hat{\theta}_i \tilde{L}(\hat{\theta}_i|\mathbf{x})}{\sum_{i=1}^{n} \tilde{L}(\hat{\theta}_i|\mathbf{x})} = \frac{\sum_{i=1}^{n} x_i \prod_{j\neq i} h(x_j - x_i)}{\sum_{i=1} \prod_{j\neq i}^{n} h(x_j - x_i)}. \tag{4.8}$$

Note that when the density h is bounded at the mode, so that $h(x_i - x_i) = h(0) = c < \infty$, the estimators (4.2) and (4.8) actually coincide.

Remark 4.1. When we define the modified likelihood at $\hat{\theta}_i$, we simply delete the observation x_i from the sample. This *leave-one-out* procedure is similar in spirit to that considered in [10], used to address the issue of unbounded density when estimating a location parameter via maximum likelihood.

With the above modification, our estimator of the location parameter θ in the symmetric gamma case (1.4) with $0 < \alpha < 1$ becomes

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i e^{-\sum_{j\neq i} |x_j - x_i|} \prod_{j\neq i} |x_j - x_i|^{\alpha-1}}{\sum_{i=1}^{n} e^{-\sum_{j\neq i} |x_j - x_i|} \prod_{j\neq i} |x_j - x_i|^{\alpha-1}}. \tag{4.9}$$

4.2. **Scale families.** Consider a scale family P_σ of distributions on $\mathbb{R}_+$, given by the PDF

$$f(x|\sigma) = g(x/\sigma)/\sigma \tag{4.10}$$

where $g(\cdot)$ is a PDF on $\mathbb{R}_+$. Many standard distributions that follow this structure, including the exponential model discussed in Section 2, satisfy the following condition:

(A) The function $u \to ug(u)$ is unimodal on $(0,\infty)$, with the mode $m \in (0,\infty)$.

It is easy to see, that if a scale family (4.10) satisfies condition (A), then the MLE of σ based on an individual value x_i is given by $\hat{\sigma}_i = x_i/m$. Further, the likelihood function evaluated at $\hat{\sigma}_i$ is then given by

$$L(\hat{\sigma}_i|\mathbf{x}) = (m/x_i)^n \prod_{j=1}^{n} g(mx_j/x_i),\ i = 1,\ldots,n, \tag{4.11}$$

assuming the function g is bounded. Thus our estimator of σ takes on the form

$$\hat{\sigma} = \frac{1}{m} \frac{\sum_{i=1}^{n} x_i^{1-n} \prod_{j=1}^{n} g(mx_j/x_i)}{\sum_{i=1}^{n} x_i^{-n} \prod_{j=1}^{n} g(mx_j/x_i)}. \tag{4.12}$$

Remark 4.2. Suppose we have a scale family of distributions on the entire real line rather than on $\mathbb{R}_+$, with the same structure (4.10) now defined through a PDF g on $\mathbb{R}$. Then, if the function g

is symmetric with respect to the origin and satisfies condition (A), the new estimator of the scale parameter is also given by (4.12), but with the $\{x_i\}$ replaced by their absolute values, so that

$$\hat{\sigma} = \frac{1}{m} \frac{\sum_{i=1}^{n} |x_i|^{1-n} \prod_{j=1}^{n} g(m|x_j/x_i|)}{\sum_{i=1}^{n} |x_i|^{-n} \prod_{j=1}^{n} g(m|x_j/x_i|)}. \tag{4.13}$$

Remark 4.3. If the scale family (4.10) is expressed through a parameter $\theta = 1/\sigma$, and the condition (A) holds for the function g, then, by Proposition 2.1, w_i's in the formula for $\hat{\theta}$ coincides with those in the formula for $\hat{\sigma}$. As a result, our method produces the following estimator of θ:

$$\hat{\theta} = m \frac{\sum_{i=1}^{n} x_i^{-1-n} \prod_{j=1}^{n} g(mx_j/x_i)}{\sum_{i=1}^{n} x_i^{-n} \prod_{j=1}^{n} g(mx_j/x_i)}. \tag{4.14}$$

When a scale family is supported on the entire real line and is expressed through $\theta = 1/\sigma$, then in the context of our previous remark, a similar adjustment applies to (4.13).

4.2.1. *Gamma distribution.* Consider a gamma distribution with known shape parameter $\alpha \in (0,\infty)$ and an unknown scale parameter $\sigma \in (0,\infty)$, given by the PDF (4.10) with

$$g(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x},\ x \in \mathbb{R}_+. \tag{4.15}$$

Routine calculations show that the condition (A) is fulfilled with $m = \alpha$. Thus, here we have $\hat{\sigma}_i = x_i/\alpha$ and the likelihood function evaluated at $\hat{\sigma}_i$ is of the form

$$L(\hat{\sigma}_i|\mathbf{x}) = \left(\frac{\alpha^\alpha}{x_i^\alpha \Gamma(\alpha)}\right)^n \left(\prod_{j=1}^{n} x_j\right)^{\alpha-1} e^{-\frac{\alpha n}{x_i}\bar{x}},\ i = 1,\ldots,n, \tag{4.16}$$

In turn, our estimator of the scale parameter becomes

$$\hat{\sigma} = \frac{1}{\alpha} \frac{\sum_{i=1}^{n} x_i^{1-\alpha n} e^{-\frac{\alpha n}{x_i}\bar{x}}}{\sum_{i=1}^{n} x_i^{-\alpha n} e^{-\frac{\alpha n}{x_i}\bar{x}}}. \tag{4.17}$$

4.2.2. *Cauchy distribution.* Consider a Cauchy distribution with the PDF

$$g(x) = \frac{1}{\pi} \frac{1}{1+x^2},\ x \in \mathbb{R}. \tag{4.18}$$

Since the function g given above is symmetric with respect to the origin and satisfies the condition (A) with $m = 1$, the above discussion (see Remark 4.2) shows that $\hat{\sigma}$ is given by (4.13). Further routine simplifications produce

$$\hat{\sigma} = \frac{\sum_{i=1}^{n} |y_i|^{1-n} \prod_{j=1}^{n} \left(1 + y_j^2/y_i^2\right)^{-1}}{\sum_{i=1}^{n} |y_i|^{-n} \prod_{j=1}^{n} \left(1 + y_j^2/y_i^2\right)^{-1}}. \tag{4.19}$$

4.2.3. *Pareto distribution.* Consider a Pareto II (Lomax) distribution (see, e.g., [1]), given by the PDF

$$f(x|\theta,\beta) = \theta\beta\left[\frac{1}{1+\theta x}\right]^{(\beta+1)},\ x\in\mathbb{R}_+, \tag{4.20}$$

where $\theta > 0$ is an unknown scale parameter and $\beta > 0$ is a given tail parameter. This family is of the form (4.10), where $\sigma = 1/\theta$ and the function g is given by the right-hand-side of (4.20) with $\theta = 1$. Straightforward algebra shows that the function g satisfies the condition (A) with $m = 1/\beta$, so that according to Remark 4.3, our estimator $\hat{\theta}$ is given by (4.14). After routine simplifications we obtain

$$\hat{\theta} = \frac{1}{\beta}\frac{\sum_{i=1}^n x_i^{-1-n}\prod_{j=1}^n\left(1+\frac{x_j}{\beta x_i}\right)^{-1-\beta}}{\sum_{i=1}^n x_i^{-n}\prod_{j=1}^n\left(1+\frac{x_j}{\beta x_i}\right)^{-1-\beta}}. \tag{4.21}$$

Thus, we obtained an estimator of the scale parameter with an explicit form. In contrast, finding the MLE of θ in this setting would require a numerical search.

4.3. **Shape/tail parameters.** Below we examine two examples other than a location or shape parameters. The first one involving a shape parameter of gamma distribution and the other one a tail parameter of Pareto distribution.

4.3.1. *Shape parameter/gamma distribution.* Without loss of generality, we consider standard gamma distribution with shape parameter $\alpha \in (0,\infty)$, given by the PDF (4.15). The derivative of the log-likelihood function based on n values $\mathbf{x} = (x_1,\ldots,x_n)$ of a random sample from this distribution takes on the form

$$l(\alpha|\mathbf{x}) = \frac{\partial}{\partial\alpha}L(\alpha|\mathbf{x}) = n\left\{\frac{1}{n}\sum_{j=1}^n \log x_j - \psi(\alpha)\right\},\ \alpha\in\mathbb{R}_+. \tag{4.22}$$

The quantity

$$\psi(\alpha) = \frac{d}{d\alpha}\log\Gamma(\alpha),\ \alpha\in\mathbb{R}_+, \tag{4.23}$$

is the digamma function, which is known to be monotonically increasing on $(0,\infty)$, with the limits of $-\infty$ and ∞ at 0 and ∞, respectively. It follows from (4.22) that there exists a unique MLE of α,

given by

$$\hat{\alpha}_{MLE} = \psi^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\log x_j\right), \tag{4.24}$$

but finding it requires a numerical search. When we apply our new method, the same analysis shows that the MLE of α based on a single data value x_i is given by $\hat{\alpha}_i = \psi^{-1}(\log x_i)$, and requires a similar numerical search as well. Thus, in this case the new estimation method does not seem to have a computational advantage over the traditional approach. However, under its slight modification, this numerical search can be avoided. Namely, as discussed at the end of the introduction, one might follow an approach involving the jackknife, which in this case can be coupled with the method of moments. Since the expectation of gamma distribution given by the PDF (4.15) is equal to α, to estimate the latter one can use the sample mean, leading to $\hat{\alpha}_{MM} = \bar{x}$. Alternatively, for each $i \in \{1,\ldots,n\}$, one can eliminate the x_i from the sample and derive a moment estimate of α based on the rest of the sample, producing

$$\hat{\alpha}_i = \frac{1}{n-1}\sum_{j\neq i} x_j. \tag{4.25}$$

The average of these would produce the jackknife estimator (see, e.g., [11]), given by (1.13) (with α in place of θ). This can be seen as the expectation of a data-driven prior distribution in a Bayesian set-up, supported on the n values of $\{\hat{\alpha}_i\}$, taken with equal probabilities. We can now go a step further, and compute the new estimator by "updating" the prior distribution via Bayes theorem and evaluate the posterior mean. Since the prior distribution converges to a point mass (concentrated at the true value of the shape parameter) as the sample size increases, one might expect that the estimator obtained through this scheme will perform well.

4.3.2. *Tail parameter/Pareto distribution.* Consider again a Pareto II (Lomax) distribution given by the PDF (4.20), where this time we are interested in estimating the tail parameter $\beta > 0$ rather than the scale parameter θ, which is assumed to be known. Since the derivative

$$\frac{\partial \log f(x_i|\theta,\beta)}{\partial \beta} = \frac{1}{\beta} - \log(1+\theta x_i) \tag{4.26}$$

is positive if and only if the value of β is less than

$$\hat{\beta}_i = \frac{1}{\log(1+\theta x_i)}, \tag{4.27}$$

the latter is the MLE of β based on a single data value x_i. When we take into account the likelihood evaluated at $\hat{\beta}_i$, we obtain an estimator of the tail parameter of the form

$$\hat{\beta} = \frac{\sum_{i=1}^{n}[\log(1+\theta x_i)]^{-1-n}\prod_{j=1}^{n}(1+\theta x_j)^{-\frac{1}{\log(1+\theta x_i)}}}{\sum_{i=1}^{n}[\log(1+\theta x_i)]^{-n}\prod_{j=1}^{n}(1+\theta x_j)^{-\frac{1}{\log(1+\theta x_i)}}}. \tag{4.28}$$

Remark 4.4. As we can see the tail parameter of Pareto II distribution can be estimated explicitly by our method, although the formula is rather complicated. However, its advantage over the MLE of β is unclear, as the latter is also available in an explicit form,

$$\hat{\beta}_{MLE} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\log(1+\theta x_i)}, \tag{4.29}$$

which can be derived along the same lines as above. Under a close inspection, it turns out that both estimators are essentially the same as those in the exponential case (cf. (2.4) given in Section 2), with obvious changes in notation, when we set $y_i = \log(1+\theta x_i)$. This is due to the fact that methods based on the likelihood function are invariant under monotonic data transformations (see Proposition 2.1). Indeed, the transformation turns a Pareto variable with scale θ and shape β into an exponential variable with scale parameter β.

4.4. **Mixtures.** Mixture models have numerous practical applications, and offer many challenging problems for a statistician (see, e.g., [4]). Here, we consider the problem of estimating the mixing parameter θ in

$$f(x|\theta) = \theta f_1(x) + (1-\theta) f_2(x), \tag{4.30}$$

where f_1 and f_2 are two given PDFs. It is well known that the likelihood equation connected with the function (4.30) has at most one root, since the second derivative of the log-likelihood function is negative (cf, [8], Example 4.7, p. 456). Consequently, ML approach to this problem should always produce a solution (if there is one), which, in general, will require a numerical search.

On the other hand, our new methodology adapted to this problem is computationally much simpler. It can be easily seen by taking the derivative, that the likelihood function (4.30) based on a single data point is either increasing or decreasing in θ, so that $\hat{\theta}_i \in \{0,1\}$. Moreover, we have

$$\hat{\theta}_i = I_{(0,\infty)}(f_1(x_i) - f_2(x_i)), \tag{4.31}$$

where I_A is an indicator function of the set A. Consequently, the likelihood $L(\hat{\theta}_i|\mathbf{x})$ takes on only two values, denoted by C_1 and C_2, where

$$C_1 = L(1|\mathbf{x}) = \prod_{i=1}^{n}[1 \cdot f_1(x_i) + (1-1) \cdot f_2(x_i)] = \prod_{i=1}^{n} f_1(x_i),$$
$$C_2 = L(0|\mathbf{x}) = \prod_{i=1}^{n}[0 \cdot f_1(x_i) + (1-0) \cdot f_2(x_i)] = \prod_{i=1}^{n} f_2(x_i). \tag{4.32}$$

Let $A = \{i : \hat{\theta}_i = 1\}$, and let k be the cardinality of A. Then, routine calculations show that our estimator is

$$\hat{\theta} = \sum_{i \in A} \frac{1 \cdot L(1|\mathbf{x})}{\sum_{j=1}^{n} L(\hat{\theta}_i|\mathbf{x})} = \frac{kC_1}{kC_1 + (n-k)C_2}. \tag{4.33}$$

Note that if the set A is empty, so that $f_1(x_i) < f_2(x_i)$ for each $i \in \{1, \ldots, n\}$, then its cardinality k is zero and so is $\hat{\theta}$. In this case the data point towards the PDF f_2 alone (no mixing). Similarly, if $f_1(x_i) > f_2(x_i)$ for each $i \in \{1, \ldots, n\}$ then the cardinality k of A equals n, and we have $\hat{\theta} = 1$ with no mixing as well. Otherwise, the estimator will be strictly between 0 and 1.

Remark 4.5. The estimator given above never produces a value outside of $\Omega = [0, 1]$. In contrast, this could be the case with the moment estimator discussed in [8] (Example 4.7, p. 456), given by

$$\tilde{\theta} = \frac{\bar{x} - \mu_2}{\mu_1 - \mu_2}, \tag{4.34}$$

where μ_i is the mean corresponding to the PDF f_i.

4.5. **Simulations.** Here we present a handful of simulation experiments that illustrate performance of the proposed estimator. Our focus is on location and scale parameters. Further examples concerning a shape parameter can be found in [6]. In all examples presented below we simulated $10,000$ samples of given size from a relevant distribution. These simulations led to the average value of the estimate as well as its mean squared error (MSE) shown in the tables below.

4.5.1. *Cauchy distribution.* Table 3 provides the results of a simulation study concerning a Cauchy distribution with unknown scale or location parameters. In each case, the other parameter is assumed to be known. When the sample size increases, the estimated values become more precise, as expected from the asymptotic results discussed in Section 3.

Our method can be adopted to a more general case of more than one unknown parameter. Here we illustrate this approach using both the parameters of a Cauchy distribution. Namely, when both θ and σ are unknown in the Cauchy case, one can proceed in an iterative manner, where

TABLE 3. Simulation results for Cauchy distribution

n	σ	$\hat{\sigma}$(MSE)	θ	$\hat{\theta}$(MSE)
2	3	19.33 (69847)	2	6.68 (108583)
5	3	4.55 (21.17)	2	1.95 (7.94)
10	3	3.67 (4.11)	2	2.00 (2.60)
50	3	3.12 (0.44)	2	2.01 (0.40)
100	3	3.07 (0.20)	2	2.01 (0.19)

at each step only one parameter is estimated. Thus, starting with an initial value $\hat{\sigma}^{(0)}$ of σ, we first estimate the parameter θ according to (4.4), where σ is set to $\hat{\sigma}^{(0)}$, leading to $\hat{\theta}^{(1)}$. Next, we re-estimate σ via (4.19), where θ is set to $\hat{\theta}^{(1)}$, leading to an updated estimate of σ, denoted by $\hat{\sigma}^{(1)}$. This procedure is then iterated, leading to final estimates of both parameters. Table 4 presented below is a result of a simulation experiment using such iterations. It shows that the resulting estimators are quite reasonable, even when the number of iterations is as low as 2. We emphasize again that this is just an illustration of our methodology, and more studies are needed for its comprehensive evaluation.

TABLE 4. Results for the Cauchy distribution when both the parameters are unknown

n	Initial σ	# of iterations	σ	θ	$\hat{\sigma}$	$\hat{\theta}$
50	20	1	2	2.11	3	2.43
50	20	2	2	1.97	3	2.86
50	20	50	2	1.97	3	2.87
500	10	1	2	2.07	3	3.05
500	10	2	2	2.07	3	3.08
500	10	50	2	2.07	3	3.08

4.5.2. *Symmetric gamma distribution.* Our second example concerns sym-gamma location model, given by the PDF (1.4). As discussed in the introduction, when the (known) shape parameter α is less than one, the likelihood function is unbounded at each of the data points, making it impossible to estimate the unknown location parameter θ by traditional maximum likelihood. Likelihood-based estimation of a location parameter when the PDF $f(x|\theta) = h(x-\theta)$ is unbounded at the mode, such as this one, was recently considered in [10]. In that paper, the authors proposed a method based on maximization of a modified likelihood,

$$L_{LOO}(\theta|\mathbf{x}) = \frac{\prod_{i=1}^{n} h(x_i-\theta)}{h\left(x_{k(\theta)}-\theta\right)}, \tag{4.35}$$

where $x_{k(\theta)}$ is the data point which is closest to θ. Thus, in this *leave-one-out* (LOO) likelihood one of the factors is removed from the likelihood, when θ enters a neighborhood of singularity. As shown in [10], the above procedure leads to consistent estimators of θ, although, in general, it requires a numerical search. It is interesting to observe that in our case of sym-gamma distribution with $\alpha < 1$, the estimator of θ is actually available in closed form. Indeed, we have the following result.

Proposition 4.6. *Let* $\mathbf{x} = (x_1, \ldots, x_n)$ *be an observed random sample from a sym-gamma distribution given by the PDF (1.4) with* $\alpha < 1$. *Further, let* $x_{(1)} < \cdots < x_{(n)}$ *be the ordered values of* $\mathbf{x}$, *and let* $y_i = (x_{(i)} + x_{(i+1)})/2$, $i = 1, \ldots, n-1$. *Then, the LOO likelihood (4.35) attains its maximum value at one of the* $\{y_i\}$.

Proof. Upon close examination of the LOO likelihood in this case, we find it to be a continuous function of θ on the real line, which is increasing on the interval $(-\infty, y_1)$, decreasing on the interval (y_{n-1}, ∞), and convex on each of the intervals (y_i, y_{i+1}), $i = 1, \ldots, n-1$. □

According to the above result, the LOO maximum likelihood estimator of θ in the sym-gamma case is given by that value of y_i that maximizes the LOO likelihood (4.35). We now compare the performance of this estimator with that of ours, which in this case is given in (4.9). Note that when the sample size is $n = 2$, then both estimators are given by $y_1 = (x_1 + x_2)/2$. For simplicity, we examine the case $n = 3$, where the estimators are no longer the same. In this case, the LOO method produces

$$\hat{\theta}_{LOO} = \operatorname{argmax} \{L_{LOO}(y_1|\mathbf{x}), L_{LOO}(y_2|\mathbf{x})\}.$$

TABLE 5. Estimated values of location parameter θ of symmetric gamma distribution based on sample size $n = 3$

α	θ (true)	$\hat{\theta}_{NEW}$(MSE)	$\hat{\theta}_{LOO}$(MSE)
0.20	5	5.0001 (0.153)	5.0005 (0.178)
0.30	5	5.0037 (0.237)	5.0043 (0.268)
0.40	5	5.0035 (0.303)	5.0040 (0.344)
0.80	5	4.9912 (0.627)	4.9924 (0.732)
0.99	5	5.0077 (0.786)	5.0111 (0.919)
0.99	35	35.0012(0.792)	34.9991 (0.927)
0.50	35	35.0055 (0.404)	35.0066 (0.468)
0.30	35	34.9952 (0.251)	34.9950 (0.285)
0.15	35	35.0019 (0.105)	35.0027 (0.118)

On the other hand, our estimator is given by (4.9) with $n = 3$. A comparison of the two methods based on $k = 10,000$ simulations is shown in Table 5. We see the results are comparable, with our proposed estimator performing slightly better.

4.5.3. *Laplace distribution.* Our final example is the Laplace distribution, which is a special case $\alpha = 1$ of the sym-gamma distribution (1.4), given by the PDF

$$f(x|\theta) = \frac{1}{2}e^{-|x-\theta|},\ x \in \mathbb{R}. \tag{4.36}$$

For this distribution, the likelihood function is maximized by the sample median, which becomes the MLE of θ (see, e.g., [9]). In our approach, we maximize the likelihood based on each x_i, leading to $\hat{\theta}_i = x_i$, which produces

$$\hat{\theta}_{NEW} = \frac{\sum_{i=1}^{n} x_i e^{-\sum_{j=1}^{n} |x_j - x_i|}}{\sum_{i=1}^{n} e^{-\sum_{j=1}^{n} |x_j - x_i|}}. \tag{4.37}$$

In our comparison, we shall also include a third estimator of θ, derived by the method of moments. Since the mean of the Laplace distribution (4.36) is equal to θ, the third estimator is simply the sample mean, $\hat{\theta}_{MME} = \bar{x}$. Table 6 shows a comparison of the three methods based on $k = 10,000$

TABLE 6. Simulation results concerning estimation of the location parameter θ of symmetric Laplace distribution.

n	θ (true)	$\hat{\theta}_{NEW}$(MSE)	$\hat{\theta}_{MME}$(MSE)	$\hat{\theta}_{MLE}$(MSE)
5	3	2.9992 (0.592)	3.0007 (0.634)	2.9990 (0.599)
11	3	3.0027 (0.368)	3.0041 (0.432)	3.0037 (0.374)
25	3	2.9993 (0.224)	2.9977 (0.281)	2.9997 (0.229)
51	3	2.9999 (0.153)	3.0009 (0.199)	2.9997 (0.155)
101	3	3.0013 (0.107)	3.0009 (0.141)	3.0015 (0.108)
501	3	3.0000 (0.045)	3.0010 (0.062)	3.0001 (0.0455)

simulations across several sample sizes n. It appears the new method is quite comparable with the MLE, and they both perform better than the method of moments.

5. APPENDIX

5.1. **Bayesian interpretation.** Let $X_1,...,X_n$ be a random sample from a distribution given by the PDF $f(x|\theta)$, where $\theta \in \Omega$ is an unknown parameter. Further, let the prior distribution π of Θ be a discrete one, concentrated on n values $a_i \in \Omega$ with equal probabilities. Then the joint PDF of $\mathbf{X} = (X_1,...,X_n)$ and Θ is given by

$$h(\mathbf{x},\theta) = \begin{cases} \frac{1}{n}\prod_{j=1}^{n} f(x_j|\theta) & \text{for } x_j \in \mathbb{R} \text{ and } \theta = a_i, i = 1,...,n \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

Consequently, the marginal PDF of $\mathbf{X}$ is

$$g_n(\mathbf{x}) = \sum_{\theta} h(\mathbf{x},\theta) = \sum_{i=1}^{n} \frac{1}{n} \prod_{j=1}^{n} f(x_j|a_i),\ x_j \in \mathbb{R}, j = 1,...,n, \tag{5.2}$$

while the conditional PDF of Θ given $\mathbf{X} = \mathbf{x}$, that is the posterior PDF of Θ, is given by

$$\pi(\theta|\mathbf{x}) = \frac{h(\mathbf{x},\theta)}{g_n(\mathbf{x})} = \frac{\prod_{j=1}^{n} f(x_j|\theta)}{\sum_{k=1}^{n}\prod_{j=1}^{n} f(x_j|a_k)} = \frac{L(\theta|\mathbf{x})}{\sum_{k=1}^{n} L(a_k|\mathbf{x})},\ \theta = a_i, i = 1,2,...,n. \tag{5.3}$$

If we now formally substitute $a_i = \hat{\theta}_i$, the posterior distribution corresponds to a random variable taking on the values $\hat{\theta}_i$ with probabilities given by the weights w_i in (1.6), and the mean of this distribution coincides with the estimator (2.2).

5.2. **Consistency argument.** In this section, we sketch formal arguments for the asymptotic behavior of $\hat{\theta}_{eb}^{(n)}$ presented in Proposition 3.1. Let us first note that by introducing $\tilde{Z}_n^0(s) = Z_n^0(s/\sqrt{I(\theta_0)n})$, the difference between an empirical Bayes and Bayes estimators can be written as follows:

$$\begin{aligned}\hat{\theta}_{eb}^{(n)} - \hat{\theta}_b^{(n)} &= \frac{\int u\, Z_n^0(u)\, d\pi^0(u)}{\int Z_n^0(u)\, d\pi^0(u)} - \frac{\int u\, Z_n^0(u)\, d\pi_n^0(u|\mathbf{x})}{\int Z_n^0(u)\, d\pi_n^0(u|\mathbf{x})} \\ &= \frac{1}{\sqrt{I(\theta_0)n}} \left(\frac{\int s\, \tilde{Z}_n^0(s)\; d\pi^0\left(s/\sqrt{I(\theta_0)n}\right)}{\int \tilde{Z}_n^0(s)\; d\pi^0\left(s/\sqrt{I(\theta_0)n}\right)} - \frac{\int s\, \tilde{Z}_n^0(s)\; d\pi_n^0\left(s/\sqrt{I(\theta_0)n}\middle|\mathbf{x}\right)}{\int \tilde{Z}_n^0(s)\; d\pi_n^0\left(s/\sqrt{I(\theta_0)n}\middle|\mathbf{x}\right)} \right).\end{aligned}$$

Thus, by denoting $\tilde{\pi}_n^0(s) = \pi^0\left(s/\sqrt{I(\theta_0)n}\right)$ and $\tilde{\pi}_n^0(s|\mathbf{x}) = \pi_n^0\left(s/\sqrt{I(\theta_0)n}\middle|\mathbf{x}\right)$, we have

$$\begin{aligned}\left|\hat{\theta}_{eb}^{(n)} - \hat{\theta}_b^{(n)}\right| \le{}& \frac{\left|\int s\, \tilde{Z}_n^0(s)\; d\left(\tilde{\pi}_n^0(s) - \tilde{\pi}_n^0\left(s|\mathbf{x}\right)\right)\right|}{\sqrt{I(\theta_0)n}\int \tilde{Z}_n^0(s)\; d\tilde{\pi}_n^0(s)} + \\ &+ \frac{\int s\, \tilde{Z}_n^0(s)\; d\tilde{\pi}_n^0\left(s|\mathbf{x}\right)}{\sqrt{I(\theta_0)n}\int \tilde{Z}_n^0(s)\, d\tilde{\pi}_n^0(s) \int \tilde{Z}_n^0(u)\, d\tilde{\pi}_n^0(s|\mathbf{x})} \left|\int \tilde{Z}_n^0(s)\; d\left(\tilde{\pi}_n^0(s) - \tilde{\pi}_n^0\left(s|\mathbf{x}\right)\right)\right|. \end{aligned} \tag{5.4}$$

Recall that the central argument for asymptotic behavior of either a Bayes or maximum likelihood estimators is the convergence of the likelihood ratio process $\tilde{Z}_n^0(s)$. Here, we just put down some informal account of the corresponding fact. Under some regularity conditions (see, e.g., [7], Chapter II, Section 1), we have the following approximation of the log-likelihood ratio process:

$$\begin{aligned}\log Z_n^0(u) &= \sum_{i=1}^n \log f(x_i|\theta_0+u) - \log f(x_i|\theta_0) \\ &\approx \sum_{i=1}^n \left[\frac{\partial \log f}{\partial \theta}(x_i|\theta_0)\right]^T u - \frac{n}{2} u^T I(\theta_0) u.\end{aligned}$$

Further, by central limit theorem, we have

$$\sum_{i=1}^n \frac{\partial \log f}{\partial \theta}(x_i|\theta_0) \approx \sqrt{nI(\theta_0)}Z,$$

where Z is a vector of IID standard normal variables of the dimension equal to the dimension of the parameter θ. One can combine these facts to approximate the likelihood ratio process through $Z_n^0(u) \approx e^{|Z|^2} e^{-|Z-\sqrt{nI(\theta_0)}u|^2/2}$, so that $\tilde{Z}_n^0(s) \approx e^{|Z|^2} e^{-|Z-s|^2/2}$.

From now on we assume that we deal with a location parameter θ which is one dimensional. Moreover, we assume that the prior π has a density p with respect to the Lebesgue measure. Then we have

$$\frac{d\tilde{\pi}_n^0}{ds}(s) = p\left(\theta_0 + s/\sqrt{I(\theta_0)n}\right)/\sqrt{I(\theta_0)n}.$$

We also assume that the density p is continuous and non-zero at θ_0. Then,

$$\begin{aligned}\sqrt{I(\theta_0)n}\int \tilde{Z}_n^0(s)\; d\tilde{\pi}_n^0(s) &= \int \tilde{Z}_n^0(s)\, p\left(\theta_0 + s/\sqrt{I(\theta_0)n}\right)\; ds \\ &\approx p(\theta_0)e^{|Z|^2}\int e^{-|Z-s|^2/2}ds = \sqrt{2\pi}p(\theta_0)e^{|Z|^2} \qquad (5.5)\end{aligned}$$

and, similarly,

$$\sqrt{I(\theta_0)n}\int s\tilde{Z}_n^0(s)\; d\tilde{\pi}_n^0(s) \approx p(\theta_0)e^{|Z|^2}\int se^{-|Z-s|^2/2}ds = 0. \qquad (5.6)$$

For a differentiable function g vanishing at $-\infty$ and a distribution function F, we have $\int g(u)\, dF(u) = \int g'(t)\bar{F}(t)\, dt$ so that

$$\begin{aligned}\left|\int \tilde{Z}_n^0(s)\; d\left(\tilde{\pi}_n^0(s) - \tilde{\pi}_n^0\left(s\middle|\mathbf{x}\right)\right)\right| &= \left|\int \frac{d\tilde{Z}_n^0}{ds}(s)\; \left(\tilde{\pi}_n^0(s) - \tilde{\pi}_n^0(s|\mathbf{x})\right) ds\right| \\ &\le \int \left|\frac{d\tilde{Z}_n^0}{ds}(s)\right|\; \left|\tilde{\pi}_n^0(s) - \tilde{\pi}_n^0(s|\mathbf{x})\right| ds \\ &\le \sup_\theta |\pi(\theta) - \pi_n(\theta|\mathbf{x})| \int \left|\frac{d\tilde{Z}_n^0}{ds}(s)\right|\; ds \\ &\approx \sup_\theta |\pi(\theta) - \pi_n(\theta|\mathbf{x})| \int |s|\, e^{-\frac{|Z-s|^2}{2}}\; ds \\ &= 2\sup_\theta |\pi(\theta) - \pi_n(\theta|\mathbf{x})|. \qquad (5.7)\end{aligned}$$

Similarly,

$$\left|\int s\tilde{Z}_n^0(s)\; d\left(\tilde{\pi}_n^0(s)-\tilde{\pi}_n^0\left(s|\mathbf{x}\right)\right)\right| \leq \tag{5.8}$$

$$\begin{aligned}
&2\sup_\theta|\pi(\theta)-\pi_n(\theta|\mathbf{x})|+\int\left|s\frac{d\tilde{Z}_n^0}{ds}(s)\right|\;\left|\tilde{\pi}_n^0(s)-\tilde{\pi}_n^0(s|\mathbf{x})\right|ds\\
&\leq \sup_\theta|\pi(\theta)-\pi_n(\theta|\mathbf{x})|\left(2+\int\left|s\frac{d\tilde{Z}_n^0}{ds}(s)\right|\;ds\right)\\
&\approx \sup_\theta|\pi(\theta)-\pi_n(\theta|\mathbf{x})|\left(2+\int s^2e^{-\frac{|Z-s|^2}{2}}\;ds\right)\\
&= \left(2+\sqrt{2\pi}\right)\sup_\theta|\pi(\theta)-\pi_n(\theta|\mathbf{x})|\,.
\end{aligned}$$

Upon substituting (5.5-5.8) into (5.4), we obtain

$$\left|\hat{\theta}_{eb}^{(n)}-\hat{\theta}_b^{(n)}\right| \lessapprox \left(\frac{\left(2+\sqrt{2\pi}\right)}{\sqrt{2\pi}p(\theta_0)e^{|Z|^2}}+\frac{o(n)}{2\pi p^2(\theta_0)e^{2|Z|^2}}\right)\sup_\theta|\pi(\theta)-\pi_n(\theta|\mathbf{x})|\,,$$

which, in view of (3.6), leads to (3.1).

ACKNOWLEDGMENT

The research of Podgórski and Kozubowski has been partially supported by the Riksbankens Jubileumsfond Grant Dnr: P13-1024:1.

REFERENCES

[1] Arnold, B.C. (1983). *Pareto Distributions*, International Cooperative Publ. House, Fairland, MD.

[2] Barnett, V.D. (1966). Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots, *Biometrika* **53**, 151-165.

[3] Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd ed., Wiley.

[4] Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*, Chapman & Hall, London.

[5] Haas, G., Lee, B. and Antle, C. (1970). Inferences for the Cauchy distribution based on maximum likelihood estimators, *Biometrika* **57**(2), 403-408.

[6] Hossain, M.M. (2014). *A Novel Estimation Method Based on Maximum Likelihood*, Thesis, University of Nevada, Reno.

[7] Ibragimov, I.A. and Khasminskii, R.Z. (1981). *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.

[8] Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed., Springer, New York.

[9] Kotz, S. Kozubowski, T.J. and and Podgórski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*, Birkhäuser Boston.

[10] Podgórski, K. and Wallin, J. (2015). Maximizing leave-one-out likelihood for the location parameter of unbounded densities, *Ann. Inst. Statist. Math.* **67**(1), 19-38.

[11] Tukey, J.W. (1958). Bias and confidence in not quite large samples, *Ann. Math. Statist.* **29**, 614.