A novel weighted likelihood estimation with empirical Bayes flavor

Krzysztof Podgórski
Department of Statistics
Lund University

September 6, 2019

Outline

Highlights

- **A set of estimators** of a parameter is combined into a **weighted average** to produce the final estimator.

Highlights

- **A set of estimators** of a parameter is combined into a **weighted average** to produce the final estimator.
- **The weights** are chosen to be **proportional to the likelihood** evaluated at the estimators.

Highlights

- **A set of estimators** of a parameter is combined into a **weighted average** to produce the final estimator.
- **The weights** are chosen to be **proportional to the likelihood** evaluated at the estimators.
- The method is presented for a set of estimators obtained by using **the maximum likelihood principle** applied to **each individual observation**.

Highlights

- **A set of estimators** of a parameter is combined into a **weighted average** to produce the final estimator.
- **The weights** are chosen to be **proportional to the likelihood** evaluated at the estimators.
- The method is presented for a set of estimators obtained by using **the maximum likelihood principle** applied to **each individual observation**.
- The approach can be interpreted as**Bayesian** with a **data driven prior**.

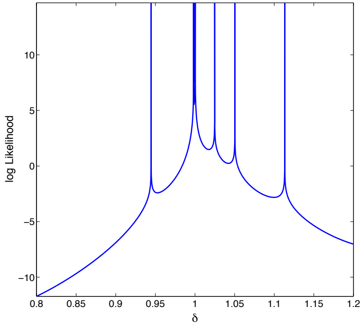
Highlights

- **A set of estimators** of a parameter is combined into a **weighted average** to produce the final estimator.
- **The weights** are chosen to be **proportional to the likelihood** evaluated at the estimators.
- The method is presented for a set of estimators obtained by using **the maximum likelihood principle** applied to **each individual observation**.
- The approach can be interpreted as**Bayesian** with a **data driven prior**.
- The estimators are **consistent, asymptotic normal, and efficient**.
- The **'posterior' distribution** automatically yields direct assessment of the performance and **accuracy of the estima**
- **Work is jointly with** – **Mobarak Hossain and Tomasz J. Kozubowski**

LUND UNIVERSITY

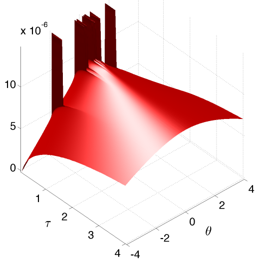
Motivating example-generalized Laplace model

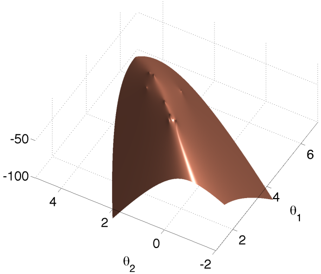
- A **multivariate generalized Laplace** model
$$\mathbf{X} = \sqrt{G}\mathbf{Z} + G\boldsymbol{\mu} + \boldsymbol{\theta},$$
$\mathbf{Z} \sim N(0, \boldsymbol{\Sigma})$, G is an independent gamma variable with the shape parameter τ.
- Multimodal likelihood can be seen as shown by examining likelihood for a sample of size ten for one dimension case with $\theta = 1$, $\sigma = 1$, $\mu = -1$, and $\tau = 0.2$



Mulitvariate/multiparameter case

- The problem becomes even more complicated for **multiparameter** or **multivariate** cases:





- *Left*: Location θ and shape τ, based a generalized Laplace distribution (with $\tau = 0.75$, $\theta = 0$, and $\mu = 0$);
- *Right*: The location $\boldsymbol{\theta}$ based on a sample of size ten of a bivariate Laplace distribution, where $\tau = 0.55$, $\boldsymbol{\mu} = (2, 3)$, $\boldsymbol{\theta} = (-1, 3)$, and $\boldsymbol{\Sigma}$ with the varian[ces] 1 and 3 and the covariance set to 1.5.

LUND UNIVERSITY

Basic idea

- Very often evaluating the MLE based on a **single sample** is not a problem (this is the case in our examples).
- Suppose that $\mathbf{x} = (x_1, \dots, x_n)$ is a sample from $f(x|\theta)$, where $\theta \in \Omega$ is an unknown (possible multivariate) parameter.
- The MLE of θ based on the ith data value is the quantity $\hat{\theta}_i = v(x_i)$.
- These individual estimators are subsequently combined as a **weighted average** to produce the final estimator

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

The weights are proportional to the likelihood.

Illustrative example

- To illustrate the proposed methodology, consider a scale parameter θ of an exponential distribution, for which the (full) likelihood function is

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

- The MLE is $\delta(\mathbf{x}) = 1/\bar{x}$.
- 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 weighted estimator is

$$\hat{\theta}(\mathbf{x}) = \frac{\sum_{k=1}^{n} x_k^{-n-1} e^{-n\bar{x}/x_k}}{\sum_{k=1}^{n} x_k^{-n} e^{-n\bar{x}/x_k}}. \quad (1)$$

- Performance based on 10,000 simulations:

n	θ	NEW $\hat{\theta}$(MSE)	MLE $\hat{\theta}$(MSE)
2	2	3.63 (45.52)	3.89 (47.41)
50	2	2.0486 (0.096)	2.0494 (0.091)
100	2	2.0195 (0.044)	2.0197 (0.042)

Outline

Bayesian setup

- Consider the Bayesian setup with some parametrized prior

$$X_i|\theta \sim f(\cdot|\theta), \;\; \Theta|\eta \sim \pi(\cdot|\eta),$$

- Typically, η is known.
- In the **empirical Bayes approach** the unknown η is estimated from the data, for example by maximizing the marginal

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

- This is a standard approach in the case of location parameter.
- The resulting estimator $\hat{\eta}$ is subsequently plugged-in into the traditional Bayesian estimator of θ.

Bayesian interpretation

- For a sample $X_1, ..., X_n$ from $f(x|\theta)$, let the prior distribution π of Θ be a discrete one, concentrated on values a_i with equal probabilities.
- The joint PDF of $\mathbf{X} = (X_1, ..., X_n)$ and Θ is given by

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

- The conditional PDF of Θ given $\mathbf{X} = \mathbf{x}$, that is the **posterior PDF of Θ**, is

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

- If $a_i = \hat{\theta}_i$, the posterior distribution corresponds to a random variable taking values $\hat{\theta}_i$'s with probabilities given by the weights w_i's
- **The mean of this posterior distribution coincides with the proposed estimator.**

Non-parametric empirical Bayes

- The previous estimation could be referred to **parametric empirical Bayesian** since the data are used to choose (estimate) the parameter of the prior.
- In contrast, in the newly proposed method, the empirical distribution of the sample approximates the entire prior distribution $\Pi(\cdot|\eta)$, which does not really have any finite dimensional parameter η – **non-parametric nature** (η is the entire distribution)
- By using empirical distribution of the data for the prior, one gets clues on possible values of θ that might have generated the sample because having an estimate of $f(\theta|\theta_0)$ as the prior, where θ_0 is the true value for the data, should be quite desirable since it assign relatively more probability to a neighborhood of θ_0.

p. 493 in Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed., Springer, New York.

"it is intuitively plausible that a close approximation to the asymptotic result will tend to be achieved more quickly (i.e. for smaller n)"

Outline

Remarks

- We note that in our ‘empirical’ Bayes formulation there is no external input of any kind with regard to the prior distribution.
- 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.
- We use the Bayesian setup primarily to establish asymptotic properties of this construction in the frequentist meaning.
- We distinguish two types of prior distributions for Θ:
  - one that does not depend on data, denoted by Π,
  - one that is data dependent (which is our case), and denoted by $\Pi_n(\cdot|\mathbf{x})$, where $\mathbf{x}$ represents the data.
- If these priors are re-centered at the true value θ_0, we write them as Π^0 and $\Pi_n^0(\cdot|\mathbf{x})$, respectively.

Frequentist theory of Bayesian estimators

- The posterior distribution and its mean are expressed by means of the likelihood ratio process,
$$Z_n^0(u) = \frac{f_n(\mathbf{x}|\theta_0 + u)}{f_n(\mathbf{x}|\theta_0)},$$
where $f_n(\mathbf{x}|\theta_0 + u)$ is the PDF of $\mathbf{X}$ given that the parameter is $\theta_0 + u$.
- The posterior mean, under the classical non-empirical prior, expresses as
$$\hat{\theta}_b^{(n)} = \frac{\int (\theta_0 + u) Z_n^0(u)\, d\Pi^0(u)}{\int Z_n^0(u)\, d\Pi^0(u)} = \theta_0 + \frac{\int u Z_n^0(u)\, d\Pi^0(u)}{\int Z_n^0(u)\, d\Pi^0(u)}.$$
- 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.
- In particular, certain regularity conditions for the IID case guarantee the asymptotic normality and efficiency of the estimator,
$$\lim_{n\to\infty} \sqrt{n}(\hat{\theta}_b^{(n)} - \theta_0) \stackrel{d}{=} N(0, \Sigma_0^2),$$
where $\Sigma_0^2 = I(\theta_0)^{-1}$ and $I(\theta_0)$ is the Fisher's information matrix.

Frequentist theory of empirical Bayesian estimators

- These results for the classical Bayes estimator do not apply directly to the new estimator since the empirical prior distribution is data dependent.
- In the important case where $\Pi_n^0(u|\mathbf{x})$ converges to a certain distribution $\Pi^0(u)$, we argue that $\hat{\theta}_{eb}^{(n)}$ inherits asymptotic properties of $\hat{\theta}_b^{(n)}$ such as asymptotic efficiency
$$\lim_{n\to\infty} \sqrt{n}(\hat{\theta}_{eb}^{(n)} - \theta_0) \overset{d}{=} N(0, \Sigma_0^2).$$
- 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.
- The results obtained can be viewed as first steps towards a more comprehensive asymptotic theory of Bayesian estimators arising in this set up.

LUND UNIVERSITY

Frequentist theory of empirical Bayesian estimators

- These results for the classical Bayes estimator do not apply directly to the new estimator since the empirical prior distribution is data dependent.
- In the important case where $\Pi_n^0(u|\mathbf{x})$ converges to a certain distribution $\Pi^0(u)$, we argue that $\hat{\theta}_{eb}^{(n)}$ inherits asymptotic properties of $\hat{\theta}_b^{(n)}$ such as asymptotic efficiency
$$\lim_{n\to\infty} \sqrt{n}(\hat{\theta}_{eb}^{(n)} - \theta_0) \stackrel{d}{=} N(0, \Sigma_0^2).$$
- 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.
- The results obtained can be viewed as first steps towards a more comprehensive asymptotic theory of Bayesian estimators arising in this set up.

The main difficulty

- Our main result and its proof depend heavily on the approach that is presented in work of Ibragimov and Khashminsky.
- 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 $\hat{\theta}_i = v(X_i)$.
- The convenience of the approach presented lies in deriving the asymptotical behavior of the likelihood ration process $\tilde{Z}_n^0(s) = Z_n^0(s/\sqrt{I(\theta_0)n})$.
- The classical and empirical cases can be written as

$$\hat{\theta}_b^{(n)} = \frac{\int(\theta_0 + u)Z_n^0(u)\, d\Pi^0(u)}{\int Z_n^0(u)\, d\Pi^0(u)} = \theta_0 + \frac{\int uZ_n^0(u)\, d\Pi^0(u)}{\int Z_n^0(u)\, d\Pi^0(u)},$$

$$\hat{\theta}_{eb}^{(n)} = \frac{\int(\theta_0 + u)Z_n^0(u)\, d\Pi_n^0(u|\mathbf{x})}{\int Z_n^0(u)\, d\Pi_n^0(u|\mathbf{x})} = \theta_0 + \frac{\int uZ_n^0(u)\, d\Pi_n^0(u|\mathbf{x})}{\int Z_n^0(u)\, d\Pi_n^0(u|\mathbf{x})}.$$

- One has to control the rate at which the *empirical prior* converges to the distribution of $v(X)$, where X is a random variable with the PDF $f(x|\theta_0)$, with θ_0 being the true value.
- Our goal was not to develop a comprehensive asymptotic theory of empirical Bayes estimators, which would be quite a challenge.

Outline

Asymptotic posterior distribution

- The asymptotic results can be utilized and interpreted to provide statistical inference based on the posterior distribution.
- The central result for this interpretation is the Bernstein-von Mises theorem, stating that, under a suitable (and non-empirical) prior, **the posterior distribution is asymptotically equal to the asymptotic normal distribution of the maximum likelihood estimator**.
- We have not pursued this theoretical development although we believe that the results holds by similar argument as in the case the asymptotics of the posterior mean.
- We illustrate this hypothesis through examples.

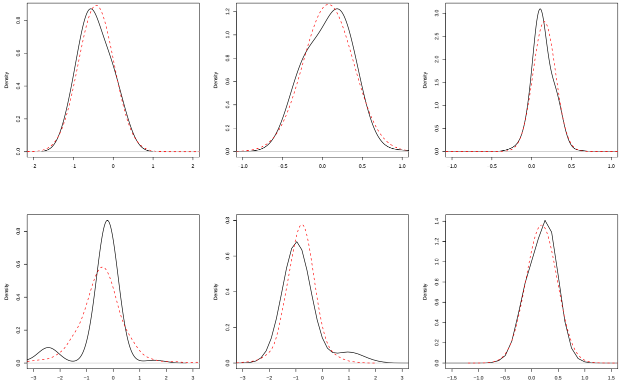
Two examples

- We start with the **Gaussian case**, the distribution of MLE follows from the classical theory.
- We use the new estimator $\hat{\theta}_{eb}$.
- The weights are used to present smoothed density estimator, representing the posterior distribution based on our empirical Bayes approach.
- By the asymptotic results, the two posterior distributions should coincide.
- We also consider the **Cauchy case**.
- Here we do not have explicit form of the MLE distribution so the graph is based on $k = 1000$ Monte Carlo simulated values.
- We compare the sampling distribution of the estimator with the posterior distribution based on a single run of the data.

LUND UNIVERSITY

Results

Comparison of the posterior distribution $P_n(\theta|\mathbf{x})$ based on the proposed approach (solid line) with: *(Top)* the MLE distribution in the Gaussian case (dashed line); *(Bottom)* the Monte Carlo simulated distribution of the estimator (dashed line). Sample sizes: 5 *(left)*, 10 *(middle)*, 50 *(right)*.



LUND UNIVERSITY

Further possible developments

- The asymptotics in the case when the empirical prior is based on the distribution of an estimator that is already consistent, for example, the leave-one-out or bootstrap distribution of an estimator.
- For example, consider the 'leave-one-out' prior, concentrated on the n estimators $\hat{\theta}_i$ calculated using the sample *without* the observation x_i.
- The multivariate location case can be treated exactly the same as the univariate one.
- For other than the location parameters one has to provide a convenient set of estimates that when given equal weights lead to a data-driven empirical prior. (The presented asymptotic result is valid in such a setup.)
- Sets of estimates can be based on maximizing likelihood based on the individual observations if the maximum is attained.
- In other cases, one can adopt other methods based on subsampling data. Investigating such empirical prior distributions for the parameters at hand is a separate problem.

Thank you!