Central European Journal of Economic Modelling and Econometrics

Bayesian Interpretation of Some *Empirical Bayes* Procedures in Hierarchical Models

Jacek Osiewalski*

Submitted: 25.11.2024, Accepted: 28.02.2025

Abstract

In modern statistics and its applications, in particular in econometrics, random parameters or latent variables are widely used, and their estimation or prediction is of interest. Under some prior assumptions, Bayes formula can be used to obtain their posterior distribution. However, on the sampling-theory grounds, the unknown constants appearing in the prior distribution are estimated using the data being actually modelled. We call such approaches quasi-Bayesian; *empirical Bayes* procedures give important examples. In this paper we propose theoretical framework that enables Bayesian validation (or interpretation) of quasi-Bayesian inference techniques. Our framework amounts to establishing a formal Bayesian model that justifies the quasi-Bayesian "posterior" as a valid posterior distribution. From the Bayesian model validating the quasi-posterior, i.e. from the joint distribution of observations and other quantities, one can deduce the true sampling model, that is the conditional distribution of observations, and the true prior (or marginal) distribution of the remaining quantities. We illustrate our approach not only by simple examples, but also by the complicated Bayesian model validating one of the basic *empirical Bayes* estimators of the multivariate normal mean. This model is in fact a non-standard joint measure that separates two subsets of the Cartesian product of the observation space and the parameter space.

Keywords: Bayesian inference, data-based "priors", Bayesian coherency

JEL Classification: C11, C51, C52

*Kraków University of Economics, Department of Econometrics and Operations Research, Kraków, Poland; e-mail: eeosiewa@cyf-kr.edu.pl; ORCID: 0000-0002-6710-6825

1 Introduction

Bayes formula for density functions of continuous random variables is a particularly important tool of Bayesian statistics, but it is not the only characteristic of this mode of statistical modelling and inference. There are two crucial features of the Bayesian approach to statistical methodology. Probabilistic representation of initial uncertainty about observations (available, missing, future), latent variables and classical parameters (unknown constants) is the main feature of Bayesian modelling. Treating all "unknowns" as random variables is closely related to the concept of subjective probability. Obeying rules of probability calculus is then the main characteristics of Bayesian inference. Obviously, Bayes formula is one of these rules (and a very useful one), but following it in isolation from other rules does not mean conducting Bayesian inference.

In modern statistics and its many subject areas (like econometrics) latent variables, random effects and other unobservable random quantities are widely used, and their estimation or prediction is usually of particular interest. Bayes formula can be used to obtain their posterior distribution, given appropriate distributional assumptions. Then, the posterior mean can be used as the estimator (or predictor), even within a non-Bayesian approach to statistics. However, the posterior distribution (and the posterior mean) of a latent variable depends on unknown constants (parameters) of the assumed marginal (or prior) distribution of this variable. The purely Bayesian solution amounts to treating all unknowns probabilistically and using probability rules on each level of the hierarchical model. However, on the sampling-theory grounds, the unknown constants are estimated using the data being actually modelled. Such approach is quite popular since 1970s under the name Empirical Bayes (EB). It uses Bayes formula for random effects, but at the same time specifies the prior hyper-parameters on the basis on the data.

Empirical Bayes methods can be described as incoherent by an orthodox Bayesian, who by coherency means following basic rules of probability. While such description is formally exact and true, it does not provide us with a deeper Bayesian understanding of incoherent inferences which are practically useful and frequently adopted in empirical research. So we propose theoretical framework that enables the purely Bayesian interpretation of incoherent, quasi-Bayesian inference techniques such as EB. Our framework amounts to establishing such formal Bayesian model that justifies a quasi-Bayesian "posterior" (resulting from some data-based "prior") as a valid posterior. From this Bayesian model, i.e. the joint distribution of observations and other quantities, which justifies the posterior in question, one can deduce (at least in principle) the true sampling model, that is the conditional distribution of observations, and the true prior (or marginal) distribution of the remaining quantities – latent variables and parameters. Since analytical derivations are possible only in very specific cases, in this paper we present only simple, illustrative examples. However, they clearly show that incoherence of quasi-Bayesian approaches leads to

posterior distributions, which formally correspond to sampling models and prior distributions different than the assumed (declared) ones. Osiewalski (2019) considered a normal statistical model with hierarchical structure, which is the starting point for explanation and justification of the EB approach. However, only normal distributions with known variances and covariances were examined in that paper – in order to use purely analytical tools and obtain closed-form solutions. A more general prior structure, that corresponds to the basic EB literature, will be considered here.
In the next section the general framework is presented and its application is illustrated using the simple problem of estimating the unknown mean of the normal distribution with known variance. Section 3 is devoted to the purely Bayesian interpretation of some basic EB estimation within the normal hierarchical model.

2 Bayesian models validating quasi-posteriors

The hierarchical structure of a parametric statistical model amounts to assuming the conditional distribution of observations described by some parametric density function

$$p\left(y \mid \theta\right) = g(y;\theta) \quad (y \in Y, \theta \in \Theta),$$

where the parameters grouped in θ are in fact latent random variables with some distribution dependent on deeper parameters (treated as unknown constants on non-Bayesian grounds); the density for random parameter vector θ is denoted as $f_0(\theta;\alpha)$, $\alpha \in A \subseteq \mathbb{R}^s$. Then the density function of the joint distribution of observations and random parameters or latent variables (with α fixed) can be written and decomposed in the following way:

$$p(y,\theta;\alpha) = p\left(y \mid \theta\right) f_0(\theta;\alpha) = g(y;\theta) f_0(\theta;\alpha) = f_1\left(\theta \mid y;\alpha\right) h(y;\alpha), \tag{1}$$

where $h(y;\alpha) = \int_\Theta g(y;\theta) f_0(\theta;\alpha) d\theta$ and $f_1\left(\theta \mid y;\alpha\right) = g(y;\theta) f_0(\theta;\alpha)/h(y;\alpha)$ are the densities of the marginal distribution of observations and the conditional distribution of latent variables, respectively. Thus, in order to make inferences on random parameters (given observations and α) Bayes formula is used:

$$f_1\left(\theta \mid y;\alpha\right) \propto g(y;\theta) f_0(\theta;\alpha).$$

Within the Bayesian approach, α would be modelled probabilistically by assuming its prior distribution. So the formal status of θ and α would be the same, and the marginal posterior obtained from the joint posterior distribution of (θ,α) would be the basis for any inference on θ. However, within the sampling-theory (non-Bayesian) statistical paradigm, no prior distribution is assumed for the deeper parameters grouped in α. Instead, they are estimated on the basis on the actual data, using some properties of the marginal distribution of observations or the maximum likelihood principle applied to the density $h(y;\alpha)$ treated as function of α (for any given y).

Then the estimate of α, e.g. $\widehat{\alpha} = \arg\max h(y;\alpha)$, $\alpha \in A$, can replace the unknown parameter vector α. By inserting $\widehat{\alpha} = a(y)$ into the posterior density of latent variables $f_1(\theta \mid y;\alpha)$ one obtains the quasi-posterior $\widehat{p}(\theta \mid y) = f_1(\theta \mid y, \widehat{\alpha})$, which enables practical inference, but is incoherent from the purely Bayesian view. Now we propose a formal Bayesian way to justify (or validate) $f_1(\theta \mid y, \widehat{\alpha})$ as a true posterior. This requires such joint distribution $\widetilde{p}(y,\theta)$ that leads to $f_1(\theta \mid y, a(y))$ as the conditional density $\widetilde{p}(\theta \mid y)$.

Definition 1. *For the hierarchical model* (1)*, any joint distribution* $\widetilde{p}(y,\theta)$ *such that* $\widetilde{p}(\theta \mid y) = f_1(\theta \mid y, a(y))$ *is called a Bayesian model validating the quasi-posterior* $f_1(\theta \mid y, \widehat{\alpha})$*, used to replace* $f_1(\theta \mid y;\alpha)$*.*

Theorem 1. *For the hierarchical model* (1) *and some positive constant* k*, the joint distribution defined by* $\widetilde{p}(y,\theta) = kg(y;\theta)f_0(\theta;\widehat{\alpha})$ *is a Bayesian model validating* $f_1(\theta \mid y, \widehat{\alpha})$*.*

Proof. For any $\alpha \in A$, the values of the four functions f_0, f_1, g, h that appear in (1) are linked by equality $g(y;\theta)f_0(\theta;\alpha) = f_1(\theta \mid y;\alpha)\,h(y;\alpha)$, no matter how α is specified. In particular, this equality holds for α replaced by its estimate $\widehat{\alpha} = a(y)$, so we have

$$\widetilde{p}(y,\theta) = kg(y;\theta)f_0(\theta;\widehat{\alpha}) = kf_1(\theta \mid y;\widehat{\alpha})\,h(y;\widehat{\alpha}), \tag{2}$$

$$\widetilde{p}(y) = \int_\Theta \widetilde{p}(y,\theta)d\theta = k\int_\Theta f_1(\theta \mid y;\widehat{\alpha})\,d\theta\; h(y;\widehat{\alpha}) = kh(y;\widehat{\alpha}); \tag{3}$$

the latter equality holds because $f_1(\theta \mid y;\alpha)$ represents a regular probability density function of θ for any value of the parameter α, in particular for its estimate $\widehat{\alpha}$. Using (2) and (3) we get

$$\widetilde{p}(\theta \mid y) = \widetilde{p}(y,\theta)/\widetilde{p}(y) = f_1(\theta \mid y;\widehat{\alpha}).$$

□

The construction described in Theorem 1 is very simple and intuitive. In order to have a Bayesian model validating the quasi-posterior $f_1(\theta \mid y;\widehat{\alpha})$ as the true posterior, it is enough to consider the product of the initial sampling density $g(y;\theta)$ and the data-based "prior" $f_0(\theta;\widehat{\alpha})$. Then, for the Bayesian model $\widetilde{p}(y,\theta) \propto g(y;\theta)f_0(\theta;\widehat{\alpha})$, the derivation of the posterior density $\widetilde{p}(\theta \mid y) = f_1(\theta \mid y;\widehat{\alpha})$ and the marginal data density $\widetilde{p}(y) \propto h(y;\widehat{\alpha})$ is straightforward. The only subtlety is that $\widetilde{p}(y) \propto h(y;a(y))$ need not be any probability density function (although $h(y;\alpha)$ is a proper probability density function for any $\alpha \in A$ fixed independently of y); this is illustrated in Example 1. The general forms of the prior density $\widetilde{p}(\theta)$ and the sampling density $\widetilde{p}(y \mid \theta)$, both implied by (2), can be written down quite easily:

$$\widetilde{p}(\theta) = k\textstyle\int_Y g(y;\theta)f_0(\theta;a(y))\,dy = k\int_Y f_1(\theta \mid y;a(y))\,h(y;a(y))\,dy,$$

$$\widetilde{p}(y \mid \theta) = \frac{\widetilde{p}(y,\theta)}{\widetilde{p}(\theta)} \propto g(y;\theta)f_0(\theta;a(y)),$$

but these densities may be non-standard and extremely difficult to characterise. Clearly, $\widetilde{p}(y \mid \theta)$ can be very different from $p(y \mid \theta) = g(y;\theta)$ assumed in the initial hierarchical model (1), and $\widetilde{p}(\theta)$ may be improper, as we show in the simple example below.

Example 1 (Estimating the mean of a normal distribution)**.** *Consider the simple hierarchical model (with normal mean following a normal prior):*

$$p(y,\mu;\alpha) = g(y;\mu)f_0(\mu;\alpha) = f_N^n\left(y \mid \mu\iota_n, cI_n\right) f_N^1(\mu \mid \alpha, v) \tag{4}$$

where $f_N^k(\cdot \mid b, A)$ *denotes the functional form of the density of the* k*-variate normal distribution with mean vector* b *and covariance matrix* A*, and* $\iota_n = (1\ 1 \ldots 1)'$*. We easily decompose the joint density into the marginal density of observations (marginal data density, MDD) and the posterior density of the parameter:*

$$p(y,\mu;\alpha) = h(y;\alpha)f_1\left(\mu \mid y;\alpha\right) = f_N^n\left(y \mid \alpha\iota_n, cI_n + v\iota_n\iota_n'\right) f_N^1(\mu \mid b_y, v_y)$$

where

$$v_y = \left(\frac{n}{c} + \frac{1}{v}\right)^{-1}, \quad b_y = \left(\frac{n}{c} + \frac{1}{v}\right)^{-1}\left(\frac{n}{c}\overline{y} + \frac{1}{v}\alpha\right), \quad \overline{y} = \frac{1}{n}\iota_n' y;$$

the sampling and prior variances (c *and* v*) are known, but* α *is unknown. If* α *is estimated by* $\widehat{\alpha} = a(y) = \overline{y}$*, then* $b_y = \overline{y}$ *and the quasi-posterior is* $f_1(\mu \mid y;\overline{y}) = f_N^1\left(\mu \,\middle|\, \overline{y}, \left(\frac{n}{c} + \frac{1}{v}\right)^{-1}\right)$*. In order to validate it, we define*

$$\widetilde{p}(y,\mu) \propto g(y;\mu)f_0\left(\mu;\overline{y}\right) = f_N^n\left(y - \mu\iota_n \mid 0, cI_n\right) f_N^1\left(\mu - \overline{y} \mid 0, v\right),$$

which decomposes in two ways, as usual:

$$\widetilde{p}(y,\mu) = \widetilde{p}\left(\mu \mid y\right)\widetilde{p}(y) = \widetilde{p}\left(y \mid \mu\right)\widetilde{p}(\mu)$$

with (i) $\widetilde{p}\left(\mu \mid y\right) = f_1(\mu \mid y;\overline{y})$ *and* $\widetilde{p}(y) \propto \exp\left(-\frac{1}{2c}y'My\right)$*, where* $M = I_n - \frac{1}{n}\iota_n\iota_n'$*;* (ii) $\widetilde{p}\left(y \mid \mu\right) = f_N^n\left(y \,\middle|\, \mu\iota_n, c\left(I_n - \frac{c}{n(c+nv)}\iota_n\iota_n'\right)\right)$ *and* $\widetilde{p}(\mu)$ *constant. Thus, in order to formally validate* $f_1(\mu \mid y;\overline{y})$ *as the posterior density, we have to use the sampling model assuming dependence (in the form of equi-correlation) and a flat, improper prior. The implicit Bayesian model, validating the quasi-posterior* $f_1(\mu \mid y;\overline{y})$ *is quite different from the declared one, characterised by independent sampling and a proper normal prior. Also note that the marginal data density in this Bayesian model,* $\widetilde{p}(y) \propto \exp\left(-\frac{1}{2c}y'My\right)$*, is improper.*

The improper marginal data density obtained in Example 1 is very interesting. Its main characteristics are easily derived due to the following Lemma.

Lemma. *Consider a change of variables, from* $y \in Y \subseteq \mathbb{R}^n$ *to* $(z, \overline{y})$*, with* $z' = (z_1 \dots z_{n-1})$*,* $z_i = y_i - \overline{y}$ *(*$i = 1, \dots, n-1$*). The density of the form* $p(y) = f(y'My)$ *corresponds then to the product of* $p(z) \propto f(z'Az)$*, where* $A = I_{n-1} + \iota_{n-1}\iota'_{n-1}$*, and the uniform density of* $\overline{y}$*.*

Proof. The Jacobian of this transformation is a constant (a number). Direct calculations show that $y'My = z'Az$. Thus $p(y) = f(y'My) \Leftrightarrow p(z, \overline{y}) \propto f(z'Az)$. So we have:

$$p(z, \overline{y}) = p(z)p(\overline{y}), \quad p(z) \propto f(z'Az), \quad p(\overline{y}) \propto 1.$$

□

From this Lemma we immediately infer that if $\widetilde{p}(y) \propto \exp\left(-\frac{1}{2c}y'My\right)$ for $y \in \mathbb{R}^n$, then $\widetilde{p}(\overline{y}) \propto 1$ and $\widetilde{p}(z) = f_N^{n-1}\left(z \mid 0, cA^{-1}\right)$, where $A^{-1} = I_{n-1} - \frac{1}{n}\iota_{n-1}\iota'_{n-1}$. So we see that $\widetilde{p}(y)$ obtained in Example 1 is the density of such σ-finite joint measure that is improper uniform for $\overline{y}$, the average value of all n observations, but it is proper normal for the deviations from the average. This marginal data density gives us an intuitive description of similarities among y_i's without specifying any information about their average level.

3 Quasi-posteriors in the basic *empirical Bayes* approach

Now we consider the following multivariate specification, which is the starting point for the basic formulation of the *empirical Bayes* approach:

$$p(y \mid \theta) = g(y; \theta) = f_N^n(y \mid \theta, cI_n), \quad f_0(\theta; \alpha) = f_N^n(\theta \mid \alpha_1\iota_n, dI_n). \tag{5}$$

We can decompose the product $g(y;\theta)f_0(\theta;\alpha)$ into $f_1(\theta \mid y; \alpha)\, h(y; \alpha)$, where

$$h(y;\alpha) = \int_{\mathbb{R}^n} g(y;\theta)f_0(\theta;\alpha)d\theta = f_N^n(y \mid \alpha_1\iota_n, (c+d)I_n) \tag{6}$$

is the density function of the marginal distribution of the observation vector (given α) and

$$f_1(\theta \mid y; \alpha) = f_N^n\left(\theta \,\middle|\, \frac{d^{-1}}{c^{-1}+d^{-1}}\alpha_1\iota_n + \frac{c^{-1}}{c^{-1}+d^{-1}}y, \frac{1}{c^{-1}+d^{-1}}I_n\right) \tag{7}$$

is the posterior density of the vector of random parameters (given α), with the mean

$$E(\theta \mid y; \alpha) = w \cdot \alpha_1\iota_n + (1-w) \cdot y, \quad w = \frac{d^{-1}}{c^{-1}+d^{-1}} = \frac{c}{c+d} \in (0,1).$$

Note that the posterior precision (the inverse of posterior variance) is the sum of the sample precision c^{-1} and the prior precision d^{-1}, and the posterior mean is a weighted average of the vector of prior means and the observation vector – with weights equal to the share of prior or sample precision in the posterior (i.e., final) precision. Thus $E(\theta \mid y;\alpha)$ is a point (in $\Theta = \mathbb{R}^n$) that lies on the line segment between $(\alpha_1\ \alpha_1 \dots \alpha_1)'$ and $(y_1\ y_2 \dots y_n)'$.
In the general formulation (5)-(7) it is assumed that the observation variance c is known. The basic *empirical Bayes* technique is formulated for unknown prior mean parameter α_1 and unknown prior variance $\alpha_2 = d$. However, we distinguish and present two cases: with $\alpha = \alpha_1$ and d known, as well as with $\alpha = (\alpha_1\ \alpha_2)'$ and unknown $d = \alpha_2$. While the latter case is closer to the *empirical Bayes* literature, the former one is much simpler and easily analysed. As we show in this section, these two cases lead to very different Bayesian models validating quasi-posteriors. The case with known d is presented in detail by Osiewalski (2019), so here we only summarise it in the following Example.

Example 2 (EB estimation of multivariate normal mean when prior variance is known)**.** *It is worth stressing that the density $f_1(\theta \mid y;\alpha)$ in (7) follows Bayes formula for any fixed α, so to this point the presented approach obeys coherence. However, the deeper parameter (here only the prior mean $\alpha = \alpha_1$) is unknown, so on the sampling theory grounds (e.g., in the EB approach) some point estimate of unknown $\alpha \in A$ is inserted into $f_1(\theta \mid y;\alpha)$, which results in the quasi-posterior $\widehat{p}(\theta \mid y) = f_1(\theta \mid y,\ \widehat{\alpha})$. For $\widehat{\alpha} = \overline{y}$ we get*

$$h(y;\widehat{\alpha}) = f_N^n(My \mid 0,(c+d)I_n), \quad f_1(\theta \mid y,\widehat{\alpha}) = f_N^n\left(\theta \,\middle|\, w\overline{y}\iota_n + (1-w)y, \frac{c\ d}{c+d}I_n\right).$$

Within the sampling theory approach, $w\overline{y}\iota_n + (1-w)y$ is a natural point estimate of the vector of random parameters. The Bayesian model validating $f_1(\theta \mid y,\widehat{\alpha})$, constructed in line with Theorem 1, takes the form

$$\widetilde{p}(y,\theta) = kg(y;\theta)f_0(\theta;\widehat{\alpha}) = kf_N^n(y \mid \theta, cI_n)\, f_N^n(\theta \mid \overline{y}\iota_n, dI_n),$$

and we seek for the sampling density $\widetilde{p}(y \mid \theta)$ and the prior density $\widetilde{p}(\theta)$ that correspond to this Bayesian model. Elementary calculations show that

$$\widetilde{p}(y,\theta) \propto f_N^n\left(y \,\middle|\, \theta, \left(\frac{1}{c}I_n + \frac{1}{dn}\iota_n\iota_n'\right)^{-1}\right) \exp\left(-\frac{1}{2d}\theta' M\theta\right),$$

$$\widetilde{p}(\theta) = \int_Y \widetilde{p}(y,\theta)dy \propto \exp\left(-\frac{1}{2d}\theta' M\theta\right),$$

$$\widetilde{p}(y \mid \theta) = \frac{\widetilde{p}(y,\theta)}{\widetilde{p}(\theta)} = f_N^n\left(y \,\middle|\, \theta, \left(\frac{1}{c}I_n + \frac{1}{dn}\iota_n\iota_n'\right)^{-1}\right).$$

The prior $\widetilde{p}(\theta)$ is an improper, σ-finite measure. It is informative, as it favours (approximate) equality $\theta_1 \approx \ldots \approx \theta_n$. In fact, it is improper uniform for the average of $\theta_1, \ldots, \theta_n$, but proper and jointly normal for the deviations from the average. The sampling density $\widetilde{p}(y \mid \theta)$ is different from $p(y \mid \theta)$ assumed in (5). The true conditional distribution is normal, like the initially declared one, but it assumes that the observations are equally correlated – instead of being independent. The true sampling covariance matrix $\widetilde{V}(y \mid \theta) = c\left(I_n - \frac{c}{n(c+d)}\iota_n\iota_n'\right)$ leads to the same correlation coefficient for each pair of observations:

$$\widetilde{Corr}(y_i, y_j \mid \theta) = -\frac{c}{(n-1)c+nd} \quad (i \neq j),$$

which tends to zero when n increases; $\widetilde{p}(y \mid \theta)$ practically coincides with $p(y \mid \theta)$ when n is sufficiently large.

The simplicity of the results in Example 2 is destroyed when the prior variance is unknown, as it is assumed in the standard EB approach, where the unbiased estimators $\overline{y}$ and $(n-3)c/\sum_{j=1}^{n}(y_j - \overline{y})^2 = (n-3)c/y'My$ are considered for α_1 and $w = c/(c+\alpha_2)$, respectively (see Morris 1983, Casella 1985). Of course, the assumption $n > 3$ holds. In order to impose the condition $w \in [0,1]$, the estimator $\widehat{w} = \min\{1, (n-3)c/y'My\}$ is taken. Finally, $\alpha = (\alpha_1\alpha_2)'$ is estimated using $\widehat{\alpha}_1 = \overline{y}$ and $\widehat{\alpha}_2 = \max\{0, y'My/(n-3) - c\}$, and the *empirical Bayes* estimator of θ takes the form $\widehat{\theta}^{EB} = \widehat{w}\overline{y}\iota_n + (1-\widehat{w})\,y$; see Casella (1985). Note that $\widehat{\theta}^{EB} = \overline{y}\iota_n$ if $\widehat{w} = 1$, or equivalently $y'My/(n-3) \leq c$, i.e. when the observation vector y is located closely enough to the point $\overline{y}\iota_n$, so that the deviations from $\overline{y}$ are small in the following sense: $z'Az/(n-3) \leq c$ (as $y'My = z'Az$, see our Lemma).
Let us split the observation set $Y = \mathbb{R}^n$ into $Y_1 = \{y \in Y : y'My/(n-3) \leq c\} = \{y \in Y : z'Mz/(n-3) \leq c, \overline{y} \in \mathbb{R}\}$ and its compliment $\overline{Y}_1 = Y \backslash Y_1$. For $y \in \overline{Y}_1$: $\widehat{\alpha}_2 > 0$, $\widehat{w} < 1$ and the EB estimate "shrinks" y towards the average value; the quasi-posterior $f_1(\theta \mid y; \widehat{\alpha})$, obtained by replacing α in (7) with $\widehat{\alpha} = (\widehat{\alpha}_1\widehat{\alpha}_2)'$ specified above, is

$$f_1(\theta \mid y; \widehat{\alpha}) = f_N^n\left(\theta \,\middle|\, \widehat{\theta}^{EB}, c\,(1-\widehat{w})\,I_n\right), \tag{8}$$

which makes all values of $\theta \in \Theta = \mathbb{R}^n$ possible. The corresponding form of $h(y; \widehat{\alpha})$ is

$$h(y; \widehat{\alpha}) = f_N^n(y \mid \widehat{\alpha}_1\iota_n, (c+\widehat{\alpha}_2)\,I_n) = f_N^n\left(y - \overline{y}\iota_n \,\middle|\, 0, \frac{y'My}{n-3}\,I_n\right) \propto y'My^{-\frac{n}{2}}.$$

For $y \in Y_1$ we obtain $\widehat{\alpha}_2 = 0$, $\widehat{w} = 1$, and the quasi-poster is just a unit point mass at $\overline{y}\iota_n$:

$$f_1(\theta \mid y; \widehat{\alpha}) = I_{\{0\}}(\theta - \overline{y}\iota_n); \tag{9}$$

the EB estimate $\widehat{\theta}^{EB}$ is located at $\overline{y}\iota_n$, where the quasi-posterior $f_1(\theta \mid y; \widehat{\alpha})$ is concentrated; so we infer that $\theta_1 = \ldots = \theta_n = \overline{y}$, and other values of θ are

precluded. Note that the data-based "prior" $f_0(\theta;\widehat{\alpha})$ corresponding to $\widehat{\alpha}_2 = 0$ is also $I_{\{0\}}(\theta - \overline{y}\iota_n)$; in this case we also obtain

$$h(y;\widehat{\alpha}) = f_N^n(y \mid \widehat{\alpha}_1\iota_n, (c+\widehat{\alpha}_2)I_n) = f_N^n(y - \overline{y}\iota_n \mid 0, cI_n) \propto \exp\left(-\frac{1}{2c}y'My\right).$$

Theorem 2. *For the hierarchical model* (5)-(7) *with unknown prior mean* α_1 *and unknown prior variance* $\alpha_2 = d$*, the Bayesian model validating the quasi-posterior in* (8) *and* (9) *is the joint measure defined as follows:*
*(*i*) for the line in* $\mathbb{R}^n$ *defined as* $\Theta_0 = \{\theta \in \mathbb{R}^n : \theta = \theta_0\iota_n,\ \theta_0 \in \mathbb{R}\}$ *and its complement* $\overline{\Theta}_0$*, assume that* $Y_1 \times \overline{\Theta}_0$ *and* $\overline{Y}_1 \times \Theta_0$ *have zero measure within this Bayesian model;*
*(*ii*) for* $(y,\theta) \in \overline{Y}_1 \times \overline{\Theta}_0$ *use* $\widetilde{p}(y,\theta) \propto y'My^{-n/2}f_N^n\left(\theta \mid \widehat{\theta}^{EB}, c(1-\widehat{w})I_n\right)$*;*
*(*iii*) for* $(y,\theta) \in Y_1 \times \Theta_0$ *use 1) improper uniform prior* $\widetilde{p}(\theta_0) \propto 1$ *for* $\theta_0 \in \mathbb{R}$ *and 2) given* $\theta = \theta_0\iota_n \in \Theta_0$*, unit mass at* $\overline{y} = \theta_0$ *and* $\widetilde{p}(z) \propto f_N^{n-1}(z \mid 0, cA^{-1})\, I_{[0,c]}\left(\frac{z'Az}{n-3}\right)$ *for deviations* z *from* $\overline{y}\iota_n = \theta_0\iota_n$*.*

Proof. From the proof of Theorem 1 we know that the Bayesian model validating $f_1(\theta \mid y;\widehat{\alpha})$ can be defined as $\widetilde{p}(y,\theta) = kf_1(\theta \mid y;\widehat{\alpha})\,h(y;\widehat{\alpha})$. Using the forms of $f_1(\theta \mid y;\widehat{\alpha})$ and $h(y;\widehat{\alpha})$, for $y \in Y_1$ we have $\widetilde{p}(y,\theta) \propto \exp\left(-\frac{1}{2c}y'My\right) I_{\{0\}}(\theta - \overline{y}\iota_n) \propto \exp\left(-\frac{1}{2c}z'Az\right) I_{\{0\}}(\theta - \overline{y}\iota_n)$, where only $\theta = \overline{y}\iota_n$ is allowed due to (9); note that $\overline{y}\iota_n \in \Theta_0$, so $\theta \in \overline{\Theta}_0$ is impossible. Thus, $y \in Y_1$ can be considered jointly only with $\theta \in \Theta_0$ and $\widetilde{p}(y,\theta)$, which defines the joint measure on $Y_1 \times \Theta_0$, can be described as using improper uniform marginal distribution $\widetilde{p}(\overline{y})$ for $\overline{y} \in \mathbb{R}$ and then, given $\overline{y}$, the unit point mass at $\theta = \overline{y}\iota_n \in \Theta_0$ and the appropriate truncated normal distribution for $z \in \mathbb{R}^{n-1}$; this is equivalent to (*iii*). For $y \in \overline{Y}_1$, the product of $f_1(\theta \mid y;\widehat{\alpha})$ in (8) and the appropriate form of $h(y;\widehat{\alpha})$ gives $\widetilde{p}(y,\theta)$ in (*ii*); since Θ_0 can be excluded from the support of (8), (*ii*) holds. Finally, we use (*i*) to extend the definition of our Bayesian model on $Y_1 \times \overline{\Theta}_0$, the set not covered above, as well as on $\overline{Y}_1 \times \Theta_0$, the subset excluded from $\overline{Y}_1 \times \Theta$. So the Bayesian model validating the quasi-posterior (8)-(9) is defined for all $(y,\theta) \in Y \times \Theta$. □

Note that the Bayesian model validating the *empirical Bayes* approach that leads to $\widehat{\theta}^{EB}$ as the estimator of the random normal mean vector θ has a very particular structure. The joint measure presented in Theorem 2 separates two subsets $Y_1 \times \Theta_0$ and $\overline{Y}_1 \times \overline{\Theta}_0$ of the Cartesian product $Y \times \Theta$. This has important consequences for inference. If the individual observations $y_1, \ldots, y_n$ are so similar that $y \in Y_1$, where Y_1 is in fact the set of points on and around the line $Y_0 = \{y \in \mathbb{R}^n : y = \overline{y}\iota_n, \overline{y} \in \mathbb{R}\}$, then we know with certainty that θ lies on the line Θ_0, at the point that corresponds to the actual data average, i.e. we know that $\theta_1 = \ldots = \theta_n = \overline{y}$. If $y_1, \ldots, y_n$ are not so close and thus $y \in \overline{Y}_1$, then θ can be located wherever in $\overline{\Theta}_0$ or even Θ, as the posterior probability that $\theta \in \Theta_0$ (or, equivalently, that $\theta_1 = \ldots = \theta_n$) is zero. Note, however, that in the case of $y \in \overline{Y}_1$ the set $\{\theta \in \mathbb{R}^n : \theta'M\theta \leq \varepsilon\}$ has non-zero

posterior probability, no matter how small ε is – and how closely to the line Θ_0 the values of θ are concentrated. While observing $y \in Y_1$ confirms the equality of all θ_i's with certainty, getting $y \in \overline{Y}_1$ need not exclude their equality, at least from practical perspective.

The different σ–finite measures appearing in points (*ii*) and (*iii*) of Theorem 2 lead to the same marginal improper uniform density for $\overline{y} \in \mathbb{R}$. In (*iii*) also the prior (improper uniform over Θ_0) and the sampling distribution (degenerate for $\overline{y}$ and truncated normal for z) are quite simple. However, the prior density $\widetilde{p}(\theta)$ and the sampling density $\widetilde{p}\,(y \mid \theta)$ for the part (*ii*) are not easy to analyse. In order to find their exact form we have to define

$$C(\theta) = \int_{\overline{Y}_1} y' M y^{-n/2} f_N^n \left(\theta \mid \widehat{\theta}^{EB}, c\,(1-\widehat{w})\, I_n\right) dy; \tag{10}$$

then $\widetilde{p}(\theta) \propto C(\theta)$ and $\widetilde{p}\,(y \mid \theta) = [C(\theta)]^{-1}\, y' M y^{-n/2} f_N^n \left(\theta \mid \widehat{\theta}^{EB}, c\,(1-\widehat{w})\, I_n\right) I_{\overline{Y}_1}(y)$. Clearly, the formal Bayesian justification of the EB approach leading to the Stein type "shrinkage" estimator relies on very strange sampling and prior distributions, quite far from the assumed normal ones. Since the "true" sampling model $\widetilde{p}\,(y \mid \theta)$ and the "true" prior $\widetilde{p}(\theta)$ correspond to the estimator $\widehat{\theta}^{EB}$ with good sampling properties, they seem worth further and deeper investigations.

Acknowledgements

This article has been finally prepared within the research project no 088/EIE/2024/POT, financed from the subsidy granted to Kraków University of Economics.

References

[1] Casella G., (1985), An introduction to empirical Bayes data analysis, *The American Statistician* 39, 83–87.

[2] Efron B., 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.

[3] Greene W. H., (2008), *Econometric Analysis* (Sixth Edition), Pearson, Upper Saddle River NJ.

[4] Morris C., (1983), Parametric empirical Bayes inference: Theory and applications (with discussion), *Journal of the American Statistical Association* 78, 47–65.

[5] Osiewalski J., (2019), Bayesian interpretation of quasi-Bayesian inference in a normal hierarchical model, [in:] *The 13th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*, C.H.Beck, 160–168, available at: http://www.konferencjazakopianska.pl/pliki/proceedings_2019/pdf/r19.pdf.