On the Bayesian Interpretation of Penalized Statistical Estimators

Jan Kalina[1,2(✉)] and Barbora Peštová[1]

[1] The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 00 Prague 8, Czech Republic
{kalina,pestova}@cs.cas.cz
[2] The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodárenskou věží 4, 182 00 Prague 8, Czech Republic

Abstract. The aim of this work is to search for intuitive interpretations of penalized statistical estimators. Penalized estimates of the parameters of three models obtained by Bayesian reasoning are explained here to correspond to the intuition. First, the paper considers Bayesian estimates of the mean and covariance matrix for the multivariate normal distribution. Second, a connection of a robust regularized version of Mahalanobis distance with Bayesian estimation is discussed. Third, regularization networks, which represent a common nonparametric tool for regression modeling, are presented as Bayesian methods as well. On the whole, selected important multivariate and/or regression models are considered and novel interpretations are formulated.

Keywords: Bayesian estimation · regularization · penalization · robustness · regression

1 Introduction

Bayesian statistical analysis plays an important role in the analysis of data in economics, biomedicine, engineering, chemistry, etc. It allows to incorporate a prior knowledge about the distribution of the parameters and the estimation is most often performed by evaluating the mean of the posterior distribution, which is obtained from the Bayes formula. Historically, the Bayes theorem for evaluating the posterior distribution first appeared in a posthumously published paper by the English theologian Rev. Thomas Bayes (1702–1761). From the beginnings, the philosophy of the Bayesian statistical data analysis was to naturally combine the results of given data with available prior information [17]. Currently, we experience a boom of Bayesian methods for the analysis of high-dimensional data [23].

Bayesian estimation allows to incorporate subjective opinion (degree of belief), to eliminate the vagueness of loosely defined concepts, or to capture unreliability (messiness) of data [2]. The prior knowledge increases the information

The research was supported by the project 21-05325S ("Modern nonparametric methods in econometrics") of the Czech Science Foundation.

L. Rutkowski et al. (Eds.): ICAISC 2023, LNAI 14126, pp. 343–352, 2023.
https://doi.org/10.1007/978-3-031-42508-0_31

available about the parameters of interest and reduces thus entropy within the considered probabilistic model. Bayesian thinking is applicable for models with uncertain parameters, if the parameters of the model are obtained with a low-precision arithmetics or are subject to measurement errors [1]. Other applications include Bayesian optimization based on the idea of surrogate modeling, Bayesian versions of quantile regression [15,16] or Bayesian classification. The Bayesian information criterion represents a popular model choice criterion for econometric models. Also in machine learning (computational intelligence), Bayesian methods have an increasing popularity (e.g. in hierarchical models [17]) with possible applications in the analysis of economic (possibly high-dimensional) data. If focusing on economic data, numerous examples of Bayesian methods were recalled in the monograph on Bayesian econometrics [4]. Other recent research includes e.g. Bayesian analysis of economic time series [25] or Bayesian methods for predicing the prices and algorithmic trading [5].

Bayesian tools may exploit non-informative priors or empirical Bayes approaches, i.e. to avoid the necessity of specifying the priors. Still, Bayesian estimation methods even for the simplest statistical models have sometimes been understood as too complicated or inaccessible. The reluctance to using Bayesian methods stems in our opinion from a lack of understanding the Bayesian methods as intuitive and natural estimation tools. Based on the classical result of Stein [28] for the location model for a multivariate normal distribution, the arithmetic mean is dominated by a shrinkage estimator obtained in a simple form as a combination of the mean with zero. In an analogous manner, shrinkage estimators obtained by shrinking standard estimates towards zero (in fact towards any arbitrary vector) in a variety of models have been shown to possess appeling properties (cf. [10]). Such estimators may often be derived in a Bayesian setup. Stein's theoretical result launched the popularity of regularization (in fact much more in machine learning than in statistics) and opened the door to the study of penalized estimators (i.e. combinations of classical methods with some prior knowledge).

The paper aims to formulate some interesting connections and interpretations related to Bayesian point estimation. The properties shown here represent carefully elaborated ideas that are straightforward but hard to find in the literature. The paper also aims to advocate the Bayesian point of view for several particular estimation tasks, where the estimates have the form of penalized methods. Starting with the simplest model, Sect. 2 is devoted to the estimation of the mean of the multivariate normal distribution. Section 3 considers the covariance matrix of normal data and the corresponding Mahalanobis distance, which is formulated here in an original form as a robustified (i.e. modified to be resistant against outliers [20]) and at the same time regularized version. Section 4 discusses regularization networks for a regression task in a nonparametric setup. Section 5 concludes the paper.

2 Estimating the Normal Mean

In this section, we discuss an interesting connection related to the Bayesian estimator of the expectation of p-variate normal distribution with a known covariance matrix. Let us consider $\mathbf{X}_1, \ldots, \mathbf{X}_n$ as a random sample of p-variate data, where $\mathbf{X}_i$ comes from the normal distribution $\mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n$. Let us assume to have available estimates $\mathbf{t}_1, \ldots, \mathbf{t}_R$ of the parameter $\boldsymbol{\mu}$, which come from some available (previously performed) study. Arithmetic means will be denoted in the usual way as

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \quad \text{and} \quad \bar{\mathbf{t}} = \frac{1}{R}\sum_{r=1}^{R} \mathbf{t}_r. \tag{1}$$

The available prior knowledge about the vector parameter $\boldsymbol{\mu}$ will be modeled here by means of the normal distribution

$$\boldsymbol{\mu} \sim \mathsf{N}_p(\bar{\mathbf{t}}, \boldsymbol{\eta}). \tag{2}$$

In this Bayesian setup, the mean of the posterior distribution of the vector $\boldsymbol{\mu}$ was expressed already in [8] in the form

$$\hat{\boldsymbol{\mu}} = (n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\eta}^{-1})^{-1}(n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{X}} + \boldsymbol{\eta}^{-1}\bar{\mathbf{t}}). \tag{3}$$

Further, we assume a specific assumption $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\mathcal{I}}$ to hold with a known $\sigma > 0$. We also assume that the precision (uncertainty) of the measurement of all the coordinates of the vector $\boldsymbol{\mu}$ is the same. The prior knowledge about the vector parameter $\boldsymbol{\mu}$ is modeled here by the normal distribution

$$\boldsymbol{\mu} \sim \mathsf{N}_p(\bar{\mathbf{t}}, \mathsf{diag}(\gamma^2, \ldots, \gamma^2)), \tag{4}$$

where diag denotes a diagonal matrix. Then, the expectation (3) of the posterior distribution of the expectation $\boldsymbol{\mu}$ corresponds to

$$\hat{\boldsymbol{\mu}} = \frac{n\gamma^2\bar{\mathbf{X}} + \sigma^2\bar{\mathbf{t}}}{n\gamma^2 + \sigma^2}, \tag{5}$$

which represents a penalized estimator alternatively expressed as

$$\hat{\boldsymbol{\mu}} = (1-\delta)\bar{\mathbf{X}} + \delta\bar{\mathbf{t}}, \quad \text{where} \quad \delta = \frac{\gamma^{-2}}{n\sigma^{-2} + \gamma^{-2}} = \frac{\sigma^2}{n\gamma^2 + \sigma^2}. \tag{6}$$

Under more particular assumptions, let us now assume a fixed value of the ratio denoted here as

$$\rho = \frac{\sigma^2}{n\gamma^2}. \tag{7}$$

This assumption may be realistic in many applications in the analysis of measurements (metrology), because σ^2/n represents the variance of each coordinate of $\bar{\mathbf{X}}$. Using a fixed (7) is realistic under the assumption of a non-contaminated

(homoscedastic) model. Naturally, the variability of the current measurements may be the same as that of the previous measurements, but still the setup (7) allows an even more realistic situation that the previously performed measurements have a lower precision (larger variability). Now, we can plug in the expression

$$\gamma^{-2} = \frac{n\rho}{\sigma^2} \tag{8}$$

to the multivariate Bayesian estimator (5). Using straightforward calculations then leads to the expression

$$\delta = \frac{R}{R+1} \quad \text{and} \quad \hat{\boldsymbol{\mu}} = \frac{1}{R+1}\bar{\mathbf{X}} + \frac{R}{R+1}\bar{\mathbf{t}}. \tag{9}$$

This represents a shrinkage estimator [28] of the vector $\boldsymbol{\mu}$ in a specific form, namely as a shrunken version of $\bar{\mathbf{X}}$ towards the average $\bar{\mathbf{t}}$. The intensity of the shrinkage is determined by the value of the coefficient (7) [7].

3 Robust Regularized Mahalanobis Distance

Let us discuss the robust regularized Mahalanobis distance; as a novel result, the distance is presented in this section as a Bayesian estimator of the population Mahalanobis distance. The novel version of the Mahalanobis distance may be useful for classification, clustering, outlier detection [6], or texture analysis in images; see also the discussion in [19] on applications to economic data. Let us assume to have a p-variate random vector $\mathbf{X}_1, \ldots, \mathbf{X}_n$ coming from the normal distribution $\mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with a known $\boldsymbol{\mu}$ and unknown matrix $\boldsymbol{\Sigma} \in \mathsf{PD}(p)$, where $\mathsf{PD}(p)$ denotes the set of all positive definite matrices of size $p \times p$. Because the classical estimates of $\boldsymbol{\Sigma}$ are vulnerable (non-robust) to the presence of outliers in the data, various robust estimates have been proposed [18].

As a particularly appealing joint estimator of both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, let us consider the recently proposed minimum weighted covariance determinant (MWCD) estimator [26]. The estimator is based on implicit weights assigned to individual observations; the user chooses only their magnitudes and the weights as such are then assigned to individual observations only after a permutation that is optimized within the computation of the estimator. The MWCD estimator is known to be robust to the presence of outliers in the data. It is in fact highly robust in terms of the breakdown point. At the same time, the influence function and asymptotic covariance matrix were evaluated for the MWCD estimator of $\boldsymbol{\mu}$ [26]. It is also important that the MWCD estimator remains highly efficient for non-contaminated data (i.e. without outliers). The MWCD estimator has appealing properties like other statistical methods based on ranks of observations [27].

We use the notation $\tilde{\mathbf{X}}_{MWCD}$ and $\mathbf{S}_{MWCD}$ to denote the MWCD-estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. We also introduce the notation

$$\mathbf{U} = n\mathbf{S}_{MWCD}. \tag{10}$$

In the following, we assume $\boldsymbol{\mu}$ to be known in order to study a Bayesian alternative of $\mathbf{S}_{MWCD}$. Assuming the data without outliers, we consider the matrix $\mathbf{U}$ to follow the Wishart distribution $W_p(\boldsymbol{\Sigma}, n)$ so that it holds $E\,\mathbf{U}/n = \Sigma$. As in [8], let us model the uncertainty about $\boldsymbol{\Sigma}$ (in fact about $\boldsymbol{\Sigma}^{-1}$) by the Wishart distribution

$$\boldsymbol{\Sigma}^{-1} \sim W_p(\nu+p-1, \boldsymbol{\Omega}^{-1}) \quad \text{for a certain } \nu > 0 \text{ and a certain } \boldsymbol{\Omega} \in \mathsf{PD}(p). \tag{11}$$

Now, $\boldsymbol{\Sigma}$ has the inverse Wishart distribution $W_p^{-1}(\nu+p-1, \boldsymbol{\Omega})$ and it holds that

$$\mathsf{E}\boldsymbol{\Sigma}^{-1} = (\nu+p-1)\boldsymbol{\Omega}^{-1} \quad \text{so that} \quad \mathsf{E}\boldsymbol{\Sigma} = \boldsymbol{\Omega}/(\nu-2). \tag{12}$$

The posterior expectation $\boldsymbol{\Sigma}$ is then equal to

$$\hat{\boldsymbol{\Sigma}} = \frac{\mathbf{U}+\boldsymbol{\Omega}}{n+\nu-2}. \tag{13}$$

The Bayesian risk of empirical Bayesian estimates of $\boldsymbol{\Sigma}$, which are based on estimating hyperparameters from the data, were studied for the given setup in the paper [11].

Specifically, let us interpret the Bayesian estimator obtained in (13). We use a different context than in Sect. 2, where a shrinkage coefficient was assumed to be fixed. Let us now consider available data from some previous study. The obtained estimates of $\boldsymbol{\Sigma}$ available from the previous study will be denoted as $\mathbf{U}_1/n, \ldots, \mathbf{U}_R/n$. In other words, the matrices $\mathbf{U}_1, \ldots, \mathbf{U}_R$ represent prior counterparts of $\mathbf{U}$. We assume that each of the previous studies used n measurements, which were performed with the same variability (precision) as the current measurements. It is natural to select $\boldsymbol{\Omega}$ and ν so that

$$\frac{1}{R}\sum_{r=1}^{R}\frac{\mathbf{U}_r}{n} = \frac{\boldsymbol{\Omega}}{\nu-2}, \tag{14}$$

i.e. to take $\boldsymbol{\Omega} = \sum_{r=1}^{R}\mathbf{U}_r$ a $\nu = Rn+2$. The estimats (13) is then obtained as an intuitive combination of the result of the measurements with the set of previous measurements. Such combination has the form of a penalized estimator

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n(R+1)}\left(\mathbf{U}+\sum_{r=1}^{R}\mathbf{U}_r\right). \tag{15}$$

Further, a robust regularized Mahalanobis distance based on the MWCD estimate of $\boldsymbol{\Sigma}$ will be defined. We consider a new observation $\mathbf{Z} \in \mathbb{R}^p$ and the task is to estimate the population Mahalanobis distance of $\mathbf{Z}$ from given data $\mathbf{X}_1, \ldots, \mathbf{X}_n$. The population version is unknown in the realistic scenario if we assume the expectation $\boldsymbol{\mu}$ of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and their covariance matrix $\boldsymbol{\Sigma}$ to be unknown. We may express the square of the distance as

$$d^2(\mathbf{Z}; \mathbf{X}_1, \ldots, \mathbf{X}_n) = (\mathbf{Z}-\bar{\mathbf{X}}_{MWCD})^T\mathbf{S}_{MWCD}^{-1}(\mathbf{Z}-\bar{\mathbf{X}}_{MWCD}). \tag{16}$$

A joint estimation of the mean and the covariance matrix in the Bayesian setup is more complicated; it is possible to consider shrinkage versions also for $\bar{\mathbf{X}}_{MWCD}$. Replacing the means by shrinkage version in the Mahalanobis distance seems common in the context of regularized LDA in biostatistics [10].

Numerical linear algebra may be exploited for an efficient computation of $\mathbf{S}_{MWCD} + \lambda \boldsymbol{\mathcal{I}}_p$ for a given λ for (16); we formulate here Algorithm 1 for this purpose with an automatically performed regularization. The notation $\mathbb{1}_A$ is used to denote the indicator of an event A. The eigenvalues used in Algorithm 1 can be easily shown to be equal to the eigenvalues of $\mathbf{S}_{MWCD} + \lambda \boldsymbol{\mathcal{I}}_p$. In addition, the algorithm uses an asymptotically optimal value of λ, which was derived already in [21] for estimating $\boldsymbol{\Sigma}$ by the classical (non-robust) but regularized form of the empirical covariance matrix.

For the context of outlier detection, we suggest to perform the commonly accepted approach to assume the regularized Mahalanobis distances to approximately follow a χ^2-distribution [3]. If we assume $\boldsymbol{\mu}$ to be known, then the squared Mahalanobis distance (16) is replaced by

$$d^2(\mathbf{Z}; \mathbf{X}_1, \ldots, \mathbf{X}_n) = (\mathbf{Z} - \boldsymbol{\mu})^T \left(\mathbf{S}_{MWCD} + \lambda \boldsymbol{\mathcal{I}}_p\right)^{-1} (\mathbf{Z} - \boldsymbol{\mu}), \tag{17}$$

where the Tikhonov regularization is applied on $\mathbf{S}_{MWCD}$. The squared distance (17) can be interpreted a Bayesian version of the MWCD-based Mahalanobis distance.

4 Regularization Networks

In this section, we explain that the connection of regularization networks [9] to Bayesian estimation is very intuitive. Regularization networks represent a class of nonlinear regression (supervised learning) tools in machine learning [24]; it deserves to be stressed that they are different from regularized neural networks [13], where the latter can be described simply as regularized versions of any models of neural networks.

Let us have the total number n observations with values of a continuous response $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and corresponding n vectors of regressors (predictors) $\mathbf{X}_1, \ldots, \mathbf{X}_n$ that are p-variate. The aim is to estimate the regression function

$$f(\mathbf{x}) = \mathsf{E}(Y | \mathbf{X} = \mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^p, \tag{24}$$

based on given data. We only assume that f exists but its shape is unknown. One possibility is to estimate f as a solution of the optimization task

$$\min_{f \in H_K} \left\{ \sum_{i=1}^{n} \left(Y_i - f(\mathbf{X}_i)\right)^2 + \lambda ||f||_{H_K} \right\}, \tag{25}$$

with a regularization parameter $\lambda > 0$, where f is searched for in a reproducing kernel Hilbert space (RKHS), i.e. a space with with reproducible kernel,

Algorithm 1. Robust regularized Mahalanobis distance

Input: Data $\mathbf{X}_1, \ldots, \mathbf{X}_n$, where $\mathbf{X}_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$

Input: A new observation $\mathbf{Z} \in \mathbb{R}^p$

Output: Robust regularized Mahalanobis distance of $\mathbf{Z}$ from $\mathbf{X}_1, \ldots, \mathbf{X}_n$ based on the MWCD estimation

1: $\bar{\mathbf{X}}_{MWCD}$:= MWCD-estimate of the expectation of $\mathbf{X}_1, \ldots, \mathbf{X}_n$

2: $\mathbf{S}_{MWCD} = \left(S_{ij}^{MWCD}\right)_{i,j=1}^{p}$:= MWCD-estimate of the covariance matrix of $\mathbf{X}_1, \ldots, \mathbf{X}_n$

3: Compute the eigendecomposition of $\mathbf{S}_{MWCD}$ as

$$\mathbf{S}_{MWCD} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T, \tag{18}$$

where

$$\mathbf{D} = diag\{\theta_1, \ldots, \theta_p\} \quad \text{and} \quad \mathbf{Q}^T = \mathbf{Q}^{-1}. \tag{19}$$

4: For

$$\begin{aligned} \hat{\lambda} &= \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} \widehat{\text{var}}(S_{ij}^{MWCD})}{\sum_{i=1}^{p}\sum_{j=1}^{p}(S_{ij}^{MWCD} - \mathbb{1}_{[i=j]})^2} \\ &= \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} \widehat{\text{var}}(S_{ij}^{MWCD})}{2\sum_{i=2}^{p}\sum_{j=1}^{i-1}\left(S_{ij}^{MWCD}\right)^2 + \sum_{i=1}^{p}(S_{ii}^{MWCD} - 1)^2}, \end{aligned} \tag{20}$$

compute

$$\mathbf{D}^* := \mathsf{diag}\{(1-\hat{\lambda})\theta_1 + \hat{\lambda}, \ldots, (1-\hat{\lambda})\theta_p + \hat{\lambda}\}. \tag{21}$$

5:

$$(\mathbf{D}^*)^{-1} := \mathsf{diag}\left\{\left((1-\hat{\lambda})\theta_1 + \hat{\lambda}\right)^{-1}, \ldots, \left((1-\hat{\lambda})\theta_p + \hat{\lambda}\right)^{-1}\right\} \tag{22}$$

6:

$$d(\mathbf{Z}; \mathbf{X}_1, \ldots, X_n) := \sqrt{(\mathbf{Z} - \bar{\mathbf{X}}_{MWCD})^T \mathbf{Q} \left(\mathbf{D}^*\right)^{-1} \mathbf{Q}^T (\mathbf{Z} - \bar{\mathbf{X}}_{MWCD})} \tag{23}$$

which corresponds to the Hilbert space of real functions on $\mathbb{R}^p$ [12]. Typically, K is now chosen as the Gaussian kernel

$$K(\mathbf{X}, \mathbf{y}) = \exp\left\{-\frac{||\mathbf{X} - \mathbf{y}||^2}{2\sigma^2}\right\}, \quad \mathbf{X}, \mathbf{y} \in \mathbb{R}^p, \tag{26}$$

with a fixed $\sigma > 0$, which can be estimated from the data. The task (25) may be expressed as

$$\min_{\alpha}\left\{||\mathbf{Y} - \mathbf{K}\boldsymbol{\alpha}||^2 + \lambda\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}\right\}, \tag{27}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ is a vector of parameters and $\mathbf{K}$ is a symmetric matrix with $K_{ij} = K(\mathbf{X}_i, \mathbf{X}_j)$ for $i, j = 1, \ldots, n$. Derivatives may be used to find out that minimum is achieved for the vector

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \left(\mathbf{K}^T\mathbf{K} + \lambda\mathbf{K}\right)^{-1}\mathbf{K}^T\mathbf{Y} \\ &= \left[(\mathbf{K} + \lambda\boldsymbol{\mathcal{I}})\mathbf{K}\right]^{-1}\mathbf{K}^T\mathbf{Y} \\ &= (\mathbf{K} + \lambda\boldsymbol{\mathcal{I}})^{-1}\mathbf{Y}, \end{aligned} \tag{28}$$

which corresponds to the ridge regression estimator for the linear regression model $\mathbf{Y} = \mathbf{K}\boldsymbol{\alpha} + \mathbf{e}$; therefore, the estimator (28) is commonly denoted as the generalized ridge estimator. At the same time, we can say that $\mathbf{K}$ is considered in a penalized version in (28). The fitted value for a new observation $\mathbf{Z} \in \mathbb{R}^p$ is then obtained according to

$$\hat{f}(\mathbf{Z}) = \sum_{i=1}^{n} \hat{\alpha}_i K(\mathbf{Z}, \mathbf{X}_i). \tag{29}$$

Bayesian Interpretation. The estimator (28) of the regression parameters may be obtained according to a Bayesian approach in the following way. We assume the random errors to follow $e \sim \mathsf{N}(0, \sigma^2\boldsymbol{\mathcal{I}})$. The prior distribution will be chosen as

$$\boldsymbol{\alpha} \sim \mathsf{N}\left(\mathbf{0}, c\mathbf{K}^{-1}\right) \quad \text{with some } c > 0. \tag{30}$$

The expectation equal to 0 is chosen here to simplify the results; if this is not realistic, the data can be transformed correspondingly. In a special situation with $c = \sigma^2/\lambda$, the mean of the posterior distribution is equal to (28).

The choice of the covariance matrix of $\boldsymbol{\alpha}$ to be directly proportional to $\mathbf{K}^{-1}$ corresponds to intuition. The covariance matrix of the parameter namely corresponds to a measure of accuracy of the prior measurements. If two particular measurements (say X_1 and X_2) are very close to each other, their element of $\mathbf{K}$ denoted as K_{12} is extraordinarily large, the corresponding element of the covariance matrix is small, i.e. we can say that the information obtained thanks to the close relationship between X_1 and X_2 is accurate. On the other hand, for two measurement very distant from each other, the corresponding element of the covariance matrix of the prior distribution is large, i.e. an individual distant (possibly outlying) measurement may be very inaccurate. To summarize, the regularization network considers the information in close (closely related) measurements to be more accurate compared to the information in outlying measurements.

5 Conclusion

The importance of Bayesian statistical estimation is without any doubt increasing hand in hand together with the increasing need to combine data from various sources. Also the availability of prior knowledge (experience) about parameters contributes to the increasing popularity of tools of Bayesian estimation. This work reveals and interprets that the Bayesian synthesis of prior and newly available information is intuitive, interpretable, mathematically elegant, and in some

important models also straightforward to compute. The interpretations formulated in the paper support the idea that existing methods have a clear interpretation and represent meaningful tools also under the Bayesian way of reasoning.

Bayesian estimation has a potential to be used in many other tasks than in those discussed in this paper. Topics omitted here include other regression methods including regularized multilayer perceptrons, which may be derived in a Bayesian setup in a natural way, Bayesian optimization, or Bayesian approaches to computational intelligence in a broader sense. Further, the Bayesian Model Averaging (BMA) allows to perform an effective model choice for basically any statistical model [14]. In recent years, intensive discussions have been focused on the choice of the prior distribution; the readers may effectively exploit weakly informative priors or non-informative (Jeffreys) priors in a variety of practical tasks.

The authors of this work believe that the future belongs to applying Bayesian principles to robust estimation (i.e. resistant to outliers) in both statistics and machine learning. The connections between Bayesian thinking and robustness, which are illustrated in Sect. 3, seem to have been discussed only rarely, such as in the paper [22] in the econometric context. We also plan to investigate Bayesian approaches to (possibly robust) regularized linear discriminant analysis for high-dimensional data [10], exploiting the ideas of Sect. 3 on the robust regularized Mahalanobis distance. Another idea is to rethink regularization approaches to multilayer perceptrons or radial basis function networks from the point of view of Bayesian statistics.

Acknowledgements. The authors would like to thank Jiří Grim and Lubomír Soukup (both ÚTIA AV ČR) for discussion about Bayesian estimation.

References

1. Beaumont, M.A.: Approximate Bayesian computation. Annu. Rev. Stat. Appl. **6**, 379–403 (2019)
2. Bryant, J., Zhang, J.L.: Bayesian Demographic Estimation and Forecasting. CRC Press, Boca Raton (2019)
3. Cerioli, A., Riani, M., Atkinson, A.C., Corbellini, A.: The power of monitoring: how to make the most of a contaminated multivariate sample. Stat. Methods Appl. **27**, 559–587 (2018)
4. Chan, J., Koop, G., Poirier, D.J., Tobias, J.L.: Bayesian Econometric Methods, 2nd edn. Cambridge University Press, Cambridge (2020)
5. Cohen, G.: Algorithmic strategies for precious metals price forecasting. Mathematics **2022**, 1134 (2022)
6. Dashdondov, K., Kim, M.H.: Mahalanobis distance based multivariate outlier detection to improve performance of hypertension prediction. Neural Process. Lett. **55**, 265–277 (2023)
7. Efron, B., Morris, C.: Stein's estimation rule and its competitors–an empirical Bayes approach. J. Am. Stat. Assoc. **68**, 117–130 (1973)
8. Evans, I.G.: Bayesian estimation of parameters of a multivariate normal distribution. J. Roy. Stat. Soc. B **27**, 279–283 (1965)

9. Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and support vector machines. Adv. Comput. Math. **13**, 1–50 (2000)
10. Guo, Y., Hastie, T., Tibshirani, R.: Regularized discriminant analysis and its application in microarrays. Biostatistics **8**, 86–100 (2007)
11. Haff, L.R.: Empirical Bayes estimation of the multivariate normal covariance matrix. Ann. Stat. **8**, 586–597 (1980)
12. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity. The Lasso and Generalizations. CRC Press, Boca Raton (2015)
13. Haykin, S.O.: Neural Networks and Learning Machines: A Comprehensive Foundation, 2nd edn. Prentice Hall, Upper Saddle River (2009)
14. Hinne, M., Gronau, Q.F., van den Bergh, D., Wagenmakers, E.J.: A conceptual introduction to Bayesian model averaging. Adv. Methods Pract. Psychol. Sci. **3**, 200–215 (2020)
15. Hlubinka, D., Šiman, M.: On generalized elliptical quantiles in the nonlinear quantile regression setup. TEST **24**, 249–264 (2015)
16. Hlubinka, D., Šiman, M.: On parametric elliptical regression quantiles. REVSTAT Stat. J. **18**, 257–280 (2020)
17. Johnson, A.A., Ott, M.Q., Dogucu, M.: Bayes Rules! An Introduction to Applied Bayesian Modeling. Chapman & Hall/CRC, Boca Raton (2022)
18. Jurečková, J., Picek, J., Schindler, M.: Robust Statistical Methods with R, 2nd edn. CRC Press, Boca Raton (2019)
19. Kalina, J.: On robust information extraction from high-dimensional data. Serb. J. Manag. **9**, 131–144 (2014)
20. Kalina, J., Tichavský, J.: On robust estimation of error variance in (highly) robust regression. Meas. Sci. Rev. **20**, 6–14 (2020)
21. Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. **88**, 365–411 (2004)
22. Pacifico, A.: Robust open Bayesian analysis: overfitting, model uncertainty, and endogeneity issues in multiple regression models. Economet. Rev. **40**, 148–176 (2021)
23. Pfarrhofer, M., Piribauer, P.: Flexible shrinkage in high-dimensional Bayesian spatial autoregressive models. Spat. Stat. **29**, 109–128 (2019)
24. Pillonetto, G.: System identification using kernel-based regularization: new insights on stability and consistency issues. Automatica **93**, 321–332 (2018)
25. Qiu, J., Jammalamadaka, S.R., Ning, N.: Multivariate time series analysis from a Bayesian machine learning perspective. Ann. Math. Artif. Intell. **88**, 1061–1082 (2020)
26. Roelant, E., Van Aelst, S., Willems, G.: The minimum weighted covariance determinant estimator. Metrika **70**, 177–204 (2009)
27. Saleh, A.K.M.E., Picek, J., Kalina, J.: R-estimation of the parameters of a multiple regression model with measurement errors. Metrika **75**, 311–328 (2012)
28. Stein, C.: Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, vol. 1, pp. 197–206 (1956)