Acta Cybernetica — *online–first paper version* — pages 1–19.

A Bayesian Framework for Regularized Estimation in Multivariate Models Integrating Approximate Computing Concepts*

Jan Kalina[ab]

Abstract

This paper discusses regularized estimators in the multivariate statistical model as tools naturally arising within a Bayesian framework. First, a link is established between Bayesian estimation and inference under parameter rounding (quantization), thereby connecting two distinct paradigms: Bayesian inference and approximate computing. Next, Bayesian estimation of the means from two independent multivariate normal samples is employed to justify shrinkage estimators, i.e., means shrunk toward the pooled mean. Finally, regularized linear discriminant analysis (LDA) is considered. Various shrinkage strategies for the mean are justified from a Bayesian perspective, and novel algorithms for their computation are proposed. The proposed methods are illustrated by numerical experiments on real and simulated data.

Keywords: multivariate data, Bayesian estimation, regularization, classification analysis, algorithms, shrinkage estimators

1 Introduction

In a variety of models of multivariate statistics and machine learning, regularized estimators may be obtained in the Bayesian statistical setup [4]. This paper is interested in finding interpretations and relationships between various types of estimators in the multivariate model including an extension for a classification model with data observed in several different groups.

Regularized (shrinkage) estimators are not only crucial for high-dimensional data analysis, where classical methods often become unstable or infeasible, but also beneficial even when working with small datasets, as they tend to improve predictive accuracy [36]. In a wide range of probabilistic models, regularization

*The research was supported by the grant 25-15490S of the Czech Science Foundation. The author would like to thank Kristýna Dvořáková, who was supported by the program Strategy AV21 "AI: Artificial Intelligence for Science and Society".

[a]The Czech Academy of Sciences, Institute of Computer Science, Prague, Czech Republic
[b]E-mail: kalina@cs.cas.cz, ORCID: 0000-0002-8491-0364

DOI: 10.14232/actacyb.316641

emerges naturally from Bayesian reasoning: prior knowledge is incorporated directly into the estimation process, resulting in estimates that are pulled (shrunk) towards a predefined reference value, typically the prior expectation [15].

A shrinkage estimator is a statistical estimator that improves estimation accuracy by combining observed data with additional information or assumptions, effectively "shrinking" raw estimates toward a predetermined target or prior value. This shrinkage reduces estimation variance, often at the cost of introducing some bias, but typically results in a lower overall mean squared error compared to unbiased estimators.

This form of regularization helps reduce variance and improve generalization performance, which is especially valuable when the data are noisy. In multivariate settings, such Bayesian-inspired regularization enhances the stability and robustness of the estimators [19]. We regard this as a natural and fruitful synthesis of Bayesian ideas and modern regularization techniques, aligning well with the goals of reliable and interpretable data analysis [13].

As a novelty, a connection between Bayesian estimators and quantized estimation, i.e. estimation contaminated by additional uncertainty or error, is established here. Moreover, we focus on shrinkage estimators for the means of independent multivariate normal samples and subsequently on regularized linear discriminant analysis, which frequently employs shrinkage techniques and has found many applications in high-dimensional data [8]. In particular, we justify the use of shrunken means within a specific Bayesian model and formulate two algorithms for regularized linear discriminant analysis exploiting linear algebra manipulations.

This paper has the following structure. Section 2 is devoted to Bayesian estimation of parameters of the multivariate normal distribution. In Section 3, an original connection between Bayesian estimation and estimation in a model with rounding (quantization) of the estimated parameters is established. In Section 4, Bayesian estimation for two independent normally distributed samples is derived, which theoretically justifies using regularized estimators in the given context and also in the context of regularized linear discriminant analysis of Section 5. Section 6 presents experiments over a real and a simulated dataset. Section 7 concludes the paper.

2 Bayesian estimation in the multivariate model

Let us assume the multivariate model

$$X_i \sim \mathsf{N}_p(\mu, \Sigma), \quad i = 1, \ldots, n, \tag{1}$$

with i.i.d. p-variate random vectors following the p-variate normal distribution with $n > p$. Let $X \in \mathbb{R}^{n \times p}$ denote the data matrix with rows $X_1, \ldots, X_n$. The covariance matrix Σ has to fulfil $\Sigma \in \mathsf{PD(p)}$, where $\mathsf{PD}(p)$ denotes the set of all positive definite symmetric matrices of size $p \times p$.

This section discusses Bayesian estimation of the mean vector $\mu \in \mathbb{R}^p$ and the covariance matrix Σ. The main purpose here is to review the Bayesian estimation of

these parameters, serving as a foundation for the subsequent sections. For a detailed description of Bayesian estimation principles, we refer the reader to the book [19].

The sample mean has the form

$$\bar{X} = (\bar{X}_1, \ldots, \bar{X}_p)^T = \frac{1}{n} X^T \hat{\jmath}_n = \frac{1}{n} \left(\hat{\jmath}_n^T X\right)^T \in \mathbb{R}^p, \quad \text{where} \quad \bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij} \tag{2}$$

and $\hat{\jmath}_n = (1, \ldots, 1)^T \in \mathbb{R}^n$. Also, the maximum likelihood estimator in (1) is the sample mean. In addition, minimizing the sum of squared residuals

$$\underset{\tilde{\mu} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (X_i - \tilde{\mu})^2 \tag{3}$$

leads again to the sample mean as the resulting estimator.

Under the assumption of a known Σ, let us consider the normal prior

$$\mu \sim \mathsf{N}_p(\theta, \eta) \tag{4}$$

with known $\theta \in \mathbb{R}^p$ and $\eta \in \mathsf{PD}(p)$. This is a conjugate prior for the multivariate normal likelihood with known covariance, meaning that the posterior distribution belongs to the same parametric family (i.e. multivariate normal). The use of a conjugate prior is advantageous because it leads to a closed-form expression for the posterior, which simplifies both analytical derivations and computational implementation of Bayesian inference. The Bayesian estimator $\hat{\mu}$ of the parameter μ obtained as the mean of the posterior distribution is a shrinkage estimator of μ shrunken towards θ. It depends on the nuisance covariance matrix Σ. The formula for the estimator, which was provided e.g. in the seminal paper [7], is given by

$$\hat{\mu} = (n\Sigma^{-1} + \eta^{-1})^{-1}(n\Sigma^{-1}\bar{X} + \eta^{-1}\theta). \tag{5}$$

Remark 1. If assuming $\eta^{-1} = c\Sigma^{-1}$ for some $c > 0$, the estimator (5) does not depend on Σ, because (5) simplifies to

$$\hat{\mu} = (1 - \delta)\bar{X} + \delta\theta, \quad \text{where} \quad \delta = \frac{c}{n + c}. \tag{6}$$

The James–Stein estimator is a biased estimator of μ with a smaller quadratic risk than the mean [18] for $p \geq 3$. Let us assume a single $X \sim \mathsf{N}_p(\mu, \sigma^2 \mathcal{I}_p)$ with a known $\sigma^2 > 0$, where $\mathcal{I}_p$ denotes the identity matrix of dimension p. The estimator is formulated as an estimator shrunken towards 0 in the form

$$\left(1 - \frac{(p-2)\sigma^2}{||X||_2^2}\right) X, \tag{7}$$

which has the structure of the Bayesian estimator in (6) obtained with a Gaussian prior with a zero expectation.

The James–Stein estimator (7) is a foundational result demonstrating the advantages of shrinkage estimation under the assumption of a known diagonal covariance

matrix with equal variances σ^2 across all coordinates. Although these assumptions may not hold in all practical scenarios, the estimator's core principles have inspired a wide range of effective regularization methods that are widely used today.

An analogous shrinkage estimator can also be defined for the sample mean of a Gaussian distribution when the covariance matrix is diagonal and known, but the coordinate-wise variances $\sigma_1^2, \ldots, \sigma_p^2$ may differ [6]. This extension relaxes the assumption of homoscedasticity and allows for heterogeneous variability across coordinates. The classical James–Stein estimator nevertheless remains a cornerstone theoretical result, providing strong justification for the use of regularized (i.e. Bayesian) estimators in multivariate settings [18].

Building on the concept of shrinkage in (7), many popular regularized estimators have been developed, including ridge regression and lasso regression. These ideas also form the basis of modern regularization techniques frequently employed in machine learning, such as bagging, boosting, and weight decay in deep learning [13]. In practice, when parameter variability such as variance is unknown, Empirical Bayes methods offer a useful alternative to the Stein estimator. They estimate hyperparameters directly from the data and enable adaptive, data-driven shrinkage [29].

If the task is to estimate the covariance matrix Σ, it is common—especially in high-dimensional applications—to apply Tikhonov regularization, replacing the empirical covariance matrix S by

$$\tilde{S} = (1 - \lambda)S + \lambda T \in \mathsf{PD}(p) \tag{8}$$

with a given regular target matrix $T \in \mathsf{PD}(p)$, depending on a tuning parameter $\lambda \in [0, 1)$. This regularized estimator $\tilde{S}$ is guaranteed to be positive definite and well-conditioned, even when S is singular or ill-conditioned due to a limited sample size or high dimensionality. This not only ensures the existence and uniqueness of subsequent solutions (e.g. in discriminant analysis), but also reduces the sensitivity of the model to noise, improves numerical stability, and mitigates overfitting.

Moreover, $\tilde{S}$ can also be interpreted from a Bayesian perspective [4], where the target matrix T corresponds to the covariance matrix of the prior distribution of the random Σ. The asymptotically optimal value of λ minimizing the mean squared error in (1) was derived in [24] for the case $T = \mathcal{I}_p$, eliminating the need for cross-validation. This result was later extended to more general choices of T in [32]. Regularized estimators of Σ that are additionally robust to outliers have been proposed for data following elliptically symmetric unimodal distributions; see e.g. [9, 3]. In practice, we often encounter situations where parameters are not estimated precisely but, for example, rounded (quantized). This leads us to consider how quantization relates to the Bayesian perspective.

3 Connection between quantization and Bayesian estimation

In this section, we formalize the connection between quantized estimation—that is, estimation involving rounding or discretization—and Bayesian estimation. Quantization introduces bias while potentially reducing variability, and it can be interpreted as a projection of the estimator onto a discrete set. Therefore, it is of interest to investigate whether quantization can be regarded as an implicit Bayesian estimator.

One common form of quantization, known as post-training rounding, has become a widely used strategy for reducing energy consumption and computational latency in statistical and machine learning models [26]. The key idea is to replace high-precision parameter estimates by their low-precision (typically integer or fixed-point) approximations while maintaining acceptable predictive performance.

Model comparison via reduced numerical precision falls under approximate computing, which studies how controlled approximations (such as rounding, noise, or perturbations) affect model behavior and reliability. In deep learning, such approximations are deliberately introduced to model small errors in learned parameters, aiming to simplify deployment on resource-constrained devices [34]. Similarly, when training is performed using low-precision arithmetic, rounding effects become inherent to the learning process and must be accounted for in both the design and evaluation of models [31].

Quantization is also gaining traction in probabilistic modeling and Bayesian inference. For example, a Bayesian regression model with quantized parameters was proposed in [40] for meteorological and hydrological applications, where the goal was to quantify how uncertainties in precipitation and temperature propagate through parameter estimates and affect predicted runoff in a snowmelt-driven watershed. More recently, numerical experiments have demonstrated how parameter quantization impacts inference in Bayesian networks [28].

We consider a multivariate model, which is a special case of (1). If the distribution of the error (e.g., due to rounding) is known, estimating $\mu \in \mathbb{R}^p$ becomes a Bayesian estimation problem. The error distribution then plays the role of a prior on μ in the original, uncontaminated model. The mathematical results highlight a conceptual link between approximate computing and Bayesian inference.

Model 1. *Let us consider the model*

$$X_i = \mu + e_i, \quad i = 1, \dots, n, \tag{9}$$

with $X_i \in \mathbb{R}^p$, a fixed $\mu \in \mathbb{R}^p$, and i.i.d. random vectors $e_1, \dots, e_n$, which follow $e_i \sim \mathsf{N}_p(0, \sigma^2 \mathcal{I}_p)$ for $i = 1, \dots, n$. Instead of the true model (9), the parameter estimation will be rather considered in the model

$$X_i = \xi + \varepsilon_i, \quad i = 1, \dots, n, \quad \text{where } \xi \text{ is obtained as } \xi = \mu + \tau \tag{10}$$

with a random vector $\tau \sim \mathsf{N}_p(0, \delta^2 \mathcal{I}_p)$ independent of $\varepsilon_1, \dots, \varepsilon_n$.

Model 2. *Let us consider the model*

$$X_i = \xi + e_i, \quad i = 1, \ldots, n, \tag{11}$$

where i.i.d. random vectors $e_1, \ldots, e_n$ *follow* $e_i \sim \mathsf{N}_p(0, \sigma^2 \mathcal{I}_p)$ *for* $i = 1, \ldots, n$ *and* $\xi \in \mathbb{R}^p$ *is a random vector following* $\xi \sim \mathsf{N}_p(\mu, \delta^2 \mathcal{I}_p)$ *with a given* $\mu \in \mathbb{R}^p$.

Theorem 1. *The posterior estimator of* ξ *in Model 1 is the same as the posterior estimator of* ξ *in Model 2.*

Proof. The idea is that a random vector ξ defined as $\xi = \mu + \tau$, where μ is a constant vector and $\tau \sim \mathcal{N}(0, \delta^2 I_p)$, satisfies $\xi \sim \mathcal{N}_p(\mu, \delta^2 I_p)$. □

It is now natural to ask what happens if Model 1 considers a prior distribution about μ instead of taking a fixed μ. In such a setup with a double uncertainty (about μ and due to the quantization), the next theorem explains that standard Bayesian estimation may be exploited combining both sources of uncertainty in an additive way.

Model 3. *Let us consider the model*

$$X_i = \mu + e_i, \quad i = 1, \ldots, n, \tag{12}$$

with $X_i \in \mathbb{R}^p$, *a random vector* $\mu \sim \mathsf{N}_p(\theta, \Psi)$ *with a fixed* $\theta \in \mathbb{R}^p$, $\Psi \in \mathsf{PD}(p)$, *and i.i.d. random vectors* $e_1, \ldots, e_n$ *following* $e_i \sim \mathsf{N}_p(0, \sigma^2 \mathcal{I}_p)$ *for* $i = 1, \ldots, n$. *Instead of the true model (12), the parameter estimation will be rather considered in the model*

$$X_i = \xi + \varepsilon_i, \quad i = 1, \ldots, n, \quad \text{where } \xi \text{ is obtained as } \xi = \mu + \tau \tag{13}$$

with a random vector $\tau \sim \mathsf{N}_p(0, \delta^2 \mathcal{I}_p)$ *independent of* μ *and* $\varepsilon_1, \ldots, \varepsilon_n$.

Model 4. *Let us consider the model*

$$X_i = \xi + e_i, \quad i = 1, \ldots, n, \tag{14}$$

where i.i.d. random vectors $e_1, \ldots, e_n$ *follow* $e_i \sim \mathsf{N}_p(0, \sigma^2 \mathcal{I}_p)$ *for* $i = 1, \ldots, n$ *and* $\xi \in \mathbb{R}^p$ *is a random vector following* $\xi \sim \mathsf{N}_p(\theta, \Psi + \delta^2 \mathcal{I}_p)$ *with a given* $\theta \in \mathbb{R}^p$ *and* $\Psi \in \mathsf{PD}(p)$.

Theorem 2. *The posterior estimator of* ξ *in Model 3 is the same as the posterior estimator of* ξ *in Model 4.*

Proof. Let us assume $\mu \sim \mathsf{N}_p(\theta, \Psi)$ and $\tau \sim \mathsf{N}_p(0, \delta^2 \mathcal{I}_p)$ independent of μ. In this case, the idea is that ξ defined as $\xi = \mu + \tau$ fulfils $\xi \sim \mathsf{N}_p(\theta, \Psi + \delta^2 \mathcal{I}_p)$. The posterior mean of ξ is the Bayes estimator under squared error loss. □

The connection between quantization and Bayesian statistics represents a significant shift in perspective. First, it marks a paradigm change by moving from viewing quantization as a purely deterministic, technical artifact to understanding it as an inherent part of probabilistic modeling. Second, it offers a unifying framework that bridges seemingly distinct concepts—quantization, rounding errors, and Bayesian inference—under a common statistical umbrella. Finally, this insight has practical implications: it suggests new ways to design algorithms that explicitly incorporate uncertainty introduced by quantization, potentially leading to more robust and interpretable models in real-world applications.

A limitation of our presented approach lies in the fact that quantization is modeled as deterministic post-training rounding, without considering its interaction with training dynamics or data structure; alternatively, rounding errors could be modeled stochastically, for example as random perturbations drawn from a uniform distribution. Next, we turn to applying Bayesian principles to a specific problem: the shrinkage estimation of group means. Shrinkage can be viewed as a form of regularization, which is naturally interpreted as a Bayesian estimate with a suitable prior.

4 Shrunken means as Bayesian estimators for two groups

Assuming two independent random samples, it is common in high-dimensional applications to consider their means shrunken towards a constant value (e.g., the pooled mean). For instance, shrinking both the means and the covariance matrix jointly was studied in [1] in the context of hypothesis testing for data from multiple groups; however, the study did not compare the tests based on shrunken means with those using standard (non-shrunken) means. Other examples include the classification task of [10] or more recent classifiers based on various specialized norms [33, 25].

In high-dimensional settings, the standard estimator of the mean often becomes highly unstable. Applying shrinkage to group means can dramatically reduce the mean squared error (MSE). Understanding the connection between shrinkage and Bayesian estimation enables the systematic derivation of shrinkage rules, such as those employed in linear discriminant analysis (LDA), which will be discussed in detail in the next section (Section 5).

An important and common scenario involves applying shrinkage across two or more estimators, thereby effectively borrowing information across different samples or groups. While shrinkage estimators are often presented in the literature as empirically motivated techniques, their Bayesian interpretation is rarely made explicit. In this work, we provide a theoretical justification for these existing shrinkage methods, which frequently appear as ad hoc solutions or are employed without clearly acknowledging their underlying Bayesian connection.

This section focuses on justifying the use of shrinkage means within a specific Bayesian framework, assuming the covariance matrix is regular. First, let us for-

mulate the result for univariate data and then we extend it to the multivariate case.

Lemma 1. *Let us assume two independent random samples*

$$X_1, \ldots, X_n \sim \mathsf{N}(\mu_x, \sigma^2) \quad \textit{and} \quad Y_1, \ldots, Y_m \sim \mathsf{N}(\mu_y, \sigma^2) \tag{15}$$

of univariate observations with a common variance $\sigma > 0$. Let us assume that the prior distributions for μ_x and μ_y are independent and of the form

$$\mu_x \sim \mathsf{N}(\theta, \gamma^2) \quad \textit{and} \quad \mu_y \sim \mathsf{N}(\theta, \gamma^2) \tag{16}$$

with $\theta \in \mathbb{R}^p$ and $\gamma > 0$. Then, estimates of μ_x and μ_y obtained as the means of the posterior distribution have the form

$$\hat{\mu}_x = \frac{n\bar{X}\sigma^{-2} + \theta\gamma^{-2}}{n\sigma^{-2} + \gamma^{-2}} \quad \textit{and} \quad \hat{\mu}_y = \frac{m\bar{Y}\sigma^{-2} + \theta\gamma^{-2}}{m\sigma^{-2} + \gamma^{-2}}. \tag{17}$$

Proof. The likelihood of the data across the two random samples, up to a multiplicative constant, is

$$f(\text{data}|\mu_x, \mu_y, \sigma^2) \propto \exp\left\{ -\frac{1}{\sigma^2} \left[\sum_{i=1}^{n} (X_i - \mu_x)^2 + \sum_{j=1}^{m} (Y_j - \mu_y)^2 \right] \right\}. \tag{18}$$

The prior distribution, if assuming the distribution of μ_x to be independent from that of μ_y, corresponds to two one-dimensional densities with the product

$$f(\mu_x) f(\mu_y) = \left(\frac{1}{2\pi\gamma^2} \right)^2 \exp\left\{ -\frac{1}{\gamma^2} \left[(\mu_x - \theta)^2 + (\mu_y - \theta)^2 \right] \right\}. \tag{19}$$

The posterior density can be expressed as $f(\mu_x, \mu_y|\text{data}) \propto$

$$\exp\left\{ -\frac{\sum_{i=1}^{n}(X_i - \mu_x)^2}{\sigma^2} - \frac{\sum_{j=1}^{m}(Y_j - \mu_y)^2}{\sigma^2} - \frac{(\mu_x - \theta)^2}{\gamma^2} - \frac{(\mu_y - \theta)^2}{\gamma^2} \right\}. \tag{20}$$

Let us search for the partial derivatives of the posterior density. Using the chain rule, the derivatives of the inside function will be put equal to 0. Solving this set of two equations for μ_x and μ_y leads us finally to

$$\hat{\mu}_x = \frac{\gamma^2 \sum_{i=1}^{n} X_i + \theta\sigma^2}{n\gamma^2 + \sigma^2} \quad \text{and} \quad \hat{\mu}_y = \frac{\gamma^2 \sum_{j=1}^{m} Y_j + \theta\sigma^2}{m\gamma^2 + \sigma^2}, \tag{21}$$

which can be equivalently reformulated as (24), i.e. the proof is concluded. □

Lemma 2. *Let us assume two independent random samples*

$$X_1, \ldots, X_n \sim \mathsf{N}_p(\mu_x, \Sigma) \quad \textit{and} \quad Y_1, \ldots, Y_m \sim \mathsf{N}_p(\mu_y, \Sigma) \tag{22}$$

of p-variate observations with $p > 1$. We assume a common covariance matrix $\Sigma \in \mathsf{PD}(p)$. Let us assume that the prior distributions for μ_x and μ_y are independent and of the form

$$\mu_x \sim \mathsf{N}_p(\theta, \Upsilon) \quad \textit{and} \quad \mu_y \sim \mathsf{N}_p(\theta, \Upsilon) \tag{23}$$

with $\theta \in \mathbb{R}^p$ and $\Upsilon \in \mathsf{PD}(p)$. Then, estimates of μ_x and μ_y obtained as the means of the posterior distribution have the form

$$\begin{aligned} \hat{\mu}_x &= (n\Sigma^{-1} + \Upsilon^{-1})^{-1}(n\Sigma^{-1}\bar{X} + \Upsilon^{-1}\theta), \\ \hat{\mu}_y &= (m\Sigma^{-1} + \Upsilon^{-1})^{-1}(m\Sigma^{-1}\bar{Y} + \Upsilon^{-1}\theta). \end{aligned} \tag{24}$$

The multivariate estimates obtained in Lemma 2 are shrunken to the common value θ, in an analogy to (17). Particularly, (24) may be reformulated in the elegant form as

$$\hat{\mu}_x = (\mathcal{I}_p - \Delta_x)\bar{X} + \Delta\theta \quad \text{and} \quad \hat{\mu}_y = (\mathcal{I}_p - \Delta_y)\bar{Y} + \delta\theta, \tag{25}$$

where the matrices $\Delta_x \in \mathbb{R}^{p\times p}$ and $\Delta_y \in \mathbb{R}^{p\times p}$ have the form

$$\Delta_x = (n\Sigma^{-1} + \Upsilon^{-1})^{-1}\Upsilon^{-1} \quad \text{and} \quad \Delta_y = (m\Sigma^{-1} + \Upsilon^{-1})^{-1}\Upsilon^{-1}. \tag{26}$$

These Bayesian estimators of μ_x and μ_y depend on Σ, just like the estimator (5) in Section 2. In the specific situation with $\Upsilon^{-1} = c\Sigma^{-1}$ with $c > 0$ and $n = m$, the estimators retain the form (25) while Δ_x and Δ_y simplify to a common matrix

$$\Delta_x = \Delta_y = \frac{c}{n+c}\mathcal{I}_p. \tag{27}$$

For practical applications, it is natural to take θ in (16) to be the pooled mean of the data across both groups. In other words, we justify here the commonly used approach based on shrinking the means in the context of LDA. The results of this section can be simply extended to the case with more than two groups. On the other hand, relying on specific distributional assumptions and the availability of a regular covariance matrix may limit the applicability of the method; furthermore, the performance and robustness of shrunken means under model misspecification remain beyond the scope of this section.

5 Regularized estimators within linear discriminant analysis

Regularized versions of LDA are studied in this section, along with novel algorithms for their computation. These methods leverage the shrinkage estimation of means discussed in Section 4, where shrinkage of the means represents one part of the overall regularization strategy. Another crucial component is the regularization of the covariance matrix. This technique is primarily applied to ensure

numerical stability and invertibility, especially in high-dimensional settings where the covariance matrix may be ill-conditioned or singular. Together, these regularization approaches enable more stable and reliable classification. In fact, LDA tends to perform poorly in high-dimensional settings; incorporating regularization enhances both its generalization capability and numerical stability, making it better suited for such challenging scenarios.

Linear discriminant analysis (LDA) is a classification method for p-variate observations

$$X_{11}, \ldots, X_{1n_1}, \ldots, X_{K1}, \ldots, X_{Kn_K} \tag{28}$$

observed in K different groups with $p > K \geq 2$. We assume the data in each of the $k = 1, \ldots, K$ groups to follow the normal $\mathsf{N}_p(\mu_k, \Sigma)$ distribution and allow the data to be high-dimensional allowing $\min_{k=1,\ldots,K} n_k < p$. If the data are high-dimensional, which is a common situation e.g. in molecular genetics, econometrics or chemistry, using a regularized estimate of Σ is typically highly beneficial [2]. While the means of each group are commonly computed as arithmetic means, their regularization might be beneficial as well. Shrinking the means towards the pooled mean across groups was used e.g. in [8], however without pointing out at the Bayesian connections discussed above.

Regularized estimation of Σ will be considered here as a replacement of the pooled empirical covariance matrix S by the shrinkage version $\tilde{S}$ (8) depending on $\lambda \in [0, 1]$. The Bayesian connections of $\tilde{S}$ were discussed already in Section 2. Let us remark that $\tilde{S}$ evaluates the covariance structure around $\bar{X}$; we are not aware of any alternative approach considering the covariance structure around a regularized mean.

Let us now formulate several possible versions of regularized means inspired by Section 4, which correspond to different forms of regularization. They depend on regularization parameters and it is natural to select the values that lead to the best classification performance; such values can be found in a cross-validation. We use the notation $\bar{X}$ for the pooled mean, $\bar{X}_k$ for the mean of the k-th group, and $\mathbb{1}$ for indicator function.

Definition 1 (Regularized means for $k = 1, \ldots, K$)**.**

1. $$\bar{X}_k^{(2)} = (1 - \delta^{(2)})\bar{X}_k + \delta^{(2)}\bar{X}, \quad \delta^{(2)} \in (0, 1), \tag{29}$$

2. $$\begin{aligned} \bar{X}_k^{(1)} &= \mathsf{sgn}(\bar{X}_k) \max\left\{|\bar{X}_k| - \delta^{(1)}, 0\right\} \\ &= \mathsf{sgn}(\bar{X}_k)\left(|\bar{X}_k| - \delta^{(1)}\right)_+, \quad \delta^{(1)} \in (0, 1), \end{aligned} \tag{30}$$

3. $$\begin{aligned} \bar{X}_k^{(0)} &= \bar{X}_k \mathbb{1}\left[|\bar{X}_k| > \delta^{(0)}\right] = \bar{X}_k \mathbb{1}\left[\bar{X}_k \notin (-\delta^{(0)}, \delta^{(0)})\right] \\ &= \begin{cases} \bar{X}_k, & |\bar{X}_k| \geq \delta^{(0)}, \\ 0, & |\bar{X}_k| \leq \delta^{(0)}. \end{cases} \end{aligned} \tag{31}$$

Both (30) and (31) may yield sparse solutions, which is particularly appealing in classification tasks involving high-dimensional data, as illustrated in image

analysis applications [22]. In particular, hard thresholding (31) may result in some coordinates being set to zero; if a given coordinate is zero across all groups k, the corresponding variable effectively drops out—this constitutes a form of feature selection. The hard thresholding in (31) is inspired by image denoising applications of [5], where the goal is to remove noise while preserving important features by setting small coefficients to zero.

None of these estimators is universally optimal. Regularization based on the L_2-norm often outperforms L_1-based methods in various statistical models [27], as it tends to work well when all variables contribute meaningfully to the prediction. Conversely, L_1-based approaches are more appropriate when only a small number of dominant variables are expected to be relevant. The hard thresholding in (31) takes a more aggressive sparsity-inducing stance by directly setting small coefficients to zero rather than continuously shrinking them, which can be particularly effective when a clear zero/non-zero distinction is desired. Moreover, in situations where continuous shrinkage of average values compromises interpretability, the hard thresholding (31) provides an appealing alternative by enforcing explicit sparsity through hard cutoffs.

The Bayesian interpretation of the soft-thresholding estimator (30), in which a Laplace prior induces shrinkage, provides a probabilistic justification for its use. In contrast, to our knowledge, no analogous Bayesian framework has been established for the hard-thresholding estimator (31), suggesting an interesting direction for future research.

To our knowledge, no version of LDA using (31) has been proposed in the literature; we introduce it here as a simplified alternative to (30), in which no shrinkage is applied when $\bar{X}_k \geq \delta^{(0)}$. The version of LDA, where only S is regularized but not the means, will be denoted here simply as RLDA. Versions with regularization applied also on the means will be denoted as RLDA(0), RLDA(1), and RLDA(2) according to the regularization type applied on the means according to Definition 1. We consider $Z \in \mathbb{R}^p$ as an observation to be classified. The classification rule assigns Z to the group that is obtained by

$$\tilde{k} := \underset{k=1,\ldots,K}{\operatorname{argmax}}\, l_k, \tag{32}$$

where l_k is the discriminant score. Prior knowledge about group membership will be denoted as π_k for the group $k = 1, \ldots, K$, where $\sum_{k=1}^{K} \pi_k = 1$. The formal definition follows.

Definition 2 (Discriminant scores of regularized LDA for $k = 1, \ldots, K$)**.**

1. *Plain RLDA*

$$\tilde{\ell}_k = \bar{X}_k^T \tilde{S}^{-1} Z - \frac{1}{2} \bar{X}_k^T \tilde{S}^{-1} \bar{X}_k + \log \pi_k. \tag{33}$$

2. *RLDA*(ω) *for* $\omega \in \{0, 1, 2\}$

$$\tilde{\ell}_k^{(\omega)} = \left(\bar{X}_k^{(\omega)}\right)^T \tilde{S}^{-1} Z - \frac{1}{2} \left(\bar{X}_k^{(\omega)}\right)^T \tilde{S}^{-1} \bar{X}_k^{(\omega)} + \log \pi_k. \tag{34}$$

A direct computation of the discriminant scores in (33) and (34) is highly inefficient due to the need to invert the empirical covariance matrix. Therefore, we present two efficient algorithms for computing regularized LDA. Both use fixed regularization parameters δ and λ that require to be optimized within cross-validation. Its advantage is also avoiding the need to specify the unknown Σ, which serves as a nuisance parameter here, and to specify hyperparameters (i.e. parameters of the underlying prior). Algorithm 1 is based on Cholesky decomposition of $\tilde{S}$. The optimal k in (34) exists with probability 1.

Algorithm 1 RLDA(2) for a general T based on Cholesky decomposition.

Require: Data (28) and $Z \in \mathbb{R}^p$.
Require: $T \in \mathsf{PD}(p)$.
Require: $\delta \in [0,1]$ and $\lambda \in [0,1]$.
Require: $\pi_1, \ldots, \pi_K$ with $\sum_{k=1}^{K} \pi_k = 1$.
Ensure: Classification decision for given $Z \in \mathbb{R}^p$.

1: Compute $\tilde{S}$ according to

$$\tilde{S} = (1-\lambda)S + \lambda T. \tag{35}$$

2: Compute the Cholesky factor L_λ of $\tilde{S}$ as

$$\tilde{S} = L_\lambda L_\lambda^T, \tag{36}$$

where L_λ is a nonsingular lower triangular matrix.

3: Compute the matrix

$$B^{\delta\lambda} = L_\lambda^{-T} \left[(1-\delta)\bar{X}_1 + \delta\bar{X} - Z, \ldots, (1-\delta)\bar{X}_K + \delta\bar{X} - Z \right], \tag{37}$$

where $L_\lambda^{-T} = (L_\lambda^{-1})^T = (L_\lambda^T)^{-1}$, and assign Z to group k, if

$$k = \underset{j=1,\ldots,K}{\operatorname{argmax}} \left\{ ||B_j^{\delta\lambda}||_2^2 + \log \pi_k \right\}, \tag{38}$$

where $||B_j^{\delta\lambda}||_2^2$ is the squared Euclidean norm of the j-th column of $B^{\delta\lambda}$.

Let us also consider a specific form of regularization, namely the ridge regularization (as denoted e.g. in [12]) in the form

$$S^* = \lambda S + (1-\lambda)\mathcal{I}_p, \quad \lambda \in [0,1]. \tag{39}$$

We use the notation

$$\tilde{X} = [X_{11} - \bar{X}, \ldots, X_{1n_1} - \bar{X}, \ldots, X_{K1} - \bar{X}, \ldots, X_{Kn_K} - \bar{X}]^T \in \mathbb{R}^{n \times p} \tag{40}$$

and express the pooled estimator S in the form $S = \tilde{X}^T\tilde{X}$.

Algorithm 1, which is based on the Cholesky decomposition, is efficient and stable for well-conditioned covariance matrices but may fail or become unstable

when the matrix is nearly singular or not positive definite. In contrast, Algorithm 2, relying on singular value decomposition, provides greater numerical stability and robustness in such challenging cases, albeit at a higher computational cost. Despite this, the numerical results produced by both algorithms are expected to be highly comparable. Both methods have similar asymptotic computational complexity of order $O(p^3)$, where p denotes the size of the covariance matrix. Nevertheless, Cholesky decomposition is typically 2 to 4 times faster than SVD in practical runtime measurements [37].

Algorithm 2 RLDA(2) for the ridge regularization.

Require: Data (28) with $n < p$ and $Z \in \mathbb{R}^p$.
Require: $\delta \in [0,1]$ and $\lambda \in [0,1]$.
Require: $\pi_1, \ldots, \pi_K$ with $\sum_{k=1}^{K} \pi_k = 1$.
Ensure: Classification decision for given $Z \in \mathbb{R}^p$.

1: Compute the matrix $\tilde{X}$ as in (40).
2: Compute the singular value decomposition of $\tilde{X}$ as

$$\tilde{X} = P\Omega Q^T, \tag{41}$$

with singular values $\{\omega_1, \ldots, \omega_n\}$ and complement these singular values with $p - n$ zero values $\omega_{n+1} = \cdots = \omega_p = 0$.
3: Estimate the variances of the columns of $\tilde{X}$, denoted by $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2$
4: For a fixed $\lambda \in [0,1]$, compute

$$D_\lambda = \mathsf{diag}\{\lambda\hat{\sigma}_1^2 + (1-\lambda), \ldots, \lambda\hat{\sigma}_p^2 + (1-\lambda)\}, \tag{42}$$

where diag denotes a diagonal matrix.
5: Compute the matrix

$$B^{\delta\lambda} = D_\lambda^{-1/2} Q^T \left[(1-\delta)\bar{X}_1 + \delta\bar{X} - Z, \ldots, (1-\delta)\bar{X}_K + \delta\bar{X} - Z\right] \tag{43}$$

and assign Z to group k, if

$$k = \underset{j=1,\ldots,K}{\operatorname{argmax}} \left\{ ||B_j^{\delta\lambda}||_2^2 + \log \pi_k \right\}. \tag{44}$$

Slight modifications can be used to adapt the algorithms for other LDA versions of Definition 2. All the described versions of LDA, which are extensions of simple regularized LDA versions of [21], are sensitive to outliers, because the mean and empirical covariance matrix are vulnerable to the presence of outliers in the data. A desirable robustification can be performed by exploiting a suitable highly robust estimator, such as the minimum weighted covariance determinant (MWCD) estimator [30, 20]. Moreover, regularization enables the extension of LDA to high-dimensional settings, including the incorporation of robust estimators such as those described in [16]. A limitation of this approach is that the decision rule remains

linear, relying on the assumption of normally distributed classes with a common covariance matrix; this may reduce classification performance when group distributions are non-Gaussian or exhibit different covariance structures.

6 Experiments

The aim of the experiments is to study the regularized LDA of Section 5 on real as well as simulated data.

As a real dataset, we use the genetic data coming from the cardiovascular genetic study performed in the Center of Biomedical Informatics in Prague. The data are gene expressions for 38 590 gene transcripts measured for 46 patients with acute myocardial infarction (AMI) and 49 control persons, previously analyzed in [20],

In a simulation study, we randomly generate two independent random samples both with $n = m = 50$ observations and $p = 1000$ variables. The first random sample is generated from $\mathsf{N}_p(0, \Sigma)$ with $\Sigma = \sigma^2[(1-c)\mathcal{I}_p + c1_p 1_p^T]$, where $\sigma = 1$, $c = 0.4$, and $1_p = (1, \ldots, 1) \in \mathbb{R}^p$. The second random samples is generated from $\mathsf{N}_p(\mu_y, \Sigma)$, where $\mu_y = (3, 3, 3, 3, 3, 0, \ldots, 0)^T \in \mathbb{R}^p$ is the vector with the first 5 coordinates being non-zero.

The methods used to analyze the two datasets are listed in Table 1. Because standard LDA cannot be computed due to a singular covariance matrix, we use regularized versions using the shrinkage targets $T_1 = \mathcal{I}_p$ or $T_2 = \sigma^2\mathcal{I}_p + \vartheta(1_p 1_p^T - \mathcal{I}_p)$ with $\vartheta = 0.15$. All our implementations use the Cholesky decomposition of the covariance matrix. If the means are also regularized, then the regularization types of Definition 1 are used. For finding the regularization parameters, 5-fold cross-validation is typically used. If only the covariance matrix is regularized, we also use the asymptotically optimal values of the regularization parameter; this was derived for T_1 as well as T_2 in [23]. For comparisons, prediction analysis of microarrays (PAM) was computed with default regularization for the means; no regularization of the covariance matrix is needed due its diagonal structure [35].

The results are presented in Table 1. PAM has a weaker classification performance than any of the regularized LDA versions. The best result is obtained in both datasets when T_2 is used as a shrinkage target and when doing the hard thresholding for the means. Results obtained with the shrinkage target T_1, which is the most common type, stay only behind. Regularizing the means improves the performance in every case compared to situations with standard means.

The choice of regularized LDA with T_2 and L_1-regularization for the means exactly corresponds to SCRDA of [10] implemented in the `rda` package of R software [11]; the best results obtained with T_2 and hard thresholding for the means overcome those of [10] remarkably. We can say that hard thresholding is most effective when the chosen regularization corresponds to a strong prior belief and the threshold is suitably selected. In the genetic dataset, it turns out that a set of a few hundred of variables is responsible for the classification between the two groups; such finding corresponds to the analysis of [20]. In the simulation, the most successful LDA versions exploit only 5 variables, which corresponds to the

Table 1: Results of the experiments of Section 6. The classification accuracy (Acc.), its standard deviation (SD) evaluated in 5-fold cross-validation, and the number of selected variables used in the classification rule are reported. Methods not performing variable selection use as many as p variables. Different methods are compared, using the regularization parameters obtained in 5-fold cross-validation (CV) or using the Ledoit-Wolf asymptotics (LW).

Classif. method	Cov. matrix T	Expect. Reg. type	Reg. param. choice	Genetic data Acc.	(SD)	# of vars	Simulated data Acc.	(SD)	# of vars
PAM	None	L_1	CV	0.69	(0.05)	541	0.75	(0.05)	51
LDA	T_1	-	CV	0.77	(0.03)	38 590	0.84	(0.03)	1000
LDA	T_1	-	LW	0.74	(0.03)	38 590	0.82	(0.03)	1000
LDA	T_2	-	CV	0.79	(0.03)	38 590	0.86	(0.03)	1000
LDA	T_2	-	LW	0.76	(0.03)	38 590	0.84	(0.03)	1000
LDA	T_1	L_2	CV	0.78	(0.02)	38 590	0.86	(0.03)	1000
LDA	T_1	L_1	CV	0.82	(0.02)	287	0.88	(0.03)	21
LDA	T_1	(31)	CV	0.83	(0.02)	260	0.88	(0.03)	14
LDA	T_2	L_2	CV	0.81	(0.02)	38 590	0.90	(0.03)	1000
LDA	T_2	L_1	CV	0.83	(0.02)	222	0.91	(0.03)	5
LDA	T_2	(31)	CV	0.86	(0.02)	213	0.92	(0.03)	5

true number of variables where the shift between the two groups occurs.

It is also worth noting that the shrinkage target T_2 precisely corresponds to the data-generating design used in the simulation; for the genetic data, T_2 is apparently a reasonable target as well, as the variables exhibit non-negligible correlations, making a diagonal covariance structure inappropriate.

The asymptotic regularization parameter values turn out to be overcome by those found in the cross-validation; this is because the approach of Ledoit-Wolf [24] does not consider regularized means and the optimality of the regularization parameter holds only asymptotically and would be relevant only for larger numbers of observations. We may conclude that not only the regularization as such, but also the way of finding the parameters are crucial for the classification performance.

7 Conclusion

Regularized (shrinkage) estimators have already been exploited in a variety of multivariate applications. Meanwhile, regularized versions of LDA represent popular tools for the analysis of high-dimensional data. Importantly, the theoretical results and proposed estimators are derived for finite sample sizes and do not rely on asymptotic approximations, which makes them directly applicable in practical scenarios. This paper justifies some regularized estimators by deriving them within specific Bayesian contexts. At the same time, regularization is known to be re-

lated to local robustness, i.e. to the known resistance of regularized estimators to local shift modification of the majority of the data [14, 39], which can be explained within the framework of robust optimization [38].

The proposed methods were illustrated by numerical experiments on both real and simulated data, supporting their practical relevance. Experiments not shown here verified that the results for Algorithm 2 are very close to those of Algorithm 1, if the ridge regularized is used.

The Bayesian approach is explained in Section 3 to allow also a different and original interpretation: it corresponds to training the parameters perturbed by errors. As a consequence, researchers interested in the effect of post-training rounding of the parameters or computing in a low-precision arithmetic should focus on Bayesian (i.e. regularized) estimation.

The Bayesian thinking used here may also be applied to a number of other multivariate contexts. Because regularized LDA is non-robust to outliers, we plan to perform future research of novel robust regularized classification methods based on the highly robust MWCD estimator [30].

Among the main limitations of the presented approach are (1) the need to specify a prior distribution, which is an inherent feature of any Bayesian framework, and (2) the sensitivity to model misspecification, a drawback stemming from the parametric nature of the model rather than from Bayesian inference itself. Nevertheless, in practice, cross-validation can be used to select regularization parameters that correspond to suitable prior choices, thus mitigating the impact of subjective prior specification. In addition, a certain degree of robustness can be achieved in robust LDA by using robust estimators of the mean and covariance matrix [17].

References

[1] Bernal, V., Bischoff, R., Guryev, V., Grzegorczyk, M., and Horvatovich, P. Exact hypothesis testing for shrinkage-based gaussian graphical models. *Bioinformatics*, 35:5011–5017, 2019. DOI: 10.1093/bioinformatics/btz357.

[2] Bose, A. and Bhattacharjee, M. *Large covariance and autocovariance matrices*. CRC Press, Boca Raton, 2020. DOI: 10.1201/9780203730652.

[3] Boudt, K., Rousseeuw, P., Vanduffel, S., and Verdonck, T. The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30:113–128, 2020. DOI: 10.1007/s11222-019-09869-x.

[4] Calvetti, D. and Somersalo, E. Inverse problems: From regularization to bayesian inference. *WIREs Computational Statistics*, 10:e1427, 2018. DOI: 10.1002/wics.1427.

[5] Davies, P. *Data analysis and approximate models: Model choice, location-scale, analysis of variance, nonparametric regression and image analysis*. Chapman & Hall/CRC, Boca Raton, 2014. DOI: 10.1201/b17146.

[6] Efron, B. Machine learning and the James–Stein estimator. *Japanese Journal of Statistics and Data Science*, 7:257–266, 2024. DOI: 10.1007/s42081-023-00209-y.

[7] Evans, I. Bayesian estimation of parameters of a multivariate normal distribution. *Journal of the Royal Statistical Society B*, 27:279–283, 1965. DOI: 10.1111/j.2517-6161.1965.tb01494.x.

[8] Fu, R., Han, M., Tian, Y., and Shi, P. Improvement motor imagery EEG classification based on sparse common spatial pattern and regularized discriminant analysis. *Journal of Neuroscience Methods*, 343:108833, 2020. DOI: 10.1016/j.jneumeth.2020.108833.

[9] Gschwandtner, M. and Filzmoser, P. Outlier detection in high dimension using regularization. In Kruse, R., Berthold, M. R., Moewes, C., Ángeles Gil, M., Grzegorzewski, P., and Hryniewicz, O., editors, *Synergies of Soft Computing and Statistics*, pages 237–244. Springer, Berlin, 2013. DOI: 10.1007/978-3-642-33042-1_26.

[10] Guo, Y., Hastie, T., and Tibshirani, R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8:86–100, 2007. DOI: 10.1093/biostatistics/kxj035.

[11] Guo, Y., Hastie, T., and Tibshirani, R. rda: Shrunken centroids regularized discriminant analysis. R package version 1.2-1, https://CRAN.R-project.org/package=rda, 2023.

[12] Hastie, T. Ridge regularization: An essential concept in data science. *Technometrics*, 62:426–433, 2020. DOI: 10.1080/00401706.2020.1791959.

[13] Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity. The lasso and generalizations*. CRC Press, Boca Raton, 2015. DOI: 10.1201/b18401.

[14] Hoffman, J., Roberts, D., and Yaida, S. Robust learning with Jacobian regularization. ArXiv:1908.02729, 2019. DOI: 10.48550/arXiv.1908.02729.

[15] Hosseini, S., Deiri, E., and Jamkhaneh, E. The Bayesian and Bayesian shrinkage estimators under square error and Al-Bayyati loss functions with right censoring scheme. *Communications in Statistics—Simulation and Computation*, 53:3234–3252, 2024. DOI: 10.1080/03610918.2022.2100906.

[16] Hubert, M., Debruyne, M., and Rousseeuw, P. Minimum covariance determinant and extensions. *WIREs Computational Statistics*, 10:e1421, 2018. DOI: 10.1002/wics.1421.

[17] Hubert, M., Raymaekers, J., and Rousseeuw, P. Robust discriminant analysis. *WIREs Computational Statistics*, 16:e70003, 2024. DOI: 10.1002/wics.70003.

[18] James, W. and Stein, C. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pages 361–379. University of California Press, Berkeley, 1961. DOI: 10.1007/978-1-4612-0919-5_30.

[19] Johnson, A., Ott, M., and Dogucu, M. *Bayes rules! An introduction to applied Bayesian modeling*. CRC Press, Boca Raton, 2022. DOI: 10.1201/9780429288340.

[20] Kalina, J. A robust pre-processing of beadchip microarray images. *Biocybernetics and Biomedical Engineering*, 38:556–563, 2018. DOI: 10.1016/j.bbe.2018.04.005.

[21] Kalina, J. and Duintjer Tebbens, J. Algorithms for regularized linear discriminant analysis. In *Proceedings of the 6th International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS '15)*, pages 128–133, Lisbon, 2015. Scitepress. DOI: 10.5220/0005234901280133.

[22] Kalina, J. and Matonoha, C. A sparse pair-preserving centroid-based supervised learning method for high-dimensional biomedical data or images. *Biocybernetics and Biomedical Engineering*, 40:774–786, 2020. DOI: 10.1016/j.bbe.2020.03.008.

[23] Ledoit, O. and Wolf, M. Honey, I shrunk the sample covariance matrix. UPF Economics and Business Working Paper 691, 2003. DOI: 10.3905/jpm.2004.110.

[24] Ledoit, O. and Wolf, M. The power of (non-)linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics*, 20:187–218, 2022. DOI: 10.1093/jjfinec/nbaa007.

[25] Li, C., Ren, P., Guo, Y., Ye, Y., and Shao, Y. Regularized linear discriminant analysis based on generalized capped $\ell_{2,q}$-norm. *Annals of Operations Research*, 339:1433–1459, 2024. DOI: 10.1007/s10479-022-04959-y.

[26] Nahshan, Y., Chmiel, B., Baskin, C., Zheltonozhskij, E., Banner, R., Bronstein, A., and Mendelson, A. Loss aware post-training quantization. *Machine Learning*, 110:3245–3262, 2021. DOI: 10.1007/s10994-021-06053-z.

[27] Piotrowski, A., Napiorkowski, J., and Piotrowska, A. Impact of deep learning-based dropout on shallow neural networks applied to stream temperature modelling. *Earth-Science Reviews*, 201:103076, 2020. DOI: 10.1016/j.earscirev.2019.103076.

[28] Ribeiro, R., Natal, J., de Campos, C., and Maciel, C. Conditional probability table limit-based quantization for Bayesian networks: Model quality, data fidelity and structure score. *Applied Intelligence*, 54:4668–4688, 2024. DOI: 10.1007/s10489-023-05153-8.

[29] Rizzelli, S., Rousseau, J., and Petrone, S. Empirical Bayes in Bayesian learning: Understanding a common practice. ArXiv:2402.19036v1, 2024. DOI: 10.48550/arXiv.2402.19036.

[30] Roelant, E., Van Aelst, S., and Willems, G. The minimum weighted covariance determinant estimator. *Metrika*, 70:177–204, 2009. DOI: 10.1007/s00184-008-0186-3.

[31] Saha, R., Srivastava, V., and Pilanci, M. Matrix compression via randomized low rank and low precision factorization. ArXiv:2310.11028, 2023. DOI: 10.48550/arXiv.2310.11028.

[32] Schäfer, J. and Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:32, 2005. DOI: 10.2202/1544-6115.1175.

[33] Sifaou, H., Kammoun, A., and Alouini, M. High-dimensional linear discriminant analysis classifier for spiked covariance model. *Journal of Machine Learning Research*, 21:1–24, 2020. DOI: 10.5555/3455716.3455828.

[34] Sze, V., Chen, Y., Yang, T., and Emer, J. *Efficient processing of deep neural networks*. Morgan & Claypool, San Rafael, 2020. DOI: 10.1007/978-3-031-01766-7.

[35] Tibshirani, R., Hastie, T., and Narasimhan, B. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 18:104–117, 2003. DOI: 10.1214/ss/1056397488.

[36] Valenta, Z. and Kalina, J. Exploiting Stein's paradox in analysing sparse data from genome-wide association studies. *Biocybernetics and Biomedical Engineering*, 35:64–67, 2015. DOI: 10.1016/j.bbe.2014.10.004.

[37] White, R. *Computational linear algebra with applications and MATLAB computations*. CRC Press, Boca Raton, 2023. DOI: 10.1201/9781003304128.

[38] Xanthopoulos, P., Pardalos, P., and Trafalis, T. *Robust data mining*. Springer, New York, 2013. DOI: 10.1007/978-1-4419-9878-1.

[39] Xie, P., Xiang, H., and Wei, Y. Randomized algorithms for total least squares problems. *Numerical Linear Algebra with Applications*, 26:e2219, 2019. DOI: 10.1002/nla.2219.

[40] Yan, X., Song, J., An, Y., and Lu, W. Uncertainty quantization of meteorological input and model parameters for hydrological modelling using a Bayesian-based integrated approach. *Hydrological Processes*, 38:e15040, 2024. DOI: 10.1002/hyp.15040.