NIH Public Access
Author Manuscript
Pattern Recognit. Author manuscript; available in PMC 2015 June 01.

Published in final edited form as:
Pattern Recognit. 2014 June 1; 47(6): 2178–2192. doi:10.1016/j.patcog.2013.11.022.

Moments and Root-Mean-Square Error of the Bayesian MMSE Estimator of Classification Error in the Gaussian Model

Amin Zollanvari[a,b,*] and **Edward R. Dougherty**[a,c]

[a]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843

[b]Department of Statistics, Texas A&M University, College Station, TX 77843

[c]Translational Genomics Research Institute (TGEN), Phoenix, AZ 85004

Abstract

The most important aspect of any classifier is its error rate, because this quantifies its predictive capacity. Thus, the accuracy of error estimation is critical. Error estimation is problematic in small-sample classifier design because the error must be estimated using the same data from which the classifier has been designed. Use of prior knowledge, in the form of a prior distribution on an uncertainty class of feature-label distributions to which the true, but unknown, feature-distribution belongs, can facilitate accurate error estimation (in the mean-square sense) in circumstances where accurate completely model-free error estimation is impossible. This paper provides analytic asymptotically exact finite-sample approximations for various performance metrics of the resulting Bayesian Minimum Mean-Square-Error (MMSE) error estimator in the case of linear discriminant analysis (LDA) in the multivariate Gaussian model. These performance metrics include the first, second, and cross moments of the Bayesian MMSE error estimator with the true error of LDA, and therefore, the Root-Mean-Square (RMS) error of the estimator. We lay down the theoretical groundwork for Kolmogorov double-asymptotics in a Bayesian setting, which enables us to derive asymptotic expressions of the desired performance metrics. From these we produce analytic finite-sample approximations and demonstrate their accuracy via numerical examples. Various examples illustrate the behavior of these approximations and their use in determining the necessary sample size to achieve a desired RMS. The Supplementary Material contains derivations for some equations and added figures.

Keywords

Linear discriminant analysis; Bayesian Minimum Mean-Square Error Estimator; Double asymptotics; Kolmogorov asymptotics; Performance metrics; RMS

1. Introduction

The most important aspect of any classifier is its error, ε, defined as the probability of misclassification, since ε quantifies the predictive capacity of the classifier. Relative to a classification rule and a given feature-label distribution, the error is a function of the

*Corresponding author: amin_zoll@neo.tamu.edu (Amin Zollanvari), edward@ece.tamu.edu (Edward R. Dougherty).

sampling distribution and as such possesses its own distribution, which characterizes the true performance of the classification rule. In practice, the error must be estimated from data by some error estimation rule yielding an estimate, $\hat{\varepsilon}$. If samples are large, then part of the data can be held out for error estimation; otherwise, the classification and error estimation rules are applied on the same set of training data, which is the situation that concerns us here. Like the true error, the estimated error is also a function of the sampling distribution. The performance of the error estimation rule is completely described by its joint distribution, (ε, $\hat{\varepsilon}$).

Three widely-used metrics for performance of an error estimator are the bias, deviation variance, and root-mean-square (RMS), given by

$$\begin{aligned} \mathrm{Bias}[\hat{\varepsilon}] &= E[\hat{\varepsilon}] - E[\varepsilon], \\ \mathrm{Var}^d[\hat{\varepsilon}] &= \mathrm{Var}(\hat{\varepsilon} - \varepsilon) = \mathrm{Var}(\varepsilon) + \mathrm{Var}(\hat{\varepsilon}) - 2\mathrm{Cov}(\varepsilon, \hat{\varepsilon}), \\ \mathrm{RMS}[\hat{\varepsilon}] &= \sqrt{E[(\varepsilon - \hat{\varepsilon})^2]} = \sqrt{E[\varepsilon^2] + E[\hat{\varepsilon}^2] - 2E[\varepsilon\hat{\varepsilon}]} = \sqrt{\mathrm{Bias}[\hat{\varepsilon}]^2 + \mathrm{Var}^d[\hat{\varepsilon}]}, \end{aligned} \tag{1}$$

respectively. The RMS (square root of mean square error, MSE) is the most important because it quantifies estimation accuracy. Bias requires only the first-order moments, whereas the deviation variance and RMS require also the second-order moments.

Historically, analytic study has mainly focused on the first marginal moment of the estimated error for linear discriminant analysis (LDA) in the Gaussian model or for multinomial discrimination [1]–[12]; however, marginal knowledge does not provide the joint probabilistic knowledge required for assessing estimation accuracy, in particular, the mixed second moment. Recent work has aimed at characterizing joint behavior. For multinomial discrimination, exact representations of the second-order moments, both marginal and mixed, for the true error and the resubstitution and leave-one-out estimators have been obtained [13]. For LDA, the exact joint distributions for both resubstitution and leave-one-out have been found in the univariate Gaussian model and approximations have been found in the multivariate model with a common known covariance matrix [14, 15]. Whereas one could utilize the approximate representations to find approximate moments via integration in the multivariate model with a common known covariance matrix, more accurate approximations, including the second-order mixed moment and the RMS, can be achieved via asymptotically exact analytic expressions using a double asymptotic approach, where both sample size (n) and dimensionality (p) approach infinity at a fixed rate between the two [16]. Finite-sample approximations from the double asymptotic method have shown to be quite accurate [16, 17, 18]. There is quite a body of work on the use of double asymptotics for the analysis of LDA and its related statistics [16, 19, 20, 21, 22, 23]. Raudys and Young provide a good review of the literature on the subject [24].

Although the theoretical underpinning of both [16] and the present paper relies on double asymptotic expansions, in which $n, p \to \infty$ at a proportional rate, our practical interest is in the finite-sample approximations corresponding to the asymptotic expansions. In [17], the accuracy of such finite-sample approximations was investigated relative to asymptotic expansions for the expected error of LDA in a Gaussian model. Several single-asymptotic expansions ($n \to \infty$) were considered, along with double-asymptotic expansions ($n, p \to \infty$) [19, 20]. The results of [17] show that the double-asymptotic approximations are significantly more accurate than the single-asymptotic approximations. In particular, even with $n/p < 3$, the double-asymptotic expansions yield “excellent approximations” while the others “falter.”

The aforementioned work is based on the assumption that a sample is drawn from a fixed feature-label distribution F, a classifier and error estimate are derived from the sample without using any knowledge concerning F, and performance is relative to F. Research dating to 1978, shows that small-sample error estimation under this paradigm tends to be inaccurate. Re-sampling methods such as cross-validation possess large deviation variance and, therefore, large RMS [9, 25]. Scientific content in the context of small-sample classification can be facilitated by prior knowledge [26, 27, 28]. There are three possibilities regarding the feature-label distribution: (1) F is known, in which case no data are needed and there is no error estimation issue; (2) nothing is known about F, there are no known RMS bounds, or those that are known are useless for small samples; and (3) F is known to belong to an uncertainty class of distributions and this knowledge can be used to either bound the RMS [16] or be used in conjunction with the training data to estimate the error of the designed classifier. If there exists a prior distribution governing the uncertainty class, then in essence we have a distributional model. Since virtually nothing can be said about the error estimate in the first two cases, for a classifier to possess scientific content we must begin with a distributional model.

Given the need for a distributional model, a natural approach is to find an optimal minimum mean-square-error (MMSE) error estimator relative to an uncertainty class Θ [27]. This results in a Bayesian approach with Θ being given a prior distribution, $\pi(\theta)$, $\theta \in \Theta$, and the sample S_n being used to construct a posterior distribution, $\pi^*(\theta)$, from which an optimal MMSE error estimator, $\hat{\varepsilon^B}$, can be derived. $\pi(\theta)$ provides a mathematical framework for both the analysis of any error estimator and the design of estimators with desirable properties or optimal performance. $\pi^*(\theta)$ provides a sample-conditioned distribution on the true classifier error, where randomness in the true error comes from uncertainty in the underlying feature-label distribution (given S_n). Finding the sample-conditioned MSE, $\text{MSE}_\theta[\hat{\varepsilon^B}|S_n]$, of an MMSE error estimator amounts to evaluating the variance of the true error conditioned on the observed sample [29]. $\text{MSE}_\theta[\hat{\varepsilon^B}|S_n] \to 0$ as $n \to \infty$ almost surely in both the discrete and Gaussian models provided in [29, 30], where closed form expressions for the sample-conditioned MSE are available.

The sample-conditioned MSE provides a measure of performance across the uncertainty class Θ for a given sample S_n. As such, it involves various sample-conditioned moments for the error estimator: $E_\theta[\hat{\varepsilon^B}|S_n]$, $E_\theta[(\hat{\varepsilon^B})^2|S_n]$, and $E_\theta[\varepsilon\hat{\varepsilon^B}|S_n]$. One could, on the other hand, consider the MSE relative to a fixed feature-label distribution in the uncertainty class and randomness relative to the sampling distribution. This would yield the feature-label-distribution-conditioned MSE, $\text{MSE}_{S_n}[\hat{\varepsilon^B}|\theta]$, and the corresponding moments: $E_{S_n}[\hat{\varepsilon^B}|\theta]$, $E_{S_n}[(\hat{\varepsilon^B})^2|\theta]$, and $E_{S_n}[\varepsilon\hat{\varepsilon^B}|\theta]$. From a classical point of view, the moments given θ are of interest as they help shed light on the performance of an estimator relative to fixed parameters of class conditional densities. Using this set of moments (i.e. given θ) we are able to compare the performance of the Bayesian MMSE error estimator to classical estimators of true error such as resubstitution and leave-one-out.

From a global perspective, to evaluate performance across both the uncertainty class and the sampling distribution requires the unconditioned MSE, $\text{MSE}_{\theta,S_n}[\hat{\varepsilon^B}]$, and corresponding moments $E_{\theta,S_n}[\hat{\varepsilon^B}]$, $E_{\theta,S_n}[(\hat{\varepsilon^B})^2]$, and $E_{\theta,S_n}[\varepsilon\hat{\varepsilon^B}]$. While both $\text{MSE}_{S_n}[\hat{\varepsilon^B}|\theta]$ and $\text{MSE}_{\theta,S_n}[\hat{\varepsilon^B}]$ have been examined via simulation studies in [27, 28, 30] for discrete and Gaussian models, our intention in the present paper is to derive double-asymptotic representations of the feature-labeled conditioned (given θ) and unconditioned MSE, along with the corresponding moments of the Bayesian MMSE error estimator for linear discriminant analysis (LDA) in the Gaussian model.

We make three modeling assumptions. As in many analytic error analysis studies, we employ stratified sampling: $n = n_0 + n_1$ sample points are selected to constitute the sample S_n in R^p, where given n, n_0 and n_1 are determined, and where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_0}$ and $\mathbf{x}_{n_0+1}, \mathbf{x}_{n_0+2}, \ldots, \mathbf{x}_{n_0+n_1}$ are randomly selected from distributions Π_0 and Π_1, respectively. Π_i possesses a multivariate Gaussian distribution $N(\boldsymbol{\mu}_i, \Sigma)$, for $i = 0, 1$. This means that the prior probabilities α_0 and $\alpha_1 = 1 - \alpha_0$ for classes 0 and 1, respectively, cannot be estimated from the sample (see [31] for a discussion of issues surrounding lack of an estimator for α_0). However, our second assumption is that α_0 and α_1 are known. This is a natural assumption for many medical classification problems. If we desire early or mid-term detection, then we are typically constrained to a small sample for which n_0 and n_1 are not random but for which α_0 and α_1 are known (estimated with extreme accuracy) on account of a large population of post-mortem examinations. The third assumption is that there is a known common covariance matrix for the classes, a common assumption in error analysis [32, 3, 5, 16]. The common covariance assumption is typical for small samples because it is well-known that LDA commonly performs better that quadratic discriminant analysis (QDA) for small samples, even if the actual covariances are different, owing to the estimation advantage of using the pooled sample covariance matrix. As for the assumption of known covariance, this assumption is typical simply owing to the mathematical difficulties of obtaining error representations for unknown covariance (we know of no unknown-covariance result for second-order representations). Indeed, the natural next step following this paper and [16] is to address the unknown covariance problem (although with it being outstanding for almost half a century, it may prove difficult).

Under our assumptions, the *Anderson W statistic* is defined by

$$W(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \mathbf{x}) = \left(\mathbf{x} - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right)^T \sum\nolimits^{-1} (\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1), \quad (2)$$

where $\overline{\mathbf{x}}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbf{x}_i$ and $\overline{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} \mathbf{x}_i$. The corresponding linear discriminant is defined by $\psi_n(\mathbf{x}) = 1$ if $W(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \mathbf{x}) \le c$ and $\psi_n(\mathbf{x}) = 0$ if $W(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \mathbf{x}) > c$, where $c = \log \frac{1-\alpha_0}{\alpha_0}$. Given sample S_n (and thus $\overline{\mathbf{x}}_0$ and $\overline{\mathbf{x}}_1$), for $i = 0, 1$, the error for ψ_n is given by $\varepsilon = \alpha_0 \varepsilon_0 + \alpha_1 \varepsilon_1$, where

$$\varepsilon_i = \Phi\left(\frac{(-1)^{i+1}\left(\boldsymbol{\mu}_i - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right)^T \sum^{-1} (\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1) + (-1)^i c}{\sqrt{(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1)^T \sum^{-1} (\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1)}}\right) \quad (3)$$

and $\Phi(.)$ denotes the cumulative distribution function of a standard normal random variable.

Raudys proposed the following approximation to the expected LDA classification error [19, 24]:

$$E_{S_n}[\varepsilon_0] = P(W(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \mathbf{x}) \le c \mid \mathbf{x} \in \Pi_0) \approx \Phi\left(\frac{-E_{S_n}[W(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0] + c}{\sqrt{\mathrm{Var}_{S_n}[W(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0]}}\right) \quad (4)$$

We provide similar approximations for error-estimation moments and prove asymptotic exactness.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

2. Bayesian MMSE Error Estimator

In the Bayesian classification framework [27, 28], it is assumed that the class-0 an class-1 conditional distributions are parameterized by θ_0 and θ_1, respectively. Therefore, assuming known α_i, the actual feature-label distribution belongs to an uncertainty class parameterized by $\theta = (\theta_0, \theta_1)$ according to a prior distribution, $\pi(\theta)$. Given a sample S_n, the Bayesian MMSE error estimator minimizes the MSE between the true error of a designed classifier, ψ_n, and an error estimate (a function of S_n and ψ_n). The expectation in the MSE is taken over the uncertainty class conditioned on S_n. Specifically, the MMSE error estimator is the expected true error, $\hat{\varepsilon}^B(S_n) = \mathrm{E}_\theta[\varepsilon(\theta)|S_n]$. The expectation given the sample is over the posterior density, $\pi^*(\theta)$. Thus, we write the Bayesian MMSE error estimator as $\hat{\varepsilon}^B = \mathrm{E}_{\pi^*}[\varepsilon]$. To facilitate analytic representations, θ_0 and θ_1 are assumed to be independent prior to observing the data. Denote the marginal priors of θ_0 and θ_1 by $\pi(\theta_0)$ and $\pi(\theta_1)$, respectively, and the corresponding posteriors by $\pi^*(\theta_0)$ and $\pi^*(\theta_1)$, respectively. Independence is preserved, i.e., $\pi^*(\theta_0, \theta_1) = \pi^*(\theta_0)\pi^*(\theta_1)$ for $i = 0, 1$ [27].

Owing to the posterior independence between θ_0 and θ_1 and the fact that α_i is known, the Bayesian MMSE error estimator can be expressed by [27]

$$\hat{\varepsilon}^B = \alpha_0 \mathrm{E}_{\pi^*}[\varepsilon_0] + \alpha_1 \mathrm{E}_{\pi^*}[\varepsilon_1] = \alpha_0 \hat{\varepsilon}_0^B + \alpha_1 \hat{\varepsilon}_1^B, \quad (5)$$

where, letting Θ_i be the parameter space of θ_i,

$$\hat{\varepsilon}_i^B = \mathrm{E}_{\pi^*}[\varepsilon_i] = \int_{\mathbf{\Theta}_i} \varepsilon_i(\theta_i)\pi^*(\theta_i)d\theta_i. \quad (6)$$

For known Σ and the prior distribution on $\boldsymbol{\mu}_i$ assumed to be Gaussian with mean $\mathbf{m}_i$ and covariance matrix Σ/ν_I, $\hat{\varepsilon}_i^B$ is given by equation (10) in [28]:

$$\hat{\varepsilon}_i^B = \Phi\left((-1)^i \frac{-\left(\mathbf{m}_i^* - \frac{\overline{\mathbf{x}}_0 + \overline{\mathbf{x}}_1}{2}\right)^T \sum^{-1} (\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1) + c}{\sqrt{(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1)^T \sum^{-1} (\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_1)}} \sqrt{\frac{\nu_i^*}{\nu_i^* + 1}} \right), \quad (7)$$

where

$$\mathbf{m}_i^* = \frac{n_i \overline{\mathbf{x}}_i + \nu_i \mathbf{m}_i}{n_i + \nu_i}, \qquad \nu_i^* = n_i + \nu_i. \quad (8)$$

and $\nu_i > 0$ is a measure of our certainty concerning the prior knowedge – the larger ν_i is the more localized the prior distribution is about $\mathbf{m}_i$. Letting $\boldsymbol{\mu} = [\boldsymbol{\mu}_0^T, \boldsymbol{\mu}_1^T]^T$, the moments that interest us are of the form $E_{S_n}[\hat{\varepsilon}^B|\boldsymbol{\mu}]$, $E_{S_n}[(\hat{\varepsilon}^B)^2|\boldsymbol{\mu}]$, and $E_{S_n}[\varepsilon\hat{\varepsilon}^B|\boldsymbol{\mu}]$, which are used to obtain $\mathrm{MSE}_{S_n}[\hat{\varepsilon}^B|\boldsymbol{\mu}]$, and $E_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}^B]$, $E_{\boldsymbol{\mu},S_n}[(\hat{\varepsilon}^B)^2]$, and $E_{\boldsymbol{\mu},S_n}[\varepsilon\hat{\varepsilon}^B]$, which are needed to characterize $\mathrm{MSE}_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}^B]$.

3. Bayesian-Kolmogorov Asymptotic Conditions

The Raudys-Kolmogorov asymptotic conditions [16] are defined on a sequence of Gaussian discrimination problems with a sequence of parameters and sample sizes: $(\boldsymbol{\mu}_{p,0}, \boldsymbol{\mu}_{p,1}, \Sigma_p, n_{p,0}, n_{p,1})$, $p = 1, 2, \ldots$, where the means and the covariance matrix are arbitrary. The common assumptions for Raudys-Kolmogorov asymptotics are $n_0 \to \infty$, $n_1 \to \infty$, $p \to \infty$, $\frac{p}{n_0} \to J_0 < \infty$, $\frac{p}{n_1} \to J < \infty$. For notational simplicity, we denote the limit under these conditions by $\lim_{p\to\infty}$. In the analysis of classical statistics related to LDA it is commonly assumed that the Mahalanobis distance, $\delta_{\boldsymbol{\mu},p}=\sqrt{(\boldsymbol{\mu}_{p,0}-\boldsymbol{\mu}_{p,1})^T\sum_p^{-1}(\boldsymbol{\mu}_{p,0}-\boldsymbol{\mu}_{p,1})}$, is finite and $\lim_{p\to\infty}\delta_{\boldsymbol{\mu},p}=\overline{\delta}_{\boldsymbol{\mu}}$ (see [22], p. 4). This condition assures existence of limits of performance metrics of the relevant statistics [16, 22].

To analyze the Bayesian MMSE error estimator, $\widehat{\varepsilon}_i^B$, we modify the sequence of Gaussian discrimination problems to:

$$(\boldsymbol{\mu}_{p,0},\boldsymbol{\mu}_{p,1},\sum\nolimits_p,n_{p,0},n_{p,1},\mathbf{m}_{p,0},\mathbf{m}_{p,1},\nu_{p,0},\nu_{p,1})\,,p=1,2,\ldots \quad (9)$$

In addition to the previous conditions, we assume that the following limits exist for *i*, *j* = 0, 1: $\lim_{p\to\infty}\mathbf{m}_{p,i}^T\sum_p^{-1}\boldsymbol{\mu}_{p,j}=\overline{\mathbf{m}_i^T\sum^{-1}\boldsymbol{\mu}_j}$, $\lim_{p\to\infty}\mathbf{m}_{p,i}^T\sum_p^{-1}\mathbf{m}_{p,j}=\overline{\mathbf{m}_i^T\sum^{-1}\mathbf{m}_j}$, and $\lim_{p\to\infty}\boldsymbol{\mu}_{p,i}^T\sum_p^{-1}\boldsymbol{\mu}_{p,j}=\overline{\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_j}$, where $\overline{\mathbf{m}_i^T\sum^{-1}\boldsymbol{\mu}_j}$, $\overline{\mathbf{m}_i^T\sum^{-1}\mathbf{m}_j}$, and $\overline{\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_j}$ are some constants to which the limits converge. In [22], fairly mild sufficient conditions are given for the existence of these limits.

We refer to all of the aforementioned conditions, along with $\nu_i\to\infty$, $\frac{\nu_i}{n_i}\to\gamma_i<\infty$, as the *Bayesian-Kolmogorov asymptotic conditions* (b.k.a.c). We denote the limit under these conditions by $\lim_{\text{b.k.a.c.}}$, which means that, for *i*, *j* = 0, 1,

$$\lim_{\text{b.k.a.c.}}(\cdot)=\lim_{\substack{p\to\infty,n_i\to\infty,\nu_i\to\infty\\ \frac{p}{n_0}\to J_0,\,\frac{p}{n_1}\to J_1,\,\frac{\nu_0}{n_0}\to\gamma_0,\,\frac{\nu_1}{n_1}\to\gamma_1\\ \gamma_i<\infty,\,J_i<\infty\\ \mathbf{m}_{p,i}^T\sum_p^{-1}\boldsymbol{\mu}_{p,j}=O(1),\ \mathbf{m}_{p,i}^T\sum_p^{-1}\boldsymbol{\mu}_{p,j}\to\overline{\mathbf{m}_i^T\sum^{-1}\boldsymbol{\mu}_j}\\ \mathbf{m}_{p,i}^T\sum_p^{-1}\mathbf{m}_{p,j}=O(1),\ \mathbf{m}_{p,i}^T\sum_p^{-1}\mathbf{m}_{p,j}\to\overline{\mathbf{m}_i^T\sum^{-1}\mathbf{m}_j}\\ \boldsymbol{\mu}_{p,i}^T\sum_p^{-1}\boldsymbol{\mu}_{p,j}=O(1),\ \boldsymbol{\mu}_{p,i}^T\sum_p^{-1}\boldsymbol{\mu}_{p,j}\to\overline{\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_j}}}(\cdot) \quad (10)$$

This limit is defined for the case where there is conditioning on a specific value of $\boldsymbol{\mu}_{p,i}$. Therefore, in this case $\boldsymbol{\mu}_{p,i}$ is not a random variable, and for each *p*, it is a vector of constants. Absent such conditioning, the sequence of discrimination problems and the above limit reduce to

$$(\sum\nolimits_p,n_{p,0},n_{p,1},\mathbf{m}_{p,0},\mathbf{m}_{p,1},\nu_{p,0},\nu_{p,1}),p=1,2,\ldots \quad (11)$$

and

$$\lim_{\text{b.k.a.c.}}(\cdot)=\lim_{\substack{p\to\infty,n_i\to\infty,\nu_i\to\infty\\ \frac{p}{n_0}\to J_0,\,\frac{p}{n_1}\to J_1,\,\frac{\nu_0}{n_0}\to\gamma_0,\,\frac{\nu_1}{n_1}\to\gamma_1\\ \gamma_i<\infty,\,J_i<\infty\\ \mathbf{m}_{p,i}^T\sum_p^{-1}\mathbf{m}_{p,j}=O(1),\ \mathbf{m}_{p,i}^T\sum_p^{-1}\mathbf{m}_{p,j}\to\overline{\mathbf{m}_i^T\sum^{-1}\mathbf{m}_j}}}(\cdot) \quad (12)$$

respectively. For notational simplicity we assume clarity from the context and do not explicitly differentiate between these conditions. We denote convergence in probability under Bayesian-Kolmogorov asymptotic conditions by "$\underset{\text{b.k.a.c.}}{\text{plim}}$"."$\underset{\text{b.k.a.c.}}{\lim}$" and "$\xrightarrow{K}$" denote ordinary convergence under Bayesian-Kolmogorov asymptotic conditions. At no risk of ambiguity, we henceforth omit the subscript "*p*" from the parameters and sample sizes in (9) or (11).

We define $\eta_{\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3,\mathbf{a}_4} = (\mathbf{a}_1 - \mathbf{a}_2)^T \Sigma^{-1}(\mathbf{a}_3 - \mathbf{a}_4)$ and, for ease of notation write $\eta_{\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_1,\mathbf{a}_2}$ as $\eta_{\mathbf{a}_1,\mathbf{a}_2}$. There are two special cases: (1) the square of the Mahalanobis distance in the space of the parameters of the unknown class conditional densities, $\delta^2_{\boldsymbol{\mu}}=\eta_{\boldsymbol{\mu}_0,\boldsymbol{\mu}_1}>0$; and (2) the square of the Mahalanobis distance in the space of prior knowledge, $\Delta^2_{\mathbf{m}}=\eta_{\mathbf{m}_0,\mathbf{m}_1}>0$, where $\mathbf{m}=[\mathbf{m}_0^T,\mathbf{m}_1^T]^T$. The conditions in (10) assure existence of $\lim_{\text{b.k.a.c}} \eta_{\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3,\mathbf{a}_4}$, where the $\mathbf{a}_j$'s can be any combination of $\mathbf{m}_i$ and $\boldsymbol{\mu}_i$, i = 0, 1. Consistent with our notations, we use $\bar{\eta}_{\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3,\mathbf{a}_4}$, $\bar{\delta}^2_{\boldsymbol{\mu}}$, and $\overline{\Delta}^2_{\mathbf{m}}$ to denote the $\lim_{\text{b.k.a.c}}$ of $\eta_{\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3,\mathbf{a}_4}$, $\delta^2_{\boldsymbol{\mu}}$, and $\Delta^2_{\mathbf{m}}$, respectively. Thus,

$$\overline{\eta}_{\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3,\mathbf{a}_4}=\overline{(\mathbf{a}_1-\mathbf{a}_2)^T\sum^{-1}(\mathbf{a}_3-\mathbf{a}_4)}=\overline{\mathbf{a}_1^T\sum^{-1}\mathbf{a}_3}-\overline{\mathbf{a}_1^T\sum^{-1}\mathbf{a}_4}-\overline{\mathbf{a}_2^T\sum^{-1}\mathbf{a}_3}+\overline{\mathbf{a}_2^T\sum^{-1}\mathbf{a}_4}. \quad (13)$$

The ratio p/n_i is an indicator of complexity for LDA (in fact, any linear classification rule): the VC dimension in this case is $p + 1$ [33]. Therefore, the conditions (10) assure the existence of the asymptotic complexity of the problem. The ratio ν_i/n_i is an indicator of relative certainty of prior knowledge to the data: the smaller ν_i/n_i, the more we rely on the data and less on our prior knowledge. Therefore, the conditions (10) state asymptotic existence of relative certainty. In the following, we let $\beta_i=\frac{\nu_i}{n_i}$, so that $\beta_i=\frac{\nu_i}{n_i}\to\gamma_i$.

4. First Moment of $\hat{\varepsilon}_{\mathbf{i}}^{\mathbf{B}}$

In this section we use the Bayesian-Kolmogorov asymptotic conditions to characterize the conditional and unconditional first moment of the Bayesian MMSE error estimator.

4.1. Conditional Expectation of $\hat{\varepsilon}_i^B$:$E_{S_n}[\hat{\varepsilon}_i^B|\boldsymbol{\mu}]$

The asymptotic (in a Bayesian-Kolmogorov sense) conditional expectation of the Bayesian MMSE error estimator is characterized in the following theorem, with the proof presented in the Appendix. Note that G_0^B, G_1^B, and D depend on $\boldsymbol{\mu}$, but to ease the notation we leave this implicit.

Theorem 1—Consider the sequence of Gaussian discrimination problems defined by (9). Then

$$\underset{b.k.a.c.}{\lim} E_{S_n}[\hat{\varepsilon}_i^B|\boldsymbol{\mu}]=\Phi\left((-1)^i\frac{-G_i^B+c}{\sqrt{D}}\right), \quad (14)$$

so that

$$\lim_{b.k.a.c.} E_{S_n}[\hat{\varepsilon}^B|\boldsymbol{\mu}]=\alpha_0\Phi\left(\frac{-G_0^B+c}{\sqrt{D}}\right)+\alpha_1\Phi\left(\frac{G_1^B-c}{\sqrt{D}}\right), \quad (15)$$

where

$$\begin{gathered} G_0^B=\tfrac{1}{2(1+\gamma_0)}\left(\gamma_0(\overline{\eta}_{\mathbf{m}_0,\boldsymbol{\mu}_1}-\overline{\eta}_{\mathbf{m}_0,\boldsymbol{\mu}_0})+\overline{\delta}_{\boldsymbol{\mu}}^2+(1-\gamma_0)J_0+(1+\gamma_0)J_1\right), \\ G_1^B=\tfrac{-1}{2(1+\gamma_1)}\left(\gamma_1(\overline{\eta}_{\mathbf{m}_1,\boldsymbol{\mu}_0}-\overline{\eta}_{\mathbf{m}_1,\boldsymbol{\mu}_1})+\overline{\delta}_{\boldsymbol{\mu}}^2+(1-\gamma_1)J_1+(1+\gamma_1)J_0\right) \\ D=\overline{\delta}_{\mu}^2+J_0+J_1. \end{gathered} \quad (16)$$

Theorem 1 suggests a finite-sample approximation:

$$E_{S_n}[\hat{\varepsilon}_0^B|\boldsymbol{\mu}] \approx \Phi\left(\frac{-G_0^{B,f}+c}{\sqrt{\delta_{\boldsymbol{\mu}}^2+\frac{p}{n_0}+\frac{p}{n_1}}}\right), \quad (17)$$

where $G_0^{B,f}$ is obtained by using the finite-sample parameters of the problem in (16), namely,

$$G_0^{B,f}=\frac{1}{2(1+\beta_0)}\left(\beta_0(\eta_{\mathbf{m}_0,\boldsymbol{\mu}_1}-\eta_{\mathbf{m}_0,\boldsymbol{\mu}_0})+\delta_{\boldsymbol{\mu}}^2+(1-\beta_0)\frac{p}{n_0}+(1+\beta_0)\frac{p}{n_1}\right). \quad (18)$$

To obtain the corresponding approximation for $E_{S_n}[\hat{\varepsilon}_1^B|\boldsymbol{\mu}]$, it suffices to use (17) by exchanging n_0 and n_1, ν_0 and ν_1, $\mathbf{m}_0$ and $\mathbf{m}_1$, and μ_0 and μ_1 in $-G_0^{B,f}$.

To obtain a Raudys-type of finite-sample approximation for the expectation of $\hat{\varepsilon}_0^B$, first note that the Gaussian distribution in (7) can be rewritten as

$$\hat{\varepsilon}_0^B=P(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}) \le c|\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z} \in \Psi_0,\boldsymbol{\mu}), \quad (19)$$

where $\mathbf{z}$ is independent of S_n, Ψ_i is a multivariate Gaussian $N(\mathbf{m}_i,\frac{(n_i+\nu_i+1)(n_i+\nu_i)}{\nu_i^2}\sum)$, and

$$U_i(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})=\left(\frac{\nu_i}{n_i+\nu_i}\mathbf{z}+\frac{n_i\overline{\mathbf{x}}_i}{n_i+\nu_i}-\frac{\overline{\mathbf{x}}_0+\overline{\mathbf{x}}_1}{2}\right)^T\sum^{-1}(\overline{\mathbf{x}}_0-\overline{\mathbf{x}}_1). \quad (20)$$

Taking the expectation of $\hat{\varepsilon}_0^B$ relative to the sampling distribution and then applying the standard normal approximation yields the Raudys-type of approximation:

$$E_{S_n}[\hat{\varepsilon}_0^B|\boldsymbol{\mu}]=P(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}) \le c|\mathbf{z} \in \Psi_0,\boldsymbol{\mu}) \approx \Phi\left(\frac{-E_{S_n,\mathbf{z}}[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z} \in \Psi_0,\boldsymbol{\mu}]+c}{\sqrt{\mathrm{Var}_{S_n,\mathbf{z}}[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z} \in \Psi_0,\boldsymbol{\mu}]}}\right). \quad (21)$$

Algebraic manipulation yields (Suppl. Section A)

$$E_{S_n}\left[\hat{\varepsilon}_0^B|\boldsymbol{\mu}\right]\approx\Phi\left(\frac{-G_0^{B,R}+c}{\sqrt{D_0^{B,R}}}\right),\quad(22)$$

where

$$G_0^{B,R}=G_0^{B,f},\quad(23)$$

with $G_0^{B,f}$ being presented in (18) and

$$\begin{aligned}D_0^{B,R}=&\delta_{\boldsymbol{\mu}}^2+\frac{\delta_{\boldsymbol{\mu}}^2}{n_0(1+\beta_0)}+\frac{\delta_{\boldsymbol{\mu}}^2}{n_1(1+\beta_0)}+\frac{\delta_{\boldsymbol{\mu}}^2}{n_0(1+\beta_0)^2}\\&+\frac{\beta_0}{(1+\beta_0)^2}\left[\frac{n_{\mathbf{m}_0,\boldsymbol{\mu}_1}-(1-\beta_0)\eta_{\mathbf{m}_0,\boldsymbol{\mu}_0}-\delta_{\boldsymbol{\mu}}^2}{n_0}+\frac{(1+\beta_0)\eta_{\mathbf{m}_0,\boldsymbol{\mu}_1}-\eta_{\mathbf{m}_0,\boldsymbol{\mu}_0}}{n_1}\right]+\frac{p}{n_0}+\frac{p}{n_1}+\frac{p}{n_0^2(1+\beta_0)}\\&+\frac{p}{n_0n_1(1+\beta_0)}+\frac{(1-\beta_0)^2p}{2n_0^2(1+\beta_0)^2}+\frac{p}{n_0n_1(1+\beta_0)^2}+\frac{p}{2n_1^2}.\end{aligned}\quad(24)$$

The corresponding approximation for $E_{S_n}\left[\hat{\varepsilon}_1^B|\boldsymbol{\mu}\right]$ is

$$E_{S_n}\left[\hat{\varepsilon}_1^B|\boldsymbol{\mu}\right]\approx\Phi\left(\frac{G_1^{B,R}-c}{\sqrt{D_1^{B,R}}}\right),\quad(25)$$

where $D_1^{B,R}$ and $G_1^{B,R}$ are obtained by exchanging n_0 and n_1, ν_0 and ν_1, $\mathbf{m}_0$ and $\mathbf{m}_1$, and $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ in $D_0^{B,R}$ and $-G_0^{B,R}$, respectively. It is straightforward to see that

$$\begin{aligned}G_0^{B,R}&\xrightarrow{K}G_0^B,\\D_0^{B,R}&\xrightarrow{K}\bar{\delta}_{\boldsymbol{\mu}}^2+J_0+J_1,\end{aligned}\quad(26)$$

with G_0^B being defined in Theorem 1. Therefore, the approximation obtained in (22) is asymptotically exact and (17) and (22) are asymptotically equivalent.

4.2. Unconditional Expectation of $\hat{\varepsilon}_i^B$:$E_{\mu,S_n}\left[\hat{\varepsilon}_i^B\right]$

We consider the unconditional expectation of $\hat{\varepsilon}_i^B$ under Bayesian-Kolmogorov asymptotics. The proof of the following theorem is presented in the Appendix.

Theorem 2—Consider the sequence of Gaussian discrimination problems defined by (11). Then

$$\lim_{b.k.a.c.}E_{\mu,S_n}\left[\hat{\varepsilon}_i^B\right]=\Phi\left((-1)^i\frac{-H_i+c}{\sqrt{F}}\right),\quad(27)$$

so that

$$\lim_{b.k.a.c.} E_{\boldsymbol{\mu},S_n}\left[\hat{\varepsilon}_i^B\right]=\alpha_0\Phi\left(\frac{-H_0+c}{\sqrt{F}}\right)+\alpha_1\Phi\left(\frac{H_1-c}{\sqrt{F}}\right), \quad (28)$$

where

$$\begin{aligned} H_0&=\tfrac{1}{2}\left(\overline{\Delta}_{\mathbf{m}}^2+J_1-J_0+\tfrac{J_0}{\gamma_0}+\tfrac{J_1}{\gamma_1}\right),\\ H_1&=-\tfrac{1}{2}\left(\overline{\Delta}_{\mathbf{m}}^2+J_0-J_1+\tfrac{J_0}{\gamma_0}+\tfrac{J_1}{\gamma_1}\right), \quad (29)\\ F&=\overline{\Delta}_{\mathbf{m}}^2+J_0+J_1+\tfrac{J_0}{\gamma_0}+\tfrac{J_1}{\gamma_1}.\end{aligned}$$

Theorem 2 suggests the finite-sample approximation

$$E_{\boldsymbol{\mu},S_n}\left[\hat{\varepsilon}_0^B\right]\approx\Phi\left(\frac{-H_0^R+c}{\sqrt{\Delta_{\mathbf{m}}^2+\frac{p}{n_0}+\frac{p}{n_1}+\frac{p}{\nu_0}+\frac{p}{\nu_1}}}\right), \quad (30)$$

where

$$H_0^R=\frac{1}{2}\left(\Delta_{\mathbf{m}}^2+\frac{p}{n_1}-\frac{p}{n_0}+\frac{p}{\nu_0}+\frac{p}{\nu_1}\right). \quad (31)$$

From (19) we can get the Raudys-type approximation:

$$E_{\boldsymbol{\mu},S_n}\left[\hat{\varepsilon}_0^B\right]=E_{\boldsymbol{\mu}}\left[P(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\le c|\mathbf{z}\in\Psi_0,\boldsymbol{\mu})\right]\approx\Phi\left(\frac{-E_{\boldsymbol{\mu},S_n,\mathbf{z}}[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|z\in\Psi_0]+c}{\sqrt{\mathrm{Var}_{\boldsymbol{\mu},S_n,\mathbf{z}}[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0]}}\right). \quad (32)$$

Some algebraic manipulations yield (Suppl. Section B)

$$E_{\boldsymbol{\mu},S_n}\left[\hat{\varepsilon}_0^B\right]\approx\Phi\left(\frac{-H_0^R+c}{\sqrt{F_0^R}}\right), \quad (33)$$

where

$$F_0^R=\left(1+\frac{1}{\nu_0}+\frac{1}{\nu_1}+\frac{1}{n_1}\right)\Delta_{\mathbf{m}}^2+p\left(\frac{1}{n_0}+\frac{1}{n_1}+\frac{1}{\nu_0}+\frac{1}{\nu_1}\right)+p\left(\frac{1}{2n_0^2}+\frac{1}{2n_1^2}+\frac{1}{2\nu_0^2}+\frac{1}{2\nu_1^2}\right)+p\left(\frac{1}{n_1\nu_0}+\frac{1}{\eta_1\nu_1}+\frac{1}{\nu_0\nu_1}\right). \quad (34)$$

It is straightforward to see that

$$\begin{aligned} H_0^R&\xrightarrow{K}H_0,\\ F_0^R&\xrightarrow{K}\overline{\Delta}_{\mathbf{m}}^2+J_0+J_1+\tfrac{J_0}{\gamma_0}+\tfrac{J_1}{\gamma_1}, \quad (35)\end{aligned}$$

with H_0 defined in Theorem 2. Hence, the approximation obtained in (33) is asymptotically exact and both (30) and (33) are asymptotically equivalent.

5. Second Moments of $\hat{\varepsilon}_i^B$

Here we employ the Bayesian-Kolmogorov asymptotic analysis to characterize the second and cross moments with the actual error, and therefore the MSE of error estimation.

5.1. Conditional Second and Cross Moments of $\hat{\varepsilon}_i^B$

Defining two i.i.d. random vectors, **z** and **z′**, yields the second moment representation

$$\begin{aligned} E_{S_n}[(\hat{\varepsilon}_0^B)^2|\boldsymbol{\mu}]&=E_{S_n}[P(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\le c|\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}\in\Psi_0,\boldsymbol{\mu})^2]\\ &=E_{S_n}\left[P(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\le c|\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}\in\Psi_0,\boldsymbol{\mu})P(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')\le c|\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}'\in\Psi_0,\boldsymbol{\mu})\right]\\ &=E_{S_n}\left[P\left(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\le c,U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')\le c|\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_0,\boldsymbol{\mu}\right)\right]\\ &=P\left(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\le c,U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')\le c|\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_0,\boldsymbol{\mu}\right), \end{aligned} \qquad (36)$$

where **z** and **z′** are independent of S_n, and Ψ_i is a multivariate Gaussian, $N(\mathbf{m}_i,\frac{(n_i+\nu_i+1)\,(n_i+\nu_i)}{\nu_i^2}\sum)$, and $U_i(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{z})$ being defined in (20).

We have the following theorem, with the proof presented in the Appendix.

Theorem 3—For the sequence of Gaussian discrimination problems in (9) and for $i, j = 0, 1$,

$$\lim_{b.k.a.c.} E_{S_n}[\hat{\varepsilon}_i^B\hat{\varepsilon}_j^B|\boldsymbol{\mu}]=\Phi\left((-1)^i\frac{-G_i^B+c}{\sqrt{D}}\right)\Phi\left((-1)^j\frac{-G_j^B+c}{\sqrt{D}}\right), \qquad (37)$$

so that

$$\lim_{b.k.a.c.} E_{S_n}[(\hat{\varepsilon}^B)^2|\boldsymbol{\mu}]=\left[\alpha_0\Phi\left(\frac{-G_0^B+c}{\sqrt{D}}\right)+\alpha_1\Phi\left(\frac{G_1^B-c}{\sqrt{D}}\right)\right]^2, \qquad (38)$$

where G_0^B, G_1^B, and D are defined in (16).

This theorem suggests the finite-sample approximation

$$E_{S_n}[(\hat{\varepsilon}_0^B)^2|\boldsymbol{\mu}]\approx\left[\Phi\left(\frac{-G_0^{B,f}+c}{\sqrt{\delta_{\boldsymbol{\mu}}^2+\frac{p}{n_0}+\frac{p}{n_1}}}\right)\right]^2, \qquad (39)$$

which is the square of the approximation (17). Corresponding approximations for $E[\hat{\varepsilon}_0^B\hat{\varepsilon}_1^B]$ and $E[(\hat{\varepsilon}_1^B)^2]$ are obtained similarly.

Similar to the proof of Theorem 3, we obtain the conditional cross moment of

$\hat{\varepsilon^B}$.

Theorem 4—Consider the sequence of Gaussian discrimination problems in (9). Then for $i, j = 0, 1$,

$$\lim_{b.k.a.c.} E_{S_n}\left[\hat{\varepsilon}_i^B \varepsilon_j | \boldsymbol{\mu}\right] = \Phi\left((-1)^i \frac{-G_i^B + c}{\sqrt{D}}\right) \Phi\left((-1)^j \frac{-G_j + c}{\sqrt{D}}\right), \quad (40)$$

so that

$$\lim_{b.k.a.c.} E_{S_n}\left[\hat{\varepsilon}^B \varepsilon | \boldsymbol{\mu}\right] = \sum_{i=0}^{1}\sum_{j=0}^{1}\left[\alpha_i \alpha_j \Phi\left((-1)^i \frac{-G_i^B + c}{\sqrt{D}}\right) \Phi\left((-1)^j \frac{-G_j + c}{\sqrt{D}}\right)\right], \quad (41)$$

where G_i^B and D are defined in (16) and G_i is defined in (47).

This theorem suggests the finite-sample approximation

$$E_{S_n}\left[\hat{\varepsilon}_0^B \varepsilon_0 | \boldsymbol{\mu}\right] \approx \Phi\left(\frac{-G_0^{B,f} + c}{\sqrt{\delta_{\boldsymbol{\mu}}^2 + \frac{p}{n_0} + \frac{p}{n_1}}}\right) \Phi\left(-\frac{1}{2}\frac{\delta_{\boldsymbol{\mu}}^2 + \frac{p}{n_1} - \frac{p}{n_0} - c}{\sqrt{\delta_{\boldsymbol{\mu}}^2 + \frac{p}{n_0} + \frac{p}{n_1}}}\right). \quad (42)$$

This is a product of (17) and the finite-sample approximation for $E_{S_n}[\varepsilon_0|\boldsymbol{\mu}]$ in [16].

A consequence of Theorems 1, 3, and 4 is that all the conditional variances and covariances are asymptotically zero:

$$\lim_{\text{b.k.a.c.}} \text{Var}_{S_n}\left(\hat{\varepsilon}^B | \boldsymbol{\mu}\right) = \lim_{\text{b.k.a.c.}} \text{Var}_{S_n}\left(\varepsilon | \boldsymbol{\mu}\right) = \lim_{\text{b.k.a.c.}} \text{Cov}_{S_n}\left(\varepsilon, \hat{\varepsilon}^B | \boldsymbol{\mu}\right) = 0. \quad (43)$$

Hence, the deviation variance is also asymptotically zero, $\lim_{\text{b.k.a.c.}} \text{Var}_{S_n}^d\left[\hat{\varepsilon}^B | \boldsymbol{\mu}\right] = 0$. Hence, defining the conditional bias as

$$\text{Bias}_{C,n}\left[\hat{\varepsilon}^B\right] = E_{S_n}\left[\hat{\varepsilon}^B - \varepsilon | \boldsymbol{\mu}\right], \quad (44)$$

the asymptotic RMS reduces to

$$\lim_{\text{b.k.a.c.}} \text{RMS}_{S_n}\left[\hat{\varepsilon}^B | \boldsymbol{\mu}\right] = \lim_{\text{b.k.a.c.}} \left|\text{Bias}_{C,n}\left[\hat{\varepsilon}^B\right]\right|. \quad (45)$$

To express the conditional bias, as proven in [16],

$$\lim_{\text{b.k.a.c.}} E_{S_n}\left[\varepsilon | \boldsymbol{\mu}\right] = \alpha_0 \Phi\left(\frac{-G_0 + c}{\sqrt{D}}\right) + \alpha_1 \Phi\left(\frac{G_1 - c}{\sqrt{D}}\right), \quad (46)$$

where

$$\begin{gathered} G_0 = \tfrac{1}{2}\left(\delta_{\boldsymbol{\mu}}^2 + J_1 - J_0\right), \\ G_1 = -\tfrac{1}{2}\left(\delta_{\boldsymbol{\mu}}^2 + J_0 - J_1\right), \\ D = \delta_{\boldsymbol{\mu}}^2 + J_0 + J_1. \end{gathered} \quad (47)$$

It follows from Theorem 1 and (46) that

$$\lim_{\text{b.k.a.c.}} \text{Bias}_{C,n}\left[\hat{\varepsilon}^B\right]=\alpha_0\left[\Phi\left(\frac{-G_0^B+c}{\sqrt{D}}\right)-\Phi\left(\frac{-G_0+c}{\sqrt{D}}\right)\right]+\alpha_1\left[\Phi\left(\frac{G_1^B+c}{\sqrt{D}}\right)-\Phi\left(\frac{G_1-c}{\sqrt{D}}\right)\right]. \quad (48)$$

Recall that the MMSE error estimator is unconditionally unbiased: $\text{Bias}_{U,n}[\hat{\varepsilon^B}] = E_{\mu,S_n}[\hat{\varepsilon^B} - \varepsilon] = 0$.

We next obtain Raudys-type approximations corresponding to Theorems 3 and 4 by utilizing the joint distribution of $U_0(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{z})$ and $U_0(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{z}')$, defined in (20), with $\mathbf{z}'$ and $\mathbf{z}'$ being independently selected from populations Ψ_0 or Ψ_1. We employ the function

$$\Phi(a,b;\rho)=\int_{-\infty}^{a}\int_{-\infty}^{b}\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{\frac{-(x^2+y^2-2\rho xy)}{2(1-\rho)^2}\right\}dx\,dy, \quad (49)$$

which is the distribution function of a joint bivariate Gaussian vector with zero means, unit variances, and correlation coefficient ρ. Note that $\Phi(a, \infty; \rho) = \Phi(a)$ and $\Phi(a, b; 0) = \Phi(a)\Phi(b)$. For simplicity of notation, we write $\Phi(a, a; \rho)$ as $\Phi(a; \rho)$. The rectangular-area probabilities involving any jointly Gaussian pair of variables (x, y) can be expressed as

$$P\,(x\le c, y\le d)=\Phi\left(\frac{c-\mu_x}{\sigma_x},\frac{d-\mu_y}{\sigma_y};\rho_{xy}\right), \quad (50)$$

with $\mu_x = E[x]$, $\mu_y = E[y]$, $\sigma_x=\sqrt{\text{Var}(x)}, \sigma_Y=\sqrt{\text{Var}(y)}$, and correlation coefficient ρ_{xy}.

Using (36), we obtain the second-order extension of (21) by

$$\begin{aligned} E_{S_n}\left[(\hat{\varepsilon}_0^B)^2|\boldsymbol{\mu}\right]&=P\left(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\le c, U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')\le c|\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_0,\boldsymbol{\mu}\right)\\ &\approx\Phi\left(\frac{-E_{S_n,\mathbf{z}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0,\boldsymbol{\mu}\right]+c}{\sqrt{\text{Var}_{S_n,\mathbf{z}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0,\boldsymbol{\mu}\right]}};\frac{\text{Cov}_{S_n,\mathbf{z}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}),U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')|\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_0,\boldsymbol{\mu}\right]}{\text{Var}_{S_n,\mathbf{z}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0,\boldsymbol{\mu}\right]}\right). \end{aligned} \quad (51)$$

Using (51), some algebraic manipulations yield

$$E_{S_n}\left[(\hat{\varepsilon}_0^B)^2|\boldsymbol{\mu}\right]\approx\Phi\left(\frac{-G_0^{B,R}+c}{\sqrt{D_0^{B,R}}};\frac{C_0^{B,R}}{D_0^{B,R}}\right), \quad (52)$$

with $G_0^{B,R}$ and $D_0^{B,R}$ being presented in (23) and (24), respectively, and

$$\begin{aligned} C_0^{B,R}&=\text{Cov}_{S_n,\mathbf{z}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}),U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')|\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_0,\boldsymbol{\mu}\right]\\ &=\frac{\beta_0}{(1+\beta_0)^2}\left[\frac{\eta_{\mathbf{m}_0,\boldsymbol{\mu}_1}-(1-\beta_0)\eta_{\mathbf{m}_0,\boldsymbol{\mu}_0}-\delta_{\boldsymbol{\mu}}^2}{n_0}+\frac{(1+\beta_0)\eta_{\mathbf{m}_0,\boldsymbol{\mu}_1}-\eta_{\mathbf{m}_0,\boldsymbol{\mu}_0}}{n_1}\right]+\frac{(1-\beta_0)^2p}{2n_0^2(1+\beta_0)^2}\\ &+\frac{p}{n_0n_1(1+\beta_0)^2}+\frac{p}{2n_1^2}+\frac{\delta_{\boldsymbol{\mu}}^2}{n_1(1+\beta_0)}+\frac{\delta_{\boldsymbol{\mu}}^2}{n_0(1+\beta_0)^2}, \end{aligned} \quad (53)$$

The proof of (53) follows by expanding $U_0(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{z})$ and $U_0(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{z}')$ from (20) and then using the set of identities in the proof of (33), i.e. equation (S.1) from Suppl. Section B. Similarly,

$$E_{S_n}\left[(\hat{\varepsilon}_1^B)^2|\boldsymbol{\mu}\right]=P(U_1(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})>c,U_1(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')>c|\mathbf{z}\in\Psi_1,\mathbf{z}'\in\Psi_1,\boldsymbol{\mu})\approx\Phi\left(\frac{G_1^{B,R}-c}{\sqrt{D_1^{B,R}}};\frac{C_1^{B,R}}{D_1^{B,R}}\right), \quad (54)$$

where $D_1^{B,R}$, $G_1^{B,R}$, and $C_1^{B,R}$ are obtained by exchanging n_0 and n_1, ν_0 and ν_1, $\mathbf{m}_0$ and $\mathbf{m}_1$, and $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, in (24), in $-G_0^{B,f}$ obtained from (18), and in (53), respectively.

Having $C_0^{B,R}\xrightarrow{K}0$ together with (26) shows that (52) is asymptotically exact, that is, asymptotically equivalent to $E_{S_n}\left[(\hat{\varepsilon}_0^B)^2|\boldsymbol{\mu}\right]$ obtained in Theorem 3. Similarly, it can be shown that

$$\begin{aligned}E_{S_n}\left[\hat{\varepsilon}_0^B\hat{\varepsilon}_1^B|\boldsymbol{\mu}\right]=&P(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\le c,-U_1(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')<-c|\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_1,\boldsymbol{\mu})\\&\approx\Phi\left(\frac{-G_0^{B,R}+c}{\sqrt{D_0^{B,R}}},\frac{G_1^{B,R}-c}{\sqrt{D_1^{B,R}}};\frac{C_{01}^{B,R}}{\sqrt{D_0^{B,R}D_1^{B,R}}}\right),\end{aligned} \quad (55)$$

where, after some algebraic manipulations we obtain

$$\begin{aligned}C_{01}^{B,R}=&\frac{1}{n_0(1+\beta_0)(1+\beta_1)}\left[\beta_0\eta_{\mathbf{m}_0,\boldsymbol{\mu}_0,\boldsymbol{\mu}_0,\boldsymbol{\mu}_1}-\beta_0\beta_1\eta_{\mathbf{m}_0,\boldsymbol{\mu}_0,\mathbf{m}_1,\boldsymbol{\mu}_0}+\beta_1\eta_{\mathbf{m}_1,\boldsymbol{\mu}_1,\boldsymbol{\mu}_1,\boldsymbol{\mu}_0}+\beta_1\delta_{\boldsymbol{\mu}}^2+\delta_{\boldsymbol{\mu}}^2\right]\\&+\frac{1}{n_1(1+\beta_0)(1+\beta_1)}\left[\beta_1\eta_{\mathbf{m}_1,\boldsymbol{\mu}_1,\boldsymbol{\mu}_1,\boldsymbol{\mu}_0}-\beta_0\beta_1\eta_{\mathbf{m}_1,\boldsymbol{\mu}_1,\mathbf{m}_0,\boldsymbol{\mu}_1}+\beta_0\eta_{\mathbf{m}_0,\boldsymbol{\mu}_0,\boldsymbol{\mu}_0,\boldsymbol{\mu}_1}+\beta_0\delta_{\boldsymbol{\mu}}^2+\delta_{\boldsymbol{\mu}}^2\right]\\&+\frac{p}{n_0n_1(1+\beta_0)(1+\beta_1)}+\frac{(1-\beta_0)p}{2n_0^2(1+\beta_0)}+\frac{(1-\beta_1)p}{2n_1^2(1+\beta_1)}.\end{aligned} \quad (56)$$

Suppl. Section C gives the proof of (56). Since $C_{01}^{B,R}\xrightarrow{K}0$, (55) is asymptotically exact, i.e. (55) becomes equivalent to the result of Theorem 3. We obtain the conditional cross moment similarly:

$$\begin{aligned}E_{S_n}\left[\hat{\varepsilon}_0^B\varepsilon_0|\boldsymbol{\mu}\right]=&P(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\le c,W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})\le c|\mathbf{z}\in\Psi_0,\mathbf{x}\in\Pi_0,\boldsymbol{\mu})\\&\approx\Phi\left(\frac{-E_{S_n,\mathbf{z}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0,\boldsymbol{\mu}\right]+c}{\sqrt{\mathrm{Var}_{S_n,\mathbf{z}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0,\boldsymbol{\mu}\right]}},\frac{-E_{S_n,\mathbf{x}}\left[W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})|\mathbf{x}\in\Pi_0,\boldsymbol{\mu}\right]+c}{\sqrt{\mathrm{Var}_{S_n,\mathbf{x}}\left[W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})|\mathbf{x}\in\Pi_0,\boldsymbol{\mu}\right]}};\right.\\&\left.\frac{\mathrm{Cov}_{S_n,\mathbf{z},\mathbf{x}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}),W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})|\mathbf{z}\in\Psi_0,\mathbf{x}\in\Pi_0,\boldsymbol{\mu}\right]}{\sqrt{V_{U_0}^CV_W^C}}\right),\end{aligned} \quad (57)$$

where

$$\begin{aligned}V_{U_0}^{\mathrm{C}}=&\mathrm{Var}_{S_n,\mathbf{z}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0,\boldsymbol{\mu}\right],\\V_W^{\mathrm{C}}=&\mathrm{Var}_{S_n,\mathbf{x}}\left[W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})|\mathbf{x}\in\Pi_0,\boldsymbol{\mu}\right],\end{aligned} \quad (58)$$

where superscript “C” denotes conditional variance. Algebraic manipulations like those leading to (53) yield

$$E_{S_n}\left[\hat{\varepsilon}_0^B\varepsilon_0|\boldsymbol{\mu}\right]\approx\Phi\left(\frac{-G_0^{B,R}+c}{\sqrt{D_0^{B,R}}}+\frac{-G_0^R+c}{\sqrt{D_0^R}};\frac{C_0^{BT,R}}{\sqrt{D_0^{B,R}D_0^R}}\right), \quad (59)$$

where

$$C_0^{BT,R}=\frac{1}{n_1\,(1+\beta_0)}\left[\delta_{\boldsymbol{\mu}}^2+\beta_0\delta_{\boldsymbol{\mu}}^2+\beta_0\eta_{\mathbf{m}_0,\boldsymbol{\mu}_0,\boldsymbol{\mu}_0,\boldsymbol{\mu}_1}\right]-\frac{(1-\beta_0)p}{2n_0^2(1+\beta_0)}+\frac{p}{2n_1^2}, \quad (60)$$

and G_0^R and D_0^R having been obtained previously in equations (49) and (50) of [16], namely,

$$\begin{aligned}G_0^R=&E_{S_n,\mathbf{x}}\left[W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})|\mathbf{x}\in\Pi_0,\boldsymbol{\mu}\right]=\frac{1}{2}\left(\delta_{\boldsymbol{\mu}}^2+\frac{p}{n_1}-\frac{p}{n_0}\right),\\ D_0^R=&\mathrm{Var}_{S_n,\mathbf{z}}\left[W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})|\mathbf{x}\in\Pi_0,\boldsymbol{\mu}\right]=\delta_{\boldsymbol{\mu}}^2+\frac{\delta_{\boldsymbol{\mu}}^2}{n_1}+p\left(\frac{1}{n_0}+\frac{1}{n_1}+\frac{1}{2n_0^2}+\frac{1}{2n_1^2}\right).\end{aligned} \quad (61)$$

Similarly, we can show that

$$E_{S_n}\left[\hat{\varepsilon}_1^B\varepsilon_1|\boldsymbol{\mu}\right]\approx\Phi\left(\frac{G_1^{B,R}-c}{\sqrt{D_1^{B,R}}},\frac{G_1^R-c}{\sqrt{D_1^R}};\frac{C_1^{BT,R}}{\sqrt{D_1^{B,R}D_1^R}}\right), \quad (62)$$

where $D_1^{B,R}$ and $G_1^{B,R}$ are obtained as in (54), and D_1^R, G_1^R, and $C_1^{BT,R}$ are obtained by exchanging n_0 and n_1 in D_0^R, $-G_0^R$, and $C_0^{BT,R}$, respectively. Similarly,

$$E_{S_n}\left[\hat{\varepsilon}_0^B\varepsilon_1|\boldsymbol{\mu}\right]\approx\Phi\left(\frac{-G_0^{B,R}+c}{\sqrt{D_0^{B,R}}},\frac{G_1^R-c}{\sqrt{D_1^R}};\frac{C_{01}^{BT,R}}{\sqrt{D_0^{B,R}D_1^R}}\right), \quad (63)$$

where

$$C_{01}^{BT,R}=\frac{1}{n_0(1+\beta_0)}\left[\delta_{\boldsymbol{\mu}}^2+\beta_0\eta_{\mathbf{m}_0,\boldsymbol{\mu}_0,\boldsymbol{\mu}_0,\boldsymbol{\mu}_1}\right]+\frac{(1-\beta_0)p}{2n_0^2(1+\beta_0)}-\frac{p}{2n_1^2}, \quad (64)$$

and

$$E_{S_n}\left[\hat{\varepsilon}_1^B\varepsilon_0|\boldsymbol{\mu}\right]\approx\Phi\left(\frac{G_1^{B,R}-c}{\sqrt{D_1^{B,R}}},\frac{-G_0^R+c}{\sqrt{D_0^R}};\frac{C_{10}^{BT,R}}{\sqrt{D_1^{B,R}D_0^R}}\right), \quad (65)$$

where $C_{10}^{BT,R}$ is obtained by exchanging n_0 and n_1, ν_0 and ν_1, $\mathbf{m}_0$ and $\mathbf{m}_1$, and $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ in $C_{01}^{BT,R}$.

We see that $C_0^{BT,R}\xrightarrow{K}0$, $C_1^{BT,R}\xrightarrow{K}0$, and $C_{01}^{BT,R}\xrightarrow{K}0$. Therefore, from (26) and the fact that $G_0^R\xrightarrow{K}\overline{\delta}_{\boldsymbol{\mu}}^2+J_1-J_0$ and $D_0^R\xrightarrow{K}\overline{\delta}_{\boldsymbol{\mu}}^2+J_0+J_1$, we see that expressions (59), (62), and (63), are all asymptotically exact (compare to Theorem 4).

5.2. Unconditional Second and Cross Moments of $\hat{\varepsilon}_i^B$

Similarly to the way (36) was obtained, we can show that

$$
\begin{aligned}
&E_{\boldsymbol{\mu},S_n}[(\hat{\varepsilon}_0^B)^2]=E_{\boldsymbol{\mu},S_n}[P(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\leq c|\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}\in\Psi_0,\boldsymbol{\mu})^2]\\
=&E_{\boldsymbol{\mu},S_n}\left[P\left(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\leq c,U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')\leq c|\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_0,\boldsymbol{\mu}\right)\right]\\
&=P\left(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\leq c,U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')\leq c|\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_0\right).
\end{aligned}
\quad (66)
$$

Similarly to the proofs of Theorem 3 and 4, we get the following theorems.

Theorem 5—Consider the sequence of Gaussian discrimination problems in (11). For i, j = 0, 1,

$$\lim_{b.k.a.c.} E_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}_i^B\hat{\varepsilon}_j^B]=\Phi\left((-1)^i\frac{-H_i+c}{\sqrt{F}}\right)\Phi\left((-1)^j\frac{-H_j+c}{\sqrt{F}}\right), \quad (67)$$

so that

$$\lim_{b.k.a.c.} E_{\boldsymbol{\mu},S_n}[(\hat{\varepsilon}^B)^2]=\left[\alpha_0\Phi\left(\frac{-H_0+c}{\sqrt{F}}\right)+\alpha_1\Phi\left(\frac{H_1-c}{\sqrt{F}}\right)\right]^2, \quad (68)$$

where H_0, H_1, and F are defined in (29).

Theorem 6—Consider the sequence of Gaussian discrimination problems in (11). For i, j = 0, 1,

$$\lim_{b.k.a.c.} E_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}_i^B\varepsilon_j]=\lim_{b.k.a.c.} E_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}_i^B\hat{\varepsilon}_j^B]=\lim_{b.k.a.c.} E_{\boldsymbol{\mu},S_n}[\varepsilon_i\varepsilon_j], \quad (69)$$

so that

$$\lim_{b.k.a.c.} E_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}^B\varepsilon]=\sum_{i=0}\sum_{j=0}\left[\alpha_i\alpha_j\Phi\left((-1)^i\frac{H_i+c}{\sqrt{F}}\right)\Phi\left((-1)^j\frac{H_j+c}{\sqrt{F}}\right)\right], \quad (70)$$

where H_0, H_1, and F are defined in (29).

Theorems 5 and 6 suggest the finite-sample approximation:

$$E_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}_0^B\hat{\varepsilon}_0^B]\approx E_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}_0^B\varepsilon_0]\approx E_{\boldsymbol{\mu},S_n}[\varepsilon_0\varepsilon_0]\approx\left[\Phi\left(-\frac{1}{2}\frac{\Delta_{\mathbf{m}}^2+\frac{p}{n_1}-\frac{p}{n_0}+\frac{p}{\nu_0}+\frac{p}{\nu_1}-c}{\sqrt{\Delta_{\mathbf{m}}^2+\frac{p}{n_0}+\frac{p}{n_1}+\frac{p}{\nu_0}+\frac{p}{\nu_1}}}\right)\right]^2. \quad (71)$$

A consequence of Theorems 2, 5, and 6 is that

$$
\begin{aligned}
\lim_{\text{b.k.a.c.}} \text{Var}^d_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}^B]=&\lim_{\text{b.k.a.c.}}|\text{Bias}_{U,n}[\hat{\varepsilon}^B]|=\lim_{\text{b.k.a.c.}}\text{Var}_{\boldsymbol{\mu},S_n}(\hat{\varepsilon}^B)=\lim_{\text{b.k.a.c.}}\text{Var}_{\boldsymbol{\mu},S_n}(\varepsilon)\\
&=\lim_{\text{b.k.a.c.}}\text{Cov}_{\mu,S_n}(\varepsilon,\hat{\varepsilon}^B)=\lim_{\text{b.k.a.c.}}\text{RMS}_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}^B]=0.
\end{aligned}
\quad (72)
$$

In [30], it was shown that $\hat{\varepsilon}^B$ is strongly consistent, meaning that $\hat{\varepsilon}^B(S_n) - \varepsilon(S_n) \to 0$ almost surely as $n \to \infty$ under rather general conditions, in particular, for the Gaussian and discrete models considered in that paper. It was also shown that $\text{MSE}_{\mu}[\hat{\varepsilon}^B|S_n] \to 0$ almost surely as n

$\to\infty$ under similar conditions. Here, we have shown that $\mathrm{MSE}_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}^B]\xrightarrow{K}0$ under conditions stated in (12). Some researchers refer to conditions of double asymptoticity as "comparable" dimensionality and sample size [20, 22]. Therefore, one may think of $\mathrm{MSE}_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}^B]\xrightarrow{K}0$ meaning that $\mathrm{MSE}_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}^B]$ is close to zero for asymptotic and comparable dimensionality, sample size, and certainty parameter.

We now consider Raudys-type approximations. Analogous to the approximation used in (51), we obtain the unconditional second moment of $\hat{\varepsilon}_0^B$:

$$E_{\boldsymbol{\mu},S_n}[(\hat{\varepsilon}_0^B)^2]\approx\Phi\left(\frac{-E_{\boldsymbol{\mu},S_n,\mathbf{z}}[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0]+c}{\sqrt{\mathrm{Var}_{\boldsymbol{\mu},S_n,\mathbf{z}}[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0]}};\frac{\mathrm{Cov}_{\boldsymbol{\mu},S_n,\mathbf{z}}[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}),U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')|\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_0]}{\mathrm{Var}_{\boldsymbol{\mu},S_n,\mathbf{z}}[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0]}\right). \quad (73)$$

Using (73) we get

$$E_{\boldsymbol{\mu},S_n}[(\hat{\varepsilon}_0^B)^2]=\Phi\left(\frac{-H_0^R+c}{\sqrt{F_0^R}};\frac{K_0^{B,R}}{F_0^R}\right) \quad (74)$$

with H_0^R and F_0^R given in (31) and (34), respectively, and

$$\begin{aligned}K_0^{B,R}=&\mathrm{Cov}_{\boldsymbol{\mu},S_n,\mathbf{z}}[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}),U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')|\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_0]\\ &=\left(\frac{1}{n_0(1+\beta_0)^2}+\frac{1}{n_1}+\frac{1}{\nu_0(1+\beta_0)^2}+\frac{1}{\nu_1}\right)\Delta_{\mathbf{m}}^2+\frac{p}{2n_0^2}\\ &+\frac{p}{2\nu_0^2}-\frac{p}{n_0\nu_0}+\frac{p}{n_1\nu_1}+\frac{p}{2n_1^2}+\frac{p}{2\nu_1^2}+\frac{p}{n_0n_1(1+\beta_0)^2}+\frac{p}{n_0\nu_1(1+\beta_0)^2}+\frac{p}{n_1\nu_0(1+\beta_0)^2},\end{aligned} \quad (75)$$

Suppl. Section D presents the proof of (75). In a similar way,

$$E_{\boldsymbol{\mu},S_n}[(\hat{\varepsilon}_1^B)^2]=\Phi\left(\frac{H_1^R-c}{\sqrt{F_1^R}};\frac{K_1^{B,R}}{F_1^R}\right), \quad (76)$$

where F_1^R, H_1^R, and $K_1^{B,R}$ are obtained by exchanging n_0 and n_1, ν_0 and ν_1, $\mathbf{m}_0$ and $\mathbf{m}_1$, and $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, in (34), in $-H_0^{B,f}$ obtained from (31), and (75), respectively.

Having $K_0^{B,R}\xrightarrow{K}0$ together with (35) makes (74) asymptotically exact. We similarly obtain

$$E_{\boldsymbol{\mu},S_n}[\hat{\varepsilon}_0^B\hat{\varepsilon}_1^B]=\Phi\left(\frac{-H_0^R+c}{\sqrt{F_0^R}},\frac{H_1^R-c}{\sqrt{F_1^R}};\frac{K_{01}^{B,R}}{\sqrt{F_0^RF_1^R}}\right), \quad (77)$$

where

$$K_{01}^{B,R}=\frac{p}{(n_0+\nu_0)\,(n_1+\nu_1)}+\frac{(n_0-\nu_0)p}{2n_0^2(n_0+\nu_0)}+\frac{(n_1-\nu_1)p}{2n_1^2(n_1+\nu_1)} +\frac{n_0n_1p}{\nu_0\nu_1(n_0+\nu_0)\,(n_1+\nu_1)}+\frac{(n_0-\nu_0)p}{2\nu_0^2(n_0+\nu_0)}+\frac{(n_1-\nu_1)p}{2\nu_1^2(n_1+\nu_1)} +\frac{1}{n_0+\nu_0}\left(1+\frac{n_0}{n_1+\nu_1}-\frac{\nu_0}{n_0}\right)\frac{p}{\nu_0}+\frac{1}{n_1+\nu_1}\left(1+\frac{n_1}{n_0+\nu_0}-\frac{\nu_1}{n_1}\right)\frac{p}{\nu_1}+\left(\frac{1}{\nu_0}+\frac{1}{\nu_1}\right)\Delta_{\mathbf{m}}^2. \quad (78)$$

Suppl. Section E presents the proof of (78). Since $K_{01}^{B,R}\xrightarrow{K}0$, (77) is asymptotically exact (compare to Theorem 5). Similar to (57) and (59), where we characterized conditional cross moments, we can get the unconditional cross moments as follows:

$$E_{\boldsymbol{\mu},S_n}\left[\hat{\varepsilon}_0^B\varepsilon_0\right]=P(U_0\left(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}\right)\le c, W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})\le c|\mathbf{z}\in\Psi_0,\mathbf{x}\in\Pi_0) \approx\Phi\left(\frac{-E_{\boldsymbol{\mu},S_n,\mathbf{z}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0\right]+c}{\sqrt{\mathrm{Var}_{\boldsymbol{\mu},S_n,\mathbf{z}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0\right]}},\frac{-E_{\boldsymbol{\mu},S_n,\mathbf{x}}\left[W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})|\mathbf{x}\in\Pi_0\right]+c}{\sqrt{\mathrm{Var}_{\boldsymbol{\mu},S_n,\mathbf{x}}\left[W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})|\mathbf{x}\in\Pi_0\right]}}; \frac{\mathrm{Cov}_{\boldsymbol{\mu},S_n,\mathbf{z},\mathbf{x}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}),W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})|\mathbf{z}\in\Psi_0,\mathbf{x}\in\Pi_0\right]}{\sqrt{V_{U_0}^U V_W^U}}\right)=\Phi\left(\frac{-H_0^R+c}{\sqrt{F_0^R}};\frac{K_0^{BT,R}}{F_0^R}\right), \quad (79)$$

where

$$V_{U_0}^{\mathrm{U}}=\mathrm{Var}_{\boldsymbol{\mu},S_n,\mathbf{z}}\left[U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})|\mathbf{z}\in\Psi_0\right], \quad V_W^{\mathrm{U}}=\mathrm{Var}_{\boldsymbol{\mu},S_n,\mathbf{x}}[W(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{x})|\mathbf{x}\in\Pi_0], \quad (80)$$

the superscript “U” representing the unconditional variance, H_0^R and F_0^R being presented in (31) and (34), respectively, and

$$K_0^{BT,R}=\left(\frac{n_0}{\nu_0(n_0+\nu_0)}+\frac{1}{n_1}+\frac{1}{\nu_1}\right)\Delta_{\mathbf{m}}^2+\frac{p}{2n_1^2}+\frac{p}{2\nu_1^2}+\frac{p}{n_1\nu_1}+\frac{n_0p}{n_1\nu_0(n_0+\nu_0)}-\frac{(n_0-\nu_0)p}{2n_0^2(n_0+\nu_0)}+\frac{(n_0-\nu_0)p}{2\nu_0^2(n_0+\nu_0)}+\frac{n_0p}{\nu_0\nu_1(n_0+\nu_0)}. \quad (81)$$

The proof of (81) is presented in Suppl. Section F. Similarly,

$$E_{\boldsymbol{\mu},S_n}\left[\hat{\varepsilon}_0^B\varepsilon_1\right]\approx\Phi\left(\frac{-H_0^R+c}{\sqrt{F_0^R}},\frac{H_1^R-c}{\sqrt{F_1^R}};\frac{K_{01}^{BT,R}}{\sqrt{F_0^RF_1^R}}\right) \quad (82)$$

where,

$$K_{01}^{BT,R}=\left(\frac{1}{\nu_0}+\frac{1}{\nu_1}\right)\Delta_{\mathbf{m}}^2+\frac{p}{2\nu_0^2}+\frac{p}{2\nu_1^2}+\frac{p}{\nu_0\nu_1}-\frac{p}{2n_0^2}-\frac{p}{2n_1^2}. \quad (83)$$

See Suppl. Section G for the proof of (83). Having $K_0^{BT,R}\xrightarrow{K}0$ and $K_{01}^{BT,R}\xrightarrow{K}0$ along with (35) makes (79) and (82) asymptotically exact (compare to Theorem 6).

5.3. Conditional and Unconditional Second Moment of ε_i

To complete the derivations and obtain the unconditional RMS of estimation, we need the conditional and unconditional second moment of the true error. The conditional second moment of the true error can be found from results in [16], which for completeness are represented here:

$$E_{S_n}\left[\varepsilon_0^2|\boldsymbol{\mu}\right] \approx \Phi\left(\frac{-G_0^R+c}{\sqrt{D_0^R}};\frac{C_0^{T,R}}{D_0^R}\right), \quad (84)$$

with G_0^R and D_0^R defined in (61),

$$E_{S_n}\left[\varepsilon_1^2|\boldsymbol{\mu}\right] \approx \Phi\left(\frac{G_1^R-c}{\sqrt{D_1^R}};\frac{C_1^{T,R}}{D_1^R}\right), \quad (85)$$

and

$$E_{S_n}\left[\varepsilon_0\varepsilon_1|\boldsymbol{\mu}\right] \approx \Phi\left(\frac{-G_0^R+c}{\sqrt{D_0^R}},\frac{G_1^R-c}{\sqrt{D_1^R}};\frac{C_{01}^{T,R}}{\sqrt{D_0^R D_1^R}}\right), \quad (86)$$

where

$$C_{01}^{T,R}=-\frac{p}{2n_0^2}-\frac{p}{2n_1^2}. \quad (87)$$

Similar to obtaining (79), we can show that

$$E_{\boldsymbol{\mu},S_n}[\varepsilon_0^2] \approx \Phi\left(\frac{-H_0^R+c}{\sqrt{F_0^R}};\frac{K_0^{T,R}}{F_0^R}\right), \quad (88)$$

with H_0^R and F_0^R given in (31) and (34), respectively, and

$$K_0^{T,R}=\left(\frac{1}{\nu_0}+\frac{1}{\nu_1}+\frac{1}{n_1}\right)\Delta_{\mathbf{m}}^2+\frac{p}{2\nu_0^2}+\frac{p}{2\nu_1^2}+\frac{p}{\nu_0\nu_1}+\frac{p}{2n_0^2}+\frac{p}{2n_1^2}+\frac{p}{n_1\nu_0}+\frac{p}{n_1\nu_1}. \quad (89)$$

Similarly,

$$E_{\boldsymbol{\mu},S_n}[\varepsilon_1^2]=\Phi\left(\frac{H_1^R-c}{\sqrt{F_1^R}};\frac{K_1^{T,R}}{F_1^R}\right), \quad (90)$$

with $K_1^{T,R}$ obtained from $K_0^{T,R}$ by exchanging n_0 and n_1, and ν_0 and ν_1. Similarly,

$$E_{\boldsymbol{\mu},S_n}[\varepsilon_0^2] \approx \Phi\left(\frac{-H_0^R+c}{\sqrt{F_0^R}},\frac{H_1^R-c}{\sqrt{F_1^R}};\frac{K_{01}^{T,R}}{\sqrt{F_0^R F_1^R}}\right), \quad (91)$$

with H_0^R and F_0^R given in (31) and (34), respectively, and

$$K_{01}^{T,R}=\left(\frac{1}{\nu_0}+\frac{1}{\nu_1}\right)\Delta_{\mathbf{m}}^2+\frac{p}{2\nu_0^2}+\frac{p}{2\nu_1^2}+\frac{p}{\nu_0\nu_1}-\frac{p}{2n_0^2}-\frac{p}{2n_1^2}. \quad (92)$$

6. Monte Carlo Comparisons

In this section we compare the asymptotically exact finite-sample approximations of the first, second and mixed moments to Monte Carlo estimations in conditional and unconditional scenarios. The following steps are used to compute the Monte Carlo estimation:

1. Define a set of hyper-parameters for the Gaussian model: $\mathbf{m}_0$, $\mathbf{m}_1$, ν_0, , ν_1, and Σ. We let Σ have diagonal elements 1 and off-diagonal elements 0.1. $\mathbf{m}_0$ and $\mathbf{m}_1$ are chosen by fixing δ_{μ}^2 ($\delta_{\mu}^2=4$, which corresponds to Bayes error 0.1586). Setting δ_{μ}^2 and Σ fixes the means μ_0 and μ_1 of the class-conditional densities. The priors, π_0 and π_1, are defined by choosing a small deviation from μ_0 and μ_1, that is, by setting $\mathbf{m}_i = \mu_i + a\mu_i$, where $a = 0.01$.
2. (unconditional case): Using π_0 and π_1, generate random realizations of μ_0 and μ_1.
3. (conditional case): Use the values of μ_0 and μ_1 obtained from Step 1.
4. For fixed Π_0 and Π_1, generate a set of training data of size n_i for class $i = 0, 1$.
5. Using the training sample, design the LDA classifier, ψ_n, using (2).
6. Compute the Bayesian MMSE error estimator, $\hat{\varepsilon}^B$, using (5) and (7).
7. Knowing μ_0 and μ_1, find the true error of ψ_n using (3).
8. Repeat Steps 3 through 6, T_1 times.
9. Repeat Steps 2 through 7, T_2 times.

In the unconditional case, we set $T_1 = T_2 = 300$ and generate 90, 000 samples. For the conditional case, we set $T_1 = 10,000$ and $T_2 = 1$, the latter because μ_0 and μ_1 are set in Step 2.

Figure 1 treats Raudys-type finite-sample approximations, including the RMS. Figure 1(a) compares the first moments obtained from equations (22) and (33). It presents $E_{S_n}[\hat{\varepsilon}^B|\mu]$ and $E_{\mu,S_n}[\hat{\varepsilon}^B]$ computed by Monte Carlo estimation and the analytical expressions. The label "FSA BE Uncond" identifies the curve of $E_{\mu,S_n}[\hat{\varepsilon}^B]$, the unconditional expected estimated error obtained from the finite-sample approximation, which according to the basic theory is equal to $E_{\mu,S_n}[\varepsilon]$. The labels "FSA BE Cond" and "FSA TE Cond" show the curves of $E_{S_n}[\hat{\varepsilon}^B|\mu]$, the conditional expected estimated error, and $E_{S_n}[\varepsilon|\mu]$, the conditional expected true error, respectively, both obtained using the analytic approximations. The curves obtained from Monte Carlo estimation are identified by "MC" labels. The analytic curves in Figure 1(a) show substantial agreement with the Monte Carlo approximation.

To obtain the second moments, $\mathrm{Var}^d[\hat{\varepsilon}]$ and $\mathrm{RMS}[\hat{\varepsilon}^B]$ as defined in (1), we use equations (52), (54), (55), (59), (63), (84), (85), (86) for the conditional case and (74), (76), (77), (79), (82), (88), (90), (91) for the unconditional case. Figures 1(b), 1(c), and 1(d) compare the Monte Carlo estimation to the finite-sample approximations obtained for second/mixed moments, $\mathrm{Var}^d[\hat{\varepsilon}]$, and $\mathrm{RMS}[\hat{\varepsilon}^B]$, respectively. The labels are interpreted similarly to those in Figure 1(a), but for the second/mixed moments instead. For example, "MC BE×TE Uncond" identifies the MC curve of $E_{\mu,S_n}[\hat{\varepsilon}^B\varepsilon]$. The Figures 1(b), 1(c), and 1(d) show that the finite-sample approximations for the conditional and unconditional second/mixed moments, variance of deviation, and RMS are quite accurate (close to the MC value).

While Figure 1 shows the accuracy of Raudys-type of finite-sample approximations, figures in the Supplementary Materials show the the comparison between the finite-sample approximations obtained directly from Theorem 1–6, i.e. equations (29), (57), (70), (73), (76), (102), and (103), to Monte Carlo estimation.

7. Examination of the Raudys-type RMS Approximation

Equations (18), (24), (53), (56), and (63) show that $\mathrm{RMS}_{S_n}[\widehat{\varepsilon^B}|\mu]$ is a function of 14 variables: p, n_0, n_1, β_0, β_1, δ_{μ}^2, $\eta_{\mathbf{m}_0,\mu_1}$, $\eta_{\mathbf{m}_0,\mu_0}$, $\eta_{\mathbf{m}_1,\mu_0}$, $\eta_{\mathbf{m}_1,\mu_1}$, $\eta_{\mathbf{m}_0,\mu_0,\mu_0,\mu_1}$, $\eta_{\mathbf{m}_0,\mu_0,\mathbf{m}_1,\mu_0}$, $\eta_{\mathbf{m}_1,\mu_1,\mathbf{m}_0,\mu_1}$, $\eta_{\mathbf{m}_1,\mu_1,\mu_1,\mu_0}$. Studying a function of this number of variables is complicated, especially because restricting some variables can constrain others. We make several simplifying assumptions to reduce the complexity. We let $n_0=n_1=\frac{n}{2}$, $\beta_0 = \beta_1 = \beta$ and assume very informative priors in which $\mathbf{m}_0 = \mu_0$ and $\mathbf{m}_1 = \mu_1$. Using these assumptions, $\mathrm{RMS}_{S_n}[\widehat{\varepsilon^B}|\mu]$ is only a function of p, n, β, and δ_{μ}^2. We let $p \in [4, 200]$, $n \in [40, 200]$, $\beta \in \{0.5, 1, 2\}$, $\delta_{\mu}^2 \in \{4, 16\}$, which means that the Bayes error is 0.158 or 0.022. Figure 2(a) shows plots of $\mathrm{RMS}_{S_n}[\widehat{\varepsilon^B}|\mu]$ as a function of p, n, β, and δ_{μ}^2. These show that for smaller distance between classes, that is, for smaller δ_{μ}^2 (larger Bayes error), the RMS is larger, and as the distance between classes increases, the RMS decreases. Furthermore, we see that in situations where very informative priors are available, i.e. $\mathbf{m}_0 = \mu_0$ and $\mathbf{m}_1 = \mu_1$, relying more on data can have a detrimental effect on RMS. Indeed, the plots in the top row (for $\beta = 0.5$) have larger RMS than the plots in the bottom row of the figure (for $\beta = 2$).

Using the RMS expressions enables finding the necessary sample size to insure a given $\mathrm{RMS}_{S_n}[\widehat{\varepsilon^B}|\mu]$ by using the same methodology as developed for the resubstitution and leave-one-out error estimators in [16, 26]. The plots in Figure 2(a) (as well as other unshown plots) show that, with $\mathbf{m}_0 = \mu_0$ and $\mathbf{m}_1 = \mu_1$, the RMS is a decreasing function of δ_{μ}^2. Therefore, the number of sample points that guarantees $\max_{\delta_{\mu}^2>0}\mathrm{RMS}_{S_n}[\widehat{\varepsilon}^B|\boldsymbol{\mu}]=\lim_{\delta_{\mu}^2\to 0}\mathrm{RMS}_{S_n}[\widehat{\varepsilon}^B|\boldsymbol{\mu}]$ being less than a predetermined value τ insures that $\mathrm{RMS}_{S_n}[\widehat{\varepsilon^B}|\mu] < \tau$, for any δ_{μ}^2. Let the desired bound be $\kappa_{\widehat{\varepsilon}}(n,p,\beta)=\lim_{\delta_{\mu}^2\to 0}\mathrm{RMS}_{S_n}[\widehat{\varepsilon}^B|\boldsymbol{\mu}]$. From equations (52), (54), (55), (59), (63), (84), (85), and (86), we can find $\kappa_{\hat{\varepsilon}}(n, p, \beta)$ and increase n until $\kappa_{\hat{\varepsilon}}(n, p, \beta) < \tau$. Table 1 ($\beta = 1$: Conditional) shows the minimum number of sample points needed to guarantee having a predetermined conditional RMS for the whole range of δ_{μ}^2 (other β shown in the Supplementary Material). A larger dimensionality, a smaller τ, and a smaller β result in a larger necessary sample size needed for having $\kappa_{\hat{\varepsilon}}(n, p, \beta) < \tau$.

Turning to the unconditional RMS, equations (34), (75), (78), (83), (89), and (92) show that $\mathrm{RMS}_{\mu,S_n}[\widehat{\varepsilon^B}]$ is a function of 6 variables: p, n_0, n_1, ν_0, ν_1, $\Delta_{\mathbf{m}}^2$. Figure 2(b) shows plots of $\mathrm{RMS}_{\mu,S_n}[\widehat{\varepsilon^B}]$ as a function of p, n, β, and $\Delta_{\mathbf{m}}^2$, assuming $n_0=n_1=\frac{n}{2}$, $\beta_0 = \beta_1 = \beta$. Note that setting the values of n and β fixes the value of $\nu_0 = \nu_1 = \nu$ in the corresponding expressions for $\mathrm{RMS}_{\mu,S_n}[\widehat{\varepsilon^B}]$. Due to the complex shape of $\mathrm{RMS}_{\mu,S_n}[\widehat{\varepsilon^B}]$, we consider a large range of n and p. The plots show that a smaller distance between prior distributions (smaller $\Delta_{\mathbf{m}}^2$) corresponds to a larger unconditional RMS of estimation. In addition, as the distance between classes increases, the RMS decreases. The plots in Figure 2(b) show that, as $\Delta_{\mathbf{m}}^2$ increases, RMS decreases. Furthermore, Figure 2(b) (and other unshown plots) demonstrate an interesting phenomenon in the shape of the RMS. In regions defined by pairs of (p, n), for each p, RMS first increases as a function of sample size and then decreases. We further observe that with fixed p, for smaller β, this "peaking phenomenon" happens for larger n. On the other hand, with fixed β, for larger p, peaking happens for larger n. These observations are presented in Figure 3, which shows curves obtained by cutting the 3D plots in the left column of Fig. 2(b) at a few dimensions. This figure shows that, for p = 900 and β = 2, adding more sample points increases RMS abruptly at first to reach a maximum value of RMS at n = 140, the point after which the RMS starts to decrease.

One may use the unconditional scenario to determine the the minimum necessary sample size for a desired $\mathrm{RMS}_{\mu,S_n}[\hat{\varepsilon}^B]$. In fact, this is the more practical way to go because in practice one does not know μ. Since the unconditional RMS shows a decreasing trend in terms of $\Delta_{\mathbf{m}}^2$, we use the previous technique to find the minimum necessary sample size to guarantee a desired unconditional RMS. Table 1 (β = 1: Unconditional) shows the minimum sample size that guarantees $\max_{\Delta_{\mathbf{m}}^2>0}\mathrm{RMS}_{\mu,S_n}[\hat{\varepsilon}^B]=\lim_{\Delta_{\mathbf{m}}^2\to 0}\mathrm{RMS}_{\mu,S_n}[\hat{\varepsilon}^B]$ being less than a predetermined value τ, i.e. insures that $\mathrm{RMS}_{\mu,S_n}[\hat{\varepsilon}^B] < \tau$ for any $\Delta_{\mathbf{m}}^2$ (other β shown in the Supplementary Material).

To examine the accuracy of the required sample size that satisfies $\kappa_{\varepsilon}(n, p, \beta) < \tau$ for both conditional and unconditional settings, we have performed a set of experiments (see Supplementary Material). The results of these experiments confirm the efficacy of Table 1 in determining the minimum sample size required to insure the RMS is less than a predetermined value τ.

8. Conclusion

Using realistic assumptions about sample size and dimensionality, standard statistical techniques are generally incapable of estimating the error of a classifier in small-sample classification. Bayesian MMSE error estimation facilitates more accurate estimation by incorporating prior knowledge. In this paper, we have characterized two sets of performance metrics for Bayesian MMSE error estimation in the case of LDA in a Gaussian model: (1) the first, second, and cross moments of the estimated and actual errors conditioned on a fixed feature-label distribution, which in turn gives us knowledge of the conditional $\mathrm{RMS}_{S_n}[\hat{\varepsilon}^B|\theta]$; and (2) the unconditional moments and, therefore, the unconditional RMS, $\mathrm{RMS}_{\theta,S_n}[\hat{\varepsilon}^B]$. We set up a series of conditions, called the Bayesian-Kolmogorov asymptotic conditions, that allow us to characterize the performance metrics of Bayesian MMSE error estimation in an asymptotic sense. The Bayesian-Kolmogorov asymptotic conditions are set up based on the assumption of increasing n, p, and certainty parameter ν, with an arbitrary constant limiting ratio between n and p, and n and ν. To our knowledge, these conditions permit, for the first time, application of Kolmogorov-type of asymptotics in a Bayesian setting. The asymptotic expressions proposed in this paper result directly in finite-sample approximations of the performance metrics. Improved finite-sample accuracy is achieved via newly proposed Raudys-type approximations. The asymptotic theory is used to prove that these approximations are, in fact, asymptotically exact under the Bayesian-Kolmogorov asymptotic conditions. Using the derived analytical expressions, we have examined performance of the Bayesian MMSE error estimator in relation to feature-label distributions, prior knowledge, sample size, and dimensionality. We have used the results to determine the minimum sample size guaranteeing a desired level of error estimation accuracy.

As noted in the Introduction, a natural next step in error estimation theory is to remove the known-covariance condition, but as also noted, this may prove to be difficult.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by the NIH grants 2R25CA090301 (Nutrition, Biostatistics, and Bioinformatics) from the National Cancer Institute.

References

1. Hills M. Allocation rules and their error rates. J Royal Statist Soc Ser B (Methodological). 1966; 28:1–31.
2. Foley D. Considerations of sample and feature size. IEEE Trans Inf Theory. 1972; IT-18:618–626.
3. Sorum MJ. Estimating the conditional probability of misclassification. Technometrics. 1971; 13:333–343.
4. McLachlan GJ. An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis. Australian Journal of Statistics. 1973; 15:210–214.
5. Moran M. On the expectation of errors of allocation associated with a linear discriminant function. Biometrika. 1975; 62:141–148.
6. Berikov VB. A priori estimates of recognition quality for discrete features. Pattern Recogn and Image Analysis. 2002; 12:235–242.
7. Berikov VB, Litvinenko AG. The influence of prior knowledge on the expected performance of a classifier. Pattern Recogn Lett. 2003; 24:2537–2548.
8. Braga-Neto U, Dougherty ER. Exact performance measures and distributions of error estimators for discrete classifiers. Pattern Recogn. 2005; 38:1799–1814.
9. Glick N. Additive estimators for probabilities of correct classification. Pattern Recogn. 1978; 10:211–222.
10. Fukunaga K, Hayes RR. Estimation of classifier performance. IEEE Trans Pattern Anal Mach Intell. 1989; 11:1087–1101.
11. Raudys, S. An Integrated Approach to Design. Springer-Verlag; London: 2001. Statistical and Neural Classifiers.
12. Zollanvari A, Braga-Neto UM, Dougherty ER. On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers. Pattern Recogn. 2009; 42:2705–2723.
13. Braga-Neto UM, Dougherty ER. Exact correlation between actual and estimated errors in discrete classification. Pattern Recogn Lett. 2010; 31:407–413.
14. Zollanvari A, Braga-Neto UM, Dougherty ER. Exact representation of the second-order moments for resubstitution and leave-one-out error estimation for linear discriminant analysis in the univariate heteroskedastic gaussian model. Pattern Recogn. 2012; 45:908–917.
15. Zollanvari A, Braga-Neto UM, Dougherty ER. Joint sampling distribution between actual and estimated classification errors for linear discriminant analysis. IEEE Trans Inf Theory. 2010; 56:784–804.
16. Zollanvari A, Braga-Neto UM, Dougherty ER. Analytic study of performance of error estimators for linear discriminant analysis. IEEE Trans Sig Proc. 2011; 59:4238–4255.
17. Wyman F, Young D, Turner D. A comparison of asymptotic error rate expansions for the sample linear discriminant function. Pattern Recogn. 1990; 23:775–783.
18. Pikelis V. Comparison of methods of computing the expected classification errors. Automat Remote Control. 1976; 5:59–63.
19. Raudys S. On the amount of a priori information in designing the classification algorithm. Technical Cybernetics. 1972; 4:168–174. In Russian.
20. Deev A. Representation of statistics of discriminant analysis and asymptotic expansion when space dimensions are comparable with sample size. Dokl Akad Nauk SSSR. 1970; 195:759–762. In Russian.
21. Fujikoshi Y. Error bounds for asymptotic approximations of the linear discriminant function when the sample size and dimensionality are large. J Multivariate Anal. 2000; 73:1–17.
22. Serdobolskii, V. Multivariate Statistical Analysis: A High-Dimensional Approach. Springer; 2000.

23. Bickel PJ, Levina E. Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. Bernoulli. 2004; 10:989–1010.
24. Raudys S, Young DM. Results in statistical discriminant analysis: A review of the former soviet union literature. J Multivariate Anal. 2004; 89:1–35.
25. Dougherty E, Sima C, Hua J, Hanczar B, Braga-Neto U. Performance of error estimators for classification. Current Bioinformatics. 2010; 5:53–67.
26. Dougherty ER, Zollanvari A, Braga-Neto UM. The illusion of distribution-free small-sample classification in genomics. Current Genomics. 2011; 12:333–341. [PubMed: 22294876]
27. Dalton L, Dougherty ER. Bayesian minimum mean-square error estimation for classification error–Part I: Definition and the Bayesian MMSE error estimator for discrete classification. IEEE Trans Sig Proc. 2011; 59:115–129.
28. Dalton L, Dougherty ER. Bayesian minimum mean-square error estimation for classification error–Part II: Linear classification of Gaussian models. IEEE Trans Sig Proc. 2011; 59:130–144.
29. Dalton L, Dougherty ER. Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error–Part I: Representation. IEEE Trans Sig Proc. 2012; 60:2575–2587.
30. Dalton L, Dougherty ER. Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error–Part II: Consistency and performance analysis. IEEE Trans Sig Proc. 2012; 60:2588–2603.
31. Anderson T. Classification by multivariate analysis. Psychometrika. 1951; 16:31–50.
32. John S. Errors in discrimination. Ann Math Statist. 1961; 32:1125–1144.
33. Devroye, L.; Gyorfi, L.; Lugosi, G. A Probabilistic Theory of Pattern Recognition. Springer; New York: 1996.
34. Kan R. From moments of sum to moments of product. J Multivariate Anal. 2008; 99:542– 554.
35. Bao Y, Ullah A. Expectation of quadratic forms in normal and nonnormal variables with econometric applications. J Statist Plann Inference. 2010; 140:1193–1205.

Appendix

Proof of Theorem 1

We explain this proof in detail as some steps will be used in later proofs. Let

$$\widehat{G}_0^B=\left(\mathbf{m}_0^*-\frac{\overline{\mathbf{x}}_0+\overline{\mathbf{x}}_1}{2}\right)^T\sum\nolimits^{-1}(\overline{\mathbf{x}}_0-\overline{\mathbf{x}}_1), \quad (93)$$

where $\mathbf{m}_0^*$ is defined in (8). Then

$$\widehat{G}_0^B=\frac{\nu_0\mathbf{m}_0^T}{n_0+\nu_0}\sum\nolimits^{-1}(\overline{\mathbf{x}}_0-\overline{\mathbf{x}}_1)+\frac{n_0-\nu_0}{2(n_0+\nu_0)}(\overline{\mathbf{x}}_0^T\sum\nolimits^{-1}\overline{\mathbf{x}}_0-\overline{\mathbf{x}}_0^T\sum\nolimits^{-1}\overline{\mathbf{x}}_1)+\frac{1}{2}(\overline{\mathbf{x}}_1^T\sum\nolimits^{-1}\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_0^T\sum\nolimits^{-1}\overline{\mathbf{x}}_1). \quad (94)$$

For $i, j = 0, 1$ and $i \neq j$, define the following random variables:

$$y_i=\mathbf{m}_i^T\sum\nolimits^{-1}(\overline{\mathbf{x}}_0-\overline{\mathbf{x}}_1), \quad z_i=\overline{\mathbf{x}}_i^T\sum\nolimits^{-1}\overline{\mathbf{x}}_i, \quad z_{ij}=\overline{\mathbf{x}}_i^T\sum\nolimits^{-1}\overline{\mathbf{x}}_j. \quad (95)$$

The variance of y_i given $\boldsymbol{\mu}$ does not depend on $\boldsymbol{\mu}$. Therefore, under the Bayesian-Kolmogorov conditions stated in (10), $\overline{\mathbf{m}_i^T\sum^{-1}\boldsymbol{\mu}_j}$ and $\overline{\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_j}$ do not appear in the limit. Only $\overline{\mathbf{m}_i^T\sum^{-1}\mathbf{m}_i}$ matters, which vanishes in the limit as follows:

$$\mathrm{Var}_{S_n}[y_i|\boldsymbol{\mu}]=\mathbf{m}_i^T(\frac{\sum^{-1}}{n_0}+\frac{\sum^{-1}}{n_1})\mathbf{m}_i \xrightarrow{K} \lim_{n_0\to\infty}\frac{\overline{\mathbf{m}_i^T\sum^{-1}\mathbf{m}_i}}{n_0}+\lim_{n_1\to\infty}\frac{\overline{\mathbf{m}_i^T\sum^{-1}\mathbf{m}_i}}{n_1}=0. \quad (96)$$

To find the variance of z_i and z_{ij} we can first transform z_i and z_{ij} to quadratic forms and then use the results of [34] to find the variance of quadratic functions of Gaussian random variables. Specifically, from [34], for $\mathbf{y} \sim N(\mu, \Sigma)$ and $\mathbf{A}$ being a symmetric positive definite matrix, $\mathrm{Var}[\mathbf{y}^T\mathbf{A}\mathbf{y}] = 2tr(\mathbf{A}\Sigma)^2 + 4\mu^T\mathbf{A}\Sigma\mathbf{A}\mu'$, with tr being the trace operator. Therefore, after some algebraic manipulations, we obtain

$$\begin{aligned}\mathrm{Var}_{S_n}[z_i|\boldsymbol{\mu}]=2\tfrac{p}{n_i^2}+4\tfrac{\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_i}{n_i} \xrightarrow{K} 2\lim_{n_i\to\infty}\tfrac{J_i}{n_i}+4\lim_{n_i\to\infty}\tfrac{\overline{\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_i}}{n_i}=0,\\ \mathrm{Var}_{S_n}[z_{ij}|\boldsymbol{\mu}]=\tfrac{p}{n_in_j}+\tfrac{\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_i}{n_j}+\tfrac{\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_j}{n_i} \xrightarrow{K} \lim_{n_j\to\infty}\tfrac{J_i}{n_j}+\lim_{n_j\to\infty}\tfrac{\overline{\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_i}}{n_j}+\lim_{n_i\to\infty}\tfrac{\overline{\boldsymbol{\mu}_j^T\sum^{-1}\boldsymbol{\mu}_j}}{n_i}=0.\end{aligned} \quad (97)$$

From the Cauchy-Schwarz inequality $(\mathrm{Cov}[x,y] \le \sqrt{\mathrm{Var}[x]\mathrm{Var}[y]})$, $\mathrm{Cov}_{S_n}[y_i,z_k|\boldsymbol{\mu}] \xrightarrow{K} 0$, $\mathrm{Cov}_{S_n}[y_i,z_{ij}|\boldsymbol{\mu}] \xrightarrow{K} 0$, and $\mathrm{Cov}_{S_n}[z_i,z_{ij}|\boldsymbol{\mu}] \xrightarrow{K} 0$ for $i, j, k = 0, 1$, $i \neq j$, Furthermore, $\frac{n_i-\nu_i}{2(n_i+\nu_i)} \xrightarrow{K} \frac{1-\gamma_i}{2(1+\gamma_i)}$ and $\frac{\nu_i}{n_i+\nu_i} \xrightarrow{K} \frac{\gamma_i}{1+\gamma_i}$. Putting this together and following the same approach for $\widehat{G}_1^B$ yields $\mathrm{Var}_{S_n}[\widehat{G}_1^B|\boldsymbol{\mu}] \xrightarrow{K} 0$. In general (via Chebyshev's inequality), $\lim_{n\to\infty}\mathrm{Var}[X_n] = 0$ implies convergence in probability of X_n to $\lim_{n\to\infty} E[X_n]$. Hence, since $\mathrm{Var}_{S_n}[\widehat{G}_1^B|\boldsymbol{\mu}] \xrightarrow{K} 0$, for $i, j = 0, 1$ and $i \neq j$,

$$\begin{aligned}\underset{\text{b.k.a.c.}}{\mathrm{plim}}\ \widehat{G}_i^B|\boldsymbol{\mu}=\lim_{\text{b.k.a.c.}} E_{S_n}[\widehat{G}_i^B|\boldsymbol{\mu}]=(-1)^i\Big[\tfrac{1}{2}\Big(\overline{\boldsymbol{\mu}_j^T\sum^{-1}\boldsymbol{\mu}_j}+J_j\Big)+\\ \tfrac{\gamma_i(\overline{\mathbf{m}_i^T\sum^{-1}\boldsymbol{\mu}_i}-\overline{\mathbf{m}_i^T\sum^{-1}\boldsymbol{\mu}_j})}{1+\gamma_i}+\tfrac{1-\gamma_i}{2(1+\gamma_i)}\Big(\overline{\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_i}+J_i\Big)-\overline{\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_j}\Big(\tfrac{1-\gamma_i}{2(1+\gamma_i)}+\tfrac{1}{2}\Big)\Big]=G_i^B.\end{aligned} \quad (98)$$

Now let

$$\widehat{D}_i=\frac{\nu_i^*+1}{\nu_i^*}(\overline{\mathbf{x}}_0-\overline{\mathbf{x}}_1)^T\sum^{-1}(\overline{\mathbf{x}}_0-\overline{\mathbf{x}}_1)=\frac{\nu_i^*+1}{\nu_i^*}\widehat{\delta}^2, \quad (99)$$

where $\widehat{\delta}^2 = (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)^T\Sigma^{-1}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)$. Similar to deriving (97) via the variance of quadratic forms of Gaussian variables, we can show

$$\mathrm{Var}_{S_n}[\widehat{\delta}^2|\boldsymbol{\mu}]=4\delta_\mu^2\left(\frac{1}{n_0}+\frac{1}{n_1}\right)+2p\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^2. \quad (100)$$

Thus,

$$\mathrm{Var}_{S_n}[\widehat{D}_i|\boldsymbol{\mu}]=\left(\frac{\nu_i^*+1}{\nu_i^*}\right)^2\mathrm{Var}_{S_n}[\widehat{\delta}^2|\boldsymbol{\mu}] \xrightarrow{K} 0. \quad (101)$$

As before, from Chebyshev's inequality it follows that

$$\underset{\text{b.k.a.c.}}{\text{plim}}\ \widehat{D}_i|\boldsymbol{\mu}=\underset{\text{b.k.a.c.}}{\lim}\ E_{S_n}[\widehat{D}_i|\boldsymbol{\mu}]=D. \quad (102)$$

By the Continuous Mapping Theorem (continuous functions preserve convergence in probability),

$$\underset{\text{b.k.a.c.}}{\text{plim}}\ \widehat{\varepsilon}_i^B|\boldsymbol{\mu}=\underset{\text{b.k.a.c.}}{\text{plim}}\ \Phi\left((-1)^i\frac{-\widehat{G}_i^B+c}{\sqrt{\widehat{D}_i}}\right)|\boldsymbol{\mu}=\Phi\left(\underset{\text{b.k.a.c.}}{\text{plim}}\ (-1)^i\frac{-\widehat{G}_i^B+c}{\sqrt{\widehat{D}_i}}|\boldsymbol{\mu}\right)=\Phi\left((-1)^i\frac{-G_i^B+c}{\sqrt{D}}\right). \quad (103)$$

The Dominated Convergence Theorem ($|X_n| \le B$, for some $B > 0$ and $X_n \to X$ in probability implies $E[X_n] \to E[X]$)), via the boundedness of φ(.) leads to completion of the proof:

$$\begin{aligned}\underset{\text{b.k.a.c.}}{\lim}\ E_{S_n}[\widehat{\varepsilon}_i^B|\boldsymbol{\mu}]&=\underset{\text{b.k.a.c.}}{\lim}\ E_{S_n}\left[\Phi\left((-1)^i\frac{-\widehat{G}_i^B+c}{\sqrt{\widehat{D}_i}}\right)|\mu\right]\\ &=E_{S_n}\left[\Phi\left(\underset{\text{b.k.a.c.}}{\text{plim}}\ (-1)^i\frac{-\widehat{G}_i^B+c}{\sqrt{\widehat{D}_i}}|\boldsymbol{\mu}\right)\right]=\underset{\text{b.k.a.c.}}{\text{plim}}\ \widehat{\varepsilon}_i^B|\boldsymbol{\mu}.\end{aligned} \quad (104)$$

Proof of Theorem 2

We first prove that $\text{Var}_{S_n}(\widehat{G}_0^B)\xrightarrow{K}0$ with $\widehat{G}_0^B$ defined in (94). To do so we use

$$\text{Var}_{\boldsymbol{\mu},S_n}[\widehat{G}_0^B]=\text{Var}_{\boldsymbol{\mu}}\left[E_{S_n}[\widehat{G}_0^B|\boldsymbol{\mu}]\right]+E_{\boldsymbol{\mu}}\left[\text{Var}_{S_n}[\widehat{G}_0^B|\boldsymbol{\mu}]\right]. \quad (105)$$

To compute the first term on the right hand side, we have

$$E_{S_n}[\widehat{G}_0^B|\boldsymbol{\mu}]=\frac{\nu_0\mathbf{m}_0^T}{n_0+\nu_0}\sum^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)+\frac{n_0-\nu_0}{2(n_0+\nu_0)}\left(\boldsymbol{\mu}_0^T\sum^{-1}\boldsymbol{\mu}_0+\frac{p}{n_0}\right)+\frac{1}{2}\left(\boldsymbol{\mu}_1^T\sum^{-1}\boldsymbol{\mu}_1+\frac{p}{n_1}\right)-\boldsymbol{\mu}_0^T\sum^{-1}\boldsymbol{\mu}_1\left(\frac{n_0}{n_0+\nu_0}\right). \quad (106)$$

For $i, j = 0, 1$ and $i \ne j$ define the following random variables:

$$y_i'=\mathbf{m}_i^T\sum^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1),\quad z_i'=\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_i,\quad z_{ij}'=\boldsymbol{\mu}_i^T\sum^{-1}\boldsymbol{\mu}_j. \quad (107)$$

The variables defined in (107) can be obtained by replacing $\bar{\mathbf{x}}_i$'s with $\boldsymbol{\mu}_i$'s in (95) and $\bar{\mathbf{x}}_i \sim N(\boldsymbol{\mu}_i, \Sigma/n_i)$ and $\boldsymbol{\mu}_i \sim N(\mathbf{m}_i, \Sigma/\nu_i)$. Replacing $\boldsymbol{\mu}_i$ with $\mathbf{m}_i$ and n_i with ν_i in (96) and (97) yields

$$\begin{aligned}&\text{Var}_{\boldsymbol{\mu}}(y_i')=\mathbf{m}_i^T\left(\frac{\sum^{-1}}{\nu_0}+\frac{\sum^{-1}}{\nu_1}\right)\mathbf{m}_i\xrightarrow{K}0,\quad \text{Var}_{\boldsymbol{\mu}}(z_i')=2\frac{p}{\nu_i^2}+4\frac{\mathbf{m}_i^T\sum^{-1}\mathbf{m}_i}{\nu_i}\xrightarrow{K}0,\\ &\text{Var}_{\boldsymbol{\mu}}(z_{ij}')=\frac{p}{\nu_i\nu_j}+\frac{\mathbf{m}_i^T\sum^{-1}\mathbf{m}_i}{\nu_j}+\frac{\mathbf{m}_j^T\sum^{-1}\mathbf{m}_j}{\nu_i}\xrightarrow{K}0.\end{aligned} \quad (108)$$

By Cauchy-Schwarz, $\text{Cov}_{\boldsymbol{\mu}}(y_i', z_k')\xrightarrow{K}0$, $\text{Cov}_{\boldsymbol{\mu}}(y_i', z_{ij}')\xrightarrow{K}0$, and $\text{Cov}_{\boldsymbol{\mu}}(z_i', z_{ij}')\xrightarrow{K}0$. Hence, $\text{Var}_{\boldsymbol{\mu}}\left[E_{S_n}[\widehat{G}_0^B|\boldsymbol{\mu}]\right]\xrightarrow{K}0$

Now consider the second term on the right hand side of (105). The covariance of a function of Gaussian random variables can be computed from results of [35]. For instance,

$$\mathrm{Cov}_{S_n}[\mathbf{a}^T\overline{\mathbf{x}}_i,\overline{\mathbf{x}}_i^T{\sum}^{-1}\overline{\mathbf{x}}_i|\boldsymbol{\mu}]=\frac{2}{n_i}\mathbf{a}^T\boldsymbol{\mu}_i. \quad (109)$$

From (109) and the independence of $\bar{\mathbf{x}}_0$ and $\bar{\mathbf{x}}_1$,

$$\mathrm{Cov}_{S_n}[\overline{\mathbf{x}}_i^T{\sum}^{-1}\overline{\mathbf{x}}_i,\overline{\mathbf{x}}_i^T{\sum}^{-1}\overline{\mathbf{x}}_i|\boldsymbol{\mu}]=\frac{2}{n_i}\mu_j^T{\sum}^{-1}\boldsymbol{\mu}_i,\ \ i\neq j \quad (110)$$

Via (108), (109), and (110), the inner variance in the second term on the right hand side of (105) is

$$\begin{array}{c}\mathrm{Var}_{S_n}[\widehat{G}_0^B|\boldsymbol{\mu}]=\frac{(n_0-\nu_0)^2p}{2n_0^2(n_0+\nu_0)^2}=\frac{n_0p}{n_1(n_0+\nu_0)^2}+\frac{p}{2n_1^2}\\ +\left(\frac{(n_0-\nu_0)^2}{n_0(n_0+\nu_0)^2}+\frac{n_0^2}{n_1(n_0+\nu_0)^2}\right)\boldsymbol{\mu}_0^T{\sum}^{-1}\boldsymbol{\mu}_0-2\left(\frac{n_0-\nu_0}{(n_0+\nu_0)^2}+\frac{n_0}{n_1(n_0+\nu_0)}\right)\boldsymbol{\mu}_0^T{\sum}^{-1}\boldsymbol{\mu}_1\\ +2\left(\frac{\nu_0(n_0-\nu_0)}{n_0(n_0+\nu_0)^2}+\frac{\nu_0n_0}{n_1(n_0+\nu_0)^2}\right)\mathbf{m}_0^T{\sum}^{-1}\boldsymbol{\mu}_0-2\left(\frac{\nu_0}{(n_0+\nu_0)^2}+\frac{\nu_0}{n_1(n_0+\nu_0)}\right)\mathbf{m}_0^T{\sum}^{-1}\boldsymbol{\mu}_1+\\ \left(\frac{n_0}{(n_0+\nu_0)^2}+\frac{1}{n_1}\right)\boldsymbol{\mu}_1^T{\sum}^{-1}\boldsymbol{\mu}_1+\frac{\nu_0^2}{(n_0+\nu_0)^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\mathbf{m}_0^T{\sum}^{-1}\mathbf{m}_0.\end{array} \quad (111)$$

Now, again from the results of [35],

$$\begin{array}{c}E_{\boldsymbol{\mu}}[\boldsymbol{\mu}_i^T{\sum}^{-1}\boldsymbol{\mu}_i]=\mathbf{m}_i^T{\sum}^{-1}\mathbf{m}_i+\frac{p}{\nu_i},\\ E_{\boldsymbol{\mu}}[\boldsymbol{\mu}_i^T{\sum}^{-1}\boldsymbol{\mu}_j]=\mathbf{m}_i^T{\sum}^{-1}\mathbf{m}_j, i\neq j.\end{array} \quad (112)$$

From (111) and (112), some algebraic manipulations yield

$$\begin{array}{c}E_{\boldsymbol{\mu}}\left[\mathrm{Var}_{S_n}[\widehat{G}_0^B|\boldsymbol{\mu}]\right]=\frac{(n_0-\nu_0)^2p}{2n_0^2(n_0+\nu_0)^2}+\frac{n_0p}{n_1(n_0+\nu_0)^2}\\ +\frac{p}{2n_1^2}+\left(\frac{(n_0-\nu_0)^2}{n_0(n_0+\nu_0)^2}+\frac{n_0^2}{n_1(n_0+\nu_0)^2}\right)\frac{p}{\nu_0}+\left(\frac{n_0}{(n_0+\nu_0)^2}+\frac{1}{n_1}\right)\frac{p}{\nu_1}+\left(\frac{n_0}{(n_0+\nu_0)^2}+\frac{1}{n_1}\right)\Delta_{\mathbf{m}}^2.\end{array} \quad (113)$$

From (113) we see that $E_{\boldsymbol{\mu}}[\mathrm{Var}_{S_n}[\widehat{G}_0^B|\boldsymbol{\mu}]] \xrightarrow{K} 0$. In sum, $Var_{\boldsymbol{\mu},S_n}[\widehat{G}_0^B] \xrightarrow{K} 0$ and similar to the use Chebyshev's inequality in the proof of Theorem 1, we get

$$\underset{\text{b.k.a.c.}}{\mathrm{plim}}\ \widehat{G}_i^B=\lim_{\text{b.k.a.c.}}E_{\boldsymbol{\mu},S_n}[\widehat{G}_i^B]\triangleq H_i, \quad (114)$$

with H_i defined in (29).

On the other hand, for $\hat{D_i}$ defined in (99) we can write

$$\mathrm{Var}_{\boldsymbol{\mu},S_n}[\widehat{D}_i]=\mathrm{Var}_{\boldsymbol{\mu}}\left[E_{S_n}[\widehat{D}_i|\boldsymbol{\mu}]\right]+E_{\boldsymbol{\mu}}\left[\mathrm{Var}_{S_n}[\widehat{D}_i|\boldsymbol{\mu}]\right]. \quad (115)$$

From similar expressions as in (112) for $\overline{\mathbf{x}}_i^T{\sum}^{-1}\overline{\mathbf{x}}_j$, we get $E_{S_n}[\widehat{\delta}^2]=\delta_{\boldsymbol{\mu}}^2+\frac{p}{n_0}+\frac{p}{n_1}$. Moreover, $\mathrm{Var}_{\boldsymbol{\mu}}[\delta_{\boldsymbol{\mu}}^2]$ is obtained from (100) by replacing n_i with ν_i, and $\delta_{\boldsymbol{\mu}}^2$ with $\Delta_{\mathbf{m}}^2$. Thus, from (99),

$$Var_{\boldsymbol{\mu}}\left[E_{S_n}[\widehat{D}_i|\boldsymbol{\mu}]\right]=\left(\frac{\nu_i^*+1}{\nu_i^*}\right)^2\left[4\Delta_{\mathbf{m}}^2\left(\frac{1}{\nu_0}+\frac{1}{\nu_1}\right)+2p\left(\frac{1}{\nu_0}+\frac{1}{\nu_0}\right)^2\right]\xrightarrow{K}0. \quad (116)$$

Furthermore, since $E_{\boldsymbol{\mu}}[\delta_{\boldsymbol{\mu}}^2]=\Delta_{\mathbf{m}}^2+\frac{p}{\nu_0}+\frac{p}{\nu_1}$, from (101),

$$E_{\boldsymbol{\mu}}\left[\mathrm{Var}_{S_n}[\widehat{D}_i|\boldsymbol{\mu}]\right]=\left(\frac{\nu_i^*+1}{\nu_i^*}\right)^2\left[4(\Delta_{\mathbf{m}}^2+\frac{p}{\nu_0}+\frac{p}{\nu_1})\left(\frac{1}{n_0}+\frac{1}{n_1}\right)+2p\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^2\right]\xrightarrow{K}0. \quad (117)$$

Hence, $Var_{\boldsymbol{\mu},S_n}[\widehat{D}_i]\xrightarrow{K}0$ and, similar to (114), we obtain

$$\underset{\text{b.k.a.c.}}{\text{plim}}\,\widehat{D}_i=\lim_{\text{b.k.a.c.}}E_{\boldsymbol{\mu},S_n}[\widehat{D}_0]=\lim_{\text{b.k.a.c.}}E_{\boldsymbol{\mu},S_n}[\widehat{D}_1]\triangleq F, \quad (118)$$

with F defined in (29). Similar to the proof of Theorem 1, by using the Continuous Mapping Theorem and the Dominated Convergence Theorem, we can show that

$$\begin{aligned}&\lim_{\text{b.k.a.c.}}E_{\boldsymbol{\mu},S_n}\left[\widehat{\varepsilon}_i^B\right]=\lim_{\text{b.k.a.c.}}E_{\boldsymbol{\mu},S_n}\left[\Phi\left((-1)^i\,\frac{-\widehat{G}_i^B+c}{\sqrt{\widehat{D}}}\right)\right]\\&=E_{\boldsymbol{\mu},S_n}\left[\Phi\left(\underset{\text{b.k.a.c.}}{\text{plim}}\,(-1)^{i+1}\frac{-\widehat{G}_i^B+c}{\sqrt{\widehat{D}}}\right)\right]=\Phi\left((-1)^i\,\frac{-H_i+c}{\sqrt{F}}\right),\end{aligned} \quad (119)$$

and the result follows.

Proof of Theorem 3

We start from

$$E_{S_n}[(\widehat{\varepsilon}_0^B)^2|\boldsymbol{\mu}]=E_{S_n}\left[P\left(U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\le c,U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')\le c|\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_0,\boldsymbol{\mu}|\right)\right], \quad (120)$$

which was shown in (36). Here we characterize the conditional probability inside $E_{S_n}[.]$. From the independence of $\mathbf{z}$, $\mathbf{z}'$, $\overline{\mathbf{x}}_0$, and $\overline{\mathbf{x}}_1$,

$$\left[\begin{array}{c}U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z})\\U_0(\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}')\end{array}\right]|\overline{\mathbf{x}}_0,\overline{\mathbf{x}}_1,\mathbf{z}\in\Psi_0,\mathbf{z}'\in\Psi_0,\boldsymbol{\mu}\sim N\left(\left[\begin{array}{c}\widehat{G}_0^B\\\widehat{G}_0^B\end{array}\right],\left[\begin{array}{cc}\widehat{D}&0\\0&\widehat{D}\end{array}\right]\right), \quad (121)$$

where here $N(.,.)$ denotes the bivariate Gaussian density function and $\widehat{G}_0^B$ and $\widehat{D}$ are defined in (94) and (99). Thus,

$$E_{S_n}[(\widehat{\varepsilon}_0^B)^2|\boldsymbol{\mu}]=\left[\Phi\left(\frac{-\widehat{G}_0^B+c}{\sqrt{\widehat{D}}}\right)\right]^2|\boldsymbol{\mu}. \quad (122)$$

Similar to the proof of Theorem 1, we get

$$\lim_{\text{b.k.a.c.}} E_{S_n}[(\hat{\varepsilon}_i^B)^2|\boldsymbol{\mu}] = \operatorname*{plim}_{\text{b.k.a.c.}} (\hat{\varepsilon}_i^B)^2|\boldsymbol{\mu} = \left(\lim_{\text{b.k.a.c.}} E_{S_n}[\hat{\varepsilon}_i^B|\boldsymbol{\mu}]\right)^2 = \left[\Phi\left((-1)^i\frac{-G_i^B+c}{\sqrt{D}}\right)\right]^2. \quad (123)$$

Similarly, we obtain $\lim_{b.k.a.c.} E[\hat{\varepsilon}_0^B \hat{\varepsilon}_1^B] = \lim_{b.k.a.c.} \hat{\varepsilon}_0^B \varepsilon_1^B$, and the results follow.

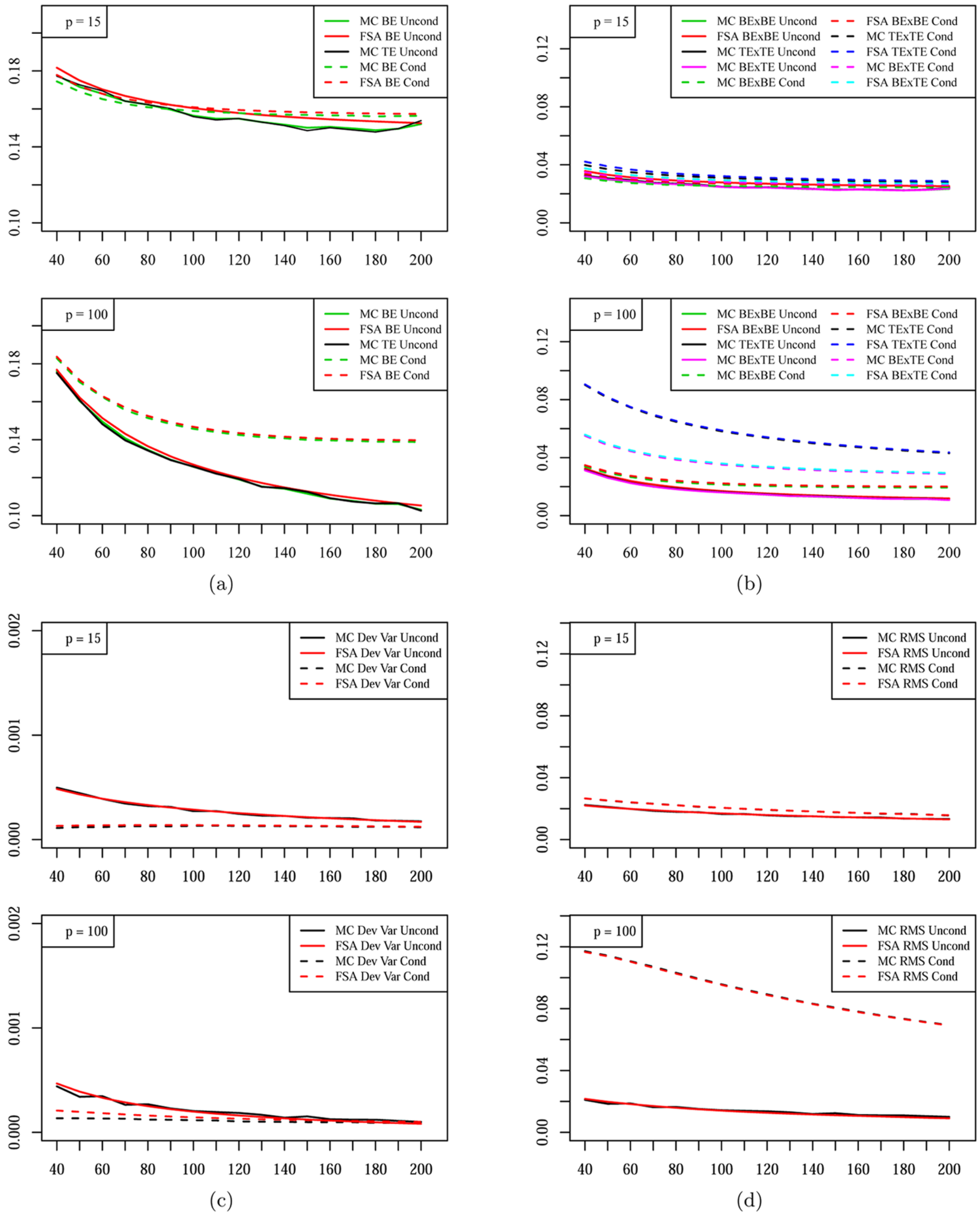

Figure 1.

Comparison of conditional and unconditional performance metrics of $\hat{\varepsilon}^B$ using asymptotically exact finite setting approximations, with Monte Carlo estimates as a function of sample size. (a) Expectations. The case of asymptotic unconditional expectation of ε is not plotted as $\hat{\varepsilon}^B$ is unconditionally unbiased; (b) Second and mixed moments; (c) Conditional variance of deviation from true error, i.e. $\mathrm{Var}^d_{S_n}[\hat{\varepsilon}^B|\boldsymbol{\mu}]$ and, unconditional variance of deviation, i.e. $\mathrm{Var}^d_{\mu,S_n}[\hat{\varepsilon}^B]$; (d) Conditional RMS of estimation, i.e. $\mathrm{RMS}_{S_n}[\hat{\varepsilon}^B|\boldsymbol{\mu}]$ and, unconditional RMS of estimation, i.e. $\mathrm{RMS}_{\mu,S_n}[\hat{\varepsilon}^B]$; (a)–(d) correspond to the same scenario in which dimension, p, is 15 and 100, $\nu_0 = \nu_1 = 50$, $\mathbf{m}_i = \mu_i + 0.01\mu_i$ with $\mu_0 = -\mu_1$, and Bayes error = 0.1586.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript
NIH-PA Author Manuscript

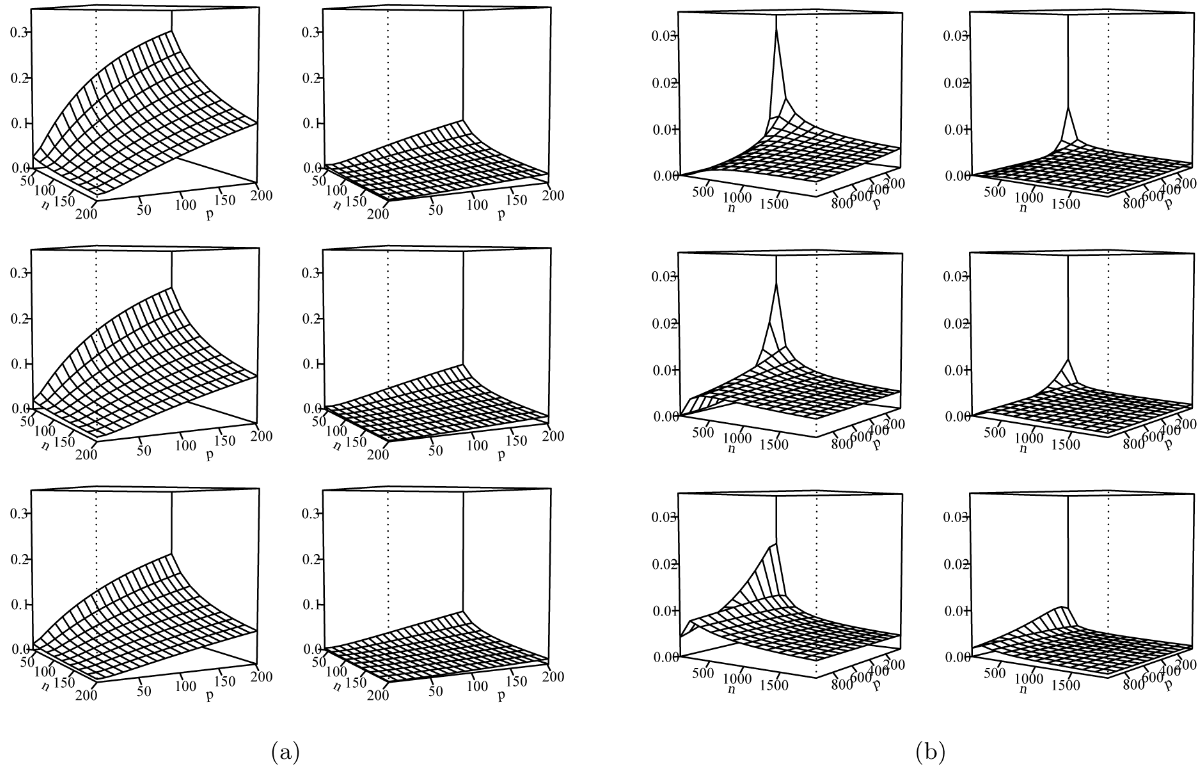

Figure 2.

(a) The conditional RMS of estimation, i.e. $\mathrm{RMS}_{S_n}[\hat{\varepsilon^{B}}|\mu]$, as a function of $p < 200$ and $n < 200$. From top to bottom, the rows correspond to $\beta = 0.5, 1, 2$, respectively. From left to right, the columns correspond to $\delta_{\mu}^{2}=4, 16$, respectively. (b) The unconditional RMS of estimation, i.e. $\mathrm{RMS}_{\mu,S_n}[\hat{\varepsilon^{B}}]$, as a function of $p < 1000$ and $n < 2000$. From top to bottom, the rows correspond to $\beta = 0.5, 1, 2$, respectively. From left to right, the columns correspond to $\Delta_{\mathbf{m}}^{2}=4, 16$, respectively.

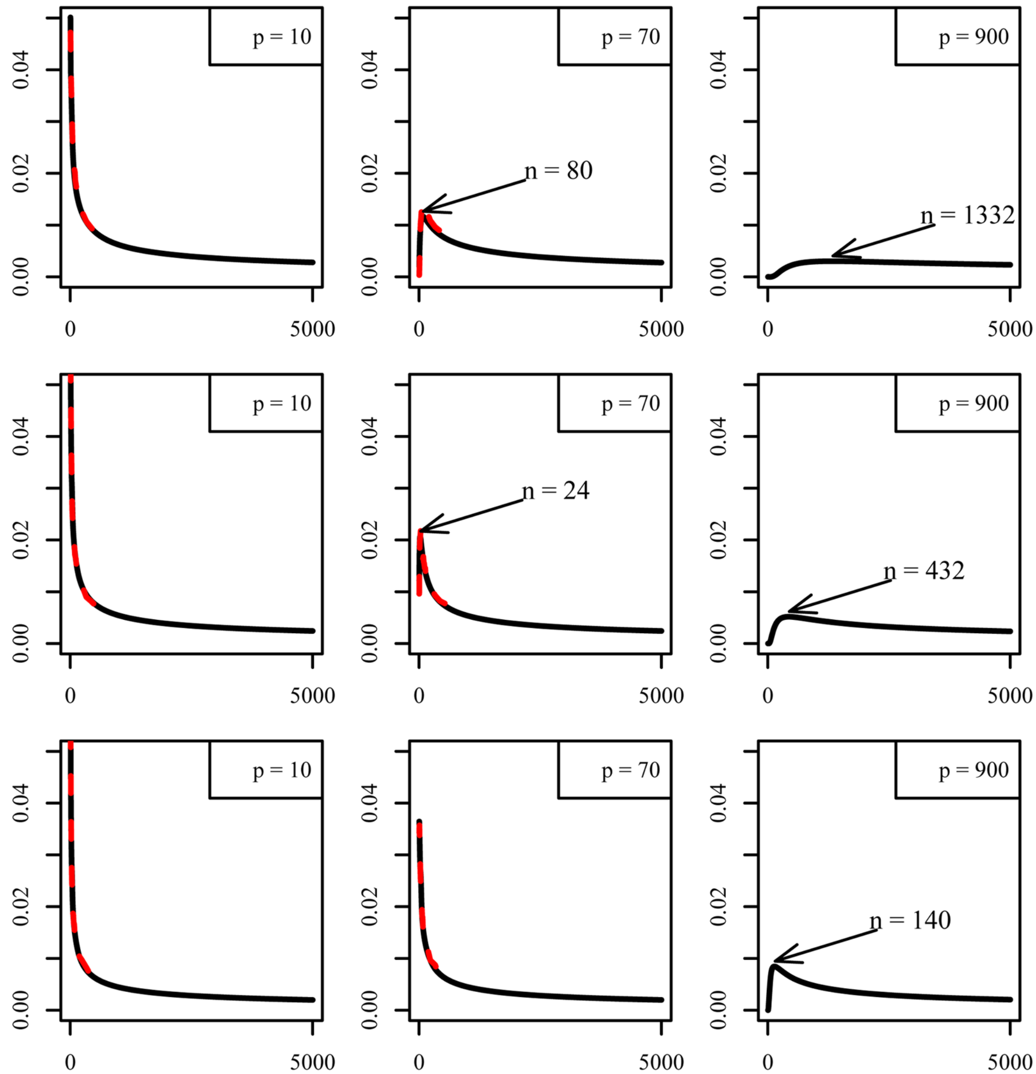

Figure 3.
$\text{RMS}_{\mu,S_n}[\hat{\varepsilon^B}]$-peaking phenomenon as a function of sample size. These plots are obtained by cutting the 3D plots in the left column of Fig. 2(b) at few dimensionality (i.e. $\Delta_{\mathbf{m}}^2=4$). From top to bottom the rows correspond to $\beta = 0.5, 1, 2$, respectively. The solid-black curves indicate $\text{RMS}_{\mu,S_n}[\hat{\varepsilon^B}]$ computed from the analytical results and the red-dashed curves show the same results computed by means of Monte Carlo simulations. Due to computational burden of estimating the curves by means of Monte Carlo studies, the simulations are limited to $n < 500$ and $p = 10, 70$.

Table 1

Minimum sample size, n, ($n_0=n_1=\frac{n}{2}$) to satisfy $\kappa_\varepsilon(n, p, \beta) < \tau$.

τ	p = 2	p = 4	p = 8	p = 16	p = 32	p = 64	p = 128
$\beta = 1$: **Conditional**							
0.1	14	22	38	70	132	256	506
0.09	18	28	48	86	164	318	626
0.08	24	36	60	110	208	404	796
0.07	32	48	80	144	272	530	1044
0.06	44	64	108	196	372	722	1424
0.05	62	94	158	284	538	1044	2056
$\beta = 1$: **Unconditional**							
0.025	108	108	106	102	92	72	2
0.02	172	170	168	164	156	138	78
0.015	308	306	304	300	292	274	236
0.01	694	694	690	686	678	662	628
0.005	2790	2786	2782	2776	2768	2752	2720