This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier’s archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/authorsrights
Moments and root-mean-square error of the Bayesian MMSE estimator of classification error in the Gaussian model

Amin Zollanvari a,b,⁎, Edward R. Dougherty a,c
a Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, United States
b Department of Statistics, Texas A&M University, College Station, TX 77843, United States
c Translational Genomics Research Institute (TGEN), Phoenix, AZ 85004, United States

1. Introduction

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.

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

Bias(ε) = E[ε] − E[ˆε],

Var[ε] = Var(ε − ˆε) = Var(ε) + Var(ˆε) − 2 Cov(ε, ˆε),

RMS(ε) = \sqrt{E(ε − ˆε)^2} = \sqrt{E[ε^2] + E[\hat{\epsilon}^2] − 2E[\hat{\epsilon}]E[\epsilon]} = \sqrt{\text{Bias}(\hat{\epsilon})^2 + \text{Var}(\hat{\epsilon})},

(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

⁎ Corresponding author at: Department of Electrical and and Computer Engineering, Texas A&M University, College Station, TX 77843, United States.
E-mail addresses: amin_zoll@neo.tamu.edu (A. Zollanvari), edward@ece.tamu.edu (E.R. Dougherty).
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 been shown to be quite accurate \([16–18]\). There is quite a body of work on the use of double asymptotics for the analysis of LDA and its related statistics \([16,19–23]\); 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 \([26–28]\). 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 \(\Theta\) \([27]\). This results in a Bayesian approach with \(\Theta\) 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, \(\hat{\Theta}\), 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, \(\text{MSE}_{\text{cond}}[\hat{\Theta}\mid S_n]\), of an MMSE error estimator amounts to evaluating the variance of the true error conditioned on the observed sample \([29]\). \(\text{MSE}_{\text{cond}}[\hat{\Theta}\mid S_n] \to 0\) 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 \(\Theta\) for a given sample \(S_n\). As such, it involves various sample-conditioned moments for the error estimator: \(E_{\Theta}[\hat{\epsilon}^2(S_n)], E_{\Theta}[\hat{\epsilon}^2(S_n)^2],\) and \(E_{\Theta}[\hat{\epsilon}^2(S_n)^2]\). 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, \(\text{MSE}_{\Theta}[\hat{\epsilon}^2\mid \Theta]\), and the corresponding moments: \(E_{\Theta}[\hat{\epsilon}^2\mid \Theta], E_{\Theta}[\hat{\epsilon}^2\mid \Theta]^2,\) and \(E_{\Theta}[\hat{\epsilon}^2\mid \Theta]^2\). From a classical point of view, the moments given \(\Theta\) 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 \(\Theta\)) 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 unconditional MSE, \(\text{MSE}_{\Theta\Theta}[\hat{\epsilon}^2]\), and corresponding moments \(E_{\Theta\Theta}[\hat{\epsilon}^2], E_{\Theta\Theta}[\hat{\epsilon}^2]^2\), and \(E_{\Theta\Theta}[\hat{\epsilon}^2]^2\). While both MSE\(_{\Theta\Theta}[\hat{\epsilon}^2\mid \Theta]\) and MSE\(_{\Theta\Theta}[\hat{\epsilon}^2]\) 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 \(\Theta\)) and unconditional MSE, along with the corresponding moments of the Bayesian MMSE error estimator for near 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 \(n_0\) and \(n_1\) are determined, and where \(x_1, x_2, \ldots, x_{n_0}\) and \(x_{n_0 + 1}, x_{n_0 + 2}, \ldots, x_{n_0 + n_1}\) are randomly selected from distributions \(P_0\) and \(P_1\), respectively. \(P_i\) possesses a multivariate Gaussian distribution \(N(\mu_i, \Sigma)\), for \(i = 0,1\). This means that the prior probabilities \(\alpha_0\) 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 \(\alpha_0\)). However, our second assumption is that \(\alpha_0\) and \(\alpha_1\) 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 \(\alpha_0\) and \(\alpha_1\) 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

\[
W(\bar{x}_0, \bar{x}_1, x) = \left( \frac{x - \frac{\bar{x}_0 + \bar{x}_1}{2}}{\bar{x}_0 - \bar{x}_1} \right)^\top \Sigma^{-1}(\bar{x}_0 - \bar{x}_1).
\]  

(2)

where \(\bar{x}_0 = (1/n_0) \sum_{i=0}^{n_0} x_i\) and \(\bar{x}_1 = (1/n_1) \sum_{i=n_0+1}^{n_0+n_1} x_i\). The corresponding linear discriminant is defined by \(\psi(x) = 1\) if \(W(\bar{x}_0, \bar{x}_1, x) \leq c\) and \(\psi(x) = 0\) if \(W(\bar{x}_0, \bar{x}_1, x) > c\), where \(c = \log(1 – \alpha_0)/\alpha_0\). Given sample \(S_n\) (and thus \(\bar{x}_0\) and \(\bar{x}_1\)), for \(i = 0,1\), the
error for \( \psi_n \) is given by 
\[
\epsilon_i = \Phi \left( \frac{(-1)^{i+1} \left( \frac{\mu_i - \mathbf{x}_0 + \mathbf{x}_1}{2} \right)^T \Sigma^{-1} (\mathbf{x}_0 - \mathbf{x}_1) + c \right) \sqrt{\mathbf{(x}_0 - \mathbf{x}_1)^T \Sigma^{-1} (\mathbf{x}_0 - \mathbf{x}_1)}
\]
and \( \Phi(\cdot) \) denotes the cumulative distribution function of a standard normal random variable.

Raudys proposed the following approximation to the expected LDA classification error [19,24]:
\[
E_{S_0}[\epsilon_0] = P(W=0|x_0, x_1, x) \leq c \in \Pi_0 = \Phi \left( \frac{-E_{S_0}[W|x_0, x_1, x] \in \Pi_0] + c}{\sqrt{\text{Var}_{S_0}[W|x_0, x_1, x] \in \Pi_0}} \right)
\]
We provide similar approximations for error-estimation moments and prove asymptotic exactness.

2. Bayesian MMSE error estimator

In the Bayesian classification framework [27,28], it is assumed that the class-0 or class-1 conditional distributions are parameterized by \( \theta_0 \) and \( \theta_1 \), respectively. Therefore, assuming known \( \alpha \), 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 \( S_0 \), the Bayesian MMSE error estimator minimizes the MSE between the true error of a designed classifier, \( \psi_n \), and an error estimate (a function of \( S_0 \) and \( \psi_n \)). The expectation in the MSE is taken over the uncertainty class conditioned on \( S_0 \). Specifically, the MMSE error estimator is the expected true error, \( \bar{\psi} = \mathbb{E}[\psi|S_0] \). The expectation given the sample is over the posterior density, \( \pi(\theta) \). Thus, we use the Bayesian MMSE error estimator as \( \bar{\psi} = \mathbb{E}_\theta[\psi] \). To facilitate analytic representations, \( \theta_0 \) and \( \theta_1 \) are assumed to be independent prior to observing the data. Denote the marginal priors of \( \theta_0 \) and \( \theta_1 \) by \( \pi(\theta_0) \) and \( \pi(\theta_1) \), respectively, and the corresponding posteriors by \( \pi^\star(\theta_0) \) and \( \pi^\star(\theta_1) \), respectively. Independence is preserved, i.e. \( \pi^\star(\theta_0, \theta_1) = \pi^\star(\theta_0) \pi^\star(\theta_1) \) for \( i = 0, 1 \), \[27\].

Owing to the posterior independence between \( \theta_0 \) and \( \theta_1 \) and the fact that \( \alpha \) is known, the Bayesian MMSE error estimator can be expressed by \[27\]:
\[
\bar{\psi} = \alpha_0 E_{\theta_0}[\psi_0] + \alpha_1 E_{\theta_1}[\psi_1] = \alpha_0 \bar{\psi}_0 + \alpha_1 \bar{\psi}_1,
\]
where, letting \( \Theta \) be the parameter space of \( \theta_i \),
\[
\bar{\psi}_i = \mathbb{E}_{\theta_i}[\psi_i] = \int_{\Theta_i} \psi_i(\theta_i) \pi^\star(\theta_i) d\theta_i.
\]
For known \( \Sigma \) and the prior distribution on \( \mu_i \), assumed to be Gaussian with mean \( \mu_i \) and covariance matrix \( \Sigma^i_{\mu_i}, \bar{\psi}_i \) is given by Eq. (10) in [28]:
\[
\bar{\psi}_i = \Phi \left( \frac{(-1)^{i+1} \left( \frac{\mu_i - \mathbf{x}_0 + \mathbf{x}_1}{2} \right)^T \Sigma^{-1} (\mathbf{x}_0 - \mathbf{x}_1) + c \right) \sqrt{\mathbf{(x}_0 - \mathbf{x}_1)^T \Sigma^{-1} (\mathbf{x}_0 - \mathbf{x}_1)}
\]
where
\[
\mu_i = \frac{n_i \mathbf{x}_0 + m_i \mathbf{x}_1}{n_i + m_i}, \quad \nu_i = n_i + m_i
\]
and \( \nu_i > 0 \) is a measure of our certainty concerning the prior knowledge – the larger the \( \nu_i \) the more the prior distribution is about \( \mu_i \). Letting \( \mu = [\mu_0^T, \mu_1^T]^T \), the moments that interest us are of the form \( E_{S_0}[\epsilon_0^2 \mu], E_{S_0}[\epsilon_0 \epsilon_1 \mu], \text{ and } E_{S_0}[\epsilon_0 \epsilon_1 \mu] \), which are used to obtain MSE\(_{S_0}[\epsilon_0^2 \mu], \text{ MSE}_{S_0}[\epsilon_0 \epsilon_1 \mu], \text{ and } \text{MSE}_{S_0}[\epsilon_0 \epsilon_1 \mu] \), which are needed to characterize MSE\(_{S_0}[\epsilon_0^2 \mu] \).

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: \( \mu_{p_0}, \mu_{p_1}, \Sigma_{p_0}, \Sigma_{p_1}, n_{p_0}, n_{p_1} \), \( p = 1, 2, \ldots \), where the means and the covariance matrix are arbitrary. The common assumptions for Raudys–Kolmogorov asymptotics are \( n_0 \to \infty, n_1 \to \infty, p \to \infty, p/n_0 \to \hat{\mu} < \infty, p/n_1 \to \hat{\mu} < \infty \). 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, \( \hat{\mu}_p = \sqrt{\left( \mu_{p_0} - \mu_{p_1} \right)^T \Sigma_{p}^{-1} \left( \mu_{p_0} - \mu_{p_1} \right)} \), is finite and \( \lim_{p \to \infty} \hat{\mu}_p = \hat{\mu} \) (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, \( \bar{\psi}_i \), we modify the sequence of Gaussian discrimination problems to \( \mu_{p_0}, \mu_{p_1}, \Sigma_{p_0}, \Sigma_{p_1}, n_{p_0}, n_{p_1}, m_{p_0}, m_{p_1}, \psi_{p_0}, \psi_{p_1} \), \( p = 1, 2, \ldots \), \( \psi_{p_0} \to \infty, \psi_{p_1} \to \infty, \psi_{p_0} \to \hat{\psi}_{p_0} < \infty, \psi_{p_1} \to \hat{\psi}_{p_1} < \infty \). In addition to the previous conditions, we assume that the following limits exist for \( i, j = 0, 1 \):
\[
\lim_{p \to \infty} m_{p_0} \Sigma_{p}^{-1} \mu_{p_0} = m \Sigma^{-1} \mu_0, \quad \lim_{p \to \infty} m_{p_1} \Sigma_{p}^{-1} \mu_{p_1} = m \Sigma^{-1} \mu_1, \quad \text{ and } \lim_{p \to \infty} m_{p} \Sigma_{p}^{-1} \mu_{p} = m \Sigma^{-1} \mu
\]
are some symbolic presentations of the constants to which the limits converge.

We refer to all of the aforementioned conditions, along with \( \nu_i \to \infty, \nu_{p_0} \to \hat{\nu}_{p_0} < \infty, \text{ as the Bayesian–Kolmogorov asymptotic conditions (b.k.a.c).} \) We denote the limit under these conditions by \( \lim_{b.k.a.c} \), which means that, for \( i, j = 0, 1 \),
\[
\lim_{b.k.a.c} \left( \frac{p \to \infty, n_0 \to \infty, \nu_i \to \infty, \hat{\nu}_p \to \hat{\nu}, \hat{\psi}_p \to \hat{\psi}}{p_0 \to \hat{\psi}_p \to \hat{\psi} \text{, where } m_i \Sigma_i^{-1} \mu_i = \Sigma_i^{-1} \mu_i, \text{ and } m_i \Sigma_i^{-1} \mu_i = \Sigma_i^{-1} \mu_i} \right)
\]
This limit is defined for the case where there is conditioning on a specific value of \( \mu_p \). Therefore, in this case \( \mu_p \) is not a random variable, and for each \( p \) it is a vector of constants. In the absence of such conditioning, the sequence of discrimination problems and the above limit reduce to
\[
\left( \Sigma_{p_0}, n_{p_0}, n_{p_1}, m_{p_0}, m_{p_1}, \psi_{p_0}, \psi_{p_1}, \psi_{p_0} \to \infty, \psi_{p_1} \to \infty \right)
\]
and
\[
\lim_{b.k.a.c} \left( \frac{p \to \infty, n_0 \to \infty, \nu_i \to \infty, \hat{\nu}_p \to \hat{\nu}, \hat{\psi}_p \to \hat{\psi}}{p_0 \to \hat{\psi}_p \to \hat{\psi} \text{, where } m_i \Sigma_i^{-1} \mu_i = \Sigma_i^{-1} \mu_i, \text{ and } m_i \Sigma_i^{-1} \mu_i = \Sigma_i^{-1} \mu_i} \right)
\]
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 \( " \lim \text{ prob. to } b.k.a.c \) and \( " \lim \) and \( " \lim \) 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 \( \eta_{a_2,a_3,a_4} = (a_1-a_2)' \Sigma^{-1}(a_3-a_4) \) and, for ease of notation write \( \eta_{a_2,a_3,a_4}, \) as \( \eta_{i} \). There are two special cases: (1) the square of the Mahalanobis distance in the space of the parameters of the unknown class conditional densities, \( D^2_m = \eta_{m,m} > 0 \); and (2) a measure of distance of prior distributions, \( \Delta^2_m = \eta_{m,m} > 0 \), where \( m = [m_0',m_1']' \). The conditions in (10) assure existence of \( \lim_{i \to \infty} \eta_{a_2,a_3,a_4} \), where the \( a_j \)'s can be any combination of \( m_0 \) and \( m_1 \), \( i = 0,1 \). Consistent with our notations, we use \( \eta_{a_2,a_3,a_4} \), \( \Delta^2_m \), and \( \Delta^2_m \) to denote the limit \( \lim_{i \to \infty} \eta_{a_2,a_3,a_4} \), \( \Delta^2_m \), and \( \Delta^2_m \), respectively. Thus,

\[
\eta_{a_2,a_3,a_4} = (a_1-a_2)' \Sigma^{-1}(a_3-a_4) = (a_1-a_2)' \Sigma^{-1}a_3 - a_2' \Sigma^{-1}a_3 + a_2' \Sigma^{-1}a_4.
\]

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 \( \nu_i/n_i \) is an indicator of relative certainty of prior knowledge to the data: the smaller the \( \nu_i/n_i \), 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 \( \beta_i = \nu_i/n_i \), so that \( \beta_i = \nu_i/n_i \to \gamma_i \).

4. First moment of \( Z_i^2 \)

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

4.1. Conditional expectation of \( Z_i^2 \): \( E_{S_i}[Z_i^2 | \mu] \)

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 \( C_0^R, C_1^R \), and \( D \) depend on \( \mu \), but to ease the notation we leave this implicit.

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

\[
\lim_{i \to \infty} E_{S_i}[Z_i^2 | \mu] = \Phi \left( \gamma_0 - \frac{C_0^R + c}{\sqrt{D}} \right),
\]

so that

\[
\lim_{i \to \infty} E_{S_i}[Z_i^2 | \mu] = \alpha_0 \Phi \left( \frac{-C_0^R + c}{\sqrt{D}} \right) + \alpha_1 \Phi \left( \frac{C_1^R - c}{\sqrt{D}} \right),
\]

where

\[
C_0^R = \frac{1}{2(1+\gamma_0)} \left( \gamma_0 \theta \eta_{m_0,m_0} - \eta_{m_0,m_0} + \beta_0 \right) + (1-\gamma_0) \theta_0 + (1+\gamma_0) J_1,
\]

\[
C_1^R = \frac{1}{2(1+\gamma_1)} \left( \gamma_1 \theta \eta_{m_1,m_1} - \eta_{m_1,m_1} + \beta_0 \right) + (1-\gamma_1) \theta_1 + (1+\gamma_1) J_0.
\]

\[
D = \delta^2_p + J_0 + J_1.
\]

**Theorem 1** suggests a finite-sample approximation:

\[
E_{S_i}[Z_i^2 | \mu] \approx D^2_m \left( -\frac{C_0^R + c}{\sqrt{D^2_m + p + \frac{p}{n_0}}} \right),
\]

where \( G_0^R \) is obtained by using the finite-sample parameters of the problem in (16), namely

\[
G_0^R = \frac{1}{2(1+\beta_0)} \left( \beta_0 \eta_{m_0,m_0} - \eta_{m_0,m_0} + \beta_0 \right) + (1-\beta_0) \frac{p}{n_0} + (1+\beta_0) \frac{p}{n_1}.
\]

To obtain the corresponding approximation for \( E_{S_i}[Z_i^2 | \mu] \), it suffices to use (17) by changing the sign of \( c \), exchanging \( n_0 \) and \( n_1 \), \( \nu_0 \) and \( \nu_1 \), \( \eta_{m_0} \) and \( \eta_{m_1} \), and \( \mu_0 \) and \( \mu_1 \). To obtain a Raudys-type of finite-sample approximation for the expectation of \( Z_i^2 \), first note that the Gaussian distribution in (7) can be rewritten as

\[
Z_i^2 = \mathcal{N}(U_0(x_0, x_1, z) + \mathcal{C}(x_0, x_1, z) \in \mathcal{Y}_0, \mu),
\]

where \( \mu \) is independent of \( S_m \), \( \mathcal{Y}_1 \) is a multivariate Gaussian \( \mathcal{N}(m_0, ((n_0/n_1+1)(n_0/n_1+1))^{-1} \Sigma) \), and

\[
U_0(x_0, x_1, z) = \left( \frac{\nu_0}{n_0/n_1+1} - z \eta_{m_0} - \eta_{m_0} x_0 + x_1 \right)' \Sigma^{-1}(x_0 - x_1).
\]

Taking the expectation of \( Z_i^2 \) relative to the sampling distribution and then applying the standard normal approximation yields the Raudys-type of approximation:

\[
E_{S_i}[Z_i^2 | \mu] = \mathcal{N}(U_0(x_0, x_1, z) \in \mathcal{Y}_0, \mu),
\]

Algebraic manipulation yields (Supplementary Section A)

\[
E_{S_i}[Z_i^2 | \mu] \approx \Phi \left( -\frac{G_0^R + c}{\sqrt{G_0^R + D^2_m}} \right).
\]

where

\[
G_0^R = C_0^R,
\]

with \( G_0^R \) being presented in (18) and

\[
D^2_m = \delta^2_p + \frac{\delta^2_p}{n_0(1+\beta_0)} + \frac{\delta^2_p}{n_1(1+\beta_0)} + \frac{\delta^2_p}{n_0(1+\beta_0)} + \left( \frac{1+\beta_0}{n_0} \right)^2 \eta_{m_0,m_0} - \left( \frac{1+\beta_0}{n_1} \right)^2 \eta_{m_0,m_0} - \left( \frac{1+\beta_0}{n_1} \right)^2 \eta_{m_0,m_0} - \left( \frac{1+\beta_0}{n_1} \right)^2 \eta_{m_0,m_0}.
\]

The corresponding approximation for \( E_{S_i}[Z_i^2 | \mu] \) is

\[
E_{S_i}[Z_i^2 | \mu] \approx \Phi \left( -\frac{C_0^R + c}{\sqrt{D^2_m}} \right).
\]

where \( D^2_m \) and \( G^R_m \) are obtained by exchanging \( n_0 \) and \( n_1 \), \( \nu_0 \) and \( \nu_1 \), \( m_0 \) and \( m_1 \), and \( \mu_0 \) and \( \mu_1 \). It is straightforward to see that

\[
C_0^R \leq C_0^R, D^2_m \leq D^2_m + J_0 + J_1.
\]
with $G_0$ 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 $\hat{\varepsilon}_i^B$: $E_{\mu,S}[\hat{\varepsilon}_i^B]$

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

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

$$\lim_{b.k.c.} E_{\mu,S}[\hat{\varepsilon}_i^B] = \Phi\left( -\frac{H_0 + c}{\sqrt{F}} \right),$$

so that

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

where

$$H_0 = \frac{1}{2} \left( \sum_{1}^2 - J_1 - J_0 + \frac{J_0}{\nu} + \frac{J_1}{\nu} \right),$$

$$H_1 = \frac{1}{2} \left( \sum_{1}^2 + J_0 - J_1 + \frac{J_0}{\nu} + \frac{J_1}{\nu} \right),$$

$$F = \sum_{1}^2 + J_0 + J_1 + \frac{J_0}{\nu} + \frac{J_1}{\nu}. \Box$$

Theorem 2 suggests the finite-sample approximation:

$$E_{\mu,S}[\hat{\varepsilon}_i^B] \approx \Phi\left( -\frac{H_0 + c}{\sqrt{F}} \right),$$

where

$$H_0^* = \frac{1}{2} \left( \sum_{1}^2 + \frac{p}{\nu_1} - \frac{p}{\nu_0} + \frac{p}{\nu_0} + \frac{p}{\nu_1} \right).$$

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

$$E_{\mu,S}[\hat{\varepsilon}_i^B] = E_{\mu,S}[U_0(\mathbf{X}_0, \mathbf{X}_1, z) \leq c(\mathbf{z} \in \mathcal{Y}_0, \mu)]$$

$$\approx \Phi\left( -\frac{E_{\mu,S}[U_0(\mathbf{X}_0, \mathbf{X}_1, z) | z \in \mathcal{Y}_0] + c}{\sqrt{\text{Var}_{\mu,S}[U_0(\mathbf{X}_0, \mathbf{X}_1, z) | z \in \mathcal{Y}_0]} \right).$$

Some algebraic manipulations yield (Supplementary Section B)

$$E_{\mu,S}[\hat{\varepsilon}_i^B] \approx \Phi\left( -\frac{H_0 + c}{\sqrt{F_0}} \right),$$

where

$$F_0^* = \left( 1 + \frac{1}{\nu_1} + \frac{1}{\nu_1} + \frac{1}{\nu_1} \right) \sum_{2} + p \left( \frac{1}{\nu_0} + \frac{1}{\nu_1} + \frac{1}{\nu_0} + \frac{1}{\nu_1} \right)$$

$$+ p \left( \frac{1}{2\nu_0} + \frac{1}{2\nu_1} + \frac{1}{2\nu_0} + \frac{1}{2\nu_1} \right) + \left( p \left( \frac{1}{\nu_1} + \frac{1}{\nu_0} + \frac{1}{\nu_1} \right) \right).$$

It is straightforward to see that

$$H_0^* \leq H_0,$$

$$F_0^* \leq F_0,$$

$$\sum_{2} - J_0 + J_1 + \frac{J_0}{\nu} + \frac{J_1}{\nu}, \Box$$

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, $\mathbf{z}$ and $\mathbf{z'}$, yields the second moment representation:

$$E_{\varepsilon^B}[\hat{\varepsilon}_i^B | \mu] = E_{\varepsilon^B}[U_0(\mathbf{X}_0, \mathbf{X}_1, \mathbf{z}) \leq c(\mathbf{z} \in \mathcal{Y}_0, \mu)]$$

$$= E_{\varepsilon^B}[U_0(\mathbf{X}_0, \mathbf{X}_1, \mathbf{z}) \leq c(\mathbf{z} \in \mathcal{Y}_0, \mu) | U_0(\mathbf{X}_0, \mathbf{X}_1, \mathbf{z'}) \leq c(\mathbf{z'} \in \mathcal{Y}_0, \mu)]$$

$$= P(U_0(\mathbf{X}_0, \mathbf{X}_1, \mathbf{z}) \leq c(\mathbf{z} \in \mathcal{Y}_0, \mathbf{z'} \in \mathcal{Y}_0, \mu), \mathbf{z} \neq \mathbf{z'}).$$

where $\mathbf{z}$ and $\mathbf{z'}$ are independent of $S_m$ and $\mathcal{Y}_0$ is a multivariate Gaussian, $N(m, ((\tau_1 + \nu_1 + 1)(\nu_1 + \nu_1)/\nu_1)^2 \Sigma)$, and $U_0(\mathbf{X}_0, \mathbf{X}_1, \mathbf{z})$ 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 $i\neq j = 1,$

$$\lim_{b.k.c.} E_{\mu,S}[\hat{\varepsilon}_i^B | \mu] = \Phi\left( -\frac{c_G^B + c}{\sqrt{D}} \right) \Phi\left( -\frac{c_G^B + c}{\sqrt{D}} \right),$$

so that

$$\lim_{b.k.c.} E_{\mu,S}[\hat{\varepsilon}_i^B | \mu] = \left[ \alpha_0 \Phi\left( -\frac{c_G^B + c}{\sqrt{D}} \right) + \alpha_1 \Phi\left( \frac{c_G^B - c}{\sqrt{D}} \right) \right]^2,$$

where $G_0^B, \gamma_1^B,$ and $D$ are defined in (16). $\Box$

This theorem suggests the finite-sample approximation:

$$E_{\varepsilon^B}[\hat{\varepsilon}_i^B | \mu] \approx \Phi\left( -\frac{c_G^B + c}{\sqrt{D}} \right),$$

which is the square of the approximation (17). Corresponding approximations for $E[I_i^B | \mu]$ and $E[I_i^B | \mu]$ are obtained similarly.

Similar to the proof of Theorem 3, we obtain the conditional cross moment of $\hat{\varepsilon}_i^B$.

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

$$\lim_{b.k.c.} E_{\mu,S}[\hat{\varepsilon}_i^B | \mu] = \Phi\left( -\frac{c_G^B + c}{\sqrt{D}} \right) \Phi\left( -\frac{c_G^B + c}{\sqrt{D}} \right),$$

so that

$$\lim_{b.k.c.} E_{\mu,S}[\hat{\varepsilon}_i^B | \mu] = \sum_{i=0}^1 \sum_{j=0}^1 \alpha_0 \alpha_j \Phi\left( -\frac{c_G^B + c}{\sqrt{D}} \right) \Phi\left( -\frac{c_G^B + c}{\sqrt{D}} \right),$$

where $G_0^B$ and $D$ are defined in (16) and $\gamma_i$ is defined in (47). $\Box$

This theorem suggests the finite-sample approximation:

$$E_{\varepsilon^B}[\hat{\varepsilon}_i^B | \mu] \approx \Phi\left( -\frac{c_G^B + c}{\sqrt{D}} \right),$$

This is a product of (17) and the finite-sample approximation for $E_{\varepsilon^B}[\hat{\varepsilon}_i^B | \mu]$. 

Hence, the deviation variance is also asymptotically zero, therefore, defining the bias as 
\[ \text{Bias}_{\gamma} (\theta^B | \mu) = E_{\gamma} (\theta^B - \epsilon | \mu). \] 

The asymptotic RMS reduces to 
\[ \text{lim RMS}_{\gamma} (\theta^B | \mu) = \text{lim} | \text{Bias}_{\gamma} (\theta^B | \mu). \] 

To express the conditional bias, as proven in [16], 
\[ \text{lim} E_{\gamma} (\theta | \mu) = \alpha_0 \Phi \left( -\frac{G_0 + \epsilon}{\sqrt{D}} \right) + \alpha_1 \Phi \left( \frac{G_1 - \epsilon}{\sqrt{D}} \right). \] 

where 
\[ G_0 = \frac{1}{2} \left( \delta^2 \mu + J_1 - J_0 \right), \]
\[ G_1 = -\frac{1}{2} \delta^2 \mu + J_0 - J_1, \]
\[ D = \delta^2 \mu + J_0 + 1. \] 

It follows from Theorem 1 and (46) that 
\[ \text{lim} \text{Bias}_{\gamma} (\theta^B | \mu) = \alpha_0 \Phi \left( -\frac{G_0 + \epsilon}{\sqrt{D}} \right) - \Phi \left( -\frac{G_0 + \epsilon}{\sqrt{D}} \right) + \alpha_1 \Phi \left( \frac{G_1 - \epsilon}{\sqrt{D}} \right) - \Phi \left( \frac{G_1 - \epsilon}{\sqrt{D}} \right). \] 

Recall that the MMSE error estimator is unconditionally unbiased: 
\[ \text{Bias}_{\gamma} (\theta^B | \mu) = E_{\gamma, \theta} (\theta^B - \mu) = 0. \]

We next obtain Raudys-type approximations corresponding to Theorems 3 and 4 by utilizing the joint distribution of \( U(x_0, x_1, z) \) and \( U(x_0, x_1, z) \), defined in (20), with \( z \) and \( z' \) being independently selected from populations \( \mathcal{P}_0 \) or \( \mathcal{P}_1 \). We employ the function 
\[ \Phi (a, b, p) = \int_a^b \int_{-\infty}^\infty \frac{1}{2\pi \sqrt{1-p^2}} \exp \left( -\frac{(x^2 + y^2 - 2pxy)}{2(1-p^2)} \right) \, dx \, dy, \] 

which is the distribution function of a joint bivariate Gaussian vector with zero means, unit variances, and correlation coefficient \( p \). Note that \( \Phi (a, b, c) = \Phi (a) \) and \( \Phi (a, b, 0) = \Phi (a) \Phi (b) \). For simplicity of notation, we write \( \Phi (a, c) \) as \( \Phi (a, c) \). The rectangular-area probabilities involving any jointly Gaussian pair of variables \( (x, y) \) can be expressed as 
\[ P(x \leq a, y \leq b) = \Phi \left( \frac{c - \mu_x}{\sigma_x}, \frac{d - \mu_y}{\sigma_y}, \rho_{xy} \right). \] 

with \( \mu_x = E(x), \mu_y = E(y), \sigma_x = \sqrt{\text{Var}(x)}, \sigma_y = \sqrt{\text{Var}(y)}, \) and correlation coefficient \( \rho_{xy} \). 

Using (36), we obtain the second-order extension of (21) by 
\[ E_{\gamma} (\theta^B | \mu) = P(U(x_0, x_1, z) \leq c, U(x_0, x_1, z) \leq c | x \in \mathcal{P}_0, z \in \mathcal{P}_0, \mu) \]
\[ \approx \Phi \left( \frac{-E_{\gamma} (U(x_0, x_1, z) | x \in \mathcal{P}_0, \mu) + c}{\sqrt{\text{Var}_{\gamma} (U(x_0, x_1, z) | x \in \mathcal{P}_0, \mu)}} \right) + c \cdot \text{Cov}_{\gamma} (U(x_0, x_1, z) | x \in \mathcal{P}_0, \mu) \] 

Where, after some algebraic manipulations we obtain 
\[ C_{\gamma} (x_0, x_1, z) = \frac{1}{\sqrt{\text{Var}(x_0, x_1, z)}} \frac{1}{\sqrt{\text{Var}(z)}} \frac{1}{\sqrt{\text{Var}(x_1, x_2, x_3)}} \] 

Supplementary Section C gives the proof of (56). Since \( C_{\gamma} (x_0, x_1, z) \) is asymptotically exact, i.e., (56) becomes equivalent to the result of Theorem 3. We obtain the conditional cross moment similarly:
theorems.

(26)

(30)

(31)

(32)

(59)

(60)

(61)

(62)

(63)

(64)

(65)

where

\[ V_{U_0} = \text{Var}_{s,x} [U_0(X_0, X_1, z) | z \in \mathcal{Y}_0, \mu] \]

\[ V_{U_0} = \text{Var}_{s,x} [W(X_0, X_1, x) | x \in \mathcal{P}_0, \mu] \]

where superscript \(^c\) denotes conditional variance. Algebraic manipulations like those leading to (53) yield

\[ E_{S_i}[\mu^R | \nu^R] \propto \Phi \left( \frac{-c_{B_i}^R + c - c_{0_1}^R + c_{0_1}^R}{\sqrt{D_{B_i}^R}} \right) \]

\[ \text{and} \ G_0^R \text{ and } D_0^R \text{ having been obtained previously in Eqs. (49) and (50) of [16]}, \]

\[ G_0^R = E_{s,x} [W(X_0, X_1, x) | x \in \mathcal{P}_0, \mu] = \frac{1}{2} \left( \sigma_{\mu}^2 + \frac{P}{m} - \frac{P}{n_0} \right) \]

\[ D_0^R = \text{Var}_{s,x} [W(X_0, X_1, x) | x \in \mathcal{P}_0, \mu] = \sigma_{\mu}^2 + \frac{P}{n_0} + \frac{1}{m} + \frac{P}{n_0} + \frac{1}{2m^2} \]

Similarly, we can show that

\[ E_{S_i}[\mu^R | \nu^R] \propto \Phi \left( \frac{-c_{B_i}^R + c - c_{0_1}^R + c_{0_1}^R}{\sqrt{D_{B_i}^R}} \right) \]

\[ \text{where} \ D_{B_i}^R \text{ and } G_{B_i}^R \text{ are obtained as in (54), and } D_{0_i}^B, G_{0_i}^B, \text{ and } c_{B_i}^R \text{ are obtained by exchanging } n_0 \text{ and } n_1 \text{ in } D_{0_i}^B, G_{0_i}^B, \text{ and } c_{B_i}^R \text{ respectively. Similarly,} \]

\[ E_{S_i}[\mu^R | \nu^R] \propto \Phi \left( \frac{-c_{B_i}^R + c - c_{0_1}^R + c_{0_1}^R}{\sqrt{D_{B_i}^R}} \right) \]

\[ \text{where} \ c_{B_i}^R = \frac{1}{n_0(1 + \beta_0)} \left[ \sigma_{\mu}^2 + \beta_0 \sigma_{\mu}^2 + \beta_0 n_{\mu_0, \mu_{B_i}} \right] + \frac{(1 - \beta_0)P}{2n_0(1 + \beta_0)} - \frac{P}{2n_1^2} \]

\[ c_{0_i}^R = \frac{1}{n_1(1 + \beta_0)} \left[ \sigma_{\mu}^2 + \beta_0 \sigma_{\mu}^2 + \beta_0 n_{\mu_0, \mu_{B_i}} \right] + \frac{(1 - \beta_0)P}{2n_0(1 + \beta_0)} - \frac{P}{2n_1^2} \]

\[ E_{S_i}[\mu^R | \nu^R] \propto \Phi \left( \frac{-c_{B_i}^R + c - c_{0_1}^R + c_{0_1}^R}{\sqrt{D_{B_i}^R}} \right) \]

\[ \text{where} \ e_{0_i}^R \text{ is obtained by exchanging } n_0 \text{ and } n_1, \nu_0 \text{ and } \nu_1, \mu_0 \text{ and } \mu_1 \text{ in } c_{0_i}^R \text{.} \]

We see that \( c_{B_i}^R > 0 \), \( c_{0_1}^R < 0 \), and \( c_{0_1}^R < 0 \). Therefore, from (26) and the fact that \( c_0^R = c_{0_1} - j_0 + D_0^R - \sigma_{\mu}^2 < 0 \), we see that expressions (59), (62), and (63) are asymptotically exact (compared to Theorem 4).

5.2. Unconditional second and cross moments of \( \epsilon_i^R \)

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

\[ E_{S_i}[\epsilon_i^R] = E_{S_i}[P[U_0(X_0, X_1, z) \leq c | X_0, X_1, z \in \mathcal{Y}_0, \mu]] \]

\[ E_{S_i}[P[U_0(X_0, X_1, z) \leq c | U_0(X_0, X_1, z), z \in \mathcal{Y}_0, \mu]] \]

\[ = P[U_0(X_0, X_1, z) \leq c, U_0(X_0, X_1, z) \leq c, U_0(X_0, X_1, z) \leq c] \]

\[ \text{Similar to the proofs of Theorems 3 and 4, we get the following theorems.} \]

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

\[ \lim_{b,k,c} E_{S_i}[\epsilon_i^R] = \Phi \left( \frac{-H_0 + c}{\sqrt{F}} \right) \Phi \left( \frac{-H_1 + c}{\sqrt{F}} \right) \]

so that

\[ \lim_{b,k,c} E_{S_i}[\epsilon_i^R] = \alpha_i \Phi \left( \frac{-H_0 + c}{\sqrt{F}} + \alpha_i \Phi \left( \frac{H_1 - c}{\sqrt{F}} \right) \right)^2 \]

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

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

\[ \lim_{b,k,c} E_{S_i}[\epsilon_i^R] = \lim_{b,k,c} E_{S_i}[\epsilon_i^R] = \lim_{b,k,c} E_{S_i}[\epsilon_i^R] \]

so that

\[ \lim_{b,k,c} E_{S_i}[\epsilon_i^R] = \sum_{i=0}^{1} \sum_{i=0}^{1} \alpha_i \alpha_i \Phi \left( \frac{-H_i + c}{\sqrt{F}} \right) \Phi \left( \frac{-H_i + c}{\sqrt{F}} \right) \]

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

Theorems 5 and 6 suggest the finite-sample approximation:

\[ E_{S_i}[\epsilon_i^R] = \frac{E_{S_i}[\epsilon_i^R] - E_{S_i}[\epsilon_i^R]}{E_{S_i}[\epsilon_i^R]} \]

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

\[ \lim_{b,k,c} \text{Var}_{S_i} [\epsilon_i^R] = \lim_{b,k,c} \text{Var}_{S_i} [\epsilon_i^R] = \lim_{b,k,c} \text{Var}_{S_i} [\epsilon_i^R] \]

\[ = \text{lim} \text{Cov}_{S_i} [\epsilon_i^R, \epsilon_i^R] = \text{lim} \text{RMS}_{S_i} [\epsilon_i^R] = 0 \]

In [30], it was shown that \( \epsilon_i^R \) is strongly consistent, meaning that \( \epsilon_i^R(S_n) - c(S_n) \to 0 \) 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 \( \text{MSE}_{S_i} [\epsilon_i^R] \to 0 \) almost surely as \( n \to \infty \) under similar conditions. Here, we have shown that \( \text{MSE}_{S_i} [\epsilon_i^R] \to 0 \) 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 \( \text{MSE}_{S_i} [\epsilon_i^R] \to 0 \) meaning that \( \text{MSE}_{S_i} [\epsilon_i^R] \) 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 \( \epsilon_i^R \):
Similarly, we similarly obtain

\[ E_{\mu_5,\nu}(\epsilon_{GR}^2) = \Phi \left( \frac{H_{\nu}^R - c \ K_{\nu0}^{GR}}{\sqrt{F_{\nu}^R}} \right), \]

(76)

where \( F_{\nu}^R \), \( H_{\nu}^R \), and \( K_{\nu0}^{GR} \) are obtained by exchanging \( n_0 \) and \( n_1 \), \( \nu_0 \) and \( \nu_1 \), \( m_0 \) and \( m_1 \), and \( \mu_0 \) and \( \mu_1 \), in (34), in \(-H_{\nu}^R \) obtained from (31), and similarly obtain

\[ E_{\mu_5,\nu}(\epsilon_{GR}^2) = \Phi \left( \frac{-H_{\nu}^R + c \ K_{\nu0}^{GR}}{\sqrt{F_{\nu}^R}} \right), \]

(77)

where

\[ K_{\nu0}^{GR} = \frac{p}{(n_0 + \nu_0)(n_1 + \nu_1)} + \frac{(n_0 - \nu_0)p}{2n_0^2(n_0 + \nu_0) + 2n_1^2(n_1 + \nu_1)} + \frac{(n_1 - \nu_1)p}{2n_1^2(n_1 + \nu_1) + 2n_0^2(n_0 + \nu_0)} + \frac{1}{n_0 + \nu_0} + \frac{n_0}{n_1 + \nu_1} + \frac{1}{n_1 + \nu_1} \left( \frac{1}{n_0 + \nu_0} + \frac{1}{n_1 + \nu_1} \right) \frac{p}{\nu_1} + \frac{1}{\nu_1} \left( \frac{1 + \nu_0}{n_1 + \nu_1} + \frac{1 + \nu_1}{n_0 + \nu_0} \right) \frac{p}{\nu_0} \]

(78)

Supplementary Section E presents the proof of (78). Since \( K_{\nu0}^{GR} < 0 \), (77) is asymptotically exact (compared to Theorem 5). Similar to (57) and (59), where we characterized conditional cross moments, we can get the unconditional cross moments as follows:

\[ E_{\mu_5,\nu}(\epsilon_{GR}^2) = E_{\mu_5,\nu}(\epsilon_{GR}^2 | x) \leq c, \ W(\Omega) \leq c | x \in \Omega, x \in \Omega) \]

\[ = \Phi \left( -E_{\mu_5,\nu}(\epsilon_{GR}^2 | x) \epsilon_{GR} \right) \]

(79)

where

\[ V_{\nu_0} = \text{Var}_{\mu_5,\nu}(\epsilon_{GR}) \leq c, \ W(\Omega) \leq c | x \in \Omega, x \in \Omega) \]

(80)

the superscript “U” representing the unconditional variance, \( H_{\nu}^R \) and \( F_{\nu}^R \) being presented in (31) and (34), respectively, and

\[ K_{\nu0}^{GR} = \left( \frac{n_0}{\nu_0} + \frac{1}{\nu_1} \right) \Delta_m^2 + \frac{p}{2n_0^2} + \frac{p}{2n_1^2} + \frac{n_0p}{n_1^2} \]

\[ + \frac{(n_0 - \nu_0)p}{2n_0^2(n_0 + \nu_0)} + \frac{(n_1 - \nu_1)p}{2n_1^2(n_1 + \nu_1)} \]

(81)

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

\[ E_{\mu_5,\nu}(\epsilon_{GR}^2) \]

(82)

where

\[ K_{\nu0}^{GR} = \frac{1}{\nu_0} + \frac{1}{\nu_1} \Delta_m^2 + \frac{p}{2n_0^2} + \frac{p}{2n_1^2} + \frac{p}{2n_0^2} + \frac{p}{2n_1^2} + \frac{p}{n_0^2} + \frac{p}{n_1^2} \]

(83)

See Supplementary Section G for the proof of (83). Having \( K_{\nu0}^{GR} < 0 \) and \( K_{\nu0}^{GR} > 0 \) along with (35) makes (79) and (82) asymptotically exact (compared to Theorem 6).

5.3. Conditional and unconditional second moment of \( \epsilon_1 \)

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:

\[ E_{\nu_0}(\epsilon_{GR}^2) \]

(84)

with \( C_{\nu_0}^R \) and \( D_{\nu_0}^R \) defined in (61),

\[ E_{\nu_0}(\epsilon_{GR}^2) \]

(85)

and

\[ E_{\nu_0}(\epsilon_{GR}^2) \]

(86)

where

\[ C_{\nu_0}^R = \frac{C_{\nu_0}^R}{\sqrt{D_{\nu_0}^R}} \]

(87)

Similar to obtaining (79), we can show that

\[ E_{\mu_5,\nu}(\epsilon_{GR}^2) \]

(88)

with \( H_{\nu_0}^R \) and \( F_{\nu_0}^R \) given in (31) and (34), respectively, and

\[ K_{\nu0}^R = \left( \frac{1}{\nu_0} + \frac{1}{\nu_1} \right) \Delta_m^2 + \frac{p}{2n_0^2} + \frac{p}{2n_1^2} + \frac{p}{2n_0^2} + \frac{p}{2n_1^2} + \frac{p}{n_0^2} + \frac{p}{n_1^2} \]

(89)

Similarly,

\[ E_{\mu_5,\nu}(\epsilon_{GR}^2) \]

(90)

with \( K_{\nu0}^R \) obtained from \( K_{\nu0}^R \) by exchanging \( n_0 \) and \( n_1 \), \( \nu_0 \) and \( \nu_1 \).
with $H_0^\beta$ and $F_0^\beta$, given in (31) and (34), respectively, and

$$K_{01}^\beta = \left( \frac{1}{\nu_0} + \frac{1}{\nu_1} \right) \Delta m_0 + \frac{p}{2\nu_0^2} + \frac{p}{2\nu_1} + \frac{p}{\nu_0 \nu_1} - \frac{p}{2\nu_0} - \frac{p}{2\nu_1}. \quad (92)$$

\section{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: $\mathbf{m}_0$, $\mathbf{m}_1$, $\nu_0$, $\nu_1$, and $\Sigma$. We let $\Sigma$ have diagonal elements 1 and off-diagonal elements 0.1. $\mathbf{m}_0$ and $\mathbf{m}_1$ are chosen by fixing $\delta_0^\beta$ ($\delta_0^\beta = 4$, which corresponds to Bayes error 0.1586). Setting $\delta_0^\beta$ and $\Sigma$ fixes the means $\mu_0$ and $\mu_1$ of the class-conditional densities (we assume $\mu_1$ has equal elements and $\mu_0 = -\mu_1$). The priors, $\pi_0$ and $\pi_1$, are defined by choosing a small deviation from $\mu_0$ and $\mu_1$, that is, by setting $\mathbf{m}_i = \mu_i + \delta \mu_i$, where $\delta = 0.01$.

![Fig. 1. Comparison of conditional and unconditional performance metrics of $i^\beta$ 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 $i$ is not plotted as $i^\beta$ is unconditionally unbiased; (b) second and mixed moments; (c) conditional variance of deviation from true error, i.e. $\text{Var}_{\nu_0}^\beta[i^\beta]$ and, unconditional variance of deviation, i.e. $\text{Var}_{\nu_0}^\beta[i^\beta]$; (d) conditional RMS of estimation, i.e. $\text{RMS}_{\text{cond}}^\beta[i^\beta]$ and, unconditional RMS of estimation, i.e. $\text{RMS}_{\text{uncond}}^\beta[i^\beta]$. (a)–(d) correspond to the same scenario in which dimension, $p$, is 15 and 100, $\nu_0 = \nu_1 = 50$, $\mu_i = 0.01 \mu_i$, with $\mu_0 = -\mu_1$, and Bayes error $= 0.1586.$]
2. (unconditional case): Using $\pi_0$ and $\pi_1$, generate random realizations of $\mu_0$ and $\mu_1$.
3. (conditional case): Use the values of $\mu_0$ and $\mu_1$, obtained from Step 1.
4. For fixed $\Pi_0$ and $\Pi_1$, generate a set of training data of size $n_i$ for class $i = 0, 1$.
5. Using the training sample, design the LDA classifier, $\psi_n$, using (2).
6. Compute the Bayesian MMSE error estimator, $\hat{\varepsilon}$, using (5) and (7).
7. Knowing $\mu_0$ and $\mu_1$, find the true error of $\psi_n$ using (3).
8. Repeat Steps 2 through 7, $T_2$ times.
9. Repeat Steps 2 through 7, $T_1$ 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 $\mu_0$ and $\mu_1$, are set in Step 2.

Fig. 1 treats Raudys-type finite-sample approximations, including the RMS. Fig. 1(a) compares the first moments obtained from Eqs. (22) and (33). It presents $E_S [\hat{\varepsilon}^2 | \mu]$ and $E_P S [\hat{\varepsilon}^2]$ computed by Monte Carlo estimation and the analytical expressions. The label “FSA BE Uncond” identifies the curve of $E_P S [\hat{\varepsilon}^2]$, the unconditional expected estimated error obtained from the finite-sample approximation, which according to the basic theory is equal to $E_P S [\varepsilon]$. The labels “FSA BE Cond” and “FSA TE Cond” show the curves of $E_S [\hat{\varepsilon}^2 | \mu]$, the conditional expected estimated error, and $E_P S [\varepsilon | \mu]$, the conditional 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 Fig. 1(a) show substantial agreement with the Monte Carlo estimation.

To obtain the second moments, $\text{Var} [\hat{\varepsilon}]$ and $\text{RMS} [\hat{\varepsilon}]$ as defined in (1), we use Eqs. (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. Fig. 1(b), (c), and (d) compares the Monte Carlo estimation to the finite-sample approximations obtained for second/mixed moments, $\text{Var} [\hat{\varepsilon}]$, and $\text{RMS} [\hat{\varepsilon}]$, respectively. The labels are interpreted similar to those in Fig. 1(a), but for the second/mixed moments instead. For example, “MC BE x TE Uncond” identifies the MC curve of $E_P S [\hat{\varepsilon}^2]$. Fig. 1(b), (c), and (d) shows 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), respectively. The curves of “FSA BExTE”, “FSA BExTE”, and “FSA TExTE” substantially overlap with each other and with their corresponding Monte-Carlo curves. For a better visibility of Fig. 1(b), we henceforth omit the curves of “FSA BExTE” and “FSA TExTE” from the plot.

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

### Table 1

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$p=2$</th>
<th>$p=4$</th>
<th>$p=8$</th>
<th>$p=16$</th>
<th>$p=32$</th>
<th>$p=64$</th>
<th>$p=128$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta = 1$: Conditional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.09</td>
<td>0.11</td>
<td>0.13</td>
<td>0.15</td>
<td>0.18</td>
<td>0.21</td>
<td>0.25</td>
</tr>
<tr>
<td>0.02</td>
<td>0.80</td>
<td>0.82</td>
<td>0.84</td>
<td>0.86</td>
<td>0.88</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td>0.03</td>
<td>1.23</td>
<td>1.25</td>
<td>1.27</td>
<td>1.29</td>
<td>1.31</td>
<td>1.33</td>
<td>1.35</td>
</tr>
<tr>
<td>0.04</td>
<td>1.66</td>
<td>1.68</td>
<td>1.70</td>
<td>1.72</td>
<td>1.74</td>
<td>1.76</td>
<td>1.78</td>
</tr>
<tr>
<td>0.05</td>
<td>2.09</td>
<td>2.11</td>
<td>2.13</td>
<td>2.15</td>
<td>2.17</td>
<td>2.19</td>
<td>2.21</td>
</tr>
<tr>
<td>0.06</td>
<td>2.52</td>
<td>2.54</td>
<td>2.56</td>
<td>2.58</td>
<td>2.60</td>
<td>2.62</td>
<td>2.64</td>
</tr>
<tr>
<td>0.07</td>
<td>2.95</td>
<td>2.97</td>
<td>2.99</td>
<td>3.01</td>
<td>3.03</td>
<td>3.05</td>
<td>3.07</td>
</tr>
<tr>
<td>0.08</td>
<td>3.38</td>
<td>3.40</td>
<td>3.42</td>
<td>3.44</td>
<td>3.46</td>
<td>3.48</td>
<td>3.50</td>
</tr>
<tr>
<td>0.09</td>
<td>3.81</td>
<td>3.83</td>
<td>3.85</td>
<td>3.87</td>
<td>3.89</td>
<td>3.91</td>
<td>3.93</td>
</tr>
<tr>
<td>0.1</td>
<td>4.24</td>
<td>4.26</td>
<td>4.28</td>
<td>4.30</td>
<td>4.32</td>
<td>4.34</td>
<td>4.36</td>
</tr>
</tbody>
</table>

| $\beta = 1$: Unconditional | | | | | | | |
| 0.005 | 7.00 | 7.02 | 7.04 | 7.06 | 7.08 | 7.10 | 7.12 |
| 0.010 | 14.01 | 14.03 | 14.05 | 14.07 | 14.09 | 14.11 | 14.13 |
| 0.020 | 28.03 | 28.05 | 28.07 | 28.09 | 28.11 | 28.13 | 28.15 |
| 0.025 | 35.04 | 35.06 | 35.08 | 35.10 | 35.12 | 35.14 | 35.16 |
| 0.030 | 42.05 | 42.07 | 42.09 | 42.11 | 42.13 | 42.15 | 42.17 |
| 0.035 | 49.06 | 49.08 | 49.10 | 49.12 | 49.14 | 49.16 | 49.18 |
| 0.040 | 56.07 | 56.09 | 56.11 | 56.13 | 56.15 | 56.17 | 56.19 |
| 0.045 | 63.08 | 63.10 | 63.12 | 63.14 | 63.16 | 63.18 | 63.20 |

![Fig. 2.](image-url) (a) The conditional RMS of estimation, i.e. $\text{RMS} [\hat{\varepsilon} | \mu]$, as a function of $p < 200$ and $n < 200$. From top to bottom, the rows correspond to $\beta = 0.5, 1, 2$. From left to right, the columns correspond to $\alpha^2 = 4, 16$. (b) The unconditional RMS of estimation, i.e. $\text{RMS} [\hat{\varepsilon}]$, as a function of $p < 1000$ and $n < 2000$. From top to bottom, the rows correspond to $\beta = 0.5, 1, 2$. From left to right, the columns correspond to $\alpha^2 = 4, 16$.}
7. Examination of the Raudys-type RMS approximation

Eqs. (18), (24), (53), (56), and (63) show that $\text{RMS}_{n}[\mathbf{b}^*|\mu]$ is a function of 14 variables: $p, n_0, \eta_1, \beta, \rho_1, \delta^2_{\mu}, \eta_{\mu,\mu}, \eta_{\mu,m_{\mu},m}, \nu_{\mu,\mu}, \nu_{\mu,m_{\mu},m}, \tau_{\mu,\mu}, \tau_{\nu,\mu}, \tau_{\nu,m_{\mu},m}, \tau_{\theta,\mu}, \tau_{\theta,m_{\mu},m}$. 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 $n_0 = n_1 = n/2$, $\beta_0 = \beta_1 = \beta$ and assume priors are centered at the unknown true means, so that $\mathbf{m}_0 = \mu_0$ and $\mathbf{m}_1 = \mu_1$. Using these assumptions, $\text{RMS}_{n}[\mathbf{b}^*|\mu]$ is only a function of $p, n, \beta$, and $\delta^2_{\mu}$. We let $p \in [4, 200]$, $n \in [40, 200]$, $\beta \in (0.5, 1, 2)$, $\delta^2_{\mu} \in (4, 16)$, which means that the Bayes error is 0.158 or 0.022. Fig. 2(a) shows plots of $\text{RMS}_{n}[\mathbf{b}^*|\mu]$ as a function of $p, n, \beta$, and $\delta^2_{\mu}$. These show that for smaller distance between classes, that is, for smaller $\delta^2_{\mu}$ (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 $\text{RMS}_{n}[\mathbf{b}^*|\mu]$ by using the same methodology as developed for the restititution and leave-one-out error estimators in [16,26]. The plots in Fig. 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 $\delta^2_{\mu}$. Therefore, the number of sample points that guarantees $\max_{\delta^2_{\mu} > 0} \text{RMS}_{n}[\mathbf{b}^*|\mu] = \lim_{\delta^2_{\mu} \to 0} \text{RMS}_{n}[\mathbf{b}^*|\mu]$ being less than a predetermined value $\tau$ insures that $\text{RMS}_{n}[\mathbf{b}^*|\mu] < \tau$, for any $\delta^2_{\mu}$. Let the desired bound be $\kappa_{\beta}(p, n, \beta) = \lim_{\delta^2_{\mu} \to 0} \text{RMS}_{n}[\mathbf{b}^*|\mu]$. From Eqs. (52), (54), (55), (59), (63), (84), (85), and (86), we can find $\kappa_{\beta}(n, p, \beta)$ and increase $n$ until $\kappa_{\beta}(n, p, \beta) < \tau$. 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 $\delta^2_{\mu}$ (other $\beta$ shown in the Supplementary Material). A larger dimensionality, a smaller $\tau$, and a smaller $\beta$ result in a larger necessary sample size needed for having $\kappa_{\beta}(n, p, \beta) < \tau$.

Turning to the unconditional RMS, Eqs. (34), (75), (78), (83), (89), and (92) show that $\text{RMS}_{n}[\mathbf{b}^*]$ is a function of 6 variables: $p, n_0, n_1, \nu_0, \nu_1, \delta^2_{\mu}$. Fig. 2(b) shows plots of $\text{RMS}_{n}[\mathbf{b}^*]$ as a function

Fig. 3. $\text{RMS}_{n}[\mathbf{b}^*]$—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. $\Delta_{\lambda}^2 = 4$). From top to bottom the rows correspond to $\beta = 0.5, 1, 2$. The solid-black curves indicate $\text{RMS}_{n}[\mathbf{b}^*]$ 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 $p = 10, 70$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)
of \( p, n, \beta \), and \( \Delta_m^2 \), assuming \( n_0 = n_1 = n/2, \beta_0 = \beta_1 = \beta \). Note that setting the values of \( n \) and \( \beta \) fixes the value of \( \nu_0 = \nu_1 = \nu \) in the corresponding expressions for \( \text{RMS}_{\text{Bayes}}[\hat{\beta}^2] \). Due to the complex shape of \( \text{RMS}_{\text{Bayes}}[\hat{\beta}^2] \), we consider a large range of \( n \) and \( p \). The plots show that a smaller distance between prior distributions (smaller \( \Delta_m^2 \)) corresponds to a larger unconditional RMS of estimation. The plots in Fig. 2(b) show that, as \( \Delta_m^2 \) increases, RMS decreases. Furthermore, Fig. 2(b) (and other unshown plots) demonstrates 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 \( \beta \), this “peaking phenomenon” happens for larger \( n \). On the other hand, with fixed \( \beta \), for larger \( p \), peaking happens for larger \( n \). These observations are presented in Fig. 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 \( \beta = 2 \), 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 minimum necessary sample size for a desired \( \text{RMS}_{\text{Bayes}}[\hat{\beta}^2] \). In fact, this is the more practical way to go because in practice one does not know \( \mu \). Since the unconditional RMS shows a decreasing trend in terms of \( \Delta_m^2 \), we use the previous technique to find the minimum necessary sample size to guarantee a desired unconditional RMS. Table 1 (\( \beta = 1 \): Unconditional) shows the minimum sample size that guarantees \( \max_{n_0 > n} \text{RMS}_{\text{Bayes}}[\hat{\beta}^2] = \lim_{n 
rightarrow 0} \text{RMS}_{\text{Bayes}}[\hat{\beta}^2] \) being less than a predetermined value \( \tau \), i.e., insures that \( \text{RMS}_{\text{Bayes}}[\hat{\beta}^2] < \tau \) for any \( \Delta_m^2 \) (other \( \beta \) shown in the Supplementary Material).

To examine the accuracy of the required sample size that satisfies \( \kappa_1(n,p,\beta) < \tau \) 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 \( \tau \).

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 RMSs, \( \text{RMS}_{\text{Bayes}}[\hat{\beta}^2] \); and (2) the unconditional moments and, therefore, the unconditional RMS, \( \text{RMS}_{\text{Bayes}}[\hat{\beta}^2] \). 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 \( \nu \), with an arbitrary constant limiting ratio between \( n \) and \( p \), and \( n \) and \( \nu \). 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.

Conflict of interest

None declared.

Acknowledgments

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

Appendix A

A.1. Proof of Theorem 1

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

\[
G_0 = \left( m_0^* - \frac{x_0 + x_1}{2} \right)^T \Sigma^{-1} (x_0 - x_1),
\]  

(93)

where \( m_0^* \) is defined in (8). Then

\[
G_0 = \frac{n_0}{(n_0 + n_1)^2} \left( x_0 - x_1 \right) + \frac{n_0 - n_1}{2(n_0 + n_1)} \left( x_0^T \Sigma^{-1} x_0 - x_1^T \Sigma^{-1} x_1 \right) + \frac{1}{2} \left( x_0^T \Sigma^{-1} x_0 - x_1^T \Sigma^{-1} x_1 \right).
\]

(94)

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

\[
y_i = m_i^T \Sigma^{-1} (x_0 - x_1), \quad z_i = x_i^T \Sigma^{-1} x_i, \quad m_i = m_i^T \Sigma^{-1} m_i
\]

(95)

The variance of \( y_i \), given \( \mu \) does not depend on \( \mu \). Therefore, under the Bayesian–Kolmogorov conditions stated in (10), \( m_i^T \Sigma^{-1} m_i \) and \( m_i^T \Sigma^{-1} m_i \) do not appear in the limit. Only \( m_i^T \Sigma^{-1} m_i \) matters, which vanishes in the limit as follows:

\[
\text{Var}_{\text{Bayes}}[y_i|\mu] = m_i^T \Sigma^{-1} m_i - \frac{m_i^T \Sigma^{-1} m_i}{n_0} \frac{n_0}{n_1} \lim_{n \rightarrow \infty} \frac{m_i^T \Sigma^{-1} m_i}{n_0} = 0.
\]

(96)

To find the variance of \( z_i \) and \( z_i \) we can first transform \( z_i \) and \( z_i \) 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 \( y \sim N(\mu, \Sigma) \) and \( A \) being a symmetric positive definite matrix, \( \text{Var}[y^T Ay] = 2\text{tr}(A\Sigma^2) + 4m^T A \Sigma m \), with \( tr \) being the trace operator. Therefore, after some algebraic manipulations, we obtain

\[
\text{Var}_{\text{Bayes}}[z_i|\mu] = \frac{2p}{n_i} + \frac{4m_i^T \Sigma^{-1} m_i}{n_1} \frac{1}{n_i} \frac{m_i^T \Sigma^{-1} m_i}{n_1} = 0.
\]
From the Cauchy–Schwarz inequality (\( \text{Cov}[x, y] \leq \sqrt{\text{Var}[x] \text{Var}[y]} \)), \( \text{Cov}_{x_i, z_i} \mu_{x_i} \rightarrow 0 \), \( \text{Cov}_{x_i, z_i} \mu_{x_i} \rightarrow 0 \), and \( \text{Cov}_{x_i, z_i} \mu_{x_i} \rightarrow 0 \) for \( i, j = 0, 1 \), \( i \neq j \). Furthermore, \( (n_1 - n_0)/2(n_1 + n_0) \mu_k(1 - \gamma)/2(1 + \gamma) \) and \( \nu_i/(n_1 + n_0) \Delta \gamma_i/(1 + \gamma) \). Putting this together and following the same approach for \( G_i \) yields \( \text{Var}_{x_i} G_{x_i} \rightarrow 0 \). In general (via Chebyshev's inequality), \( \text{lim}_{n \rightarrow \infty} \text{Var}[x_i] = 0 \) implies convergence in probability of \( X_i \) to \( \lim_{n \rightarrow \infty} E[X_i] \). Hence, since \( \text{Var}_{x_i} G_{x_i} \rightarrow 0 \), for \( i, j = 0, 1 \) and \( i \neq j \),

\[
\text{plim } G_{x_i} = \lim_{n \rightarrow \infty} E_{x_i} G_{x_i} = \left( -1 \right) \mathbf{1}_{\mu_i} + \frac{\text{Var}(\mu_1 - \mu_2)}{\frac{1}{n_1} + \frac{1}{n_1}} + \frac{\nu_i}{\frac{1}{n_1} + \frac{1}{n_1} + \frac{1}{n_1} + \frac{1}{n_1}} = \mu_i. \tag{98}
\]

Now let

\[
\delta_i = \mathbf{1}_{\mu_i} + \frac{\text{Var}(\mu_1 - \mu_2)}{\frac{1}{n_1} + \frac{1}{n_1}} + \frac{\nu_i}{\frac{1}{n_1} + \frac{1}{n_1} + \frac{1}{n_1} + \frac{1}{n_1}}, \tag{99}
\]

where \( \delta_i = (x_i - x_0)^\top \Sigma^{-1}(x_i - x_0)^\top \). Similar to deriving (97) via the variance of quadratic forms of Gaussian variables, we can show

\[
\text{Var}_{x_i}(\delta_i^2) = 4 \left[ \frac{1}{n_0} + \frac{1}{n_1} + \frac{1}{n_1} \right]^2. \tag{100}
\]

Thus,

\[
\text{Var}_{x_i}(\delta_i^2) \rightarrow 0. \tag{101}
\]

As before, from Chebyshev's inequality it follows that

\[
\text{plim } D_{x_i} = \lim_{n \rightarrow \infty} E_{x_i} D_{x_i} = D. \tag{102}
\]

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

\[
\text{plim } G_{x_i} = \lim_{n \rightarrow \infty} E_{x_i} G_{x_i} = D. \tag{103}
\]

From (103) we have

\[
\frac{1}{\sqrt{D_i}} \rightarrow \frac{1}{\sqrt{D_i}} = \frac{1}{\sqrt{D_i}} \rightarrow \frac{1}{\sqrt{D_i}} = \frac{1}{\sqrt{D_i}}. \tag{104}
\]

with \( \delta(.) \) being the delta function and \( \Delta \) showing convergence in distribution. Having this along with boundedness and continuity of \( \Phi(.) \) allow one to apply the Helly-Bray lemma to write

\[
\text{Var}_{x_i}[z_i(\mu)] = p \left[ \frac{\text{Var}(\mu_1 - \mu_2)}{\frac{1}{n_1} + \frac{1}{n_1}} + \frac{\nu_i}{\frac{1}{n_1} + \frac{1}{n_1} + \frac{1}{n_1} + \frac{1}{n_1}} \right] \lim_{n \rightarrow \infty} \frac{1}{n} = 0. \tag{97}
\]
From (113) we see that $E_p [\text{Var}_b (\tilde{G}^p | \theta)] \leq 0$. In sum, $\text{Var}_b (\tilde{G}^p) \leq 0$ and similar to the use of Chebyshev’s inequality in the proof of Theorem 1, we get

$$
\text{plim}_{b,k,c} G^p = \lim_{b,k,c} E_p [\text{Var}_b (\tilde{G}^p | \theta)] = H_0,
$$

with $H_0$ defined in (29).

On the other hand, for $\tilde{D}_1$ defined in (99) we can write

$$
\text{Var}_b (\tilde{D}_1) = \text{Var}_b (E_p [\tilde{D}_1 | \theta]) + E_p [\text{Var}_b (\tilde{D}_1 | \theta)].
$$

From (115), similar expressions as in (112) for $\tilde{x}^T \Sigma^{-1} \tilde{x}$, we get $E_p [\tilde{D}_1^2] = \Delta^2_p + p/n_0 + p/n_1$. Moreover, $\text{Var}_b (\tilde{D}_1^2)$ is obtained from (100) by replacing $n_i$ with $n_0$ and $n_1$, with $\Delta^2_p$. Thus, from (99),

$$
\text{Var}_b (E_p [\tilde{D}_1 | \theta]) = \frac{(\mu^2 + 1)}{k^2} \left[ 2\Delta^2_p + p/n_0 + p/n_1 \right] = \mu_0 = 0.
$$

Furthermore, since $E_p [\tilde{D}_1^2] = \Delta^2_p + p/n_0 + p/n_1$, from (101),

$$
E_p [\text{Var}_b (\tilde{D}_1 | \theta)] = \frac{(\mu^2 + 1)}{k^2} \left[ 2\Delta^2_p + p/n_0 + p/n_1 \right] = \mu_0 = 0.
$$

Hence, $\text{Var}_b (\tilde{D}_1) \leq 0$ and, similar to (110), we obtain

$$
\text{plim}_{b,k,c} \tilde{D}_1 = \lim_{b,k,c} E_p [\tilde{D}_1 | \theta] = \lim_{b,k,c} E_p [\tilde{D}_1 | \theta] = F,
$$

with $F$ defined in (29). Similar to the proof of Theorem 1, by using the Continuous Mapping Theorem and the Helly-Bray lemma we can show that

$$
\text{lim}_{b,k,c} E_p [\tilde{G}^p] = \text{lim}_{b,k,c} E_p [\tilde{D}_1],
$$

and the result follows. \(\square\)

### Appendix B. Supplementary materials

Supplementary data associated with this paper can be found in the online version at [http://dx.doi.org/10.1016/j.patcog.2013.11.022](http://dx.doi.org/10.1016/j.patcog.2013.11.022).

### References


Amin Zollanvari received Ph.D. in Electrical and Computer Engineering from Texas A&M University, College Station, TX, in 2010. He held a postdoctoral position in Harvard Medical School and Brigham and Women’s Hospital, Boston MA, from 2010 to 2012. He is currently an assistant research scientist in Electrical and Computer Engineering Department and Department of Statistics at Texas A&M University, College Station, TX. His research interests include statistical pattern recognition, random matrix theory, and genomic signal processing.

Edward R. Dougherty is a Professor in the Department of Electrical and Computer Engineering at Texas A&M University in College Station, TX, where he holds the Robert M. Kennedy 26 Chair in Electrical Engineering and is Director of the Genomic Signal Processing Laboratory. He is also Co-Director of the Computational Biology Division of the Translational Genomics Research Institute in Phoenix, AZ. He holds a Ph.D. in mathematics from Rutgers University and an M.S. in Computer Science from Stevens Institute of Technology, and has been awarded the Doctor Honoris Causa by the Tampere University of Technology in Finland. He has been elected fellow to both IEEE and SPIE, has received the SPIE Presidents Award, and served as the editor of the SPIE/IS&T Journal of Electronic Imaging.