ANALYSIS AND OPTIMIZATION OF CLASSIFIER ERROR ESTIMATOR PERFORMANCE WITHIN A BAYESIAN MODELING FRAMEWORK

A Dissertation

by

LORI ANNE DALTON

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2012

Major Subject: Electrical Engineering

ANALYSIS AND OPTIMIZATION OF CLASSIFIER ERROR ESTIMATOR PERFORMANCE WITHIN A BAYESIAN MODELING FRAMEWORK

A Dissertation

by

LORI ANNE DALTON

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Edward R. Dougherty
Committee Members,	Ulisses M. Braga-Neto
	Aniruddha Datta
	Jean-Francois Chamberland
	Ivan Ivanov
Head of Department,	Costas N. Georghiades

May 2012

Major Subject: Electrical Engineering

ABSTRACT

Analysis and Optimization of Classifier Error Estimator Performance within a Bayesian Modeling Framework. (May 2012)

Lori Anne Dalton, B.S., Texas A&M University;
M.S., Texas A&M University
Chair of Advisory Committee: Dr. Edward R. Dougherty

With the advent of high-throughput genomic and proteomic technologies, in conjunction with the difficulty in obtaining even moderately sized samples, small-sample classifier design has become a major issue in the biological and medical communities. Training-data error estimation becomes mandatory, yet none of the popular error estimation techniques have been rigorously designed via statistical inference or optimization. In this investigation, we place classifier error estimation in a framework of minimum mean-square error (MMSE) signal estimation in the presence of uncertainty, where uncertainty is relative to a prior over a family of distributions. This results in a Bayesian approach to error estimation that is optimal and unbiased relative to the model. The prior addresses a trade-off between estimator robustness (modeling assumptions) and accuracy.

Closed-form representations for Bayesian error estimators are provided for two important models: discrete classification with Dirichlet priors (the discrete model) and linear classification of Gaussian distributions with fixed, scaled identity or arbitrary covariances and conjugate priors (the Gaussian model). We examine robustness to false modeling assumptions and demonstrate that Bayesian error estimators perform especially well for moderate true errors.

The Bayesian modeling framework naturally gives rise to a practical expected measure of performance for arbitrary error estimators: the sample-conditioned mean-square error (MSE). Closed-form expressions are provided for both Bayesian models. We examine the consistency of Bayesian error estimation and illustrate a salient application in censored sampling, where sample points are collected one at a time until the conditional MSE reaches a stopping criterion.

Finally, we address applications for gene-expression microarray data, including the suitability of the Gaussian model, a methodology for calibrating normal-inverse-Wishart priors from unused data, and an approximation method for non-linear classification. Arbitrary error estimators may also be optimally calibrated on the fly using a calibration function found off-line for an assumed Bayesian model, sample size, classification rule, and error estimation rule.

In contrast to classical data-driven methods, the Bayesian model proposed here facilitates both the rigorous optimization and analysis of classifier error estimation, exploiting both the assumed model and observed data. Important applications include, but are not limited to, cancer diagnosis and any small-sample classification problem.

To my parents and brother

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Prof. Edward R. Dougherty, for his guidance and dedication to teaching and high-quality research. As my mentor, he has truly had a profound impact on my life.

I would also like to thank Prof. Ulisses M. Braga-Neto, Prof. Aniruddha Datta, Prof. Jean-Francois Chamberland and Prof. Ivan Ivanov for serving on my committee and for their constructive advice. Thanks also go to Prof. Costas N. Georghiades, Ms. Tammy Carda and all of the faculty and staff at Texas A&M that have guided and supported me throughout the course of my research and made my time at Texas A&M University a great experience. I also thank former and current students in the Genomic Signal Processing Lab who have been a generous source of help and encouragement.

Last but not least, I would like to thank my parents and brother for their encouragement, patience and love. I would have been lost without them.

NOMENCLATURE

RMS	Root-mean-square
MSE	Mean-square error
MMSE	Minimum mean-square error
$f_{\mathbf{X},Y}\left(\mathbf{x},y\right)$	Feature-label distribution
$f_{\mathbf{X}\|Y}\left(\mathbf{x}\|y\right)$	Class-conditional distribution
$\delta\left(x\right)$	Generalized delta functional
$f_{\theta}\left(\mathbf{x},y\right)$	Parameterized feature-label distribution
$f_{\theta_y}\left(\mathbf{x}\|y\right)$	Parameterized class-conditional distribution
$B\left(\alpha,\beta\right)$	Beta function
$\mathbf{I}_E$	Indicator function, equal to one if E is true and zero otherwise
$\Gamma\left(\alpha\right)$	Gamma function
δ_i	Kronecker delta function
$f_{\mu,\Sigma}(\mathbf{x})$	Multivariate Gaussian distribution, mean μ and covariance Σ
I_D	$D\times D$ identity matrix
$\Phi\left(x\right)$	Unit normal Gaussian cumulative distribution function
$I\left(x;a,b\right)$	Regularized incomplete beta function
$f_G(x;\alpha,\beta)$	Inverse-gamma distribution, shape α and scale β
Γ_D	Multivariate gamma function
$f_W(\Sigma;S,\kappa)$	Inverse-Wishart distribution, inverse scale matrix S and degrees of freedom κ
$0_{a\times b}$	All zero $a\times b$ matrix
!!	Double factorial
$\mathrm{F}_1\left(a;b,b';c;z,z'\right)$	Appell's hypergeometric function of the first kind

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

Although classification itself has received a great deal of attention, in the form of rule design algorithms such as discriminant analysis and neural networks as well as feature-selection methods, error estimation accuracy represents the salient epistemological issue in classification: model validity [1, 2]. When designing a classifier, error estimation is a critical issue since the error estimate quantifies the predictive capacity of the classifier and prediction is the basis of scientific validation.

The problem of classifier error estimation for small samples has become critical in the past decade with the explosion of interest in molecular biomarker classification for phenotypic discrimination, especially in genomic signal processing [3]. Much attention has been paid to cancer, where classification can be between different kinds of cancer, different stages of tumor development, or various other differences. In response to the flood of high-dimensional gene expression, protein expression, and sequence data from new high-throughput genomic and proteomic technologies, hundreds of papers have appeared relating to biomarkers, the vast majority of which have small samples, even 20 or 30. Table 1 provides a flavor of the situation in gene-expression-based cancer classification, where the table gives the cancer type, classification problem, sample size, and error estimator. The question is: Do these papers contain scientific knowledge [4]? The answer depends on the performance of the error estimator.

In applications where sample data are abundant and cheap to acquire, we can partition the observed data into training and testing samples, the classifier being determined by a classification rule acting on the training sample and the classifier error

The journal model is *IEEE Transactions on Automatic Control.*

Table 1. Summary of cancer classification studies based on less than 50 sample points

Sample Size	Error Estimator	Classification Problem
30	LOOCV	glioblastoma multiforme VS normal [5]
14	LOOCV	endometrial tumors VS ovarian tumors [6]
43	10-fold CV	lymph node metastasis VS non-metastasis [7]
24	LOOCV	adenomas VS follicular carcinomas [8]
18	6-fold CV	colon cancer: stage II VS stage III [9]
20	LOOCV	ovarian tumors: stage I VS stage III or IV [10]
32	LOOCV	primary gastric cancer: lymph node-positive VS node-negative [11]
12	LOOCV	colon cancer: stage II VS stage III [9]
46	LOOCV	breast cancer: high-risk and low-risk patients [12]
33	10-fold CV	acute myeloid leukemia: good VS poor response to induction chemotherapy [13]
38	10-fold CV	paediatric rhabdomyosarcomas (RMS): embryonal (ERMS) VS alveolar (ARMS) [14]
26	10-fold CV	oral squamous cell carcinoma VS normal specimens [15]
19	3-fold CV	gliomas: glioblastomas VS oligodendrogliomas [16]
24	LOOCV	stage II colon cancer: metachronous metastasis VS non-metastasis [17]

estimated by the error rate on the testing sample, without significantly degrading the quality of the classifier or the accuracy of the error estimator. However, error estimation becomes problematic in a small-sample setting since splitting the data results in poor classifier design and so the sample is not split: all of the data is used for design (training) and the error is estimated on the same data. Absent any guarantee on estimation accuracy, it becomes necessary to carefully study the relationship between the true error of a classifier and an error estimator within a probabilistic framework.

A number of training-data error estimators have been proposed in the past, leave-one-out and cross-validation being two popular options that are "distribution-free" in the sense that their computation does not require any distributional knowledge. They are intuitively conceived and asymptotically converge to the true error, however they are supported by little or no validation for small samples. Another question arises: How can we quantify, or even optimize, the small-sample validity of such error estimators?

When an error estimate is reported, it implicitly carries with it the properties of the error estimator and these properties characterize the goodness of the estimate [2]. Full information is contained in the joint distribution between the estimated error, $\widehat{\varepsilon}$, and the true error, ε, of a classifier. Perhaps the single most useful measure of error estimation accuracy is the mean-square error (MSE) between the estimated and true errors, which is the expected square deviation of the estimate from the true error. We also use the root-mean-square (RMS) error, which is the square root of the MSE:

$$\mathrm{RMS}(\widehat{\varepsilon}) = \sqrt{\mathrm{E}[(\widehat{\varepsilon} - \varepsilon)^2]}. \tag{1.1}$$

Being that RMS is the square root of MSE, we will use the two interchangeably, with MSE being used mainly in the equations to avoid square roots. The RMS can also be expressed in terms of bias and deviation variance,

$$\mathrm{RMS}(\widehat{\varepsilon}) = \sqrt{\mathrm{Var}_{\mathrm{dev}}(\widehat{\varepsilon}) + \mathrm{Bias}(\widehat{\varepsilon})^2},$$

where

$$\mathrm{Bias}(\widehat{\varepsilon}) = \mathrm{E}[\widehat{\varepsilon} - \varepsilon] \quad \text{and} \quad \mathrm{Var}_{\mathrm{dev}}(\widehat{\varepsilon}) = \mathrm{Var}(\widehat{\varepsilon} - \varepsilon).$$

The classical interpretation for the expectation in (1.1) conditions on a fixed feature-label distribution and averages performance over the corresponding random sampling procedure. In this work we will develop theory based on a Bayesian interpretation, which conditions on the actual observed sample and averages over all distributions in a Bayesian model. We will clarify the distinction between these interpretations, with emphasis on the implications of the Bayesian interpretation.

A. Classical Error Estimator Analysis: Conditioning on a Fixed Distribution

Historically, analytic study has focused on the first and second marginal moments of true and estimated errors with fixed distributions, either for multinomial discrimination or linear discriminant analysis (LDA) over Gaussian distributions [18, 19, 20, 21, 22, 23, 24]. A summary of such results is available in [25].

That being said, marginal knowledge regarding the error estimator does not provide the kind of joint probabilistic knowledge required for the assessment of estimation accuracy. Such characterizations of classical point-based error estimators, either in the form of a joint density or RMS, are much more recent. For multinomial discrimination, joint distributions of the true error with the resubstitution and leave-one-out cross-validation error estimators for the discrete histogram rule, found using complete enumeration, were only published in 2005 [26], and exact representations of both marginal and mixed second-order moments in 2010 [27]. For LDA, in the univariate Gaussian model the exact marginal distributions for both the resubstitution and leave-one-out estimators have been found, and in the multivariate model with a common known covariance matrix quasi-binomial approximations to the distributions of the resubstitution and leave-one-out estimators were discovered in 2009 [28]. The exact joint distribution between true and estimated errors for LDA with both resubstitution and leave-one-out in the univariate Gaussian model, and approximate joint distributions in the multivariate model with a common known covariance matrix were also found in 2010 [25]. Regarding the RMS, whereas one could utilize approximate representations of the joint density in the multivariate model with a common known covariance matrix to find approximate moments via integration, 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 and dimensionality approach infinity at a fixed rate between the two [29]. Such finite-sample approximations from the double asymptotic method have long been known to show good accuracy [30, 31].

Since performance is averaged over the sampling distribution, both the classifier and its true error are random, being evaluated from different samples. Hence a weakness of the classical approach is that it can only provide insight for a classification rule, not for the actual observed sample or trained classifier. Indeed, the classical approach does not address performance for a fixed sample at all because, absent an underlying framework, nothing is known given a single sample. Further, it is somewhat paradoxical to consider performance on a fixed distribution, since it is unknown in practice. Indeed, if we knew the underlying distribution it would not be necessary to train the classifier or estimate its error in the first place, since the optimal classifier and its true error could be determined exactly.

Given that the actual feature-label distribution is unknown in practice, the classical approach is usually applied in one of two extremes. The first is to estimate the feature-label distribution from the data and take this fixed distribution to be true, hoping that the results are robust. This approach is problematic because small-sample density estimation can be even more difficult than error estimation.

The other extreme would be to avoid distributional assumptions altogether and employ distribution-free bounds on the RMS. In this case, very little, or perhaps nothing, can be said about the precision of the estimate. Further, in the rare instances in which performance bounds are known in the absence of any assumptions on the feature-label distribution, these bounds are so loose as to be virtually worthless for small samples [32]. For instance, consider the following distribution-free RMS bound for the leave-one-out error estimator with the discrete histogram rule and tie-breaking in the direction of class 0 [33]:

$$\mathrm{RMS}(\widehat{\varepsilon}_{\mathrm{loo}}|F) \leq \sqrt{\frac{1}{n}\left(1+\frac{6}{e}\right)+\frac{6}{\sqrt{\pi\,(n-1)}}}, \tag{1.2}$$

where F represents the true feature-label distribution and n is the sample size. This bound is almost useless for small samples; for $n = 200$ it is 0.506. As another example, consider the following bound for leave-one-out with k-nearest-neighbor (kNN) classification and random tie-breaking [34]:

$$\mathrm{RMS}(\widehat{\varepsilon}_{\mathrm{loo}}|F) \leq \sqrt{\frac{1}{n}+\frac{24}{n}\sqrt{\frac{k}{2\pi}}}.$$

With 3NN classification and a sample size of $n = 100$, this bound is 0.353, which is again useless. Although such bounds guarantee good performance for large-samples, a model-free approach for small-samples would leave us without a measure of error-estimation accuracy, thereby rendering the resulting classifier model, classifier and error estimate, epistemologically unsound.

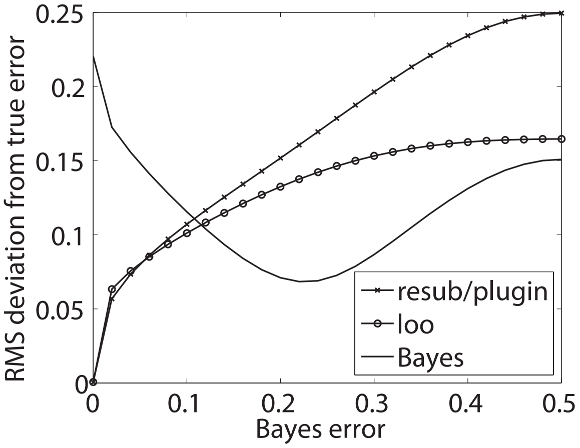


Fig. 1. RMS of three error estimators (y-axis) with respect to Bayes error (x-axis) for the discrete model ($b = 8$ bins, class probability $c = 0.5$, sample size $n = 20$).

Let us now consider bounds on the RMS when there are partial distributional assumptions. If we assume that the distribution comes from an uncertainty class of distributions, $\mathcal{F}$, and we have an expression for the RMS for each distribution in $\mathcal{F}$, then to be assured that the RMS is bounded by some desired level of accuracy, say λ, we require that $\max_{F\in\mathcal{F}} \text{RMS}(F) \leq \lambda$. We may then, for instance, determine a required sample size to insure that $\max_{F\in\mathcal{F}} \text{RMS}(F) \leq \lambda$. If we do not assume an uncertainty class as prior knowledge, then we cannot practically bound the RMS.

The RMS graphs in Fig. 1 represent a synthetic Monte-Carlo simulation for discrete classification with $b = 8$ bins, a class of bin probabilities (Zipf distributions defined in [26], mapping each Bayes error to specific distributions), sample size $n = 20$ and the discrete histogram rule. Resubstitution and leave-one-out are shown, along with a Bayesian error estimator with flat priors defined in the next chapter. Leave-one-out performs well below the bound (1.2); even with $n = 20$ the worst case performance for the Zipf model is 0.25. Moreover, if one wishes to bound the RMS of leave-one-out to a useful degree, one need only assume some maximum Bayes error.

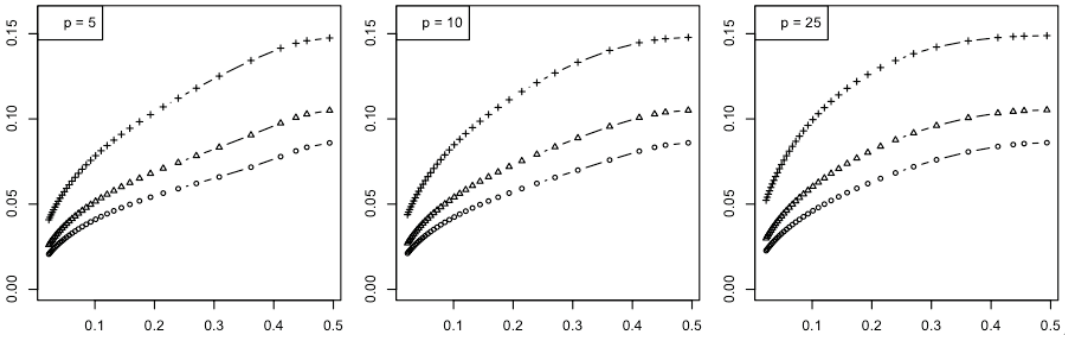


Fig. 2. RMS of leave-one-out (y-axis) with respect to Bayes error (x-axis) for LDA. The left, middle and right plots represent 5, 10 and 25 features, respectively. Within each subplot, lines marked with (+) represent 20 samples, (△) 40 samples and (O) 60 samples.

In the next example, the feature-label distribution consists of two equally probable Gaussian class-conditional densities sharing a known covariance matrix. For the LDA classification rule, we possess an analytic representation of the joint distribution of the true error with the leave-one-out estimator [25]. Figure 2 shows the exact RMS to be a one-to-one increasing function of the Bayes error for dimensions 5, 10 and 25 and sample sizes $n = 20$, 40 and 60. In this model, where the Bayes error is a function of the distance between means of each class, in all cases the maximum RMS is bounded and does not exceed 0.15, even with only 20 sample points. And as before, to bound the RMS below some tolerance, one need only assume a maximum Bayes error, or equivalently a minimum distance between the means. This kind of behavior, where the RMS of leave-one-out is tolerable when the Bayes error is small, is often observed–indeed, we will see this throughout our simulations–but it has only been quantified in a small number of cases [27, 25].

The point of these examples is that in practice, the distribution-free application of any error estimator is an illusion [32]. Even though the computation of an error estimator may be purely data-driven, for instance by counting, which is without an obvious connection to the underlying distributions, its performance is certainly not. For instance, leave-one-out in both Figs. 1 and 2 operates best with low Bayes errors, which is quite typical, so that its use in a small-sample setting implicitly assumes low Bayes error, at least if one is assuming some degree of accuracy. The upshot of all this is that if an error estimator is going to be used in a small-sample setting, there must be modeling assumptions to ensure that the RMS is acceptable and the classifier valid. And if this is the case, why not confront the necessity of assumptions and fully integrate them into the analysis and design process? This is exactly what is done in the Bayesian approach.

B. Bayesian Error Estimator Analysis: Conditioning on a Fixed Sample

Having recognized that modeling assumptions (an uncertainty class) must be postulated when the sample is small to achieve an acceptable RMS, we can go a step further and assume a prior distribution on the uncertainty class, resulting in a Bayesian modeling framework. The transition from an unstructured uncertainty class to a prior distribution governing the parameters defining the uncertainty class is not uncommon in signal processing. For instance, assuming uncertainty in the second-order statistics of a random process originally led to a minimax theory of robust optimal linear filtering [35, 36, 37], whereas subsequently a prior distribution was assumed to govern the uncertainty class, thereby leading to a Bayesian theory of robust linear filtering [38]. In genomic signal processing, the first analysis of robust control for gene regulatory networks assumed an uncertainty class without a prior distribution, thereby resulting in a minimax theory of robust control [39]; subsequently it was assumed that a prior distribution governed the uncertainty class and a Bayesian theory of robust control was developed [40] (see also [41]).

Bayesian frameworks define a mathematical foundation for both the analysis of arbitrary error estimators and the design of estimators with desirable properties or optimal performance relative to a family of distributions, conditioned on the actual observed sample. To summarize, the Bayesian modeling framework parameterizes the feature-label distribution by a parameter, θ, and then assigns "prior" distributions to θ that quantify the initial uncertainty we have about the distribution before observing the data. We have the option of using either a non-informative prior or supplementing the classification problem with expert information in an informative prior, to either make the problem tractable or improve performance when the sample size is small.

The observed sample is used to update the prior to a "posterior" on the distribution parameters, which represents information about the true distribution combined from the prior and data. In essence, the Bayesian model quantifies the information we have about the distribution, but only to an extent, admitting that we do not know the underlying distributions perfectly and that we can not estimate them reliably because there is not enough data. This is in contrast to the extreme approaches in classical error estimator analysis, which either assume perfect knowledge of the distribution or avoid distributional assumptions altogether in favor of distribution-free bounds.

Given a fixed sample and classifier the error estimator is simply fixed. Thus, in a Bayesian approach the sample-conditioned distribution of the true error contains the full information about error estimator accuracy, where randomness stems from the posterior uncertainty in the underlying feature-label distribution. We will consider only moments of the true error (for a fixed sample and classifier), and in particular the expectation and variance. Throughout this work, we will focus on two important Bayesian modeling frameworks: multinomial distributions with Dirichlet priors and arbitrary classification (henceforth referred to as the discrete model) and multivariate Gaussian distributions with fixed, scaled identity or arbitrary independent covariance matrices, a general class of conjugate priors and arbitrary linear classification (the Gaussian model). When arbitrary covariance matrices are used, the priors are normal-inverse-Wishart distributions. Both discrete classification and LDA in the Gaussian model are classical problems; indeed, the form of the LDA classifier and the distribution of the true error go back to [42] and [43], respectively.

1. Mean of the True Error: The Bayesian Error Estimator

Bayesian error estimation is defined to be the sample-conditioned minimum mean-square error (MMSE) estimate of the true error, which, under weak regulatory assumptions, is given by the first moment of the true error conditioned on the observed sample, where the expectation is taken over the posterior distribution of θ. It is a training data error estimator that is a function of the entire observed sample (and implicitly the designed classifier). Not only are Bayesian error estimators defined to have optimal RMS performance for a fixed sample relative to the posterior, but they enjoy several other advantages: they are unbiased, they are evaluated relative to a fixed classifier without the need for surrogate classifiers, they are independent of the feature-selection method, which is part of the classification rule, and they can be customized via the priors to target certain properties, for example, to optimize performance in moderately difficult classification problems that are typical in biomedicine with Bayes errors in the mid range. We define the Bayesian MMSE error estimator and discuss its properties in Chapter II. We also provide closed-form representations in both the discrete model (Chapter III) and Gaussian model (Chapter IV). Work in these chapters are originally from [44] and [45]. In the discrete case, we examine performance with the discrete histogram rule when compared to classical point-based estimators. Simulations in the Gaussian case are extensive, with a particular emphasis on robustness to false modeling assumptions. As a whole, this work pushes the study of error estimation ahead by placing it in a rigorous optimization setting rather than relying on *ad hoc* “intuitive” estimation rules.

Optimization is not completely new to classifier error estimation. Under the assumption that the error estimator is a linear combination of counting estimators, the weights have been optimized relative to a given feature-label distribution and classification rule [46]. Here, however, we do not wish to impose a form on the estimator, nor do we wish to assume a known feature-label distribution.

Bayesian modeling frameworks for classification are also not completely new, although we know of no work in recent years. Average Bayes error and the average true error of discrete histogram classifiers have been addressed by assuming fixed class probabilities and a uniform prior over the bin probabilities, resulting in a performance measure dependent on only sample and bin size [47, 48]. Although this work applies a prior to an uncertainty class of distributions, the average true error first averages over all samples drawn from a fixed distribution and then averages over all distributions, so that a posterior or conditioning on the sample alone were not considered.

In the 1960s, two papers made small forays in Bayesian modeling for error estimation. In [49], a Bayesian error estimator is given for the univariate Gaussian model with known covariance matrices. In [50], the problem is addressed in the multivariate Gaussian model for a particular linear classification rule based on Fishers discriminant for a common unknown covariance matrix and known class probabilities by using a specific prior on the means and the inverse of the covariance matrix. In neither case were the properties or performance of these estimators considered. Here we derive the Bayesian MMSE error estimator for an arbitrary linear classification rule in the multivariate Gaussian model for both known and unknown independent covariance matrices and both known and unknown class probabilities. We use a more general class of priors on the means and an intermediate parameter that allows us to impose structure on the covariance matrices.

Work in [51] uses a Bayesian approach to address confidence intervals for classification error rates; a beta prior is assigned to the true error directly and updated to a posterior by conditioning on the size of the sample and number of misclassified training points. One issue arises: how can we define a sensible prior on the true error? Related work by [52] considers confidence intervals, as well as the expected true error conditioned on an error estimate. There, the feature-label distribution is modeled as Gaussian or mixed-Gaussian with fixed means and scalable covariance matrices, where the Bayes error of the feature-label distribution is assigned a beta prior scaled between 0 and 0.25, indirectly corresponding to a distribution on the scale for the covariances used in the model. There is no updating to a posterior. The Bayesian framework utilized here is distinct from these works because we define a prior on the feature-label distribution itself, which is the most fundamental state of nature in a classification problem. Also, posteriors utilize the full information in the sample, not just the number of misclassified points. Furthermore, the current work will be founded on a deeper theory, including analytical representations of the MSE performance for arbitrary error estimators conditioned on the sample and the consistency of Bayesian error estimation in both the discrete and Gaussian models.

We also address practical considerations for the application of Bayesian error estimation in microarray data analysis in Chapter VII, originally from [53]. There, a method-of-moments approach is proposed to calibrate priors using features from the microarray data set that are discarded by feature selection. In addition, a toolbox of code implementing closed-form solutions for the Gaussian model with linear classifiers, as well as a Monte-Carlo approximation for the Gaussian model with non-linear classification, are provided. Bayesian error estimation is shown to have improved performance relative to classical error estimation schemes when applying the proposed calibration and estimation techniques to real biological data.

Chapter VIII presents a method of optimally calibrating arbitrary error estimators under a Bayesian framework, originally from [54], which may be very practical when closed-form representations are not available for the optimal Bayesian error estimator. Performance improvement for calibrated error estimators can be significant compared to their classical counterparts.

2. Variance of the True Error: The Sample-Conditioned MSE of the Bayesian Error Estimator

Although the Bayesian error estimator minimizing MSE has been solved in the discrete and Gaussian models, the MSE itself was not explicitly derived. This is addressed by the sample-conditioned MSE of Bayesian error estimators, which, we will show using the orthogonality principle, is equivalent to the variance of the true error conditioned on the sample. Uncertainty in the MSE is relative to the parameters in the feature-label distribution conditioned on the sample, which is fundamentally different from the classical approach relative to the sampling distribution for a fixed feature-label distribution. Under the Bayesian model, the sample conditions the uncertainty, and different samples condition it to different extents.

Consider a typical application, where we are given a specific sample to train a classifier. We are interested in estimating the error rate of our designed classifier, as well as the validity and properties of this estimate. Bayesian frameworks not only enable us to find an MMSE estimate of the classifier's true error, but also make it possible to study the performance of an error estimate conditioned on the precise sample, trained classifier and computed error estimate in hand. In contrast, classical analysis cannot be applied in this way because it only addresses average performance of a classification scheme over a sampling distribution. Thus, by taking into consideration a family of distributions and reporting the exact performance using the best knowledge available on the parameters of the distribution, the posterior probabilities, the new concept of a sample-conditioned MSE becomes a more practical measure of estimation accuracy falling out of the Bayesian approach.

Closed-form solutions for the sample-conditioned MSE in both the discrete model and Gaussian model are available in Chapter V, originally from [55], and Monte-Carlo approximation methods for the Gaussian model with non-linear classification are also discussed. Furthermore, the exact MSE for arbitrary error estimators falls out naturally. That is, if $\widehat{\varepsilon}$ is an arbitrary error estimator and $\widehat{\varepsilon}_{\mathrm{BEE}}$ is the Bayesian error estimator with correct priors, then the sample-conditioned MSE of $\widehat{\varepsilon}$ may be decomposed into the MSE of the Bayesian error estimator plus an easily calculable positive residual term:

$$\mathrm{MSE}(\widehat{\varepsilon}\,|S_n) = \mathrm{MSE}(\widehat{\varepsilon}_{\mathrm{BEE}}|S_n) + (\widehat{\varepsilon}_{\mathrm{BEE}}\,(S_n) - \widehat{\varepsilon}\,(S_n))^2,$$

where S_n is a sample of size n. This clearly illustrates the optimality of the Bayesian error estimator, and shows how the closed-form analytical results presented here may be easily applied for any error estimator under the Bayesian model.

C. Consistency and Censored Sampling

As we observe sample points, our uncertainty in the feature-label distribution should converge to a certainty on the true distribution, and in Chapter VI, which covers work originally from [56], we show that the posteriors indeed converge to delta functions on the true parameters for both the discrete and Gaussian models. Convergence may be faster with more informative priors, but convergence is assured as long as the prior has mass on any neighborhood of the true distribution.

One may then ask if classical frequentist consistency holds for Bayesian error estimators on fixed distributions, that is if the estimated error converges to the true error in some sense. Indeed, we show that it does for all true distributions in both the discrete and Gaussian models. Hence, frequentist consistency is not exclusive to distribution-free error estimators, which insist on being blind about the feature-label distribution while Bayesian error estimators confront the necessity of distributional knowledge in small-sample settings.

Not only may we observe convergence in the error estimator, but we expect the sample-conditioned RMS converges to zero as well. For example, suppose we have a sequence of sample points indexed by n, drawn from an unknown fixed distribution. Starting with the first, say, $n = 10$ points in this sequence, we may calculate the RMS of the Bayesian error estimator and find it to be relatively high. Although the prior is fixed, as we observe more sample points, the posterior distribution of the parameters will become tighter around the true distribution parameters. In this way, the Bayesian error estimate will be closer to the true error (both are changing since the sample is changing), and this will be reflected in the RMS. Thus, although the RMS is calculated for a fixed sample of size n, as we increase n by acquiring more sample points, the RMS will tend to zero if the true distribution is in the family of distributions considered in our model.

With this motivation we also prove that the sample-conditioned MSE converges to zero in probability for all distributions in both the discrete and Gaussian models as we increase sample size. This suggests an important application in censored sampling, where sample points are collected one at a time until the conditional MSE reaches an acceptable level. Finally, we provide several simulation studies on the general behavior of the conditional MSE, including practical examples with censored sampling.

CHAPTER II

MODELING[*]

A. Classification

In this section we define the classification setting. Sample spaces will be denoted with a calligraphy style, such as $\mathcal{X}$, vectors with a boldface style, such as $\mathbf{x}$, and random variables with capital letters, such as Y, or if it's also a vector, $\mathbf{X}$.

Confining ourselves to binary class labels, classification involves a feature vector $\mathbf{X}$ on a sample space $\mathcal{X}$ (two examples are a simple discrete set of bins and a continuous space $\mathcal{X} = \mathbb{R}^D$ with D features), a binary random variable Y (corresponding to class labels 0 or 1), and a function (classifier) $\psi : \mathcal{X} \rightarrow \{0, 1\}$ for which $\psi(\mathbf{X})$ is to predict Y. The joint behavior of $\mathbf{X}$ and Y is governed by a feature-label distribution $f_{\mathbf{X},Y}(\mathbf{x}, y)$, and we denote the class-conditional distributions by $f_{\mathbf{X}|Y}(\mathbf{x}|y)$. The *a priori* probabilities for the classes are defined by $c = \mathrm{P}(Y = 0)$ with $\mathrm{P}(Y = 1) = 1-c$.

The error, ε, of ψ is the probability of erroneous classification, namely, $\varepsilon = P(\psi(\mathbf{X}) \neq Y)$. This true error is relative to a feature-label distribution $f_{\mathbf{X},Y}$, and it equals the expected absolute difference between the label and classifier prediction, $E[|Y - \psi(\mathbf{X})|]$. It can also be decomposed as

$$\varepsilon = c\varepsilon^0 + (1 - c)\varepsilon^1, \tag{2.1}$$

[*] Part of this chapter is reprinted with permission from "Bayesian Minimum Mean-Square Error Estimation for Classification Error–Part I: Definition and the Bayesian MMSE Error Estimator for Discrete Classification" by L. A. Dalton and E. R. Dougherty, 2011, *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 115–129, Copyright 2011 by IEEE.

where

$$\varepsilon^0 = \mathrm{P}(\psi(\mathbf{X}) = 1|Y = 0) = \int_{\psi(\mathbf{X})=1} f_{\mathbf{X}|Y}(\mathbf{x}|0)d\mathbf{x}$$

is the probability of an element from class 0 being wrongly classified (which we may think of as the error contributed by class 0). Similarly $\varepsilon^1 = \mathrm{P}(\psi(\mathbf{X}) = 0|Y = 1)$.

In practice, the feature-label distribution is usually unknown, so that a classifier and its error are generally discovered via classification and error estimation rules. We assume a supervised sampling process modeled by n independent and identically distributed (i.i.d.) draws from the feature-label distribution. We denote a size n random sample of pairs by $S_n = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)\}$, where each pair is governed by the feature-label distribution, $f_{\mathbf{X},Y}$. A classification rule is a function on the sample that yields a good classifier, that is, a mapping of the form $\Psi : [\mathcal{X} \times \{0,1\}]^n \to \{0,1\}^{\mathcal{X}}$, where $\{0,1\}^{\mathcal{X}}$ is the family of $\{0,1\}$-valued functions on $\mathcal{X}$. Given a specific sample (realization) of S_n, we obtain a designed classifier $\psi_n = \Psi(S_n)$, where we have added a subscript n to emphasize that a classification rule is really a sequence depending on n. Similarly, we write the true error of the designed classifier as ε_n. n_0 and n_1 are the numbers of sample points from classes 0 and 1, respectively, and we denote the samples from class y by $\mathbf{x}_i^y$, $i = 1, ..., n_y$. An error estimate, $\widehat{\varepsilon}$, of ε_n is determined by an estimation rule $\Xi : [\mathcal{X} \times \{0,1\}]^n \to [0,1]$, with an estimator being a function of the random sample, $\widehat{\varepsilon} = \Xi(S_n)$.

Throughout this work, we will use four popular classification rules. The first is the discrete histogram rule for multinomial discrimination, which is essentially a majority vote in each discrete bin. For classification of continuous variables, we will use linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and k-nearest-neighbor (kNN) classification. LDA is a simple linear classification rule, often very effective in small-sample settings [57, 58]. It was developed by Wald [42] based on the “Fisher discriminant” [59], and given the form known today by Anderson [60]. LDA approaches the problem of classification by assuming that the class-conditional distributions are both Gaussian with identical full rank covariances in each class, where the Bayes optimal solution is a linear classifier. We obtain the LDA classifier by plugging in the estimated sample means for each class and a pooled sample covariance into the model. QDA is similar except the covariance of each class is not necessarily identical, and the Bayes optimal solution becomes quadratic [61]. The estimated mean and covariance of each class is substituted for the true values in the model. Finally, a kNN classifier is a non-parametric rule that classifies future points based on a majority vote from the k nearest training examples in the feature space [62].

B. Classical Classifier Error Estimators

Many commonly used training-data error estimators, including resubstitution, leave-one-out, cross-validation and bootstrap, are based on counting points. The resubstitution (also called “resub”) error estimate, $\widehat{\varepsilon}_{\text{resub}}$, is the error rate of the designed classifier on the training data.

In cross-validation (“cv”) [63, 64], the sample is randomly partitioned into k folds (subsets). At each stage of the procedure, one fold is left out, a surrogate classifier is designed on the remaining folds, and its error is estimated on the left-out fold. The cross-validation estimate, $\widehat{\varepsilon}_{\text{cv}}$, of the misclassification error of the original classifier trained on the full data set is estimated by the average surrogate errors on the left-out folds. This process may be repeated some number of times and the average taken as the cross-validation estimate. In our implementation, we use $k = 5$ folds and 5 repetitions with different partitions. The leave-one-out (“loo”) error estimate is a special case of cross-validation where $k = n$, that is where each fold contains a single point. In the case of leave-one-out, there is no randomness to fold generation because there is only one possible partition of folds; however, when $k < n$ evaluating all combinations of partitions is computationally prohibitive so in this case partitions are randomly chosen to make the estimation.

Counting estimators generally perform poorly for small samples owing to bias or variance. Resubstitution tends to be optimistically biased, often severely. Leave-one-out is close to unbiased, and more generally cross-validation is close to being unbiased if k is not too small. However, leave-one-out and cross-validation tend to have a large variance for small samples [65, 33, 66] and also to be poorly correlated with the actual error [67], the two combining to create a large RMS for small samples. For a review of error estimation performance, see [68].

The basic bootstrap zero estimator, $\widehat{\varepsilon}_{\mathrm{b0}}$, generates B bootstrap samples, $S_n^{(i)}$, $i = 1, \ldots, B$, each consisting of n equally-likely draws with replacement from the original sample, S_n [69]. Each bootstrap sample is used to design a classifier whose error is estimated by the error rate on $S_n - S_n^{(i)}$. The bootstrap zero estimator is the average of these errors for $i = 1, \ldots, B$. Like cross-validation, this error estimator is randomized because of the randomly selected bootstrap samples, and also tends to be pessimistic because the expected bootstrap sample size is only $0.632n$. In our simulations, we use the popular 0.632 bootstrap ("boot") error estimator with $B = 100$, which attempts to correct the pessimistic bias of the bootstrap zero estimator with optimistically biased resubstitution [70]. In particular, $\widehat{\varepsilon}_{\mathrm{boot}} = (1 - 0.632)\widehat{\varepsilon}_{\mathrm{resub}} + 0.632\widehat{\varepsilon}_{\mathrm{b0}}$.

Bolstered ("bol") error estimation associates a bolstering kernel (density) with each sample point to spread the mass so that a point contributes to the bolstered error estimate based on its distance from the classifier decision boundary, thereby smoothing counting estimators and balancing bias and variance. If the kernel f_i is used for point i, the bolstered error estimator is given by

$$\widehat{\varepsilon}_{\text{bol}} = \frac{1}{n}\left(\sum_{i=1}^{n_0}\int_{\psi_n(\mathbf{x})=1} f_i(\mathbf{x}-\mathbf{x}_i^0)d\mathbf{x} + \sum_{i=1}^{n_1}\int_{\psi_n(\mathbf{x})=0} f_i(\mathbf{x}-\mathbf{x}_i^1)d\mathbf{x}\right).$$

We use spherical Gaussian kernels with the same variance used for all points in a class. The kernel variances are determined by the method proposed in [71].

A classical estimator not based on counting is the plug-in rule, $\widehat{\varepsilon}_{\text{plug-in}}$, which assumes a parameterized model for the class-conditional distributions, $f_{\mathbf{X}|Y}(\mathbf{x}|y)$. The distribution parameters are estimated from the data and $\widehat{\varepsilon}_{\text{plug-in}}$ is the classifier error for the resulting feature-label distribution. The plug-in rule is the only model-dependent classical error estimator presented here, but is known to perform poorly for small samples owing to poor estimation of the model parameters even if the model assumption is accurate, and performance degrades further with model inaccuracy. Potentially, however, model-based estimation can be beneficial because the model is a form of prior knowledge that facilitates estimation (if it is accurate).

C. The Bayesian Modeling Framework

Classical error estimation methods, such as cross-validation and bootstrap, are typically heuristic counting methods that are "model-free" in the sense that their evaluation does not utilize modeling assumptions. In contrast, Bayesian error estimation uses modeling assumptions in a Bayesian framework to quantify the uncertainty in our knowledge of the feature-label distribution parameters. We begin by reviewing classical MMSE estimation in a general filtering framework.

1. Optimal MSE Estimation in the Presence of Uncertainty

We approach error estimation from a classical filtering perspective: find an

MMSE estimator of the error. To motivate our approach, consider finding a MMSE estimator (filter), $\widehat{g}(Y)$, of a function of two random variables, $g(X,Y)$, based on observing only Y; that is, minimize $\mathrm{E}_{X,Y}[|g(X,Y) - g(Y)|^2]$ over all Borel measurable functions $g(Y)$. It is well known that the optimal estimator,

$$\widehat{g} = \arg\min_{g} \mathrm{E}_{X,Y}[|g(X,Y) - g(Y)|^2] \tag{2.2}$$

is given by the conditional expectation

$$\widehat{g}(Y) = \mathrm{E}_X[g(X,Y)|Y]. \tag{2.3}$$

Moreover, $\widehat{g}(Y)$ is an unbiased estimator over the distribution, $f(x,y)$, of (X,Y), namely,

$$\mathrm{E}_{X,Y}[\widehat{g}(Y)] = \mathrm{E}_{X,Y}[g(X,Y)]. \tag{2.4}$$

The fact that $\widehat{g}(Y)$ is an unbiased MMSE estimator of $g(X,Y)$ over $f(x,y)$ does not tell us how well $\widehat{g}(Y)$ estimates $g(\bar{x},Y)$ for some specific value $X = \bar{x}$. This has to do with the expected difference

$$\begin{aligned} \mathrm{E}_Y[|g(\bar{x},Y) - \widehat{g}(Y)|^2] &= \mathrm{E}_Y\left[\left|g(\bar{x},Y) - \int g(x,Y)f(x|Y)dx\right|^2\right] \\ &= \mathrm{E}_Y\left[\left|\int g(x,Y)[f(x|Y) - \delta(x-\bar{x})]dx\right|^2\right], \end{aligned}$$

where $\delta(x)$ is the generalized delta function. Bringing the absolute value inside the integral yields

$$\mathrm{E}_Y[|g(\bar{x},Y) - \widehat{g}(Y)|^2] \leq \mathrm{E}_Y\left[\left(\int |g(x,Y)||f(x|Y) - \delta(x-\bar{x})|dx\right)^2\right],$$

which reveals that the accuracy of the estimate at a point, $\bar{x}$, depends upon the degree to which the mass of the conditional distribution for X given Y is concentrated at $\bar{x}$ on average for Y. If we replace the single random variable Y by a sequence $\{Y_n\}$ of random variables such that $f(x|Y_n) \to \delta(x - \bar{x})$ in a suitable sense (this not being the place to go into the convergence of generalized functions), then we are assured that $\widehat{g}(Y_n) - g(\bar{x}, Y_n) \to 0$ in the mean-square sense.

The conditional distribution $f(x|Y)$ characterizes uncertainty with regard to $\bar{x}$. We desire $\widehat{g}(Y)$ to estimate $g(\bar{x}, Y)$ but are uncertain of $\bar{x}$; we can obtain an unbiased MMSE estimator for $g(X, Y)$, which means good performance across all possible values of X relative to the distribution of X and Y, but the performance of that estimator for a particular value $X = \bar{x}$ depends on the concentration of the conditional mass of X relative to $\bar{x}$.

2. Definition of the Bayesian Error Estimator

We apply MMSE estimation theory to error estimation, in which case the uncertainty will manifest itself in a Bayesian framework relative to a space of feature-label distributions and random samples. The random variable X is replaced by a random variable θ governed by a specified "prior" distribution, $\pi(\theta)$, where each θ corresponds to a feature-label distribution parameterized by θ and denoted $f_\theta(\mathbf{x}, y)$. The random variable Y_n is replaced by a random sample S_n, and we set

$$g(X, Y) = \varepsilon_n(\theta, S_n),$$

which is the true error on f_θ of the designed classifier, ψ_n. In this scenario, $\widehat{g}(Y)$ becomes the error estimator

$$\widehat{\varepsilon}(S_n) = \mathrm{E}_\theta[\varepsilon_n(\theta, S_n)|S_n], \tag{2.5}$$

which we call the "Bayesian MMSE error estimator." The conditional distribution, $f(x|Y)$, becomes the "posterior" distribution $\pi^*(\theta|S_n)$, which for simplicity we often write as simply $\pi^*(\theta)$, tacitly keeping in mind conditioning on the sample. In this light, we will write the Bayesian MMSE error estimator as

$$\widehat{\varepsilon} = \mathrm{E}_{\pi^*}[\varepsilon_n(\theta)],$$

but one should keep in mind that this is short-hand for $\widehat{\varepsilon}(S_n)$ expressed in (2.5). In general we will use E_{π^*} to denote a conditional expectation given the sample.

In the case of classification, θ is a random vector composed of three parts: $\theta = [c, \theta_0, \theta_1]$ where c is the class probability for class 0, θ_0 contains the parameters of the class-0 conditional distribution and θ_1 contains the parameters of the class-1 conditional distribution. We also define $\mathbf{\Theta}_y$, $y \in \{0, 1\}$, to be the parameter space containing all permitted values for θ_y, and write the class-conditional distributions as $f_{\theta_y}(\mathbf{x}|y)$ to emphasize that they are parameterized. The marginal prior density of the class probability is denoted $\pi(c)$ and that of the class-conditional distributions are denoted $\pi(\theta_0)$ and $\pi(\theta_1)$. In using common Bayesian terminology, we also refer to these prior distributions as "prior probabilities."

As discussed in the previous section in a general setting, the Bayesian MMSE error estimate is not guaranteed to be the optimal error estimate for any particular feature-label distribution (the true error being the best estimate and perfect), but for a given sample, and assuming the parameterized model and prior probabilities, it is both optimal on average with respect to MSE (and therefore RMS) and unbiased when averaged over all parameters and samples. These implications apply for any classification rule as long as the classifier is fixed given the sample.

To facilitate analytic representations, we assume that c, θ_0 and θ_1 are all independent prior to observing the data. This assumption carries limitations. For instance, we cannot assume Gaussian distributions with the same unknown covariance for both classes, nor can we use the same parameter in both classes. However, this assumption will ultimately allow us to separate the Bayesian error estimator into components representing the error contributed by each class.

3. Prior and Posterior Probabilities

The prior probabilities, $\pi(c)$ and $\pi(\theta_y)$, may be based on objective and subjective data. It is up to the investigator to consider the nature of the problem at hand and to choose an appropriate model [72]. In genomic applications based on microarray experiments, it may be possible to define priors based on the aggregate behavior of microarray samples by incorporating data from experiments using similar instrumentation and techniques. One might also take advantage of any prior theoretical knowledge about the data. Another approach is to use objective priors, which are useful for simplifying equations or if one wishes to avoid using subjective data. Even in many classical problems, there is no universal agreement in the "right" prior to use. Based on our preceding filter analysis, we would like $\pi^*(\theta)$ to be close to $\delta(\theta - \bar{\theta})$, where $\bar{\theta}$ corresponds to the actual feature-label distribution from which the data have come, but we do not know $\bar{\theta}$ and an overzealous effort to concentrate the conditional mass at a particular value of θ can have detrimental effects if that value is far from $\bar{\theta}$.

Once $\pi(c)$, $\pi(\theta_0)$ and $\pi(\theta_1)$ have been established, we use the data to find the joint posterior density for all parameters. By the product rule,

$$\pi^*(\theta) = f\left(c, \theta_0, \theta_1 | S_n\right) = f(c|S_n, \theta_0, \theta_1) f(\theta_0|S_n, \theta_1) f(\theta_1|S_n).$$

Given n_0, c is independent from the sample values and the distribution parameters for each class. Hence,

$$f(c|S_n, \theta_0, \theta_1) = f(c|n_0, \left\{\mathbf{x}_i^0\right\}_1^{n_0}, \left\{\mathbf{x}_i^1\right\}_1^{n_1}, \theta_0, \theta_1) = f(c|n_0).$$

Given n_0, the sample and distribution parameters for class 1 are independent from the sample and distribution parameters for class 0. Thus,

$$\begin{aligned} f(\theta_0|S_n,\theta_1) &= f(\theta_0|n_0,\left\{\mathbf{x}_i^0\right\}_1^{n_0},\left\{\mathbf{x}_i^1\right\}_1^{n_1},\theta_1) \\ &= \frac{f(\theta_0,\{\mathbf{x}_i^0\}_1^{n_0}|n_0,\{\mathbf{x}_i^1\}_1^{n_1},\theta_1)}{f(\{\mathbf{x}_i^0\}_1^{n_0}|n_0,\{\mathbf{x}_i^1\}_1^{n_1},\theta_1)} \\ &= \frac{f(\theta_0,\{\mathbf{x}_i^0\}_1^{n_0}|n_0)}{f(\{\mathbf{x}_i^0\}_1^{n_0}|n_0)} \\ &= f(\theta_0|n_0,\left\{\mathbf{x}_i^0\right\}_1^{n_0}) \\ &= f(\theta_0|\left\{\mathbf{x}_i^0\right\}_1^{n_0}). \end{aligned}$$

In the last line, we assume that knowledge of n_0 is implied in the notation $\{\mathbf{x}_i^0\}_1^{n_0}$. Given n_1, analogous statements apply and

$$f(\theta_1|S_n) = f(\theta_1|n_1,\left\{\mathbf{x}_i^0\right\}_1^{n_0},\left\{\mathbf{x}_i^1\right\}_1^{n_1}) = f(\theta_1|\left\{\mathbf{x}_i^1\right\}_1^{n_1}).$$

As before, we assume that knowledge of n_1 is implied in the notation $\{\mathbf{x}_i^1\}_1^{n_1}$. Combining these results, we have that c, θ_0 and θ_1 remain independent posterior to observing the data:

$$\pi^*(\theta) = f\left(c|n_0\right) f\left(\theta_0|\left\{\mathbf{x}_i^0\right\}_1^{n_0}\right) f\left(\theta_1|\left\{\mathbf{x}_i^1\right\}_1^{n_1}\right) = \pi^*(c)\pi^*(\theta_0)\pi^*(\theta_1),$$

where $\pi^*(c)$, $\pi^*(\theta_0)$ and $\pi^*(\theta_1)$ are the marginal posterior densities for the parameters c, θ_0 and θ_1.

For the class prior probabilities, we only need to consider the size of each class:

$$\pi^*(c) = f(c|n_0) \propto \pi(c)f(n_0|c) \propto \pi(c)c^{n_0}(1-c)^{n_1}, \tag{2.6}$$

where we have taken advantage of the fact that given c, n_0 has a binomial(n, c) distribution. We present three useful models for the prior distributions of the *a priori* class probabilities: beta, uniform, and known. As we will see, the Bayesian MMSE error estimator requires only the posterior expectation, $\mathrm{E}_{\pi^*}[c]$.

If we assume the prior distribution for c is beta(α^0, α^1) distributed, then the posterior distribution for c can be simplified from (2.6). From this beta-binomial model, $\pi^*(c)$ is beta$(\alpha^0 + n_0, \alpha^1 + n_1)$ distributed:

$$\pi^*(c) = \frac{c^{\alpha^0+n_0-1}\,(1-c)^{\alpha^1+n_1-1}}{B(\alpha^0+n_0, \alpha^1+n_1)},$$

where B is the beta function. The expectation of this distribution is given by [73],

$$\mathrm{E}_{\pi^*}[c] = \frac{\alpha^0+n_0}{\alpha^0+\alpha^1+n}.$$

In the special case where we have uniform priors that assume initially all parameters between 0 and 1 are equally likely, we have $\alpha^0 = \alpha^1 = 1$, and

$$\pi^*(c) = \frac{(n+1)!}{n_0!n_1!}c^{n_0}\,(1-c)^{n_1}\,, \tag{2.7}$$

$$\mathrm{E}_{\pi^*}[c] = \frac{n_0+1}{n+2}. \tag{2.8}$$

Finally, to apply a known prior we define the parameter c to have a trivial sample space with one point. Then, the expectation is simply the known value for c, regardless of the data. Note if stratified sampling is used, c is essentially given in the data and $\mathrm{E}_{\pi^*}[c] = \frac{n_0}{n}$.

When finding the posterior probabilities for the class-conditional distribution parameters, we need only consider the sample points from the corresponding class. We find $\pi^*(\theta_y)$ using Bayes' rule:

$$\begin{aligned}\pi^*(\theta_y) &= f(\theta_y|\,\{\mathbf{x}_i^y\}_1^{n_y})\\ &\propto \pi(\theta_y)f(\{\mathbf{x}_i^y\}_1^{n_y}\,|\theta_y)\\ &= \pi(\theta_y)\prod_{i=1}^{n_y} f_{\theta_y}(\mathbf{x}_i^y|y),\end{aligned} \tag{2.9}$$

where the constant of proportionality can be found by normalizing the integral of $\pi^*(\theta_y)$ to 1. The term $f(\{\mathbf{x}_i^y\}_1^{n_y} | \theta_y)$ is called the "likelihood function."

Although we call $\pi(\theta_y)$ the "prior probabilities," they are not required to be valid density functions. In particular, the priors are called "improper" if the integral of $\pi(\theta_y)$ is infinite, i.e., if $\pi(\theta_y)$ induces a σ-finite measure but not a finite probability measure. Such priors can be used to represent uniform weight for all parameters in an unbounded range, rather than truncating the range of each parameter to a finite range. When improper priors are used, Bayes' rule does not apply so we take (2.9) as a definition, but normalize the posterior distributions to have a unit integral as usual. The use of improper priors for error estimation is justified in Section II.C.5. Whether one decides it is appropriate to use improper priors or not, in all cases it is mandatory that the posterior is a valid probability density. If the prior is proper, then the posterior is also guaranteed to be proper.

4. Evaluating the Bayesian MMSE Error Estimator

Since the underlying feature-label distribution is parameterized by θ, the true error of ψ_n can be written as,

$$\varepsilon_n(\theta, S_n) = c\varepsilon_n^0(\theta_0, S_n) + (1-c)\varepsilon_n^1(\theta_1, S_n),$$

where we have explicitly indicated the dependence of ε_n and ε_n^y on the distribution parameters and the sample/classifier. Owing to the posterior independence between c, θ_0 and θ_1, and since ε_n^y is a function of θ_y only, the Bayesian MMSE error estimator can be expressed as

$$\begin{aligned}\widehat{\varepsilon}\,(S_n) &= \mathrm{E}_\theta[\varepsilon_n(\theta, S_n)|S_n] \\ &= \mathrm{E}_\theta[c\varepsilon_n^0\,(\theta_0, S_n) + (1-c)\varepsilon_n^1\,(\theta_1, S_n)\,|S_n] \\ &= \mathrm{E}_c[c|S_n]\mathrm{E}_{\theta_0}[\varepsilon_n^0\,(\theta_0, S_n)\,|S_n] + (1 - \mathrm{E}_c[c|S_n])\,\mathrm{E}_{\theta_1}[\varepsilon_n^1\,(\theta_1, S_n)\,|S_n].\end{aligned}$$

We apply the shorthand notation introduced earlier in the definition of the Bayesian error estimator:

$$\begin{aligned}\widehat{\varepsilon} &= \mathrm{E}_{\pi^*}[c]\mathrm{E}_{\pi^*}[\varepsilon_n^0(\theta_0)] + (1 - \mathrm{E}_{\pi^*}[c])\,\mathrm{E}_{\pi^*}[\varepsilon_n^1(\theta_1)] \\ &= \mathrm{E}_{\pi^*}[c]\widehat{\varepsilon}^0 + (1 - \mathrm{E}_{\pi^*}[c])\,\widehat{\varepsilon}^1, \end{aligned} \tag{2.10}$$

where we have defined $\widehat{\varepsilon}^y = \mathrm{E}_{\pi^*}[\varepsilon_n^y(\theta_y)]$, which may be viewed as the posterior expectation for the true error contributed by class y. Note that we have suppressed dependence on the sample in several quantities to avoid cumbersome notation, for instance we sometimes write $\varepsilon_n(\theta)$ instead of $\varepsilon_n(\theta, S_n)$, $\varepsilon_n^y(\theta_y)$ instead of $\varepsilon_n^y(\theta_y, S_n)$, $\widehat{\varepsilon}$ instead of $\widehat{\varepsilon}(S_n)$, and $\widehat{\varepsilon}^y$ instead of $\widehat{\varepsilon}^y(S_n)$. However, the reader should keep in mind that these quantities are always functions of the sample. If any of the prior probabilities are improper, this is called the "generalized Bayesian MMSE error estimator." Also, the Bayesian error estimator is a training data error estimator, meaning that no sample points are held out for error estimation and the entire sample set is used to estimate the true error.

$\mathrm{E}_{\pi^*}[c]$ depends on our prior assumptions about the class probability. For example, if we assume flat priors for c and apply (2.8), then

$$\widehat{\varepsilon} = \frac{n_0+1}{n+2}\widehat{\varepsilon}^0 + \frac{n_1+1}{n+2}\widehat{\varepsilon}^1.$$

For a fixed classifier, ε_n^y is a deterministic function of θ_y and

$$\widehat{\varepsilon}^y = \mathrm{E}_{\pi^*}[\varepsilon_n^y(\theta_y)] = \int_{\boldsymbol{\Theta}_y} \varepsilon_n^y(\theta_y)\pi^*(\theta_y)d\theta_y. \tag{2.11}$$

The solution to this integral depends on the classifier/classification rule through $\varepsilon_n^y(\theta_y)$ and the Bayesian model and posterior through $\pi^*(\theta_y)$.

We will derive closed-form solutions for the discrete and Gaussian models in Chapters III and IV, respectively. When closed-form solutions are not available, $\widehat{\varepsilon}^y$ may be approximated from (2.11) using Monte-Carlo integral approximation as discussed in Chapter VII. Once $\widehat{\varepsilon}^y$ has been found for each class, we find $\mathrm{E}_{\pi^*}[c]$ according to our prior model for c and refer to (2.10) for the complete Bayesian error estimator.

5. On Improper Priors

The Bayesian community is currently divided on the validity of improper priors. A notable example suggesting that improper priors should be avoided completely comes from [74], which presents “marginalization paradoxes” and points a finger at the use of improper priors as the cause. At the same time, these claims and demonstrations have been refuted by many, for example Jaynes’ response in [75] explains that there is no marginalization paradox, and that the controversy stems from an improper use of notation and failure to capture what information is known at different stages of a problem.

In many cases, the posterior distribution (or a Bayesian estimate) obtained from an improper prior is equivalent to a limit of posterior distributions (or Bayesian estimates) from some sequence of proper prior distributions [76, 77, 78], however extra care must be taken to ensure that the resulting posterior density can be normalized and makes sense. Here, we justify the use of improper priors for error estimation, where we are primarily interested in evaluating $\widehat{\varepsilon}^y = \mathrm{E}_{\pi^*}[\varepsilon_n^y(\theta_y)]$. Using improper priors directly amounts to evaluating the ratio

$$\mathrm{E}_{\pi^*}[\varepsilon_n^y(\theta_y)] = \frac{\int_{\boldsymbol{\Theta}_y} \varepsilon_n^y(\theta_y)\pi(\theta_y) \prod_{i=1}^{n_y} f_{\mu_y,\Sigma_y}(\mathbf{x}_i^y|y)d\theta_y}{\int_{\boldsymbol{\Theta}_y} \pi(\theta_y) \prod_{i=1}^{n_y} f_{\mu_y,\Sigma_y}(\mathbf{x}_i^y|y)d\theta_y}.$$

However, a more mathematically sound approach is to use a sequence of proper priors indexed by positive integers k, $\pi_k(\theta_y)$, that converge in some sense to the improper priors $\pi(\theta_y)$. We would then evaluate the limit of the ratio,

$$\lim_{k\to\infty} \frac{\int_{\boldsymbol{\Theta}_y} \varepsilon_n^y(\theta_y)\pi_k(\theta_y) \prod_{i=1}^{n_y} f_{\mu_y,\Sigma_y}(\mathbf{x}_i^y|y)d\theta_y}{\int_{\boldsymbol{\Theta}_y} \pi_k(\theta_y) \prod_{i=1}^{n_y} f_{\mu_y,\Sigma_y}(\mathbf{x}_i^y|y)d\theta_y}.$$

Suppose there exists a sequence of proper priors, $\pi_k(\theta_y) = A_k\pi(\theta_y)\mathbf{I}_{\theta_y\in B_k}$, where A_k is the normalization constant (which is always finite) and B_k is a sequence of bounded, increasing sets that cover the sample space. For example, with a flat prior over a parameter space $\boldsymbol{\Theta}_y = \mathbb{R}^D$, we may choose B_k to be the open ball centered at zero with radius k. Then the correct approach to find a Bayesian error estimator leads to

$$\lim_{k\to\infty} \frac{\int_{\boldsymbol{\Theta}_y} \varepsilon_n^y(\theta_y)\pi(\theta_y)I_{\theta_y\in B_k} \prod_{i=1}^{n_y} f_{\mu_y,\Sigma_y}(\mathbf{x}_i^y|y)d\theta_y}{\int_{\boldsymbol{\Theta}_y} \pi(\theta_y)I_{\theta_y\in B_k} \prod_{i=1}^{n_y} f_{\mu_y,\Sigma_y}(\mathbf{x}_i^y|y)d\theta_y}.$$

As long as the limits for the numerator and denominator exist and the denominator is non-zero with a non-zero limit (both are verified if the posterior obtained from our improper priors can be normalized), we may take the limit in the numerator and denominator separately. In addition, the Monotone Convergence Theorem applies since all terms are positive and the integrands are increasing with respect to k. Once we bring the limits inside the integrals, the indicator functions are removed and we obtain exactly the same result as we would by starting with improper priors, with the added caution to verify that the posterior densities can be normalized.

CHAPTER III

BAYESIAN MMSE ERROR ESTIMATION—DISCRETE CLASSIFICATION*

A. The Discrete Model

We next illustrate the Bayesian error estimator applied to the discrete classification setting. Discrete classification, also called categorical classification or multinomial discrimination [33, 79, 80, 81] is very important in several applications, particularly in biology, economics, psychology, and social science [80]. In a general discrete classification problem, the sample space is discrete with b bins. Let p_i and q_i, $i = 1, ..., b$, be the class-conditional probabilities for each bin for class 0 and 1, respectively. Similarly, let U_i and V_i, $i = 1, ..., b$, be the number of samples observed in each bin for class 0 and 1, respectively. The U_i's and V_i's are outcomes of a multinomial sampling distribution with parameters $\{p_i\}_1^b$ and $\{q_i\}_1^b$, respectively. The class sizes are $n_0 = \sum_{i=1}^b U_i$ and $n_1 = \sum_{i=1}^b V_i$. A classifier in the discrete setting assigns each bin to a class, so $\psi_n : \{1, \ldots, b\} \to \{0, 1\}$. This classifier may be trained using the discrete histogram classification rule but this is not necessary.

The true error of a classifier ψ_n is given by $\varepsilon_n = c\varepsilon_n^0 + (1-c)\varepsilon_n^1$ from (2.1), where

$$\varepsilon_n^0 = \sum_{i=1}^{b} p_i \mathbf{I}_{\psi_n(i)=1} \quad \text{and} \quad \varepsilon_n^1 = \sum_{i=1}^{b} q_i \mathbf{I}_{\psi_n(i)=0}, \tag{3.1}$$

and where $\mathbf{I}_E$ is an indicator function equal to one if E is true and zero otherwise.

*Part of this chapter is reprinted with permission from "Bayesian Minimum Mean-Square Error Estimation for Classification Error–Part I: Definition and the Bayesian MMSE Error Estimator for Discrete Classification" by L. A. Dalton and E. R. Dougherty, 2011, *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 115–129, Copyright 2011 by IEEE.

Many classical error estimators can be simplified considerably for the discrete problem. For one, resubstitution is the same as the plug-in rule. More details on the discrete error estimation problem are available in [26] and [82].

We consider two models. The first is a simple problem where we derive step-by-step the Bayesian error estimator for uniform priors and $b = 2$ bins. The second generalizes the previous model by considering an arbitrary number of bins with Dirichlet priors. When no information is available about the bin probabilities, the second model can be applied with uniform priors as a special case.

1. Uniform Priors and $b = 2$

In a binary problem with $b = 2$, define p to be the probability for bin 1 in class 0 and let q be the probability for bin 1 in class 1, i.e., $p = p_1 = 1 - p_2$ and $q = q_1 = 1 - q_2$. In this case, p and q completely model the distributions, so we define $\theta_0 = p$ and $\theta_1 = q$ with the parameter spaces $\mathbf{\Theta}_y = [0, 1]$.

If we assign uniform prior distributions for p and q, the posterior probabilities are straightforward to find using a method analogous to that used to find (2.7). In particular, we have that $\pi^*(p)$ and $\pi^*(q)$ are beta distributions:

$$\pi^*(p) = \frac{(n_0+1)!}{U_1!U_2!} p^{U_1} \left(1-p\right)^{U_2} \quad \text{and} \quad \pi^*(q) = \frac{(n_1+1)!}{V_1!V_2!} q^{V_1} \left(1-q\right)^{V_2}.$$

To find the Bayesian MMSE error estimator, we simplify the posterior expected true error contributed by each class from (2.11). For class 0 we obtain

$$\begin{aligned}\widehat{\varepsilon}^0 &= \int_0^1 \varepsilon_n^0(p)\pi^*(p)dp \\ &= \frac{(n_0+1)!}{U_1!U_2!}\mathbf{I}_{\psi_n(1)=1}\int_0^1 p^{U_1+1}\left(1-p\right)^{U_2}dp + \frac{(n_0+1)!}{U_1!U_2!}\mathbf{I}_{\psi_n(2)=1}\int_0^1 p^{U_1}\left(1-p\right)^{U_2+1}dp \\ &= \frac{U_1+1}{n_0+2}\mathbf{I}_{\psi_n(1)=1} + \frac{U_2+1}{n_0+2}\mathbf{I}_{\psi_n(2)=1}.\end{aligned}$$

Similarly, for class 1,

$$\widehat{\varepsilon}^1 = \frac{V_1+1}{n_1+2}\mathbf{I}_{\psi_n(1)=0} + \frac{V_2+1}{n_1+2}\mathbf{I}_{\psi_n(2)=0}.$$

Combining these results using (2.10), we obtain the Bayesian MMSE error estimate. Note these results apply for any fixed discrete classification rule.

2. Dirichlet Priors and Arbitrary Bin Size

We now extend the result in the previous section by applying a more general conjugate prior to the problem with an arbitrary number of bins. The bin probabilities, $\{p_i\}_1^b$ and $\{q_i\}_1^b$, are both members of the "standard $(b-1)$-simplex," which is the set of all sequences of length b whose terms are nonnegative and add to one. Define the parameters for each class to contain all but one bin probability, i.e., $\theta_0 = [p_1, p_2, \ldots, p_{b-1}]$ and $\theta_1 = [q_1, q_2, \ldots, q_{b-1}]$. With this model, each parameter space is defined as the set of all valid bin probabilities, for example $[p_1, p_2, \ldots, p_{b-1}] \in \mathbf{\Theta}_0$ if and only if $0 \leq p_i \leq 1$ for $i = 1, \ldots, b-1$ and $\sum_{i=1}^{b-1} p_i \leq 1$. Given θ_0, the last bin probability is defined by $p_b = 1 - \sum_{i=1}^{b-1} p_i$.

The conjugate prior for the multinomial distribution used to model the bin probabilities in either class is given by a generalized beta distribution known as the Dirichlet distribution:

$$\pi(\theta_0) \propto \prod_{i=1}^{b} p_i^{\alpha_i^0 - 1} \quad \text{and} \quad \pi(\theta_1) \propto \prod_{i=1}^{b} q_i^{\alpha_i^1 - 1},$$

where we require the hyperparameters α_i^y, $i = 1, \ldots, b$, to satisfy $\alpha_i^y > 0$. If $\alpha_i^y = 1$ for all bins, $i = 1, \ldots, b$, and both classes, $y = 0$ and $y = 1$, we obtain uniform priors. Furthermore, the Dirichlet prior for class y is mathematically equivalent to a likelihood resulting from $\sum_{i=1}^{b} \alpha_i^y$ class y observations, with α_i^y observations in bin i. As we increase a specific α_i^y, it is as if we bias the corresponding bin with α_i^y samples from the corresponding class before ever observing the data.

Rather than working with the bin probabilities directly, it is easier to solve the integrals with ordered bin dividers. In other words, we define the linear one-to-one change of variables,

$$a^0_{(i)} = \begin{cases} 0 & \text{if } i = 0, \\ \sum_{k=1}^{i} p_k & \text{if } i = 1, \ldots, b-1, \\ 1 & \text{if } i = b, \end{cases}$$

and define $a^1_{(i)}$ similarly using q_k. We use the subscript (i) instead of just i to emphasize that the $a^y_{(i)}$ are ordered so that $0 \leq a^y_{(1)} \leq \ldots \leq a^y_{(b-1)} \leq 1$. The bin probabilities are determined by the partitions the $a^y_{(i)}$ make in the interval $[0, 1]$, i.e., $p_i = a^0_{(i)} - a^0_{(i-1)}$ for all i. The Jacobean determinant of this transformation is one, so integrals over the p_i may be converted to integrals over the $a^0_{(i)}$ by simply replacing p_i with $a^0_{(i)} - a^0_{(i-1)}$ and defining the new integration region characterized by $0 \leq a^y_{(1)} \leq \ldots \leq a^y_{(b-1)} \leq 1$. To find the posterior probability of parameters θ_0 and θ_1 and the Bayesian MMSE error estimator itself, we will use the following lemma.

Lemma 1. *Let $b \geq 2$ be an integer and let $U_i > -1$ be real numbers for $i = 1, \ldots, b$. Define $a_{(0)} \equiv 0$ and $a_{(b)} \equiv 1$. Then,*

$$\int_0^1 \int_0^{a_{(b-1)}} \cdots \int_0^{a_{(2)}} \prod_{i=1}^{b} \left(a_{(i)} - a_{(i-1)}\right)^{U_i} da_{(1)} \ldots da_{(b-2)} da_{(b-1)} = \frac{\prod_{k=1}^{b} \Gamma\left(U_k + 1\right)}{\Gamma\left(\sum_{i=1}^{b} U_i + b\right)},$$

where Γ is the gamma function.

Proof. Define M to be the value of this integral. Note that,

$$M = \int_0^1 \left(1 - a_{(b-1)}\right)^{U_b} \cdots \int_0^{a_{(3)}} \left(a_{(3)} - a_{(2)}\right)^{U_3} \int_0^{a_{(2)}} \left(a_{(2)} - a_{(1)}\right)^{U_2} \left(a_{(1)}\right)^{U_1} da_{(1)} \ldots da_{(b-1)}. \tag{3.2}$$

For $k = 2, \ldots, b$ also define:

$$N_k \equiv \int_0^{a_{(k)}} \left(a_{(k)} - a_{(k-1)}\right)^{U_k} \left(a_{(k-1)}\right)^{\sum_{i=1}^{k-1} U_i + k - 2} da_{(k-1)}.$$

Substitute $x = a_{(k-1)}/a_{(k)}$ and note that,

$$\begin{aligned} N_k &= \int_0^1 \left(a_{(k)} - a_{(k)}x\right)^{U_k} \left(a_{(k)}x\right)^{\sum_{i=1}^{k-1} U_i + k - 2} a_{(k)} dx \\ &= \left(a_{(k)}\right)^{\sum_{i=1}^{k} U_i + (k+1) - 2} \int_0^1 (1-x)^{U_k} (x)^{\sum_{i=1}^{k-1} U_i + k - 2} dx \\ &= \left(a_{(k)}\right)^{\sum_{i=1}^{k} U_i + (k+1) - 2} B\left(\sum_{i=1}^{k-1} U_i + k - 1, U_k + 1\right), \end{aligned}$$

where the last integral is essentially the definition of the beta function, B.

In (3.2), the innermost integral is N_2. After evaluating it and pulling out the constant beta function, the new innermost integral is exactly N_3. Using induction, we repeat this for all $b - 1$ integrals to obtain the desired result:

$$\begin{aligned} M &= \prod_{k=2}^{b} B\left(\sum_{i=1}^{k-1} U_i + k - 1, U_k + 1\right) \\ &= \prod_{k=2}^{b} \frac{\Gamma\left(\sum_{i=1}^{k-1} U_i + k - 1\right) \Gamma\left(U_k + 1\right)}{\Gamma\left(\sum_{i=1}^{k} U_i + k\right)} \\ &= \frac{\prod_{k=1}^{b} \Gamma\left(U_k + 1\right)}{\Gamma\left(\sum_{i=1}^{b} U_i + b\right)}. \end{aligned}$$

□

We focus on the posterior of class 0 first. $f_{\theta_0}(\mathbf{x}_i^0|0)$ is equal to the bin probability corresponding to bin $\mathbf{x}_i^0$, thus we have the likelihood function,

$$\prod_{i=1}^{n_0} f_{\theta_0}(\mathbf{x}_i^0|0) = \prod_{i=1}^{b} p_i^{U_i} = \prod_{i=1}^{b} \left(a_{(i)}^0 - a_{(i-1)}^0\right)^{U_i}.$$

The posterior parameter density is still proportional to the product of the bin prob-

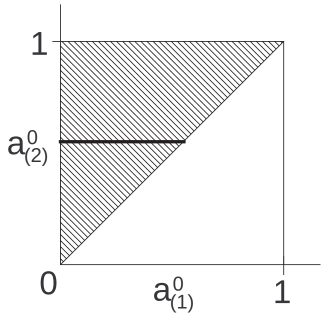


Fig. 3. Region of integration in $\mathrm{E}_{\pi^*}[\varepsilon^0_n(\theta_0)]$ for $b = 3$.

abilities:

$$\pi^*(\theta_0) \propto \prod_{i=1}^{b} p_i^{\alpha_i^0+U_i-1} = \prod_{i=1}^{b} \left(a^0_{(i)} - a^0_{(i-1)}\right)^{\alpha_i^0+U_i-1}. \tag{3.3}$$

The posterior for θ_1 is similar. Thus, $\pi^*(\theta_0)$ and $\pi^*(\theta_1)$ are also Dirichlet distributions with updated hyperparameters, $\alpha_i^0 + U_i$ and $\alpha_i^1 + V_i$ [83].

To find the proportionality constant in $\pi^*(\theta_0)$, we must be careful with the region of integration to force the bin dividers to be ordered. An example of this region is shown in Fig. 3 for $b = 3$. We proceed by letting the last bin divider, $a^0_{(b-1)}$, vary freely between 0 and 1. Once this divider is fixed, the next smallest bin divider can vary from 0 to $a^0_{(b-1)}$, and this continues until we reach the first bin divider, $a^0_{(1)}$, which can vary from 0 to $a^0_{(2)}$. The proportionality constant for $\pi^*(\theta_0)$ can be found by applying Lemma 1 to (3.3) to obtain

$$\pi^*(\theta_0) = \frac{\Gamma\left(n_0 + \sum_{i=1}^{b} \alpha_i^0\right)}{\prod_{k=1}^{b} \Gamma\left(U_k + \alpha_k^0\right)} \prod_{i=1}^{b} \left(a^0_{(i)} - a^0_{(i-1)}\right)^{U_i+\alpha_i^0-1}. \tag{3.4}$$

The Bayesian MMSE error estimate contributed by class 0 is found from (2.11):

$$\widehat{\varepsilon}^0 = \int_0^1 \ldots \int_0^{a_{(2)}} \varepsilon^0_n(\theta_0)\pi^*(\theta_0) da^0_{(1)} \ldots da^0_{(b-1)}.$$

The true error for class 0, $\varepsilon_n^0(\theta_0)$, is given by (3.1), and the posterior parameter density, $\pi^*(\theta_0)$, has been found in (3.4). Using the same region of integration as before, with an example illustrated in Fig. 3 for $b = 3$, the true error for class 0 is

$$\begin{aligned}
\widehat{\varepsilon}^0 &= \int_0^1 \int_0^{a^0_{(b-1)}} \cdots \int_0^{a^0_{(3)}} \int_0^{a^0_{(2)}} \varepsilon_n^0(\theta_0)\pi^*(\theta_0) da^0_{(1)} da^0_{(2)} \ldots da^0_{(b-2)} da^0_{(b-1)} \\
&= \int_0^1 \cdots \int_0^{a^0_{(2)}} \left(\sum_{j=1}^{b} \left(a^0_{(j)} - a^0_{(j-1)}\right) \mathbf{I}_{\psi_n(j)=1} \right) \\
&\quad \times \left(\frac{\Gamma\left(n_0 + \sum_{i=1}^b \alpha_i^0\right)}{\prod_{k=1}^b \Gamma\left(U_k + \alpha_k^0\right)} \prod_{i=1}^{b} \left(a^0_{(i)} - a^0_{(i-1)}\right)^{U_i + \alpha_i^0 - 1} \right) da^0_{(1)} \ldots da^0_{(b-1)} \\
&= \sum_{j=1}^{b} \frac{\Gamma\left(n_0 + \sum_{i=1}^b \alpha_i^0\right)}{\prod_{k=1}^b \Gamma\left(U_k + \alpha_k^0\right)} \mathbf{I}_{\psi_n(j)=1} \\
&\quad \times \int_0^1 \cdots \int_0^{a^0_{(2)}} \prod_{i=1}^{b} \left(a^0_{(i)} - a^0_{(i-1)}\right)^{U_i + \alpha_i^0 - 1 + \delta_{i-j}} da^0_{(1)} \ldots da^0_{(b-1)},
\end{aligned}$$

where δ_i is the Kronecker delta function, equal to 1 if $i = 0$ and 0 otherwise. These integrals can also be solved using Lemma 1. We obtain

$$\begin{aligned}
\widehat{\varepsilon}^0 &= \sum_{j=1}^{b} \frac{\Gamma\left(n_0 + \sum_{i=1}^b \alpha_i^0\right)}{\prod_{k=1}^b \Gamma\left(U_k + \alpha_k^0\right)} \mathbf{I}_{\psi_n(j)=1} \frac{\prod_{k=1}^b \Gamma\left(U_k + \alpha_k^0 + \delta_{k-j}\right)}{\Gamma\left(\sum_{i=1}^b (U_i + \alpha_i^0 - 1 + \delta_{i-j}) + b\right)} \\
&= \sum_{j=1}^{b} \frac{\Gamma\left(n_0 + \sum_{i=1}^b \alpha_i^0\right)}{\prod_{k=1}^b \Gamma\left(U_k + \alpha_k^0\right)} \mathbf{I}_{\psi_n(j)=1} \frac{\left(U_j + \alpha_j^0\right) \prod_{k=1}^b \Gamma\left(U_k + \alpha_k^0\right)}{\left(n_0 + \sum_{i=1}^b \alpha_i^0\right) \Gamma\left(n_0 + \sum_{i=1}^b \alpha_i^0\right)} \\
&= \sum_{j=1}^{b} \frac{U_j + \alpha_j^0}{n_0 + \sum_{i=1}^b \alpha_i^0} \mathbf{I}_{\psi_n(j)=1}.
\end{aligned} \tag{3.5}$$

The proof for class 1 is identical, except we replace $a^0_{(k)}$ with $a^1_{(k)}$, U_i with V_i and n_0 with n_1. In the end we obtain

$$\widehat{\varepsilon}^1 = \sum_{j=1}^{b} \frac{V_j + \alpha_j^1}{n_1 + \sum_{i=1}^b \alpha_i^1} \mathbf{I}_{\psi_n(j)=0}. \tag{3.6}$$

In the special case where we have uniform priors for the bin probabilities ($\alpha_i^y = 1$

for all i and y) and uniform c, the Bayesian MMSE error estimate is:

$$\widehat{\varepsilon} = \frac{n_0+1}{n+2}\left(\sum_{i=1}^{b} \frac{U_i+1}{n_0+b}\mathbf{I}_{\psi_n(i)=1}\right) + \frac{n_1+1}{n+2}\left(\sum_{i=1}^{b} \frac{V_i+1}{n_1+b}\mathbf{I}_{\psi_n(i)=0}\right).$$

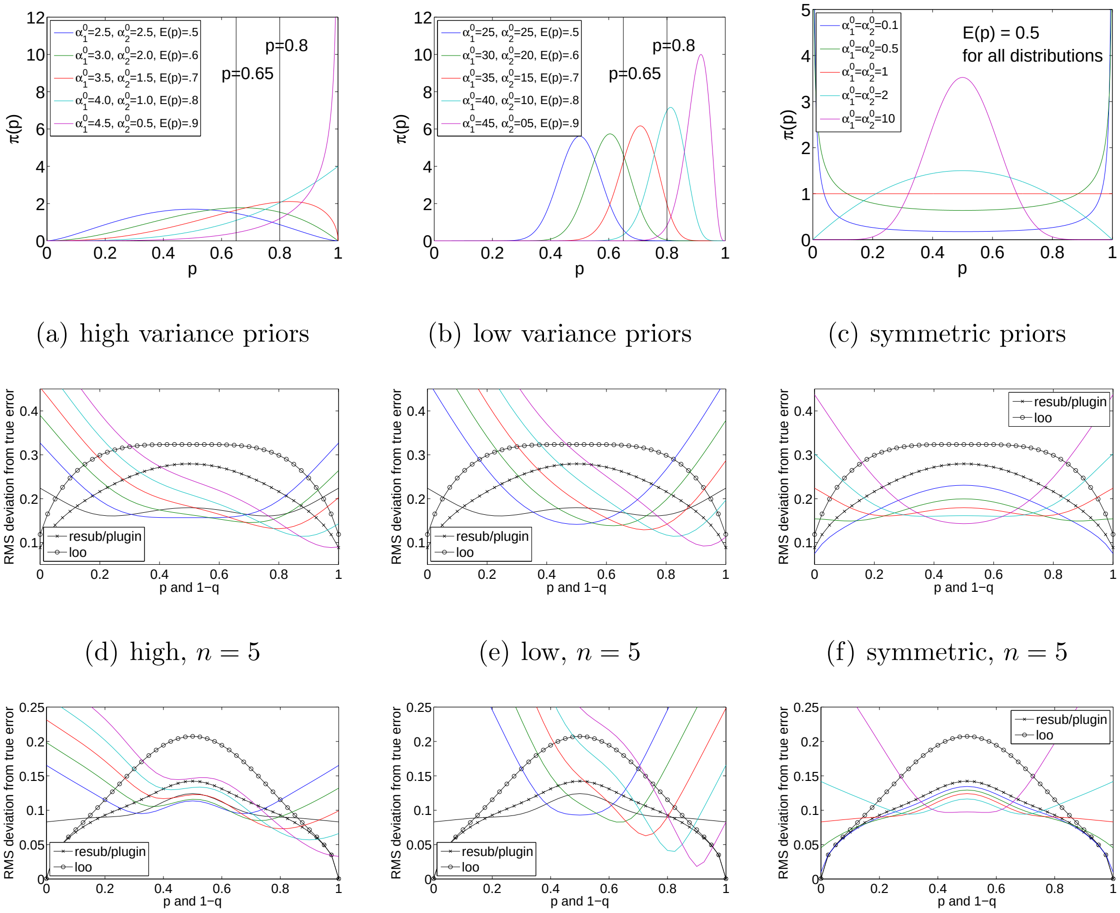
These results agree with the Bayesian error estimator for uniform priors and $b = 2$ found in the previous section. Also, comparing the true error in (3.1) with the above equation, we may view this Bayesian MMSE error estimator as a plug-in rule with $\frac{U_i+1}{n_0+b}$ as the estimate for p_i, $\frac{V_i+1}{n_1+b}$ as the estimate for q_i, and $\frac{n_0+1}{n+2}$ as the estimate of c.

B. Performance and Robustness

This section includes three simulation studies on Bayesian error estimators for discrete models. In the first study, we observe the performance of Bayesian error estimators for two bins and different beta prior distributions for the bin probabilities. By studying beta priors that target specific values for p and q, we will observe the benefits of informative priors and assess the robustness of discrete Bayesian error estimators to poor prior distribution modeling. In the second study, we present performance of Bayesian error estimators with uniform priors for an arbitrary number of bins. These simulations show how and when Bayesian error estimators improve on the resubstitution and leave-one-out error estimators, especially as we increase the number of bins. Finally, we conclude this section with performance results with respect to bias and deviation variance.

1. Beta Priors and $b = 2$

In each simulation, we fix the bin size to $b = 2$, the true distribution ($c = 0.5$, $p \in [0, 1]$ and $q = 1 - p$) and the sample size. The Bayes error, or the optimal true error obtained from the optimal classifier (not to be confused with Bayesian error



(a) high variance priors

(b) low variance priors

(c) symmetric priors

(d) high, $n = 5$

(e) low, $n = 5$

(f) symmetric, $n = 5$

(g) high, $n = 20$

(h) low, $n = 20$

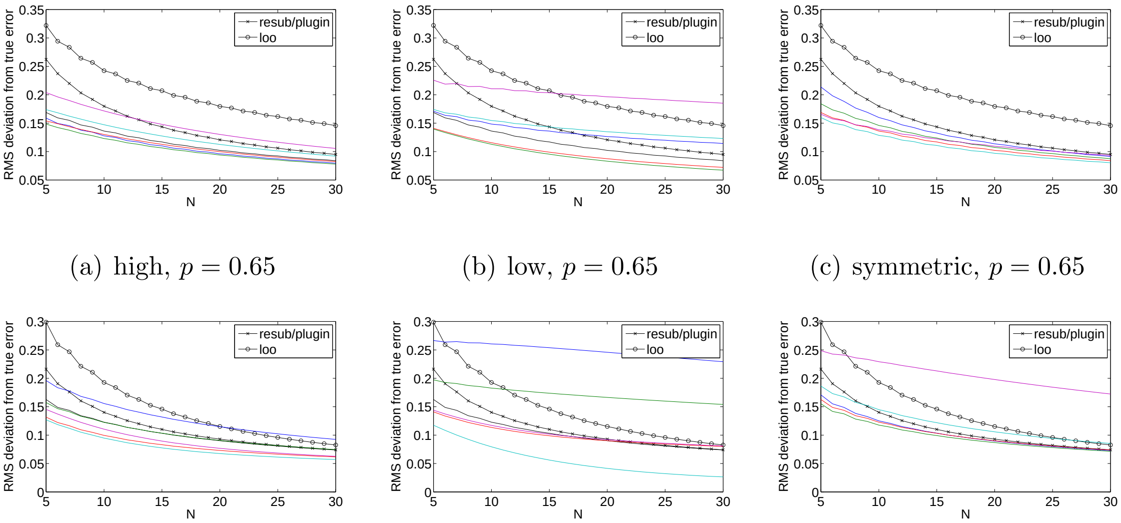
(i) symmetric, $n = 20$

Fig. 4. High-variance, low-variance and symmetric priors centered at $p = 0.5$ versus p along with RMS deviation from true error for discrete classification versus p ($b = 2$, $c = 0.5$). Colored graphs represent Bayesian error estimators with different beta priors, which are color coded with the distributions labeled and shown at the top of each column.

estimators), is simply $\min(p, q)$. We generate a random non-stratified sample by first determining the sample size for each class using a binomial(n, c) experiment and then assign each sample point a bin number according to the distribution of its class. The sample is then used to train a histogram classifier where the class assigned to each bin is determined by a majority vote. The true error is calculated using the known distribution parameters, where the same sample used for classifier design are used to find resubstitution, leave-one-out, and several Bayesian error estimates for the designed classifier. With the true error and estimated error found, we finally have the squared deviation of each estimate with respect to the true error. This process is repeated 10,000,000 times to find a Monte-Carlo approximation for the RMS deviation from true error for each error estimator. All results are presented in Figs. 4 and 5.

For all Bayesian error estimators, c is assumed to have a uniform prior. In each simulation, all Bayesian error estimators utilize slightly different priors, defined by different hyperparameters α_i^y. Since we always set $q = 1 - p$, given fixed priors for p we choose priors for q that are the same but flipped about 0.5, i.e., $\alpha_1^1 = \alpha_2^0$ and $\alpha_2^1 = \alpha_1^0$ so that $\mathrm{E}_\pi[q] = 1 - \mathrm{E}_\pi[p]$. The top row of Fig. 4 shows several beta distributions used as priors for p, each defining a different Bayesian error estimator. Part (a) contains five beta distributions representing priors with varying means ($\mathrm{E}_\pi[p] = 0.5$, 0.6, 0.7, 0.8 and 0.9) and relatively high variance. Part (b) is similar, except with tighter variances. Part (c) shows several symmetric beta distributions centered at $p = 0.5$ (including the uniform prior in red) with varying degrees of bias toward middle versus edge values of p. In all priors in part (c), the α_i^y are equal for all i.

The graphs below the priors in Fig. 4 present RMS deviation from true error versus the true distributions, p, for the error estimators corresponding to these priors. Figure 5 is similar, but provides performance versus sample size. In all RMS graphs of



(a) high, $p = 0.65$ (b) low, $p = 0.65$ (c) symmetric, $p = 0.65$

(d) high, $p = 0.8$ (e) low, $p = 0.8$ (f) symmetric, $p = 0.8$

Fig. 5. RMS deviation from true error for discrete classification versus sample size ($b = 2$, $c = 0.5$). Colored graphs represent Bayesian error estimators with different beta priors, which are color coded with the distributions labeled and shown at the top of each column.

both figures, each point represents a fixed sample size and true distribution (c, p and q). The graphs are color coded to aid in matching priors to their corresponding error estimator, for example if we pick the high-variance red prior in the upper left graph of Fig. 4, the performance of the Bayesian error estimator using this prior is also shown in red in all the graphs in the same column. For comparison, the resubstitution, leave-one-out, and Bayesian error estimator with uniform priors are also included in all RMS graphs.

In the first row of Fig. 5, we fix the true distributions at $p = 0.65$ and $q = 0.35$ and observe performance as we increase the sample size. Similarly, in the second row the true distributions are fixed at $p = 0.8$ and $q = 0.2$. Naturally, these simulations show that priors with a high density around the true distributions have better performance and tend to converge more quickly to the true error. For example, in Fig 4(b) the light blue prior (with $\mathrm{E}_\pi[p] = 0.8$) matches the distribution $p = 0.8$ very well, and this is reflected in the RMS of Fig. 5(e), where we observe remarkable performance. On the other hand, when our priors have a small density around the true distributions, performance can be quite poor compared to resubstitution and leave-one-out and converge very slowly as we observe more samples. See for example the dark blue prior (with $\mathrm{E}_\pi[p] = 0.5$) in Fig 5(e).

Figure 5 also suggests that, whereas, low-variance priors can have excellent performance as well as the potential for catastrophic results, high-variance priors tend to give safer results by avoiding catastrophic behavior at the expense of performance. This is clear by comparing Fig. 5(d), which uses high variance priors and exhibits a fairly tight range of performance, with Fig. 5(e), where there is a wider range of results.

The second and third rows of Fig. 4 show performance with sample sizes $n = 5$ and $n = 20$, respectively, as a function of p. These illustrate how each prior performs as the true distributions vary. In all cases, performance is best in the ranges of p and q well represented in the prior distributions, but outside this range results can be poor. This is best seen in Fig. 4(h), where the RMS curves move to the right as the priors move right.

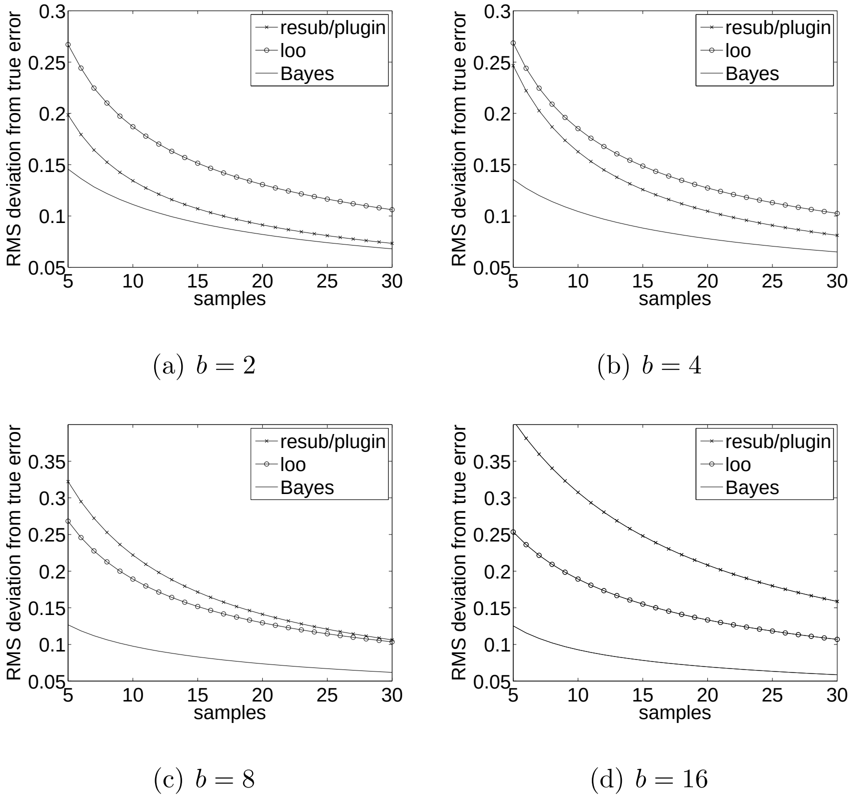
The RMS graphs in Fig. 4 reinforce the notion that narrow priors offer better performance if they are within the targeted range of parameters, but performance outside this range is reciprocally worse. For example, in Fig. 4(h) note how the curves dip very low (good performance when in range) but are narrow (away from this range performance rapidly deteriorates) compared to the corresponding graphs using high-variance priors in Fig. 4(g).

In the right column, Figs. 4(f) and 4(i) show that the uniform prior tends to have poor performance near the edges where p is close to 0 or 1. We will discuss this phenomenon in more detail in the next section, but here these graphs show that if one has a strong belief that p is near 0 or 1, then the uniform prior can be corrected to improve performance. For instance, if we use the dark blue prior in Fig. 4(c), then the new error estimator no longer has a problem near the edges in Figs. 4(f) and 4(i), however performance near $p = 0.5$ is sacrificed. With this prior, note performance of the Bayesian error estimator becomes similar to that of resubstitution; the difference is mostly due to the estimation of c, where resubstitution effectively uses $\frac{n_0}{n}$ and the Bayesian error estimator uses $\frac{n_0+1}{n+2}$.

2. Uniform Priors

This section treats the RMS performance of Bayesian error estimators with non-informative uniform priors for an arbitrary number of bins. As before, we use a histogram classification rule and non-stratified sampling throughout.

Figure 6 gives the average RMS deviation from true error, as a function of sample size, over all distributions in the model with uniform priors for the bin probabilities and c. To generate these graphs, the true distributions and c were randomly selected, a collection of random samples was randomly generated according to the current distributions, and the square deviation from true error was calculated for each error estimator. This was repeated to obtain Monte-Carlo approximations of the RMS for each error estimator. The figure indicates that the Bayesian error estimator has excellent average performance for each fixed n. Indeed, it is optimal according to (2.2). The Bayesian MMSE error estimator shows great improvement over resubstitution and leave-one-out, especially for small samples or a large number of bins. Note also, as has been demonstrated analytically for discrete histogram classification, resubstitution is superior to leave-one-out for small numbers of bins but poorer for large



(a) $b = 2$

(b) $b = 4$

(c) $b = 8$

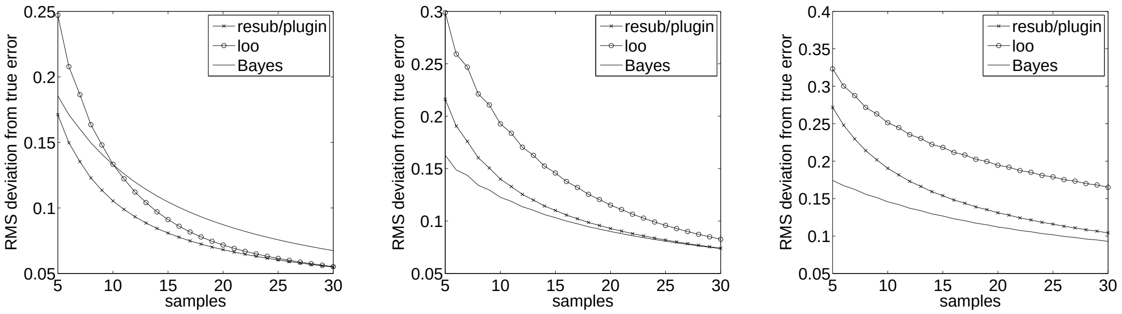
(d) $b = 16$

Fig. 6. RMS deviation from true error for discrete classification and uniform priors with respect to sample size (c and bin probabilities uniform).

numbers (on account of increasing bias) [26].

In the remaining plots in this section, the distributions are fixed with $c = 0.5$. We use the Zipf model, or power law model from [26], where $p_i \propto i^{-\alpha}$ and $q_i = p_{b-i+1}$, $i = 1, \ldots, b$. The parameter $\alpha \geq 0$ is a free parameter used to target a specific Bayes error, where larger α corresponds to smaller Bayes error.

Figure 7 shows RMS as a function of sample size for bin size 2 and Bayes errors 0.1, 0.2 and 0.4. Figures 8, 9 and 10 present analogous results for bin sizes 4, 8 and 16, respectively. Figure 11 shows RMS as a function of Bayes error for bin size 2 and sample sizes 5 and 20. Figures 12, 13 and 14 present analogous results for bin sizes

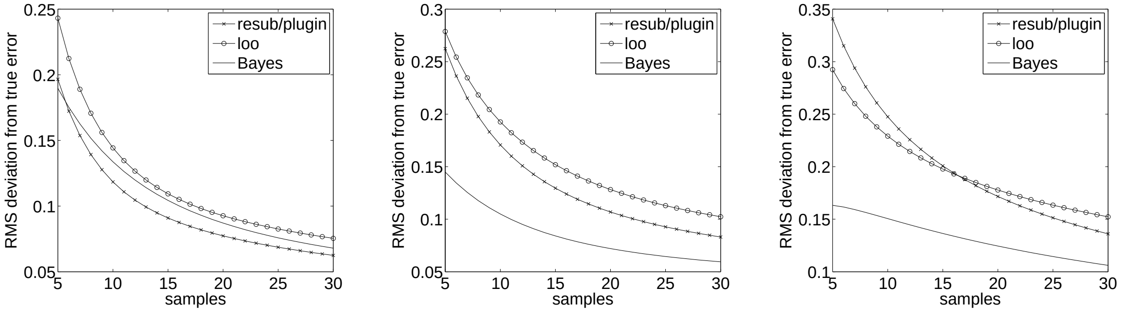


(a) Bayes error = 0.1

(b) Bayes error = 0.2

(c) Bayes error = 0.4

Fig. 7. RMS deviation from true error for discrete classification and uniform priors with respect to sample size ($c = 0.5$, $b = 2$).

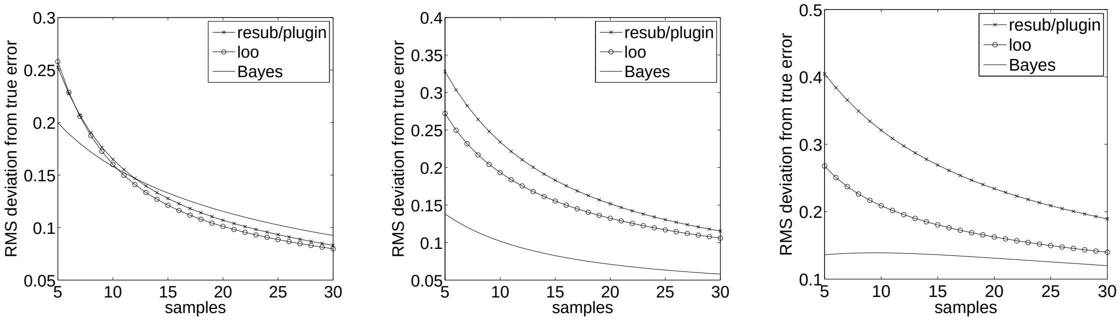


(a) Bayes error = 0.1

(b) Bayes error = 0.2

(c) Bayes error = 0.4

Fig. 8. RMS deviation from true error for discrete classification and uniform priors with respect to sample size ($c = 0.5$, $b = 4$).

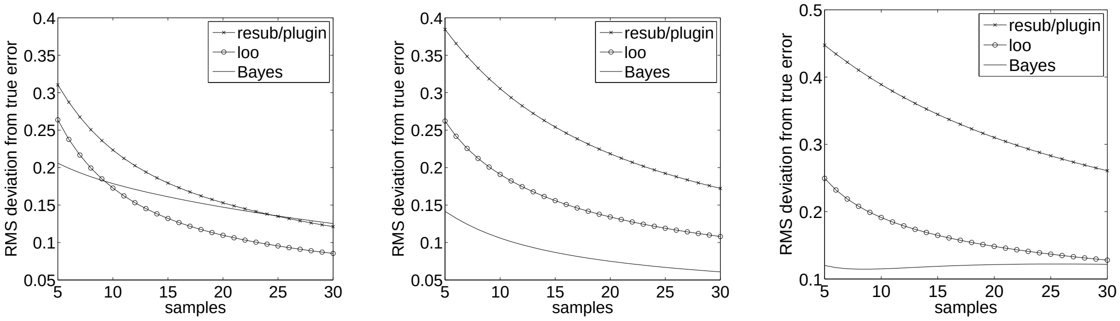


(a) Bayes error = 0.1

(b) Bayes error = 0.2

(c) Bayes error = 0.4

Fig. 9. RMS deviation from true error for discrete classification and uniform priors with respect to sample size ($c = 0.5$, $b = 8$).

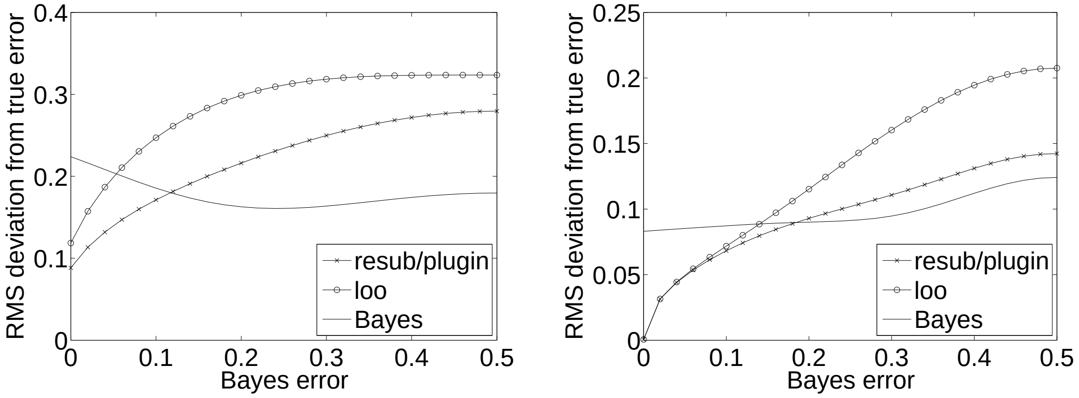


(a) Bayes error = 0.1

(b) Bayes error = 0.2

(c) Bayes error = 0.4

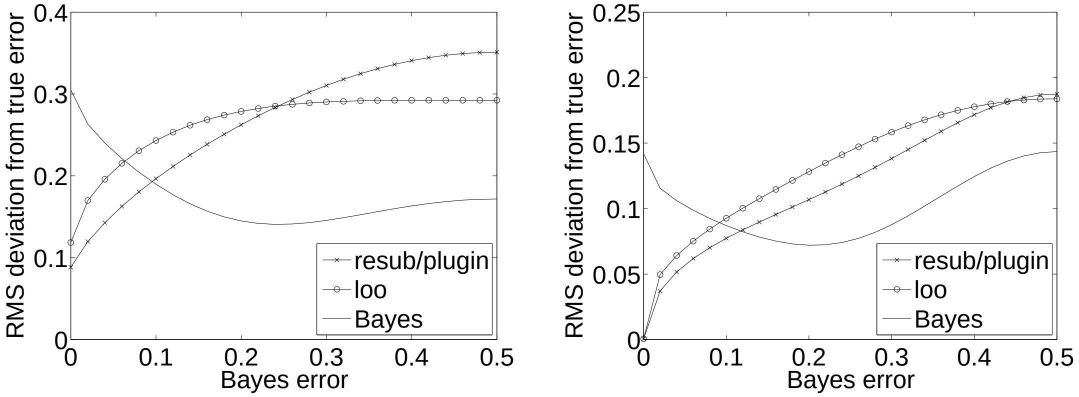
Fig. 10. RMS deviation from true error for discrete classification and uniform priors with respect to sample size ($c = 0.5$, $b = 16$).



(a) $n = 5$

(b) $n = 20$

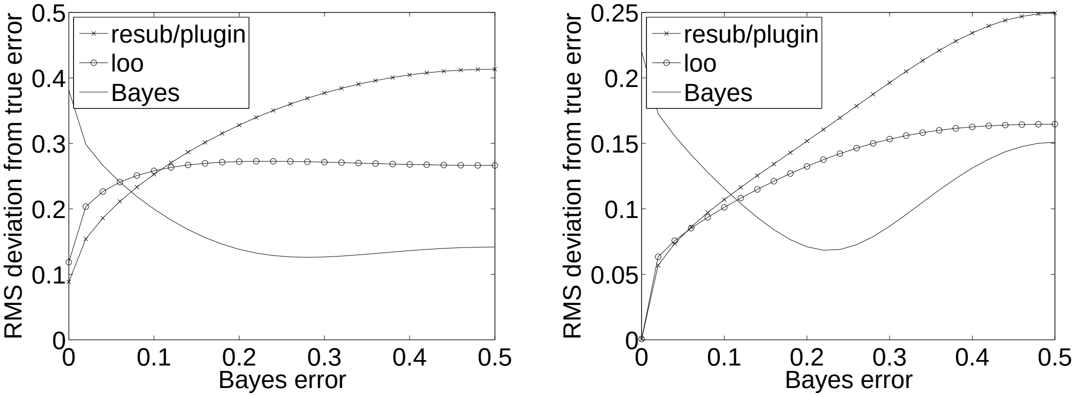
Fig. 11. RMS deviation from true error for discrete classification and uniform priors with respect to Bayes error ($c = 0.5$, $b = 2$).



(a) $n = 5$

(b) $n = 20$

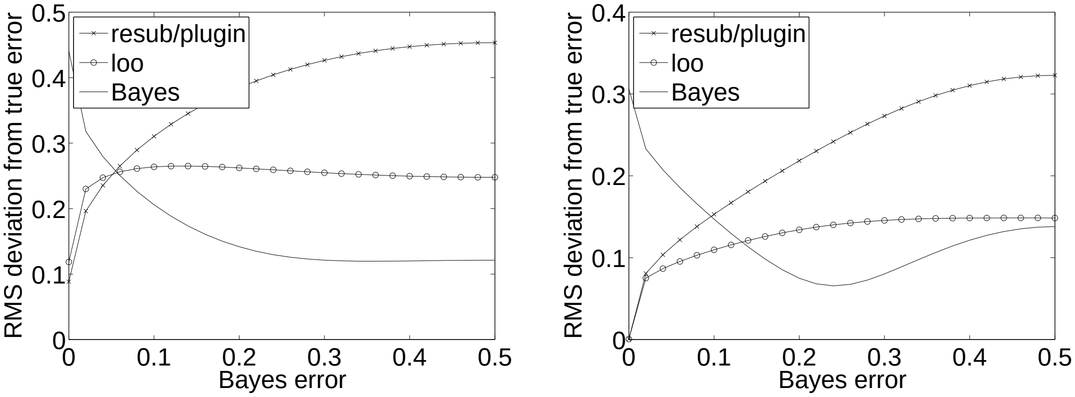
Fig. 12. RMS deviation from true error for discrete classification and uniform priors with respect to Bayes error ($c = 0.5$, $b = 4$).



(a) $n = 5$

(b) $n = 20$

Fig. 13. RMS deviation from true error for discrete classification and uniform priors with respect to Bayes error ($c = 0.5$, $b = 8$).



(a) $n = 5$

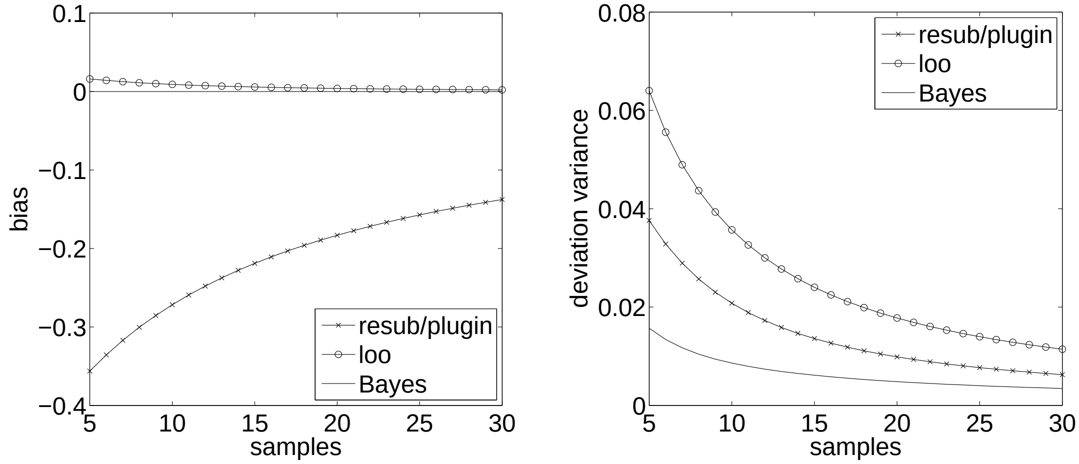
(b) $n = 20$

Fig. 14. RMS deviation from true error for discrete classification and uniform priors with respect to Bayes error ($c = 0.5$, $b = 16$).

4, 8 and 16, respectively. Increasing Bayes error corresponds to increasingly difficult classification. As noted in [26], discrete classification for these bin sizes corresponds to regulatory rule design in binary gene regulatory networks.

In sum, these graphs show that performance for Bayesian error estimation is superior to resubstitution and leave-one-out for most distributions and tends to be especially favorable with moderate to high Bayes errors and smaller sample sizes. From Figs. 11 through 14, it appears that the performance of Bayesian MMSE error estimation tends to be more uniform across all distributions, while the other error estimators, especially resubstitution, favor a small Bayes error. Bayesian MMSE error estimators are guaranteed to be optimal on average over the ensemble of parameterized distributions modeled with respect to the given priors; however, they are not guaranteed to be optimal for a specific distribution, and a clear weakness of these error estimators occurs when the Bayes error is very small.

To explain this latter phenomenon, suppose the true distributions are perfectly separated by the bins, for instance, $p_1 = 1$ and $q_b = 1$, thereby giving a Bayes error of zero. If we observe 5 samples from each class, these will be perfectly separated into the two bins and the histogram classifier will assign the correct class to each bin. Resubstitution and leave-one-out will both give estimates of 0, which is correct; however, since the true distribution is unknown, the Bayesian error estimator considers the possibility that the bin probabilities are non-trivial. This improves the average performance, but not for cases with zero (or very small) Bayes error. Of course, if it is suspected before the experiment that the Bayes error is very low, or if any additional information about the parameters is available to incorporate into the priors, we can improve the Bayesian MMSE error estimator using informed priors as demonstrated in Section III.B.1 with beta priors.



(a) bias

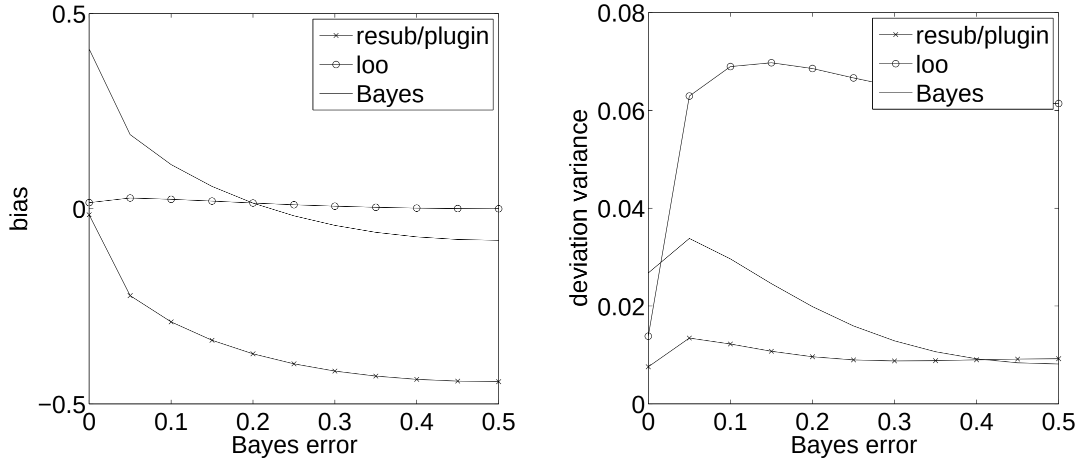
(b) deviation variance

Fig. 15. Bias and deviation variance from true error for discrete classification and uniform priors versus sample size (c and bin probabilities uniform, $b = 16$).

3. Bias and Variance

Figure 15 examines bias and deviation variance versus sample size for a 16 bin problem, averaged over both samples and a uniform prior on the distributions. Resubstitution, leave-one-out, and the Bayesian error estimator with uniform priors are shown. Recall that, according to (2.4), the Bayesian MMSE error estimator is unbiased when averaged over all distributions in the model and all possible samples from these distributions, regardless of the classification rule. We see the unbiasedness of the Bayesian MMSE error estimator in Fig. 15(a), which shows the average bias over all distributions and samples with respect to sample size. Figure 15(b) also shows the significant small-sample advantage in average deviation variance of the Bayesian MMSE error estimator relative to leave-one-out and resubstitution.

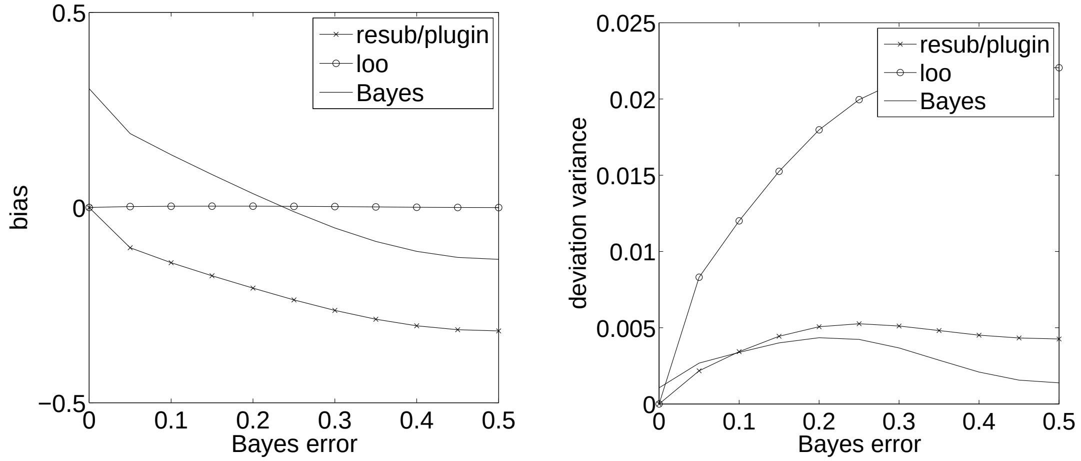
Figures 16 and 17 present bias and deviation variance versus Bayes error for 16 bins, $c = 0.5$, and the same error estimators with $n = 5$ and $n = 20$, respectively.



(a) bias

(b) deviation variance

Fig. 16. Bias and deviation variance from true error for discrete classification and uniform priors versus Bayes error ($c = 0.5$, $b = 16$, $n = 5$).



(a) bias

(b) deviation variance

Fig. 17. Bias and deviation variance from true error for discrete classification and uniform priors versus Bayes error ($c = 0.5$, $b = 16$, $n = 20$).

In these figures, each point represents a fixed distribution, using the same model described previously (i.e., the Zipf model with $c = 0.5$). Notice that leave-one-out is nearly unbiased with a very large deviation variance, while resubstitution is quite optimistically biased with a much lower deviation variance. In contrast, the Bayesian error estimator is pessimistically biased when the true classes are well separated (low Bayes error), but tends to be optimistically biased when the true classes are highly mixed together (high Bayes error). This correlates with our previous RMS graphs, where the performance of the error estimator is usually best with moderate Bayes error. At the same time, the deviation variance often rivals that of resubstitution.

C. Discussion

We have defined the Bayesian MMSE error estimator for classification, discussed some of its properties, derived its analytic representation for discrete classification, and considered its performance. In the next chapter, we will derive and study the Bayesian MMSE error estimator for linear classification in the Gaussian model, including an application to genomic cancer classification. Before closing we would like to comment on three background issues.

The entire development of the Bayesian MMSE estimator is based on the expectation of (2.3) involving the function $g(X, Y)$. In this case, the expectation is conditioned on the sample and therefore yields a function of the sample as occurs in (2.5). This kind of expectation, when unconditioned, plays a fundamental role in robust classification and, more generally, in robust filter design. The theory of optimal robust filtering dates back to the 1970s where the problem was to design a linear filter that is optimal across an uncertainty class, $\mathcal{P} = \{P_\theta\}$, of random processes, i.e., a robust Wiener filter. The problem was originally posed in a minimax framework:

find the filter that minimizes the maximum error across the uncertainty class [84, 85]. By putting a probability measure $\pi(\theta)$ on the space $\{P_\theta\}$, thereby giving more weight to more likely processes, robust filtering, both linear and nonlinear, was put in a Bayesian framework by defining the *Bayesian robust filter* to be the optimal filter, ξ_φ, for the process P_φ that minimizes $\mathrm{E}_\theta[g(\theta,\varphi)]$, where $g(\theta,\varphi)$ is the MSE error for the filter ξ_φ applied to the process P_θ [86, 38]. Optimal Bayesian robust filtering was applied to classification by letting $\mathcal{F} = \{F_\theta\}$ be a space of feature-label distributions and, given a sample S_n, defining the *Bayesian robust classifier* for classification rule Ψ to be the designed classifier ψ_φ for the feature-label distribution F_φ that minimizes $\mathrm{E}_\theta[g(\theta,\varphi,S_n)]$, where $g(\theta,\varphi,S_n)$ is the classifier error for the classifier ψ_φ applied to F_θ [87]. More recently, the concept of Bayesian robustness has been extended to finding a robust controller across a space of ergodic Markov chains, in particular, gene regulatory networks [40]. All of these approaches optimize operator behavior across a space of distributions for a given error estimator. In defining the Bayesian MMSE error estimator, we have viewed the problem from a reverse perspective: optimize the error estimator across a space of distributions for a given operator.

Model uncertainty leads naturally to a Bayesian approach in the context of optimal filtering. More generally, Bayesian estimation involves a loss function (MSE being one possibility) and minimization of the expected value of the loss function. For parameter estimation, the most direct Bayesian approach is to assume a prior distribution for the parameter and then optimize relative to the corresponding posterior distribution. In our case, that would mean postulating a prior distribution for the true error directly. However, since given the classifier the true error is known for a known feature-label distribution, the uncertainty naturally arises in regard to the feature-label distribution and, as we have seen, this fits naturally within the filter theory of Section II.B.

A key aspect of the Bayesian MMSE error estimator is that it can improve error estimation via the assumption of a prior distribution; on the other hand, it is claimed that cross-validation is advantageous because it requires no prior distribution to compute the estimate. While this is true, assumptions are needed to insure acceptable performance. Consider Fig. 14(a). Clearly, an RMS exceeding 0.2 renders the error estimator virtually useless. In this sense, leave-one-out is only useful for Bayes error less than 0.02. Hence, there must be a prior assumption to this effect, or else why is it being used? On the other hand, RMS for the Bayesian MMSE error estimator is below 0.2 for Bayes error exceeding 0.1. It is useful over a much wider range. Moreover, whereas we explicitly know this range because the assumptions are explicit, the assumptions required for leave-one-out to be useful are typically not specified, so that they remain implicit and the meaningfulness of the estimate is unknown.

CHAPTER IV

BAYESIAN MMSE ERROR ESTIMATION—LINEAR CLASSIFICATION OF GAUSSIAN DISTRIBUTIONS*

A. The Gaussian Model

We will derive closed-form Bayesian MMSE error estimators for the Gaussian model. Each sample point is a column vector of D multivariate Gaussian features, so that the sample space is $\mathbb{R}^D$ with D dimensions. For each class, labeled $y = 0$ or $y = 1$, assume a Gaussian distribution with parameters $\theta_y = [\mu_y, \Lambda_y]$, where μ_y is the mean of the class-conditional distribution and Λ_y is a collection of parameters that determine the covariance of the class, Σ_y (we make a distinction to enable us to impose a structure on the covariance). The parameter space of μ_y is $\mathbb{R}^D$, and the parameter space of Λ_y, denoted $\mathbf{\Lambda}_y$, must be carefully defined to permit only valid covariance matrices. We will sometimes write Σ_y without explicitly showing its dependence on Λ_y, that is, we simply write Σ_y instead of $\Sigma_y(\Lambda_y)$. A multivariate Gaussian distribution with mean μ and covariance Σ is denoted by $f_{\mu,\Sigma}(\mathbf{x})$, so that the parameterized class-conditional distributions are $f_{\theta_y}(\mathbf{x}|y) = f_{\mu_y,\Sigma_y}(\mathbf{x})$.

We will consider three covariance models: a fixed covariance ($\Sigma_y = \Lambda_y$ is known perfectly), a scaled identity covariance having features that are uncorrelated with equal variances ($\Lambda_y = \sigma_y^2$ is a scaler and $\Sigma_y = \sigma_y^2 I_D$, where I_D is the $D \times D$ identity

[*] Reprinted with permission from "Bayesian Minimum Mean-Square Error Estimation for Classification Error–Part II: The Bayesian MMSE Error Estimator for Linear Classification of Gaussian Distributions" by L. A. Dalton and E. R. Dougherty, 2011, *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 130–144, Copyright 2011 by IEEE.

matrix) and an arbitrary covariance ($\Sigma_y = \Lambda_y$ can be any valid covariance matrix). Note that a different covariance model may be used for each class.

The sample mean and covariance matrices are found using the usual formulas:

$$\widehat{\mu}_y = \tfrac{1}{n_y} \textstyle\sum_{i=1}^{n_y} \mathbf{x}_i^y \quad \text{and} \quad \widehat{\Sigma}_y = \tfrac{1}{n_y - 1} \textstyle\sum_{i=1}^{n_y} (\mathbf{x}_i^y - \widehat{\mu}_y)(\mathbf{x}_i^y - \widehat{\mu}_y)^T. \tag{4.1}$$

We assume $\widehat{\Sigma}_y$ is nonsingular. If $\widehat{\Sigma}_y$ is singular, then $\pi^*(\theta_y)$ is not trivial to find and we will not go through the details here. Alternatively, one may also convert the classifier and the distribution for class y to a problem in smaller dimensions where $\widehat{\Sigma}_y$ is nonsingular, but this is not an equivalent approach since the class-conditional densities will effectively be restricted to a smaller subspace.

1. Prior Parameter Densities

Considering one class at a time, we assume Σ_y is invertible with probability 1, and for invertible Σ_y our priors are of the form:

$$\begin{aligned} \pi(\theta_y) \propto & |\Sigma_y|^{-(\kappa+D+1)/2} \exp\left(-\tfrac{1}{2}\text{trace}\left(S\Sigma_y^{-1}\right)\right) \\ & \times |\Sigma_y|^{-1/2} \exp\left(-\tfrac{\nu}{2}(\mu_y - \mathbf{m})^T \Sigma_y^{-1} (\mu_y - \mathbf{m})\right), \end{aligned} \tag{4.2}$$

where in general we minimally require the hyperparameters κ to be a real number (we will show that restricting κ to be an integer will permit us to utilize a closed form solution for the Bayesian MMSE error estimator), S to be a non-negative definite $D \times D$ matrix, ν to be a real number, and $\mathbf{m}$ to be a length D real vector. Note that we can have different priors for both classes, but since the analysis for each class can be done independently we will not make a distinction between the notation for hyperparameters in either class.

The hyperparameter $\mathbf{m}$ can be viewed as a target for the mean, where the larger ν is the more localized the prior is about $\mathbf{m}$. Similarly, S can be viewed as a target for the shape of the covariance, although the actual expected variance may be scaled. For instance, in the arbitrary covariance model where $\Sigma_y = \Lambda_y$, this prior is a normal-inverse-Wishart distribution, which is the conjugate prior for the mean and covariance when sampling from normal distributions [77, 88], and $\mathrm{E}_\pi[\Sigma_y] = S/(\kappa - D - 1)$. If S is scaled appropriately, then the larger κ is the less the covariance, Σ_y, is allowed to wiggle. At the same time, increasing κ while fixing the other hyperparameters defines a prior favoring smaller $|\Sigma_y|$.

Requirements for a proper prior depend on the definition of Λ_y, for example, in the arbitrary covariance model we require $\kappa > D-1$, S positive definite, and $\nu > 0$ to guarantee a proper normal-inverse-Wishart prior. That being said, in the Gaussian model we will use improper priors freely in our analysis, as long as the posterior is proper. Some useful examples of improper priors occur when $S = 0$ and $\nu = 0$. In this case, our prior has the form

$$\pi(\theta_y) \propto |\Sigma_y|^{-(\kappa+D+2)/2}. \tag{4.3}$$

If $\kappa + D + 2 = 0$, we obtain flat priors used by Laplace [89]. Alternatively, if $\Lambda_y = \Sigma_y$, then with $\kappa = 0$ we obtain Jeffreys' rule prior, which is designed to be invariant to differentiable one-to-one transformations of the parameters [90, 91], and with $\kappa = -1$ we obtain independence Jeffreys' prior, which uses the same principle as the Jeffreys' rule prior but also treats the mean and covariance matrix as independent parameters.

2. Posterior Parameter Densities

For fixed κ, S, ν and $\mathbf{m}$, the posterior probabilities of the distribution parameters are found from (2.9). After some simplification, we have

$$\pi^*(\theta_y) \propto \pi(\theta_y)|\Sigma_y|^{\frac{-n_y}{2}} \exp\left(-\tfrac{1}{2}\mathrm{trace}((n_y-1)\widehat{\Sigma}_y\Sigma_y^{-1}) - \tfrac{n_y}{2}(\mu_y - \widehat{\mu}_y)^T\Sigma_y^{-1}(\mu_y - \widehat{\mu}_y)\right).$$

Our prior has a similar form to this expression and can be merged with the rest of the equation, giving

$$\pi^*(\theta_y) \propto |\Sigma_y|^{-(\kappa+n_y+D+1)/2} \exp\left(-\frac{1}{2}\text{trace}\left(\left((n_y-1)\widehat{\Sigma}_y + S\right)\Sigma_y^{-1}\right)\right)$$
$$\times |\Sigma_y|^{-1/2} \exp\Bigg(-\frac{1}{2}\Big(n_y(\mu_y-\widehat{\mu}_y)^T\Sigma_y^{-1}(\mu_y-\widehat{\mu}_y) + \nu(\mu_y-\mathbf{m})^T\Sigma_y^{-1}(\mu_y-\mathbf{m})\Big)\Bigg).$$

Furthermore, as long as either $\nu + n_y > 0$ we have

$$\begin{aligned}
&n_y(\mu_y-\widehat{\mu}_y)^T\Sigma_y^{-1}(\mu_y-\widehat{\mu}_y) + \nu(\mu_y-\mathbf{m})^T\Sigma_y^{-1}(\mu_y-\mathbf{m})\\
=&(n_y+\nu)\left(\mu_y - \frac{n_y\widehat{\mu}_y+\nu\mathbf{m}}{n_y+\nu}\right)^T \Sigma_y^{-1}\left(\mu_y - \frac{n_y\widehat{\mu}_y+\nu\mathbf{m}}{n_y+\nu}\right)\\
&+\frac{n_y\nu}{n_y+\nu}(\widehat{\mu}_y-\mathbf{m})^T\Sigma_y^{-1}(\widehat{\mu}_y-\mathbf{m}).
\end{aligned}$$

This leads us finally to the posterior density, which has the same form as the prior:

$$\begin{aligned}
\pi^*(\theta_y) \propto &|\Sigma_y|^{-(\kappa^*+D+1)/2} \exp\left(-\frac{1}{2}\text{trace}\left(S^*\Sigma_y^{-1}\right)\right)\\
&\times|\Sigma_y|^{-1/2}\exp\left(-\frac{\nu^*}{2}\left(\mu_y-\mathbf{m}^*\right)^T\Sigma_y^{-1}\left(\mu_y-\mathbf{m}^*\right)\right)
\end{aligned} \tag{4.4}$$

where

$$\begin{aligned}
\kappa^* &= \kappa + n_y,\\
S^* &= (n_y-1)\widehat{\Sigma}_y + S + \frac{n_y\nu}{n_y+\nu}(\widehat{\mu}_y-\mathbf{m})(\widehat{\mu}_y-\mathbf{m})^T,\\
\nu^* &= \nu + n_y,\\
\mathbf{m}^* &= \frac{n_y\widehat{\mu}_y+\nu\mathbf{m}}{n_y+\nu}.
\end{aligned}$$

These hyperparameters may be viewed as being updated after observing the data. Similar results have been found in [77]. Note that the choice of Λ_y will effect the proportionality constant in $\pi^*(\theta_y)$. We may also write the posterior probability in (4.4) as

$$\pi^*(\theta_y) = \pi^*(\mu_y|\Lambda_y)\pi^*(\Lambda_y),$$

where

$$\begin{aligned} \pi^*(\mu_y|\Lambda_y) &= f_{\mathbf{m}^*,\Sigma_y/\nu^*}(\mu_y), \\ \pi^*(\Lambda_y) &\propto |\Sigma_y|^{-(\kappa^*+D+1)/2} \exp\left(-\frac{1}{2}\text{trace}\left(S^*\Sigma_y^{-1}\right)\right). \end{aligned}$$

Thus, for a fixed covariance matrix the posterior density for the mean, $\pi^*(\mu_y|\Lambda_y)$, is Gaussian. We will see that we require $\nu^* = \nu + n_y > 0$ in all models considered, so $\pi^*(\mu_y|\Lambda_y)$ is always proper. The validity of $\pi^*(\Lambda_y)$ depends on the definition of Λ_y, which will be covered in detail in later sections. Although it is not mandatory for the prior to be a proper density (e.g., in the general covariance model where $\Sigma_y = \Lambda_y$, recall that the prior is proper if $\kappa > D - 1$, S positive definite, and $\nu > 0$), it is crucial for the posterior to be proper (e.g., in the general covariance model we must have $\kappa^* > D - 1$, S^* positive definite, and $\nu^* > 0$).

3. The Bayesian Error Estimator for Linear Classifiers

The posterior expectations for ε_n^y used to find the Bayesian estimator follow from (2.11):

$$\widehat{\varepsilon}^y = \int_{\mathbf{\Lambda}_y} \int_{\mathbb{R}^D} \varepsilon_n^y(\mu_y, \Lambda_y)\pi^*(\mu_y|\Lambda_y)d\mu_y\pi^*(\Lambda_y)d\Lambda_y. \tag{4.5}$$

Suppose the classifier discriminant is linear in form, i.e.,

$$\psi_n(\mathbf{x}) = \begin{cases} 0 & \text{if } g(\mathbf{x}) \leq 0, \\ 1 & \text{if } g(\mathbf{x}) > 0, \end{cases} \tag{4.6}$$

where $g(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ with some constant vector $\mathbf{a}$ and constant scalar b, and we allow this classifier to be any function of the observed samples. With fixed distribution parameters and non-zero $\mathbf{a}$, the true error for this classifier applied to a class y Gaussian distribution with mean μ_y and covariance Σ_y is given by

$$\varepsilon_n^y = \Phi\left(\frac{(-1)^y g(\mu_y)}{\sqrt{\mathbf{a}^T \Sigma_y \mathbf{a}}}\right), \tag{4.7}$$

where Φ is the unit normal Gaussian cumulative distribution function [43]. If $\mathbf{a} = 0$, that is if the designed classifier is constant, then the true error, ε_n^y, is deterministically zero or one, depending on the sign of b, so that the Bayesian error estimator can be found deterministically from (2.10). Hence, in the remainder of this chapter we assume $\mathbf{a} \neq 0$.

Interestingly, the Bayesian error estimator simplifies to a function of just the sample mean and covariance, not the individual sample points themselves. In this sense, the Bayesian error estimation rule boils down to the quality of the parameter estimates, just like the plug-in rule. The difference is that it optimally processes these parameters to find the MMSE error estimate. The plug-in rule is intuitive, but really an arbitrary method based on the hope that parameter estimates will be close to the true ones.

In the remainder of this section, we consider the effect of applying priors to different transformations of the covariance matrix. For instance, in the scaled identity covariance model Λ_y contains the variances in Σ_y rather than standard deviations, and in the arbitrary covariance model Λ_y contains the covariance matrix itself, rather than the precision matrix (the inverse covariance matrix) or parameters from a decomposition of the covariance matrix. We will demonstrate how such transformations can result in Bayesian error estimators of the same form. In particular, we will show that Bayesian error estimators derived for the arbitrary covariance model using the covariance matrix itself for Λ_y are of the same form as estimators derived using a statistic based on the Cholesky decomposition (in one dimension this is equivalent to using the standard deviation rather than variance).

Consider two different priors. The first is defined with respect to the covariance matrix itself, i.e., $\Sigma_y = \Lambda_y$ and $\mathbf{\Lambda}_y$ contains all positive definite matrices. For this prior, we write the posterior density as $\pi_1^*(\Sigma_y, \kappa)$ and the Bayesian error estimator as $\widehat{\varepsilon}_1^y(\kappa)$, to emphasize the value of κ. In the second prior, we let $\Sigma_y = \Lambda_y \Lambda_y^T$, where $\mathbf{\Lambda}_y$ is the set of all invertible lower triangular matrices. The Jacobean determinant of this transformation is determined by $d\Lambda_y = |\Sigma_y|^{-1/2} d\Sigma_y$ [92] and Λ_y is an invertible lower triangular matrix if and only if Σ_y is positive definite. For this prior, we denote the posterior density of Λ_y by $\pi_2^*(\Lambda_y, \kappa)$ and the Bayesian error estimator as $\widehat{\varepsilon}_2^y(\kappa)$.

Observe when we normalize $\pi_2^*(\Lambda_y, \kappa)$,

$$\begin{aligned}\int_{\mathbf{\Lambda}_y} \pi_2^*(\Lambda_y, \kappa) d\Lambda_y &= \int_{\mathbf{\Lambda}_y} |\Lambda_y \Lambda_y^T|^{-\frac{\kappa^*+D+1}{2}} \exp\left(-\frac{1}{2}\text{trace}\left(S^*(\Lambda_y \Lambda_y^T)^{-1}\right)\right) d\Lambda_y \\ &= \int_{\Sigma_y > 0} |\Sigma_y|^{-\frac{\kappa^*+D+1}{2}} \exp\left(-\frac{1}{2}\text{trace}\left(S^* \Sigma_y^{-1}\right)\right) |\Sigma_y|^{-\frac{1}{2}} d\Sigma_y \\ &= \int_{\Sigma_y > 0} \pi_1^*(\Sigma_y, \kappa + 1) d\Sigma_y.\end{aligned}$$

In other words, we obtain the same normalization constant as we would for a prior defined with respect to the covariance matrix with κ increased by 1. In fact, $\pi_2^*(\Lambda_y, \kappa) = \pi_1^*(\Sigma_y, \kappa + 1)|\Sigma_y|^{1/2}$. The Bayesian error estimator for the second prior is thus

$$\begin{aligned}\widehat{\varepsilon}_2^y(\kappa) &= \int_{\mathbf{\Lambda}_y} f\left(\Lambda_y \Lambda_y^T\right) \pi_2^*(\Lambda_y, \kappa) d\Lambda_y \\ &= \int_{\Sigma_y > 0} f\left(\Sigma_y\right) \pi_1^*(\Sigma_y, \kappa + 1) d\Sigma_y = \widehat{\varepsilon}_1^y(\kappa + 1),\end{aligned}$$

where f is the inner integral in (4.5), which can be expressed as a function of the covariance, Σ_y. In other words, the Bayesian error estimator using a Cholesky decomposition of Σ_y for Λ_y is exactly the same as the Bayesian error estimator obtained using $\Lambda_y = \Sigma_y$, with a slight modification of κ.

4. Solution for Fixed Covariance

We first consider the Bayesian error estimator for the fixed (invertible) covariance model with arbitrary linear classification. Equivalently, we seek a closed-form solution for the inner integral in (4.5). This is solved analytically in the following lemma.

Lemma 2. *Let $y \in \{0, 1\}$ be a class label and let $\nu^* > 0$. Also let $\mathbf{m}^* \in \mathbb{R}^D$ be a mean vector with $D \geq 1$ features, Σ be an invertible covariance matrix, and $g(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$, where $\mathbf{a} \in \mathbb{R}^D$ is a non-zero length D vector and $b \in \mathbb{R}$ is a scalar. Then,*

$$\int_{\mathbb{R}^D} \Phi\left(\frac{(-1)^y g(\mu)}{\sqrt{\mathbf{a}^T\Sigma\mathbf{a}}}\right) f_{\mathbf{m}^*,\Sigma/\nu^*}(\mu) d\mu = \Phi\left(\frac{(-1)^y g(\mathbf{m}^*)}{\sqrt{\mathbf{a}^T\Sigma\mathbf{a}}}\sqrt{\frac{\nu^*}{\nu^*+1}}\right),$$

where $f_{\mu,\Sigma}$ is a Gaussian density with mean μ and covariance Σ.

Proof. Call this integral M. We have that,

$$M = \int_{\mathbb{R}^D} \Phi\left(\frac{(-1)^y g(\mu)}{\sqrt{\mathbf{a}^T\Sigma\mathbf{a}}}\right) \frac{\nu^{*\frac{D}{2}}}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{\nu^*}{2}(\mu - \mathbf{m}^*)^T\Sigma^{-1}(\mu - \mathbf{m}^*)\right) d\mu.$$

Since Σ is an invertible covariance matrix, we can use singular value decomposition to write $\Sigma = WW^T$ with $|\Sigma| = |W|^2$. Next consider the linear change of variables, $\mathbf{z} = \sqrt{\nu^*}W^{-1}(\mu - \mathbf{m}^*)$. We have that,

$$M = \int_{\mathbb{R}^D} \Phi\left(\frac{(-1)^y\left(\frac{1}{\sqrt{\nu^*}}\mathbf{a}^TW\mathbf{z} + \mathbf{a}^T\mathbf{m}^* + b\right)}{\sqrt{\mathbf{a}^T\Sigma\mathbf{a}}}\right) \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{\mathbf{z}^T\mathbf{z}}{2}\right) d\mathbf{z}.$$

Define $\bar{\mathbf{a}} = \frac{(-1)^y W^T\mathbf{a}}{\sqrt{\nu^*}\sqrt{\mathbf{a}^T\Sigma\mathbf{a}}}$ and $\bar{b} = \frac{(-1)^y g(\mathbf{m}^*)}{\sqrt{\mathbf{a}^T\Sigma\mathbf{a}}}$, and note that $\|\bar{\mathbf{a}}\|^2 = \frac{1}{\nu^*}$. Then,

$$\begin{aligned} M &= \int_{\mathbb{R}^D} \Phi\left(\bar{\mathbf{a}}^T\mathbf{z} + \bar{b}\right) \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{\mathbf{z}^T\mathbf{z}}{2}\right) d\mathbf{z} \\ &= \int_{\mathbb{R}^D} \int_{x<\bar{\mathbf{a}}^T\mathbf{z}+\bar{b}} \frac{1}{(2\pi)^{\frac{D+1}{2}}} \exp\left(-\frac{x^2 + \mathbf{z}^T\mathbf{z}}{2}\right) dx d\mathbf{z}. \end{aligned}$$

This is the integral of a $D + 1$ dimensional multivariate Gaussian distribution on one side of a hyperplane, which is equivalent to the well known true error of a linear classifier contributed by a single Gaussian class given in (4.7). In this case, we have the classifier $\bar{g}(\mathbf{z}, x) = \bar{\mathbf{a}}^T\mathbf{z} - x + \bar{b}$ applied to a class 0 Gaussian distribution with zero mean and identity covariance. Hence,

$$M = \Phi\left(\frac{\bar{b}}{\sqrt{\|\bar{\mathbf{a}}\|^2 + 1}}\right) = \Phi\left(\frac{(-1)^y g\left(\mathbf{m}^*\right)}{\sqrt{\mathbf{a}^T\Sigma\mathbf{a}}}\sqrt{\frac{\nu^*}{\nu^* + 1}}\right),$$

as desired. □

Thus, the Bayesian error estimator, or expected error, for a Gaussian class with fixed covariance and a linear classifier is given by

$$\widehat{\varepsilon}^y = \int_{\mathbb{R}^D} \Phi\left(\frac{(-1)^y g(\mu_y)}{\sqrt{\mathbf{a}^T\Sigma_y\mathbf{a}}}\right) f_{\mathbf{m}^*, \Sigma_y/\nu^*}(\mu_y)d\mu_y = \Phi\left(\frac{(-1)^y g(\mathbf{m}^*)}{\sqrt{\mathbf{a}^T\Sigma_y\mathbf{a}}}\sqrt{\frac{\nu^*}{\nu^* + 1}}\right). \quad (4.8)$$

This equation suggests that averaging over the means simply applies a factor of $\sqrt{\frac{\nu^*}{\nu^*+1}}$ inside Φ. Since this factor is always less than 1, and for a good classifier $(-1)^y g(\mathbf{m}^*)$ tends to be negative, this suggests that the plug-in rule is pessimistic, and presents the proper way to correct it.

5. Solution for Scaled Identity Covariance

Having solved the Bayesian error estimator for fixed covariance, Bayesian error estimators for random covariance models can now be reduced to the following integral over the covariance parameter only:

$$\widehat{\varepsilon}^y = \int_{\mathbf{\Lambda}_y} \Phi\left(\frac{(-1)^y g(\mathbf{m}^*)}{\sqrt{\mathbf{a}^T\Sigma_y\mathbf{a}}}\sqrt{\frac{\nu^*}{\nu^* + 1}}\right)\pi^*(\Lambda_y)d\Lambda_y. \quad (4.9)$$

We now assume Λ_y contains only one parameter, $\Lambda_y = \sigma_y^2$. We define the parameter space $\mathbf{\Lambda}_y = [0, \infty)$ and $\Sigma_y = \sigma_y^2 I_D$. This simplification of the covariance matrix is most useful in cases with a very small sample, where estimating the entire covariance matrix is not reliable.

In this case, the posterior density $\pi^*(\sigma_y^2)$ is an inverse-gamma distribution:

$$\pi^*(\sigma_y^2) = \frac{1}{\Gamma(\alpha)}\beta^\alpha \frac{1}{(\sigma_y^2)^{\alpha+1}} \exp\left(-\frac{\beta}{\sigma_y^2}\right),$$

where $\alpha > 0$ and $\beta > 0$ are given by

$$\begin{aligned} \alpha &= \frac{(\kappa^* + D + 1)D}{2} - 1, \\ \beta &= \frac{1}{2}\text{trace}\left(S^*\right). \end{aligned}$$

If $\alpha \leq 0$ or $\beta \leq 0$, then the posterior distribution is not valid and cannot be used to find the Bayesian error estimate. Problems normalizing the posterior density occur because we have used improper priors, which are a convenience. In these troublesome cases, either a larger sample or better prior is needed to proceed [75].

If the posterior is valid, the expected error $\widehat{\varepsilon}^y$ from (4.9) is exactly the integral in the following lemma.

Lemma 3. *Let $A \in \mathbb{R}$, $\alpha > 0$, and $\beta > 0$. Also let $f_G(x; \alpha, \beta)$ be an inverse-gamma distribution with shape parameter α and scale parameter β. Then,*

$$\int_0^\infty \Phi\left(\frac{A}{\sqrt{z}}\right) f_G(z; \alpha, \beta) dz = \frac{1}{2}\left(1 + \text{sgn}(A) I\left(\frac{A^2}{A^2 + 2\beta}; \frac{1}{2}, \alpha\right)\right),$$

where $I(x; a, b)$ is the regularized incomplete beta function.

Proof. Call this integral M. Observe,

$$\begin{aligned} M &= \frac{1}{\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \int_{-\frac{A}{\sqrt{z}}}^\infty \frac{1}{z^{\alpha+1}} \exp\left(-\frac{x^2}{2} - \frac{\beta}{z}\right) dx dz \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{\Gamma(\alpha) 2^\alpha} \int_0^\infty \int_{-\frac{Ay}{\sqrt{2\beta}}}^\infty y^{2\alpha-1} \exp\left(-\frac{x^2 + y^2}{2}\right) dx dy. \end{aligned}$$

The last line follows from the change of variables $y^2/2 = \beta/z$.

We next convert to polar coordinates with $x = r\cos\theta$ and $y = r\sin\theta$. The limits of integration are determined by three cases, depending on the sign of A. These are all considered simultaneously by defining,

$$\begin{aligned}\theta_0 &= \begin{cases} \arctan\left(\frac{\sqrt{2\beta}}{|A|}\right) & \text{if } A < 0 \\ \frac{\pi}{2} & \text{if } A = 0 \\ \pi - \arctan\left(\frac{\sqrt{2\beta}}{|A|}\right) & \text{if } A > 0 \end{cases} \\ &= \arctan\left(\frac{A}{\sqrt{2\beta}}\right) + \frac{\pi}{2},\end{aligned}$$

where the second equality follows using the identity $\arctan x + \arctan 1/x = \pi/2$ for $x > 0$. We have

$$\begin{aligned} M &= \sqrt{\frac{2}{\pi}} \frac{1}{\Gamma(\alpha)} \frac{1}{2^\alpha} \int_0^{\theta_0} \int_0^\infty (r\sin\theta)^{2\alpha-1} \exp\left(-\frac{r^2}{2}\right) r dr d\theta \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{\Gamma(\alpha)} \frac{1}{2^\alpha} \int_0^{\theta_0} \sin^{2\alpha-1}\theta d\theta \int_0^\infty r^{2\alpha} \exp\left(-\frac{r^2}{2}\right) dr \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{\Gamma(\alpha)} \frac{1}{2^\alpha} \int_0^{\theta_0} \sin^{2\alpha-1}\theta d\theta \; 2^{\alpha-1/2}\Gamma\left(\alpha+\frac{1}{2}\right) \\ &= \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\Gamma(\alpha)} \int_0^{\theta_0} \sin^{2\alpha-1}\theta d\theta. \end{aligned}$$

The integral over r was solved by noting that it contains a chi distribution with $2\alpha+1$ degrees of freedom. The remaining integral over θ can be written in terms of the regularized incomplete beta function. In particular, we have

$$\begin{aligned} M &= \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\Gamma(\alpha)} \frac{\sqrt{\pi}}{2} \frac{\Gamma(\alpha)}{\Gamma\left(\alpha+\frac{1}{2}\right)} \left(1 - \operatorname{sgn}(\cos\theta_0) I\left(\cos^2\theta_0; \frac{1}{2}, \alpha\right)\right) \\ &= \frac{1}{2}\left(1 + \operatorname{sgn}(A) I\left(\frac{A^2}{A^2+2\beta}; \frac{1}{2}, \alpha\right)\right), \end{aligned}$$

where we have used that $\cos^2\theta_0 = \frac{A^2}{A^2+2\beta}$. □

Thus, the Bayesian error estimator for the Gaussian model assuming scaled identity covariances can be simplified to

$$\widehat{\varepsilon}^y = \frac{1}{2}\left(1 + \operatorname{sgn}(A) I\left(\frac{A^2}{A^2 + \operatorname{trace}(S^*)}; \frac{1}{2}, \frac{(\kappa^* + D + 1)D}{2} - 1\right)\right), \tag{4.10}$$

where

$$A = \frac{(-1)^y g(\mathbf{m}^*)}{\|\mathbf{a}\|}\sqrt{\frac{\nu^*}{\nu^* + 1}}.$$

A closed-form representation for the regularized incomplete beta function, $I(x; a, b)$, is provided in Section IV.A.7 for cases when κ is an integer.

6. Solution for General Covariance

We now consider the general covariance model, where we define $\Lambda_y = \Sigma_y$ and $\mathbf{\Lambda}_y$ contains all positive definite matrices. In this case, $\pi^*(\Sigma_y)$ is an inverse-Wishart distribution [93, 94]:

$$\pi^*(\Sigma_y) = \frac{|S^*|^{\kappa^*/2}|\Sigma_y|^{-\frac{\kappa^*+D+1}{2}}}{2^{\kappa^* D/2}\Gamma_D(\kappa^*/2)} \exp\left(-\frac{1}{2}\operatorname{trace}\left(S^*\Sigma_y^{-1}\right)\right),$$

where Γ_D is the multivariate gamma function and for a proper posterior we require S^* to be positive definite (which is true when $\widehat{\Sigma}_y$ is invertible) and $\kappa^* > D-1$. In one dimension, this is also an inverse-gamma distribution. If $\kappa^* \leq D-1$ then one should seek a proper prior distribution, obtain a larger sample, or simplify the form of the covariance matrix (for instance by assuming identity covariances as in the previous section) to proceed.

If the posterior is valid, the Bayesian error estimator for linear classifiers is found from (4.9) in the following lemma.

Lemma 4. *Let $A \in \mathbb{R}$, $\mathbf{a} \in \mathbb{R}^D$ be a non-zero column vector, $\kappa^* > D-1$ be a real number, and S^* be a positive definite $D \times D$ matrix. Also let $f_W(\Sigma; S^*, \kappa^*)$ be an* *inverse-Wishart distribution with parameters S^* and κ^*. Then*

$$\begin{aligned}\int_{\Sigma>0} &\Phi\left(\frac{A}{\sqrt{\mathbf{a}^T\Sigma\mathbf{a}}}\right) f_W(\Sigma; S^*, \kappa^*) d\Sigma \\ &= \frac{1}{2}\left(1 + \operatorname{sgn}(A) I\left(\frac{A^2}{A^2 + \mathbf{a}^T S^* \mathbf{a}}; \frac{1}{2}, \frac{\kappa^* - D + 1}{2}\right)\right),\end{aligned}$$

where the integration is over all positive definite matrices.

Proof. Call this integral M, and define the following matrix:

$$B = \left[\begin{array}{c|c} \multicolumn{2}{c}{\mathbf{a}^T} \\ \hline 0_{D-1\times 1} & I_{D-1}. \end{array}\right].$$

Since $\mathbf{a}$ is non-zero, with a simple reordering of the dimensions we can guarantee $a_1 \neq 0$. The value of $\mathbf{a}^T S^* \mathbf{a}$ is unchanged by such a redefinition, so without loss of generality assume B is invertible.

Next define a change of variables, $Y = B\Sigma B^T$. Since B is invertible, Y is positive definite if and only if Σ is also. Furthermore, the Jacobean determinant of this transformation is $|B|^{D+1}$ [95, 92]. Note $\mathbf{a}^T\Sigma\mathbf{a} = y_{11}$, where the subscript 11 indexes the upper left element of a matrix, and we have:

$$\begin{aligned} M &= \int_{Y>0} \Phi\left(\frac{A}{\sqrt{y_{11}}}\right) f_W(B^{-1}Y(B^T)^{-1}; S^*, \kappa^*) \frac{dY}{|B|^{D+1}} \\ &= \int_{Y>0} \Phi\left(\frac{A}{\sqrt{y_{11}}}\right) f_W(Y; BS^*B^T, \kappa^*) dY. \end{aligned}$$

Since $\Phi\left(\frac{A}{\sqrt{y_{11}}}\right)$ now depends on only one parameter in Y, the other parameters can be integrated out. It can be shown that for any inverse-Wishart random variable, X, with density $f_W(X; S^*, \kappa^*)$, the marginal distribution of x_{11} is also inverse-Wishart with density $f_W(x_{11}; s^*_{11}, \kappa^* - D + 1)$ [96]. In one dimension, this is equivalent to the inverse-gamma distribution $f_G(x_{11}; (\kappa^* - D + 1)/2, s^*_{11}/2)$. In this case, $(BS^*B^T)_{11} = \mathbf{a}^T S^* \mathbf{a}$, so

$$M = \int_0^\infty \Phi\left(\frac{A}{\sqrt{y_{11}}}\right) f_G\left(y_{11}; \frac{\kappa^* - D + 1}{2}, \frac{\mathbf{a}^T S^* \mathbf{a}}{2}\right) dy_{11}.$$

Next apply Lemma 3, and we have

$$M = \frac{1}{2}\left(1 + \operatorname{sgn}(A) I\left(\frac{A^2}{A^2 + \mathbf{a}^T S^* \mathbf{a}}; \frac{1}{2}, \frac{\kappa^* - D + 1}{2}\right)\right). \qquad \square$$

Thus, the Bayesian error estimator in the case of general covariances is given by

$$\widehat{\varepsilon}^y = \frac{1}{2}\left(1 + \operatorname{sgn}(A) I\left(\frac{A^2}{A^2 + \mathbf{a}^T S^* \mathbf{a}}; \frac{1}{2}, \frac{\kappa^* - D + 1}{2}\right)\right), \tag{4.11}$$

where

$$A = (-1)^y g(\mathbf{m}^*) \sqrt{\frac{\nu^*}{\nu^* + 1}}.$$

If κ^* (or κ) is an integer, then a closed-form representation for the Bayesian error estimator is available in the next section.

7. Closed Form Representation for the I Function

The regularized incomplete beta function is defined by

$$I\left(x; a, b\right) = \frac{1}{B(a, b)} \int_0^x t^{a-1}(1-t)^{b-1} dt$$

for $0 \leq x \leq 1$, $a > 0$ and $b > 0$, where the beta function $B(a, b)$ normalizes I so that $I\left(1; a, b\right) = 1$. In our application, note that we only need to evaluate $I\left(x; \frac{1}{2}, b\right)$ for $0 \leq x < 1$ and $b > 0$.

Although this integral does not have a closed-form solution for arbitrary parameters, in the following lemma we provide exact expressions for $I\left(x; \frac{1}{2}, \frac{N}{2}\right)$ for positive integers N. Restricting b to be an integer or half integer, which in all cases equivalently restricts κ to be an integer, guarantees that these equations may be applied, so that Bayesian error estimators for the Gaussian model with linear classification may be evaluated exactly using finite sums of common single variable functions.

Lemma 5. *Let N be a positive integer. Then $I\left(1; \frac{1}{2}, \frac{N}{2}\right) = 1$ and for any real number $0 \leq x < 1$,*

$$I\left(x; \frac{1}{2}, \frac{N}{2}\right) = \begin{cases} \frac{2}{\pi}\arcsin\left(\sqrt{x}\right) & \text{if } N = 1, \\ \frac{2}{\pi}\arcsin\left(\sqrt{x}\right) + \frac{2}{\pi}\sqrt{x}\sum_{k=1}^{\frac{N-1}{2}} \frac{(2k-2)!!}{(2k-1)!!}(1-x)^{k-\frac{1}{2}} & \text{for } N > 1 \text{ odd}, \\ \sqrt{x}\sum_{k=0}^{\frac{N-2}{2}} \frac{(2k-1)!!}{(2k)!!}(1-x)^{k} & \text{for } N > 1 \text{ even}, \end{cases} \tag{4.12}$$

where !! is the double factorial.

Proof. $I(1; a, b) = 1$ is a property of the regularized incomplete beta function for all $a, b > 0$. For $0 \leq x < 1$, we have that

$$I\left(x; \frac{1}{2}, \frac{N}{2}\right) = \frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{N}{2}\right)}\int_0^x t^{-\frac{1}{2}}(1-t)^{\frac{N-2}{2}}dt.$$

Using the substitution $\sin\theta = \sqrt{t}$, we have

$$\begin{aligned} I\left(x; \frac{1}{2}, \frac{N}{2}\right) &= \frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{N}{2}\right)}\int_{\arcsin\sqrt{0}}^{\arcsin\sqrt{x}} \frac{1}{\sin\theta}\left(\cos^{N-2}\theta\right)2\sin\theta\cos\theta d\theta \\ &= 2\frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{N}{2}\right)}\int_0^{\arcsin\sqrt{x}} \cos^{N-1}\theta d\theta. \end{aligned}$$

For $0 \leq \alpha < \pi/2$ and $k \geq 0$, define

$$M_k(\alpha) \equiv \int_0^\alpha \cos^k\theta d\theta.$$

Using integration by parts or integration tables, it is well known that

$$M_k\left(\alpha\right) = \begin{cases} \alpha & \text{if } k = 0, \\ \sin\alpha & \text{if } k = 1, \\ \frac{k-1}{k} M_{k-2}\left(\alpha\right) + \dfrac{\sin\alpha\cos^{k-1}\alpha}{k} & \text{if } k > 1. \end{cases}$$

The claim for $N = 1$ is easy to verify using the case $k = 0$ above.

For $n > 0$, we apply a recursion using the equation for $k > 1$. If the recursion is applied $i > 0$ times such that $n - 2i \geq 0$, then

$$\begin{aligned} M_n\left(\alpha\right) = & \frac{(n-1)!!}{(n-2i-1)!!} \frac{(n-2i)!!}{n!!} M_{n-2i}\left(\alpha\right) \\ & + \sum_{k=1}^{i} \frac{(n-1)!!}{(n-2k+1)!!} \frac{(n-2k)!!}{n!!} \sin\alpha\cos^{n-2k+1}\alpha \\ = & \frac{(n-1)!!(n-2i)!!}{n!!(n-2i-1)!!} M_{n-2i}\left(\alpha\right) \\ & + \frac{(n-1)!!}{n!!} \sin\alpha \sum_{k=1}^{i} \frac{(n-2k)!!}{(n-2k+1)!!} \cos^{n-2k+1}\alpha. \end{aligned}$$

In particular, for n even we may repeat the recursion $i = n/2$ times to obtain

$$\begin{aligned} M_n\left(\alpha\right) &= \frac{(n-1)!!(0)!!}{n!!(-1)!!} M_0\left(\alpha\right) + \frac{(n-1)!!}{n!!} \sin\alpha \sum_{k=1}^{n/2} \frac{(n-2k)!!}{(n-2k+1)!!} \cos^{n-2k+1}\alpha \\ &= \frac{(n-1)!!}{n!!} \left(\alpha + \sin\alpha \sum_{k=1}^{n/2} \frac{(n-2k)!!}{(n-2k+1)!!} \cos^{n-2k+1}\alpha \right) \\ &= \frac{(n-1)!!}{n!!} \left(\alpha + \sin\alpha \sum_{k=1}^{n/2} \frac{(2k-2)!!}{(2k-1)!!} \cos^{2k-1}\alpha \right), \end{aligned}$$

and if n is odd we may repeat the recursion $i = (n-1)/2$ times to obtain

$$\begin{aligned}
M_n(\alpha) &= \frac{(n-1)!!(1)!!}{n!!(0)!!} M_1(\alpha) + \frac{(n-1)!!}{n!!} \sin\alpha \sum_{k=1}^{(n-1)/2} \frac{(n-2k)!!}{(n-2k+1)!!} \cos^{n-2k+1}\alpha \\
&= \frac{(n-1)!!}{n!!} \sin\alpha \left(1 + \sum_{k=1}^{(n-1)/2} \frac{(n-2k)!!}{(n-2k+1)!!} \cos^{n-2k+1}\alpha \right) \\
&= \frac{(n-1)!!}{n!!} \sin\alpha \left(1 + \sum_{k=1}^{(n-1)/2} \frac{(2k-1)!!}{(2k)!!} \cos^{2k}\alpha \right) \\
&= \frac{(n-1)!!}{n!!} \sin\alpha \sum_{k=0}^{(n-1)/2} \frac{(2k-1)!!}{(2k)!!} \cos^{2k}\alpha,
\end{aligned}$$

where in each case we have redefined the indices of the sums in reverse order.

Returning to the original problem, we have for odd $N > 1$,

$$\frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{N}{2}\right)} = \frac{2^{N-1}\left(\frac{N-1}{2}\right)!\left(\frac{N-1}{2}\right)!}{\pi(N-2)!} = \frac{2^{\frac{N-1}{2}}\left(\frac{N-1}{2}\right)!}{\pi(N-2)!!} = \frac{(N-1)!!}{\pi(N-2)!!}$$

and

$$\begin{aligned}
I\left(x; \frac{1}{2}, \frac{N}{2}\right) &= 2\frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{N}{2}\right)} M_{N-1}\left(\arcsin\sqrt{x}\right) \\
&= 2\frac{(N-1)!!}{\pi(N-2)!!}\frac{(N-2)!!}{(N-1)!!} \\
&\quad \times \left(\arcsin\sqrt{x} + \sqrt{x} \sum_{k=1}^{(N-1)/2} \frac{(2k-2)!!}{(2k-1)!!} \left(\sqrt{1-x}\right)^{2k-1} \right) \\
&= \frac{2}{\pi}\arcsin\sqrt{x} + \frac{2}{\pi}\sqrt{x} \sum_{k=1}^{\frac{N-1}{2}} \frac{(2k-2)!!}{(2k-1)!!}(1-x)^{k-\frac{1}{2}}.
\end{aligned}$$

Finally, for even $N > 1$,

$$\frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{N}{2}\right)} = \frac{N!}{2^N\left(\frac{N}{2}\right)!\left(\frac{N-2}{2}\right)!} = \frac{(N-1)!!}{2^{\frac{N}{2}}\left(\frac{N-2}{2}\right)!} = \frac{(N-1)!!}{2(N-2)!!}$$

and

$$\begin{aligned} I\left(x; \frac{1}{2}, \frac{N}{2}\right) &= 2\frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{N}{2}\right)} M_{N-1}\left(\arcsin\sqrt{x}\right) \\ &= 2\frac{(N-1)!!}{2\,(N-2)!!}\frac{(N-2)!!}{(N-1)!!}\sqrt{x}\sum_{k=0}^{(N-2)/2}\frac{(2k-1)!!}{(2k)!!}\left(\sqrt{1-x}\right)^{2k} \\ &= \sqrt{x}\sum_{k=0}^{\frac{N-2}{2}}\frac{(2k-1)!!}{(2k)!!}(1-x)^{k}. \qquad \square \end{aligned}$$

B. Performance and Robustness

We next present several simulation studies examining various aspects of performance for Bayesian MMSE error estimators in the Gaussian model. In the first section, we provide performance results for Bayesian error estimators that correctly assume circular Gaussian distributions, thus demonstrating performance under true modeling assumptions. The next section then simulates the same Bayesian error estimators under non-circular Gaussian distributions, which is intended to show the performance under false circular Gaussian modeling assumptions.

In the third section, we graph performance under Johnson distributions, which are outside the assumed Gaussian model. These simulations show how robust Bayesian error estimators are relative to the Gaussian assumption. This is important in practice since we cannot guarantee Gaussianity. We show that performance does require nearly Gaussian distributions, but there is some degree of flexibility (in skewness and kurtosis). This section is followed by a presentation of empirical performance on real data from a breast cancer study.

Finally, in the last section we present an example demonstrating the average performance for a Bayesian error estimator over all distributions using proper priors. We show that performance is superior over all sample sizes on average, and also verify that these error estimators are unbiased under correct modeling assumptions.

1. Performance in True Circular Gaussian Modeling Assumptions

In this section, we provide several synthetic Monte-Carlo simulation results comparing error estimators under circular Gaussian distributions. In all simulations, the mean of class 0 is fixed at $\mu_0 = [0, 0, \ldots, 0]$ and the mean of class 1 at $\mu_1 = [1, 0, \ldots, 0]$. Throughout most of this chapter, the covariance of each class is chosen to make the distributions mirror images with respect to the hyperplane between the two means. This plane is the optimal linear classifier and the classifier designed from the data is meant to approximate it. In this section, the covariance of both classes are scaled identity matrices, with the same scaling factor, denoted σ^2, in both classes, i.e., $\Sigma_0 = \Sigma_1 = \sigma^2 I_D$. The scale of the covariance matrix, σ^2, is used to control Bayes error, where a low Bayes error corresponds to a small variance and high Bayes error to high variance.

We fix $c = 0.5$ and generate a random sample by first determining the sample size for each class using a binomial(n, c) experiment. Each sample point is then assigned a vector according to the Gaussian distribution of its class. The sample is used to train an LDA classifier defined by

$$\mathbf{a} = \widehat{\Sigma}^{-1} \left(\widehat{\mu}_1 - \widehat{\mu}_0\right) \quad \text{and} \quad b = -\frac{1}{2}\mathbf{a}^T \left(\widehat{\mu}_1 + \widehat{\mu}_0\right) + \ln \frac{n_1}{n_0},$$

where the pooled covariance matrix, $\widehat{\Sigma}$, is given by

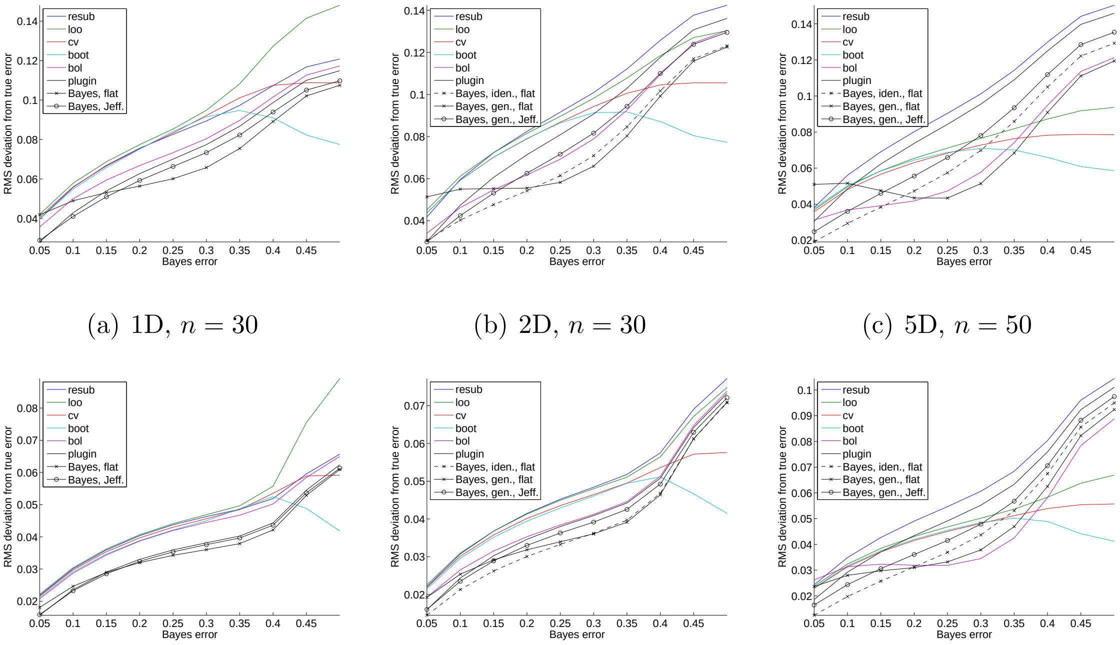
$$\widehat{\Sigma} = \frac{(n_0 - 1)\widehat{\Sigma}_0 + (n_1 - 1)\widehat{\Sigma}_1}{n_0 + n_1 - 2}.$$

The estimates of the mean and covariance for each class are the usual ones given in (4.1), and the true error of this classifier is calculated via (4.7) using the fixed true distribution parameters.

The same sample is used to find 5 non-parametric error estimates (resubstitution, leave-one-out, cross-validation, 0.632 bootstrap, and bolstered resubstitution) and the plug-in error estimate for the designed classifier, and ultimately the squared deviation of each estimate with respect to the true error. The plug-in estimate is computed using the usual estimates of the mean and covariance and the *a priori* class probability estimate $\widehat{c} = \frac{n_0}{n}$.

Up to three Bayesian MMSE error estimators are also evaluated, using the simple improper priors in (4.3) with $S = 0$ and $\nu = 0$ ($\mathbf{m}$ does not matter because $\nu = 0$). Two of these assume general covariances, one with $\kappa + D + 2 = 0$ (flat priors) and one with $\kappa = 0$ (Jeffreys' rule prior), and the last assumes scaled identity covariances with $\kappa + D + 2 = 0$ (flat priors). In cases with only one feature, the Bayesian error estimators assuming scaled identity covariances are the same as the ones assuming general covariances, so only two Bayesian error estimators are provided. Since closed-form equations are available from (4.12), these error estimates can be computed very quickly, and this entire process is repeated 100,000 times to find a Monte-Carlo approximation for the RMS deviation from the true error for each error estimator. For all Bayesian error estimators, in the event where the number of samples in one class is so small that the posteriors used to find the Bayesian error estimator cannot be normalized, κ is increased until the posterior is valid.

Figure 18 shows the RMS error of all error estimators with respect to Bayes error. We see that the Bayesian MMSE error estimator for general covariances using a flat prior is best for distributions with moderate Bayes error, but poor for very small or large Bayes error. A similar result was found in the discrete classification problem. Bolstered resubstitution is very competitive with the Bayesian error estimator for general covariances and flat priors, especially in higher dimensions, and it is also very flexible since it can be applied fairly easily to any classifier; however, keep in mind that



(a) 1D, $n = 30$

(b) 2D, $n = 30$

(c) 5D, $n = 50$

(d) 1D, $n = 100$

(e) 2D, $n = 100$

(f) 5D, $n = 100$

Fig. 18. RMS deviation from true error for Gaussian distributions with respect to Bayes error.

bolstering is known to perform particularly well with circular densities (uncorrelated equal variance features) like those in this example.

The Bayesian error estimator for general covariances using Jeffreys' rule prior ($\kappa = 0$) shifts performance in favor of lower Bayes error. Recall from the form of the priors in (4.3) that a larger κ will put more weight on covariances with a small determinant (usually corresponding to a small Bayes error) and less weight on those with a large determinant (usually corresponding to a large Bayes error). If the Bayes error is indeed very small, then the Bayesian error estimator using Jeffreys' rule prior is usually the best followed by the plug-in rule, which performs exceptionally well because the sample mean and sample variance are very accurate even with a small sample.

Finally, regarding Fig. 18 note that with a larger number of features the Bayesian error estimator assuming scaled identity covariances tends to be better than the one assuming general covariances with $\kappa = 0$ over the entire range of Bayes error. This makes clear the benefit of using more constrained assumptions, as long as the assumptions are correct.

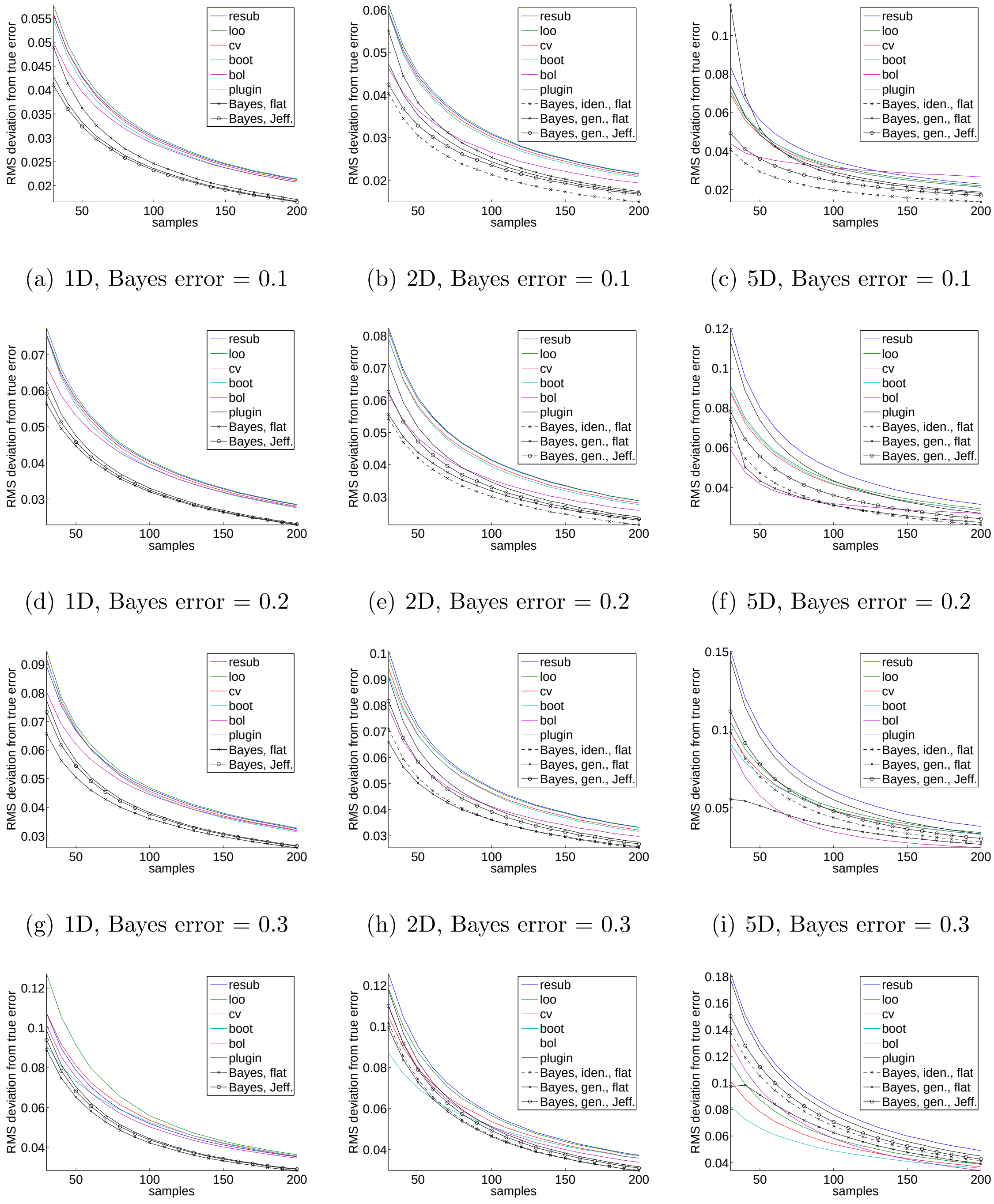
We graph RMS error with respect to sample size in Fig. 19 for 1, 2 and 5 dimensions, and Bayes errors of 0.1, 0.2, 0.3 and 0.4. Graphs like these can be used to determine the sample size needed to guarantee a certain RMS. As the sample size increases, the parametrically based error estimators (the plug-in rule and Bayesian MMSE error estimators) tend to converge to zero much more quickly than the distribution-free error estimators. For example, all of the simulations using one feature (the left column of Fig. 19) clearly separate the parametric and distribution-free error estimators. This is not surprising since for a large sample the sample parameter estimates tend to be very accurate.

Bayesian MMSE error estimators can improve greatly on traditional error estimators. For only one feature, the benefit is clear, especially for moderate Bayes error like in parts (d) and (g) of Fig. 19. In higher dimensions, there are many options to constrain the covariance matrix and choose different priors, so the picture is more complex.

2. Robustness to False Circular Gaussian Modeling Assumptions

The Bayesian MMSE error estimator assuming identity covariances performs very well in many cases in Section IV.B.1, but in these simulations the identity covariance assumption is correct. We consider two examples to investigate robustness relative to the inaccuracy of this assumption.

For the first example, define ρ to be the correlation coefficient for class 0 in a two



(a) 1D, Bayes error = 0.1

(b) 2D, Bayes error = 0.1

(c) 5D, Bayes error = 0.1

(d) 1D, Bayes error = 0.2

(e) 2D, Bayes error = 0.2

(f) 5D, Bayes error = 0.2

(g) 1D, Bayes error = 0.3

(h) 2D, Bayes error = 0.3

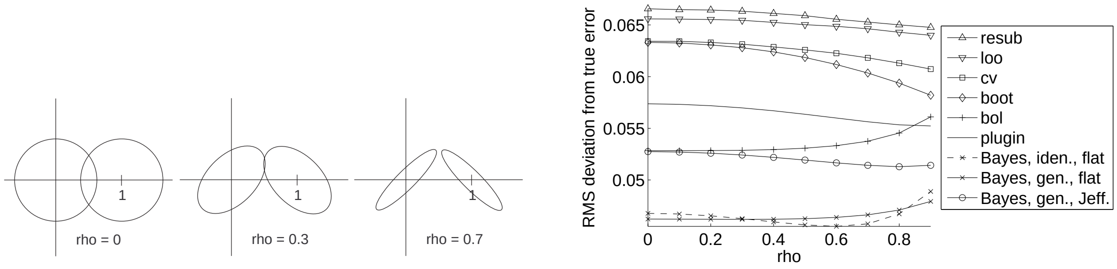
(i) 5D, Bayes error = 0.3

(j) 1D, Bayes error = 0.4

(k) 2D, Bayes error = 0.4

(l) 5D, Bayes error = 0.4

Fig. 19. RMS deviation from true error for Gaussian distributions with respect to sample size.



(a) Distributions used in RMS graphs

(b) RMS deviation from true error

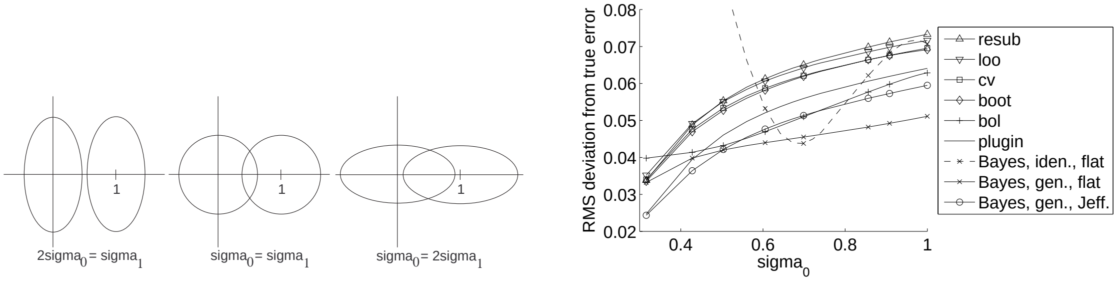
Fig. 20. Gaussian distributions varying correlation (2D, $\sigma = 0.7413$, $n = 50$).

feature problem. The correlation coefficient for class 1 is $-\rho$ to ensure mirror image distributions. Thus, the covariance matrices are given by

$$\Sigma_0 = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \quad \text{and} \quad \Sigma_1 = \begin{bmatrix} \sigma^2 & -\rho\sigma^2 \\ -\rho\sigma^2 & \sigma^2 \end{bmatrix}.$$

Illustrations of the distributions used in this experiment are shown in Fig. 20(a) and simulation results are shown in Fig. 20(b). For the simulations, we fix $\sigma^2 = 0.7413^2$, which corresponds to a Bayes error of 0.25 when there is no correlation. The Bayesian error estimators assuming general covariances are not affected by correlation very much, and interestingly the performance of the error estimator assuming identity covariances is also fairly robust to correlation in this particular model, although some degradation can be seen for $\rho > 0.8$. Meanwhile, bolstering also appears to be somewhat negatively affected by high correlation, probably owing to the use of spherical kernels when the true distributions are skewed.

In Fig. 21, we present a second experiment using different variances for each



(a) Distributions used in RMS graphs

(b) RMS deviation from true error

Fig. 21. Gaussian distributions varying σ_0 (2D, $\sigma = 0.7413$, $n = 50$).

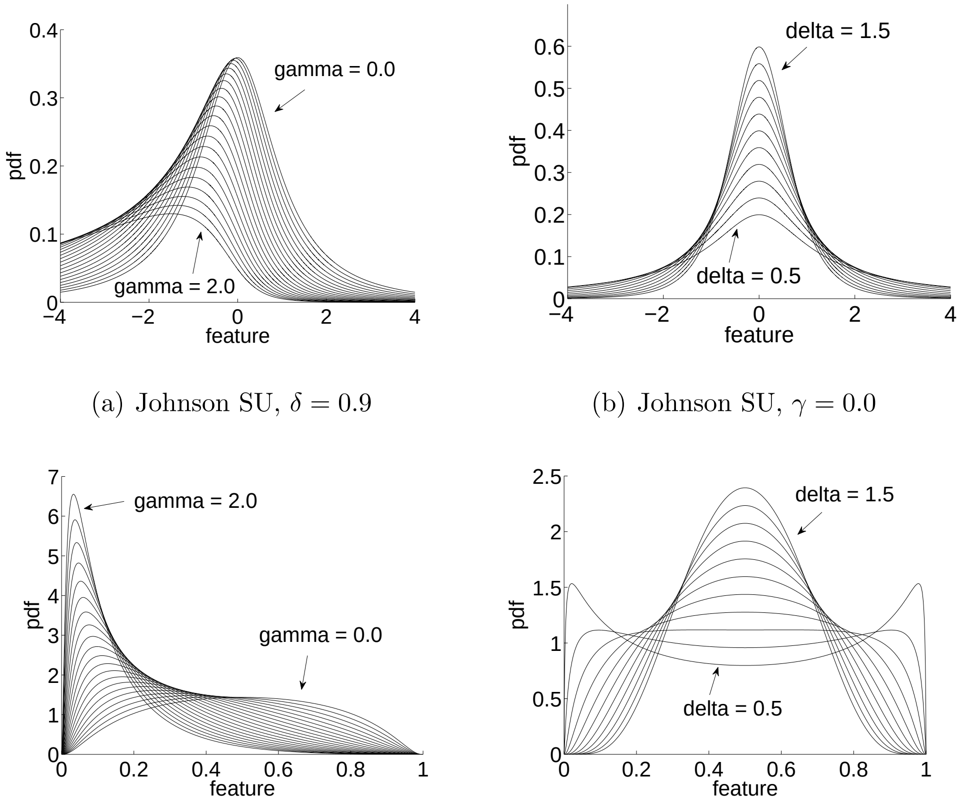
feature. The covariances are given by

$$\Sigma_0 = \Sigma_1 = \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix},$$

and we fix the average variance between the classes so that $\frac{1}{2}(\sigma_0^2 + \sigma_1^2) = 0.7413^2$. When $\sigma_0^2 = \sigma_1^2$, the Bayes error of the classification problem is again 0.25. These simulations show that the Bayesian error estimator assuming identity covariances can be highly sensitive to unbalanced features, however this problem may be alleviated by normalizing the raw data.

3. Robustness to False Gaussian Modeling Assumptions

Since Bayesian error estimators depend on parametric models of the true distributions, one may apply a Kolmogorov-Smirnov normality test or other hypothesis test to discern if a sample deviates substantially from being Gaussian; nevertheless, the actual distribution is very unlikely to be truly Gaussian, so we need to investigate robustness relative to the Gaussian assumption. To explore this issue in a systematic setting, we have applied Bayesian MMSE error estimators to Johnson distributions



(a) Johnson SU, $\delta = 0.9$

(b) Johnson SU, $\gamma = 0.0$

(c) Johnson SB, $\delta = 0.9$

(d) Johnson SB, $\gamma = 0.0$

Fig. 22. Johnson Distributions with one parameter fixed and the other varying in increments of 0.1 ($\eta = 0$, $\lambda = 1$).

in 1 dimension. Johnson distributions are a flexible family of distributions with four free parameters, including mean and variance [97, 98]. There are two main classes in the Johnson system of distributions: Johnson SU (for unbounded) and Johnson SB (for bounded). The normal and log-normal distributions are also considered classes in this system, and in fact they are limiting cases of the SU and SB distributions.

The Johnson system can be summarized as follows. If Z is a unit normal random variable, then X is Johnson if $(Z - \gamma)/\delta = f((X - \eta)/\lambda)$, where f is a simple function satisfying some desirable properties such as monotonicity [97, 98]. For log-normal distributions, $f(y) = \log(y)$; for Johnson SU distributions, $f(y) = \sinh^{-1}(y)$; and for Johnson SB distributions, $f(y) = \log(y/(1-y)) = 2\tanh^{-1}(2y-1)$. For reference, example graphs of these distributions are given in Fig. 22. Johnson SU distributions are always unimodal, while SB distributions can also be bimodal. In particular, an SB distribution is bimodal if $\delta < \frac{1}{\sqrt{2}}$ and $|\gamma| < \delta^{-1}\sqrt{1-2\delta^2} - 2\delta\tanh^{-1}\sqrt{1-2\delta^2}$.

The parameters γ and δ control the shape of the Johnson distribution and together essentially determine its skewness and kurtosis, which are normalized third and fourth moments. In particular, skewness is equal to μ_3/σ^3 and kurtosis is μ_4/σ^4, where μ_n is the nth mean-adjusted moment of a random variable and $\sigma^2 = \mu_2$ is the variance. Skewness and kurtosis are very useful statistics to measure normality; Gaussian distributions always have a skewness of 0 and kurtosis of 3. For Johnson distributions, skewness is more influenced by γ and kurtosis by δ, but the relationship is not exclusive. Once the shape of the distribution is determined, η and λ are chosen to fix the mean and variance.

Figure 23 illustrates the values of skewness and kurtosis obtainable within the Johnson family. The region below the log-normal line can be achieved with Johnson SU distributions, while the region above can be achieved with Johnson SB distributions. In fact, the normal, log-normal, SU and SB systems uniquely cover the entire obtainable region of the skewness/kurtosis plane, so there is just one distribution corresponding to each skewness/kurtosis pair. For all distributions, kurtosis $\geq$ skewness^2 + 1, where equality corresponds to a two point distribution (taking on two values, one with probability p and the other with $1-p$).

In this figure, $\gamma = 0$ corresponds to points on the left axis. The dotted diagonal lines represent skewness and kurtosis obtainable with SU distributions and fixed values of δ. As we increase δ, these lines move up in an almost parallel manner. As we increase γ, kurtosis increases along with skewness until we converge to a point on the

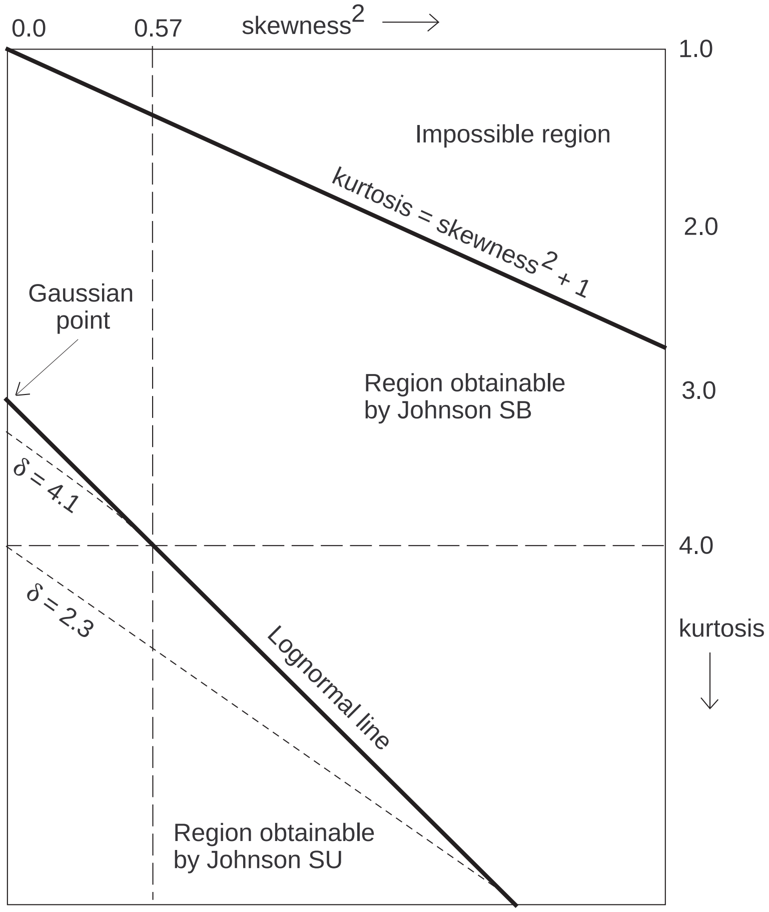
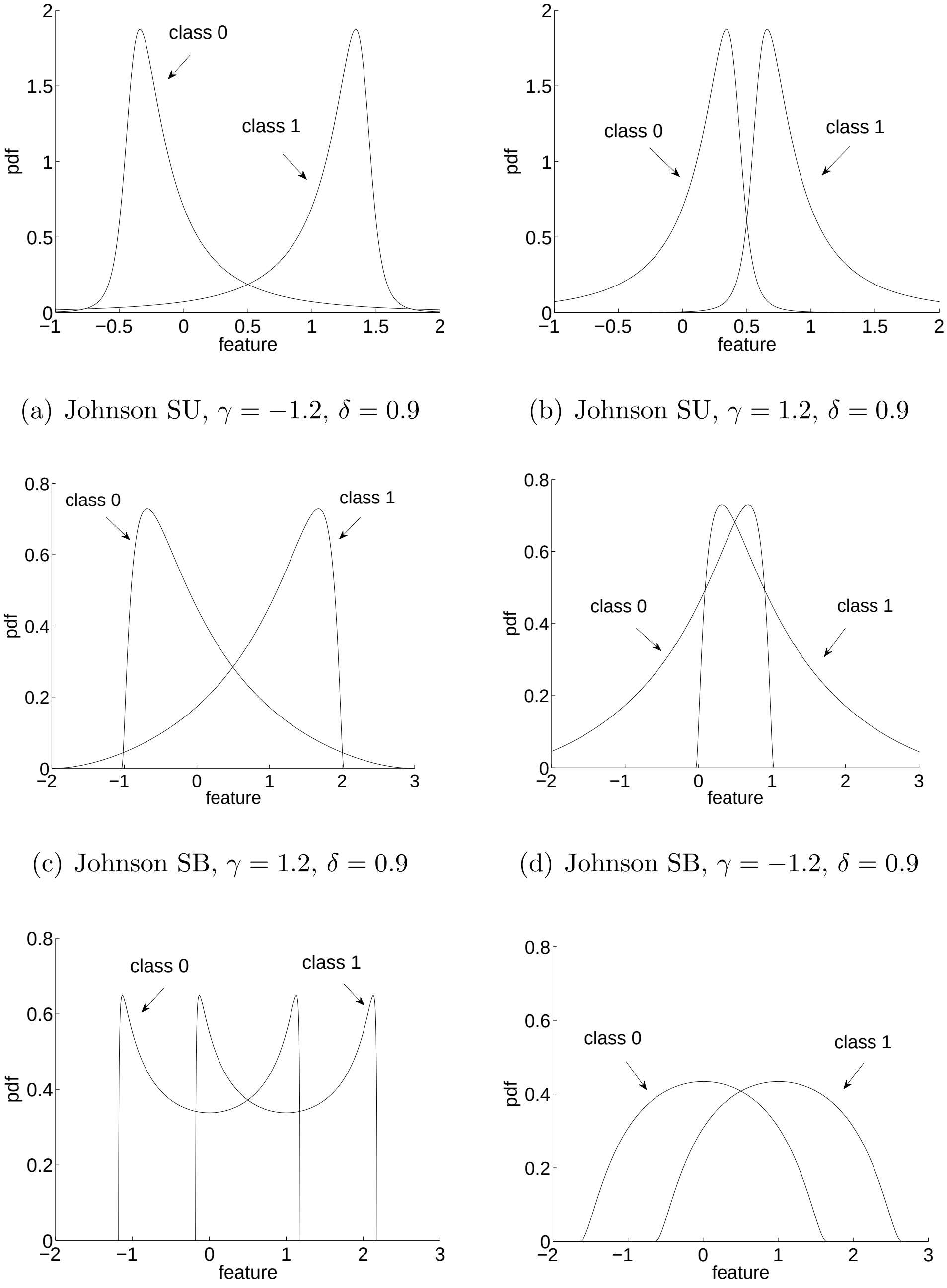


Fig. 23. Skewness and kurtosis obtainable regions for Johnson distributions.

log-normal line. As a quick example, suppose we fix kurtosis at 4.0. We must have $\delta > 2.3$, which is limited by the worst case where $\gamma = 0$. Also, with SU distributions we can only obtain a maximum skewness of about 0.75 (or square skewness of 0.57), which is achieved using $\delta \approx 4.1$ and γ very large.

The simulation procedure in this section is the same as that in Section IV.B.1, except the sample points are each assigned a Johnson distributed value rather than Gaussian. We use mirror images of the same Johnson distribution for both classes;



(a) Johnson SU, $\gamma = -1.2$, $\delta = 0.9$

(b) Johnson SU, $\gamma = 1.2$, $\delta = 0.9$

(c) Johnson SB, $\gamma = 1.2$, $\delta = 0.9$

(d) Johnson SB, $\gamma = -1.2$, $\delta = 0.9$

(e) Johnson SB, $\gamma = 0.0$, $\delta = 0.5$

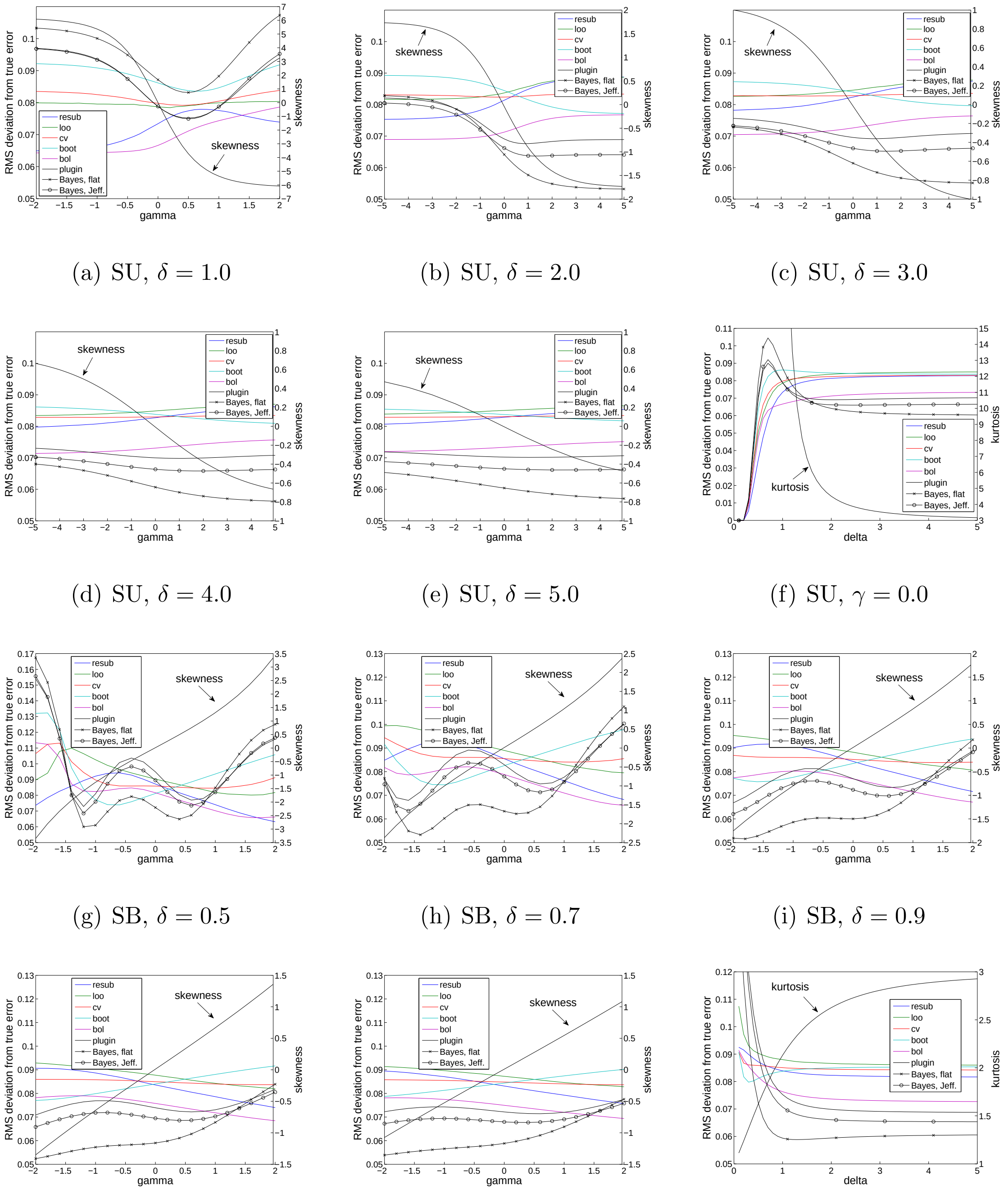
(f) Johnson SB, $\gamma = 0.0$, $\delta = 0.9$

Fig. 24. Two class problems with Johnson distributions (1D, $\sigma^2 = 0.7413^2$).

examples are shown in Fig. 24. In the following, the parameters γ and δ refer to that of class 0, while class 1 has the same δ and negative γ. Meanwhile, for each class η and λ are selected to give the appropriate mean and covariance. The sample size is fixed at $n = 30$, the means are always fixed at $\mu_0 = 0$ and $\mu_1 = 1$, and the covariances are fixed at $\sigma^2 = 0.7413^2$, which corresponds to a Bayes error of 0.25 for the Gaussian distribution. From Fig. 18(a), note with one feature, $n = 30$ and Gaussian distributions with a Bayes error of 0.25 that the Bayesian error estimators using the flat prior and Jeffreys' rule prior perform quite well with RMSs of about 0.060 and 0.066, respectively. These are followed by the plug-in rule with an RMS of 0.070 and bolstering with an RMS of 0.073. We wish to observe whether the Bayesian error estimation remains superior after distorting the skewness and kurtosis of the original Gaussian distributions using Johnson distributions.

Figures 25(a) through 25(f) show the RMS of all error estimators for various Johnson SU distributions, and Figs. 25(g) through 25(l) show analogous graphs for Johnson SB distributions. In each sub-figure, we fix either δ or γ and vary the other parameter to observe a slice of the performance behavior. The scale for the RMS error of all error estimators is provided on the left axis as usual, and a graph of either skewness (when δ is fixed) or kurtosis (when γ is fixed) as also been added and labeled with a arrow, with the scale shown on the right axis. These skewness and kurtosis graphs help illustrate the non-Gaussianity of the distributions represented by each point.

Figure 25(f) presents a simulation observing the effect of δ (which has more influence on kurtosis) with SU distributions and $\gamma = 0$. For $\gamma = 0$ there is no skewness, and this graph shows that the Bayesian error estimator with flat priors requires δ to be at least 1.5 before it surpasses all of the other error estimators (in this case the next best is bolstering). This corresponds to a kurtosis of about 7.0. A similar graph of



(a) SU, $\delta = 1.0$

(b) SU, $\delta = 2.0$

(c) SU, $\delta = 3.0$

(d) SU, $\delta = 4.0$

(e) SU, $\delta = 5.0$

(f) SU, $\gamma = 0.0$

(g) SB, $\delta = 0.5$

(h) SB, $\delta = 0.7$

(i) SB, $\delta = 0.9$

(j) SB, $\delta = 1.1$

(k) SB, $\delta = 1.3$

(l) SB, $\gamma = 0.0$

Fig. 25. RMS deviation from true error for Johnson SU and SB distributions (1D, $\sigma = 0.7413$, $n = 30$). Right axis show skewness, or in (f) and (l) kurtosis.

performance with Johnson SB distributions and $\gamma = 0$ is given in Fig. 25(l), in which the same error estimator is the best whenever $\delta > 0.4$, corresponding to kurtosis greater than about 1.5. So although Gaussian distributions have a kurtosis of 3.0, in this example the Bayesian MMSE error estimator is still better than all of the other error estimators whenever there is no skewness and kurtosis is between 1.5 and 7.0.

Interestingly, performance can actually improve as we move away from Gaussianity. For example, although it appears in Fig. 25(a) that the Bayesian error estimators dip in the middle when $\delta = 1.0$ (which is expected since $\gamma = 0$ for Gaussian distributions), for larger δ the RMS of the Bayesian estimators seem to monotonically decrease with γ, as in Fig. 25(b), suggesting that they favor negative skewness (positive γ) where the classes are skewed away from each other. Simulations with Johnson SB distributions also appear to favor slight negative skewness (negative γ), although RMS graphs are not monotonic.

Finally, in Fig. 26 we present a graph summarizing the performance of Bayesian error estimators on Johnson distributions with respect to the skewness and kurtosis of class 0. The skewness-kurtosis plane shown in this figure is essentially the same as that illustrated in Fig. 23, but also showing two sides to distinguish between positive and negative skewness. Note performance for either positive or negative skewness is distinct: when class 0 has positive skewness the distributions are skewed toward each other (for mirror image distributions the kurtosis of class 1 is the same but skewness is negative), and similarly when class 0 has negative skewness the distributions are skewed away from each other. Each of the dots in Fig. 26 represent fixed class-conditional Johnson distributions, for example the pairs shown in Fig. 24. As before, we fix $\sigma^2 = 0.7413^2$, corresponding to a Bayes error of 0.25 for the Gaussian distribution (which has a skewness 0 and kurtosis 3). All of the performance results shown in Fig. 25 were included, along with a battery of several other simulations

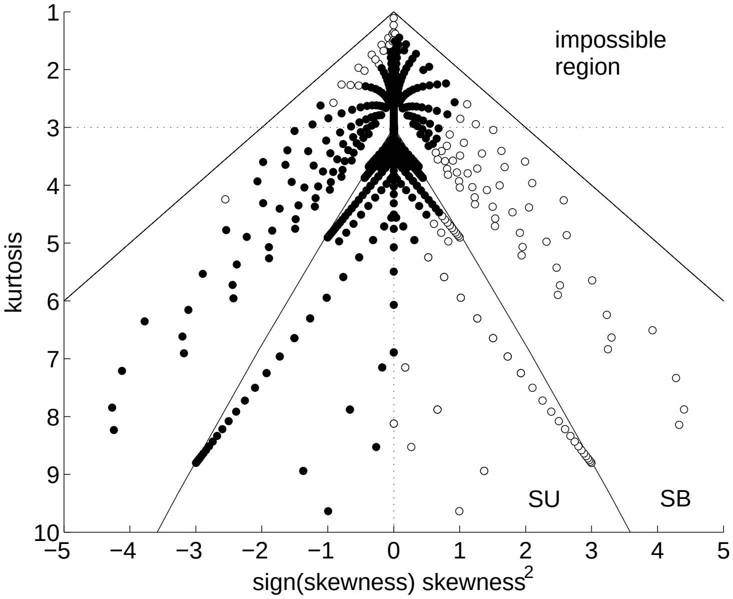


Fig. 26. RMS deviation from true error for Johnson distributions varying both skewness and kurtosis (1D, $\sigma = 0.7413$, $n = 30$). Black dots are where the Bayesian MMSE error estimator is best, white dots are where any other error estimator is best.

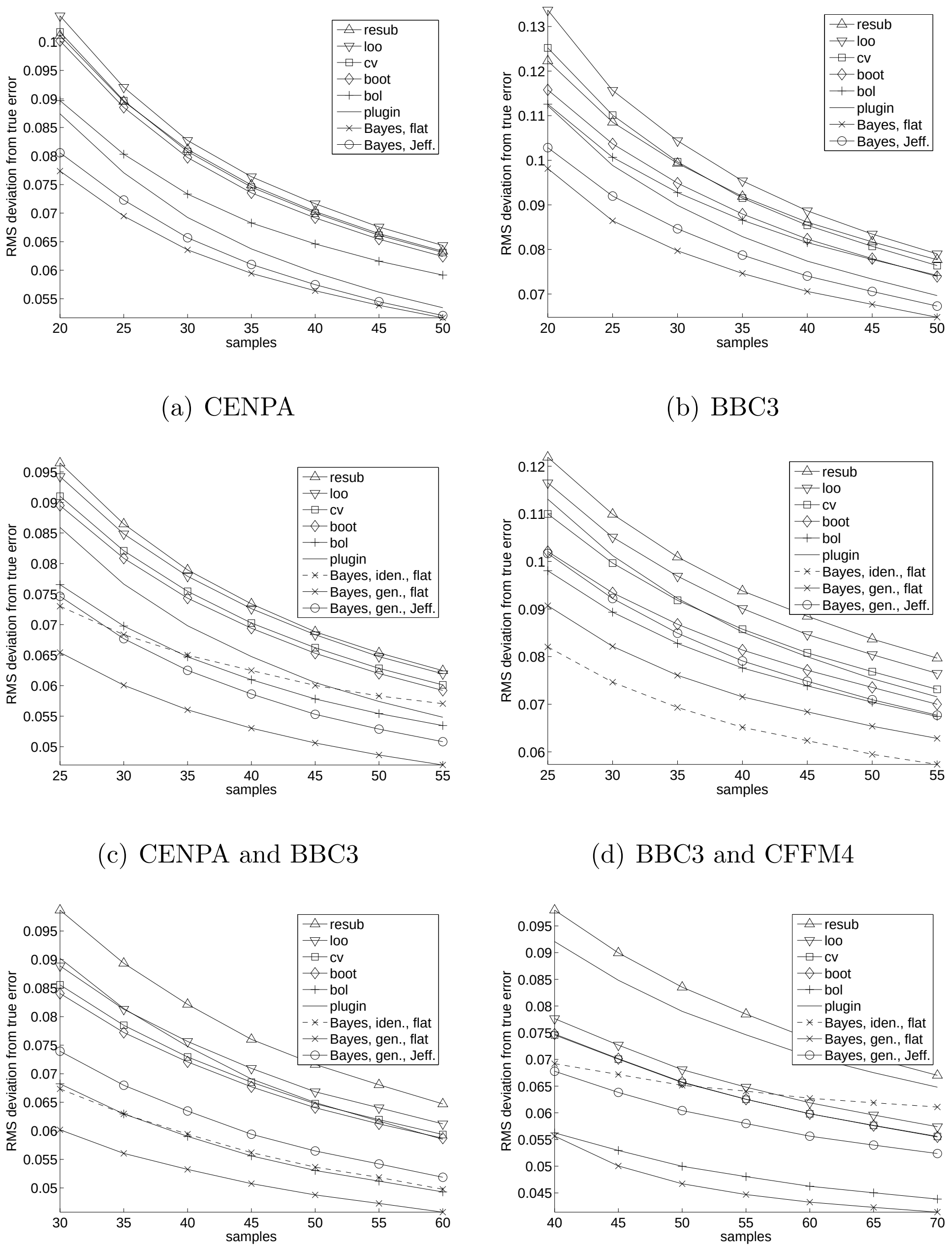
covering different ranges for γ and δ.

Black dots in Fig. 26 represent distributions where the Bayesian MMSE error estimator with flat priors performs better than all of the other six standard error estimators, while white dots pessimistically represent distributions where any other error estimator was better. With one feature, $n = 30$ and $\sigma^2 = 0.7413^2$, the black dots cover a relatively large range of skewness and kurtosis (especially with negative skewness), indicating that Bayesian error estimators can be used relatively reliably even if the true distributions are not perfectly Gaussian. Similar graphs or studies may be used to determine an "acceptable" region for Gaussian modeling assumptions, which may be useful for designing hypothesis tests. However performance in this graph depends heavily on the simulation settings, for instance notice in Fig. 18(a) with one feature, $n = 30$ and a Bayes error of 0.45 that the Bayesian error estimator is not the best error estimator even for Gaussian distributions, let alone Johnson distributions.

4. Performance on Real Breast Cancer Data

We have applied the three non-informative Bayesian error estimators from the previous sections to normalized gene-expression measurements from a breast cancer study [99]. The study included 295 sample points, with 180 assigned to class 0 (good prognosis) and 115 in class 1 (bad prognosis). From the original 295 points, we randomly draw a non-stratified training sample of size n and use the remaining sample points as holdout data to approximate the true error. This process is repeated 100,000 times to estimate the average RMS deviation of each error estimator from the true error. In this analysis, we consider several combinations of 5 genes picked in [28]: CENPA, BBC3, CFFM4, TGFB3 and DKFZP564D0462. For all feature sets considered, a multivariate Shapiro-Wilk test applied to the full data set does not reject Gaussianity over either of the classes at a 95% significance level.

Performance for sample sizes between 20 and 70 are shown in Fig. 27. The Bayesian error estimator assuming general covariances with flat priors usually performs quite well compared to the other error estimators and the error estimator assuming general covariances with Jeffreys' rule prior ($\kappa = 0$) is a decent performer, especially for a small number of dimensions. That the flat prior seems to perform better than Jeffreys' rule prior is likely due to a fairly high Bayes error in these cases; the flat prior defines a smaller κ ($\kappa = -D-2$ versus $\kappa = 0$) and is therefore better for higher Bayes errors. If it is supposed before the experiment that the Bayes error is in some range, this information can be used to select which prior is more appropriate



(a) CENPA

(b) BBC3

(c) CENPA and BBC3

(d) BBC3 and CFFM4

(e) CENPA, BBC3 and CFFM4

(f) all 5 genes

Fig. 27. RMS deviation from true error for empirical measurements from a breast cancer study.

to use. In two or more dimensions, the Bayesian error estimator assuming identity covariances sometimes performs very well, as in Fig. 27(d); however, its performance advantage can be lost as sample size grows, as in Fig. 27(c). Contributing factors are that there can be large differences between the variances of the features, and no attempt has been made to avoid correlation.

5. Average Performance Using Proper Priors

Finally, we present an example illustrating the average performance of Bayesian error estimators over all distributions in a model. To average over all distributions we require proper priors, so for the sake of demonstration we will use a carefully designed proper prior in this section rather than the improper priors used previously. Define $\Lambda_y = \Sigma_y$ to allow general covariances, and for both classes define the prior hyperparameters $\kappa = \nu = 5D$ and $S = (\kappa - D - 1)\, 0.7413^2 I_D$. For class 0 also define $\mathbf{m} = [0, 0, \ldots, 0]$, and for class 1 define $\mathbf{m} = [1, 0, \ldots, 0]$. For each class, this prior is always proper and can be interpreted as the information available if we have observed 5 samples per feature before the experiment with sample mean $\mathbf{m}$ and covariance $0.7413^2 I_D$, in the sense that this would be the posterior distribution if we had started with a uniform prior and then observed this sample. In addition, we assume a uniform distribution for the class probabilities, c.

We randomly generate 100,000 feature-label distributions–each determined by a random class probability, c, and a set of means and covariances, μ_y and Σ_y for $y \in \{0, 1\}$, which were generated independently for each class according to the distribution of the priors in (4.2). For each fixed feature-label distribution, we generated 10 sets of samples, each used to train a classifier. The true error, all classical error estimator used before and the Bayesian error estimator with correct priors are evaluated as usual. These results were all averaged to produce Monte-Carlo approximations of

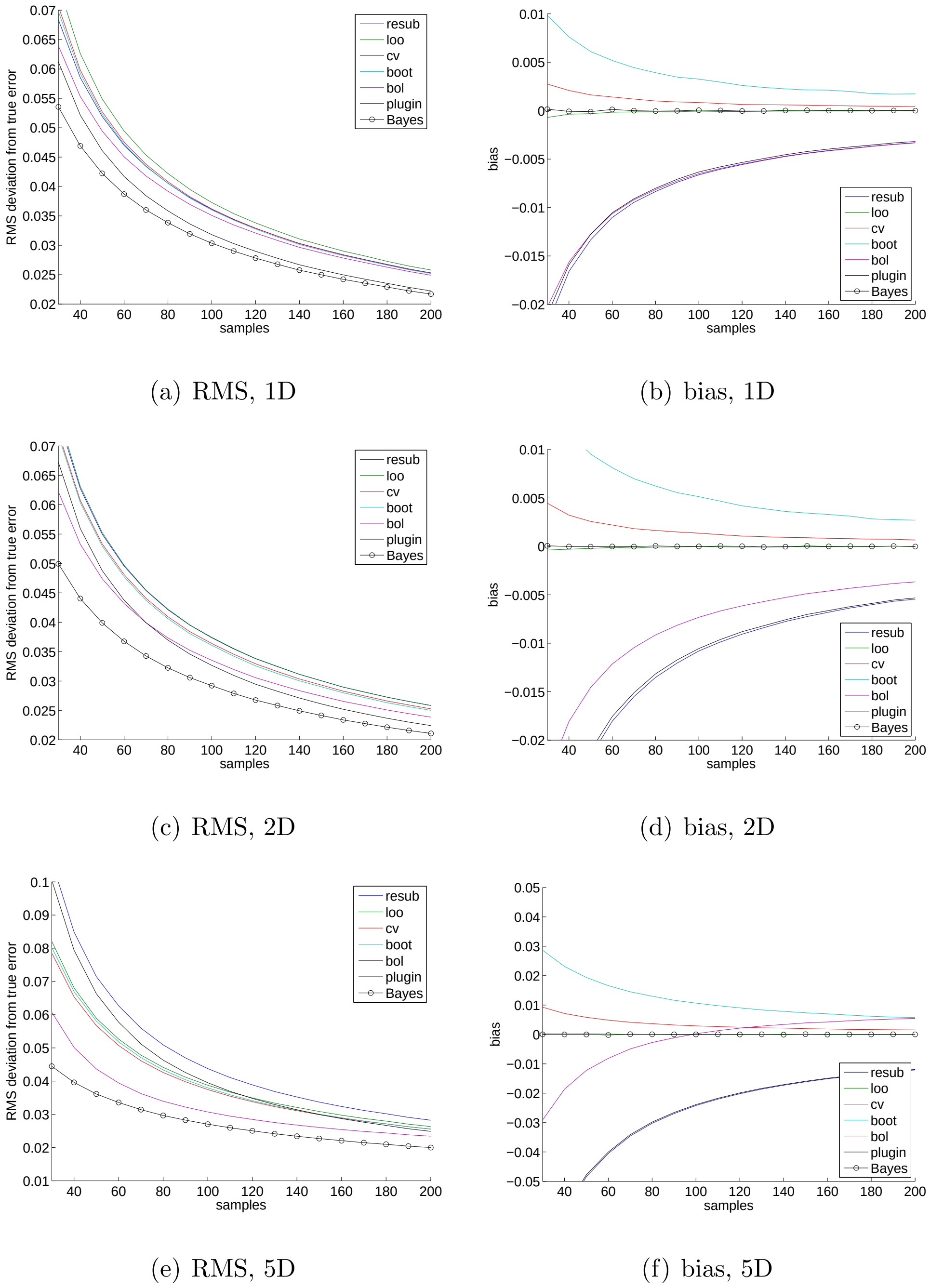
RMS and bias over all distributions and sample sets for 1, 2 and 5 features, as shown in Fig. 28.

These graphs validate that Bayesian error estimators, when averaged over all distributions in the parameterized family and assuming the specified priors are true, have optimal RMS performance and are unbiased for each sample size. In fact, the performance of the Bayesian error estimator improves significantly relative to the other error estimators as we increase the number of features. However, these results only speak for average performance over all feature-label distributions with respect to a specific prior; RMS and bias can both be poor for specific distributions.

C. Discussion

In this chapter, we have presented closed-form expressions for Bayesian MMSE error estimators applied to Gaussian distributions with a very general class of priors and linear classification. Simulation results show that even non-informative Bayesian error estimators can improve significantly upon traditional error estimators. Furthermore, since most performance results reported here utilize non-informative priors, there is potential to improve results further by tailoring the priors for the experiment at hand. We have also provided simulation results for Johnson distributions, which show that Bayesian error estimators are fairly robust to false modeling assumptions; nevertheless, for the sake of prudence this error estimator should be used in conjunction with hypothesis tests or a thorough examination of the problem to verify the appropriateness of the modeling assumptions.

Robustness is a crucial issue for Bayesian error estimation because performance can be seriously degraded when the feature-label distribution corresponding to the data is not contained within the family of distributions covered by the model. This



(a) RMS, 1D

(b) bias, 1D

(c) RMS, 2D

(d) bias, 2D

(e) RMS, 5D

(f) bias, 5D

Fig. 28. RMS deviation from true error and bias for linear classification of Gaussian distributions, averaged over all distributions and samples using a proper prior.

issue is especially problematic in the case of small samples, precisely the situation in which Bayesian error estimation can be most beneficial. But, as noted in the conclusion of the previous chapter, “model-free” error estimators are only model-free in the sense that no model is used in their calculation. In fact, their performance is strongly dependent on the feature-label distribution so that their use is not model-free. In the case of Bayesian error estimators, modeling assumptions are explicit so that it is possible to obtain concrete answers to questions regarding optimality and performance bounds, whereas for “model-free” error estimators it is typically the case that nothing is known of the validity of the estimate. Moreover, if we are willing to add an extra step to the error estimation process, where we define a model and test the observed sample for fitness in the model, then we can mitigate concern regarding model assumptions and obtain a superior error estimator. A key aspect of this work is that it directly confronts the necessity of assumptions by stating them outright. In this way, Bayesian error estimators rigorously address the trade-off between accuracy (closeness to the true error) and robustness (modeling assumptions).

That being said, there remains the critical practical issue of defining an appropriate model and level of robustness for a given experiment and sample size. In our Bayesian approach, assumptions can be made on several levels. At the highest level, we can define a larger or smaller family of distributions to consider in the model. A few important factors to consider in this stage are model validity (are the samples sufficiently Gaussian?), the number of degrees of freedom (parameters) in the model that can be handled given the sample size, and the availability of a closed-form solution. Once a model has been determined, we can restrict the parameter space to reduce the number of degrees of freedom, as we have in the Gaussian model assuming scaled identity covariance matrices. Finally, the investigator has the option to tune the prior probabilities of the distribution parameters to take advantage of prior knowledge or otherwise manipulate the probability density of the parameters. Non-informative priors generate a more robust estimator, though with a higher Bayesian expected loss. Alternatively, informed priors may not be as robust but offer decreased expected loss as long as one has fairly accurate knowledge concerning the model parameters.

CHAPTER V

EXACT SAMPLE-CONDITIONED MSE PERFORMANCE OF BAYESIAN MMSE ERROR ESTIMATORS*

A. Definition of the Sample-Conditioned MSE

There are two sources of randomness in the Bayesian model. The first is the sample, which also randomizes the designed classifier and its true error. Almost all current results on error estimator performance are averaged over random samples, which demonstrates performance relative to a fixed classification rule. The second source of randomness, which is the focus of this work, is uncertainty in the underlying feature-label distribution. The Bayesian error estimator addresses the second source of randomness, naturally giving rise to a practical expected measure of performance given a fixed sample and classifier.

We fix the sample and consider the conditional MSE, which is exactly the objective function optimized by the Bayesian MMSE error estimator. According to MMSE estimation theory, we may apply the orthogonality principle:

$$\begin{aligned}
\mathrm{MSE}(\widehat{\varepsilon}|S_n) &= \mathrm{E}_\theta\left[(\varepsilon_n(\theta)-\widehat{\varepsilon})^2|S_n\right] \\
&= \mathrm{E}_\theta\left[(\varepsilon_n(\theta)-\widehat{\varepsilon})\varepsilon_n(\theta)|S_n\right] + \mathrm{E}_\theta\left[(\varepsilon_n(\theta)-\widehat{\varepsilon})\widehat{\varepsilon}|S_n\right] \\
&= \mathrm{E}_\theta\left[(\varepsilon_n(\theta)-\widehat{\varepsilon})\varepsilon_n(\theta)|S_n\right]
\end{aligned}$$

*Reprinted with permission from “Exact Sample Conditioned MSE Performance of the Bayesian MMSE Estimator for Classification Error–Part I: Representation” by L. A. Dalton and E. R. Dougherty, 2012, *IEEE Transactions on Signal Processing*, in press, Copyright 2012 by IEEE.

$$
\begin{aligned}
&= \mathrm{E}_{\theta}\left[(\varepsilon_n(\theta))^2|S_n\right] - (\widehat{\varepsilon})^2 \\
&= \mathrm{Var}_{\theta}\left(\varepsilon_n(\theta)|S_n\right),
\end{aligned}
$$

where we have used the definition of the Bayesian error estimator given in (2.5) and suppressed dependence on the sample in $\varepsilon_n(\theta)$ and $\widehat{\varepsilon}$ to avoid cumbersome notation. That is, the conditional MSE of the Bayesian error estimator is equivalent to the variance of the true error. Thanks to the posterior independence between c, θ_0 and θ_1, we may expand this, via the basic variance identity, to

$$
\begin{aligned}
\mathrm{MSE}(\widehat{\varepsilon}|S_n) &= \mathrm{Var}_{c,\theta_0,\theta_1}\left(c\varepsilon_n^0(\theta_0) + (1-c)\varepsilon_n^1(\theta_1)|S_n\right) \\
&= \mathrm{Var}_c\left(\mathrm{E}_{\theta_0,\theta_1}\left[c\varepsilon_n^0(\theta_0) + (1-c)\varepsilon_n^1(\theta_1)|c, S_n\right]|S_n\right) \\
&\quad + \mathrm{E}_c\left[\mathrm{Var}_{\theta_0,\theta_1}\left(c\varepsilon_n^0(\theta_0) + (1-c)\varepsilon_n^1(\theta_1)|c, S_n\right)|S_n\right].
\end{aligned}
$$

Further decomposing the inner expectation and variance, we have

$$
\begin{aligned}
\mathrm{MSE}(\widehat{\varepsilon}|S_n) &= \mathrm{Var}_c\left(c\widehat{\varepsilon}^0 + (1-c)\widehat{\varepsilon}^1|S_n\right) \\
&\quad + \mathrm{E}_c\left[c^2\mathrm{Var}_{\theta_0}\left(\varepsilon_n^0(\theta_0)|S_n\right) + (1-c)^2\mathrm{Var}_{\theta_1}\left(\varepsilon_n^1(\theta_1)|S_n\right)|S_n\right] \\
&= \mathrm{Var}_{\pi^*}(c)\left(\widehat{\varepsilon}^0 - \widehat{\varepsilon}^1\right)^2 \\
&\quad + \mathrm{E}_{\pi^*}\left[c^2\right]\mathrm{Var}_{\pi^*}\left(\varepsilon_n^0(\theta_0)\right) + \mathrm{E}_{\pi^*}\left[(1-c)^2\right]\mathrm{Var}_{\pi^*}\left(\varepsilon_n^1(\theta_1)\right), \qquad (5.1)
\end{aligned}
$$

where $\widehat{\varepsilon}^0$ and $\widehat{\varepsilon}^1$ are defined in (2.11), and in the last line we have employed our shorthand notation for expectations conditioned on the sample. Therefore, finding the MSE of the Bayesian error estimator boils down to finding the posterior variance of ε_n^0 and ε_n^1. Furthermore, since $\mathrm{Var}_{\pi^*}\left(\varepsilon_n^y(\theta_y)\right) = \mathrm{E}_{\pi^*}\left[(\varepsilon_n^y(\theta_y))^2\right] - (\widehat{\varepsilon}^y)^2$,

$$
\begin{aligned}
\mathrm{MSE}(\widehat{\varepsilon}|S_n) &= -2\mathrm{Var}_{\pi^*}(c)\,\widehat{\varepsilon}^0\widehat{\varepsilon}^1 - \left(\mathrm{E}_{\pi^*}[c]\right)^2(\widehat{\varepsilon}^0)^2 - \left(\mathrm{E}_{\pi^*}[1-c]\right)^2(\widehat{\varepsilon}^1)^2 \\
&\quad + \mathrm{E}_{\pi^*}\left[c^2\right]\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^0(\theta_0)\right)^2\right] + \mathrm{E}_{\pi^*}\left[(1-c)^2\right]\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^1(\theta_1)\right)^2\right]. \qquad (5.2)
\end{aligned}
$$

The variance and expectations related to the variable c depend on our prior model for c, but are straightforward to find analytically. For example, if the prior distribution of c is beta with hyperparameters α^0 and α^1, which holds with $\alpha^0 = \alpha^1 = 1$ when c has a uniform prior, then the posterior of c is also beta with hyperparameters $\alpha^0 + n_0$ and $\alpha^1 + n_1$ and,

$$E_{\pi^*}\left[c\right] = \frac{\alpha^0 + n_0}{\alpha^0 + \alpha^1 + n}, \tag{5.3}$$

$$E_{\pi^*}\left[1 - c\right] = \frac{\alpha^1 + n_1}{\alpha^0 + \alpha^1 + n}, \tag{5.4}$$

$$E_{\pi^*}\left[c^2\right] = \frac{(\alpha^0 + n_0)(\alpha^0 + n_0 + 1)}{(\alpha^0 + \alpha^1 + n)(\alpha^0 + \alpha^1 + n + 1)}, \tag{5.5}$$

$$E_{\pi^*}\left[(1 - c)^2\right] = \frac{(\alpha^1 + n_1)(\alpha^1 + n_1 + 1)}{(\alpha^0 + \alpha^1 + n)(\alpha^0 + \alpha^1 + n + 1)}, \tag{5.6}$$

$$\mathrm{Var}_{\pi^*}\left(c\right) = \frac{(\alpha^0 + n_0)(\alpha^1 + n_1)}{(\alpha^0 + \alpha^1 + n)^2(\alpha^0 + \alpha^1 + n + 1)}. \tag{5.7}$$

Hence,

$$\begin{aligned}\mathrm{MSE}(\widehat{\varepsilon}|S_n) = &-\frac{2(\alpha^0 + n_0)(\alpha^1 + n_1)}{(\alpha^0 + \alpha^1 + n)^2(\alpha^0 + \alpha^1 + n + 1)}\widehat{\varepsilon}^0\widehat{\varepsilon}^1 \\ &- \frac{(\alpha^0 + n_0)^2}{(\alpha^0 + \alpha^1 + n)^2}(\widehat{\varepsilon}^0)^2 - \frac{(\alpha^1 + n_1)^2}{(\alpha^0 + \alpha^1 + n)^2}(\widehat{\varepsilon}^1)^2 \\ &+ \frac{(\alpha^0 + n_0)(\alpha^0 + n_0 + 1)}{(\alpha^0 + \alpha^1 + n)(\alpha^0 + \alpha^1 + n + 1)}\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^0(\theta_0)\right)^2\right] \\ &+ \frac{(\alpha^1 + n_1)(\alpha^1 + n_1 + 1)}{(\alpha^0 + \alpha^1 + n)(\alpha^0 + \alpha^1 + n + 1)}\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^1(\theta_1)\right)^2\right].\end{aligned}$$

Therefore, the conditional MSE for fixed samples is solved if we can find the first moment of the true error used in the definition of the Bayesian error estimator, $\widehat{\varepsilon}^y = \mathrm{E}_{\pi^*}\left[\varepsilon_n^y(\theta_y)\right]$, and the second moment, $\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^y(\theta_y)\right)^2\right]$, for both classes, $y \in \{0, 1\}$.

Having evaluated the conditional MSE of the Bayesian error estimator, it is easy to find analogous results for an arbitrary error estimate. Let $\widehat{\varepsilon}_\bullet$ be a constant number representing an error estimate evaluated from the given sample. Then,

$$\begin{aligned}\mathrm{MSE}(\widehat{\varepsilon}_{\bullet}|S_n) &= \mathrm{E}_{\theta}\left[(\varepsilon_n(\theta) - \widehat{\varepsilon}_{\bullet})^2|S_n\right] \\ &= \mathrm{E}_{\theta}\left[(\varepsilon_n(\theta) - \widehat{\varepsilon} + \widehat{\varepsilon} - \widehat{\varepsilon}_{\bullet})^2|S_n\right] \\ &= \mathrm{E}_{\theta}\left[(\varepsilon_n(\theta) - \widehat{\varepsilon})^2|S_n\right] + 2\left(\widehat{\varepsilon} - \widehat{\varepsilon}_{\bullet}\right)\mathrm{E}_{\theta}\left[\varepsilon_n(\theta) - \widehat{\varepsilon}|S_n\right] + (\widehat{\varepsilon} - \widehat{\varepsilon}_{\bullet})^2 \\ &= \mathrm{MSE}(\widehat{\varepsilon}|S_n) + (\widehat{\varepsilon} - \widehat{\varepsilon}_{\bullet})^2, \qquad (5.8)\end{aligned}$$

the last equality following from (2.5). Thus, if we solve the conditional MSE of the Bayesian error estimator, $\mathrm{MSE}(\widehat{\varepsilon}|S_n)$, it is trivial to evaluate the conditional MSE of any error estimator, $\mathrm{MSE}(\widehat{\varepsilon}_{\bullet}|S_n)$, under the Bayesian model. Further, (5.8) clearly shows that the conditional MSE of the Bayesian error estimator lower bounds the conditional MSE of any other error estimator.

B. The Discrete Model

We first solve the conditional MSE for the discrete classification problem defined in Chapter III with b bins and Dirichlet priors. It has been shown that the posteriors, $\pi^*(\theta_0)$ and $\pi^*(\theta_1)$, are Dirichlet distributions with updated hyperparameters $\alpha_i^0 + U_i$ and $\alpha_i^1 + V_i$ [83]. Furthermore, from (3.5) and (3.6),

$$\begin{aligned}\widehat{\varepsilon}^0 &= \sum_{j=1}^{b} \frac{U_j + \alpha_j^0}{n_0 + \sum_{i=1}^{b} \alpha_i^0} \mathbf{I}_{\psi_n(j)=1}, \\ \widehat{\varepsilon}^1 &= \sum_{j=1}^{b} \frac{V_j + \alpha_j^1}{n_1 + \sum_{i=1}^{b} \alpha_i^1} \mathbf{I}_{\psi_n(j)=0}.\end{aligned}$$

Following a similar method as that used to derive $\widehat{\varepsilon}^0$ and $\widehat{\varepsilon}^1$, we may also evaluate the second moments of the true errors contributed by each class. In particular, for class 0,

$$\begin{aligned}
\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^0(\theta_0)\right)^2\right] &= \int_{\boldsymbol{\Theta}_0} \left(\varepsilon_n^0(\theta_0)\right)^2 \pi^*(\theta_0) d\theta_0 \\
&= \int_0^1 \cdots \int_0^{a^0_{(2)}} \left(\sum_{j=1}^{b} \left(a^0_{(j)} - a^0_{(j-1)}\right) \mathbf{I}_{\psi_n(j)=1} \right)^2 \\
&\quad \times \left(\frac{\Gamma\left(n_0 + \sum_{i=1}^b \alpha_i^0\right)}{\prod_{i=1}^b \Gamma\left(U_i + \alpha_i^0\right)} \prod_{i=1}^{b} \left(a^0_{(i)} - a^0_{(i-1)}\right)^{U_i + \alpha_i^0 - 1} \right) da^0_{(1)} \ldots da^0_{(b-1)} \\
&= \frac{\Gamma\left(n_0 + \sum_{i=1}^b \alpha_i^0\right)}{\prod_{i=1}^b \Gamma\left(U_i + \alpha_i^0\right)} \sum_{j=1}^{b} \sum_{k=1}^{b} \mathbf{I}_{\psi_n(j)=1} \mathbf{I}_{\psi_n(k)=1} \\
&\quad \times \int_0^1 \cdots \int_0^{a^0_{(2)}} \prod_{i=1}^{b} \left(a^0_{(i)} - a^0_{(i-1)}\right)^{U_i + \alpha_i^0 - 1 + \delta_{i-j} + \delta_{i-k}} da^0_{(1)} \ldots da^0_{(b-1)}.
\end{aligned}$$

The integral in the last line has been solved in Lemma 1, thus,

$$\begin{aligned}
\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^0(\theta_0)\right)^2\right] &= \frac{\Gamma\left(n_0 + \sum_{i=1}^b \alpha_i^0\right)}{\prod_{i=1}^b \Gamma\left(U_i + \alpha_i^0\right)} \sum_{j=1}^{b} \sum_{k=1}^{b} \mathbf{I}_{\psi_n(j)=1} \mathbf{I}_{\psi_n(k)=1} \\
&\quad \times \frac{\prod_{i=1}^b \Gamma\left(U_i + \alpha_i^0 + \delta_{i-j} + \delta_{i-k}\right)}{\Gamma\left(b + \sum_{i=1}^b \left(U_i + \alpha_i^0 - 1 + \delta_{i-j} + \delta_{i-k}\right)\right)} \\
&= \frac{\Gamma\left(n_0 + \sum_{i=1}^b \alpha_i^0\right)}{\prod_{i=1}^b \Gamma\left(U_i + \alpha_i^0\right)} \sum_{j=1}^{b} \sum_{k=1}^{b} \mathbf{I}_{\psi_n(j)=1} \mathbf{I}_{\psi_n(k)=1} \\
&\quad \times \frac{\left(U_k + \alpha_k^0 + \delta_{k-j}\right)\left(U_j + \alpha_j^0\right) \prod_{i=1}^b \Gamma\left(U_i + \alpha_i^0\right)}{\left(1 + n_0 + \sum_{i=1}^{b} \alpha_i^0\right)\left(n_0 + \sum_{i=1}^{b} \alpha_i^0\right) \Gamma\left(n_0 + \sum_{i=1}^{b} \alpha_i^0\right)} \\
&= \sum_{j=1}^{b} \sum_{k=1}^{b} \mathbf{I}_{\psi_n(j)=1} \mathbf{I}_{\psi_n(k)=1} \frac{\left(U_k + \alpha_k^0 + \delta_{k-j}\right)\left(U_j + \alpha_j^0\right)}{\left(1 + n_0 + \sum_{i=1}^b \alpha_i^0\right)\left(n_0 + \sum_{i=1}^b \alpha_i^0\right)},
\end{aligned}$$

where the second equality follows from properties of the gamma function. Finally, we simplify this expression to obtain

$$
\begin{aligned}
\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^0(\theta_0)\right)^2\right] &= \frac{\displaystyle\sum_{j=1}^{b} \mathbf{I}_{\psi_n(j)=1}\left(U_j+\alpha_j^0\right)\sum_{k=1}^{b} \mathbf{I}_{\psi_n(k)=1}\left(U_k+\alpha_k^0+\delta_{k-j}\right)}{\left(1+n_0+\sum_{i=1}^{b}\alpha_i^0\right)\left(n_0+\sum_{i=1}^{b}\alpha_i^0\right)} \\
&= \frac{1+\sum_{j=1}^{b}\mathbf{I}_{\psi_n(j)=1}\left(U_j+\alpha_j^0\right)}{1+n_0+\sum_{i=1}^{b}\alpha_i^0} \times \frac{\sum_{j=1}^{b}\mathbf{I}_{\psi_n(j)=1}\left(U_j+\alpha_j^0\right)}{n_0+\sum_{i=1}^{b}\alpha_i^0}. \qquad (5.9)
\end{aligned}
$$

Similar results can be found for class 1:

$$
\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^1(\theta_1)\right)^2\right] = \frac{1+\sum_{j=1}^{b}\mathbf{I}_{\psi_n(j)=0}\left(V_j+\alpha_j^1\right)}{1+n_1+\sum_{i=1}^{b}\alpha_i^1} \times \frac{\sum_{j=1}^{b}\mathbf{I}_{\psi_n(j)=0}\left(V_j+\alpha_j^1\right)}{n_1+\sum_{i=1}^{b}\alpha_i^1}. \qquad (5.10)
$$

Combining equations (3.5), (3.6), (5.9) and (5.10) with (5.2) specifies the conditional MSE of the Bayesian error estimator in the discrete model.

C. The Gaussian Model with Linear Classification

We next consider the Gaussian models defined in Section IV. If the designed classifier is constant, that is, if $\mathbf{a} = 0$, then the true error, ε_n^y, is deterministically zero or one, depending on the sign of b. In this special case, the conditional MSE is found trivially:

$$
\begin{aligned}
\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^0\left(\theta_0\right)\right)^2\right] &= \widehat{\varepsilon}^0 = \varepsilon_n^0 = \mathbf{I}_{b>0}, \\
\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^1\left(\theta_1\right)\right)^2\right] &= \widehat{\varepsilon}^1 = \varepsilon_n^1 = \mathbf{I}_{b\leq 0},
\end{aligned}
$$

so that from (5.1) we have

$$
\mathrm{MSE}(\widehat{\varepsilon}|S_n) = \mathrm{Var}_{\pi^*}\left(c\right),
$$

which is the posterior variance of the *a priori* class probability. In the remainder of this section we assume $\mathbf{a} \neq 0$.

We will present closed-form expressions for the conditional MSE of Bayesian error estimators under Gaussian distributions with linear classification for all three covariance models. The second moments we require in (5.2) may be written as,

$$
\begin{aligned}
\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^y\left(\theta_y\right)\right)^2\right] &= \int_{\mathbf{\Theta}_y} \left(\varepsilon_n^y(\theta_y)\right)^2 \pi^*(\theta_y) d\theta_y \\
&= \int_{\mathbf{\Lambda}_y} \int_{\mathbb{R}^D} \left(\varepsilon_n^y(\mu_y, \Lambda_y)\right)^2 \pi^*(\mu_y|\Lambda_y) d\mu_y \pi^*(\Lambda_y) d\Lambda_y. \qquad (5.11)
\end{aligned}
$$

1. Solution for Fixed Covariance

For a fixed (invertible) covariance, Σ_y, we require $\nu^* > 0$ to ensure that the posterior, $\pi^*(\mu_y|\Lambda_y)$, is proper. From (4.8) we have

$$\widehat{\varepsilon}^y = \Phi\left(d\right)$$

where

$$d = \frac{(-1)^y g\left(\mathbf{m}^*\right)}{\sqrt{\mathbf{a}^T \Sigma_y \mathbf{a}}} \sqrt{\frac{\nu^*}{\nu^* + 1}}. \qquad (5.12)$$

To find the conditional MSE, the outer integral in the definition of the second moment (5.11) is not needed in the fixed covariance model. We need only solve the inner integral, which is given by,

$$
\begin{aligned}
\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^y\left(\theta_y\right)\right)^2\right] &= \int_{\mathbb{R}^D} \left(\varepsilon_n^y(\mu_y, \Lambda_y)\right)^2 \pi^*(\mu_y|\Lambda_y) d\mu_y \\
&= \int_{\mathbb{R}^D} \left(\Phi\left(\frac{(-1)^y g(\mu_y)}{\sqrt{\mathbf{a}^T \Sigma_y \mathbf{a}}}\right)\right)^2 f_{\mathbf{m}^*, \Sigma_y/\nu^*}(\mu_y) d\mu_y.
\end{aligned}
$$

This integral is simplified to a well-behaved single integral in Lemma 6.

Lemma 6. *Let $y \in \{0, 1\}$ be a class label and let $\nu^* > 0$. Also let $\mathbf{m}^* \in \mathbb{R}^D$ be a mean vector with $D \geq 1$ features, Σ be an invertible covariance matrix, and $g(\mathbf{x}) = \mathbf{a}^T \mathbf{x} + b$,* *where* $\mathbf{a} \in \mathbb{R}^D$ *is a non-zero length* D *vector and* $b \in \mathbb{R}$ *is a scalar. Then,*

$$\int_{\mathbb{R}^D} \left(\Phi \left(\frac{(-1)^y g(\mu)}{\sqrt{\mathbf{a}^T \Sigma \mathbf{a}}} \right) \right)^2 f_{\mathbf{m}^*, \Sigma/\nu^*}(\mu) d\mu$$
$$= \mathbf{I}_{\{d>0\}} \left(2\Phi\left(d\right) - 1 \right) + \frac{1}{\pi} \int_0^{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}} \right)} \exp\left(-\frac{d^2}{2\sin^2\theta} \right) d\theta,$$

where $f_{\mu,\Sigma}$ *is a Gaussian density with mean* μ *and covariance* Σ*,* $\mathbf{I}_{\{d>0\}}$ *is an indicator function equal to one if* $d > 0$ *and zero otherwise, and*

$$d = \frac{(-1)^y g\left(\boldsymbol{m}^*\right)}{\sqrt{\boldsymbol{a}^T \Sigma \boldsymbol{a}}} \sqrt{\frac{\nu^*}{\nu^*+1}}.$$

Proof. Call this integral M. We have that,

$$M = \int_{\mathbb{R}^D} \left(\Phi \left(\frac{(-1)^y g(\mu)}{\sqrt{\mathbf{a}^T \Sigma \mathbf{a}}} \right) \right)^2 \frac{\nu^{*\frac{D}{2}}}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{\nu^*}{2} (\mu - \mathbf{m}^*)^T \Sigma^{-1} (\mu - \mathbf{m}^*) \right) d\mu.$$

Since Σ is an invertible covariance matrix, we can use singular value decomposition to write $\Sigma = WW^T$ with $|\Sigma| = |W|^2$. Next consider the linear change of variables, $\mathbf{z} = \sqrt{\nu^*} W^{-1} (\mu - \mathbf{m}^*)$. We have that,

$$M = \int_{\mathbb{R}^D} \left(\Phi \left(\frac{(-1)^y \left(\frac{1}{\sqrt{\nu^*}} \mathbf{a}^T W \mathbf{z} + \mathbf{a}^T \mathbf{m}^* + b \right)}{\sqrt{\mathbf{a}^T \Sigma \mathbf{a}}} \right) \right)^2 \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{\mathbf{z}^T \mathbf{z}}{2} \right) d\mathbf{z}.$$

Define $\bar{\mathbf{a}} = \frac{(-1)^y W^T \mathbf{a}}{\sqrt{\nu^*} \sqrt{\mathbf{a}^T \Sigma \mathbf{a}}}$ and $\bar{b} = \frac{(-1)^y g(\mathbf{m}^*)}{\sqrt{\mathbf{a}^T \Sigma \mathbf{a}}}$, and note that $\|\bar{\mathbf{a}}\|^2 = \frac{1}{\nu^*}$. Then,

$$\begin{aligned}
M &= \int_{\mathbb{R}^D} \left(\Phi\left(\bar{\mathbf{a}}^T \mathbf{z} + \bar{b} \right) \right)^2 \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{\mathbf{z}^T \mathbf{z}}{2} \right) d\mathbf{z} \\
&= \int_{\mathbb{R}^D} \int_{-\infty}^{\bar{\mathbf{a}}^T \mathbf{z} + \bar{b}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right) dx \\
&\quad \times \int_{-\infty}^{\bar{\mathbf{a}}^T \mathbf{z} + \bar{b}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2} \right) dy \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{\mathbf{z}^T \mathbf{z}}{2} \right) d\mathbf{z} \\
&= \int_{\mathbb{R}^D} \int_{-\infty}^{\bar{\mathbf{a}}^T \mathbf{z} + \bar{b}} \int_{-\infty}^{\bar{\mathbf{a}}^T \mathbf{z} + \bar{b}} \frac{1}{(2\pi)^{\frac{D+2}{2}}} \exp\left(-\frac{x^2 + y^2 + \mathbf{z}^T \mathbf{z}}{2} \right) dx dy d\mathbf{z}.
\end{aligned}$$

Next consider a change of variables $\mathbf{w} = R\mathbf{z}$, where R rotates the vector $\bar{\mathbf{a}}$ to the vector $\left(\frac{1}{\sqrt{\nu^*}}, 0, \ldots, 0\right)$. Since R is a rotation matrix, $\det(R) = 1$ and $R^T R$ is an identity matrix. Let the first element in the vector $\mathbf{w}$ be called w. Then this integral simplifies to:

$$\begin{aligned} M &= \int_{\mathbb{R}^D} \int_{-\infty}^{\bar{\mathbf{a}}^T R^T \mathbf{w} + \bar{b}} \int_{-\infty}^{\bar{\mathbf{a}}^T R^T \mathbf{w} + \bar{b}} \frac{1}{(2\pi)^{\frac{D+2}{2}}} \exp\left(-\frac{x^2 + y^2 + \mathbf{w}^T \mathbf{w}}{2}\right) dx dy d\mathbf{w} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{1}{\sqrt{\nu^*}} w + \bar{b}} \int_{-\infty}^{\frac{1}{\sqrt{\nu^*}} w + \bar{b}} \frac{1}{(2\pi)^{\frac{3}{2}}} \exp\left(-\frac{x^2 + y^2 + w^2}{2}\right) dx dy dw. \end{aligned}$$

This reduces the problem to a three dimensional space.

Now consider the following rotation of the coordinate system:

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = \begin{bmatrix} -\frac{\sqrt{2\nu^*}}{2\sqrt{\nu^*+2}} & -\frac{\sqrt{2\nu^*}}{2\sqrt{\nu^*+2}} & \frac{\sqrt{2}}{\sqrt{\nu^*+2}} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ \frac{1}{\sqrt{\nu^*+2}} & \frac{1}{\sqrt{\nu^*+2}} & \frac{\sqrt{\nu^*}}{\sqrt{\nu^*+2}} \end{bmatrix} \begin{bmatrix} x \\ y \\ w \end{bmatrix}.$$

This rotates the vector $(x, y, w) = \left(1, 1, \sqrt{\nu^*}\right)$ to the vector $(x', y', w') = \left(0, 0, \sqrt{\nu^* + 2}\right)$. To determine the new region of integration, note in the (x, y, w) coordinate system the region of integration is defined by two restrictions: $x < \frac{1}{\sqrt{\nu^*}} w + \bar{b}$ and $y < \frac{1}{\sqrt{\nu^*}} w + \bar{b}$. In the new coordinate system, the first restriction is

$$-\frac{\sqrt{2\nu^*}}{2\sqrt{\nu^* + 2}} x' + \frac{\sqrt{2}}{2} y' + \frac{1}{\sqrt{\nu^* + 2}} w' < \frac{1}{\sqrt{\nu^*}} \left(\frac{\sqrt{2}}{\sqrt{\nu^* + 2}} x' + \frac{\sqrt{\nu^*}}{\sqrt{\nu^* + 2}} w'\right) + \bar{b}.$$

Equivalently,

$$y' < \left(\frac{\sqrt{\nu^* + 2}}{\sqrt{\nu^*}}\right) x' + \sqrt{2}\, \bar{b}.$$

And similarly for the other restriction,

$$-y' < \left(\frac{\sqrt{\nu^* + 2}}{\sqrt{\nu^*}}\right) x' + \sqrt{2}\, \bar{b}.$$

We have designed our new coordinate system to make the variable w' independent from these restrictions. Hence, it may be integrated out of our original integral, which may be simplified to

$$M = \int_{-\infty}^{\infty} \int_{\frac{\sqrt{\nu^*}}{\sqrt{\nu^*+2}}\left(|y'|-\sqrt{2}\,\bar{b}\right)}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{x'^2+y'^2}{2}\right) dx'dy'$$

$$= 2\int_{0}^{\infty} \int_{\frac{\sqrt{\nu^*}}{\sqrt{\nu^*+2}}\left(y'-\sqrt{2}\,\bar{b}\right)}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{x'^2+y'^2}{2}\right) dx'dy'. \qquad (5.13)$$

If $\bar{b} \leq 0$, then we convert to polar coordinates, (r, θ), using

$$x' = r\cos\left(\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right) - \theta\right),$$

$$y' = r\sin\left(\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right) - \theta\right),$$

to obtain

$$M = \frac{1}{\pi}\int_{0}^{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)} \int_{-\frac{\sqrt{\nu^*}\,\bar{b}}{\sqrt{\nu^*+1}\sin\theta}}^{\infty} \exp\left(-\frac{r^2}{2}\right) r dr d\theta.$$

Let $u = \frac{r^2}{2}$. Then finally,

$$M = \frac{1}{\pi}\int_{0}^{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)} \int_{\frac{\nu^*\,\bar{b}^2}{2(\nu^*+1)\sin^2\theta}}^{\infty} \exp\left(-u\right) du d\theta$$

$$= \frac{1}{\pi}\int_{0}^{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)} \exp\left(-\frac{\nu^*\,\bar{b}^2}{2\left(\nu^*+1\right)\sin^2\theta}\right) d\theta.$$

On the other hand, if $\bar{b} > 0$, then from (5.13),

$$M = \frac{1}{\pi}\int_{0}^{\infty} \int_{\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}y'}^{\infty} \exp\left(-\frac{x'^2+y'^2}{2}\right) dx'dy'$$

$$+ \frac{1}{\pi}\int_{0}^{\infty} \int_{\frac{\sqrt{\nu^*}}{\sqrt{\nu^*+2}}\left(y'-\sqrt{2}\,\bar{b}\right)}^{\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}y'} \exp\left(-\frac{x'^2+y'^2}{2}\right) dx'dy'.$$

The first integral is easily solved using the result for $\bar{b} \leq 0$, and for the second integral we use the same polar transformation and u-substitution as before,

$$\begin{aligned}M &= \frac{1}{\pi}\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)+\frac{1}{\pi}\int_{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)-\pi}^{0}\int_{0}^{-\frac{\sqrt{\nu^*}\,\bar{b}}{\sqrt{\nu^*+1}\sin\theta}}\exp\left(-\frac{r^2}{2}\right)rdrd\theta\\
&= \frac{1}{\pi}\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)+\frac{1}{\pi}\int_{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)-\pi}^{0}\int_{0}^{\frac{\nu^*\,\bar{b}^2}{2(\nu^*+1)\sin^2\theta}}\exp\left(-u\right)dud\theta\\
&= \frac{1}{\pi}\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)+\frac{1}{\pi}\int_{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)-\pi}^{0}\left(1-\exp\left(-\frac{\nu^*\,\bar{b}^2}{2\left(\nu^*+1\right)\sin^2\theta}\right)\right)d\theta\\
&= 1-\frac{1}{\pi}\int_{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)-\pi}^{0}\exp\left(-\frac{\nu^*\,\bar{b}^2}{2\left(\nu^*+1\right)\sin^2\theta}\right)d\theta.\end{aligned}$$

This may be simplified by realizing that a component of this integral is equivalent to an alternate representation for the Gaussian CDF function [100]. We first break the integral into two parts, and then use symmetry in the integrand to simplify the result.

$$\begin{aligned}M &= 1-\frac{1}{\pi}\int_{-\pi}^{0}\exp\left(-\frac{\nu^*\,\bar{b}^2}{2\left(\nu^*+1\right)\sin^2\theta}\right)d\theta\\
&\quad+\frac{1}{\pi}\int_{-\pi}^{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)-\pi}\exp\left(-\frac{\nu^*\,\bar{b}^2}{2\left(\nu^*+1\right)\sin^2\theta}\right)d\theta\\
&= 1-\frac{1}{\pi}\int_{0}^{\pi}\exp\left(-\frac{\nu^*\,\bar{b}^2}{2\left(\nu^*+1\right)\sin^2\theta}\right)d\theta\\
&\quad+\frac{1}{\pi}\int_{0}^{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)}\exp\left(-\frac{\nu^*\,\bar{b}^2}{2\left(\nu^*+1\right)\sin^2\theta}\right)d\theta\\
&= 2\Phi\left(\frac{\sqrt{\nu^*}\,\bar{b}}{\sqrt{\nu^*+1}}\right)-1+\frac{1}{\pi}\int_{0}^{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)}\exp\left(-\frac{\nu^*\,\bar{b}^2}{2\left(\nu^*+1\right)\sin^2\theta}\right)d\theta. \qquad \square\end{aligned}$$

Thus, we have

$$\mathrm{E}_{\pi^*}\left[\left(\varepsilon_n^y\left(\theta_y\right)\right)^2\right]=\mathbf{I}_{\{d>0\}}\left(2\Phi\left(d\right)-1\right)+\frac{1}{\pi}\int_{0}^{\arctan\left(\frac{\sqrt{\nu^*+2}}{\sqrt{\nu^*}}\right)}\exp\left(-\frac{d^2}{2\sin^2\theta}\right)d\theta, \quad (5.14)$$

where d is defined in (5.12). Combining (4.8) and (5.14) with (5.2) defines the sample-conditioned MSE of the Bayesian error estimator under the fixed covariance model.

2. Solution for Scaled Identity Covariance

In this model, we assume Σ_y is a scaled identity covariance matrix, that is, $\Lambda_y = \sigma^2$ and $\Sigma_y = \sigma^2 I_D$. Under this model, it has been shown that $\pi^*(\sigma^2)$ has an inverse-gamma distribution with parameters

$$\begin{aligned} \alpha &= \frac{(\kappa^* + D + 1)\, D}{2} - 1, \\ \beta &= \frac{1}{2}\text{trace}\,(S^*)\,. \end{aligned}$$

Hence, we require $\nu^* > 0$ to ensure that $\pi(\mu_y|\Lambda_y)$ is proper and, additionally, we require $\alpha > 0$ and $\beta > 0$ to ensure that $\pi^*(\sigma^2)$ is proper or, equivalently, $(\kappa^* + D + 1)\, D > 2$, and S^* must be positive definite.

The Bayesian error estimator is given in (4.10):

$$\widehat{\varepsilon}^y = \frac{1}{2}\left(1 + \text{sgn}(A) I\left(\frac{A^2}{A^2 + \text{trace}\,(S^*)}; \frac{1}{2}, \frac{(\kappa^* + D + 1)\, D}{2} - 1\right)\right),$$

where

$$A = \frac{(-1)^y g(\mathbf{m}^*)}{\|\mathbf{a}\|}\sqrt{\frac{\nu^*}{\nu^* + 1}}.$$

To evaluate the second moment of the true error for scaled identity covariances, we use the previous result from Lemma 6 for the inner integral, so that (5.11) is precisely the integral solved in Lemma 7.

Lemma 7. *Let $A \in \mathbb{R}$, $\pi/4 < B < \pi/2$, $\alpha > 0$, and $\beta > 0$. Let $f_G(x; \alpha, \beta)$ be an inverse-gamma distribution with shape parameter α and scale parameter β, and*

$\mathbf{I}_{\{A>0\}}$ *be an indicator function equal to one if* $A>0$ *and zero otherwise. Then,*

$$\int_0^\infty \left(\mathbf{I}_{\{A>0\}}\left(2\Phi\left(\frac{A}{\sqrt{z}}\right)-1\right)+\frac{1}{\pi}\int_0^B \exp\left(-\frac{A^2}{2z\sin^2\theta}\right)d\theta\right)f_G(z;\alpha,\beta)dz$$
$$=\mathbf{I}_{\{A>0\}}I\left(\frac{A^2}{A^2+2\beta};\frac{1}{2},\alpha\right)+R\left(\sin^2 B,\frac{A^2}{2\beta};\alpha\right),$$

where $I(x;a,b)$ *is the regularized incomplete beta function, defined for* $0\le x\le 1$, $a>0$ *and* $b>0$, *and* R *is given by an Appell hypergeometric function,* F_1, *such that* $R\left(x,0;a\right)=\frac{1}{\pi}\arcsin\left(\sqrt{x}\right)$ *and*

$$R\left(x,y;a\right)=\frac{\sqrt{y}}{\pi(2a+1)}\left(\frac{x}{x+y}\right)^{a+\frac{1}{2}}\mathrm{F}_1\left(a+\frac{1}{2};\frac{1}{2},1;a+\frac{3}{2};\frac{x\left(y+1\right)}{x+y},\frac{x}{x+y}\right) \tag{5.15}$$

for $a>0$, $0<x<1$ *and* $y>0$.

Proof. Call this integral M. When $A=0$, it is easy to show that $M=B/\pi$. For $A\neq 0$, we obtain,

$$\begin{aligned}M&=\int_0^\infty \mathbf{I}_{\{A>0\}}\left(2\Phi\left(\frac{A}{\sqrt{z}}\right)-1\right)f_G(z;\alpha,\beta)dz\\&\quad+\frac{1}{\pi}\int_0^\infty\int_0^B \exp\left(-\frac{A^2}{2z\sin^2\theta}\right)d\theta f_G(z;\alpha,\beta)dz\\&=\mathbf{I}_{\{A>0\}}\left(2\int_0^\infty \Phi\left(\frac{A}{\sqrt{z}}\right)f_G(z;\alpha,\beta)dz-1\right)\\&\quad+\frac{1}{\pi}\int_0^\infty\int_0^B \exp\left(-\frac{A^2}{2z\sin^2\theta}\right)d\theta f_G(z;\alpha,\beta)dz.\end{aligned}$$

The integral in the first term has already been solved in Lemma 3. We have,

$$\begin{aligned}M&=\mathbf{I}_{\{A>0\}}\operatorname{sgn}\left(A\right)I\left(\frac{A^2}{A^2+2\beta};\frac{1}{2},\alpha\right)+\frac{1}{\pi}\int_0^\infty\int_0^B \exp\left(-\frac{A^2}{2z\sin^2\theta}\right)d\theta f_G(z;\alpha,\beta)dz\\&=\mathbf{I}_{\{A>0\}}I\left(\frac{A^2}{A^2+2\beta};\frac{1}{2},\alpha\right)+\frac{1}{\pi}\int_0^B\int_0^\infty \exp\left(-\frac{A^2}{2z\sin^2\theta}\right)f_G(z;\alpha,\beta)dzd\theta.\end{aligned} \tag{5.16}$$

This intermediate result will be used in Lemma 8.

We next focus on the inner integral in the second term. Call this integral N. We have,

$$\begin{aligned}
N &= \int_0^\infty \exp\left(-\frac{A^2}{2z\sin^2\theta}\right)\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{1}{z^{\alpha+1}}\exp\left(-\frac{\beta}{z}\right)dz \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \frac{1}{z^{\alpha+1}}\exp\left(-\left(\beta+\frac{A^2}{2\sin^2\theta}\right)\frac{1}{z}\right)dz \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha)}{\left(\beta+\frac{A^2}{2\sin^2\theta}\right)^\alpha} \\
&= \left(\frac{\sin^2\theta}{\sin^2\theta+\frac{A^2}{2\beta}}\right)^\alpha,
\end{aligned}$$

where we have solved this integral by noting it is essentially an inverse-gamma distribution. Thus our original integral is,

$$M = \mathbf{I}_{\{A>0\}} I\left(\frac{A^2}{A^2+2\beta};\frac{1}{2},\alpha\right) + \frac{1}{\pi}\int_0^B \left(\frac{\sin^2\theta}{\sin^2\theta+\frac{A^2}{2\beta}}\right)^\alpha d\theta. \tag{5.17}$$

For the final integral, consider the substitution $u = \frac{\sin^2\theta}{\sin^2 B}$. We have,

$$\begin{aligned}
&\int_0^B \left(\frac{\sin^2\theta}{\sin^2\theta+\frac{A^2}{2\beta}}\right)^\alpha d\theta \\
&= \int_0^1 \left(\frac{u\sin^2 B}{u\sin^2 B+\frac{A^2}{2\beta}}\right)^\alpha \frac{\sin B}{2}u^{-1/2}(1-u\sin^2 B)^{-1/2}du \\
&= \frac{\sin^{2\alpha+1}B}{2}\left(\frac{2\beta}{A^2}\right)^\alpha \int_0^1 u^{\alpha-1/2}(1-u\sin^2 B)^{-1/2}\left(1+u\frac{2\beta\sin^2 B}{A^2}\right)^{-\alpha}du.
\end{aligned}$$

This is essentially a one-dimensional Euler-type integral representation of Appell's hypergeometric function, F_1. In other words,

$$\begin{aligned}
&\int_0^B \left(\frac{\sin^2\theta}{\sin^2\theta+\frac{A^2}{2\beta}}\right)^\alpha d\theta \\
&= \frac{\sin^{2\alpha+1}B}{2\alpha+1}\left(\frac{2\beta}{A^2}\right)^\alpha \mathrm{F}_1\left(\alpha+\frac{1}{2};\frac{1}{2},\alpha;\alpha+\frac{3}{2};\sin^2 B,-\frac{2\beta\sin^2 B}{A^2}\right).
\end{aligned}$$

Finally, from the identity [101]

$$\mathrm{F}_1\left(a;b,b';c;z,z'\right)=(1-z')^{-a}\mathrm{F}_1\left(a;b,c-b-b';c;\frac{z-z'}{1-z'},-\frac{z'}{1-z'}\right)$$

we have

$$\int_0^B\left(\frac{\sin^2\theta}{\sin^2\theta+\frac{A^2}{2\beta}}\right)^{\alpha}d\theta$$
$$=\frac{\sqrt{\frac{A^2}{2\beta}}}{2\alpha+1}\left(\frac{\sin^2 B}{\sin^2 B+\frac{A^2}{2\beta}}\right)^{\alpha+\frac{1}{2}}\mathrm{F}_1\left(\alpha+\frac{1}{2};\frac{1}{2},1;\alpha+\frac{3}{2};\frac{\left(\frac{A^2}{2\beta}+1\right)\sin^2 B}{\sin^2 B+\frac{A^2}{2\beta}},\frac{\sin^2 B}{\sin^2 B+\frac{A^2}{2\beta}}\right)$$
$$=\pi R\left(\sin^2 B,\frac{A^2}{2\beta};\alpha\right)$$

Combining this result with (5.17) completes the proof. □

Thus, the sample-conditioned MSE for scaled identity covariances is,

$$\begin{aligned}\mathrm{E}_{\pi^*}\left[(\varepsilon_n^y(\theta_y))^2\right]=\mathbf{I}_{\{A>0\}}I\left(\frac{A^2}{A^2+\mathrm{trace}\,(S^*)};\frac{1}{2},\frac{(\kappa^*+D+1)\,D}{2}-1\right)\\+R\left(\frac{\nu^*+2}{2(\nu^*+1)},\frac{A^2}{\mathrm{trace}\,(S^*)};\frac{(\kappa^*+D+1)\,D}{2}-1\right),\end{aligned}\tag{5.18}$$

where R is defined in Lemma 7. Combining (4.10) and (5.18) with (5.2) defines the conditional MSE of the Bayesian error estimator under the scaled identity covariance model. Closed-form expressions for both I and R for integer or half-integer values of κ are discussed in Sections IV.A.7 and V.C.4, respectively.

3. Solution for General Covariance

Finally, in the general covariance model we assume $\Sigma_y=\Lambda_y$, that is, Σ_y is an arbitrary covariance matrix, and that the parameter space $\mathbf{\Lambda}_y$ contains all positive definite matrices. In this case, $\pi^*\left(\Sigma_y\right)$ is an inverse-Wishart distribution:

$$\pi^*(\Sigma_y) = \frac{|S^*|^{\kappa^*/2}}{2^{\kappa^* D/2}\Gamma_D(\kappa^*/2)}|\Sigma_y|^{-(\kappa^*+D+1)/2} \exp\left(-\frac{1}{2}\text{trace}\left(S^*\Sigma_y^{-1}\right)\right).$$

For a proper posterior, we require $\nu^* > 0$, $\kappa^* > D-1$ and S^* positive definite. It has been shown in (4.11) that

$$\widehat{\varepsilon}^y = \frac{1}{2}\left(1 + \operatorname{sgn}(A) I\left(\frac{A^2}{A^2 + \mathbf{a}^T S^* \mathbf{a}}; \frac{1}{2}, \frac{\kappa^* - D + 1}{2}\right)\right),$$

where

$$A = (-1)^y g(\mathbf{m}^*)\sqrt{\frac{\nu^*}{\nu^*+1}}.$$

Using the same method from the previous section, we evaluate the second moment of the true error for arbitrary covariances using the previous result from Lemma 6 for the inner integral in (5.11). This is solved in Lemma 8 below.

Lemma 8. *Let $A \in \mathbb{R}$, $\pi/4 < B < \pi/2$, $\mathbf{a} \in \mathbb{R}^D$ be a non-zero column vector, $\kappa^* > D - 1$, and S^* be a positive definite $D \times D$ matrix. Also let $f_W(\Sigma; S^*, \kappa^*)$ be an inverse-Wishart distribution with parameters S^* and κ^* and $\mathbf{I}_{\{A>0\}}$ be an indicator function equal to one if $A > 0$ and zero otherwise. Then*

$$\begin{aligned}\int_{\Sigma>0} &\left(\mathbf{I}_{\{A>0\}}\left(2\Phi\left(\frac{A}{\sqrt{\boldsymbol{a}^T\Sigma\boldsymbol{a}}}\right) - 1\right)\right. \\ &\qquad\left. + \frac{1}{\pi}\int_0^B \exp\left(-\frac{A^2}{\left(2\sin^2\theta\right)\boldsymbol{a}^T\Sigma\boldsymbol{a}}\right)d\theta\right) f_W(\Sigma; S^*, \kappa^*)d\Sigma \\ &= \mathbf{I}_{\{A>0\}} I\left(\frac{A^2}{A^2 + \boldsymbol{a}^T S^* \boldsymbol{a}}; \frac{1}{2}, \frac{\kappa^* - D + 1}{2}\right) + R\left(\sin^2 B, \frac{A^2}{\boldsymbol{a}^T S^* \boldsymbol{a}}; \frac{\kappa^* - D + 1}{2}\right),\end{aligned}$$

where the outer integration is over all positive definite matrices, $I(x; a, b)$ is the regularized incomplete beta function, and $R(x, y; a)$ is defined in the statement of Lemma 7.

Proof. Call this integral M. If $A = 0$, it is easy to show that $M = B/\pi$. Otherwise, if $A \neq 0$ then we have,

$$\begin{aligned}
M &= \int_{\Sigma>0} \mathbf{I}_{\{A>0\}} \left(2\Phi\left(\frac{A}{\sqrt{\mathbf{a}^T\Sigma\mathbf{a}}}\right) - 1\right) f_W(\Sigma; S^*, \kappa^*) d\Sigma \\
&\quad + \frac{1}{\pi}\int_{\Sigma>0}\int_0^B \exp\left(-\frac{A^2}{\left(2\sin^2\theta\right)\mathbf{a}^T\Sigma\mathbf{a}}\right) d\theta f_W(\Sigma; S^*, \kappa^*) d\Sigma \\
&= \mathbf{I}_{\{A>0\}} \left(2\int_{\Sigma>0} \Phi\left(\frac{A}{\sqrt{\mathbf{a}^T\Sigma\mathbf{a}}}\right) f_W(\Sigma; S^*, \kappa^*) d\Sigma - 1\right) \\
&\quad + \frac{1}{\pi}\int_{\Sigma>0}\int_0^B \exp\left(-\frac{A^2}{\left(2\sin^2\theta\right)\mathbf{a}^T\Sigma\mathbf{a}}\right) d\theta f_W(\Sigma; S^*, \kappa^*) d\Sigma.
\end{aligned}$$

The integral in the first term has been solved in Lemma 4. We have that

$$\begin{aligned}
M &= \mathbf{I}_{\{A>0\}} \operatorname{sgn}(A)\, I\left(\frac{A^2}{A^2 + \mathbf{a}^T S^* \mathbf{a}}; \frac{1}{2}, \frac{\kappa^* - D + 1}{2}\right) \\
&\quad + \frac{1}{\pi}\int_{\Sigma>0}\int_0^B \exp\left(-\frac{A^2}{\left(2\sin^2\theta\right)\mathbf{a}^T\Sigma\mathbf{a}}\right) d\theta f_W(\Sigma; S^*, \kappa^*) d\Sigma \\
&= \mathbf{I}_{\{A>0\}} I\left(\frac{A^2}{A^2 + \mathbf{a}^T S^* \mathbf{a}}; \frac{1}{2}, \frac{\kappa^* - D + 1}{2}\right) \\
&\quad + \frac{1}{\pi}\int_0^B\int_{\Sigma>0} \exp\left(-\frac{A^2}{\left(2\sin^2\theta\right)\mathbf{a}^T\Sigma\mathbf{a}}\right) f_W(\Sigma; S^*, \kappa^*) d\Sigma d\theta.
\end{aligned}$$

Define the following constant matrix:

$$C = \left[\begin{array}{c|c} \multicolumn{2}{c}{\mathbf{a}^T} \\ \hline 0_{D-1\times 1} & I_{D-1}. \end{array}\right].$$

Since $\mathbf{a}$ is non-zero, with a simple reordering of the dimensions we can guarantee $a_1 \neq 0$. The value of $\mathbf{a}^T S^* \mathbf{a}$ is unchanged by such a redefinition, so without loss of generality assume C is invertible. Consider the change of variables, $Y = C\Sigma C^T$. Since C is invertible, Y is positive definite if and only if Σ is also. Furthermore, the Jacobean determinant of this transformation is $|C|^{D+1}$ [95, 92]. Note $\mathbf{a}^T\Sigma\mathbf{a} = y_{11}$, where the subscript 11 indexes the upper left element of a matrix, and we have:

$$\begin{aligned}M &= \mathbf{I}_{\{A>0\}} I\left(\frac{A^2}{A^2+\mathbf{a}^T S^* \mathbf{a}}; \frac{1}{2}, \frac{\kappa^* - D + 1}{2}\right) \\ &\quad + \frac{1}{\pi}\int_0^B \int_{Y>0} \exp\left(-\frac{A^2}{2y_{11}\sin^2\theta}\right) f_W(C^{-1}Y(C^T)^{-1}; S^*, \kappa^*)\frac{1}{|C|^{D+1}} dY d\theta \\ &= \mathbf{I}_{\{A>0\}} I\left(\frac{A^2}{A^2+\mathbf{a}^T S^* \mathbf{a}}; \frac{1}{2}, \frac{\kappa^* - D + 1}{2}\right) \\ &\quad + \frac{1}{\pi}\int_0^B \int_{Y>0} \exp\left(-\frac{A^2}{2y_{11}\sin^2\theta}\right) f_W(Y; CS^*C^T, \kappa^*) dY d\theta.\end{aligned}$$

Since the integrand now depends on only one parameter in Y, namely y_{11}, the other parameters can be integrated out. It can be shown that for any inverse-Wishart random variable, X, with density $f_W(X; A, m)$, the marginal distribution of x_{11} is also an inverse-Wishart distribution with density $f_W(x_{11}; a_{11}, m - D + 1)$ [96]. In one dimension, this is equivalent to the inverse-gamma distribution $f_G(x_{11}; (m-D+1)/2, a_{11}/2)$. In this case, $(CS^*C^T)_{11} = \mathbf{a}^T S^* \mathbf{a}$, so

$$\begin{aligned}M &= \mathbf{I}_{\{A>0\}} I\left(\frac{A^2}{A^2+\mathbf{a}^T S^* \mathbf{a}}; \frac{1}{2}, \frac{\kappa^* - D + 1}{2}\right) \\ &\quad + \frac{1}{\pi}\int_0^B \int_0^\infty \exp\left(-\frac{A^2}{2y_{11}\sin^2\theta}\right) f_G\left(y_{11}; \frac{\kappa^* - D + 1}{2}, \frac{\mathbf{a}^T S^* \mathbf{a}}{2}\right) dy_{11} d\theta \\ &= \mathbf{I}_{\{A>0\}} I\left(\frac{A^2}{A^2+2\beta}; \frac{1}{2}, \alpha\right) + \frac{1}{\pi}\int_0^B \int_0^\infty \exp\left(-\frac{A^2}{2y_{11}\sin^2\theta}\right) f_G\left(y_{11}; \alpha, \beta\right) dy_{11} d\theta,\end{aligned}$$

where we have defined

$$\begin{aligned}\alpha &= \frac{\kappa^* - D + 1}{2}, \\ \beta &= \frac{\mathbf{a}^T S^* \mathbf{a}}{2}.\end{aligned}$$

Note $\alpha > 0$, $\beta > 0$, and this integral is exactly the same as (5.16) so we apply Lemma 7 to complete the proof. □

Thus, the sample-conditioned MSE for arbitrary covariances is

$$\begin{aligned}\mathrm{E}_{\pi^*}\left[(\varepsilon_n^y(\theta_y))^2\right] = \mathbf{I}_{\{A>0\}} I\left(\frac{A^2}{A^2+\mathbf{a}^T S^* \mathbf{a}}; \frac{1}{2}, \frac{\kappa^* - D + 1}{2}\right) \\ + R\left(\frac{\nu^*+2}{2(\nu^*+1)}, \frac{A^2}{\mathbf{a}^T S^* \mathbf{a}}; \frac{\kappa^* - D + 1}{2}\right).\end{aligned} \tag{5.19}$$

Combining (4.11) and (5.19) with (5.2) defines the conditional MSE of the Bayesian error estimator under the general covariance model. Again note that closed form expressions for both I and R for integer or half-integer values of κ are discussed in Sections IV.A.7 and V.C.4, respectively.

4. Closed Form Representation for the R Function

The solutions proposed in the previous sections utilize two Euler integrals. The first is the regularized incomplete beta function, which is discussed in Section IV.A.7. A closed form solution for $I\left(x; \frac{1}{2}, \frac{N}{2}\right)$ was found for $0 \leq x \leq 1$ and positive integers N in (4.12).

The second integral is the function $R(x, y; a)$, defined for $a > 0$, $0 < x < 1$ and $y \geq 0$ and given by $R(x, 0; a) = \frac{1}{\pi}\arcsin(\sqrt{x})$ for $y = 0$ and (5.15) for $y > 0$. The definition of R uses the Appell hypergeometric function F_1 with an Euler-type integral representation,

$$\mathrm{F}_1(a; b, b'; c; z, z') = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \int_0^1 t^{a-1}(1-t)^{c-a-1}(1-zt)^{-b}(1-z't)^{-b'} dt,$$

defined for $|z| < 1$, $|z'| < 1$, and $0 < a < c$.

Although this integral does not have a closed-form solution for arbitrary parameters, in Lemma 9 below we provide exact closed-form expressions for $R\left(x, y; \frac{N}{2}\right)$ for $0 < x < 1$, $y \geq 0$ and positive integers N. Restricting a to be an integer or half integer, or equivalently restricting κ to be an integer, guarantees that these equations may be applied, so that both Bayesian error estimators and their conditional MSE for the Gaussian model with linear classification may be evaluated exactly using finite sums of common single variable functions.

Lemma 9. *Let N be a positive integer, $0 < x < 1$ and $y \geq 0$. Then the function $R\left(x, y; \frac{N}{2}\right)$ defined in the statement of Lemma 7 can be expressed as,*

$$R\left(x, y; \frac{N}{2}\right) = \begin{cases} r(x,y) & \text{if } N = 1, \\ r(x,y) - \frac{\sqrt{y}}{\pi} \sum_{i=1}^{\frac{N-1}{2}} \frac{(2i-2)!!}{(2i-1)!!} \left(\frac{1}{y+1}\right)^{i} & \\ \qquad \times \left(1 - I\left(\frac{y(1-x)}{x+y}; \frac{1}{2}, i\right)\right) & \text{if } N > 1 \text{ is odd}, \\ \frac{1}{\pi} \arcsin\left(\sqrt{x}\right) - \frac{\sqrt{y}}{2} \sum_{i=0}^{\frac{N-2}{2}} \frac{(2i-1)!!}{(2i)!!} \left(\frac{1}{y+1}\right)^{i+\frac{1}{2}} & \\ \qquad \times \left(1 - I\left(\frac{y(1-x)}{x+y}; \frac{1}{2}, i + \frac{1}{2}\right)\right) & \text{if } N > 1 \text{ is even}, \end{cases}$$

where

$$r(x,y) = \frac{1}{\pi} \arcsin\left(\sqrt{\frac{x+y}{1+y}}\right) - \frac{1}{\pi} \arctan\left(\sqrt{y}\right)$$

and we may apply (4.12) to evaluate the regularized incomplete beta function, I, in closed-form.

Proof. If $y = 0$, then we have $R(x, 0; a) = \frac{1}{\pi} \arcsin\left(\sqrt{x}\right)$. The solution for R in the statement of this lemma applies for this case. For $y \neq 0$, to solve R for half integer values we first focus on the Appell function, F_1. Define $w = \frac{x(y+1)}{x+y}$ and $z = \frac{x}{x+y}$, and note that $0 < z < w < 1$. For any real number a, we have the definition,

$$\mathrm{F}_1\left(a+1; \frac{1}{2}, 1; a+2; w, z\right) = (a+1) \int_0^1 u^a (1 - wu)^{-1/2} (1 - zu)^{-1} \, du.$$

With some manipulation we have,

$$\begin{aligned}
&\mathrm{F}_1\left(a+1;\frac{1}{2},1;a+2;w,z\right)\\
&\quad=-\frac{a+1}{z}\int_0^1 u^{a-1}(-zu)\left(1-wu\right)^{-1/2}\left(1-zu\right)^{-1}du\\
&\quad=-\frac{a+1}{z}\left(\int_0^1 u^{a-1}(-zu)\left(1-wu\right)^{-1/2}\left(1-zu\right)^{-1}du\right.\\
&\qquad\left.+\int_0^1 u^{a-1}\left(1-wu\right)^{-1/2}\left(1-zu\right)^{-1}du-\int_0^1 u^{a-1}\left(1-wu\right)^{-1/2}\left(1-zu\right)^{-1}du\right)\\
&\quad=-\frac{a+1}{z}\left(\int_0^1 u^{a-1}\left(1-wu\right)^{-1/2}du-\int_0^1 u^{a-1}\left(1-wu\right)^{-1/2}\left(1-zu\right)^{-1}du\right).
\end{aligned}$$

In the first integral, let $v=wu$. We have,

$$\begin{aligned}
&\mathrm{F}_1\left(a+1;\frac{1}{2},1;a+2;w,z\right)\\
&\quad=-\frac{a+1}{z}\left(w^{-a}\int_0^w v^{a-1}\left(1-v\right)^{-1/2}dv-\int_0^1 u^{a-1}\left(1-wu\right)^{-1/2}\left(1-zu\right)^{-1}du\right).
\end{aligned}$$

The first integral is an incomplete beta function, and the second is again an Appell function, so that

$$\begin{aligned}
&\mathrm{F}_1\left(a+1;\frac{1}{2},1;a+2;w,z\right)\\
&\quad=-\frac{a+1}{zw^a}B\left(a,\frac{1}{2}\right)I\left(w;a,\frac{1}{2}\right)+\frac{a+1}{az}\mathrm{F}_1\left(a;\frac{1}{2},1;a+1;w,z\right).
\end{aligned}$$

A property of the regularized incomplete beta function is $I\left(x;a,b\right)=1-I\left(1-x;b,a\right)$, hence,

$$\begin{aligned}
&\mathrm{F}_1\left(a+1;\frac{1}{2},1;a+2;w,z\right)\\
&\quad=\frac{a+1}{az}\mathrm{F}_1\left(a;\frac{1}{2},1;a+1;w,z\right)-\frac{a+1}{zw^a}B\left(a,\frac{1}{2}\right)\left(1-I\left(1-w;\frac{1}{2},a\right)\right).
\end{aligned}$$

By induction, for any positive integer k,

$$\mathrm{F}_1\left(a+k;\frac{1}{2},1;a+k+1;w,z\right) = \frac{a+k}{az^k}\mathrm{F}_1\left(a;\frac{1}{2},1;a+1;w,z\right) - \frac{a+k}{w^a z^k}\sum_{i=0}^{k-1}\left(\frac{z}{w}\right)^i B\left(a+i,\frac{1}{2}\right)\left(1-I\left(1-w;\frac{1}{2},a+i\right)\right).$$

We apply this to the definition of R in the statement of Lemma 7 to decompose R into one of two Appell functions with known solutions. In particular,

$$R\left(x,y;\frac{N}{2}\right) = \frac{\sqrt{y}}{\pi(N+1)}z^{\frac{N+1}{2}} \times \begin{cases} \mathrm{F}_1\left(1;\frac{1}{2},1;2;w,z\right) & \text{if } N=1,\\ \dfrac{N+1}{2z^{\frac{N-1}{2}}}\mathrm{F}_1\left(1;\dfrac{1}{2},1;2;w,z\right) - \dfrac{N+1}{2wz^{\frac{N-1}{2}}} & \\ \quad\times\displaystyle\sum_{i=0}^{\frac{N-3}{2}}\left(\frac{z}{w}\right)^i B\left(i+1,\frac{1}{2}\right)\left(1-I\left(1-w;\frac{1}{2},i+1\right)\right) & \text{if } N>1 \text{ is odd},\\ \dfrac{N+1}{z^{\frac{N}{2}}}\mathrm{F}_1\left(\dfrac{1}{2};\dfrac{1}{2},1;\dfrac{3}{2};w,z\right) - \dfrac{N+1}{2w^{\frac{1}{2}}z^{\frac{N}{2}}} & \\ \quad\times\displaystyle\sum_{i=0}^{\frac{N-2}{2}}\left(\frac{z}{w}\right)^i B\left(i+\frac{1}{2},\frac{1}{2}\right)\left(1-I\left(1-w;\frac{1}{2},i+\frac{1}{2}\right)\right) & \text{if } N>1 \text{ is even}.\end{cases}$$

After some simplification,

$$R\left(x,y;\frac{N}{2}\right) = \frac{\sqrt{y}}{2\pi} \times \begin{cases} z\mathrm{F}_1\left(1;\frac{1}{2},1;2;w,z\right) & \text{if } N=1,\\ z\mathrm{F}_1\left(1;\frac{1}{2},1;2;w,z\right) & \\ \quad-\displaystyle\sum_{i=0}^{\frac{N-3}{2}}\left(\frac{z}{w}\right)^{i+1} B\left(i+1,\frac{1}{2}\right)\left(1-I\left(1-w;\frac{1}{2},i+1\right)\right) & \text{if } N>1 \text{ is odd},\\ 2\sqrt{z}\mathrm{F}_1\left(\frac{1}{2};\frac{1}{2},1;\frac{3}{2};w,z\right) & \\ \quad-\displaystyle\sum_{i=0}^{\frac{N-2}{2}}\left(\frac{z}{w}\right)^{i+\frac{1}{2}} B\left(i+\frac{1}{2},\frac{1}{2}\right)\left(1-I\left(1-w;\frac{1}{2},i+\frac{1}{2}\right)\right) & \text{if } N>1 \text{ is even}.\end{cases}$$

Finally, to evaluate R it can be shown that

$$\mathrm{F}_1\left(1; \frac{1}{2}, 1; 2; w, z\right) = \frac{2}{\sqrt{z(w-z)}}\left(\arctan\left(\sqrt{\frac{z}{w-z}}\right) - \arctan\left(\sqrt{\frac{z(1-w)}{w-z}}\right)\right)$$

and

$$\mathrm{F}_1\left(\frac{1}{2}; \frac{1}{2}, 1; \frac{3}{2}; w, z\right) = \frac{1}{\sqrt{w-z}} \arctan\left(\sqrt{\frac{w-z}{1-w}}\right).$$

With further simplification, we obtain the result in the statement of the lemma. □

D. Discussion

Perhaps the most important advantage of Bayesian error estimation is that its mathematical framework naturally gives rise to the sample-conditioned MSE performance of any arbitrary error estimate, where uncertainty is modeled relative to the unknown distribution parameters. Prior to this work, RMS for non-hold-out error estimators has always been considered by averaging over the sampling distribution, and nothing could be said about performance for a particular sample. In contrast, the conditional RMS proposed in this chapter formally defines a very practical measure of the expected performance of an error estimate given a fixed sample.

In the next chapter we shall characterize the consistency of the Bayesian error estimator, conditioned upon the sample, and demonstrate consistency for both the discrete and Gaussian models under very mild assumptions. We will show how the sample-conditioned RMS can used for censored sampling, thereby conditioning the sample size on the desired accuracy of the error estimator, and we will apply censored sampling to genomic classification. We will also present simulations to examine the performance characteristics of Bayesian error estimation in relation to the prior distribution and sample size.

CHAPTER VI

CONSISTENCY AND SAMPLE-CONDITIONED MSE PERFORMANCE ANALYSIS[*]

A. Consistency in a Bayesian Framework

A key issue in any estimation scheme is consistency: as more data are collected, will the estimate of a parameter converge to its true value? In our case, it is important to determine for which parameters a Bayesian estimator is consistent. Hence in this section we will be interested in frequentist asymptotics, which concern behavior with respect to a fixed parameter and its sampling distribution.

Suppose that $\theta \in \boldsymbol{\Theta}$ parameterizes a distribution of interest and that $\overline{\theta} \in \boldsymbol{\Theta}$ is the unknown true parameter, where $\boldsymbol{\Theta}$ is the parameter space. Further, let S_∞ represent an infinite sample drawn from the true distribution and S_n denote the first n observations of this sample. The sampling distribution will be specified in the subscript of probabilities and expectations using a notation of the form "$S_\infty|\overline{\theta}$."

A sequence of estimators, $\widehat{\varepsilon}_n(S_n)$, of a sequence of functions of the parameter, $\varepsilon_n(\theta, S_n)$, is said to be weakly consistent at $\overline{\theta}$ if $\widehat{\varepsilon}_n(S_n) - \varepsilon_n(\overline{\theta}, S_n) \to 0$ in probability. If this is true for all $\overline{\theta} \in \boldsymbol{\Theta}$, then we say that $\widehat{\varepsilon}_n(S_n)$ is weakly consistent. L^2 consistency is defined by convergence in the mean-square:

$$\mathrm{E}_{S_n|\overline{\theta}}\left[(\widehat{\varepsilon}_n(S_n) - \varepsilon_n(\overline{\theta}, S_n))^2\right] \to 0.$$

[*] Reprinted with permission from "Exact Sample Conditioned MSE Performance of the Bayesian MMSE Estimator for Classification Error–Part II: Consistency and Performance Analysis" by L. A. Dalton and E. R. Dougherty, 2012, *IEEE Transactions on Signal Processing*, in press, Copyright 2012 by IEEE.

L^2 consistency implies weak consistency. Strong consistency is defined by almost sure convergence:

$$\mathrm{P}_{S_\infty|\overline{\theta}}\left(\widehat{\varepsilon}_n(S_n) - \varepsilon_n(\overline{\theta}, S_n) \to 0\right) = 1. \tag{6.1}$$

If $\widehat{\varepsilon}_n(S_n) - \varepsilon_n(\overline{\theta}, S_n)$ is bounded, which is always true for classifier error estimation, then strong consistency implies L^2 consistency by the Dominated Convergence Theorem. We are also interested in showing that for all $\overline{\theta} \in \mathbf{\Theta}$, $\mathrm{MSE}(\widehat{\varepsilon}_n(S_n)|S_n) \to 0$ (a.s.), or more precisely,

$$\mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{E}_{\theta|S_n}\left[(\widehat{\varepsilon}_n(S_n) - \varepsilon_n(\theta, S_n))^2\right] \to 0\right) = 1. \tag{6.2}$$

We refer to this property as “conditional MSE convergence.”

For Bayesian error estimators, we will see that strong consistency is equivalent to the expected true error converging to the actual true error (a.s.), while conditional MSE convergence is equivalent to the variance of the true error converging to 0 (a.s.). The combination of these two notions of convergence is a strong property for an estimator. Note the similarity between the expectation in (6.2) and in the definition of L^2 consistency. The difference is that in L^2 consistency the expectation is over a sampling distribution for a fixed parameter, whereas in (6.2) it is over a posterior distribution of the parameter for a fixed sample. We will prove (6.1) and (6.2) assuming fairly weak conditions on the model and classification rule.

1. Convergence of Posteriors to Delta Functions

It is essential in our proof to show that the Bayes posterior of the parameter converges in some sense to a delta function on the true parameter. Note in particular that this is a property of the posterior distribution, whereas the preceding definitions of consistency are properties of the estimator itself, which in the case of Bayesian MMSE estimation is only the expected value of the posterior.

We formalize this concept with weak* consistency and to do so we require a few comments regarding measure theory. Assume the sample space, $\mathcal{X}$, and the parameter space, $\mathbf{\Theta}$, are Borel subsets of complete separable metric spaces, each being endowed with the induced σ-algebra from the Borel σ-algebra on its respective metric space. In the discrete model with bin probabilities p_i and q_i, $\mathbf{\Theta} = \{[c, p_1, ..., p_{b-1}, q_1, ..., q_{b-1}] : c, p_i, q_i \in [0,1], i = 1, \ldots, b-1, \sum_{i=1}^{b-1} p_i \leq 1, \sum_{i=1}^{b-1} q_i \leq 1\}$, so $\mathbf{\Theta} \subset \mathbb{R} \times \mathbb{R}^{b-1} \times \mathbb{R}^{b-1}$, which is a normed space for which we use the L^1-norm. Letting $\mathcal{B}$ be the Borel σ-algebra on $\mathbb{R} \times \mathbb{R}^{b-1} \times \mathbb{R}^{b-1}$, $\mathbf{\Theta} \in \mathcal{B}$ and the σ-algebra on $\mathbf{\Theta}$ is the induced σ-algebra $\mathcal{B}_{\mathbf{\Theta}} = \{\mathbf{\Theta} \cap A : A \in \mathcal{B}\}$. In the Gaussian model, $\mathbf{\Theta} = \{[c, \mu_0, \Sigma_0, \mu_1, \Sigma_1] : c \in [0,1], \mu_0, \mu_1 \in \mathbb{R}^D, \Sigma_0$ and Σ_1 are $D \times D$ invertible matrices$\} \subset S = \mathbb{R} \times \mathbb{R}^D \times \mathbb{R}^{D^2} \times \mathbb{R}^D \times \mathbb{R}^{D^2}$, which is a normed space, for which we use the L^1-norm. $\mathbf{\Theta}$ lies in the Borel σ-algebra on S and the σ-algebra on $\mathbf{\Theta}$ is defined in the same manner as in the discrete model. If λ_n and λ are probability measures on $\mathbf{\Theta}$, then $\lambda_n \to \lambda$ weak* (that is, in the weak* topology on the space of all probability measures over $\mathbf{\Theta}$) if and only if $\int f d\lambda_n \to \int f d\lambda$ for all bounded continuous functions f on $\mathbf{\Theta}$. Further, if δ_θ is a point mass at $\theta \in \mathbf{\Theta}$, then it can be shown that $\lambda_n \to \delta_\theta$ weak* if and only if $\lambda_n(U) \to 1$ for every neighborhood U of θ.

Bayesian modeling parameterizes a family of probability measures, $\{F_\theta : \theta \in \mathbf{\Theta}\}$, on $\mathcal{X}$. For a fixed true parameter, $\overline{\theta}$, and assuming an i.i.d. sampling process, we denote the sampling distribution by $F_{\overline{\theta}}^\infty$, which is an infinite product measure on $\mathcal{X}^\infty$. We say that the Bayes posterior of θ is weak* consistent at $\overline{\theta} \in \mathbf{\Theta}$ if the posterior probability of the parameter converges weak* to $\delta_{\overline{\theta}}$ for $F_{\overline{\theta}}^\infty$-almost all sequences. In other words, if for all bounded continuous functions f on $\mathbf{\Theta}$,

$$\mathrm{P}_{S_\infty|\overline{\theta}} \left(\mathrm{E}_{\theta|S_n} [f(\theta)] \to f(\overline{\theta}) \right) = 1. \tag{6.3}$$

Equivalently, we require the posterior probability (given a fixed sample) of any neighborhood, U, of the true parameter, $\overline{\theta}$, to converge to 1 almost surely with respect to the sampling distribution, i.e.,

$$\mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{P}_{\theta|S_n}(U) \to 1\right) = 1. \tag{6.4}$$

The posterior is called weak* consistent if it is weak* consistent for every $\overline{\theta} \in \mathbf{\Theta}$.

We now establish that the Bayes posteriors of c, θ_0 and θ_1 are weak* consistent for both discrete and Gaussian models (in the usual topologies). Throughout, we assume proper priors on these parameters, and that the priors have positive mass on every open set. If the underlying probability mechanism in a Bayesian estimation problem has only a finite number of possible outcomes, e.g., flipping a coin, and the prior probability does not exclude any neighborhood of the true parameter as impossible, it has long been known that posteriors are weak* consistent [102, 103]. Thus, if the Bayes prior of the *a priori* probability of the classes, c, has a beta distribution, which has positive mass in every open interval in $[0, 1]$, then the posterior is weak* consistent. Likewise, since sample points in our discrete classification model also have a finite number of possible outcomes, the posteriors of θ_0 and θ_1 are weak* consistent as n_0 and n_1 go to infinity, respectively.

In a general Bayesian estimation problem with a proper prior on a finite dimensional parameter space, as long as the true data distribution is included in the parameterized family of distributions and some regularity conditions hold, notably that the likelihood is a bounded continuous function of the parameter that is not underidentified (i.e., not flat for a range of values of the parameter) and the true parameter is not excluded by the prior as impossible or on the boundary of the parameter space, then the posterior distribution of the parameter approaches a normal distribution centered at the true mean with variance proportional to $1/n$ as $n \to \infty$ [83]. These regularity conditions hold in our Gaussian model for both classes, $y \in \{0, 1\}$, hence the posterior of θ_y is weak* consistent as n_y goes to infinity.

Owing to the weak* consistency of posteriors for c, θ_0 and θ_1 in the discrete and Gaussian models, for any bounded continuous function f on $\boldsymbol{\Theta}$, (6.3) holds for all $\overline{\theta} = [\overline{c}, \overline{\theta}_0, \overline{\theta}_1] \in \boldsymbol{\Theta}$.

2. Sufficient Conditions for the Consistency of Bayesian Error Estimation

Given a true parameter, $\overline{\theta}$, and a fixed infinite sample, for each n suppose that the true error function, $\varepsilon_n(\theta, S_n)$, is a real measurable function on the parameter space. Define $f_n(\theta, S_n) = \varepsilon_n(\theta, S_n) - \varepsilon_n(\overline{\theta}, S_n)$. Note that the actual true error, $\varepsilon_n(\overline{\theta}, S_n)$, is a constant, and $f_n(\overline{\theta}, S_n) = 0$. Since $\widehat{\varepsilon}_n(S_n) = \mathrm{E}_{\theta|S_n}[\varepsilon_n(\theta, S_n)]$ for the Bayesian error estimator, to prove strong consistency we must show

$$\mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{E}_{\theta|S_n}[f_n(\theta, S_n)] \to 0\right) = 1,$$

and for conditional MSE convergence we must show

$$\begin{aligned}
&\mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{E}_{\theta|S_n}\left[\left(\mathrm{E}_{\theta|S_n}[\varepsilon_n(\theta, S_n)] - \varepsilon_n(\theta, S_n)\right)^2\right] \to 0\right)\\
&= \mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{E}_{\theta|S_n}\left[\left(\mathrm{E}_{\theta|S_n}\left[\varepsilon_n(\theta, S_n) - \varepsilon_n(\overline{\theta}, S_n)\right] - \varepsilon_n(\theta, S_n) + \varepsilon_n(\overline{\theta}, S_n)\right)^2\right] \to 0\right)\\
&= \mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{E}_{\theta|S_n}\left[\left(\mathrm{E}_{\theta|S_n}[f_n(\theta, S_n)] - f_n(\theta, S_n)\right)^2\right] \to 0\right)\\
&= \mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{E}_{\theta|S_n}\left[f_n^2(\theta, S_n)\right] - \left(\mathrm{E}_{\theta|S_n}[f_n(\theta, S_n)]\right)^2 \to 0\right)\\
&= 1.
\end{aligned}$$

Hence, both forms of convergence are proved if for any true parameter $\overline{\theta}$ and both $i = 1$ and $i = 2$,

$$\mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{E}_{\theta|S_n}\left[f_n^i(\theta, S_n)\right] \to 0\right) = 1. \tag{6.5}$$

If the classifier in our original classification problem is fixed, and hence the true error is fixed, then we may define error functions independent of the sample, i.e., $\varepsilon(\theta) = \varepsilon_n(\theta, S_n)$ and $f(\theta) = f_n(\theta, S_n) = \varepsilon(\theta) - \varepsilon(\overline{\theta})$. If the true error function, $\varepsilon(\theta)$, is continuous (as in our discrete model and Gaussian model with linear classification), then (6.5) follows directly from (6.3), which is the definition of the weak* convergence of the posteriors of the parameters.

When applying a classification rule, the classifier and true error may change for each n. Hence, (6.3) cannot be applied directly because $f_n^i(\theta, S_n)$ depends on the sample. To proceed, we place restrictions on the Bayesian model and classification rule. The next two theorems prove that the Bayesian error estimator is both strongly consistent and conditional MSE convergent as long as the true error functions, $\varepsilon_n(\theta, S_n)$, form equicontinuous sets for fixed samples and the posterior is weak* consistent.

Theorem 10. *Let $\overline{\theta} \in \boldsymbol{\Theta}$ represent an unknown true parameter and let $F(S_\infty) = \{f_n(\bullet, S_n)\}_{n=1}^{\infty}$ be a uniformly bounded collection of measurable functions associated with the sample S_∞, where $f_n(\bullet, S_n) : \boldsymbol{\Theta} \to \mathbb{R}$ and $|f_n(\bullet, S_n)| \leq \frac{1}{2}M(S_\infty)$ for each $n \in \mathbb{N}$. If $F(S_\infty)$ is equicontinuous at $\overline{\theta}$ (almost surely with respect to the sampling distribution for $\overline{\theta}$) and the posterior of θ is weak* consistent at $\overline{\theta}$, then*

$$\mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{E}_{\theta|S_n}\left[f_n(\theta, S_n)\right] - f_n(\overline{\theta}, S_n) \to 0\right) = 1.$$

Proof. We begin by examining the probability of interest.

$$\begin{aligned}
&\mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{E}_{\theta|S_n}\left[f_n(\theta, S_n) - f_n(\overline{\theta}, S_n)\right] \to 0\right) \\
&= \mathrm{P}_{S_\infty|\overline{\theta}}\left(\left|\mathrm{E}_{\theta|S_n}\left[f_n(\theta, S_n) - f_n(\overline{\theta}, S_n)\right]\right| \to 0\right) \\
&\geq \mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{E}_{\theta|S_n}\left[\ |f_n(\theta, S_n) - f_n(\overline{\theta}, S_n)|\ \right] \to 0\right).
\end{aligned}$$

Let $d_{\boldsymbol{\Theta}}$ be the metric associated with $\boldsymbol{\Theta}$. For fixed S_∞ and $\epsilon > 0$, if equicontinuity holds for $F(S_\infty)$, there is a $\delta > 0$ such that $|f_n(\theta, S_n) - f_n(\overline{\theta}, S_n)| < \epsilon$ for all $f_n \in$

$F(S_\infty)$ whenever $d_{\boldsymbol{\Theta}}(\theta,\overline{\theta}) < \delta$. Hence,

$$\begin{aligned}
\mathrm{E}_{\theta|S_n}\left[\ |f_n(\theta,S_n) - f_n(\overline{\theta},S_n)|\ \right] &= \mathrm{E}_{\theta|S_n}\left[\ |f_n(\theta,S_n) - f_n(\overline{\theta},S_n)|\mathbf{I}_{d_{\boldsymbol{\Theta}}(\theta,\overline{\theta})<\delta}\ \right] \\
&\quad + \mathrm{E}_{\theta|S_n}\left[\ |f_n(\theta,S_n) - f_n(\overline{\theta},S_n)|\mathbf{I}_{d_{\boldsymbol{\Theta}}(\theta,\overline{\theta})\geq\delta}\ \right] \\
&\leq \mathrm{E}_{\theta|S_n}\left[\epsilon\mathbf{I}_{d_{\boldsymbol{\Theta}}(\theta,\overline{\theta})<\delta}\right] + \mathrm{E}_{\theta|S_n}\left[M(S_\infty)\mathbf{I}_{d_{\boldsymbol{\Theta}}(\theta,\overline{\theta})\geq\delta}\right] \\
&= \epsilon\mathrm{E}_{\theta|S_n}\left[\mathbf{I}_{d_{\boldsymbol{\Theta}}(\theta,\overline{\theta})<\delta}\right] + M(S_\infty)\mathrm{E}_{\theta|S_n}\left[\mathbf{I}_{d_{\boldsymbol{\Theta}}(\theta,\overline{\theta})\geq\delta}\right] \\
&= \epsilon\mathrm{P}_{\theta|S_n}\left(d_{\boldsymbol{\Theta}}(\theta,\overline{\theta}) < \delta\right) + M(S_\infty)\mathrm{P}_{\theta|S_n}\left(d_{\boldsymbol{\Theta}}(\theta,\overline{\theta}) \geq \delta\right).
\end{aligned}$$

From the weak* consistency of the posterior of θ at $\overline{\theta}$, (6.4) holds and we have,

$$\begin{aligned}
\limsup_{n\to\infty} \mathrm{E}_{\theta|S_n}\left[\ |f_n(\theta,S_n) - f_n(\overline{\theta},S_n)|\ \right] &\leq \epsilon \limsup_{n\to\infty} \mathrm{P}_{\theta|S_n}\left(d_{\boldsymbol{\Theta}}(\theta,\overline{\theta}) < \delta\right) \\
&\quad + M(S_\infty)\limsup_{n\to\infty} \mathrm{P}_{\theta|S_n}\left(d_{\boldsymbol{\Theta}}(\theta,\overline{\theta}) \geq \delta\right) \\
&\overset{\text{a.s.}}{=} \epsilon \cdot 1 + M(S_\infty)\cdot 0 = \epsilon.
\end{aligned}$$

Finally, since this is (almost surely) true for all $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathrm{E}_{\theta|S_n}\left[\ |f_n(\theta,S_n) - f_n(\overline{\theta},S_n)|\ \right] \overset{\text{a.s.}}{=} 0,$$

so that the probabilities at the beginning of this proof must all be 1. □

Theorem 11. *Given a Bayesian model and classification rule, if for both $y = 0$ and $y = 1$ we have that $F^y(S_\infty) = \{\varepsilon_n^y(\bullet,S_n)\}_{n=1}^\infty$ is equicontinuous at $\overline{\theta}_y$ (almost surely with respect to the sampling distribution for $\overline{\theta}_y$) for every $\overline{\theta}_y \in \boldsymbol{\Theta}_y$ and the posterior of θ is weak* consistent, then the resulting Bayesian error estimator is both strongly consistent and conditional MSE convergent.*

Proof. We may decompose the true error of a classifier by $\varepsilon_n(\theta,S_n) = c\varepsilon_n^0(\theta_0,S_n) + (1-c)\varepsilon_n^1(\theta_1,S_n)$, and it is not hard to show that $F(S_\infty) = \{\varepsilon_n(\bullet,S_n)\}_{n=1}^\infty$ is also (a.s.) equicontinuous at every $\overline{\theta} = [\overline{c},\overline{\theta}_0,\overline{\theta}_1] \in \boldsymbol{\Theta} = [0,1]\times\boldsymbol{\Theta}_0\times\boldsymbol{\Theta}_1$. Define

$f_n(\theta, S_n) = \varepsilon_n(\theta, S_n) - \varepsilon_n(\overline{\theta}, S_n)$, and note $|f_n(\theta, S_n)| \leq 1$. Since $\{f_n(\bullet, S_n)\}_{n=1}^{\infty}$ and $\{f_n^2(\bullet, S_n)\}_{n=1}^{\infty}$ are also (a.s.) equicontinuous at every $\overline{\theta} \in \mathbf{\Theta}$, by Theorem 10,

$$\mathrm{P}_{S_\infty|\overline{\theta}}\left(\mathrm{E}_{\theta|S_n}\left[f_n^i(\theta, S_n)\right] \to 0\right) = 1$$

for both $i = 1$ and $i = 2$. □

3. Consistency of Bayesian Error Estimation in the Discrete and Gaussian Models

Equicontinuity essentially guarantees that the true errors for designed classifiers are somewhat "robust" near the true parameter. Loosely speaking, with equicontinuity we can (almost surely) find a neighborhood, U, of the true parameter such that the error of all classifiers (for any sample size) at any parameter in U is as close as desired to the true error. This property is only a sufficient condition for consistency but it usually holds. Indeed, the following two theorems prove that it holds for both the discrete and Gaussian Bayesian models. Combining these results with Theorem 11, the Bayesian error estimator is strongly consistent and conditional MSE convergent for both the discrete model with any classification rule and the Gaussian model with any linear classification rule, under our assumptions.

Theorem 12. *In the discrete Bayesian model with any classification rule, $F^y(S_\infty) = \{\varepsilon_n^y(\bullet, S_n)\}_{n=1}^{\infty}$ is equicontinuous at every $\overline{\theta}_y \in \mathbf{\Theta}_y$ for both $y = 0$ and $y = 1$.*

Proof. This is a slightly stronger proof than required in Theorem 11, since equicontinuity is always true for any sample. Also, we need not specify a particular classification rule; any sequence of classifiers may be applied at each n.

In a b bin model, suppose we obtain the sequence of classifiers $\psi_n : \{1, \ldots, b\} \to \{0, 1\}$ from a given sample. The error of classifier ψ_n contributed by class 0 at parameter $\theta_0 = [p_1, \ldots p_{b-1}] \in \boldsymbol{\Theta}_0$ is

$$\varepsilon_n^0(\theta_0, S_n) = \sum_{i=1}^{b} p_i \mathbf{I}_{\psi_n(i)=1}.$$

For any fixed sample, S_∞, fixed true parameter $\overline{\theta}_0 = [\overline{p}_1, \ldots, \overline{p}_{b-1}]$ and any $\theta_0 = [p_1, \ldots, p_{b-1}]$,

$$\begin{aligned} |\varepsilon_n^0(\theta_0, S_n) - \varepsilon_n^0(\overline{\theta}_0, S_n)| &= \left| \sum_{i=1}^{b} (p_i - \overline{p}_i)\, \mathbf{I}_{\psi_n(i)=1} \right| \\ &= \left| \sum_{i=1}^{b-1} (p_i - \overline{p}_i)\, \mathbf{I}_{\psi_n(i)=1} - \sum_{i=1}^{b-1} (p_i - \overline{p}_i)\, \mathbf{I}_{\psi_n(b)=1} \right| \\ &\leq 2 \sum_{i=1}^{b-1} |p_i - \overline{p}_i| = 2\|\theta_0 - \overline{\theta}_0\|. \end{aligned}$$

Since $\overline{\theta}_0$ was arbitrary, $F^0(S_\infty)$ is equicontinuous. Similarly, we may show that $F^1(S_\infty) = \{\sum_{i=1}^{b} q_i \mathbf{I}_{\psi_n(i)=0}\}_{n=1}^{\infty}$ is equicontinuous, which completes the proof. □

Theorem 13. *In the Gaussian Bayesian model with D features and any linear classification rule, $F^y(S_\infty) = \{\varepsilon_n^y(\bullet, S_n)\}_{n=1}^{\infty}$ is equicontinuous at every $\overline{\theta}_y \in \boldsymbol{\Theta}_y$ for both $y = 0$ and $y = 1$.*

Proof. Given S_∞, suppose we obtain a sequence of linear classifiers $\psi_n : \mathbb{R}^D \to \{0, 1\}$ of the form (4.6) with discriminant functions $g_n(\mathbf{x}) = \mathbf{a}_n^T \mathbf{x} + b_n$ defined by vectors $\mathbf{a}_n$ and constants b_n. If $\mathbf{a}_n = 0$ for some n, then the classifier and classifier errors are constant. In this case, $|\varepsilon_n^y(\theta_y, S_n) - \varepsilon_n^y(\overline{\theta}_y, S_n)| = 0$ for all $\theta_y, \overline{\theta}_y \in \boldsymbol{\Theta}_y$, so this classifier does not effect the equicontinuity of $F^y(S_\infty)$. Hence, without loss of generality we assume $\mathbf{a}_n \neq 0$, so that the error of classifier ψ_n contributed by class y at parameter $\theta_y = [\mu_y, \Sigma_y]$ is given by

$$\varepsilon_n^y(\theta_y, S_n) = \Phi\left(\frac{(-1)^y g_n(\mu_y)}{\sqrt{\mathbf{a}_n^T \Sigma_y \mathbf{a}_n}} \right).$$

Since scaling g_n does not effect the decision of classifier ψ_n and $\mathbf{a}_n \neq 0$, without loss of generality we also assume g_n is normalized so that $\max_i |(\mathbf{a}_n)_i| = 1$ for all n, where $(\mathbf{a}_n)_i$ is the ith element of $\mathbf{a}_n$.

Treating both classes at the same time, it is enough to show that $\{g_n(\mu)\}_{n=1}^{\infty}$ is equicontinuous at every $\overline{\mu} \in \mathbb{R}^D$ and $\{\mathbf{a}_n^T \Sigma \mathbf{a}_n\}_{n=1}^{\infty}$ is equicontinuous at every positive definite $\overline{\Sigma}$ (considering one fixed $\overline{\Sigma}$ at a time, by positive definiteness $\mathbf{a}_n^T \overline{\Sigma} \mathbf{a}_n > 0$). For any fixed but arbitrary $\overline{\mu} = [\overline{\mu}_1, \ldots, \overline{\mu}_D]$ and any μ,

$$\begin{aligned} |g_n(\mu) - g_n(\overline{\mu})| &= \left| \sum_{i=1}^{D} (\mathbf{a}_n)_i \left(\mu_i - \overline{\mu}_i\right) \right| \\ &\leq \max_i |(\mathbf{a}_n)_i| \sum_{i=1}^{D} |\mu_i - \overline{\mu}_i| \\ &= \|\mu - \overline{\mu}\|. \end{aligned}$$

This proves that $\{g_n(\mu)\}_{n=1}^{\infty}$ is equicontinuous. For any fixed $\overline{\Sigma}$, we denote $\overline{\sigma}_{ij}$ as its ith row, jth column element and we use similar notation for an arbitrary matrix, Σ. Then,

$$\begin{aligned} |\mathbf{a}_n^T \Sigma \mathbf{a}_n - \mathbf{a}_n^T \overline{\Sigma} \mathbf{a}_n| &= \left|\mathbf{a}_n^T \left(\Sigma - \overline{\Sigma}\right) \mathbf{a}_n\right| \\ &= \left| \sum_{i=1}^{D} \sum_{j=1}^{D} (\mathbf{a}_n)_i (\mathbf{a}_n)_j \left(\sigma_{ij} - \overline{\sigma}_{ij}\right) \right| \\ &\leq \max_i |(\mathbf{a}_n)_i|^2 \sum_{i=1}^{D} \sum_{j=1}^{D} |\sigma_{ij} - \overline{\sigma}_{ij}| \\ &= \|\Sigma - \overline{\Sigma}\|. \end{aligned}$$

Hence, $\{\mathbf{a}_n^T \Sigma \mathbf{a}_n\}_{n=1}^{\infty}$ is equicontinuous. □

B. RMS Bound for the Discrete Model

In the previous section on consistency, we have proven that $\mathrm{MSE}(\widehat{\varepsilon}|S_n) \to 0$ as $n \to \infty$ (almost surely relative to the sampling process) for the discrete model. However, we can go one step further using the formulas derived in the previous chapter to find an upper bound on the conditional MSE as a function of only the sample size under fairly general assumptions. In the discrete model, noting that $\mathrm{Var}_{\pi^*}(\varepsilon_n^0(\theta_0)) = \mathrm{E}_{\pi^*}[(\varepsilon_n^0(\theta_0))^2] - (\widehat{\varepsilon}^0)^2$, we apply (5.9) and after some simplification we have

$$\begin{aligned}\mathrm{Var}_{\pi^*}\left(\varepsilon_n^0(\theta_0)\right) &= \left(\frac{\left(n_0 + \sum_{i=1}^b \alpha_i^0\right)\widehat{\varepsilon}^0 + 1}{n_0 + \sum_{i=1}^b \alpha_i^0 + 1}\right)\widehat{\varepsilon}^0 - \left(\widehat{\varepsilon}^0\right)^2 \\ &= \frac{\widehat{\varepsilon}^0\left(1 - \widehat{\varepsilon}^0\right)}{n_0 + \sum_{i=1}^b \alpha_i^0 + 1}.\end{aligned}$$

Analogous results follow for class 1:

$$\mathrm{Var}_{\pi^*}\left(\varepsilon_n^1(\theta_1)\right) = \frac{\widehat{\varepsilon}^1\left(1 - \widehat{\varepsilon}^1\right)}{n_1 + \sum_{i=1}^b \alpha_i^1 + 1}.$$

Plugging these in (5.1) and applying the beta prior/posterior model for c,

$$\begin{aligned}\mathrm{MSE}(\widehat{\varepsilon}|S_n) = &\frac{(\alpha^0 + n_0)(\alpha^1 + n_1)}{(\alpha^0 + \alpha^1 + n)^2(\alpha^0 + \alpha^1 + n + 1)}(\widehat{\varepsilon}^0 - \widehat{\varepsilon}^1)^2 \\ &+ \frac{(\alpha^0 + n_0)(\alpha^0 + n_0 + 1)}{(\alpha^0 + \alpha^1 + n)(\alpha^0 + \alpha^1 + n + 1)} \times \frac{\widehat{\varepsilon}^0\left(1 - \widehat{\varepsilon}^0\right)}{n_0 + \displaystyle\sum_{i=1}^b \alpha_i^0 + 1} \\ &+ \frac{(\alpha^1 + n_1)(\alpha^1 + n_1 + 1)}{(\alpha^0 + \alpha^1 + n)(\alpha^0 + \alpha^1 + n + 1)} \times \frac{\widehat{\varepsilon}^1\left(1 - \widehat{\varepsilon}^1\right)}{n_1 + \displaystyle\sum_{i=1}^b \alpha_i^1 + 1}.\end{aligned}$$

From this, it is clear that $\mathrm{MSE}(\widehat{\varepsilon}|S_n)$ indeed converges to zero (and these results apply for any classification rule). In particular, as long as $\alpha^0 \leq \sum_{i=1}^b \alpha_i^0$ and $\alpha^1 \leq \sum_{i=1}^b \alpha_i^1$, which is often the case,

$$\begin{aligned}\mathrm{MSE}(\widehat{\varepsilon}|S_n) &\leq \frac{1}{\alpha^0+\alpha^1+n+1}\left(\frac{\alpha^0+n_0}{\alpha^0+\alpha^1+n}\cdot\frac{\alpha^1+n_1}{\alpha^0+\alpha^1+n}(\widehat{\varepsilon}^0-\widehat{\varepsilon}^1)^2\right.\\ &\qquad\left.+\frac{\alpha^0+n_0}{\alpha^0+\alpha^1+n}\widehat{\varepsilon}^0\left(1-\widehat{\varepsilon}^0\right)+\frac{\alpha^1+n_1}{\alpha^0+\alpha^1+n}\widehat{\varepsilon}^1\left(1-\widehat{\varepsilon}^1\right)\right)\\ &= \frac{1}{\alpha^0+\alpha^1+n+1}\Big(E_{\pi^*}\left[c\right]E_{\pi^*}\left[1-c\right](\widehat{\varepsilon}^0-\widehat{\varepsilon}^1)^2\\ &\qquad+E_{\pi^*}\left[c\right]\widehat{\varepsilon}^0\left(1-\widehat{\varepsilon}^0\right)+E_{\pi^*}\left[1-c\right]\widehat{\varepsilon}^1\left(1-\widehat{\varepsilon}^1\right)\Big),\end{aligned}$$

where we have used $E_{\pi^*}\left[c\right]=(\alpha^0+n_0)/(\alpha^0+\alpha^1+n)$ and $E_{\pi^*}\left[1-c\right]=(\alpha^1+n_1)/(\alpha^0+\alpha^1+n)$. To help simplify this equation further, define $x=\widehat{\varepsilon}^0$, $y=\widehat{\varepsilon}^1$ and $z=E_{\pi^*}\left[c\right]$. Then

$$\begin{aligned}\mathrm{MSE}(\widehat{\varepsilon}|S_n) &\leq \frac{z\left(1-z\right)\left(x-y\right)^2+zx\left(1-x\right)+\left(1-z\right)y\left(1-y\right)}{\alpha^0+\alpha^1+n+1}\\ &= \frac{zx+\left(1-z\right)y-\left(zx+\left(1-z\right)y\right)^2}{\alpha^0+\alpha^1+n+1}.\end{aligned}$$

From (2.10) note that $\widehat{\varepsilon}=zx+\left(1-z\right)y$, and also note that $0\leq\widehat{\varepsilon}\leq 1$. Hence,

$$\mathrm{MSE}(\widehat{\varepsilon}|S_n)\leq\frac{\widehat{\varepsilon}-\left(\widehat{\varepsilon}\right)^2}{\alpha^0+\alpha^1+n+1}\leq\frac{1}{4(\alpha^0+\alpha^1+n+1)},$$

where in the last inequality we have used the fact that $w-w^2=w(1-w)\leq 1/4$ whenever $0\leq w\leq 1$. Thus, the conditional RMS of the Bayesian error estimator for any discrete classifier, averaged over all feature-label distributions with beta priors on c and Dirichlet priors on the bin probabilities such that $\alpha^0\leq\sum_{i=1}^{b}\alpha_i^0$ and $\alpha^1\leq\sum_{i=1}^{b}\alpha_i^1$, satisfies

$$\mathrm{RMS}(\widehat{\varepsilon}|S_n)\leq\sqrt{\frac{1}{4n}}. \tag{6.6}$$

Since this bound is only a function of the sample size, it holds if we remove the conditioning on S_n.

For comparison, we consider a remarkably similar holdout bound. If the data are split between training and test data, where the classifier is designed on the training data and classifier error is estimated on the test data, then we have the distribution-free bound

$$\mathrm{RMS}(\widehat{\varepsilon}_{\mathrm{holdout}}|S_{n-m}, c, \theta_0, \theta_1) \leq \sqrt{\frac{1}{4m}}, \tag{6.7}$$

where m is the size of the test sample and S_{n-m} is the training sample [33] . Note that uncertainty here stems from the sampling distribution of the test sample. In any case, the bound is still true if we remove the conditioning. The RMS bound on the Bayesian error estimator is always lower than that of the holdout estimate, which is a testament to the power of modeling assumptions. Moreover, as $m \to n$ for full holdout, the holdout bound converges down to the Bayesian estimate bound.

C. Performance

All synthetic data simulations in this chapter implement a Bayesian model, where we assume known fixed priors, generate random feature-label distributions, and finally generate random samples for each fixed feature-label distribution. Unless otherwise indicated, experiments use a fixed sample size. A summary of the simulation method for fixed sample size experiments is shown in Fig. 29, which lists the general steps and flow of information. The steps are as follows:

- Step 1: Define a fixed set of hyperparameters specifying a specific set of (proper) priors.
  - Define α^0 and α^1 for the prior of c.
  - In a discrete model, define $\alpha_1^0, \ldots, \alpha_b^0$ for the prior of θ_0, and $\alpha_1^1, \ldots, \alpha_b^1$ for the prior of θ_1.
  - In a Gaussian model, define κ, S, ν and $\mathbf{m}$ for both classes.

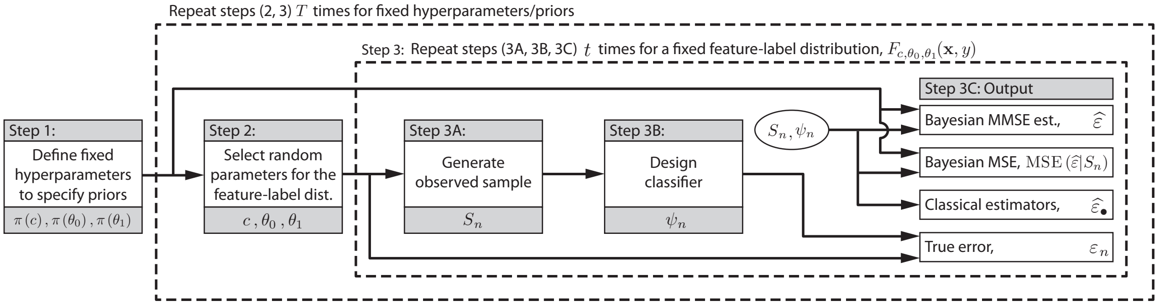

Repeat steps (2, 3) T times for fixed hyperparameters/priors
Step 3: Repeat steps (3A, 3B, 3C) t times for a fixed feature-label distribution, $F_{c,\theta_0,\theta_1}(\mathbf{x}, y)$
Step 1:
Define fixed hyperparameters to specify priors
$\pi(c), \pi(\theta_0), \pi(\theta_1)$
Step 2:
Select random parameters for the feature-label dist.
c, θ_0, θ_1
Step 3A:
Generate observed sample
S_n
Step 3B:
Design classifier
ψ_n
S_n, ψ_n
Step 3C: Output
Bayesian MMSE est., $\widehat{\varepsilon}$
Bayesian MSE, $\mathrm{MSE}(\widehat{\varepsilon}|S_n)$
Classical estimators, $\widehat{\varepsilon}_\bullet$
True error, ε_n

Fig. 29. Simulation methodology for a Bayesian framework with fixed sample size.

- Step 2: Using the priors, generate a random realization of the parameters, $[c, \theta_0, \theta_1]$, corresponding to a fixed feature-label distribution, $F_{c,\theta_0,\theta_1}(\mathbf{x}, y)$.

- Step 3A: Generate a training sample of fixed sample size from the feature-label distribution.

- Step 3B: Design a classifier from the training sample.

- Step 3C: Collect output variables.

  - Compute the Bayesian MMSE error estimator, $\widehat{\varepsilon}$, from the sample, classifier and priors.

  - Compute the Bayesian conditional MSE, $\mathrm{MSE}(\widehat{\varepsilon}|S_n)$, from the sample, classifier and priors.

  - Compute classical error estimators from the sample and classifier.

  - Compute the exact true error from the classifier and true distribution.

Step 2 is repeated T times, to generate T different feature-label distributions. For a fixed feature-label distribution, step 3 (steps 3A through 3C) is repeated t times to obtain t samples and sets of output. In total, each simulation using the model in Fig. 29 will produce $t \times T$ sets of output results.

Some simulation studies will use a censored sampling procedure (to be explained) in place of step 3; nevertheless, all experiments produce the same four quantities in each iteration. From these we compute related results. For instance, although we only evaluate the conditional MSE of the Bayesian error estimator, we may use (5.8) to compute the conditional MSE of any classical error estimator for each iteration. Also, it is possible to approximate the unconditional MSE (averaged over both the feature-label distribution and the sampling distribution) for any error estimator, $\widehat{\varepsilon}_{\bullet}$, using one of two methods:

- Semi-analytical unconditional MSE: average MSE $(\widehat{\varepsilon}_{\bullet}|S_n)$ over iterations/samples.
- Empirical unconditional MSE: compute $(\varepsilon_n - \widehat{\varepsilon}_{\bullet})^2$ for each iteration/sample and average.

The empirical RMS and semi-analytical RMS are the square roots of the empirical MSE and semi-analytical MSE, respectively. We use the semi-analytical unconditional MSE unless otherwise indicated.

We present five simulation studies to demonstrate the power of prior knowledge and modeling assumptions, as well as practical applications of Bayesian error estimation and conditional MSE.

- Bayesian Error Estimation Versus Holdout Error Estimation: this is inspired by the similarity between the performance bounds (6.6) and (6.7).
- Discrete Model with Synthetic Data: here we demonstrate how the theoretical conditional RMS provides practical performance results for small samples.

These are in contrast with distribution free RMS bounds, which are so loose as to be useless for small samples.

- Gaussian Model with Synthetic Data and Fixed Sample Size: these simulations illustrate that different samples condition RMS performance to different extents, and that models using more informative, or "tighter," priors have better RMS performance.

- Gaussian Model with Synthetic Data and Censored Sampling: here we examine a useful application in which sample points are added one at a time until reaching a desired conditional RMS.

- Gaussian Model with Real Breast Cancer Data and Censored Sampling: we provide a detailed example of censored sampling using real breast cancer data.

1. Bayesian Error Estimation Versus Holdout Error Estimation

We use the fixed sample size methodology outlined in Fig. 29 with a discrete model and fixed bin size, b. In step 1, where we define a fixed prior model for c, θ_0 and θ_1, we assume $\alpha^0 = \alpha^1 = 1$ so that the *a priori* probability of both classes is uniformly distributed between 0 and 1. We also assume the bin probabilities of class 0 and 1 have Dirichlet priors given by the hyperparameters $\alpha_i^0 \propto 2b - 2i + 1$ and $\alpha_i^1 \propto 2i - 1$, where the α_i^y are normalized such that $\sum_{i=1}^{b} \alpha_i^y = b$ for both $y \in \{0, 1\}$. Essentially, class 0 tends to assign more weight to bins with a low index, while class 1 assigns a higher weight to bins with a high index. Note that these priors satisfy $\alpha^0 \leq \sum_{i=1}^{b} \alpha_i^0$ and $\alpha^1 \leq \sum_{i=1}^{b} \alpha_i^1$.

In step 2, we generate a random c from the uniform distribution and generate random bin probabilities from our Dirichlet priors by first generating $2b$ independent gamma distributed random variables, $\gamma_i^y \sim \text{gamma}\,(\alpha_i^y)$, $i = 1, \ldots, b$ and $y \in \{0, 1\}$.

The bin probabilities are then given by

$$p_i = \frac{\gamma_i^0}{\sum_{i=1}^b \gamma_i^0} \quad \text{and} \quad q_i = \frac{\gamma_i^1}{\sum_{i=1}^b \gamma_i^1}. \tag{6.8}$$

Having defined a fixed feature-label distribution, we generate a random sample with fixed sample size, n, in step 3A. To do this, the sample size of class 0, n_0, is determined using a binomial(c, n) experiment, and we set $n_1 = n - n_0$. Then n_0 points are drawn from the discrete distribution $\{p_i\}_1^b$ and n_1 points are drawn from the discrete distribution $\{q_i\}_1^b$, resulting in n non-stratified sample points.

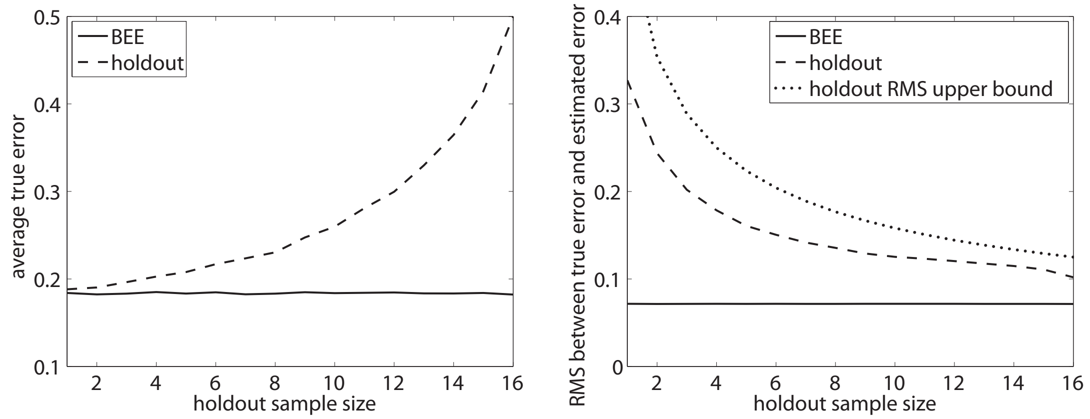
In this study, we are interested in the Bayesian error estimator, which is a full sample error estimator, and the holdout estimator, which partitions the sample into training and testing data sets. For a fair comparison, we will treat both error estimators as separate experiments, each with the same full sample size but with different classifiers designed from different training samples.

To compute the Bayesian error estimator, the full set of labeled sample points is used to train a discrete histogram classifier in step 3B, which uses a majority vote to assign a class to each bin and breaks ties toward class 0. In step 3C, the Bayesian error estimator is found from the full sample, classifier and the same prior probabilities used in the data model (i.e., the correct prior) by evaluating (2.10) with $\mathrm{E}_{\pi^*}[c] = (n_0 + 1)/(n + 2)$, $\widehat{\varepsilon}^0$ defined in (3.5), and $\widehat{\varepsilon}^1$ defined in (3.6). This Bayesian error estimator is theoretically optimal in our Bayesian framework in the mean-square sense. We also evaluate the theoretical RMS conditioned on the sample for the Bayesian error estimate from (5.2) with moments of c defined in (5.3) through (5.7) for $\alpha^0 = \alpha^1 = 1$, $\widehat{\varepsilon}^0$ and $\widehat{\varepsilon}^1$ given in (3.5) and (3.6), and $\mathrm{E}_{\pi^*}[(\varepsilon_n^0(\theta_0))^2]$ and $\mathrm{E}_{\pi^*}[(\varepsilon_n^1(\theta_1))^2]$ given in (5.9) and (5.10). The exact true error of the designed classifier is also computed from the classifier and true distribution, and no other error estimators are computed in this experiment.

To compute the holdout error estimator, the sample is partitioned into training and holdout subsets, where the proportion of points from each class in the holdout set is kept as close as possible to the original sample. The training subset is used to find a discrete histogram classifier with the same classification rule as before and the holdout estimate is the proportion of classification errors on the holdout subset. The exact true error of the designed classifier is also found, but no Bayesian estimates are computed.

Both experiments are run for each sample and the sampling procedure is repeated $t = 10,000$ times for each fixed feature-label distribution. We also generate $T = 10,000$ feature-label distributions (corresponding to randomly selected parameters), for a total of 100 million samples. The sample sizes for each experiment are chosen so that the expected true error of the classifier trained on the full sample is 0.25 when $c = 0.5$ is fixed. Note that the true error here will be somewhat smaller than 0.25, since in these experiments c is uniform. The experiments have been run with different values of b from 2 to 16 and different values of n from 10 to 30.

The results shown in Fig. 30 for $b = 8$ with $n = 16$ are typical, where part (a) shows the expected true error and part (b) shows the RMS between the true and estimated errors, both as a function of the holdout sample size. As expected, the average true error of the classifier in the holdout experiment decreases and converges to the average true error of the classifier trained from the full sample as the holdout sample size decreases. In addition, the RMS performance of the Bayesian error estimator consistently surpasses that of the holdout error estimator, as suggested by the RMS bounds given in (6.6) and (6.7). Thus, under a Bayesian model not only does using the full sample to train the classifier result in a lower true error, but we can achieve better RMS performance using training-data error estimation than we would by holding out the entire sample for error estimation.



(a) average true error

(b) RMS performance

Fig. 30. Comparison of the holdout error estimator and Bayesian error estimator with correct priors with respect to the holdout sample size for a discrete model with $b = 8$ bins and fixed sample size $n = 16$.

2. Discrete Model with Synthetic Data

We again use the fixed sample size methodology outlined in Fig. 29 with a discrete model and fixed bin size, b; however, in step 1 where we define a fixed prior model for c, θ_0 and θ_1, we assume that the *a priori* probability of both classes is known and fixed at 0.5, rather than being uniform, so that both classes are equally likely. For the bin probabilities of class 0 and 1, we assume the same Dirichlet priors as before with hyperparameters $\alpha_i^0 \propto 2b - 2i + 1$ and $\alpha_i^1 \propto 2i - 1$, where the α_i^y are normalized such that $\sum_{i=1}^{b} \alpha_i^y = b$ for both $y \in \{0, 1\}$. For step 2, c is already fixed, and we generate random bin probabilities from our Dirichlet priors using the same method as described in the previous section, that is by first generating $2b$ independent gamma distributed random variables, and then defining p_i and q_i by normalizing these gamma random variables according to (6.8).

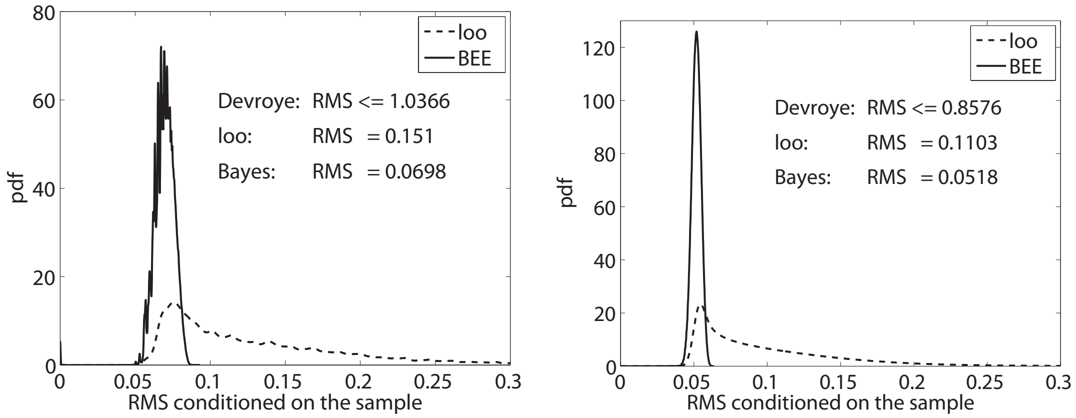
Once the feature-label distribution has been specified by the parameters c, p_i

and q_i, in step 3A we generate a non-stratified random sample with fixed sample size, n. The sample size, n_0, of class 0 is determined from a binomial(c, n) experiment and the sample size of class 1 is set to $n_1 = n - n_0$. Then, the corresponding number of sample points in each class is randomly generated according to the realized bin probabilities. That is, n_0 points are drawn from the discrete distribution $\{p_i\}_1^b$ and n_1 points are drawn from the discrete distribution $\{q_i\}_1^b$, resulting in n non-stratified training points. Although the classes are equally likely, the actual number of sample points from each class may not be the same. In step 3B, these labeled sample points are used to train a discrete histogram classifier, which uses a majority vote to assign a class to each bin and breaks ties toward class 0.

Subsequently in step 3C, the true error of the classifier is computed exactly and the training data are used to evaluate the classical leave-one-out training-data error estimator. We also evaluate a Bayesian error estimator with the same prior probabilities as the data model (i.e., the correct prior). As before, we use (2.10) to evaluate the Bayesian MMSE error estimator, this time with $\mathrm{E}_{\pi^*}[c] = 0.5$. We also evaluate the theoretical RMS conditioned on the sample for the Bayesian error estimator from (5.2), this time with moments of c given by $E_{\pi^*}[c] = E_{\pi^*}[1-c] = 0.5$, $E_{\pi^*}[c^2] = E_{\pi^*}[(1-c)^2] = 0.25$, and $\mathrm{Var}_{\pi^*}(c) = 0$. The conditional RMS for the leave-one-out error estimator is computed from (5.8).

In each simulation iteration, the true error, both error estimates, and their conditional RMS's are recorded. The sampling procedure is repeated $t = 1,000$ times for each fixed feature-label distribution, with $T = 10,000$ feature-label distributions, for a total of ten million samples.

Figures 31(a) and 31(b) show the probability densities of the conditional RMS for both the leave-one-out and Bayesian error estimators with settings $b = 8, n = 16$ and $b = 16, n = 30$, respectively. The sample sizes for each experiment are the same



(a) $b = 8$, $n = 16$

(b) $b = 16$, $n = 30$

Fig. 31. Probability densities for the conditional RMS of the leave-one-out and Bayesian error estimators with correct priors. The sample sizes for each experiment were chosen so that the expected true error is 0.25. The unconditional RMS for both error estimators is also shown, as well as Devroye's distribution free bound.

as in the previous section, chosen so that the expected true error is 0.25. Within each plot, we also show the unconditional semi-analytical RMS of both the leave-one-out and Bayesian error estimators, as well as the distribution free RMS bound on the leave-one-out error estimator for the discrete histogram rule in (1.2). Note that the jaggedness in part (a) is not due to poor density estimation or Monte-Carlo approximation, but rather is caused by the discrete nature of the problem. In particular, the expressions for $\widehat{\varepsilon}^0$, $\widehat{\varepsilon}^0$, $\mathrm{E}_{\pi^*}[(\varepsilon_n^0(\theta_0))^2]$, and $\mathrm{E}_{\pi^*}[(\varepsilon_n^1(\theta_1))^2]$ in (3.5) through (5.10) can take on only a finite set of values, which is especially small for a small number of bins or sample points. In both parts of Fig. 31 (as well as in other unshown plots for different values of b and n), the density of the conditional RMS for the Bayesian error estimator is much tighter than that of leave-one-out. See for example Fig. 31(b), where the conditional RMS of the Bayesian error estimator tends to be very close to 0.05, whereas the leave-one-out error estimator has a long tail with substantial mass between 0.05 and 0.2. Furthermore, the conditional RMS for the Bayesian error estimator is concentrated on lower values of RMS, so much so that in all cases the unconditional RMS of the Bayesian error estimator is less than half that of leave-one-out.

Without any kind of modeling assumptions, distribution-free bounds on the unconditional RMS are too loose to be useful. In fact, Devroye's bound from (1.2) is greater than 0.85 in both subplots of Fig. 31. On the other hand, a Bayesian framework permits us to obtain exact expressions for the RMS conditioned on the sample for both the Bayesian error estimator and any other error estimation rule.

3. Gaussian Model with Synthetic Data and Fixed Sample Size

We next evaluate the performance of Bayesian error estimators on synthetic Gaussian data with LDA classification and a fixed sample size, n. We again use the fixed sample size methodology outlined in Fig. 29, this time with a Gaussian model assuming arbitrary covariances.

In step 1, we assume the *a priori* probability of both classes is known and fixed at 0.5. For the class-conditional distribution parameters, we consider three priors: "low-information," "medium-information," and "high-information" priors, with hyperparameters defined in Table 2. All priors are proper probability densities and are designed to emulate prior knowledge in normalized microarray expression data (see Chapter VII for more information about priors for microarray data). For each prior model, the parameter $\mathbf{m}$ for class 1 has been calibrated to give an expected true error of 0.25 with one feature. The low information prior is closer to a flat non-informative prior and models a setting where our knowledge about the distribution parameters is less certain. Conversely, the high information prior has a relatively tight distri-

Table 2. "Low-information," "medium-information" and "high-information" priors used in the Gaussian model for conditional MSE experiments

Hyperparameter	Low-info prior	Medium-info prior	High-info prior
a priori prob., c	fixed at 0.5	fixed at 0.5	fixed at 0.5
κ, class 0 and 1	$3D$	$9D$	$54D$
S, class 0 and 1	$0.03(\kappa - D - 1)I_D$	$0.03(\kappa - D - 1)I_D$	$0.03(\kappa - D - 1)I_D$
ν, class 0	$6D$	$18D$	$108D$
ν, class 1	$3D$	$9D$	$54D$
$\mathbf{m}$, class 0	$[0, 0, \ldots, 0]$	$[0, 0, \ldots, 0]$	$[0, 0, \ldots, 0]$
$\mathbf{m}$, class 1	$-0.1719[1, 1, \ldots, 1]$	$-0.2281[1, 1, \ldots, 1]$	$-0.2406[1, 1, \ldots, 1]$

bution around the expected parameters and models a situation where we have more certainty. The amount of information in each prior is reflected in the values of κ and ν, which increase as the amount of information in the prior increases.

Since c is fixed at 0.5, in step 2 we only need to generate a random mean and covariance for both classes, μ_0, Σ_0, μ_1, and Σ_1, according to the specified priors. For each class, we first generate a random covariance according to the inverse-Wishart distribution $\pi\left(\Sigma_y\right)$ using methods in [104]. Conditioned by the covariance, we generate a random mean from the Gaussian distribution $\pi\left(\mu_y|\Sigma_y\right) = f_{\mathbf{m},\Sigma_y/\nu}\left(\mu_y\right)$, resulting in a normal-inverse-Wishart distributed mean and covariance pair. The parameters for class 0 are generated independently from those of class 1.

In step 3A, once the feature-label distribution has been specified by the parameters c, μ_0, Σ_0, μ_1, and Σ_1, the sample size, n_0, of class 0 is selected from a binomial(c, n) experiment and $n_1 = n - n_0$. The corresponding number of sample points in each class is generated according to Gaussian(μ_y, Σ_y) distributions. In this way, we generate n non-stratified labeled training points (so that the number of sample points from each class may be different). These labeled sample points are used to train an LDA classifier in step 3B, where no feature selection is involved. In step 3C, the true error of the classifier is computed exactly and the training data are also used to evaluate the classical 5-fold cross-validation training-data error estimator. We also compute a Bayesian error estimator with the same prior probabilities as the data model (the correct prior) from (2.10) with $\mathrm{E}_{\pi^*}[c] = 0.5$ and $\widehat{\varepsilon}^y$ defined in (4.11). Since the classifier is linear, the Bayesian error estimator may be computed exactly using a closed form solution. We also evaluate the theoretical RMS conditioned on the sample for the Bayesian error estimator, using (5.2) with moments of c given by $E_{\pi^*}[c] = E_{\pi^*}[1-c] = 0.5$, $E_{\pi^*}[c^2] = E_{\pi^*}[(1-c)^2] = 0.25$, and $\mathrm{Var}_{\pi^*}(c) = 0$, as well as $\widehat{\varepsilon}^y$ defined in (4.11) and $\mathrm{E}_{\pi^*}[(\varepsilon_n^y(\theta_y))^2]$ defined in (5.19). The conditional RMS for the cross-validation error estimator is computed from (5.8).

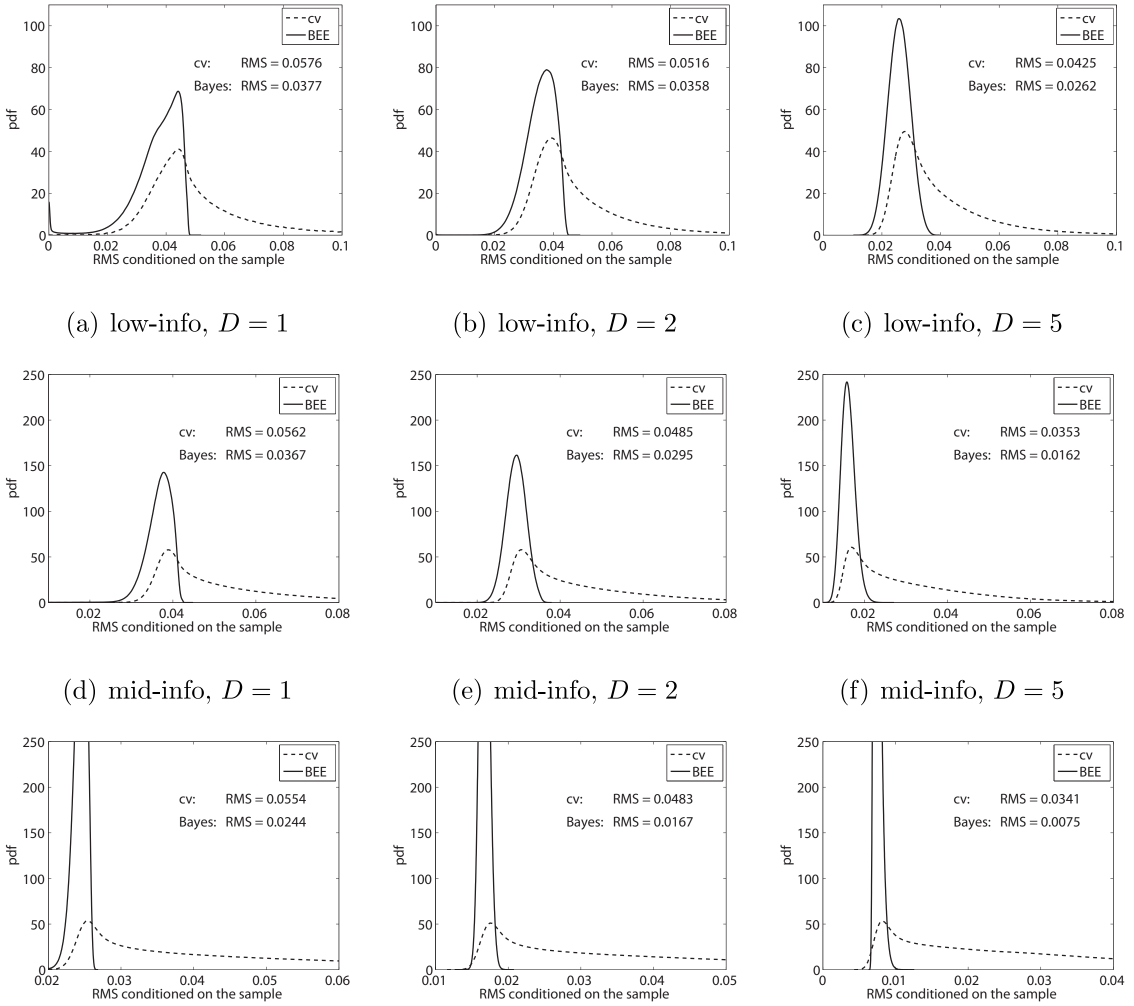
In each iteration the true error, both error estimates, and their conditional RMS's are recorded. The sampling procedure is repeated $t = 1,000$ times for each fixed feature-label distribution, with $T = 10,000$ feature-label distributions, for a total of $t \times T =$ ten million samples.

Table 3 shows the accuracy of the analytical formulas for conditional RMS under nine models using $n = 60$ with different priors (low, medium and high) and feature sizes ($D = 1$, 2, and 5). There is close agreement between the semi-analytical RMS and empirical RMS of the Bayesian error estimator with correct priors. The table also provides the average true errors of each model.

Figure 32 shows the estimated densities of the conditional RMS, found from the conditional RMS values recorded in each iteration of the experiment, for both the cross-validation and Bayesian error estimators with the low, medium and high information priors corresponding to each row. These figures contain the same nine models listed in Table 3 for $n = 60$ sample points. The semi-analytical unconditional

Table 3. Comparison of the semi-analytical unconditional RMS and the empirical RMS for Bayesian error estimators in nine models

Simulation settings	Expected true error	Semi-analytical RMS	Empirical RMS	Difference
low-info, $n = 60,\ D = 1$	0.2473811329	0.0377405474	0.0377457558	5.208×10^{-6}
low-info, $n = 60,\ D = 2$	0.1998595625	0.0358154105	0.0358465148	3.110×10^{-5}
low-info, $n = 60,\ D = 5$	0.1156264112	0.0261524313	0.0262421408	8.971×10^{-5}
mid-info, $n = 60,\ D = 1$	0.2517011284	0.0366527118	0.0365686031	8.411×10^{-5}
mid-info, $n = 60,\ D = 2$	0.1728722842	0.0294667188	0.0294413566	2.536×10^{-5}
mid-info, $n = 60,\ D = 5$	0.0767712062	0.0161898912	0.0162939187	1.040×10^{-4}
high-info, $n = 60,\ D = 1$	0.2483633891	0.0243576546	0.0242240517	1.336×10^{-4}
high-info, $n = 60,\ D = 2$	0.1698410618	0.0167361018	0.0164930674	2.430×10^{-4}
high-info, $n = 60,\ D = 5$	0.0713487938	0.0075466705	0.0075414393	5.231×10^{-6}



(a) low-info, $D = 1$

(b) low-info, $D = 2$

(c) low-info, $D = 5$

(d) mid-info, $D = 1$

(e) mid-info, $D = 2$

(f) mid-info, $D = 5$

(g) high-info, $D = 1$

(h) high-info, $D = 2$

(i) high-info, $D = 5$

Fig. 32. Probability densities for the conditional RMS of the cross-validation and Bayesian error estimators with correct priors and sample size $n = 60$. The unconditional RMS for both error estimators is also indicated.

RMS for each error estimator is also printed in each graph for reference. The high variance of these distributions illustrates that different samples condition the RMS to different extents. For example, in Fig. 32(a) the expected true error is 0.25 and the conditional RMS of the optimal Bayesian error estimator ranges between about 0 and 0.05, depending on the actual observed sample. Meanwhile, the conditional RMS for cross-validation has a much higher variance and is shifted to the right, which is expected since the conditional RMS of the Bayesian error estimator is optimal. Further, the distributions for the conditional RMS of the Bayesian error estimator with high-information priors have a very low variance and are shifted to the left relative to the low information prior, demonstrating that models using more informative, or "tighter," priors have better RMS performance.

4. Gaussian Model with Synthetic Data and Censored Sampling

We now apply the conditional RMS to censored sampling with synthetic data from our Gaussian model with arbitrary covariance matrices. Steps 1 and 2 of the experimental design outlined in Fig. 29 remain exactly the same, that is, we still define a fixed set of hyperparameters (for either the low, medium or high-information prior) and use these priors to generate random feature-label distributions. However, the sampling procedure in step 3 is modified to use censored sampling, as shown in Fig. 33. Instead of fixing the sample size ahead of time, we collect sample points one at a time until the conditional MSE reaches a stopping criterion in the form of a desired conditional RMS.

Since steps 1 and 2 are unchanged, we begin with step 3A. Once the feature-label distribution parameters have been determined, we draw a small initial training sample from the feature-label distribution. The training sample is initialized with $3D$ sample points in each class, for a total of $6D$ sample points. In step 3B, we design an

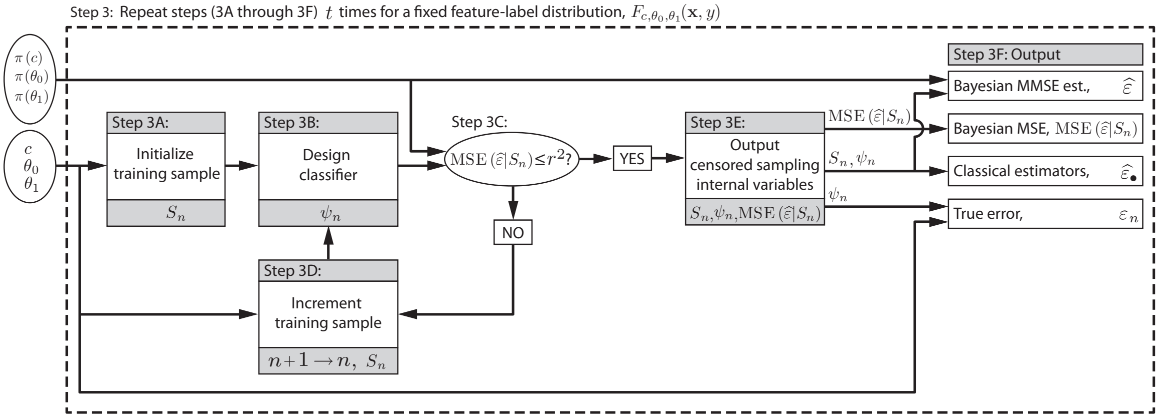

Step 3: Repeat steps (3A through 3F) t times for a fixed feature-label distribution, $F_{c,\theta_0,\theta_1}(\mathbf{x}, y)$
$\pi(c)$
$\pi(\theta_0)$
$\pi(\theta_1)$
c
θ_0
θ_1
Step 3A:
Initialize training sample
S_n
Step 3B:
Design classifier
ψ_n
Step 3C:
$\mathrm{MSE}(\widehat{\varepsilon}|S_n) \leq r^2$?
YES
NO
Step 3D:
Increment training sample
$n+1 \rightarrow n,\ S_n$
Step 3E:
Output censored sampling internal variables
$S_n, \psi_n, \mathrm{MSE}(\widehat{\varepsilon}|S_n)$
$\mathrm{MSE}(\widehat{\varepsilon}|S_n)$
S_n, ψ_n
ψ_n
Step 3F: Output
Bayesian MMSE est., $\widehat{\varepsilon}$
Bayesian MSE, $\mathrm{MSE}(\widehat{\varepsilon}|S_n)$
Classical estimators, $\widehat{\varepsilon}_{\bullet}$
True error, ε_n

Fig. 33. Simulation methodology for a Bayesian framework with censored sampling.

LDA classifier on the initial training sample with no feature selection. In step 3C, we check the current conditional MSE for the initial training sample. If $\mathrm{MSE}(\widehat{\varepsilon}|S_n) > r^2$, for some fixed constant, r, representing the desired RMS (which will be specified shortly), then we append a new point to the current sample in step 3D. To do this, we first establish the label of the new sample point from an independent Bernoulli(c) experiment, and then draw the sample point from the corresponding class-conditional distribution. We then design a new classifier (step 3B) and check the conditional MSE again (step 3C). This is repeated until $\mathrm{MSE}(\widehat{\varepsilon}|S_n) \leq r^2$, in which case we stop the sampling procedure, because we have reached the desired MSE, and move on to step 3E. The sample size is different in each trial because the conditional MSE depends on the actual data obtained from sampling. The consistency of Bayesian error estimation guarantees that $\mathrm{MSE}(\widehat{\varepsilon}|S_n)$ will eventually reach the stopping criterion, so that censored sampling may work to any degree desired.

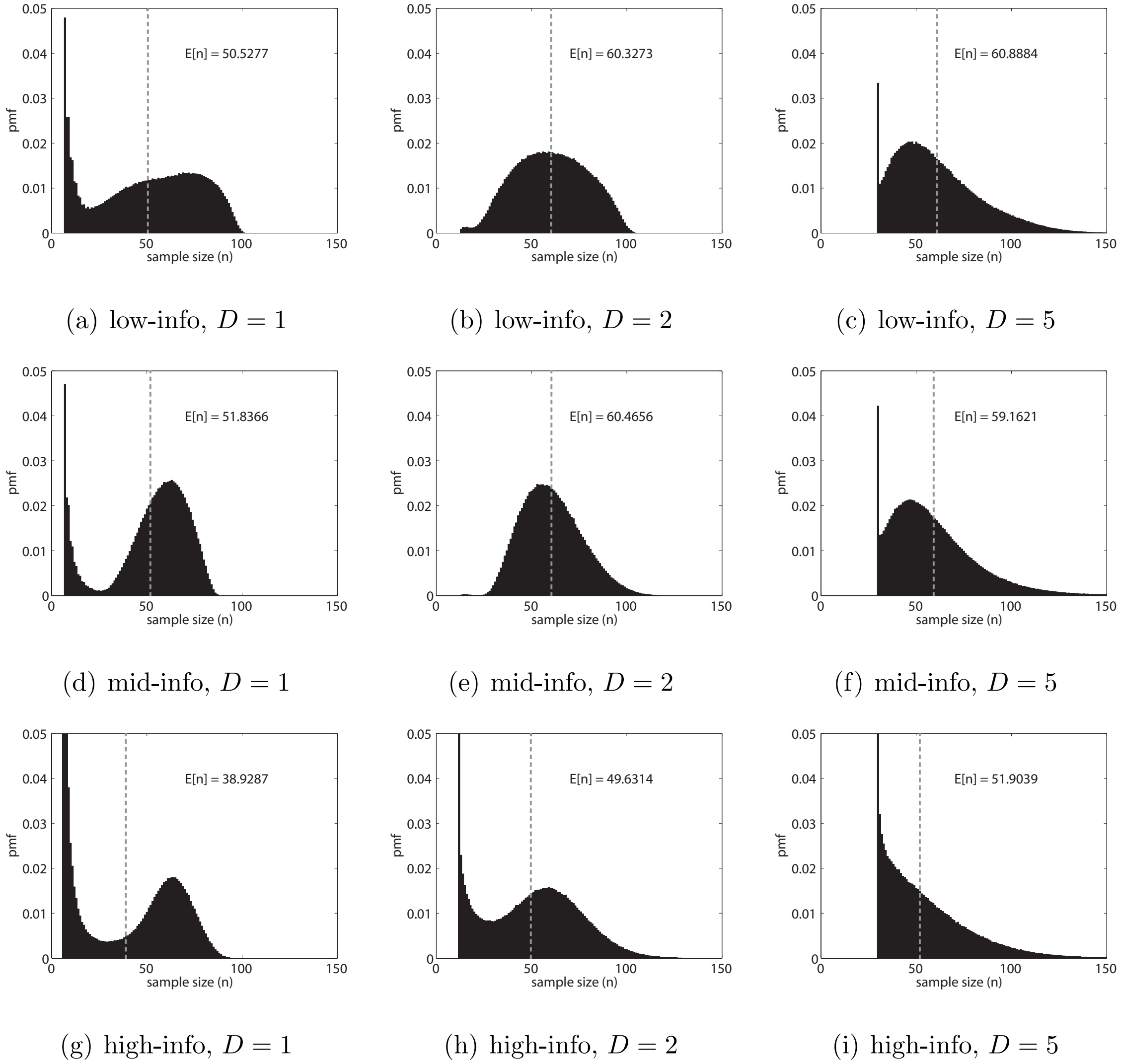
Having completed the sampling procedure, in step 3E we collect three internal variables, including the final censored sample, the classifier designed from the final censored sample, and the conditional MSE computed from the final censored sample.

From these, in step 3F we find the exact true error (from the classifier and the true distribution), a 5-fold cross-validation estimate (from the censored sample) and a Bayesian error estimate (from the censored sample, classifier and correct priors) exactly as in the fixed sample size experiment. The conditional MSE need not be computed again, since it has already been found in the censored sampling procedure. Step 3 is repeated $t = 1,000$ times for each fixed feature-label distribution, with $T = 1,000$ random feature-label distributions for a total of $t \times T =$ one million samples.

It remains to specify a desired RMS, r, for each experiment. In this study, we apply censored sampling to each of our original nine models (low, medium and high-information priors with $D = 1, 2,$ and 5). For each model, the desired conditional RMS of the Bayesian error estimator is set to the semi-analytical RMS reported in Table 3 for the fixed sample experiments with $n = 60$.

Distributions of the sample size obtained in the censored sampling experiments are shown in Fig. 34 with the low, medium and high-information priors corresponding to each row. The means of the distributions are indicated with vertical dotted lines, and spikes seen on the left side of some subplots, for example in Fig. 34(f), are caused because the censored sample size starts at $6D$ and any mass of the probability density for smaller sample sizes is concentrated at this value. For reference, a summary of simulation results for each of the nine censored sampling experiments is provided in Table 4.

In all cases, the RMS with censored sampling is slightly less than the RMS with fixed sampling, which is expected since the conditional RMS with censored sampling is upper bounded for each final sample in the censored sampling process. Further, note that in the worst case the expected sample size is only slightly larger than 60, especially for mid or low-information priors and higher dimensions. Cases where the



(a) low-info, $D = 1$

(b) low-info, $D = 2$

(c) low-info, $D = 5$

(d) mid-info, $D = 1$

(e) mid-info, $D = 2$

(f) mid-info, $D = 5$

(g) high-info, $D = 1$

(h) high-info, $D = 2$

(i) high-info, $D = 5$

Fig. 34. Density of sample size when using censored sampling with correct priors. For each subplot, the desired conditional RMS of the Bayesian error estimator is set to the semi-analytical RMS reported in Table 3 for sample size $n = 60$. The vertical dotted line indicates the mean sample size.

Table 4. Comparison of the semi-analytical unconditional RMS of the Bayesian error estimator for fixed sample size experiments and censored sampling experiments

Simulation settings	Average true error		Semi-analytical RMS		Average sample size
($n = 60$)	Fixed sample	Censored sample	Fixed sample	Censored sample	Censored sample
low-info, $D = 1$	0.2473811329	0.2800682853	0.0377405474	0.0344056212	50.527672
low-info, $D = 2$	0.1998595625	0.2024042627	0.0358154105	0.0354051929	60.327343
low-info, $D = 5$	0.1156264112	0.1143676417	0.0261524313	0.0259373140	60.888388
mid-info, $D = 1$	0.2517011284	0.2746156724	0.0366527118	0.0344869815	51.836600
mid-info, $D = 2$	0.1728722842	0.1723720336	0.0294667188	0.0293098530	60.465612
mid-info, $D = 5$	0.0767712062	0.0757449018	0.0161898912	0.0160699276	59.162091
high-info, $D = 1$	0.2483633891	0.2881658720	0.0243576546	0.0231655105	38.928677
high-info, $D = 2$	0.1698410618	0.1743022396	0.0167361018	0.0166196726	49.631354
high-info, $D = 5$	0.0713487938	0.0692469271	0.0075466705	0.0074534698	51.903893

average sample size is slightly larger than 60 may be explained by a fundamental tradeoff between sample size and RMS, where in this case the RMS is slightly lower. On the other hand, the high-information prior has an expected sample size significantly smaller than 60. A key point is that the distributions in Fig. 34 have very wide supports, illustrating that the sample significantly conditions the RMS.

Note that one need take caution when using a smaller sample size because the classifier does not take advantage of the information in the prior and the true error of the classifier may increase. This effect may be alleviated by adding an additional condition to stop collecting samples once the Bayesian error estimate itself (the expected true error) also reaches a desired threshold.

Even when the fixed and censored sample experiments have essentially the same unconditional RMS and average sample size, recall from the previous section that the conditional RMS in the fixed sample size experiment has a high variance. In contrast, censored sampling experiments enjoy a nearly fixed conditional RMS for each censored sample. Hence, censored sampling provides the same RMS and average sample size or better, while also guaranteeing a specified conditional RMS for each final sample in the censored sampling process. We are exploiting a duality between RMS and sample size: if we fix sample size, we observe in Fig. 32 that the conditional RMS has a large variance, but if we fix RMS, in Fig. 34 the sample size has a large variance.

5. Gaussian Model with Real Breast Cancer Data and Censored Sampling

In this section, we apply censored sampling to classification using genomic data but before doing so we need to explain the difference in the simulation methodology used for real data and that for synthetic data. Heretofore we have employed two randomizations: randomization of the feature-label distribution (fixed for an iteration) and randomization of the samples (from the selected feature-label distribution). In effect, each iteration involves the assumption of a (randomly selected) "true" distribution and, since we want a global performance analysis not dependent on any specifically assumed "true" distribution, we average over all distributions and samples. Now, suppose we want to consider performance for a specific true distribution, as would be the case if we are considering samples from a real-data distribution. Then we would not indulge in the randomization of the feature-label distribution; rather, we would fix it and only average over the samples. The prior distribution would still be involved because it plays a role in error estimation and the computation of MSE $(\widehat{\varepsilon}|S_n)$, but we are no longer interested in averaging performance across the prior distribution. This is precisely the approach taken in this section. The simulation methodology, outlined in Fig. 35, is similar to the censored sampling experiments in Section VI.C.4; however, since there is a fixed true feature-label distribution, we do not simulate steps 1 or 2 in Fig. 29. We also only consider the empirical RMS method in accessing performance relative to the data set.

Proceeding, we apply censored sampling to normalized gene-expression measurements from the same breast cancer study [99] used in Section IV.B.4. The data set includes 295 sample points, each with a 70 feature gene profile. 180 points are assigned to class 0 (good prognosis) and 115 to class 1 (bad prognosis). We choose conservative non-informative priors for the Bayesian estimator. In particular, we assume c is uniform from 0 to 1, and that the priors for both classes are improper flat distributions such that $\pi(\theta_0) = \pi(\theta_1) \propto 1$.

In step A, we randomly select an initial sample from the data set without replacement. The training sample is initialized with $6D$ stratified sample points, where the ratio of points from each class is kept as close as possible to that of the original data set. In step B, we design an LDA classifier on the initial training sample. To simplify the analysis, the classifier is designed from fixed feature sets: $\{\text{CENPA}\}$ for $D = 1$,

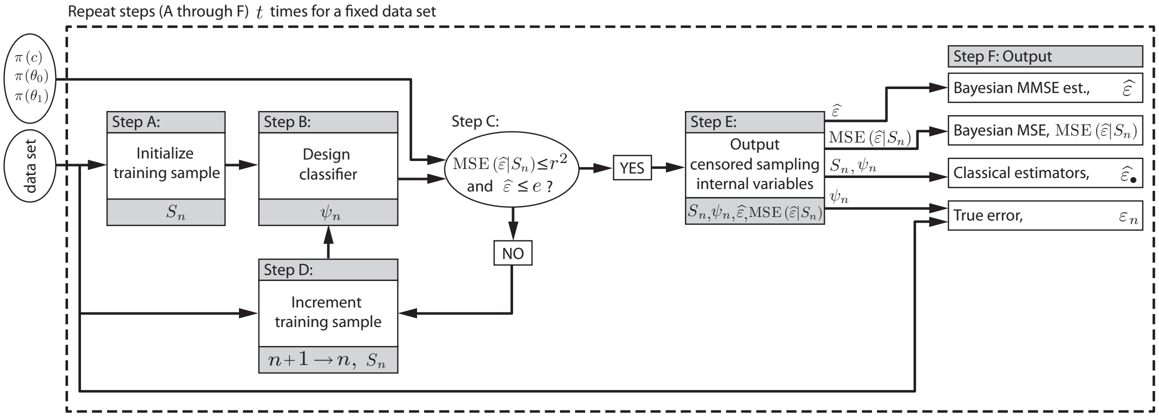

Repeat steps (A through F) t times for a fixed data set
$\pi(c)$
$\pi(\theta_0)$
$\pi(\theta_1)$
data set
Step A:
Initialize training sample
S_n
Step B:
Design classifier
ψ_n
Step C:
$\mathrm{MSE}(\widehat{\varepsilon}|S_n) \leq r^2$ and $\widehat{\varepsilon} \leq e$?
YES
NO
Step D:
Increment training sample
$n+1 \rightarrow n,\ S_n$
Step E:
Output censored sampling internal variables
$S_n, \psi_n, \widehat{\varepsilon}, \mathrm{MSE}(\widehat{\varepsilon}|S_n)$
$\widehat{\varepsilon}$
$\mathrm{MSE}(\widehat{\varepsilon}|S_n)$
S_n, ψ_n
ψ_n
Step F: Output
Bayesian MMSE est., $\widehat{\varepsilon}$
Bayesian MSE, $\mathrm{MSE}(\widehat{\varepsilon}|S_n)$
Classical estimators, $\widehat{\varepsilon}_\bullet$
True error, ε_n

Fig. 35. Simulation methodology for censored sampling with real data.

$\{\mathrm{CENPA}, \mathrm{BBC3}\}$ for $D = 2$ and $\{\mathrm{CENPA}, \mathrm{BBC3}, \mathrm{CFFM4}, \mathrm{TGFB3}, \mathrm{DKFZP564D0462}\}$ for $D = 5$. For all feature sets considered, a multivariate Shapiro-Wilk test applied to the full data set does not reject Gaussianity over either of the classes at a 95% significance level [28]. Although we do not implement a feature selection scheme, one can be applied as part of the classifier design in step B.

Assuming flat priors, in step C we evaluate the Bayesian error estimate (the expected true error) as well as the conditional MSE of the Bayesian error estimate for the initial sample. Letting $r = 0.05$ and $e = 0.30$ be the maximum acceptable RMS and error, respectively, if $\mathrm{MSE}(\widehat{\varepsilon}|S_n) > r^2$ or $\widehat{\varepsilon} > e$, then we append a new point to the current sample in step D, which is selected randomly from remaining points in the data set independently of the label and without replacement. We then design a new classifier (step B) and check the conditional MSE and expected true error again (step C).

Ideally, this is repeated until $\mathrm{MSE}(\widehat{\varepsilon}|S_n) \leq r^2$ and $\widehat{\varepsilon} \leq e$, in which case we stop the sampling procedure because we have reached our desired MSE and acceptable error and move on to step E. The consistency of Bayesian error estimation guarantees that $\mathrm{MSE}(\widehat{\varepsilon}|S_n)$ will eventually reach the stopping criterion (assuming the true distributions are truly Gaussian) and, assuming the classification rule is consistent, $\widehat{\varepsilon}$ is also guaranteed to reach its stopping criterion so long as the optimal linear classifier has error less than the acceptable error. That being said, because we need an accurate estimate of the true error in the simulation, if convergence is too slow, then we stop the sampling procedure at $n = 100$ to ensure there are enough data points left over to obtain an accurate holdout estimate of the true error. In practice, if, after a large amount of sampling, $\widehat{\varepsilon}$ does not fall below e, then we simply assume that we cannot achieve an acceptable classification error for the problem at hand.

Having completed the sampling procedure, in step E we collect four internal variables: the final censored sample, its corresponding classifier, the Bayesian error estimate, and the conditional MSE. In step F we approximate the true error of the classifier using (holdout) points remaining in the data set (after censored sampling). The Bayesian error estimator and conditional MSE need not be computed again, since they have already been found in the censored sampling procedure. This entire process is repeated $t = 100,000$ times.

In Table 5 we provide a detailed example of the censored sampling procedure from a single iteration of an experiment with $D = 1$. As sample points are added, the expected true error of the classifier tends to decrease, while the conditional MSE decreases almost monotonically. We list the actual sample points in the initial sample (4 in class 0 and 2 in class 1), along with the initial Bayesian error estimate and conditional MSE. These are followed by the sample points added in each repetition of the procedure, along with the current Bayesian error estimate and conditional MSE computed as each point is added. Finally, in this example we stop at a sample size of 37 because the stopping criteria are satisfied: $\widehat{\varepsilon} = 0.149821 \leq 0.30$ and $\mathrm{RMS}(\widehat{\varepsilon}|S_n) = 0.049262 \leq 0.05$. The approximate true error of the designed classifier, found using

Table 5. Censored sampling example for real breast cancer data experiments with flat priors

Index	Label	Point	$\widehat{\varepsilon}$	RMS($\widehat{\varepsilon}\|S_n$)
Initial sample:				
1	class 0	0.309		
2	class 0	-0.127		
3	class 0	0.153		
4	class 0	0.473		
5	class 1	-0.485		
6	class 1	0.160	0.360119	0.243913
Appended sample points:				
7	class 0	-0.114	0.304501	0.218271
8	class 0	0.357	0.253987	0.196361
9	class 1	-0.399	0.293964	0.219222
10	class 1	-0.718	0.190709	0.120224
11	class 0	0.391	0.162752	0.108650
12	class 1	-0.909	0.141007	0.095877
Appended sample points (cont'd):				
13	class 0	0.355	0.120434	0.086282
14	class 0	0.385	0.104277	0.078103
15	class 1	-0.273	0.114879	0.076103
⋮	⋮	⋮	⋮	⋮
30	class 0	-0.180	0.167546	0.057499
31	class 0	0.655	0.165431	0.056250
32	class 0	0.284	0.157773	0.054302
33	class 0	0.616	0.153618	0.052873
34	class 1	-0.429	0.151831	0.051866
35	class 0	0.160	0.146605	0.050339
36	class 1	-0.250	0.152377	0.050352
37	class 0	0.038	0.149821	0.049262
Approximate true error:			0.197674	

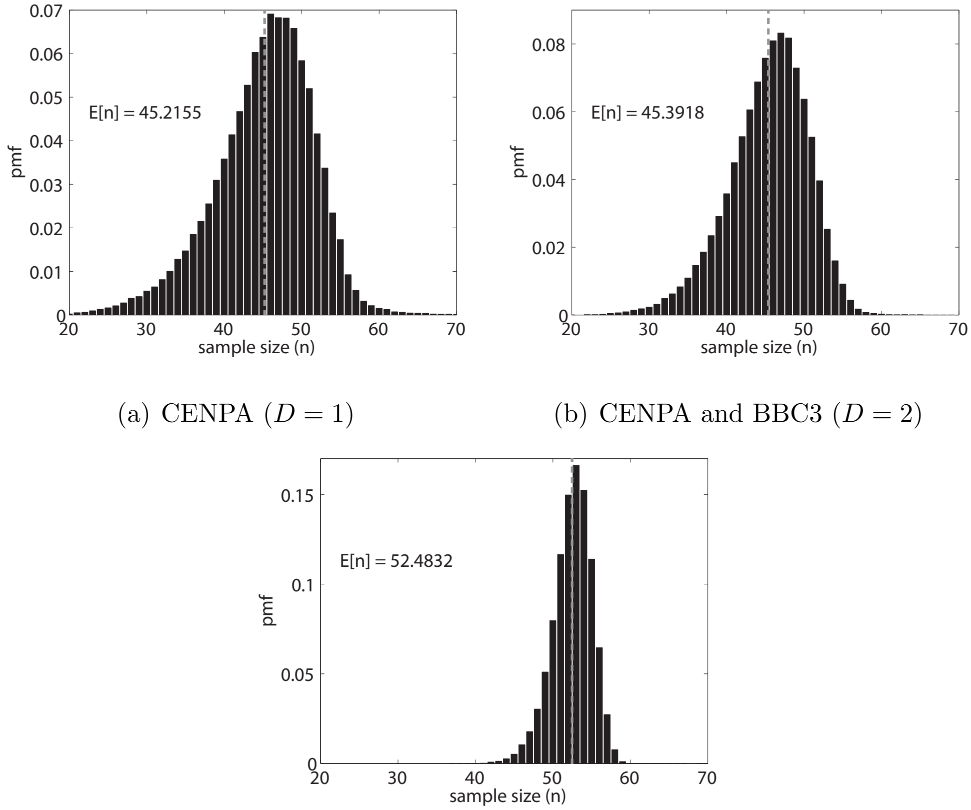
Table 6. Simulation results for real breast cancer data with censored sampling and flat priors

Features	Average n	Average $\widehat{\varepsilon}$	Bias	Empirical RMS
$D = 1$	45.21553	0.2044588007	9.478×10^{-4}	0.0543953165
$D = 2$	45.39178	0.1973571098	-1.541×10^{-3}	0.0471570662
$D = 5$	52.48325	0.2004421915	-4.537×10^{-3}	0.0462898593

the holdout sample points, is 0.197674.

Average simulation results are shown in Table 6. Note that the empirical RMS is very close to our desired RMS, $r = 0.05$. Since the average Bayesian error estimate is much less than our desired maximum error of 0.30, in most cases this bound was met well before the RMS bound. There is no guarantee, for a fixed distribution with censored sampling, that the empirical RMS (which in this case is essentially the RMS conditioned on the distribution) will be bounded by the desired RMS (which bounds the RMS conditioned on any particular censored sample), in fact it could be either higher or lower as reflected in Table 6. This is because the RMS conditioned on the sample, for any individual sample, is not comparable to the RMS conditioned on the distribution. The empirical RMS being bounded by the desired RMS is only guaranteed when the empirical RMS is found by averaging over all distributions in the model.

Finally, we provide a distribution of the sample size in each experiment in Fig. 36. Even though in this experiment all samples are drawn from the same distribution, we observe a relatively large range of sample sizes, though the variance of the sample size is much smaller for a higher number of (fixed) features. This may be caused by the increased average sample size, possibly because larger samples drawn from a relatively small real data set are more likely to have common points, or larger samples



(a) CENPA ($D = 1$)

(b) CENPA and BBC3 ($D = 2$)

(c) all 5 genes ($D = 5$)

Fig. 36. Density of sample size when using censored sampling with empirical measurements from a breast cancer study. Both classes have improper non-informative priors. The vertical dotted line indicates the mean sample size.

are more likely to faithfully represent the true distribution with posteriors closer to delta functions on the true parameters.

These results again suggest that different samples condition the RMS to different extents, even when samples are drawn from the same distribution. Hence, using the conditional RMS to produce a censored sample with precisely the RMS necessary for the experiment at hand can be a very attractive and economical sampling method.

D. Discussion

Although Bayesian error estimators are not distribution-free, frequentist consistency still holds for Bayesian error estimation in both the discrete model and Gaussian model with linear classifiers for all distributions in the parameterized model family. We have also analytically characterized the accuracy advantage of Bayesian error estimation over holdout, thereby showing that the use of prior knowledge can simultaneously provide better classification performance and better error estimation.

Not only may we observe convergence in the error estimator, but we expect the sample-conditioned RMS converges to zero as well. This suggests an important application in censored sampling, where sample points are collected one at a time until the conditional MSE reaches an acceptable level, thereby guaranteeing a desired error-estimation accuracy with minimal sampling cost.

Extensive simulations presented in this chapter examine RMS performance characteristics of Bayesian error estimation relative to the priors for both fixed sample and censored sample experiments. Two main realizations emerge from the new sample-conditioned MSE. First, under Bayesian models the sample conditions the uncertainty, and different samples condition it to different extents. Second, models using more informative, or “tighter,” priors have better RMS performance.

CHAPTER VII

APPLICATION OF BAYESIAN MMSE ERROR ESTIMATION TO GENE-EXPRESSION MICROARRAY DATA*

Two practical problems naturally arise in Bayesian error estimation. First, how does one arrive at a prior distribution governing the model? This issue arises in any Bayesian approach, and the current chapter proposes a method to calibrate priors using discarded microarray data. The second issue is the availability of analytic expressions for Bayesian MMSE estimators. Although the Bayesian error estimator has been solved in both the discrete and Gaussian models, here we also demonstrate how to approximate Bayesian error estimators when closed-form representations are not available. While we are not advocating the abandonment of analytic methods, it is practically useful to have software that can evaluate Bayesian MMSE estimators via Monte-Carlo methods. Currently, approximation is necessary in the Gaussian model when using a non-linear classifier, since a closed form solution is not known. Software is publicly available at http://gsp.tamu.edu/Publications/supplementary/dalton11a.

A. Modeling Microarray Data

We assume two classes and require the training sample to consist of normalized log-ratios. Thus, use of normalization schemes such as total intensity normalization or the LOESS method, which are popular transformations before high-level analysis is applied, are required. Log-transformed gene expression values have nearly Gaussian class-conditional distributions (with unknown parameters) [105, 106]. To further val-

*Reprinted with permission from “Application of the Bayesian MMSE Estimator for Classification Error to Gene Expression Microarray Data” by L. A. Dalton and E. R. Dougherty, 2011, *Bioinformatics*, vol. 27, no. 13, pp. 1822–1831, Copyright 2011 by Oxford University Press.

idate a Gaussian modeling assumption, during feature selection we will permit only features that pass a Shapiro-Wilk Gaussianity test. Note that Bayesian error estimators designed under the Gaussian model were shown in Chapter IV to be robust in the sense that performance is still good when the true distributions are Johnson distributions, which are a class of non-Gaussian distributions with four free parameters to control mean, variance, skewness and kurtosis.

Normal-inverse-Wishart priors compose a flexible class of distributions with many degrees of freedom to facilitate the calibration of priors for gene-expression microarrays. Further, this family of priors possesses a fast closed-form solution when used with linear classification. In problems where the Gaussian model applies and one wishes to use a linear classifier, the benefit one might gain by having more control over the prior is not worth the much greater amount of time required to run an integral approximation code and the effort of designing a specialized model, especially for small samples where one cannot afford a very complex model anyway. Hence, we focus on calibrating normal-inverse-Wishart priors.

Assuming the parameters between classes are fairly independent, we have justified the assumptions posed in the definition of the Bayesian error estimator, the others being that the class-conditional distributions are relatively Gaussian and that normal-inverse-Wishart priors are adequate for representing prior knowledge. We are left to devise a method of calibrating priors for the mean and covariance of each class.

B. Implementation of Exact and Approximate Bayesian Error Estimators

Throughout this chapter, we will assume the Gaussian model with arbitrary covariance matrices defined in Section IV.A.6, so that the prior and posterior of θ_y are normal-inverse-Wishart distributions. We fix the hyperparameters for the priors of each class and use the observed sample to update the hyperparameters of the posteriors. We also check that these posteriors are valid density functions, and if they are not, by default the code reports the error contributed by that class to be 0.5. Note that the Bayesian error estimator is most useful in a small sample setting, but the sample size must not be so small that the posterior is not a valid density function. This may happen, for instance, if we use a flat prior with $\kappa + D + 2 = 0$ and the sample size for class y is $n_y \leq 2D + 1$, so that $\kappa^* = \kappa + n_y \leq D - 1$. In such cases, the Bayesian error estimator is meaningless because the available information is not sufficient for estimation, but generally there are also too few sample points for any error estimator to provide meaningful results.

Given valid normal-inverse-Wishart posteriors, the closed form Bayesian error estimator in Equation (4.11) for linear classification is easily evaluated. For arbitrary classifiers, we approximate the Bayesian error estimator in Equation (2.11) with a Monte-Carlo approach. For each class, we generate a random mean and covariance pair according to the specified posterior normal-inverse-Wishart distribution. Several algorithms for generating normal-inverse-Wishart distributed multivariate sample points are available, for example see [104]. For each mean and covariance pair, the true error contributed by the class for the designed classifier is approximated by generating 10,000 sample points from the Gaussian distribution having the specified mean and covariance, and finding the error of these sample points on the classifier. The Bayesian error estimator is computed by averaging these true errors over 2,500 random sets of mean and covariance pairs.

A toolbox of C code for Bayesian error estimation is publicly available. This includes the exact Bayesian error estimator for linear classifiers, the approximation code described above for arbitrary classifiers, a three-stage feature selection algorithm discussed in the next section, as well as code implementing the method of generating priors described in Section VII.D. Simulations demonstrating the accuracy of this approximation with synthetic data and LDA classification are available in the supplementary material of [53].

C. Feature Selection

We use a three-stage feature selection method based on the t-test and a Gaussianity test to reduce the original feature set to D features. Since this work is not focused on optimizing a classification scheme, but rather on investigating the performance of error estimators, this feature selection scheme is intended to be a simple possible scheme to produce highly differentially expressed Gaussian features.

In the first stage, only highly differentially expressed features or features with a high likelihood of biological significance are selected. These may be selected by a t-test or based on biological knowledge. This stage reduces the number of features from tens of thousands to a few hundred. The second stage applies a Shapiro-Wilk hypothesis test [107] on each feature of each class. Only features passing the Shapiro-Wilk test with 95% confidence in both classes are used, unless there are not enough features passing the test, in which case we select a fixed number of features with the highest sum of the Shapiro-Wilk test statistics in each class. In the final stage of feature selection, we reduce the feature set to D features. This is done either by applying a t-test if it has not already been applied in the first stage, or by using the same t-test statistics from the first stage to pick the D most differentially expressed Gaussian features.

This implementation employs classifier independent feature selection schemes, such as the t-test and Shapiro-Wilk test. However, even for classifier dependent schemes, once the feature selection and classification schemes have been implemented, the Bayesian error estimator may be calculated as a deterministic function of the fixed classifier. This is in contrast to cross-validation, which uses surrogate classifiers to estimate the error of the designed classifier.

D. Estimating Prior Hyperparameters

When calibrating priors for microarrays, what data should be used and how? With the explosion of microarray experimentation over the last decade, the genomics community has amassed an enormous database of gene expression data, and trends in the entire history of microarray experimentation could be used to find a prior, perhaps conditioned on a particular organism, tissue, gene and/or type of abnormality, depending on the nature of the experiment at hand. However, different microarray experiments are currently very difficult to compare, although there have been some recent efforts to normalize and integrate different data sets [106].

The method employed here uses discarded gene expression data, consisting of a subset of the features from the microarray data that are not used for classification, to calibrate the priors of the Bayesian error estimator. Though these features are not used in the actual classifier, they may implicitly contain useful calibration information such as the varying concentrations of DNA material used in each microarray, background intensities and other characteristics of the digitized images of a microarray slide. And although calibration requires a large amount of data and in microarray gene expression analysis we typically expect a very small sample setting, the huge number of discarded features ensures that there is enough data for a successful calibration of the hyperparameters.

It is possible to define a prior on the entire feature set and to compute the Bayesian error estimator over the reduced feature set based on the marginal distribution of this prior on only the selected features. However, the following approach directly defines a prior on only the selected features.

Consider one class at a time, $y = 0$ or $y = 1$. To simplify notation, in this section we write μ instead of μ_y and Σ instead of Σ_y. We essentially use a method of moments approach to calibrate the hyperparameters; however, estimating a vector $\mathbf{m}$ and matrix S may be problematic for a small number of sample points, so to simplify the analysis we assume the following structure on these hyperparameters:

$$\mathbf{m} = m\,[1, 1, \ldots, 1]^T\,,$$

$$S = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix},$$

where m is a real number, $\sigma^2 \geq 0$, and $-1 \leq \rho \leq 1$. This structure is justified because prior to observing the data, there is no reason to think that any feature, or pair of features, should have distinct properties. With this simplification, our problem is now reduced to estimating five scalers for each class: ν, m, κ, σ^2 and ρ.

In the first stage of a method of moments approach, we find the theoretical first and second moments of the random variables μ and Σ (random because of the prior distribution applied to them) in terms of the hyperparameters we wish to estimate. Throughout the remainder of this section, a subscript i represents the ith element of a vector, and a subscript jk represents the jth row, kth column element of a matrix.

First consider the parameter Σ, with a marginal prior having an inverse-Wishart distribution with hyperparameters κ and S. The mean of this distribution is well known [108],

$$\mathrm{E}[\Sigma] = \frac{S}{\kappa - D - 1},$$

and given the previously defined structure on S, we obtain

$$\sigma^2 = (\kappa - D - 1)\mathrm{E}[\Sigma_{11}], \tag{7.1}$$

$$\rho = \frac{\mathrm{E}[\Sigma_{12}]}{\mathrm{E}[\Sigma_{11}]}. \tag{7.2}$$

Due to our imposed structure, only $\mathrm{E}[\Sigma_{11}]$ and $\mathrm{E}[\Sigma_{12}]$ are needed.

The variance of the jth diagonal element in inverse-Wishart distributed Σ may be expressed as

$$\mathrm{Var}\,(\Sigma_{jj}) = \frac{2(S_{jj})^2}{(\kappa - D - 1)^2(\kappa - D - 3)} = \frac{2(\mathrm{E}[\Sigma_{11}])^2}{\kappa - D - 3},$$

where we have applied Equation (7.1) in the second equality. Solving for κ,

$$\kappa = \frac{2(\mathrm{E}[\Sigma_{11}])^2}{\mathrm{Var}\,(\Sigma_{11})} + D + 3. \tag{7.3}$$

We next consider the mean, μ, which is parameterized by the hyperparameters ν and m. The marginal distribution of the mean is a multivariate Student's t-distribution given by [108]:

$$\pi(\mu) = \frac{\Gamma\left(\frac{\kappa+1}{2}\right)}{\Gamma\left(\frac{\kappa-D+1}{2}\right)} \sqrt{\frac{\nu^D}{\pi^D} \frac{|S|^{-1}}{(1 + \nu(\mu - \mathbf{m})^T S^{-1}(\mu - \mathbf{m}))^{\kappa+1}}}.$$

The mean and covariance of this distribution are well known:

$$\mathrm{E}[\mu] = \mathbf{m},$$

$$\mathrm{Var}\,(\mu) = \frac{S}{(\kappa - D - 1)\nu} = \frac{\mathrm{E}[\Sigma]}{\nu}.$$

With the assumed structure on $\mathbf{m}$, we obtain

$$m = \mathrm{E}[\mu_1], \tag{7.4}$$

$$\nu = \frac{\mathrm{E}[\Sigma_{11}]}{\mathrm{Var}\,(\mu_1)}. \tag{7.5}$$

Finally, our objective is to approximate the expectations in Equations (7.1) through (7.5) using calibration features left out of the classification scheme. Suppose the calibration data for the current class consists of n sample points with $E \gg D$ features. Let $\widehat{\mu}^E$ be the sample mean and $\widehat{\Sigma}^E$ be the sample covariance matrix of the complete set of E features in the calibration data. From these we wish to find several sample moments of μ and Σ in our original D feature problem, that is, to find $\widehat{\mathrm{E}}[\mu_1]$, $\widehat{\mathrm{Var}}\,(\mu_1)$, $\widehat{\mathrm{E}}[\Sigma_{11}]$, $\widehat{\mathrm{E}}[\Sigma_{12}]$ and $\widehat{\mathrm{Var}}\,(\Sigma_{11})$, where the hats indicate the sample moment of the corresponding quantity. All of these are scaler quantities.

To compress the set of E features in the calibration data to solve an estimation problem on just D features, and ultimately to find these scaler sample moments in a balanced way, we emulate the feature selection process by assuming the selected features are drawn uniformly. Since any of the E features is equally likely to be selected as the ith feature, the sample mean of the mean of the ith feature, $\widehat{\mathrm{E}}[\mu_i]$, is computed as the average of the sample means of all E features in the calibration data. This result is the same for all i, and we use $\widehat{\mathrm{E}}[\mu_1]$ to represent all features. In particular,

$$\widehat{\mathrm{E}}[\mu_1] = \frac{1}{E}\sum_{i=1}^{E} \widehat{\mu}_i^E. \tag{7.6}$$

Thanks to uniform feature selection, all other moments may be balanced over all features or any pair of distinct features. The remaining sample moments are obtained in a similar manner:

$$\widehat{\mathrm{Var}}\,(\mu_1) = \frac{1}{E-1}\sum_{i=1}^{E} \left(\widehat{\mu}_i^E - \widehat{\mathrm{E}}[\mu_1]\right)^2, \tag{7.7}$$

$$\widehat{\mathrm{E}}[\Sigma_{11}] = \frac{1}{E}\sum_{i=1}^{E} \widehat{\Sigma}_{ii}^E, \tag{7.8}$$

$$\widehat{\mathrm{E}}[\Sigma_{12}] = \frac{2}{E(E-1)} \sum_{i=2}^{E} \sum_{j=1}^{i-1} \widehat{\Sigma}_{ij}^{E}, \tag{7.9}$$

$$\widehat{\mathrm{Var}}\,(\Sigma_{11}) = \frac{1}{E-1} \sum_{i=1}^{E} \left(\widehat{\Sigma}_{ii}^{E} - \widehat{\mathrm{E}}[\Sigma_{11}] \right)^2 . \tag{7.10}$$

Here, $\widehat{\mathrm{Var}}\,(\mu_1)$ represents the variance of each feature in the mean. We also have $\widehat{\mathrm{E}}[\Sigma_{11}]$ and $\widehat{\mathrm{E}}[\Sigma_{12}]$ representing the sample mean of diagonal elements and off-diagonal elements in Σ, respectively. Finally, $\widehat{\mathrm{Var}}\,(\Sigma_{11})$ is the sample variance of the diagonal elements in Σ.

Plugging our sample moments into Equations (7.1) through (7.5), we obtain

$$\sigma^2 = 2\widehat{\mathrm{E}}[\Sigma_{11}] \left(\frac{(\widehat{\mathrm{E}}[\Sigma_{11}])^2}{\widehat{\mathrm{Var}}\,(\Sigma_{11})} + 1 \right), \tag{7.11}$$

$$\rho = \frac{\widehat{\mathrm{E}}[\Sigma_{12}]}{\widehat{\mathrm{E}}[\Sigma_{11}]}, \tag{7.12}$$

$$\kappa = \frac{2(\widehat{\mathrm{E}}[\Sigma_{11}])^2}{\widehat{\mathrm{Var}}\,(\Sigma_{11})} + D + 3, \tag{7.13}$$

$$m = \widehat{\mathrm{E}}[\mu_1], \tag{7.14}$$

$$\nu = \frac{\widehat{\mathrm{E}}[\Sigma_{11}]}{\widehat{\mathrm{Var}}\,(\mu_1)}. \tag{7.15}$$

Note Equation (7.3) for κ was plugged into Equation (7.1) to obtain the final σ^2.

In sum, calibration for the prior hyperparameters is defined by Equations (7.11) through (7.15), the sample moments being given in Equations (7.6) through (7.10). The estimates of κ and ν can be unstable, since they rely on second moments, $\widehat{\mathrm{Var}}\,(\Sigma_{11})$ and $\widehat{\mathrm{Var}}\,(\mu_1)$, in a denominator. These parameters can be made more stable by discarding outliers when computing the sample moments. Herein, we discard the 10% of the $\widehat{\mu}_i^E$ with largest magnitude and the 10% of the $\widehat{\Sigma}_{ii}^E$ with largest value.

This method is one of many possible approaches; for simplicity and to avoid an over-defined system of equations, we do not incorporate the covariance between distinct features in μ (that is $\mathrm{Cov}\,(\mu)_{12}$), the variance of off-diagonal elements in Σ (that is $\mathrm{Var}\,(\Sigma_{12})$), or the correlation between distinct elements in Σ, though it may be possible to use these to improve the estimates of the hyperparameters. It may also be feasible to use other estimation methods, such as maximum likelihood. Furthermore, the method proposed here to calibrate the priors is a purely data driven technique for easy and general application to microarray experiments. Ideally, the best way to calibrate priors would be to incorporate data and biological knowledge specific to the particular features selected for classification.

E. Performance

We present two sets of results demonstrating good performance of Bayesian error estimators, one on synthetic high dimensional data with three-stage feature selection and a second based on breast cancer data with two stages of feature selection.

1. Gaussian Model with High-dimensional Synthetic Data

In this section, we apply our Bayesian prior estimation method to synthetic high-dimensional microarray data. We use the same synthetic data model provided in [109], which models many observations made in microarray expression based studies, including blocked covariance matrices to model groups of interacting variables with negligible interactions between groups.

Our model emulates a full feature-label distribution with 20,000 total features. Features are categorized as either “markers” or “non-markers.” Markers represent features that have different class-conditional distributions in the two classes and are further divided into two subtypes: global markers and heterogeneous markers. Non-markers have the same distributions for both classes and thus have no discriminatory

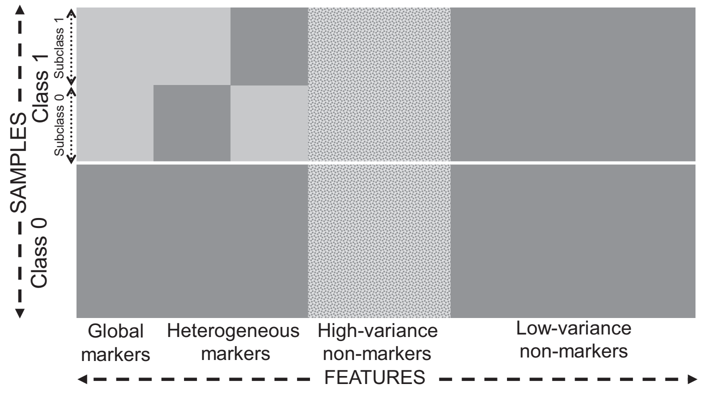


Fig. 37. Different feature types in constructing the high-dimensional synthetic data model.

power, and are also divided into two subtypes: high-variance non-markers and low-variance non-markers. A summary of the feature types is shown in Figure 37.

Twenty features are global markers, which are homogeneous in each class. In particular, the set of all global markers in class y has a Gaussian distribution with mean μ_y^{gm} and covariance matrix Σ_y^{gm}.

Within class 1, we assume each sample point belongs to one of two equally likely subclasses named 0 and 1, representing different stages or subtypes of cancer. Each subclass is associated with fifty heterogeneous markers, which are jointly Gaussian with mean μ_1^{hm} and covariance Σ_1^{hm}. Sample points associated with the other subclass have the same distribution as class 0, which is Gaussian with mean μ_0^{hm} and covariance Σ_0^{hm}. Each heterogeneous marker may only be associated with one subclass, thus there are 100 total heterogeneous markers in the model.

We simplify the model by assuming μ_y^{gm} and μ_y^{hm} have the form $m_y \times (1, 1, \ldots, 1)$ for fixed scalers m_y. We assume Σ_y^{gm} and Σ_y^{hm} have the form $\sigma_y^2 \Sigma$, where σ_y^2 are constants and Σ has a block covariance structure, i.e.,

$$\Sigma = \begin{bmatrix} \Sigma_\rho & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_\rho \end{bmatrix},$$

with Σ_ρ being a 5×5 matrix with 1 on the diagonal and $\rho = 0.8$ off the diagonal. That is, we group markers into blocks of 5 features, where the blocks are independent from each other, and the markers within each block are correlated with a relatively high correlation coefficient to emulate a pathway.

We generate 2,000 high-variance non-marker features, which have independent mixed Gaussian distributions given by $pN(m_0, \sigma_0^2) + (1-p)N(m_1, \sigma_1^2)$, where m_i and σ_i^2 are the same scalers defined for markers and $N(m_i, \sigma_i^2)$ is a normal random variable with mean m_i and variance σ_i^2. The random variable p is selected independently for each feature with a uniform distribution over $[0, 1]$ and is applied to all sample points of both classes for the given feature. These features can be viewed as genes regulated by mechanisms unrelated to those that regulate the class-0 and class-1 phenotypes. The remaining features are low-variance non-marker features, each having independent univariate Gaussian distributions with mean m_0 and variance σ_0^2.

In this model, heterogeneous markers are Gaussian within each sub-class, but the class-conditional distribution for class 1 is a mixed Gaussian distribution (mixing the distributions of the sub-classes), and is thus not Gaussian. Further, the high-variance features are also mixed Gaussian distributions, so this model incorporates both Gaussian and non-Gaussian features to challenge the Shapiro-Wilk Gaussianity test in the feature selection scheme.

To simplify our simulations, we set the *a priori* probability of both classes to 0.5 and fix the parameters $m_0 = 0$ and $m_1 = 1$. We also define a single parameter

Table 7. Synthetic high-dimensional data model parameters

Parameters	Values/description
Total features	20,000
Global markers	20
Subclasses in class 1	2
Heterogeneous markers	50 per subclass (100 total)
High-variance features	2,000
Low-variance features	17,880
Mean	$m_0 = 0$, $m_1 = 1$
Variances	$\sigma^2 = \sigma_0^2 = \sigma_1^2$ (controls Bayes error)
Block size	5
Block correlation	0.8
a priori prob. of class 0	0.5

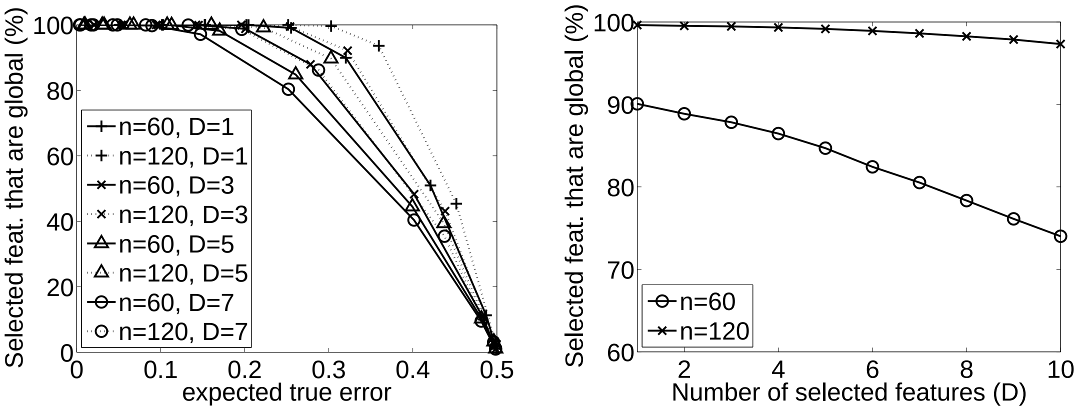
$\sigma^2 = \sigma_0^2 = \sigma_1^2$, which specifies the difficulty of the classification problem. A summary of our synthetic high-dimensional data model parameters is given in Table 7. In all simulations, the values for σ^2 are chosen so that a single global feature (note that all global features are identical) has a specific Bayes error. We call this the "Bayes error" in the remainder of this section, and it is given by $\varepsilon^* = \Phi\left(-1/(2\sigma)\right)$, where Φ is the unit normal Gaussian cumulative distribution function, so for instance, we use $\sigma = 0.9537$ for a Bayes error of 0.3.

Under this high-dimensional model, we run several Monte-Carlo simulations. In each experiment we fix the training sample size, n, the number of selected features, D, and the difficulty of the classification problem via σ. The synthetically generated samples are non-stratified, meaning that in each iteration the sample size of each class is not fixed but determined by a binomial$(0.5, n)$ experiment, and the corresponding sample points are randomly generated according to the distributions defined for each class.

Once the sample has been generated, we apply the three-stage feature selection scheme outlined in Section VII.C. In the first stage, we apply a t-test to obtain 1,000 highly differentially expressed features by removing most non-informative features. In the second stage, we apply a Shapiro-Wilk Gaussianity test and eliminate features that do not pass the test with 95% confidence. The number of features output in this stage is variable. If there are not at least thirty features that pass the test, then we return the thirty features with the highest sum of the Shapiro-Wilk test statistics for both classes. In the final stage, we use the same t-test values computed before to obtain the final set of D highly differentially expressed Gaussian features, which will be used to design our classifier. The $1,000 - D$ features that pass the first stage of feature selection but are not used for classification are saved as calibration data.

The feature selected training data are then used to train an LDA classifier. With the classifier fixed, 5,000 testing points are drawn from exactly the same distribution as the training data and used expressly to approximate the true error. Subsequently, several training-data error estimators are computed, including leave-one-out (loo), 5-fold cross-validation (cv), 0.632 bootstrap (boot), and bolstered resubstitution (bol). Two Bayesian error estimators are also applied, one with flat non-informative priors defined by $\pi\left(\theta_y\right) = 1$ (the flat Bayesian error estimator), and the other with priors calibrated as described in Section VII.D (the calibrated Bayesian error estimator). Since the classifier is linear, these Bayesian error estimators are computed exactly. This entire process is repeated 120,000 times to approximate the RMS deviations from the true error for each error estimator.

We first analyze the quality of features selected by the three-stage feature selection algorithm. Figure 38(a) shows the percentage of selected features that are

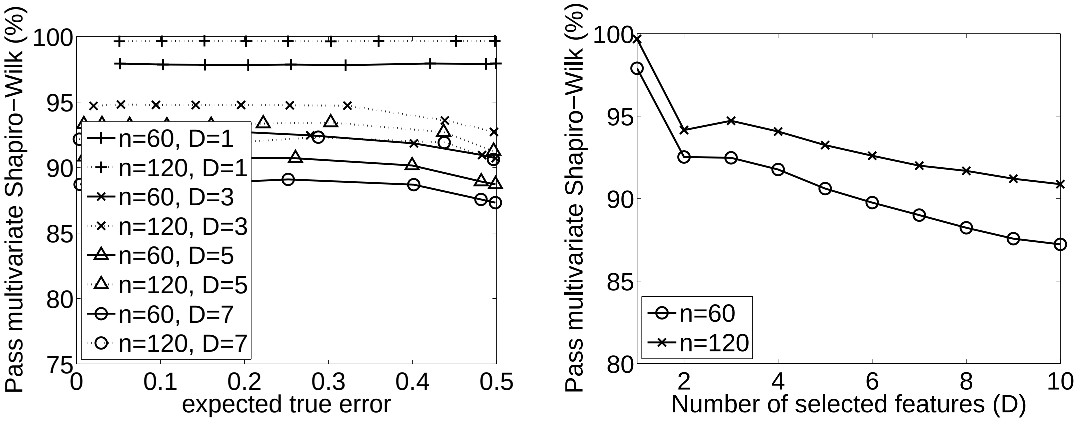


(a) vs. expected true error

(b) vs. feature size, Bayes error = 0.3

Fig. 38. Percentage of three-stage selected features that are global features in the synthetic high-dimensional data model.

global features with respect to the expected true error of the designed classifier. We would like to graph performance with respect to Bayes error, which is a more pure measure of the difficulty of a classification problem, but evaluating Bayes error on our high-dimensional model is difficult and it may not be close to the true error of the designed classifier. Hence, in our graphs we focus on performance with respect to expected true error. Similarly, Figure 38(b) graphs against feature size with a fixed Bayes error of 0.3. Recall that this model uses 20,000 features, of which only 20 are global features that most effectively discriminate the classes. As long as the feature size is reasonable given the difficulty of the problem (expected true error and sample size), this percentage is quite large. However, in Figure 38(b) for sample size 60 we see that a feature size larger than 7 will result in less than 80% of the selected features being global features. This illustrates the necessity of restricting feature size in a small sample setting, and is consistent with earlier studies showing the difficulty of finding good feature sets when the number of features is large and the sample is



(a) vs. expected true error

(b) vs. feature size, Bayes error = 0.3

Fig. 39. Percentage of three-stage selected features that are not rejected by a multivariate Shapiro-Wilk test on either class at a 95% significance level with the synthetic high-dimensional data model.

small [110, 111].

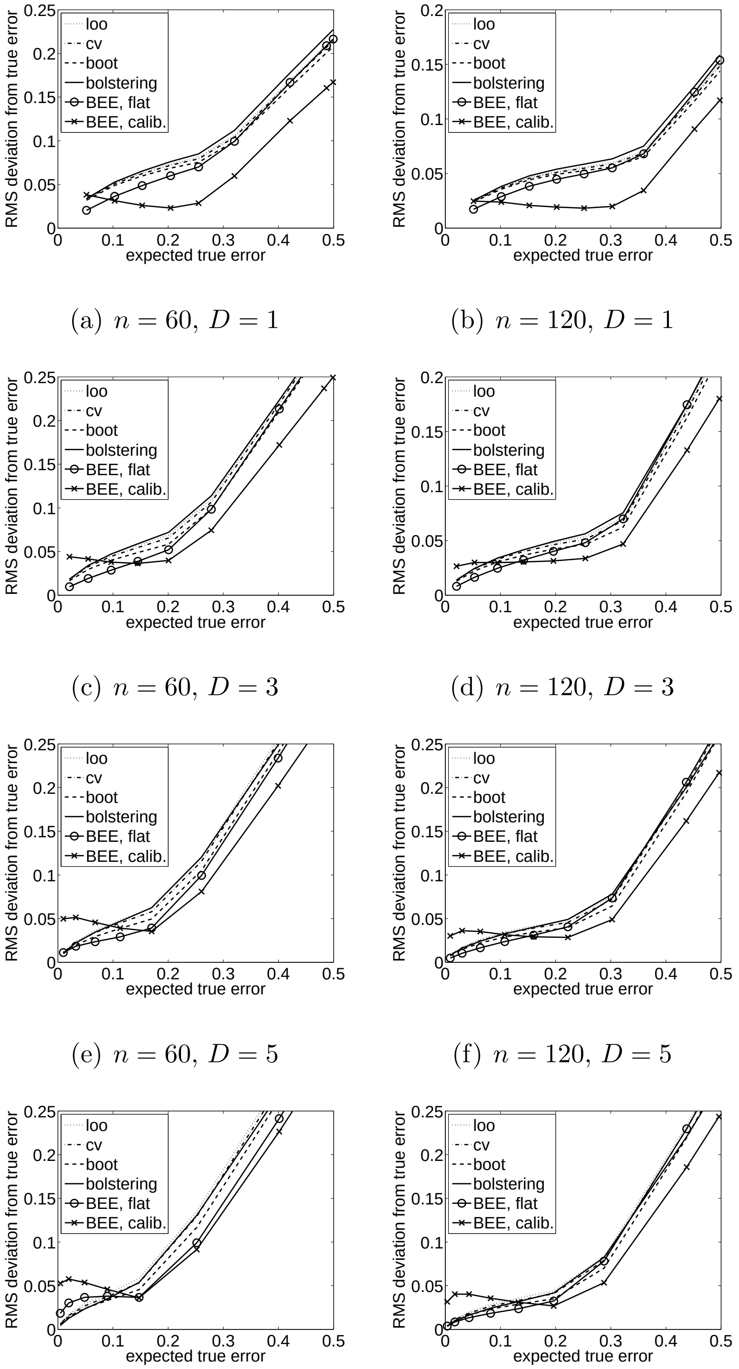
The graphs in Figure 39 show the percentage of selected feature sets that are not rejected by a multivariate Shapiro-Wilk test on either class at a 95% significance level. There are several multivariate Gaussianity tests based on the Shapiro-Wilk statistic. We used [112], which generalizes the classical univariate Shapiro-Wilk test to the multivariate case by transforming the data into a set of approximately independent standard normal random variables, and essentially summing up the standard Shapiro-Wilk statistic on each dimension. The results show that even though the three-stage feature selection algorithm only uses a univariate Gaussianity test, and univariate normality does not imply multivariate normality, the resulting feature set still tends to have a high probability of passing the multivariate Gaussianity test.

We next turn our attention to the RMS performance of error estimators under our synthetic high-dimensional model, where a summary of all simulation settings are available in Table 8. Our first battery of simulations in Figure 40 shows RMS deviation from true error for all error estimators with respect to expected true error for LDA classification with 1, 3, 5, or 7 selected features and either 60 or 120 sample points. Given the sample sizes, it is prudent to keep the number of selected features small to have satisfactory feature selection [110] and to avoid the peaking phenomena [113, 109]. Lines marked with 'o' represent the Bayesian error estimator with flat priors, and lines marked with 'x' represent the Bayesian error estimator with the calibrated priors. The key point in these graphs is that the calibrated Bayesian error estimator has best performance in the mid and high range. For an expected true error of about 0.25 and $n = 60$, the RMS for the calibrated Bayesian error estimator outperforms 5-fold cross-validation for $D = 1, 3, 5$ and 7 by 0.0507, 0.0300, 0.0335, and 0.0379, respectively, representing 64, 32, 30, and 29 percent decrease in RMS, respectively. For $n = 120$, the decrease in RMS for $D = 1, 3, 5$ and 7 is 0.0366, 0.0175, 0.0192, and 0.0198, respectively, for 67, 34, 35, and 33 percent decrease in RMS, respectively. All other error estimators typically have best performance for low expected true errors, with the flat Bayesian error estimator having even better performance than the classical error estimation schemes. Indeed, all graphs except Figure 40(g) demonstrate that either the flat or calibrated Bayesian error estimator is the best scheme over the whole range of expected true error.

Our next set of graphs in Figure 41 show simulation results with respect to feature size. For reference, graphs of the expected true error for these simulations are shown in Figure 42. Calibrated priors provide the best performance, except when combining large feature and small sample sizes, in which case a flat prior performs best. In fact, performance of the calibrated Bayesian error estimator in Figure 41 tends to be best precisely in the rage of feature sizes with the highest percentage of global features and the lowest true errors. For example, the calibrated Bayesian

Table 8. Data model and classification settings for simulations with synthetic high-dimensional data

Data model	Classifier	Sample size		Feature selection				Calibration	Iteration
Bayes error		Training	Test	Initial	1^{st} t-test	Shapiro-Wilk	2^{nd} t-test		
0.05 to 0.45	LDA	$n = 60$	5000	20000	1000	95% confidence	$D = 1$	$1000 - D$	120000
0.05 to 0.45	LDA	$n = 60$	5000	20000	1000	95% confidence	$D = 3$	$1000 - D$	120000
0.05 to 0.45	LDA	$n = 60$	5000	20000	1000	95% confidence	$D = 5$	$1000 - D$	120000
0.05 to 0.45	LDA	$n = 60$	5000	20000	1000	95% confidence	$D = 7$	$1000 - D$	120000
0.05 to 0.45	LDA	$n = 120$	5000	20000	1000	95% confidence	$D = 1$	$1000 - D$	120000
0.05 to 0.45	LDA	$n = 120$	5000	20000	1000	95% confidence	$D = 3$	$1000 - D$	120000
0.05 to 0.45	LDA	$n = 120$	5000	20000	1000	95% confidence	$D = 5$	$1000 - D$	120000
0.05 to 0.45	LDA	$n = 120$	5000	20000	1000	95% confidence	$D = 7$	$1000 - D$	120000
0.3	LDA	$n = 60$	5000	20000	1000	95% confidence	1 to 10	$1000 - D$	120000
0.3	LDA	$n = 120$	5000	20000	1000	95% confidence	1 to 10	$1000 - D$	120000



(a) $n = 60$, $D = 1$

(b) $n = 120$, $D = 1$

(c) $n = 60$, $D = 3$

(d) $n = 120$, $D = 3$

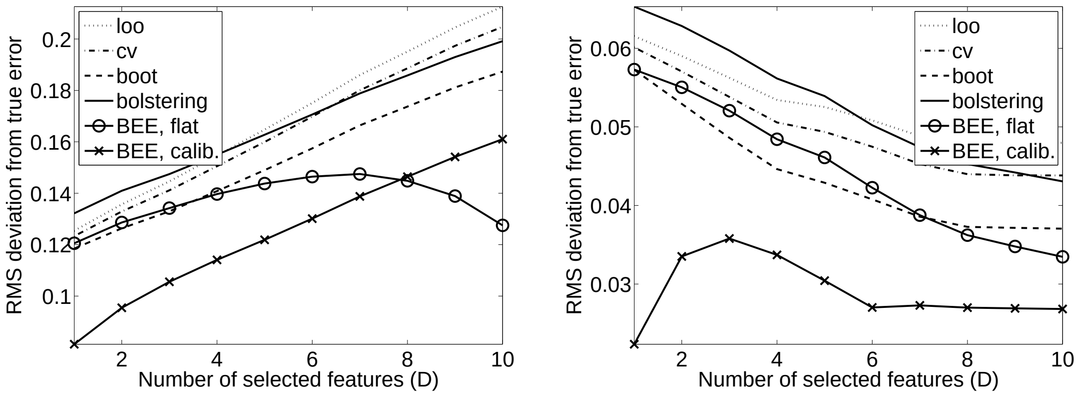
(e) $n = 60$, $D = 5$

(f) $n = 120$, $D = 5$

(g) $n = 60$, $D = 7$

(h) $n = 120$, $D = 7$

Fig. 40. RMS deviation from true error for the synthetic high-dimensional data model with LDA classification versus expected true error.



(a) $n = 60$, Bayes error = 0.3

(b) $n = 120$, Bayes error = 0.3

Fig. 41. RMS deviation from true error for the synthetic high-dimensional data model with LDA classification versus feature size.

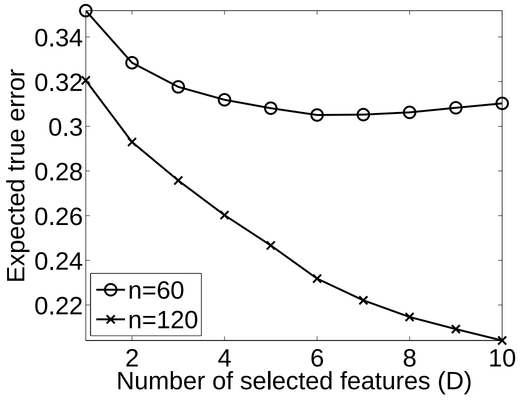


Fig. 42. Expected true error for the synthetic high-dimensional data model with LDA classification versus feature size, Bayes error = 0.3.

error estimator in Figure 41(a) for sample of size 60 has the best performance up to 7 features, where in Figure 38(b) the percentage of selected features being global is greater than about 80% and in Figure 42 the true error has started to level off. Note, also, the consistently superior performance of the calibrated Bayesian error estimator over the non-Bayesian estimators for $n = 60$; indeed, throughout the range of feature sizes, the calibrated Bayesian error estimator has an RMS at least 0.0263 smaller than the best performing non-Bayesian error estimator, which represents an improvement of at least 14 percent.

Note the upward RMS trend in Figure 41(a) and the downward trend in Figure 41(b) for the non-Bayesian error estimators. Although it can be dangerous to generalize about the behavior of error estimators, let us at least conjecture. We see in Figure 42 that the true error is large for $n = 60$, with little improvement as we increase the number of features and, in fact, increasing true error as the number of features passes 7, which is a clear sign of the peaking phenomenon. Thus, for $n = 60$, adding features creates a more difficult estimation problem that is not offset by easing error estimation on account of small true errors. On the other hand, in Figure 42 we see a fast reduction of true error for $n = 120$ as more features are added, thereby greatly easing the error estimation problem and resulting in the declining RMS trend in Figure 41(b). While these comments apply directly to the non-Bayesian error estimators they apply to the Bayesian estimators relative to their change of slope. The flat Bayesian error estimator is relatively constant in Figure 41(a) but falls along with the non-Bayesian error estimators in Figure 41(b), whereas the calibrated Bayesian error estimator consistently rises in Figure 41(a) but remains relatively flat in Figure 41(b).

2. Gaussian Model Applied to Real Breast Cancer Data

We next apply Bayesian error estimation to the normalized gene-expression measurements from the same breast cancer study [99] used in Section IV.B.4. This study used 295 sample points, with 180 assigned to class 0 (good prognosis) and 115 in class 1 (bad prognosis), and provides a 70 feature prognosis profile. From the original 295 points, we randomly draw a non-stratified training sample of size n. Since the number of features in the data set is relatively small, we apply only the last two stages of our feature selection scheme in Section VII.C. The first stage selects features passing a Shapiro-Wilk Gaussianity test with 95% confidence and must report at least D features, while the second stage selects D features with the highest t-test statistic. The $70 - D$ features not used for classification are retained as calibration data for Bayesian error estimation. After feature selection, we train an LDA, QDA or 3NN classifier.

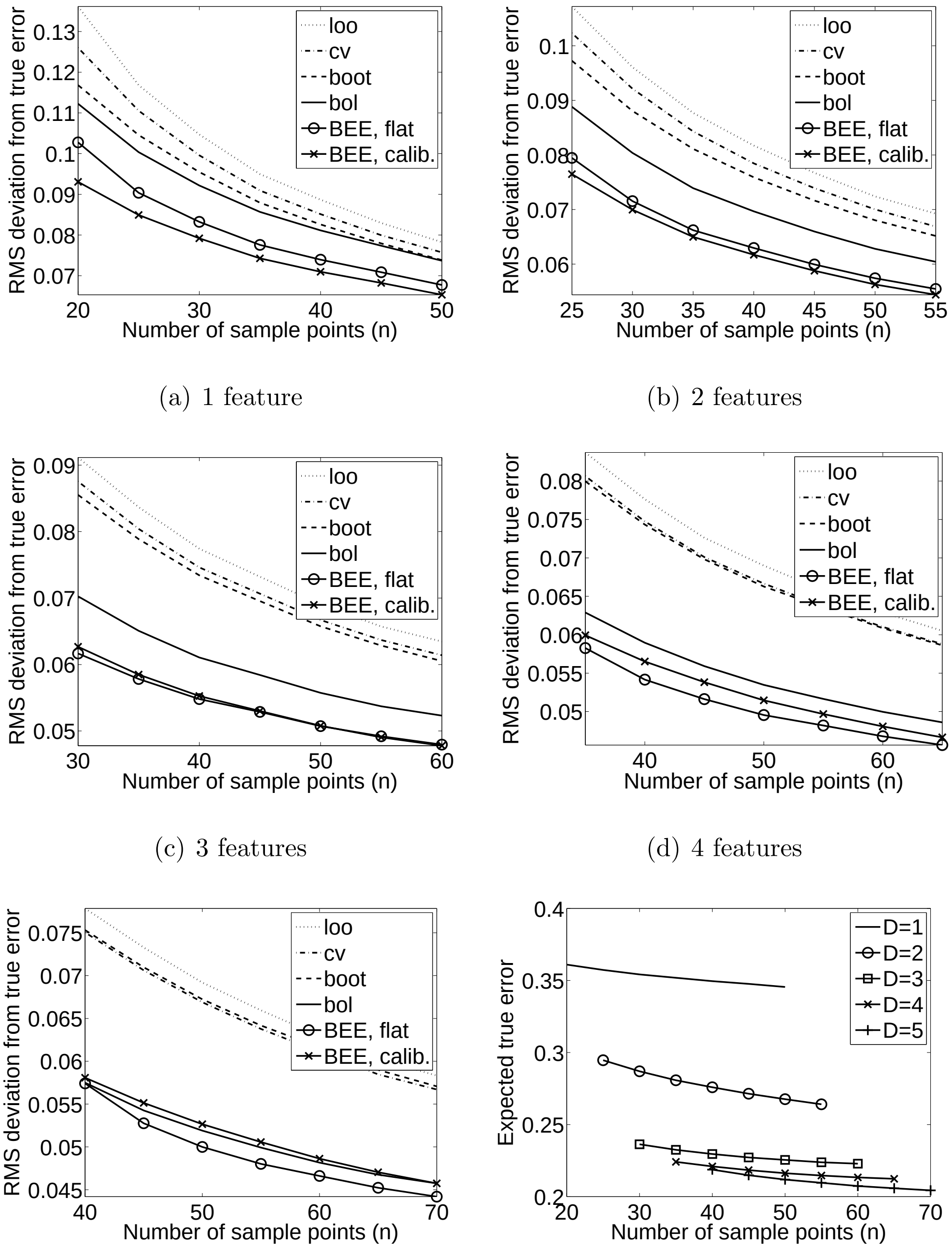
The remaining sample points are used as holdout data to approximate the true error of the designed classifier. The previously considered error estimators are also evaluated from the training samples (except in the case of 3NN where semi-bolstering is used instead of bolstering owing to its superior performance for 3NN [71]), along with exact Bayesian error estimators (for LDA) or approximate Bayesian error estimators (for QDA and 3NN). Both flat and calibrated priors are applied. This process is repeated either 100,000 times (for LDA) or 10,000 times (for QDA and 3NN) to estimate the average RMS deviation of each error estimator from the true error.

The priors are calibrated as discussed in Section VII.D. A typical prior with 2 features and 40 sample points is $\nu = 16.80$, $m = -0.004$, $\kappa = 12$, $\sigma^2/(\kappa - D - 1) = 0.042$ and $\rho = 0.020$ for class 0, and $\nu = 2.78$, $m = -0.068$, $\kappa = 10$, $\sigma^2/(\kappa - D - 1) = 0.024$ and $\rho = 0.073$ for class 1. These indicate that the good-prognosis class (0) has a distribution with a more concentrated mean (since ν is much larger) and the mean is close to 0, which is expected since the data has been normalized. On the other hand, κ is fairly large for both classes, suggesting that the variance of each feature in either class is probably close to the prior expected variance, $\sigma^2/(\kappa - D - 1)$. Interestingly, the variance is a bit larger for class 0 and ρ is usually small but positive.

Figures 43, 44 and 45 provide simulation results for LDA, QDA and 3NN, respectively. Each figure contains subplots representing fixed feature sizes between one and five, and one figure showing the expected true error for all simulations with the corresponding classifier. A summary of the simulation settings is shown in Table 9. The uniform prior performs well over a wide range of sample and feature sizes, and generally shows significant improvement over the classical error estimators. Prior calibration can have even more pronounced improvement, especially for small feature sets. And although the uniform prior often performs better than the calibrated prior for high feature sizes, see for example Figure 43(e) for 5 features, we observe in Figure 43(f) that true error does not improve much, and may actually get worse, for as little as 5 features. This may indicate that when there is not enough calibration data for good prior design, there is also insufficient data for good classifier design.

F. Discussion

Our synthetic data simulations demonstrate the power of prior knowledge in two ways: we may assume a low Bayes error by using a flat prior and outperform the classical error estimators where they perform best, or we may calibrate a prior, even using purely data driven methods, and obtain superior performance in the mid range of Bayes errors. Also note that for moderately difficult classification problems which are typical in a small sample biological setting, the mid range is precisely where



(a) 1 feature

(b) 2 features

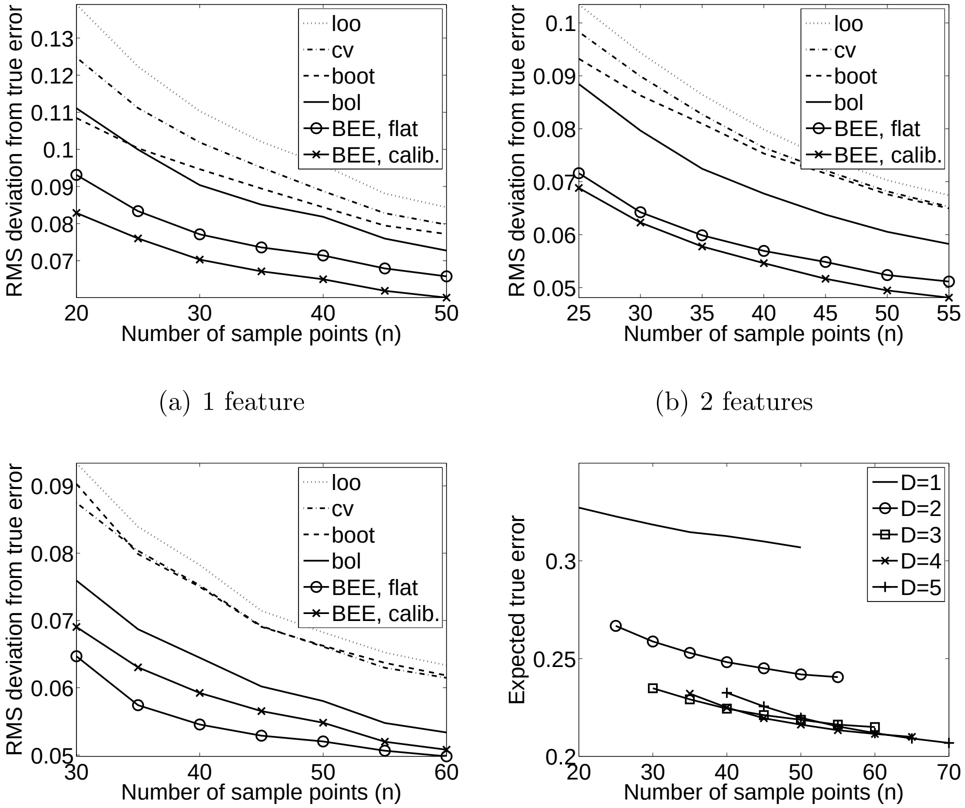
(c) 3 features

(d) 4 features

(e) 5 features

(f) expected true error

Fig. 43. RMS deviation from true error and expected true error with LDA classification of empirical measurements from a breast cancer study.



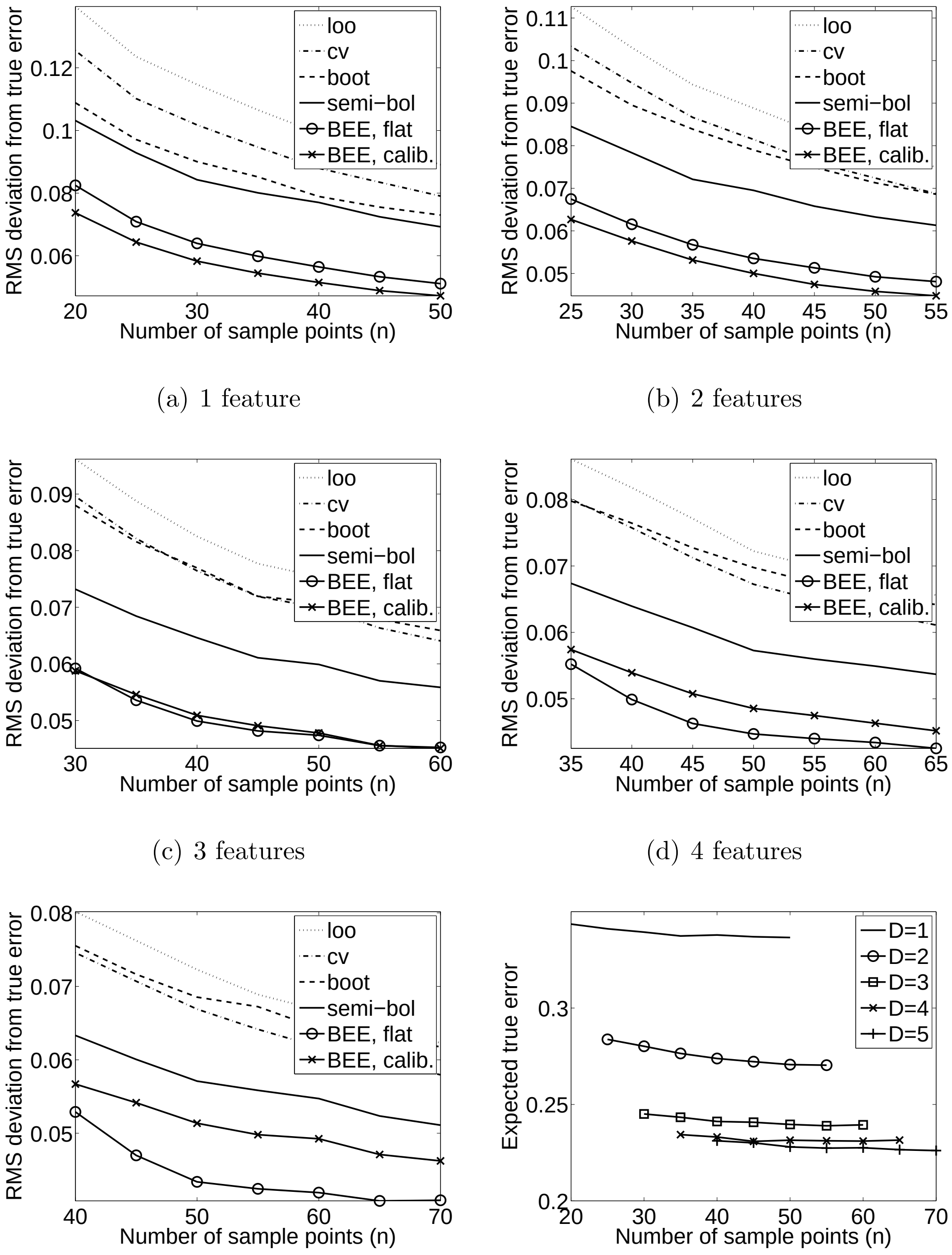
(a) 1 feature

(b) 2 features

(c) 3 features

(d) expected true error

Fig. 44. RMS deviation from true error and expected true error with QDA classification of empirical measurements from a breast cancer study.



(a) 1 feature

(b) 2 features

(c) 3 features

(d) 4 features

(e) 5 features

(f) expected true error

Fig. 45. RMS deviation from true error and expected true error with 3NN classification of empirical measurements from a breast cancer study.

Table 9. Classification schemes and settings for simulations with real high-dimensional breast cancer data

Classifier	Sample size		Feature selection			Calibration	Iteration
	Training	Test	Initial	Shapiro-Wilk	2^{nd} t-test		
LDA	20 to 50	$295 - n$	70	95% confidence	$D = 1$	$70 - D$	100000
LDA	25 to 55	$295 - n$	70	95% confidence	$D = 2$	$70 - D$	100000
LDA	30 to 60	$295 - n$	70	95% confidence	$D = 3$	$70 - D$	100000
LDA	35 to 65	$295 - n$	70	95% confidence	$D = 4$	$70 - D$	100000
LDA	40 to 70	$295 - n$	70	95% confidence	$D = 5$	$70 - D$	100000
QDA	20 to 50	$295 - n$	70	95% confidence	$D = 1$	$70 - D$	10000
QDA	25 to 55	$295 - n$	70	95% confidence	$D = 2$	$70 - D$	10000
QDA	30 to 60	$295 - n$	70	95% confidence	$D = 3$	$70 - D$	10000
3NN	20 to 50	$295 - n$	70	95% confidence	$D = 1$	$70 - D$	10000
3NN	25 to 55	$295 - n$	70	95% confidence	$D = 2$	$70 - D$	10000
3NN	30 to 60	$295 - n$	70	95% confidence	$D = 3$	$70 - D$	10000
3NN	35 to 65	$295 - n$	70	95% confidence	$D = 4$	$70 - D$	10000
3NN	40 to 70	$295 - n$	70	95% confidence	$D = 5$	$70 - D$	10000

training data error estimation is needed. One might argue that there is a risk with postulating a low-Bayes-error prior since, although it will show excellent performance if the Bayes error is truly low, it will suffer for large Bayes errors. In Figure 40, not only does performance deteriorate with increasing Bayes error for the Bayesian MMSE estimator, so too does the performance of cross-validation. This should not be surprising because the use of cross-validation presupposes that the Bayes error is small because its performance seriously degrades for increasing Bayes error. This behavior, noted more than 30 years ago in a simple 1-dimensional Gaussian model [65], has been demonstrated via large-simulations for both the discrete and Gaussian models, and has been analytically proven in the Gaussian model [25]. In other words, unless one is not interested in error estimator performance, use of cross-validation carries with it implicitly assumed prior knowledge. If one knows that the Bayes error is low, then why not define a prior model based on this assumption to design a Bayesian error estimator with even better performance?

CHAPTER VIII

BAYESIAN MMSE CALIBRATION OF CLASSIFIER ERROR ESTIMATORS AND CONCLUSION*

When it is reasonable to assume a Bayesian framework but an analytical or closed-form Bayesian error estimator is not available, it may be approximated using Monte-Carlo methods. That being said, approximating a Bayesian error estimator is much more computationally intensive than classical counting methods and may be infeasible. To address this, we propose a new method of optimally calibrating arbitrary error estimators within Bayesian frameworks. Assuming a fixed sample size, fixed classification and error estimation schemes, and a set of priors for the distribution parameters, this is done in two steps. First, we compute a calibration function mapping error estimates (from the specified error estimation rule) to their calibrated values off-line according to the assumed model. Second, in all future experiments a practitioner may perform classification and error estimation in the usual way, but at the last step use the calibration function as a simple lookup table to calibrate the final error estimate on the fly.

The calibration function is defined to be the MMSE estimate of the true error of a classifier designed from the assumed classification scheme, given an observed error estimate. Equivalently, this is the expected true error conditioned on the observed error estimate, where uncertainty in the expectation stems from our uncertainty in both the feature-label distribution and the sample. This is similar to Bayesian MMSE error estimation itself, which is equivalent to the expected true error of a designed classifier conditioned on the entire observed sample, except that the calibrated error

*Reprinted from *Pattern Recognition*, vol. 45, no. 6, L. A. Dalton and E. R. Dougherty, “Optimal MSE Calibration of Classifier Error Estimators Under Bayesian Models,” pp. 2308–2320, Copyright 2012, with permission from Elsevier.

estimator conditions on only the observed error estimate. In other words, both error estimators minimize MSE in the same assumed Bayesian model, but the Bayesian error estimator has the benefit of the entire sample, which is an array of n sample points with D features each, and the MMSE calibrated error estimator uses only a single statistic (a lossy function of the observed sample) containing information about the true error. Also, a basic property of both Bayesian and calibrated error estimators is that they are unbiased relative to the true error. However, since the MMSE calibrated error estimate averages true errors over all samples producing the observed error estimate, the sample and classifier are not fixed as they are in Bayesian error estimation, where conditioning is on the sample itself.

A. Optimal Calibration of Arbitrary Error Estimators

An optimal calibration function is associated with four assumptions: a fixed sample size n, a Bayesian model with a proper prior $\pi\left(\theta\right) = \pi\left(c\right)\pi\left(\theta_0\right)\pi\left(\theta_1\right)$, a fixed classification rule (including possibly a feature selection scheme), and a fixed (uncalibrated) error estimation rule with estimates denoted by $\widehat{\varepsilon}_{\mathrm{UEE}}$. Given these assumptions, the optimal MMSE calibration function is the expected true error conditioned on the observed error estimate,

$$\begin{aligned} \mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{UEE}}] &= \int_0^1 \varepsilon_n f\left(\varepsilon_n|\widehat{\varepsilon}_{\mathrm{UEE}}\right) d\varepsilon_n \\ &= \frac{\int_0^1 \varepsilon_n f\left(\varepsilon_n, \widehat{\varepsilon}_{\mathrm{UEE}}\right) d\varepsilon_n}{f\left(\widehat{\varepsilon}_{\mathrm{UEE}}\right)}, \end{aligned} \tag{8.1}$$

where $f\left(\varepsilon_n, \widehat{\varepsilon}_{\mathrm{UEE}}\right)$ is the unconditional joint density between the true and estimated errors and $f\left(\widehat{\varepsilon}_{\mathrm{UEE}}\right)$ is the unconditional marginal density of the estimated error. Viewed as a function of $\widehat{\varepsilon}_{\mathrm{UEE}}$, this expectation is called the "MMSE calibration function." It may be used to calibrate any error estimator to have optimal MSE performance for the assumed model. Evaluated at a particular value of $\widehat{\varepsilon}_{\text{UEE}}$, it is called the "MMSE calibrated error estimate" and will be denoted by $\widehat{\varepsilon}_{\text{CEE}}$. As noted in the Introduction, calibrated error estimators are unbiased. To wit, according to a basic property of conditional expectation,

$$\mathrm{E}[\widehat{\varepsilon}_{\text{CEE}}] = \mathrm{E}[\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\text{UEE}}]] = \mathrm{E}[\varepsilon_n].$$

If an analytical representation for the joint density between true and estimated errors for fixed distributions, $f\left(\varepsilon_n, \widehat{\varepsilon}_{\text{UEE}}|\theta\right)$, is available, then

$$f\left(\varepsilon_n, \widehat{\varepsilon}_{\text{UEE}}\right) = \int_{\mathbf{\Theta}} f\left(\varepsilon_n, \widehat{\varepsilon}_{\text{UEE}}|\theta\right) \pi\left(\theta\right) d\theta, \tag{8.2}$$

where $\mathbf{\Theta}$ is the parameter space of θ. $f\left(\widehat{\varepsilon}_{\text{UEE}}\right)$ may either be found directly from $f\left(\varepsilon_n, \widehat{\varepsilon}_{\text{UEE}}\right)$ or from analytical representations of $f\left(\widehat{\varepsilon}_{\text{UEE}}|\theta\right)$ via

$$f\left(\widehat{\varepsilon}_{\text{UEE}}\right) = \int_{\mathbf{\Theta}} f\left(\widehat{\varepsilon}_{\text{UEE}}|\theta\right) \pi\left(\theta\right) d\theta. \tag{8.3}$$

From (8.2), it is clear that $f\left(\varepsilon_n, \widehat{\varepsilon}_{\text{UEE}}\right)$ utilizes all of our modeling assumptions, including the classification rule (because different classifiers will have different true errors), the error estimation rule, and the Bayesian prior.

If analytical results for $f\left(\varepsilon_n, \widehat{\varepsilon}_{\text{UEE}}|\theta\right)$ and $f\left(\widehat{\varepsilon}_{\text{UEE}}|\theta\right)$ are not available, then $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\text{UEE}}]$ may be found via Monte-Carlo approximation by simulating the model and classification procedure to generate a large collection of true and estimated error pairs. The MMSE calibration function may then be approximated by either estimating the joint density $f\left(\varepsilon_n, \widehat{\varepsilon}_{\text{UEE}}\right)$ or by simply partitioning error estimates into bins and then finding the corresponding average true error for estimated errors falling in each bin. An example is discussed using synthetic data in Section VIII.C.

Even though calibrated error estimation is suboptimal compared to Bayesian error estimation, it has several practical advantages:

1. Given the four necessary assumptions with any classification/error estimation architecture, a calibration function may be found off-line with straightforward Monte-Carlo approximation.

2. Analytical solutions may be derived using independent theoretical work on representations for $f\left(\varepsilon_n, \widehat{\varepsilon}_{\mathrm{UEE}}|\theta\right)$ and $f\left(\widehat{\varepsilon}_{\mathrm{UEE}}|\theta\right)$.

3. Once a calibration function has been established, it may be applied by post-processing a final error estimate with a simple lookup table.

B. On Ideal Regression

Given an arbitrary error estimation rule, $\widehat{\varepsilon}$, the non-linear regression between the true and estimated errors is represented by $g(\widehat{\varepsilon}) = \mathrm{E}[\varepsilon_n|\widehat{\varepsilon}\,]$. If $\widehat{\varepsilon} = \widehat{\varepsilon}_{\mathrm{UEE}}$ is a basic error estimate, then g is the calibration function mapping error estimates to their calibrated values. We say that an error estimator has "ideal regression" if $g(\widehat{\varepsilon}) = \mathrm{E}[\varepsilon_n|\widehat{\varepsilon}\,] = \widehat{\varepsilon}$ (almost surely).

In this section, we prove that both calibrated and Bayesian error estimators have ideal regression. The following theorem and corollary actually prove a more general result using a measure-theoretic definition of conditional expectation based on the Radon-Nikodym Theorem [114]. The measure theoretic definition conditions on an entire sub-sigma-algebra, so that the conditional expectation is viewed as a function or a random variable itself.

Theorem 14. *Consider a probability space $(\Omega, \mathcal{A}, P)$. Let X be any $\mathcal{A}$-measurable function whose integral exists and $\mathcal{B}$ be a σ-algebra contained in $\mathcal{A}$. Then,*

$$\mathrm{E}[X|\mathrm{E}[X|\mathcal{B}]] = \mathrm{E}[X|\mathcal{B}] \textit{ almost surely.}$$

Proof. Let $P_{\mathcal{B}}$ be the restriction of P to $\mathcal{B}$. By definition, the conditional expectation of X given $\mathcal{B}$, $\mathrm{E}[X|\mathcal{B}]$, is a $\mathcal{B}$-measurable function, defined up to $P_{\mathcal{B}}$ measure zero, by

$$\int_B \mathrm{E}[X|\mathcal{B}]dP_{\mathcal{B}} = \int_B XdP \tag{8.4}$$

for any $B \in \mathcal{B}$, where the existence of $\mathrm{E}[X|\mathcal{B}]$ is guaranteed by the Radon-Nikodym Theorem because $P_{\mathcal{B}}$ is absolutely continuous with respect to P. Since $\mathrm{E}[X|\mathcal{B}]$ is $\mathcal{B}$-measurable, the σ-algebra $\mathcal{C}$ generated by $\mathrm{E}[X|\mathcal{B}]$ is a sub-algebra of $\mathcal{B}$, and therefore a sub-algebra of $\mathcal{A}$. Hence, by definition the conditional expectation of X given $\mathcal{C}$, $\mathrm{E}[X|\mathcal{C}]$, is a $\mathcal{C}$-measurable function, defined up to $P_{\mathcal{C}}$ measure zero, by

$$\int_C \mathrm{E}[X|\mathcal{C}]dP_{\mathcal{C}} = \int_C XdP \tag{8.5}$$

for any $C \in \mathcal{C}$. Since $\mathcal{C} \subseteq \mathcal{B}$, (8.4) and (8.5) imply that

$$\int_C \mathrm{E}[X|\mathcal{B}]dP_{\mathcal{C}} = \int_C \mathrm{E}[X|\mathcal{C}]dP_{\mathcal{C}}$$

for any $C \in \mathcal{C}$. Hence, $\mathrm{E}[X|\mathcal{B}] = \mathrm{E}[X|\mathcal{C}]$ almost surely relative to $P_{\mathcal{C}}$. Q.E.D. □

Corollary 15. *Consider a probability space $(\Omega, \mathcal{A}, P)$ and let X be an integrable random variable and Y be a random vector. Then,*

$$\mathrm{E}[X|\mathrm{E}[X|Y]] = \mathrm{E}[X|Y]$$

almost surely.

Proof. Let $\mathcal{B}$ be the σ-algebra generated by Y. Then in Theorem 14 $\mathrm{E}[X|\mathcal{B}]$ becomes $\mathrm{E}[X|Y]$, $\mathcal{C}$ becomes the σ-algebra generated by $\mathrm{E}[X|Y]$, and $\mathrm{E}[X|\mathcal{C}]$ becomes $\mathrm{E}[X|\mathrm{E}[X|Y]]$. □

Note $X = \varepsilon_n$ is a random variable, which is integrable since the true error is bounded. If we let $Y = \widehat{\varepsilon}_{\mathrm{UEE}}$ be an uncalibrated error estimator, by Corollary 15 we have $\mathrm{E}[\varepsilon_n|\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{UEE}}]] = \mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{UEE}}]$. Since a calibrated error estimator is itself a conditional expectation given by $\widehat{\varepsilon}_{\mathrm{CEE}} = \mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{UEE}}]$,

$$\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{CEE}}] = \mathrm{E}\left[\varepsilon_n|\mathrm{E}\left[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{UEE}}\right]\right] = \mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{UEE}}] = \widehat{\varepsilon}_{\mathrm{CEE}}.$$

Hence, calibrated error estimators have ideal regression.

Similarly, if we let $Y = S_n$ be the entire observed sample, by Corollary 15 $\mathrm{E}[\varepsilon_n|\mathrm{E}[\varepsilon_n|S_n]] = \mathrm{E}[\varepsilon_n|S_n]$. Denoting the Bayesian error estimator by $\widehat{\varepsilon}_{\mathrm{MMSE}}$, we have $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{MMSE}}] = \widehat{\varepsilon}_{\mathrm{MMSE}}$, proving that Bayesian error estimators also have ideal regression. We will observe that joint density plots generated from Monte-Carlo simulations for calibrated error estimators and Bayesian error estimators indeed appear to have ideal regression.

C. Performance

In the following synthetic data simulations we assume a fixed sample size and known priors, generate random feature-label distributions, and generate random samples for each fixed feature-label distribution. A summary of the simulation methodology is shown in Fig. 46, which lists the general steps and flow of information. Throughout this section, we maintain the notation where $\widehat{\varepsilon}_{\mathrm{UEE}}$ is an uncalibrated error estimator, $\widehat{\varepsilon}_{\mathrm{CEE}}$ is a calibrated error estimator, and $\widehat{\varepsilon}_{\mathrm{MMSE}}$ is a Bayesian error estimator. We also use $\widehat{\varepsilon}$ in formulas that may be applied to all three types of error estimators.

1. Gaussian Model with LDA and Synthetic Data

In this section we evaluate the performance of MMSE calibrated error estimation using synthetic data from the Gaussian model with arbitrary covariance matrices defined in Section IV.A.6. We assume a fixed sample size, n, and LDA classification,

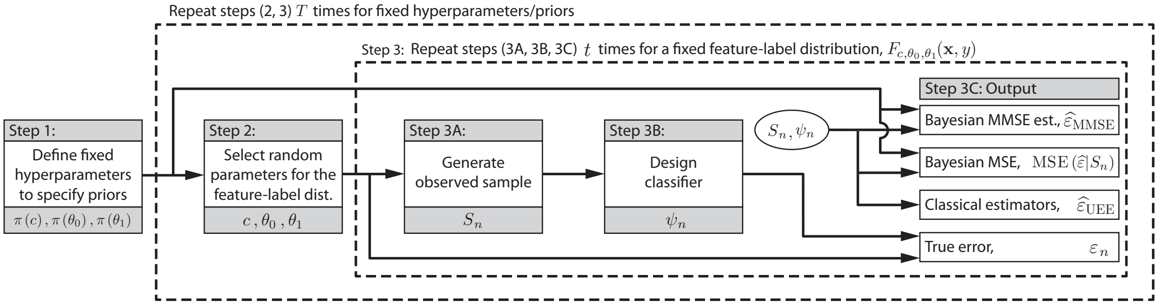

Repeat steps (2, 3) T times for fixed hyperparameters/priors
Step 3: Repeat steps (3A, 3B, 3C) t times for a fixed feature-label distribution, $F_{c,\theta_0,\theta_1}(\mathbf{x}, y)$
Step 1:
Define fixed hyperparameters to specify priors
$\pi(c), \pi(\theta_0), \pi(\theta_1)$
Step 2:
Select random parameters for the feature-label dist.
c, θ_0, θ_1
Step 3A:
Generate observed sample
S_n
Step 3B:
Design classifier
ψ_n
S_n, ψ_n
Step 3C: Output
Bayesian MMSE est., $\widehat{\varepsilon}_{\mathrm{MMSE}}$
Bayesian MSE, $\mathrm{MSE}(\widehat{\varepsilon}|S_n)$
Classical estimators, $\widehat{\varepsilon}_{\mathrm{UEE}}$
True error, ε_n

Fig. 46. Synthetic data simulation methodology for a Bayesian framework with fixed sample size.

where closed form solutions for the Bayesian error estimator and the RMS conditioned on the sample for arbitrary error estimators are both available.

In step 1 of Fig. 46, we specify one of three normal-inverse-Wishart priors: the "low-information," "medium-information" or "high-information" prior, with hyperparameters defined in Table 10. The *a priori* probability of class 0 is assumed to be known, with $c = 0.5$. All priors are proper probability densities designed to emulate prior knowledge in normalized microarray expression data, where class 0 is considered to represent a "good" prognosis (see Chapter VII for more information about priors for microarray data). The low information prior is closer to a flat non-informative prior and models a setting where our knowledge about the distribution parameters is less certain. Conversely, the high information prior has a relatively tight distribution around the expected parameters and models a situation where we have more certainty about the feature-label distribution. In general, the amount of information in each prior is reflected in the values of κ and ν, which increase as the amount of information in the prior increases. For each prior model, the parameter S for each class was inspired by the average variance of all features for both classes in the real breast cancer data set provided in [99]. We do not attempt to model differences between

Table 10. "Low-information," "medium-information" and "high-information" priors used in the Gaussian model for optimal calibration experiments

Hyperparameter	Low-info prior	Medium-info prior	High-info prior
a priori prob., c	fixed at 0.5	fixed at 0.5	fixed at 0.5
κ, class 0 and 1	$3D$	$9D$	$54D$
S, class 0 and 1	$0.03(\kappa - D - 1)I_D$	$0.03(\kappa - D - 1)I_D$	$0.03(\kappa - D - 1)I_D$
ν, class 0	$6D$	$18D$	$108D$
ν, class 1	$1D$	$3D$	$18D$
$\mathbf{m}$, class 0	$[0, 0, \ldots, 0]$	$[0, 0, \ldots, 0]$	$[0, 0, \ldots, 0]$
$\mathbf{m}$, class 1	$-0.1210[1, 1, \ldots, 1]$	$-0.1925[1, 1, \ldots, 1]$	$-0.2000[1, 1, \ldots, 1]$

the variances of the classes or any correlations, however these will be considered in the posterior using the observed data. The parameter $\mathbf{m}$ for class 0 (representing the expected mean of the class) is set to zero, which approximates the effect of data normalization, and $\mathbf{m}$ for class 1 has been adjusted to give an expected true error of about 0.28 with one feature.

In step 2 we generate random feature-label distribution parameters from the chosen prior. With $c = 0.5$ fixed, we need only realizations of μ_0, Σ_0, μ_1 and Σ_1. For each class, we select a random covariance according to the inverse-Wishart distribution with parameters κ and S, $\pi(\Sigma_y)$, using methods in [104]. Conditioned on the covariance, we generate a random mean using the Gaussian distribution $\pi(\mu_y|\Sigma_y) = f_{\mathbf{m},\Sigma_y/\nu}(\mu_y)$, resulting in a normal-inverse-Wishart distributed mean and covariance pair. The parameters for each class are generated independently.

In step 3A, we generate a training sample of sample size n (n even) from the realized feature-label distribution. The sample sizes of both classes are fixed at $n_0 = n_1 = n/2$. The corresponding number of sample points in each class, $y \in \{0, 1\}$, are synthetically produced according to Gaussian f_{μ_y,Σ_y} class-conditional distributions. In this way, we generate n stratified labeled training points. Next, these labeled sample points are used to train an LDA classifier in step 3B. No feature selection is involved.

In step 3C, we collect several output variables, including the exact true error, ε_n, from the classifier and true distribution, and the Bayesian MMSE error estimator, $\widehat{\varepsilon}_{\mathrm{MMSE}}$, from the sample, classifier and (correct) priors. To aid performance analysis, from the sample, classifier and priors we also find the MSE of the Bayesian error estimator conditioned on the sample defined in Chapter V by

$$\mathrm{MSE}\,(\widehat{\varepsilon}_{\mathrm{MMSE}}|S_n) = \mathrm{E}[(\varepsilon_n - \widehat{\varepsilon}_{\mathrm{MMSE}})^2|S_n].$$

The Bayesian error estimator is theoretically optimal in the MSE sense. Since the classifier is linear, both $\widehat{\varepsilon}_{\mathrm{MMSE}}$ and $\mathrm{MSE}\,(\widehat{\varepsilon}_{\mathrm{MMSE}}|S_n)$ may be computed exactly using closed form expressions. The training data and classifier are also used to evaluate several classical training-data error estimators: 5-fold cross-validation, 0.632 bootstrap and bolstered resubstitution. The conditional MSE of any error estimator, $\widehat{\varepsilon}$, may be evaluated off-line for each iteration from (5.8):

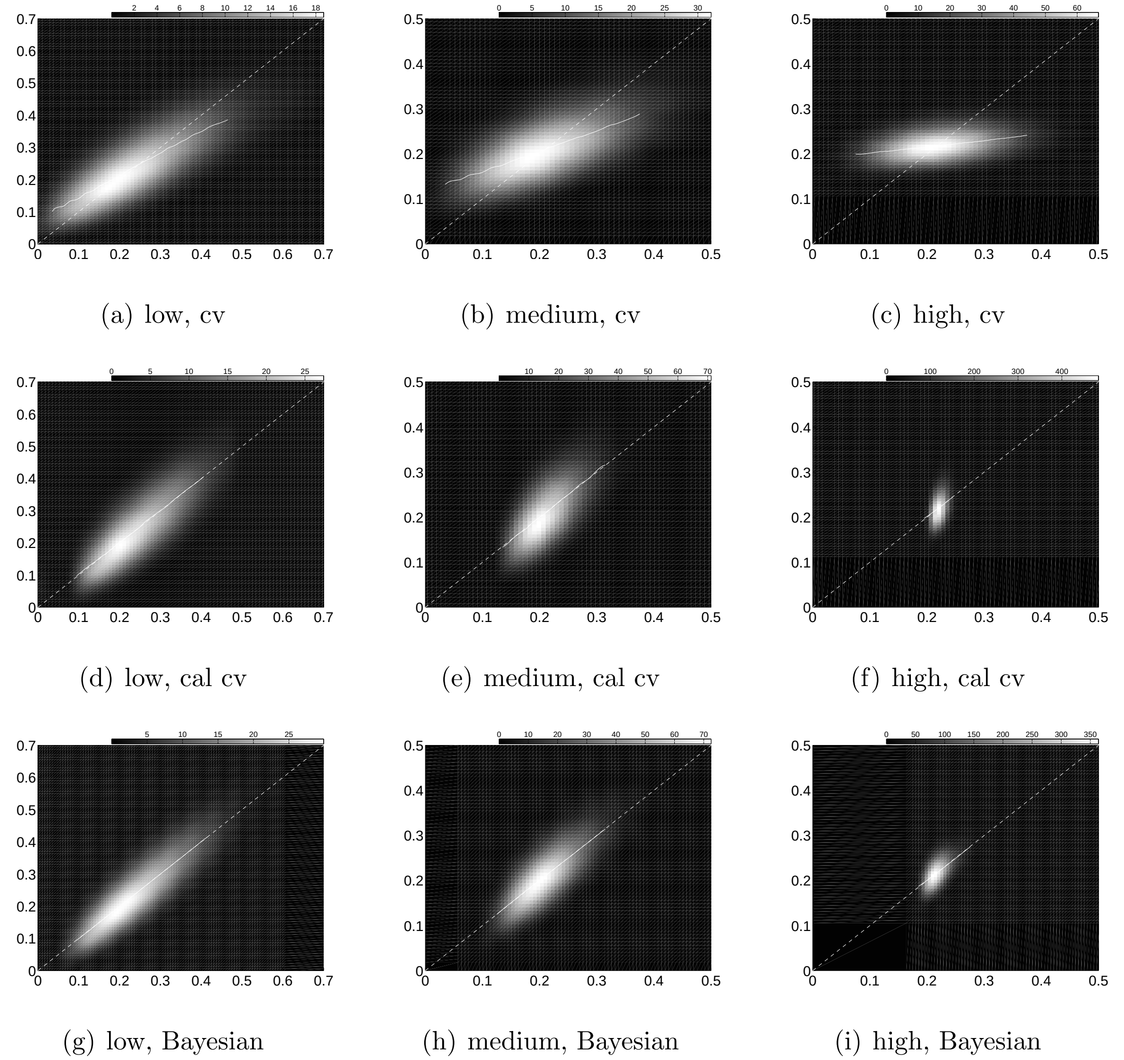
$$\mathrm{MSE}(\widehat{\varepsilon}|S_n) = \mathrm{MSE}(\widehat{\varepsilon}_{\mathrm{MMSE}}|S_n) + (\widehat{\varepsilon}_{\mathrm{MMSE}} - \widehat{\varepsilon})^2.$$

For each fixed feature-label distribution, steps 3A through 3C (collectively called step 3) are repeated $t = 1,000$ times to obtain t samples and sets of output. Further, step 2 is repeated $T = 10,000$ times for T different feature-label distributions (corresponding to the randomly selected parameters). In total, each simulation produces $t \times T = 10$ million samples and sets of output results.

After the simulation is complete, the synthetically generated true and estimated error pairs are used to estimate four joint densities, $f\,(\varepsilon_n, \widehat{\varepsilon}_{\mathrm{MMSE}})$ and $f\,(\varepsilon_n, \widehat{\varepsilon}_{\mathrm{UEE}})$, where $\widehat{\varepsilon}_{\mathrm{UEE}}$ can be cross-validation, bootstrap or bolstering. We use a bivariate Gaussian kernel density estimation method. For each non-Bayesian error estimator, we also find the expected true error conditioned on the error estimate, $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{UEE}}]$. This expectation is defined in (8.1), but we approximate it by uniformly partitioning the interval $[0, 1]$ into 500 bins and averaging the true errors corresponding to error estimates that fall in each bin. Moreover, the average true error is only found for bins with at least 100 points; otherwise, the bin is considered "rare" and the lookup table simply leaves the error estimate unchanged (an identity mapping). The result is a calibration function (a lookup table) mapping each of the 500 bins to a corresponding expected true error.

Once a lookup table has been generated for each error estimator, the entire experiment is repeated again using the same prior model, classification rule, and classical training-data error estimators; however, at the end of each iteration in step 3C, this time we apply the corresponding MMSE calibration lookup tables to each non-Bayesian error estimator. We also report the exact true error and Bayesian sample-conditioned MSE again, but the Bayesian error estimator is not needed since it is not calibrated and performance would be identical to the original experiment. As before, the procedure is iterated $t = 1,000$ times for each fixed feature-label distribution for $T = 10,000$ sets of feature-label distribution parameters.

Figure 47 shows the estimated joint densities between the true error (y-axis) and three error estimators (x-axis) for $D = 2$ and $n = 30$ sample points. Low, medium and high-information priors are shown left to right, with expected true errors 0.2494, 0.2153 and 0.2194, respectively. Cross-validation is shown in the top row, calibrated cross-validation in the middle row and the optimal Bayesian error estimator in the bottom row. Within each sub-figure, the dashed white line represents the ideal case where an error estimate equals the true error, and the solid white line is the expected



(a) low, cv

(b) medium, cv

(c) high, cv

(d) low, cal cv

(e) medium, cal cv

(f) high, cal cv

(g) low, Bayesian

(h) medium, Bayesian

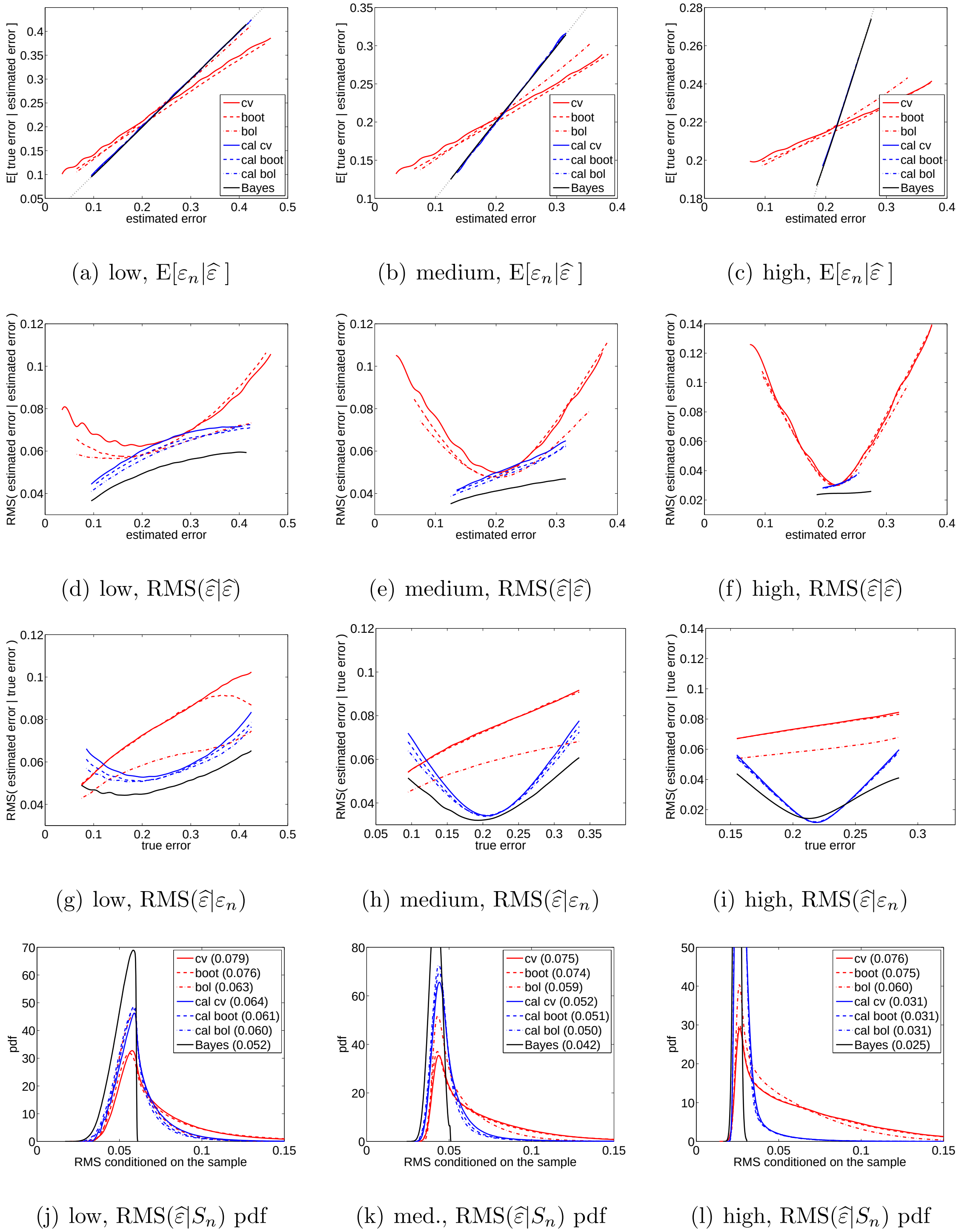
(i) high, Bayesian

Fig. 47. Joint distributions between true errors (y-axis) and error estimators (x-axis) for the Gaussian model with $D = 2$, $n = 30$ and LDA. Low, medium and high-information priors are shown left to right, with expected true errors 0.2494, 0.2153 and 0.2194, respectively. Cross-validation, calibrated cross–validation and Bayesian error estimation with correct priors are shown in the top, middle and bottom rows, respectively. White areas indicate a higher density, where the scale for each plot is shown in the upper right.

true error conditioned on the error estimator, $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{UEE}}]$, $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{CEE}}]$, or $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{MMSE}}]$. To avoid misleading results from rare observations of the error estimate, the estimated error is partitioned into 100 bins and the expected true error is only shown for bins with at least $t \times T/100 = 100,000$ points. Similar results were found for $D = 1$ and $D = 5$, which are provided in the supplementary material of [54].

The results illustrate good performance for calibrated cross-validation, relative to classical cross-validation. Analogous plots for the bootstrap and bolstered error estimators are not shown but similar. Classical cross-validation has decent regression with the true error for the low-information prior, but much less regression for higher information priors. See for instance Fig. 47(c), where the regression is nearly flat. On the other hand, calibrated cross-validation, like Bayesian error estimation, has ideal regression with the true error in all plots, which is consistent with the theory presented in Section VIII.B.

Figure 48 shows four different kinds of performance results for $D = 2$ and $n = 30$: expected true error given estimated error, conditional RMS given estimated error, conditional RMS given true error, and probability densities for the sample-conditioned RMS. Left, middle and right columns contain plots for low, medium and high-information priors, respectively. In all sub-figures, the Bayesian error estimator is shown in black, the three classical error estimators considered (cross-validation, bootstrap and bolstering) are in red, and the corresponding calibrated error estimators are in blue. Legends for all sub-figures are the same and shown in the top and bottom rows. Similar results were found for $D = 1$ and $D = 5$, which are provided in the supplementary material of [54]. In general, although the Bayesian error estimator may have slightly better MSE performance, calibrated cross-validation is easy to implement and still offers a significant improvement over classical cross-validation within the proposed Bayesian models, especially for higher information priors.



(a) low, $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}\,]$ (b) medium, $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}\,]$ (c) high, $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}\,]$

(d) low, $\mathrm{RMS}(\widehat{\varepsilon}|\widehat{\varepsilon})$ (e) medium, $\mathrm{RMS}(\widehat{\varepsilon}|\widehat{\varepsilon})$ (f) high, $\mathrm{RMS}(\widehat{\varepsilon}|\widehat{\varepsilon})$

(g) low, $\mathrm{RMS}(\widehat{\varepsilon}|\varepsilon_n)$ (h) medium, $\mathrm{RMS}(\widehat{\varepsilon}|\varepsilon_n)$ (i) high, $\mathrm{RMS}(\widehat{\varepsilon}|\varepsilon_n)$

(j) low, $\mathrm{RMS}(\widehat{\varepsilon}|S_n)$ pdf (k) med., $\mathrm{RMS}(\widehat{\varepsilon}|S_n)$ pdf (l) high, $\mathrm{RMS}(\widehat{\varepsilon}|S_n)$ pdf

Fig. 48. Conditional RMS performance for Gaussian models ($D = 2$, $n = 30$, LDA).

The top row in Fig. 48 shows the expected true error conditioned on the error estimate. For cross-validation, these are the same plots presented with the joint density graphs, and as before the dotted diagonal line represents an ideal error estimator equal to the true error. The second row shows RMS for each error estimator conditioned on the error estimate itself, which, by definition, is given by

$$\mathrm{RMS}[\widehat{\varepsilon}\,|\widehat{\varepsilon}\,] = \sqrt{\mathrm{E}\left[(\varepsilon_n - \widehat{\varepsilon})^2\,|\widehat{\varepsilon}\,\right]},$$

and the third row shows RMS conditioned on the true error,

$$\mathrm{RMS}[\widehat{\varepsilon}\,|\varepsilon_n] = \sqrt{\mathrm{E}\left[(\varepsilon_n - \widehat{\varepsilon})^2\,|\varepsilon_n\right]},$$

where $\widehat{\varepsilon}$ equals $\widehat{\varepsilon}_{\mathrm{UEE}}$, $\widehat{\varepsilon}_{\mathrm{CEE}}$, or $\widehat{\varepsilon}_{\mathrm{MMSE}}$. These graphs indicate error estimation accuracy for fixed error estimates and fixed true errors, respectively. Finally, the bottom row has probability densities for the RMS conditioned on the sample for each error estimator, that is, estimated densities for the root of the values computed for $\mathrm{MSE}(\widehat{\varepsilon}|S_n)$ over all samples. For comparison, legends in the bottom row also show the unconditional RMS for all error estimators (averaged over both distributions and samples).

All simulations in the top row of Fig. 48 again demonstrate that the expected true error conditioned on calibrated error estimators aligns with the ideal dashed diagonal line, as they must. Furthermore, the RMS conditioned on calibrated error estimators is significantly improved relative to their uncalibrated counterparts, usually tracking just above the Bayesian error estimator. Figure 48(d) is typical, where for the low information prior the RMS conditioned on calibrated error estimators is almost uniformly lower.

The RMS conditioned on uncalibrated error estimators tends to have a “V” shape, achieving a minimum RMS for a very small window of estimated errors. The

RMS conditioned on a low estimated error tends to be high because the error estimator is usually low-biased, and conditioning on a high estimated error tends to result in a high RMS because the error estimator is high-biased. The error estimate where the RMS is minimized approximately corresponds to the point where the expected true error conditioned on the error estimate crosses the ideal dotted line. This is seen for example in Fig. 48(b) for the medium-information prior, where the expected true error for all uncalibrated error estimators crosses the ideal dotted line just above 0.2 and in Fig. 48(e) they all have minimum RMS just above 0.2. Note in a small-sample setting without modeling assumptions, this window where the estimated error is most accurate is unknown, in contrast to Bayesian modeling where these graphs demonstrate how to find the optimal window. Furthermore, the error-estimate-conditioned RMS of calibrated error estimators and Bayesian error estimators tend to monotonically increase, so that the accuracy of error estimation is usually known to be higher when the estimated error is low.

Figure 48(g) for the low-information prior is a very typical representative for the behavior of the RMS conditioned on true errors. Uncalibrated error estimators tend to be best for low true errors, which is consistent with many previous studies on error estimation accuracy [65, 29, 115]. Bayesian error estimators are usually best for moderate true errors where small-sample classification is most interesting, as observed in Chapter III. This is also true for calibrated error estimators, which have true-error-conditioned RMS plots usually tracking just above the Bayesian error estimator.

Although the unconditional RMS for Bayesian error estimators are guaranteed to be optimal (within the assumed model), in some cases the conditional RMS of calibrated error estimators can actually exceed that of the Bayesian error estimator for some small ranges of the true error, as in Fig. 48(i) for true errors around

0.22. Furthermore, although the unconditional RMS for calibrated error estimators are guaranteed to be lower than their uncalibrated counterparts, uncalibrated error estimators can even outperform Bayesian error estimators in the same way, as in Fig. 48(g) where uncalibrated bolstering has the best RMS for true errors less then 0.1. This is possible because Bayesian error estimators are only guaranteed to be optimal given a fixed sample, and calibrated error estimators are only guaranteed to be optimal for a fixed observed error estimate, whereas there is no guarantee of optimality for fixed distributions or any arbitrary class of distributions (e.g., the class of distributions having a specified true error).

The distribution of the RMS conditioned on the sample for calibrated error estimators tends to have more mass towards lower values of RMS than uncalibrated error estimators, with the Bayesian error estimator being even more shifted to the left. This is evident for example in Fig. 48(k), where the unconditional RMS indicated in the legend of this graph for calibrated error estimators (at most 0.052) is always lower than that of the uncalibrated estimators (at best 0.059), with the Bayesian error estimator having optimal RMS (at 0.042).

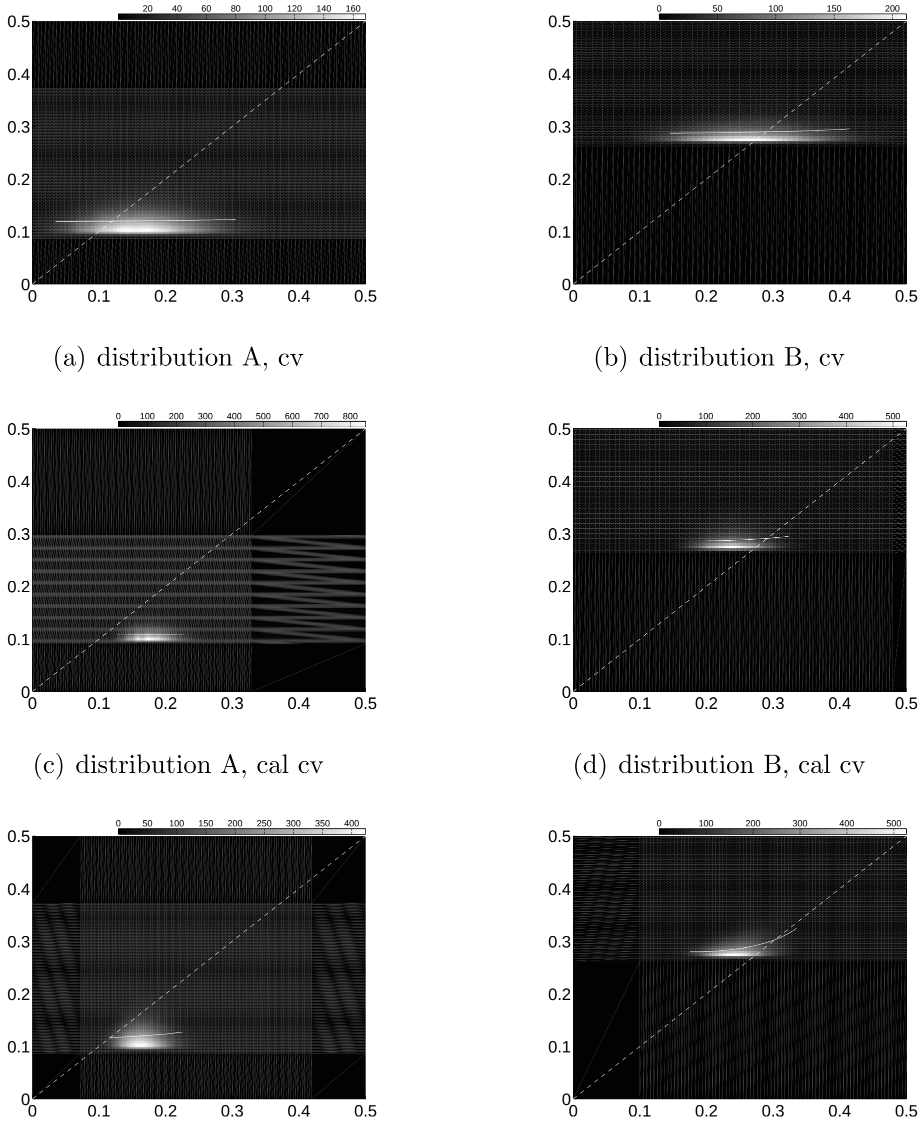
The performances of all calibrated error estimators tend to be very close relative to each other. For example, all blue curves in Fig. 48 have almost the same performance, with perhaps the calibrated bolstered error estimator performing slightly better than the others. This phenomenon may be due to a fundamental limit in the amount of information available from classical counting and even bolstered counting error estimators. Further, there is a gap in the performance between the optimal Bayesian error estimator and the calibrated error estimators. This may be because calibrated error estimators average the expected true error over all samples producing the observed error estimate, so that performance must be averaged over different trained classifiers, whereas the Bayesian error estimator is always evaluated directly on the actual designed classifier. If averaging over random classifiers from a classification rule introduces additional uncertainty in the estimation problem, the RMS performance of calibrated error estimators may be inherently bounded some distance from the optimal Bayesian error estimator.

We next illustrate performance for fixed distributions. For the purposes of demonstration we consider two distributions, named distribution “A” and distribution “B,” drawn from the medium-information prior with $D = 2$ and provided in Table 11. Since we are interested in only a single fixed distribution at a time, we perform a new experiment using only step 3 in Fig. 46 with the medium-information prior and the same classification rule and classical, calibrated and Bayesian error estimators as before. We collect only the true error and error estimates, and repeat the procedure $t = 1,000,000$ times.

Figure 49 shows the estimated joint densities between the true error (y-axis) and three error estimators (x-axis) for distributions A (left) and B (right) with $n = 30$ sample points. Cross-validation is shown in the top row, calibrated cross-validation in the middle row and the optimal Bayesian error estimator in the bottom row. As before, the dashed white line in each sub-figure represents the ideal case where an error estimate equals the true error and the solid white line is the expected true error conditioned on the error estimator (nonlinear regression). The estimated error is partitioned into 100 bins and the expected true error is only shown for bins with at least $t/100 = 10,000$ points. Joint density graphs for fixed distributions typically exhibit very little regression [67, 68] and we witness that phenomenon here, especially in regard to uncalibrated and calibrated cross-validation. While both exhibit virtually no regression, calibrated cross-validation has much less variation. The Bayesian estimator has some regression. The lack of regression in Fig. 49 is in contrast to joint densities for Bayesian models, which achieve regression by spreading flat joint density

Table 11. Fixed distributions “A” and “B” from the medium-information prior used in the Gaussian model with $D = 2$

Hyperparameter	Distribution A	Distribution B
a priori probability, c	0.5	0.5
μ_0 (mean of class 0)	$[0.000, -0.068]$	$[-0.001, 0.015]$
μ_1 (mean of class 1)	$[-0.324, -0.230]$	$[-0.173, -0.114]$
diagonal of Σ_0 (variance of class 0 features)	$[0.018, 0.029]$	$[0.039, 0.024]$
off-diagonal of Σ_0 (covariance between class 0 features)	0.000	-0.003
diagonal of Σ_1 (variance of class 1 features)	$[0.019, 0.020]$	$[0.033, 0.049]$
off-diagonal of Σ_1 (covariance between class 1 features)	-0.001	-0.009
expected true error	0.11	0.29



(a) distribution A, cv

(b) distribution B, cv

(c) distribution A, cal cv

(d) distribution B, cal cv

(e) distribution A, Bayesian

(f) distribution B, Bayesian

Fig. 49. Joint distributions between true errors (y-axis) and error estimators (x-axis) for fixed distributions from the Gaussian model with $D = 2$, $n = 30$ and LDA. Distribution A is shown on the left and distribution B on the right. Cross-validation, calibrated cross-validation and Bayesian error estimation with medium-information priors are shown in the top, middle and bottom rows, respectively. White areas indicate a higher density, where the scale for each plot is shown in the upper right.

graphs, like those in Fig. 49, across all densities in a Bayesian model according to a prior distribution (Fig. 47). When considering the distributions and regression lines in Figs. 47 and 49, one needs to keep in mind the difference in their settings.

2. Gaussian Model with LDA Applied to Real Breast Cancer Data

To implement a Bayesian analysis on a given data set using a Gaussian model, a practitioner should select features passing a Gaussianity test (such as the Shapiro-Wilk test) or verify that the selected feature set in a given data set is approximately Gaussian. The next step is to determine priors for the distribution parameters, including a prior for c (uniform, beta or fixed) and normal-inverse-Wishart hyperparameters for each class. Note that the calibration scheme described here requires a proper prior because (8.2) and (8.3) are only valid if $\pi(\theta)$ is proper, and Monte-Carlo methods are based on generating random parameters from valid distributions in step 2 of Fig. 46. Thus, rather than a flat "non-informative" prior, we will use a low information prior similar to the one in Table 10 for calibration. In any case, once a proper prior is established, along with a sample size, classification rule and error estimator, one may use the methods described previously with synthetic data to find the corresponding calibration function. This may then be applied to an estimate of the true error based on real data to obtain a calibrated error estimate.

That being said, demonstrating RMS performance for real data is difficult because any data set essentially represents a single realization of the distribution parameters. This is not a new problem or a consequence of the theory of calibrated error estimation, but rather an inherent difficulty that always persists with real data analysis. If we want to consider performance for a specific true distribution, then we cannot indulge in the randomization of the feature-label distribution in step 2 of Fig. 46; rather, we would fix it and only average over the samples. The prior distri-

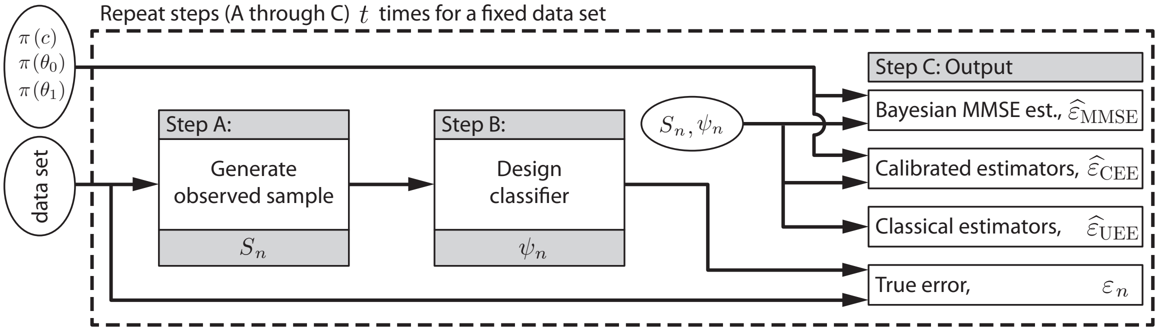

Repeat steps (A through C) t times for a fixed data set
$\pi(c)$
$\pi(\theta_0)$
$\pi(\theta_1)$
data set
Step A:
Generate observed sample
S_n
Step B:
Design classifier
ψ_n
S_n, ψ_n
Step C: Output
Bayesian MMSE est., $\widehat{\varepsilon}_{\text{MMSE}}$
Calibrated estimators, $\widehat{\varepsilon}_{\text{CEE}}$
Classical estimators, $\widehat{\varepsilon}_{\text{UEE}}$
True error, ε_n

Fig. 50. Real data simulation methodology for a Bayesian framework with fixed sample size.

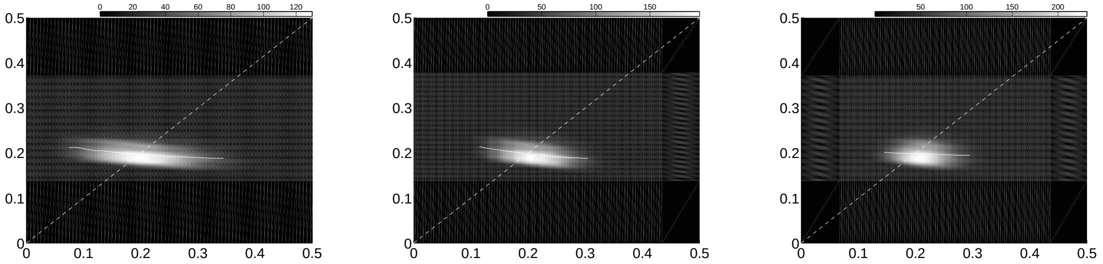
bution would still be involved in error estimation, but we are no longer interested in averaging performance across the prior distribution. This is precisely the approach taken in this section with the performance analysis methodology outlined in Fig. 50. It is similar to the synthetic data methodology in Fig. 46, except that we do not simulate steps 1 or 2.

Our real data set is the same normalized gene-expression measurements from a breast cancer study [99] used in Section IV.B.4. The data set includes 295 sample points, each with a 70 feature gene profile. 180 points are assigned to class 0 (good prognosis) and 115 to class 1 (bad prognosis).

In step A, we randomly select a stratified sample of size n, where the ratio of points from each class is kept as close as possible to that of the original data set. For $n = 30$, 18 points are in class 0 and 12 points are in class 1. In step B, we design an LDA classifier on the initial training sample. The classifier is designed from fixed feature sets: $\{\text{CENPA}\}$ for $D = 1$, $\{\text{CENPA}, \text{BBC3}\}$ for $D = 2$ and $\{\text{CENPA}, \text{BBC3}, \text{CFFM4}, \text{TGFB3}, \text{DKFZP564D0462}\}$ for $D = 5$. These have previously been shown to perform reasonably well on the full data set and a multivariate Shapiro-Wilk test applied to the full data set does not reject Gaussianity over either of the classes at a 95% significance level [28]. Although we do not implement a feature selection scheme here, one can be applied as part of the classifier design in step B.

In step C we approximate the true error of the classifier using holdout points remaining in the data set, and compute three classical error estimates, including 5-fold cross-validation, bootstrap and bolstering. Using a modified low-information prior with c fixed at 0.61 instead of 0.5 (corresponding to the proportion of sample points in class 0), we evaluate calibration functions using exactly the same method described in the synthetic data study of Section VIII.C.1. Using these calibration functions, we compute three calibrated error estimates corresponding to each classical error estimator. Finally, we evaluate two Bayesian error estimators: one using the modified low-information prior and the other using a flat non-informative prior where c is uniform from 0 to 1 and the priors for both classes are improper flat distributions such that $\pi(\theta_0) = \pi(\theta_1) \propto 1$. All together, we evaluate eight error estimators, and the entire sampling, classification and error estimation process is repeated $t = 1,000,000$ times.

Figure 51 shows the estimated joint densities between the approximate true error (y-axis) and three error estimators (x-axis) for $D = 2$ and $n = 30$ sample points. Cross-validation is shown in part (a), calibrated cross-validation in part (b) and the low-information Bayesian error estimator in part (c). As before, the dashed white line represents the ideal case where an error estimate equals the true error, and the solid white line is the expected true error conditioned on the error estimator, $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}\,]$. To avoid misleading results from rare observations of the error estimate, the estimated error is partitioned into 100 bins and the expected true error is only shown for bins with at least $t \times T/100 = 10,000$ points. Similar plots for $D = 1$ and $D = 5$ are available in the supplementary material of [54]. Also, the average true errors and unconditioned RMS performance results are shown in Table 12 for $n = 30$ and $D = 1$,



(a) cv

(b) cal cv

(c) low-info Bayesian

Fig. 51. Joint distributions between true errors (y-axis) and error estimators (x-axis) for real data with $D = 2$, $n = 30$ and LDA. Cross-validation, calibrated cross-validation and Bayesian error estimation with low-information priors are shown left to right. White areas indicate a higher density, where the scale for each plot is shown in the upper right.

2 and 5.

The joint densities in Fig. 51 have almost no regression, which is similar to the fixed-distribution graphs in Fig. 49. This is because these simulations are based on a single data set representing a single realization of the distribution parameters and, as noted previously, lack of regression is common in such a situation. Indeed, here we even see slightly negative regression. This is not an abberation; it has been theoretically shown that negative correlation can occur for a standard model [27]. In Table 12, we observe results that are similar to the synthetic data results. Bayesian error estimators typically perform best, with the low-information prior performing better than the flat prior. Also, calibrated error estimators generally outperform their uncalibrated counterparts, each having similar RMS performance regardless of the underlying error estimation rule.

Table 12. Average true error and the RMS performance of error estimators on real breast cancer data for optimal calibration experiments with fixed feature sets and $n = 30$

Features	Average true error	RMS - Uncalibrated			RMS - Calibrated			RMS - Bayesian	
		cv	boot	bol	cv	boot	bol	low-info	flat
$D = 1$	0.1988	0.077	0.074	0.069	0.061	0.060	0.063	0.053	0.061
$D = 2$	0.1978	0.078	0.076	0.064	0.057	0.055	0.055	0.047	0.058
$D = 5$	0.2148	0.086	0.088	0.059	0.051	0.051	0.053	0.051	0.095

3. Gaussian Model with 3NN and Synthetic Data

In this section, we again evaluate the performance of MMSE calibrated error estimation on synthetic Gaussian models with arbitrary covariance matrices and fixed sample size, this time for the 3-nearest-neighbor (3NN) classification rule. In this case, closed form solutions for the Bayesian error estimator and the MSE conditioned on the sample for arbitrary error estimators are not available and must be approximated using Monte-Carlo methods.

The simulation methodology is based on Fig. 46 with a few modifications. Since calculations involved in 3NN classification require more computation time, in step 1 we use only the medium-information prior shown in Table 10 to demonstrate that results for 3NN are similar to those obtained for LDA. As before, c is assumed to be known and fixed at 0.5. Step 2 is performed exactly as before, where we select random parameters, μ_0, Σ_0, μ_1 and Σ_1, from the normal-inverse-Wishart medium-information priors.

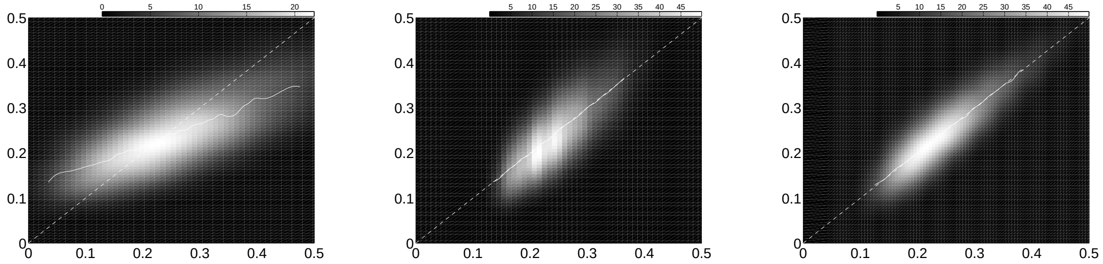
Data generation in step 3 is also unchanged. We draw n stratified labeled training points, $n/2$ being from class $y \in \{0, 1\}$ with f_{μ_y,Σ_y} class-conditional distributions. We then apply the 3NN classification rule to the training data in step 3B. As before, no feature selection is involved.

In step 3C, we implement several changes to work with the new 3NN classifier. First, the true error, ε_n, is now approximated by independently generating 100,000 labeled data points from the same distribution as the training data and evaluating the proportion of points mislabeled by the classifier. Second, since the classifier is non-linear, the Bayesian MMSE error estimator and theoretical MSE of the Bayesian error estimator conditioned on the sample, $\text{MSE}\,(\widehat{\varepsilon}_{\text{MMSE}}|S_n)$, are approximated using a Monte-Carlo approach. In particular, for each iteration and class $y \in \{0, 1\}$, we generate 10,000 mean and covariance pairs from the corresponding normal-inverse-Wishart posterior. The first and second moments of the true error contributed by class y, $\widehat{\varepsilon}^y = E_{\pi^*}[\varepsilon_n^y(\theta_y)]$ and $E_{\pi^*}[(\varepsilon_n^y(\theta_y))^2]$, are then approximated by averaging $\varepsilon_n^y(\theta_y)$ and $(\varepsilon_n^y(\theta_y))^2$ for our 3NN classifier on each of the 10,000 distributions. The Bayesian error estimator is then approximated from (2.10), and $\mathrm{MSE}\,(\widehat{\varepsilon}_{\mathrm{MMSE}}|S_n)$ is found using formulas derived from the definition in Chapter V. Although the Bayesian error estimator evaluated here is approximately optimal in the mean-square sense, this part of the code is by far the most time-consuming. See Chapter VII for more details on Monte-Carlo approximation in Bayesian error estimation.

Finally, the training data and classifier are used to evaluate several classical training-data error estimators, including resubstitution, 5-fold cross-validation, 0.632 bootstrap and semi-bolstered resubstitution. The conditional MSE of each of these error estimators is evaluated off-line from (5.8).

Step 3 is executed only $t = 1$ time for each fixed feature-label distribution, and step 2 is repeated $T = 100,000$ times for T different feature-label distributions. In total, each simulation produces $t \times T = 100,000$ samples and sets of output results.

After the simulation is complete, the $t \times T = 100,000$ synthetically generated true and estimated error pairs are used to estimate five joint densities, $f\,(\varepsilon_n, \widehat{\varepsilon}_{\mathrm{MMSE}})$ and $f\,(\varepsilon_n, \widehat{\varepsilon}_{\mathrm{UEE}})$, where $\widehat{\varepsilon}_{\mathrm{UEE}}$ can be resubstitution, cross-validation, bootstrap or semi-bolstering. We use the same bivariate Gaussian kernel density estimation method as before. For each non-Bayesian error estimator, we also find the expected true error conditioned on the error estimate, $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{UEE}}]$. Since the number of error pairs is smaller in the 3NN simulation, this time we approximate it by uniformly partitioning the interval $[0, 1]$ into only 100 bins and averaging the true errors corresponding to each bin. Also, the average true error is only found for bins with at least 10 points, otherwise the bin is too rare and the lookup table leaves the error estimate unchanged.



(a) cv

(b) cal cv

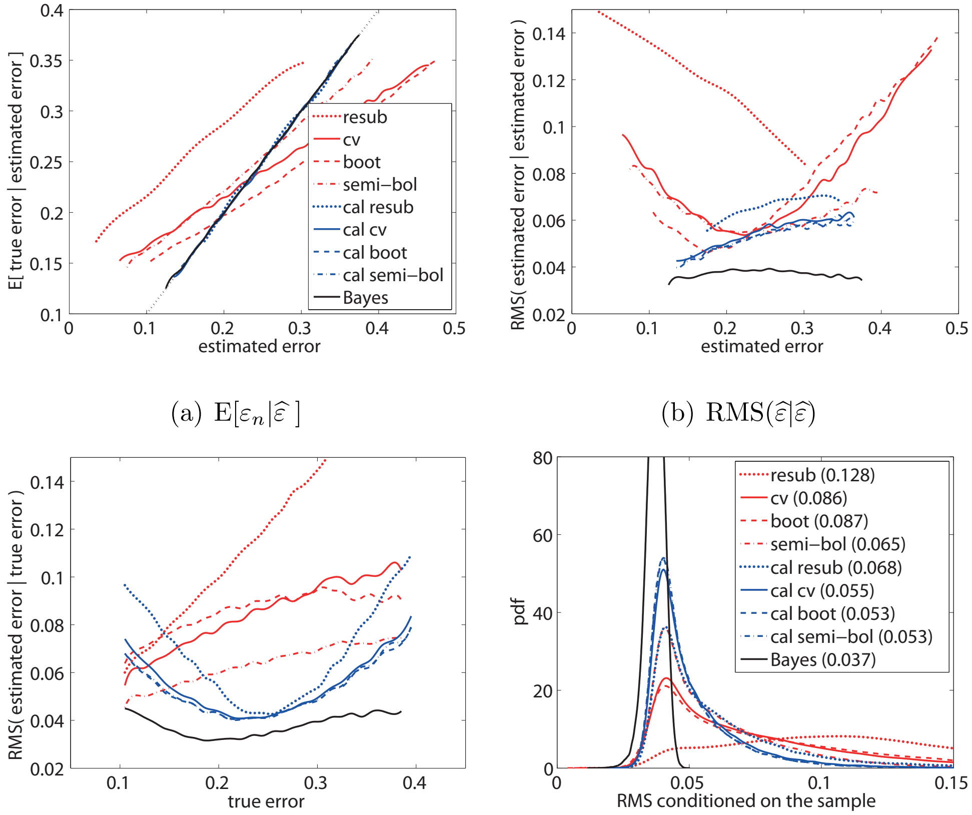
(c) Bayesian

Fig. 52. Joint distributions between true errors (y-axis) and error estimators (x-axis) for the medium-information Gaussian model with $D = 2$, $n = 30$ and 3NN. The expected true error is 0.2153. Cross-validation, calibrated cross-validation and Bayesian error estimation with correct priors are shown left to right. White areas indicate a higher density, where the scale for each plot is shown in the upper right.

The result is a calibration function mapping each of the 100 bins to a corresponding expected true error.

Once a lookup table has been generated for each error estimator, the entire experiment is repeated again using the same prior model, classification rule, and classical training-data error estimators. However, at the end of each iteration in step 3C, this time we apply the corresponding MMSE calibration lookup tables to each non-Bayesian error estimator. As in the LDA experiments, we also report the approximate true error and Bayesian sample-conditioned MSE again, but not the Bayesian error estimator. Also, only $t = 1$ training sample is drawn from each fixed feature-label distribution for $T = 100,000$ sets of feature-label distribution parameters.

Figure 52 shows the joint density between the true error (y-axis) and estimated error (x-axis) for cross-validation, calibrated cross-validation and Bayesian error estimation with $D = 2$ and $n = 30$ sample points, medium-information priors and



(a) $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}\,]$

(b) $\mathrm{RMS}(\widehat{\varepsilon}|\widehat{\varepsilon})$

(c) $\mathrm{RMS}(\widehat{\varepsilon}|\varepsilon_n)$

(d) $\mathrm{RMS}(\widehat{\varepsilon}|S_n)$ pdf

Fig. 53. Conditional RMS performance for the medium-information Gaussian model ($D = 2$, $n = 30$, 3NN).

3NN classification. This figure is analogous to the middle column of Fig. 47 for LDA classification. Cross-validation is shown in part (a), calibrated cross-validation in part (b) and the optimal Bayesian error estimator in part (c). The solid white line is the expected true error conditioned on the error estimator. To avoid misleading results from rare observations, the estimated error is partitioned into 100 bins and the expected true error is only shown for bins with at least $t \times T/100 = 1,000$ points. Although the number of iterations with 3NN is only $t \times T = 100,000$, the joint density plots for 3NN are clearly similar to those for LDA.

Figure 53 for 3NN is analogous to the middle column of Fig. 48 for LDA, and presents four different kinds of performance results for the medium-information prior with $D = 2$ and $n = 30$. In all sub-figures, the Bayesian error estimator is shown in black, the four classical error estimators considered (resubstitution, cross-validation, bootstrap and bolstering) are in red, and the corresponding calibrated error estimators are in blue. Legends for all figures are the same and shown in two of the sub-figures. Results for 3NN are again very similar to LDA. Resubstitution in particular improves dramatically. For example in Fig. 53(a) we see that it is very low biased for 3NN; indeed, for resubstitution $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{UEE}}] > \widehat{\varepsilon}_{\mathrm{UEE}}$ for all values of $\widehat{\varepsilon}_{\mathrm{UEE}}$. On the other hand, as must be the case, $\mathrm{E}[\varepsilon_n|\widehat{\varepsilon}_{\mathrm{CEE}}] = \widehat{\varepsilon}_{\mathrm{CEE}}$.

D. Discussion

Given a fixed sample size, classification rule, error estimation rule, and Bayesian framework with priors, MMSE calibrated error estimation offers a method to optimize the performance of the specified error estimator. A primary point is that it becomes possible to take advantage of modeling assumptions offered from a Bayesian framework for any classification and error estimation rule pair, especially when closed-form analytical solutions for the Bayesian error estimator are not available. The calibration function itself may be found in a conceptually straightforward manner via Monte-Carlo simulations, where the modeling assumptions are used to emulate the entire classification procedure and collect true and estimated error pairs for joint density estimation. Although discovering a calibration function is somewhat computationally involved, once found it may be kept in a database for use any time the modeling assumptions are employed. Furthermore, since calibration functions are essentially lookup tables, they may be easily applied with almost no changes in any classification and error estimation procedures, coding infrastructure, or simulation methodology.

Let us close by noting that, while the requirement of a Bayesian framework for calibration might at first glance seem constraining, when confronted with small sample sizes one really has very little other choice if accurate error estimation is to be achieved: accurate distribution-free small-sample error estimation is virtually impossible [32].

REFERENCES

[1] E. R. Dougherty, A. Datta, and C. Sima, "Research issues in genomic signal processing," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 46–68, 2005.

[2] E. R. Dougherty and U. Braga-Neto, "Epistemology of computational biology: Mathematical models and experimental prediction as the basis of their validity," *Biological Systems*, vol. 14, no. 1, pp. 65–90, 2006.

[3] I. Shmulevich and E. R. Dougherty, *Genomic Signal Processing.* Princeton: Princeton University Press, 2001.

[4] E. R. Dougherty, "On the epistemological crisis in genomics," *Current Genomics*, vol. 9, no. 2, pp. 69–79, 2008.

[5] O. Persson, M. Krogh, L. H. Saal, E. Englund, J. Liu, R. Parsons, N. Mandahl, A. Borg, B. Widegren, and L. G. Salford, "Microarray analysis of gliomas reveals chromosomal position-associated gene expression patterns and identifies potential immunotherapy targets," *J. Neurooncol.*, vol. 85, pp. 11–24, 2007.

[6] A. Guirguis, E. Elishaev, S. H. Oh, G. C. Tseng, K. Zorn, and J. A. DeLoia, "Use of gene expression profiles to stage concurrent endometrioid tumors of the endometrium and ovary," *Gynecol. Oncol.*, vol. 108, pp. 370–376, 2008.

[7] T. J. Kim, J. J. Choi, W. Y. Kim, C. H. Choi, J. W. Lee, D. S. Bae, D. S. Son, J. Kim, B. K. Park, G. Ahn, E. Y. Cho, and B. G. Kim, "Gene expression profiling for the prediction of lymph node metastasis in patients with cervical cancer," *Cancer Sci.*, vol. 99, pp. 31–38, 2008.

[8] B. S. Stolf, M. M. Santos, D. F. Simao, J. P. Diaz, E. B. Cristo, R. Hirata, M. P. Curado, E. J. Neves, L. P. Kowalski, and A. F. Carvalho, "Class distinction between follicular adenomas and follicular carcinomas of the thyroid gland on the basis of their signature expression," *Cancer*, vol. 106, pp. 1891–1900, 2006.

[9] A. Barrier, A. Lemoine, P. Y. Boelle, C. Tse, D. Brault, F. Chiappini, J. Breittschneider, F. Lacaine, S. Houry, M. Huguier, M. J. Van der Laan, T. Speed, B. Debuire, A. Flahault, and S. Dudoit, "Colon cancer prognosis prediction by gene expression profiling," *Oncogene*, vol. 24, pp. 6155–6164, 2005.

[10] F. De Smet, N. L. Pochet, K. Engelen, T. Van Gorp, P. Van Hummelen, K. Marchal, F. Amant, D. Timmerman, B. L. De Moor, and I. B. Vergote, "Predicting the clinical behavior of ovarian cancer from gene expression profiles," *Int. J. Gynecol. Cancer*, vol. 16 Suppl. 1, pp. 147–151, 2006.

[11] A. Marchet, S. Mocellin, C. Belluco, A. Ambrosi, F. DeMarchi, E. Mammano, M. Digito, A. Leon, A. D'Arrigo, M. Lise, and D. Nitti, "Gene expression profile of primary gastric cancer: towards the prediction of lymph node status," *Ann. Surg. Oncol.*, vol. 14, pp. 1058–1064, 2007.

[12] E. Karlsson, U. Delle, A. Danielsson, B. Olsson, F. Abel, P. Karlsson, and K. Helou, "Gene expression variation to predict 10-year survival in lymph-node-negative breast cancer," *BMC Cancer*, vol. 8, p. 254, 2008.

[13] M. Heuser, L. U. Wingen, D. Steinemann, G. Cario, N. von Neuhoff, M. Tauscher, L. Bullinger, J. Krauter, G. Heil, H. Dhner, B. Schlegelberger, and A. Ganser, "Gene-expression profiles and their association with drug resistance in adult acute myeloid leukemia," *Haematologica*, vol. 90, pp. 1484–1492, 2005.

[14] M. La, E. H. Ahn, G. E. Mercado, S. Chuai, M. Edgar, B. R. Pawel, A. Olshen, F. G. Barr, and M. Ladanyi, "Global gene expression profiling of pax-fkhr fusion-positive alveolar and pax-fkhr fusion-negative embryonal rhabdomyosarcomas," *J. Pathol.*, vol. 212, pp. 143–151, 2007.

[15] A. F. Ziober, K. R. Patel, F. Alawi, P. Gimotty, R. S. Weber, M. M. Feldman, A. A. Chalian, G. S. Weinstein, J. Hunt, and B. L. Ziober, "Identification of a gene signature for rapid screening of oral squamous cell carcinoma," *Clin. Cancer Res.*, vol. 12, pp. 5960–5971, 2006.

[16] G. N. Fuller, C. Mircean, I. Tabus, E. Taylor, R. Sawaya, J. M. Bruner, I. Shmulevich, and W. Zhang, "Molecular voting for glioma classification reflecting heterogeneity in the continuum of cancer progression," *Oncol. Rep.*, vol. 14, pp. 651–656, 2005.

[17] A. Barrier, F. Roser, P. Y. Bolle, B. Franc, C. Tse, D. Brault, F. Lacaine, S. Houry, P. Callard, C. Penna, B. Debuire, A. Flahault, S. Dudoit, and A. Lemoine, "Prognosis of stage ii colon cancer by non-neoplastic mucosa gene expression profiling," *Oncogene*, vol. 26, pp. 2642–2648, 2007.

[18] M. Hills, "Allocation rules and their error rates," *J. Royal Statistical Society. Series B, Methodological*, vol. 28, no. 1, pp. 1–31, 1966.

[19] D. Foley, "Considerations of sample and feature size," *IEEE Trans. Inf. Theory*, vol. 18, no. 5, pp. 618–626, 1972.

[20] M. J. Sorum, "Estimating the expected probability of misclassification for a rule based on the linear discriminant function: univariate normal case," *Technometrics*, vol. 15, pp. 329–339, 1973.

[21] G. J. McLachlan, "An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis," *Australian J. Statistics*, vol. 15, pp. 210–214, 1973.

[22] M. Moran, "On the expectation of errors of allocation associated with a linear discriminant function," *Biometrika*, vol. 62, no. 1, pp. 141–148, 1975.

[23] M. Goldstein and E. Wolf, "On the problem of bias in multinomial classification," *Biometrics*, vol. 33, pp. 325–31, June 1977.

[24] A. Davison and P. Hall, "On the bias and variability of bootstrap and crossvalidation estimates of error rates in discrimination problems," *Biometrica*, vol. 79, pp. 274–284, 1992.

[25] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, "On the joint sampling distribution between the actual classification error and the resubstitution and leave-one-out error estimators for linear classifiers," *IEEE Trans. Inf. Theory*, vol. 56, no. 2, pp. 784–804, 2010.

[26] U. Braga-Neto and E. R. Dougherty, "Exact performance of error estimators for discrete classifiers," *Pattern Recogn.*, vol. 38, no. 11, pp. 1799–1814, November 2005.

[27] ——, "Exact correlation between actual and estimated errors in discrete classification," *Pattern Recogn. Letters*, vol. 31, no. 5, pp. 407–412, April 2010.

[28] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, "On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers," *Pattern Recogn.*, vol. 42, no. 11, pp. 2705–2723, November 2009.

[29] ——, “Analytic study of performance of error estimators for linear discriminant analysis,” *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4238–4255, Sep. 2011.

[30] F. Wyman, D. Young, and D. Turner, “A comparison of asymptotic error rate expansions for the sample linear discriminant function,” *Pattern Recogn.*, vol. 23, pp. 775–783, 1990.

[31] V. Pikelis, “Comparison of methods of computing the expected classification errors,” *Automatic Remote Control*, vol. 5, pp. 59–63, 1976.

[32] E. R. Dougherty, A. Zollanvari, and U. M. Braga-Neto, “The illusion of distribution-free small-sample classification in genomics,” *Current Genomics*, vol. 12, no. 5, pp. 333–341, August 2011.

[33] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition.* New York: Springer-Verlag, 1996.

[34] L. Devroye and T. Wagner, “Distribution-free inequalities for the deleted and hold-out error estimates,” *IEEE Trans. Inf. Theory*, vol. 25, no. 2, pp. 202–207, 1979.

[35] H. V. Poor, “On robust Wiener filtering,” *IEEE Trans. Automatic Control*, vol. 25, no. 4, pp. 531–536, 1980.

[36] K. S. Vastola and H. V. Poor, “Robust Wiener-Kolmogorov theory,” *IEEE Trans. Inf. Theory*, vol. 30, no. 2, pp. 316–327, 1984.

[37] S. Verdu and H. V. Poor, “On minimax robustness: a general approach and applications,” *IEEE Trans. Inf. Theory*, vol. 30, no. 2, pp. 328–340, 1984.

[38] A. M. Grigoryan and E. R. Dougherty, “Bayesian robust optimal linear filters,” *Signal Processing*, vol. 81, no. 12, pp. 2503 – 2521, December 2001.

[39] R. Pal, E. R. Dougherty, and A. Datta, "Robust intervention in probabilistic boolean networks," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1280 – 1294, March 2008.

[40] R. Pal, A. Datta, and E. R. Dougherty, "Bayesian robustness in the control of gene regulatory networks," *IEEE Trans. Signal Process.*, vol. 57, no. 9, pp. 3667–3678, 2009.

[41] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Operations Research*, vol. 53, no. 5, pp. 780–798, 2005.

[42] A. Wald, "On a statistical problem arising in the classificaion of an individual into one of two groups," *Annals Math. Stat.*, vol. 15, no. 2, pp. 145–162, June 1944.

[43] R. Sitgreaves, "Some results on the distribution of the w-classification statistic," in *Studies in Item Analysis and Prediction*, January ed., H. Solomon, Ed. Stanford University Press, 1961, pp. 241–251.

[44] L. A. Dalton and E. R. Dougherty, "Bayesian minimum mean-square error estimation for classification error–Part I: Definition and the Bayesian MMSE error estimator for discrete classification," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 115–129, January 2011.

[45] ——, "Bayesian minimum mean-square error estimation for classification error–Part II: The Bayesian MMSE error estimator for linear classification of Gaussian distributions," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 130–144, January 2011.

[46] C. Sima and E. R. Dougherty, "Optimal convex error estimators for classification," *Pattern Recogn.*, vol. 39, no. 6, pp. 1763–1780, 2006.

[47] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[48] ——, "Number of pattern classifier design samples per class," *IEEE Trans. Inf. Theory*, vol. 15, no. 5, pp. 615–618, 1969.

[49] M. J. Sorum, "Estimating the probability of misclassification," Ph.D. dissertation, University of Minnesota, Minneapolis, December 1968.

[50] S. Geisser, "Estimation associated with linear discriminants," *Annals Math. Stat.*, vol. 38, no. 3, pp. 807–817, June 1967.

[51] J. K. Martin and D. S. Hirschberg, "Small sample statistics for classification error rates II: Confidence intervals and significance tests," Technical Report No. 96-22, Irvine, CA:University of California, 1996.

[52] Q. Xu, J. Hua, U. Braga-Neto, Z. Xiong, E. Suh, and E. R. Dougherty, "Confidence intervals for the true classification error conditioned on the estimated error," *Technology in Cancer Research and Treatment*, vol. 5, no. 6, pp. 579–589, December 2006.

[53] L. A. Dalton and E. R. Dougherty, "Application of the Bayesian MMSE estimator for classification error to gene expression microarray data," *Bioinformatics*, vol. 27, no. 13, pp. 1822–1831, 2011.

[54] ——, "Optimal MSE calibration of classifier error estimators under Bayesian models," *Pattern Recogn.*, vol. 45, no. 6, pp. 2308–2320, June 2012.

[55] ——, “Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error–Part I: Representation,” *IEEE Trans. Signal Process.*, doi:10.1109/TSP.2012.2184101, in press, 2012.

[56] ——, “Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error–Part II: Consistency and performance analysis,” *IEEE Trans. Signal Process.*, doi:10.1109/TSP.2012.2184102, in press, 2012.

[57] D. Hand, “Classifier technology and the illusion of progress,” *Statistic. Sci.*, vol. 21, pp. 1–14, 2006.

[58] S. Attoor and E. Dougherty, “Classifier performance as a function of distributional complexity,” *Pattern Recogn.*, vol. 37, no. 8, pp. 1629–1640, 2004.

[59] R. Fisher, “The use of multiple measurements in taxonomic problems,” *Ann. Eugen.*, vol. 7, pp. 179–188, 1936.

[60] T. Anderson, “Classification by multivariate analysis,” *Psychometrika*, vol. 16, pp. 31–50, 1951.

[61] S. Geisser, “Posterior odds for multivaraiate normal classifications,” *J. Royal Statistical Society. Series B*, vol. 26, pp. 69–76, 1964.

[62] T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, 1967.

[63] A. Lunts and V. Brailovsky, “Evaluation of attributes obtained in statistical decision rules,” *Engineering Cybernetics*, vol. 3, pp. 98–109, 1967.

[64] M. Stone, “Cross-validatory choice and assessment of statistical predictions,” *J. Royal Statistical Society*, vol. 36, pp. 111–147, 1974.

[65] N. Glick, "Additive estimators for probabilites of correct classification," *Pattern Recogn.*, vol. 10, no. 3, pp. 211–222, January 1978.

[66] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification," *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[67] B. Hanczar, J. Hua, and E. R. Dougherty, "Decorrelation of the true and estimated classifier errors in high-dimensional settings," *EURASIP J. Bioinformatics and Systems Biology*, vol. 2007, January 2007, article ID 38473, 12 pages.

[68] E. R. Dougherty, C. Sima, J. Hua, B. Hanczar, and U. M. Braga-Neto, "Performance of error estimators for classification," *Current Bioinformatics*, vol. 5, no. 1, pp. 53–67, March 2010.

[69] B. Efron, "Bootstrap methods: another look at the jackknife," *Annals Stat.*, vol. 7, pp. 1–26, 1979.

[70] ——, "Estimating the error rate of a prediction rule: improvement on cross validation," *J. American Statistical Assoc.*, vol. 78, pp. 316–331, 1983.

[71] U. Braga-Neto and E. R. Dougherty, "Bolstered error estimation," *Pattern Recogn.*, vol. 37, no. 6, pp. 1267–1281, June 2004.

[72] E. T. Jaynes, "Prior probabilities," *IEEE Trans. Systems Science and Cybernetics*, vol. 4, no. 3, pp. 227–241, September 1968.

[73] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, 4th ed. Pacific Grove, CA: Brooks/Cole, 1995.

[74] A. P. Dawid, M. Stone, and J. V. Zidek, "Marginalization paradoxes in Bayesian and structural inference (with discussion)," *J. Royal Statistical Society. Series B, Methodological*, vol. 35, no. 2, pp. 189–233, 1973.

[75] E. T. Jaynes, *Probability Theory: The Logic of Science.* Cambridge, UK: Cambridge University Press, 2003.

[76] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag, 1985.

[77] M. H. DeGroot, *Optimal Statistical Decisions.* New York: McGraw-Hill, 1970.

[78] H. Akaike, "The interpretation of improper prior distributions as limits of data dependent proper prior distributions," *J. Royal Statistical Society. Series B, Methodological*, vol. 42, no. 1, pp. 46–52, 1980.

[79] N. Glick, "Sample-based multinomial classification," *Biometrics*, vol. 29, no. 2, pp. 241–256, 1973.

[80] M. Goldstein and W. R. Dillon, *Discrete Discriminant Analysis.* New York: Wiley, 1978.

[81] M. Hills, "Discrimination and allocation with discrete data," *J. Royal Statistical Society. Series C, Applied Statistics*, vol. 16, no. 3, pp. 237–250, 1967.

[82] U. Braga-Neto, "Classification and error estimation for discrete data," *Current Genomics*, vol. 10, no. 7, pp. 446–462, November 2009.

[83] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2004.

[84] V. P. Kuznetsov, “Stable detection when the signal and spectrum of normal noise are inaccurately known,” *Telecommunications and Radio Engineering*, vol. 30-31, pp. 58–64, 1976.

[85] S. A. Kassam and T. I. Lim, “Robust wiener filters,” *J. Franklin Institute*, vol. 304, no. 415, pp. 171–185, October/November 1977.

[86] A. M. Grigoryan and E. R. Dougherty, “Design and analysis of robust optimal binary filters in the context of a prior distribution for the states of nature,” *Mathematical Imaging and Vision*, vol. 11, no. 3, pp. 239–254, December 1999.

[87] E. R. Dougherty, J. Hua, Z. Xiong, and Y. Chen, “Optimal robust classifiers,” *Pattern Recogn.*, vol. 38, no. 10, pp. 1520–1532, 2005.

[88] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory.* Cambridge, MA: MIT Press, 1961.

[89] P. S. de Laplace, *Théorie Analytique des Probabilitiés.* Paris: Courceir, 1812.

[90] H. Jeffreys, “An invariant form for the prior probability in estimation problems,” *Proc. Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, September 1946.

[91] ——, *Theory of Probability.* London: Oxford University Press, 1961.

[92] A. M. Mathai and H. J. Haubold, *Special Functions for Applied Scientists.* New York: Springer, 2008.

[93] A. O’Hagan and J. Forster, *Kendalls Advanced Theory of Statistics, Volume 2B: Bayesian Inference*, 2nd ed. London: Hodder Arnold, 2004.

[94] J. T. K. Kanti V. Mardia and J. M. Bibby, *Multivariate Analysis.* London: Academic Press, 1979.

[95] S. F. Arnold, *The Theory of Linear Models and Multivariate Analysis.* New York: John Wiley & Sons, 1981.

[96] K. E. Muller and P. W. Stewart, *Linear model theory: univariate, multivariate, and mixed models.* Hoboken, NJ: John Wiley & Sons, 2006.

[97] N. L. Johnson, "Systems of frequency curves generated by methods of translation," *Biometrika*, vol. 36, no. 1-2, pp. 149–176, June 1949.

[98] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, 2nd ed., ser. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons, 1994, vol. 1.

[99] M. J. van de Vijver, Y. D. He, L. J. van 't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, "A gene-expression signature as a predictor of survival in breast cancer," *New England J. Medicine*, vol. 347, no. 25, pp. 1999–2009, December 2002.

[100] J. W. Craig, "A new, simple and exact result for calculating the probability of error for two-dimensional signal constellations," in *Proc. Military Communications Conference (MILCOM '91)*, vol. 2, McLean, VA, November 1991, pp. 571–575.

[101] L. J. Slater, *Ganeralized Hypergeometric Functions.* Cambridge, UK: Cambridge University Press, 1966.

[102] D. A. Freedman, “On the asymptotic behavior of Bayes’ estimates in the discrete case,” *Annals Math. Stat.*, vol. 34, no. 4, pp. 1386–1403, December 1963.

[103] P. Diaconis and D. Freedman, “On the consistency of Bayes estimates,” *Annals Stat.*, vol. 14, no. 1, pp. 1–26, March 1986.

[104] M. E. Johnson, *Multivariate Statistical Simulation*, ser. Wiley Series in Applied Probability and Statistics. New York: John Wiley and Sons, 1987.

[105] D. C. Hoyle, M. Rattray, R. Jupp, and A. Brass, “Making sense of microarray data distributions,” *Bioinformatics*, vol. 18, no. 4, pp. 576–584, 2002.

[106] R. Autio, S. Kilpinen, M. Saarela, O. Kallioniemi, S. Hautaniemi, and J. Astola, “Comparison of affymetrix data normalization methods using 6,926 experiments across five array generations,” *BMC Bioinformatics*, vol. 10, supplement 1, article S24, 2009.

[107] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality (complete samples),” *Biometrika*, vol. 3, no. 52, pp. 591–611, 1965.

[108] D. B. Rowe, *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing.* Boca Raton, FL: Chapman & Hall/CRC, 2003.

[109] J. Hua, W. D. Tembe, and E. R. Dougherty, “Performance of feature-selection methods in the classification of high-dimension data,” *Pattern Recogn.*, vol. 42, pp. 409–424, 2009.

[110] C. Sima and E. R. Dougherty, “What should be expected from feature selection in small-sample settings,” *Bioinformatics*, vol. 22, no. 19, pp. 2430–2436, 2006.

[111] E. R. Dougherty, J. Hua, and C. Sima, “Performance of feature selection methods,” *Current Genomics*, vol. 10, no. 6, pp. 365–374, 2009.

[112] J. A. Villasenor Alvaa and E. G. Estradaa, “A generalization of shapiro-wilk’s test for multivariate normality,” *Communications in Statistics - Theory and Methods*, vol. 38, no. 11, p. 1870 1883, 2009.

[113] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, “Optimal number of features as a function of sample size for various classification rules,” *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, 2005.

[114] M. Loève, *Probability theory, Volume 2*, 4th ed., ser. Graduate texts in mathematics. New York: Springer-Verlag, 1978.

[115] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, “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.*, vol. 45, no. 2, pp. 908–917, February 2012.

VITA

Lori Anne Dalton received the B.S. and M.S. degrees in electrical engineering at Texas A&M University, College Station, in May 2001 and December 2002, respectively, specializing in wireless communications. She received the Ph.D. degree in electrical engineering at Texas A&M University in May 2012 while working in the Genomic Signal Processing Lab. Her current research interests include pattern recognition, classification, clustering, estimation, optimization and signal processing, with emphasis in bioinformatics and systems biology applications.

She was awarded the NSF Graduate Research Fellowship in 2001, the Association of Former Students Distinguished Graduate Student Masters Research Award in 2003, the Philanthropic Educational Organization (P.E.O.) Scholar Award in 2010 and the Second Place Top Oral Presentation by Students at MCBIOS 2011.

Dr. Dalton may be reached at the Department of Electrical and Computer Engineering, c/o Dr. Edward R. Dougherty, Texas A&M University, 111D Zachry Engineering Bldg., College Station, TX 77843. Her email is ldalton@tamu.edu.