Dalton and Yousefi *EURASIP Journal on Bioinformatics and Systems Biology* (2015) 2015:8
DOI 10.1186/s13637-015-0028-3

EURASIP Journal on Bioinformatics and Systems Biology
a SpringerOpen Journal

RESEARCH Open Access

On optimal Bayesian classification and risk estimation under multiple classes

Lori A. Dalton[1,2*] and Mohammadmahdi R. Yousefi[1]

Abstract

A recently proposed optimal Bayesian classification paradigm addresses optimal error rate analysis for small-sample discrimination, including optimal classifiers, optimal error estimators, and error estimation analysis tools with respect to the probability of misclassification under binary classes. Here, we address multi-class problems and optimal expected risk with respect to a given risk function, which are common settings in bioinformatics. We present Bayesian risk estimators (BRE) under arbitrary classifiers, the mean-square error (MSE) of arbitrary risk estimators under arbitrary classifiers, and optimal Bayesian risk classifiers (OBRC). We provide analytic expressions for these tools under several discrete and Gaussian models and present a new methodology to approximate the BRE and MSE when analytic expressions are not available. Of particular note, we present analytic forms for the MSE under Gaussian models with homoscedastic covariances, which are new even in binary classification.

Keywords: Risk estimation; Multi-class classification; Bayesian estimation; Genomics; Minimum mean-square error; Small samples

1 Introduction

Classification in biomedicine is often constrained by small samples so that understanding properties of the error rate is critical to ensure the scientific validity of a designed classifier. While classifier performance is typically evaluated by employing distribution-free training-data error estimators such as cross-validation, leave-one-out, or bootstrap, a number of studies have demonstrated that these methods are highly problematic in small-sample settings [1]. Under real data and even under simple synthetic Gaussian models, cross-validation has been shown to suffer from a large variance [2] and often has nearly zero correlation, or even negative correlation, with the true error [3, 4]. Among other problems, this directly leads to severely optimistic reporting biases when selecting the best results among several datasets [5] or when selecting the best classification rule among several candidates [6] and difficulties with performance reproducibility [7].

Furthermore, there are typically no accuracy guarantees for error estimators when applied under small samples. *Distribution-free* bounds on the *mean-square error* (MSE) or its square root, the *root-mean-square* (RMS), of an error estimator with respect to the true error rate are typically either unavailable or unhelpful under small samples. Consider leave-one-out error estimation for a discrete histogram rule that breaks ties with class 0. The following is a distribution-free RMS bound [8]:

$$\mathrm{RMS}(\widehat{\varepsilon}_{\mathrm{loo}}(\mathcal{S}) \mid \theta) \leq \sqrt{\frac{1+6/e}{n} + \frac{6}{\sqrt{\pi\,(n-1)}}}, \qquad (1)$$

where $\mathcal{S}$ is a random sample, θ is a feature-label distribution, and n is the sample size. To guarantee an RMS less than 0.5 for all distributions, this bound indicates that a sample size of at least $n = 209$ would be required. Typically, the error of a classifier should be between 0 and 0.5 so that an RMS of 0.5 is trivially guaranteed.

Rather than a distribution-free approach, recent work takes a Bayesian approach to address these problems. The idea is to assume the true distributions characterizing classes in the population are members of an uncertainty class of models. We also assume that members of the uncertainty class are weighted by a *prior* distribution,

*Correspondence: dalton@ece.osu.edu
[1]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA
[2]Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA

and after observing a sample, we update the prior to a *posterior* distribution. For a given classifier we may find an optimal MSE error estimator, called a *Bayesian error estimator* (BEE) [9, 10] and evaluate the MSE of any arbitrary error estimator [11, 12]. These quantities are found by conditioning on the sample in hand and averaging with respect to the unknown population distribution via the posterior, rather than by conditioning on the distribution and averaging over random samples as in (1). Not only does the Bayesian framework supply more powerful error estimators, but the sample-conditioned MSE allows us to evaluate the accuracy of error estimation. The Bayesian framework also facilitates *optimal Bayesian classification* (OBC), which provides decision boundaries to minimize the BEE [13, 14]. In this way, the Bayesian framework can be used to optimize both error estimation and classification.

Classifier design and analysis in the Bayesian framework have previously been developed for binary classification with respect to the probability of misclassification. However, it is common in small-sample classification problems to be faced with classification under multiple classes and for different types of error to be associated with different levels of risk or loss. A few classical classification algorithms naturally permit multiple classes and arbitrary loss functions; for example, a *plug-in rule* takes the functional form for an optimal Bayes decision rule under a given modeling assumption and substitutes sample estimates of model parameters in place of the true parameters. This can be done with linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) for multiple classes with arbitrary loss functions, which essentially assume that the underlying class-conditional densities are Gaussian with equal or unequal covariances, respectively. Most training-data error estimation methods, for instance, cross-validation, can also be generalized to handle multiple classes and arbitrary loss functions. However, it is expected that the same difficulties encountered under binary classes with simple zero-one loss functions (where the expected risk reduces to the probability of misclassification) will carry over to the more general setting, as they have in ROC curve estimation [15].

Support vector machines (SVM) are inherently binary but can be adapted to incorporate penalties that influence risk by implementing slack terms or applying a shrinkage or robustifying objective function [16, 17]. It is also common to construct multi-class classifiers from binary classifiers using the popular "one-versus-all" or "all-versus-all" strategies [18]. The former method builds several binary classifiers by discriminating one class, in turn, against all others, and at a given test point reports the class corresponding to the highest classification score. The latter discriminates between each combination of pairs of classes and reports a majority vote. However, it is unclear how one may assess the precise effect of these adaptations on the expected risk.

We are thus motivated to generalize the BEE, sample-conditioned MSE, and OBC to treat multiple classes with arbitrary loss functions. We will present analogous concepts of *Bayesian risk estimation* (BRE), the sample-conditioned MSE for risk estimators, and *optimal Bayesian risk classification* (OBRC). We will show that the BRE and OBRC can be represented in the same form as the expected risk and Bayes decision rule with unknown true densities replaced by *effective densities*. This approach is distinct from the simple plug-in rule discussed earlier, since the form of the effective densities may not be the same as the individual densities represented in the uncertainty class. We will also develop an interpretation of the conditional MSE based on an *effective joint density*, which is new even under binary classes with a zero-one loss function.

Furthermore, we will provide analytic solutions under several models: discrete spaces with Dirichlet priors (discrete models) and Gaussian distributions with known, independent scaled identity, independent arbitrary, homoscedastic scaled identity, and homoscedastic arbitrary covariance models, all with conjugate priors (Gaussian models). We provide expressions for the BRE and conditional MSE for arbitrary classification in the discrete model and binary linear classification in the Gaussian model. The analytic form that we provide for the MSE of arbitrary error estimators under homoscedastic models is completely new without an analog in prior work under binary classification and zero-one loss. For models in which an analytic form for the BRE and conditional MSE are unavailable, for instance, under multi-class or non-linear classification in the Gaussian model, we also discuss efficient methods to approximate these quantities. In particular, we present a new computationally efficient method to approximate the conditional MSE based on the effective joint density.

2 Notation

We denote random quantities with capital letters, e.g., Y; realizations of random variables with lowercase letters, e.g., y; and vectors in bold, e.g., $\mathbf{X}$ and $\mathbf{x}$. Matrices will generally be in bold upper case, e.g., $\mathbf{S}$. Spaces will be denoted by a stylized font, e.g., $\mathcal{X}$. Distributions with conditioning will be made clear through the function arguments; for instance, we write the distribution of $\mathbf{X}$ given Y as $f(\mathbf{x} \mid y)$. The probability space of expectations will be made clear by denoting random quantities in the expectation and conditioning, e.g., the expectation of Y conditioned on the random variable $\mathbf{X}$ and the event $\mathbf{C} = \mathbf{c}$ is denoted by $\mathrm{E}[Y \mid \mathbf{X}, \mathbf{c}]$. When the region of integration in an integral is omitted then this region is the whole space. Any exceptions in notation will be defined throughout.

3 Bayes decision theory

We next review concepts from classical Bayes decision theory. Consider a classification problem in which we are to predict one of M classes, $y = 0, \ldots, M-1$, from a sample drawn in feature space $\mathcal{X}$. Let $\mathbf{X}$ and Y denote a random feature vector and its corresponding random label. Let $f(y \mid \mathbf{c})$ be the probability mass function of Y, parameterized by a vector $\mathbf{c}$, and for each y, let $f(\mathbf{x} \mid y, \theta_y)$ be the *class-y-conditional density* of $\mathbf{X}$, parameterized by a vector θ_y. The full feature-label distribution is parameterized by $\mathbf{c}$ and $\theta = \{\theta_0, \ldots, \theta_{M-1}\}$.

Let $\lambda(i, y)$ be a loss function quantifying a penalty in predicting label i when the true label is y. The *conditional risk* in predicting label i for a given point, $\mathbf{x}$, is defined as

$$\begin{aligned} R(i, \mathbf{x}, \mathbf{c}, \theta) &= \mathrm{E}[\, \lambda(i, Y) \mid \mathbf{x}, \mathbf{c}, \theta] \\ &= \sum_{y=0}^{M-1} \lambda(i, y) f(y \mid \mathbf{x}, \mathbf{c}, \theta) \\ &= \frac{\sum_{y=0}^{M-1} \lambda(i, y) f(y \mid \mathbf{c}) f(\mathbf{x} \mid y, \theta_y)}{\sum_{y=0}^{M-1} f(y \mid \mathbf{c}) f(\mathbf{x} \mid y, \theta_y)}. \end{aligned} \tag{2}$$

The *expected risk* of a given classification rule, $\psi : \mathcal{X} \to \{0, \ldots, M-1\}$, is given by

$$\begin{aligned} R(\psi, \mathbf{c}, \theta) &= \mathrm{E}[\, R(\psi(\mathbf{X}), \mathbf{X}, \mathbf{c}, \theta) \mid \mathbf{c}, \theta] \\ &= \sum_{y=0}^{M-1} \sum_{i=0}^{M-1} \lambda(i, y) f(y \mid \mathbf{c}) \varepsilon^{i,y}(\psi, \theta_y), \end{aligned} \tag{3}$$

where

$$\varepsilon^{i,y}(\psi, \theta_y) = \int_{\Gamma_i} f(\mathbf{x} \mid y, \theta_y) d\mathbf{x} = \mathrm{P}(\mathbf{X} \in \Gamma_i \mid y, \theta_y) \tag{4}$$

is the probability that a class-y point will be assigned class i by the classifier ψ, and the $\Gamma_i = \{\mathbf{x} \in \mathcal{X} : \psi(\mathbf{x}) = i\}$ partition the sample space into decision regions.

A *Bayes decision rule* (BDR) minimizes expected risk or, equivalently, the conditional risk at each fixed point $\mathbf{x}$:

$$\psi_{\mathrm{BDR}}(\mathbf{x}) = \arg \min_{i \in \{0, \ldots, M-1\}} R(i, \mathbf{x}, \mathbf{c}, \theta). \tag{5}$$

By convention, we break ties with the lowest index, $i \in \{0, \ldots, M-1\}$, minimizing $R(i, \mathbf{x}, \mathbf{c}, \theta)$.

4 Optimal Bayesian risk classification

In practice, the feature-label distribution is unknown so that we must train a classifier and estimate risk or error with data. The Bayesian framework resolves this by assuming the true feature-label distribution is a member of a parameterized uncertainty class. In particular, assume that $\mathbf{c}$ is the probability mass function of Y, that is, $\mathbf{c} = \{c_0, \ldots, c_{M-1}\} \in \Delta^{M-1}$, where $f(y \mid \mathbf{c}) = c_y$ and Δ^{M-1} is the standard $M-1$ simplex defined by $c_y \in [0, 1]$ for $y \in \{0, \ldots, M-1\}$ and $\sum_{y=0}^{M-1} c_y = 1$. Also assume $\theta_y \in \mathcal{T}_y$ for some parameter space $\mathcal{T}_y$, and $\theta \in \mathcal{T} = \mathcal{T}_0 \times \ldots \times \mathcal{T}_{M-1}$. Let $\mathbf{C}$ and Θ denote random vectors for parameters $\mathbf{c}$ and θ, respectively. Finally, assume $\mathbf{C}$ and Θ are independent prior to observing data and assign *prior* probabilities, $\pi(\mathbf{c})$ and $\pi(\theta)$.

Priors quantify uncertainty we have about the distribution before observing the data. Although non-informative priors may be used as long as the posterior is normalizable, informative priors can supplement the classification problem with information to improve performance when the sample size is small. This is key for problems with limited or expensive data. Under mild regularity conditions, as we observe sample points, this uncertainty converges to a certainty on the true distribution parameters, where more informative priors may lead to faster convergence [12]. For small samples, the performance of Bayesian methods depends heavily on the choice of prior. Performance tends to be modest but more robust with a non-informative or weakly informative prior. Conversely, informative priors offer the potential for great performance improvement, but if the true population distribution is not well represented in the prior, then performance may be poor. This trade-off is acceptable as long as the prior is an accurate reflection of available scientific knowledge so that one is reasonably sure that catastrophic results will not occur. If multiple models are scientifically reasonable but result in different inferences, and if it is not possible to determine which model is best from data or prior knowledge, then the range of inferences must be considered [19]. For the sake of illustration, in simulations, we will utilize either low-information priors or a simple prior construction method for microarray data, although modeling and prior construction remain important problems [20].

Let S be a sample, that is, a realization of n independent labeled points drawn from $\mathcal{X}$. Also let $\mathbf{x}_i^y$ denote the ith sample point in class y and n_y denote the number of sample points observed from class y. Given a sample, the priors are updated to *posterior* densities:

$$f(\mathbf{c}, \theta \mid S) \propto \pi(\mathbf{c}) \pi(\theta) \prod_{y=0}^{M-1} \prod_{i=1}^{n_y} f(\mathbf{x}_i^y, y \mid \mathbf{c}, \theta_y), \tag{6}$$

where the product on the right is the usual *likelihood function*. Since $f(\mathbf{x}_i^y, y \mid \mathbf{c}, \theta_y) = c_y f(\mathbf{x}_i^y \mid y, \theta_y)$, we may write $f(\mathbf{c}, \theta \mid S) = f(\mathbf{c} \mid S) f(\theta \mid S)$, where

$$f(\mathbf{c} \mid S) \propto \pi(\mathbf{c}) \prod_{y=0}^{M-1} (c_y)^{n_y} \tag{7}$$

and

$$f(\theta \mid S) \propto \pi(\theta) \prod_{y=0}^{M-1} \prod_{i=1}^{n_y} f(\mathbf{x}_i^y \mid y, \theta_y) \tag{8}$$

are marginal posteriors of $\mathbf{C}$ and Θ. Thus, independence between $\mathbf{C}$ and Θ is preserved in the posterior. Constants of proportionality are found by normalizing the integral of posteriors to 1. When the prior density is proper, this all follows from Bayes' rule; otherwise, (7) and (8) are taken as definitions, where we require posteriors to be proper.

$f(\mathbf{c} \mid S)$ depends on the prior and sampling method used. For instance, if $\mathbf{C}$ is known, then $\pi(\mathbf{c})$ and $f(\mathbf{c} \mid S)$ are both point masses at the known value of $\mathbf{C}$. Under separate sampling, in which the number of sample points in each class is fixed to an arbitrary value prior to sampling, $f(\mathbf{c} \mid S) = \pi(\mathbf{c})$. Under random sampling, the sample size is fixed at n and the number of points observed from each class is determined by independent draws from the feature-label distribution. Given a Dirichlet prior on $\mathbf{C}$ with hyperparameters $\alpha = \{\alpha_0, \ldots, \alpha_{M-1}\}$, a special case being $\alpha_0 = \ldots = \alpha_{M-1} = 1$ for a uniform distribution on Δ^{M-1}, then under random sampling the posterior on $\mathbf{C}$ is still Dirichlet with hyperparameters $\alpha_y^* = \alpha_y + n_y$. Defining $\alpha_+^* = \sum_{i=0}^{M-1} \alpha_i^*$, we also have for $y \neq z$,

$$\mathrm{E}\left[C_y \mid S\right] = \frac{\alpha_y^*}{\alpha_+^*}, \tag{9}$$

$$\mathrm{E}\left[C_y^2 \mid S\right] = \frac{\alpha_y^* \left(1 + \alpha_y^*\right)}{\alpha_+^* \left(1 + \alpha_+^*\right)}, \tag{10}$$

$$\mathrm{E}\left[C_y C_z \mid S\right] = \frac{\alpha_y^* \alpha_z^*}{\alpha_+^* \left(1 + \alpha_+^*\right)}. \tag{11}$$

4.1 Bayesian risk estimation

We define the BRE to be the minimum mean-square error (MMSE) estimate of the expected risk or, equivalently, the conditional expectation of the expected risk given observations. Given a sample, S, and a classifier, ψ, that is not informed by θ, thanks to posterior independence between $\mathbf{C}$ and Θ, the BRE is given by,

$$\begin{aligned}\widehat{R}(\psi, S) &= \mathrm{E}[\, R(\psi, \mathbf{C}, \Theta) \mid S] \\ &= \sum_{y=0}^{M-1} \sum_{i=0}^{M-1} \lambda(i, y) \mathrm{E}[f(y \mid \mathbf{C}) \mid S]\, \mathrm{E}[\, \varepsilon^{i,y}\, (\psi, \Theta)\ \mid S]\,.\end{aligned} \tag{12}$$

If we assume that $\{\mathbf{X}, Y\}$ and $\mathcal{S}$ are independent given $\mathbf{C}$ and Θ, then

$$\begin{aligned} f(y \mid S) &= \int f(y \mid \mathbf{c}) f(\mathbf{c} \mid S) d\mathbf{c} \\ &= \mathrm{E}\left[f(y \mid \mathbf{C}) \mid S\right], \end{aligned} \tag{13}$$

$$\begin{aligned} f\left(\mathbf{x} \mid y, S\right) &= \int f\left(\mathbf{x} \mid y, \theta_y\right) f\left(\theta_y \mid S\right) d\theta_y \\ &= \mathrm{E}[f(\mathbf{x} \mid y, \Theta_y) \mid S]\,. \end{aligned} \tag{14}$$

We may thus write the BRE in (12) as

$$\widehat{R}(\psi, S) = \sum_{y=0}^{M-1} \sum_{i=0}^{M-1} \lambda(i, y) f(y \mid S) \widehat{\varepsilon}^{i,y}(\psi, S), \tag{15}$$

where $\widehat{\varepsilon}^{i,y}(\psi, S) = \mathrm{E}[\, \varepsilon^{i,y}\, (\psi, \Theta)\ \mid S]$ is the posterior probability of assigning a class-y point to class i,

$$\begin{aligned} \widehat{\varepsilon}^{i,y}(\psi, S) &= \mathrm{E}\left[\int_{\Gamma_i} f\left(\mathbf{x} \mid y, \Theta_y\right) d\mathbf{x} \,\middle|\, S\right] \\ &= \int_{\Gamma_i} \mathrm{E}\left[f\left(\mathbf{x} \mid y, \Theta_y\right) \mid S\right] d\mathbf{x} \\ &= \int_{\Gamma_i} f\left(\mathbf{x} \mid y, S\right) d\mathbf{x} \qquad (16) \\ &= \mathrm{P}\left(\mathbf{X} \in \Gamma_i \mid y, S\right). \qquad (17) \end{aligned}$$

The second equality follows from Fubini's theorem, and in the last equality, $\mathbf{X}$ is a random vector drawn from the density in the integrand of (16). We also have $f(y \mid S) = \mathrm{E}\left[C_y \mid S\right]$, which depends on the prior for $\mathbf{C}$ and is easily found, for instance, from (9) under Dirichlet posteriors. Comparing (3) and (15), observe that $f(y \mid S)$ and $f(\mathbf{x} \mid y, S)$ play roles analogous to $f(y \mid \mathbf{c})$ and $f(\mathbf{x} \mid y, \theta_y)$ in Bayes decision theory. We thus call $f\left(\mathbf{x} \mid y, S\right)$ the *effective class-y conditional density* or simply the *effective density*.

Whereas the BRE addresses overall classifier performance across the entire sample space, $\mathcal{X}$, we may also consider classification at a fixed point, $\mathbf{x} \in \mathcal{X}$. We define the *Bayesian conditional risk estimator* (BCRE) for class $i \in \{0, \ldots, M-1\}$ at point $\mathbf{x} \in \mathcal{X}$ to be the MMSE estimate of the conditional risk:

$$\begin{aligned}\widehat{R}(i, \mathbf{x}, S) &= \mathrm{E}[\, R(i, \mathbf{x}, \mathbf{C}, \Theta) \mid S] \\ &= \sum_{y=0}^{M-1} \lambda(i, y) \mathrm{E}[f(y \mid \mathbf{x}, \mathbf{C}, \Theta) \mid S]\,. \end{aligned} \tag{18}$$

Again assuming $\{\mathbf{X}, Y\}$ and $\mathcal{S}$ are independent given $\mathbf{C}$ and Θ, and if we further assume $\mathbf{X}$ is independent from $\mathbf{C}$, Θ, and $\mathcal{S}$, then,

$$\begin{aligned} \mathrm{E}[f(y \mid \mathbf{x}, \mathbf{C}, \Theta) \mid S] &= \int f(y \mid \mathbf{x}, \mathbf{c}, \theta) f(\mathbf{c}, \theta \mid S) d\mathbf{c} d\theta \\ &= \int f(y, \mathbf{c}, \theta \mid \mathbf{x}, S) d\mathbf{c} d\theta \\ &= f(y \mid \mathbf{x}, S). \end{aligned}$$

Applying Bayes' rule,

$$f(y \mid \mathbf{x}, S) = \frac{f(y \mid S) f(\mathbf{x} \mid y, S)}{\sum_{y=0}^{M-1} f(y \mid S) f(\mathbf{x} \mid y, S)}, \tag{19}$$

and applying this to (18), we have

$$\widehat{R}(i, \mathbf{x}, S) = \frac{\sum_{y=0}^{M-1} \lambda(i, y) f(y \mid S) f(\mathbf{x} \mid y, S)}{\sum_{y=0}^{M-1} f(y \mid S) f(\mathbf{x} \mid y, S)}. \tag{20}$$

This is analogous to (2) in Bayes decision theory. Furthermore, given a classifier ψ,

$$\begin{aligned}\mathrm{E}\left[\widehat{R}(\psi(\mathbf{X}),\mathbf{X},S)\,\middle|\,S\right] &= \sum_{i=0}^{M-1}\int_{\Gamma_i}\widehat{R}(i,\mathbf{X},S)f(\mathbf{x}\,|\,S)d\mathbf{x}\\ &= \widehat{R}(\psi,S),\end{aligned}$$

where $f(\mathbf{x}\,|\,S) = \sum_{y=0}^{M-1} f(y\,|\,S)f(\mathbf{x}\,|\,y,S)$ is the marginal distribution of $\mathbf{x}$ given S. Hence, the BRE of ψ is the mean of the BCRE across the sample space.

For binary classification, $\widehat{\varepsilon}^{i,y}(\psi,S)$ has been solved in closed form as components of the BEE for both discrete models under arbitrary classifiers and Gaussian models under linear classifiers, so the BRE with an arbitrary loss function is available in closed form for both of these models. When closed-form solutions for $\widehat{\varepsilon}^{i,y}(\psi,S)$ are not available, from (17), $\widehat{\varepsilon}^{i,y}(\psi,S)$ may be approximated for all i and a given fixed y by drawing a large synthetic sample from $f(\mathbf{x}\,|\,y,S)$ and evaluating the proportion of points assigned class i. The final approximate BRE can be found by plugging the approximate $\widehat{\varepsilon}^{i,y}(\psi,S)$ for each y and i into (15).

A number of practical considerations for BEEs addressed under binary classification naturally carry over to multiple classes, including robustness to false modeling assumptions [9, 10] and a prior calibration method for microarray data analysis using features discarded by feature selection and a method-of-moments approach [21]. Furthermore, classical frequentist consistency holds for BREs on fixed distributions in the parameterized family owing to the convergence of posteriors in both the discrete and Gaussian models [12].

4.2 Optimal Bayesian risk classification

We define the OBRC to minimize the BRE, that is,

$$\psi_{\mathrm{OBRC}} = \arg\inf_{\psi\in\mathcal{C}}\widehat{R}(\psi,S), \tag{21}$$

where $\mathcal{C}$ is a family of classifiers. If $\mathcal{C}$ is the set of all classifiers with measurable decision regions, it can be shown that ψ_{OBRC} exists and is given for any $\mathbf{x}\in\mathcal{X}$ by

$$\psi_{\mathrm{OBRC}}(\mathbf{x}) = \arg\min_{i\in\{0,\ldots,M-1\}}\widehat{R}(i,\mathbf{x},S). \tag{22}$$

Analogously to the relationship between the BRE and expected risk, the OBRC has the same functional form as the BDR with $f(y\,|\,S)$ substituted for the true class probability, $f(y\,|\,\mathbf{c})$, and $f(\mathbf{x}\,|\,y,S)$ substituted for the true density, $f\left(\mathbf{x}\,|\,y,\theta_y\right)$, for all y. Closed-form OBRC are available for any model in which $f(\mathbf{x}\,|\,y,S)$ has been found, including discrete and Gaussian models [13]. A number of important properties also carry over, including invariance to invertible transformations, pointwise convergence to the Bayes classifier, and robustness to false modeling assumptions.

4.3 Sample-conditioned MSE of risk estimation

In a typical small-sample classification scenario, a classifier is trained from data and a risk estimate found for the true risk of this classifier. A key question arises: How close is the risk estimate to the actual risk? A Bayesian approach answers this question with the sample-conditioned MSE of the BRE relative to the true expected risk:

$$\begin{aligned}\mathrm{MSE}(\widehat{R}(\psi,S)\,|\,S) &= \mathrm{E}\left[\left(R(\psi,\mathbf{C},\Theta)-\widehat{R}(\psi,S)\right)^2\,\middle|\,S\right]\\ &= \mathrm{Var}\left(R(\psi,\mathbf{C},\Theta)\,|\,S\right).\end{aligned} \tag{23}$$

This MSE is precisely the quantity that the BRE minimizes, and it quantifies the accuracy of $\widehat{R}$ as an estimator of R, conditioned on the actual sample in hand. Thanks to posterior independence between $\mathbf{C}$ and Θ, it can be decomposed:

$$\begin{aligned}&\mathrm{MSE}(\widehat{R}(\psi,S)\,|\,S)\\ &= \left(\sum_{y=0}^{M-1}\sum_{z=0}^{M-1}\sum_{i=0}^{M-1}\sum_{j=0}^{M-1}\lambda(i,y)\lambda(j,z)\mathrm{E}\left[C_yC_z\,|\,S\right]\right.\\ &\qquad\left.\times\,\mathrm{E}\left[\varepsilon^{i,y}(\psi,\Theta_y)\varepsilon^{j,z}(\psi,\Theta_z)\,|\,S\right]\right)-(\widehat{R}(\psi,S))^2,\end{aligned} \tag{24}$$

where we have applied (3) in (23) and noted $\mathrm{E}\left[f(y\,|\,\mathbf{C})f(z\,|\,\mathbf{C})\,|\,S\right] = \mathrm{E}\left[C_yC_z\,|\,S\right]$. Second-order moments of C_y depend on our prior for $\mathbf{C}$ and can be found, for instance, from (10) and (11) under Dirichlet posteriors. Hence, evaluating the conditional MSE of the BRE boils down to evaluating the BRE itself, $\widehat{R}(\psi,S)$, and evaluating expressions of the form $\mathrm{E}\left[\varepsilon^{i,y}(\psi,\Theta_y)\varepsilon^{j,z}(\psi,\Theta_z)\,|\,S\right]$. Furthermore, if we additionally assume $\Theta_0,\ldots,\Theta_{M-1}$ are pairwise independent, then when $y\neq z$,

$$\mathrm{E}\left[\varepsilon^{i,y}(\psi,\Theta_y)\varepsilon^{j,z}(\psi,\Theta_z)\,|\,S\right] = \widehat{\varepsilon}^{i,y}(\psi,S)\widehat{\varepsilon}^{j,z}(\psi,S), \tag{25}$$

where $\widehat{\varepsilon}^{i,y}(\psi,S)$, given in (16), is a component of the BRE. The conditional MSE of an arbitrary risk estimate, $\widehat{R}_\bullet(\psi,S)$, is also of interest and may be easily found from the BRE and the MSE of the BRE:

$$\begin{aligned}&\mathrm{MSE}(\widehat{R}_\bullet(\psi,S)\,|\,S)\\ &= \mathrm{E}\left[\left(R(\psi,\mathbf{C},\Theta)-\widehat{R}_\bullet(\psi,S)\right)^2\,\middle|\,S\right]\\ &= \mathrm{MSE}(\widehat{R}(\psi,S)\,|\,S)+(\widehat{R}(\psi,S)-\widehat{R}_\bullet(\psi,S))^2.\end{aligned} \tag{26}$$

In this form, the optimality of the BRE is clear.

For binary classification with zero-one loss, the sample-conditioned MSE of the BRE converges to zero almost surely as sample size increases, for both discrete models under arbitrary classifiers and Gaussian models with independent covariances under linear classifiers [12].

Closed-form expressions for the MSE are available in these models. In this work, we extend this to multi-class discrimination under discrete models and binary linear classification under homoscedastic Gaussian models. For cases where closed-form solutions are unavailable, in the next section, we present a method to approximate the MSE.

4.4 Efficient computation

The following new interpretation for $\mathrm{E}\left[\varepsilon^{i,y}(\psi,\Theta_y)\,\varepsilon^{j,z}(\psi,\Theta_z)\mid S\right]$ is useful in both deriving analytic forms for and approximating the MSE. From (4),

$$\begin{aligned}&\mathrm{E}\left[\varepsilon^{i,y}(\psi,\Theta_y)\varepsilon^{j,z}(\psi,\Theta_z)\mid S\right]\\&=\int_{\mathcal{T}}\int_{\Gamma_i} f(\mathbf{x}\mid y,\theta_y)d\mathbf{x}\int_{\Gamma_j} f(\mathbf{w}\mid z,\theta_z)d\mathbf{w}f(\theta\mid S)d\theta\\&=\int_{\Gamma_i}\int_{\Gamma_j}\int_{\mathcal{T}} f(\mathbf{x}\mid y,\theta_y)f(\mathbf{w}\mid z,\theta_z)f(\theta\mid S)d\theta d\mathbf{w}d\mathbf{x},\end{aligned}\tag{27}$$

where we have again applied Fubini's theorem. Further, we may write

$$\begin{aligned}&\mathrm{E}\left[\varepsilon^{i,y}(\psi,\Theta_y)\varepsilon^{j,z}(\psi,\Theta_z)\mid S\right]\\&\qquad=\int_{\Gamma_i}\int_{\Gamma_j} f(\mathbf{x},\mathbf{w}\mid y,z,S)d\mathbf{w}d\mathbf{x}&(28)\\&\qquad=\mathrm{P}\left(\mathbf{X}\in\Gamma_i\cap\mathbf{W}\in\Gamma_j\mid y,z,S\right),&(29)\end{aligned}$$

where $\mathbf{X}$ and $\mathbf{W}$ are random vectors drawn from an *effective joint density*, defined using similar independence assumptions as in (14):

$$f(\mathbf{x},\mathbf{w}\mid y,z,S)=\int f(\mathbf{x}\mid y,\theta_y)f(\mathbf{w}\mid z,\theta_z)f(\theta\mid S)d\theta.\tag{30}$$

The marginal densities of $\mathbf{X}$ and $\mathbf{W}$ under $f(\mathbf{x},\mathbf{w}\mid y,z,S)$ are precisely the effective density, i.e.,

$$\begin{aligned}&\int_{\mathcal{X}} f(\mathbf{x},\mathbf{w}\mid y,z,S)d\mathbf{w}\\&=\int_{\mathcal{X}}\int_{\mathcal{T}} f(\mathbf{x}\mid y,\theta_y)f(\mathbf{w}\mid z,\theta_z)f(\theta\mid S)d\theta d\mathbf{w}\\&=\int_{\mathcal{T}} f(\mathbf{x}\mid y,\theta_y)\int_{\mathcal{X}} f(\mathbf{w}\mid z,\theta_z)d\mathbf{w}f(\theta\mid S)d\theta\\&=\int_{\mathcal{T}_y} f(\mathbf{x}\mid y,\theta_y)f(\theta_y\mid S)d\theta_y\\&=f(\mathbf{x}\mid y,S),\end{aligned}$$

where $f(\theta_y\mid S)$ is the marginal posterior density of Θ_y. Further, we have an *effective conditional density* of $\mathbf{W}$ given $\mathbf{X}$,

$$\begin{aligned}f(\mathbf{w}\mid\mathbf{x},y,z,S)&=\frac{f(\mathbf{x},\mathbf{w}\mid y,z,S)}{f(\mathbf{x}\mid y,S)}\\&=\int f(\mathbf{w}\mid z,\theta_z)\frac{f(\mathbf{x}\mid y,\theta_y)f(\theta\mid S)}{\int f(\mathbf{x}\mid y,\theta'_y)f(\theta'_y\mid S)d\theta'_y}d\theta\\&=\int f(\mathbf{w}\mid z,\theta_z)f(\theta\mid S\cup\{\mathbf{x},y\})d\theta\\&=f(\mathbf{w}\mid z,S\cup\{\mathbf{x},y\}),\end{aligned}$$

where we have used the fact that the fractional term in the integrand of the second equality is of the same form as the posterior defined in (8), updated with a new independent sample point with feature vector $\mathbf{x}$ and class y. Hence, the effective joint density may be easily found, once the effective density is known. Furthermore, from (29), we may approximate $\mathrm{E}\left[\varepsilon^{i,y}(\psi,\Theta_y)\varepsilon^{j,z}(\psi,\Theta_z)\mid S\right]$ by drawing a large synthetic sample from $f(\mathbf{x}\mid y,S)$, drawing a single point, $\mathbf{w}$, from the effective conditional density $f(\mathbf{w}\mid z,S\cup\{\mathbf{x},y\})$ for each $\mathbf{x}$, and evaluating the proportion of pairs, $(\mathbf{x},\mathbf{w})$, for which $\mathbf{x}\in\Gamma_i$ and $\mathbf{w}\in\Gamma_j$. Additionally, since $\mathbf{x}$ is marginally governed by the effective density, from (17) we may approximate $\widehat{\varepsilon}^{i,y}(\psi,S)$ by evaluating the proportion of $\mathbf{x}$ in Γ_i.

Evaluating the OBRC, BRE, and conditional MSE requires obtaining $\mathrm{E}[C_y\mid S]$, $\mathrm{E}[C_y^2\mid S]$ and $\mathrm{E}[C_yC_z\mid S]$ based on the posterior for $\mathbf{C}$ and finding the effective density, $f(\mathbf{x}\mid y,S)$, and the effective joint density, $f(\mathbf{x},\mathbf{w}\mid y,z,S)$, based on the posterior for Θ. At a fixed point, $\mathbf{x}$, one may then evaluate the posterior probability of each class, $f(y\mid\mathbf{x},S)$, from (19) and the BCRE from (20). The OBRC is then found from (22) or, equivalently, by choosing the class, i, that minimizes $\sum_{y=0}^{M-1}\lambda(i,y)\mathrm{E}[C_y\mid S]f(\mathbf{x}\mid y,S)$. For any classifier, the BRE is given by (15) with $\widehat{\varepsilon}^{i,y}(\psi,S)$ given by (16) (or equivalently (17)) using the effective density, $f(\mathbf{x}\mid y,S)$. The MSE of the BRE is then given by (24), where $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,z}(\Theta_z)\mid S\right]$ is given by (25) when $\Theta_0,\ldots,\Theta_{M-1}$ are pairwise independent and $y\neq z$, and $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,z}(\Theta_z)\mid S\right]$ is otherwise found from (28) (or equivalently (29)) using the effective joint density, $f(\mathbf{x},\mathbf{w}\mid y,z,S)$. The MSE of an arbitrary risk estimator can also be found from (26) using the BRE and the MSE for the BRE. We summarize these tools for several discrete and Gaussian models in Appendices 1, 2, and 3 by providing the effective density, the effective joint density (or a related density), $\widehat{\varepsilon}^{i,y}(\psi,S)$, and $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,z}(\Theta_z)\mid S\right]$.

5 Simulation setup and results

In the this section, we examine several synthetic data simulations, where random distributions and samples are generated from a low-information prior, and demonstrate the performance gain and optimality of Bayesian methods

within the Bayesian framework. We also examine performance with informed priors in two real datasets.

5.1 Classification rules

We consider five classification rules: OBRC, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), linear support vector machine (L-SVM), and radial basis function SVM (RBF-SVM). We will implement OBRC under Gaussian models. We used built-in MATLAB functions to implement LDA and QDA. For a collection of binary-labeled training sample points, an SVM classifier finds a maximal margin hyperplane based on a well-behaved optimization objective function and a set of constraints. When the data are not perfectly linearly separable, introduction of slack variables in the optimization procedure leads to *soft margin* classifiers for which mislabeled sample points are allowed. The resulting hyperplane in the feature space is called L-SVM. Alternatively, the underlying feature space can be transformed to a higher dimensional space where the data becomes linearly separable. The equivalent classifier back in the original feature space will generally be non-linear [22, 23]. When the kernel function is a Gaussian radial basis function, we call the corresponding classifier RBF-SVM. We used the package LIBSVM, which, by default, implements a *one-versus-one* approach for multi-class classification [24]. Since SVM classifiers optimize relative to their own objective function (for example, hinge loss), rather than expected risk, we exclude them from our analysis when using a non-zero-one loss function.

For all classification rules, we calculate the true risk defined in (3) and (4). We find the exact value if a formula is available; otherwise, we use a test sample of at least 10,000 points generated from the true feature-label distributions, stratified relative to the true class prior probabilities. This will yield an approximation of the true risk with $\text{RMS} \leq 1/\sqrt{4 \times 10,000} = 0.005$ [8].

5.2 Risk estimation rules

We consider four risk estimation methods: BRE, 10-fold cross-validation (CV), leave-one-out (LOO), and 0.632 bootstrap (boot). When we do not have closed-form formulae for calculating the BRE, we approximate it by drawing a sample of 1,000,000 points from the effective density of each class. In CV, the training data, S, is randomly partitioned into 10 stratified folds, $S^{(i)}$ for $i = 1, 2, \ldots, 10$. Each fold, in turn, is held out of the classifier design step as the test set, and a surrogate classifier is designed on the remaining folds, $S \setminus S^{(i)}$, as the training set. The risk of each surrogate classifier is estimated using $S^{(i)}$. The resulting risk values from all surrogate classifiers are then averaged to get the CV estimate. To reduce "internal variance" arising from random selection of the partitions, we average the CV estimates over 10 repetitions (10 randomly generated partitions over S). If the number of folds equals the sample size, n, then each fold consists of a single point and we get the LOO risk estimation.

Bootstrap risk estimators are calculated using bootstrap samples of size n, where in each bootstrap sample, points are drawn, with replacement, from the original training dataset. A surrogate classifier is designed on the bootstrap sample and its risk estimated using sample points left out of the bootstrap sample. The basic bootstrap estimator is the expectation of this risk with respect to the bootstrap sampling distribution. The expectation is usually approximated by Monte Carlo repetitions (100 in our simulations) over a number of independent bootstrap samples. It is known that this estimate is high biased. To reduce bias, the 0.632 bootstrap reports a linear combination of this estimate, with weight 0.632, and the low-biased resubstitution risk estimate, with weight 0.368 [25–27].

Under linear classification, the sample-conditioned MSE from (24) is found analytically by evaluating $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,y}(\Theta_y) \mid S\right]$ from (52), plugging in the appropriate values for k and γ^2 depending on the covariance model, and $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,z}(\Theta_z) \mid S\right]$ for $z \neq y$ are found via (25) for independent and (53) for homoscedastic covariance models, plugging in appropriate values for k and γ^2. When analytic forms are not available, the sample-conditioned MSE is approximated as follows. In independent covariance models, for each sample point generated to approximate the BRE, we draw a single point from the effective conditional density with $y = z$, giving 1,000,000 sample point pairs to approximate $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,y}(\Theta_y) \mid S\right]$ for each y. In homoscedastic covariance models, to find the BRE, we have 1,000,000 points available from the effective density for each y. We generate an additional $1,000,000 \times (M - 1)$ synthetic points for each y, thus allocating 1,000,000 synthetic points for each combination of y and z. For each of these points, we draw a single point from the effective conditional density of a class-z point given a class-y point. For each y and z, the corresponding 1,000,000 point pairs are used to approximate $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,z}(\Theta_z) \mid S\right]$.

5.3 Synthetic data

In synthetic data simulations, we assume all classes are equally likely and that the data is stratified, giving an equal number of sample points from each class. We further assume Gaussian feature-label distributions. Table 1 lists all models and prior distributions used. We implement both a low number of features ($D = 2$) and a high number of features ($D = 20$), with independent arbitrary, homoscedastic arbitrary, and independent identity covariance priors. Under each type of prior, we consider classification under a non-zero-one loss function for binary classification and a zero-one loss function for multiple classes. For each prior model and a fixed sample size,

Table 1 Synthetic data classification settings and prior models

	D	M	$\nu_0, \ldots, \nu_{M-1}$	$\mathbf{m}_0, \ldots, \mathbf{m}_{M-1}$	κ_y (k_y)	$\frac{\mathbf{S}_y}{k_y-2}$	Prior (cov.)	λ
Model 1	2	2	12, 2	$\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0.5\\0.5\end{bmatrix}$	6 (5)	0.3 I_2	Indep. arbit.	$\begin{bmatrix}0 & 2\\1 & 0\end{bmatrix}$
Model 2	2	2	12, 2	$\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0.5\\0.5\end{bmatrix}$	6 (5)	0.3 I_2	Homo. arbit.	$\begin{bmatrix}0 & 2\\1 & 0\end{bmatrix}$
Model 3	2	5	12, 2, 2, 2, 2	$\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}-1\\-1\end{bmatrix}, \begin{bmatrix}1\\-1\end{bmatrix}, \begin{bmatrix}-1\\1\end{bmatrix}$	6 (5)	0.3 I_2	Indep. arbit.	0–1 loss
Model 4	2	5	12, 2, 2, 2, 2	$\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}-1\\-1\end{bmatrix}, \begin{bmatrix}1\\-1\end{bmatrix}, \begin{bmatrix}-1\\1\end{bmatrix}$	6 (5)	0.3 I_2	Homo. arbit.	0–1 loss
Model 5	20	2	12, 2	$0_{20}, (0.05)_{20}$	−20.65 (5)	0.3 I_2	Indep. iden.	$\begin{bmatrix}0 & 2\\1 & 0\end{bmatrix}$
Model 6	20	2	20, 20	$0_{20}, 0_{20}$	−20.65 (5)	0.3 I_{20}	Indep. iden.	$\begin{bmatrix}0 & 2\\1 & 0\end{bmatrix}$
Model 7	20	5	12, 2, 2, 2, 2	$0_{20}, (0.1)_{20}, (-0.1)_{20}, \begin{bmatrix}(0.1)_{10}\\(-0.1)_{10}\end{bmatrix}, \begin{bmatrix}(-0.1)_{10}\\(0.1)_{10}\end{bmatrix}$	−20.65 (5)	0.3 I_{20}	Indep. iden.	0–1 loss
Model 8	20	5	20, 20, 20, 20, 20	$0_{20}, 0_{20}, 0_{20}, 0_{20}, 0_{20}$	−20.65 (5)	0.3 I_{20}	Indep. iden.	0–1 loss

0_k and $(a)_k$ represent all-zero and all-a column vectors of length k, respectively

we evaluate classification performance in a Monte Carlo estimation loop with 10,000 iterations. In each iteration, we follow a two-step procedure for sample generation: (1) generate random feature-label distribution parameters from the prior (each serving as the true underlying feature-label distribution) and (2) generate a random sample of size n from this fixed feature-label distribution. The generated random sample is used to train classifiers and evaluate their true risk. In the non-zero-one loss case, we also estimate risk and evaluate its accuracy using the performance metrics discussed earlier. We vary the sample size throughout and analyze its effect on performance.

5.4 Real data

We consider two real datasets. The first is a breast cancer dataset containing 295 sample points [28], which will be used to demonstrate binary classification under a non-zero-one loss function. The second is composed of five different cancer types from The Cancer Genome Atlas (TCGA) project, which demonstrates multi-class classification under zero-one loss.

In all real-data simulations, we assume that c_y is known and equal to the proportion of class-y sample points in the whole dataset. We form a Monte Carlo estimation loop to evaluate classification and risk estimation, where we iterate 1000 times with the breast cancer dataset and 10,000 times with the TCGA dataset. In each iteration, we obtain a stratified training sample of size n, i.e., we select a subset of the original dataset, keeping the proportion of points in class y as close as possible to c_y for every y. We use these training points to design several classifiers, while the remaining sample points are used as holdout data to approximate the true risk of each designed classifier. For the breast cancer dataset, we also use the training data to estimate risk and find the sample-conditioned MSE of the BRE. We vary sample size and analyze its effect on performance.

To implement Bayesian methods, we assume Gaussian distributions with arbitrary independent covariances in all real-data simulations. We calibrate hyperparameters, defined in Appendix 2, using a variant of the method-of-moments approach presented in [21]. In particular, we construct a calibration dataset from features not used to train the classifier and set $\nu_y = s_y/t_y$, $\kappa_y = 2(s_y^2/u_y) + D + 3$, $\mathbf{m}_y = [m_y, \ldots, m_y]$, and $\mathbf{S}_y = (\kappa_y - D - 1)s_y\mathbf{I}_D$, where m_y is the mean of the means of features among class-y points of the calibration dataset, and s_y is the mean of the variances of features in class y. t_y is the variance of the means of features in class y, where the 10 % of the means with the largest absolute value are discarded. Likewise, u_y is the variance of the variances of features in class y, where the 10 % of the variances with the largest value are discarded.

In the breast cancer data, 180 patients are assigned to class 0 (good prognosis) and 115 to class 1 (bad prognosis) in a 70-feature prognosis profile. A correct prognosis is associated with 0 loss, wrongly declaring a good prognosis incurs a loss of 1, and wrongly declaring a bad prognosis incurs a loss of 2. We use pre-selected features for classifier training, originally published in [29]. When $D = 2$, these features are CENPA and BBC3, and when $D = 5$, we

also add CFFM4, TGFB3, and DKFZP564D0462. Rather than discard the $70 - D$ features not used for classification, we use these features to calibrate priors using the method-of-moments approach described above.

For our second dataset, we downloaded level-3 microarray data from the TCGA data portal for five different kinds of cancers: breast invasive carcinoma (BRCA) with 593 sample points, colon adenocarcinoma (COAD) with 174 sample points, kidney renal clear cell carcinoma (KIRC) with 72 sample points, lung squamous cell carcinoma (LUSC) with 155 sample points, and ovarian serous cystadenocarcinoma (OV) with 562 sample points. We pooled all the sample points into a single dataset, removed features with missing values in any cancer type (17,016 features remained out of 17,814), and quantile-normalized the data with the median of the ranked values. We preselect features for classifier training and prior calibration using the full dataset and one of two methods, which both operate in two phases: in phase 1, we pass $D+100$ features, and in phase 2, we select D features from those passing phase 1. The D features passing both phases are used for classifier training, and the features passing phase 1 but not phase 2 are used for prior calibration. The first feature selection method (FS-1) passes features that minimize a score evaluating separation between classes in phase 1 and selects features that minimize a score evaluating Gaussianity of the classes in phase 2. To evaluate separation between classes in phase 1, for each pair of classes, we obtain t-test p-values for each feature and rank these across all features, low p-values being assigned a lower rank, and finally, we report the rank product score for each feature over all 10 pairs of classes. To evaluate Gaussianity in phase 2, for each class, we rank Shapiro-Wilk test p-values across all features passing phase 1, high p-values being assigned a lower rank, and report the rank product score for each feature across all five classes. The second feature selection method (FS-2) passes features minimizing the rank product score from Shapiro-Wilk tests applied to all 17,016 features in phase 1, and in phase 2, we select D features from those passing phase 1 using sequential forward search (SFS) with LDA classification and resubstitution risk as the optimization criterion.

5.5 Discussion

Models 1 and 2 focus on the effect of risk on classification and risk estimation performance. In Fig. 1, we evaluate the performance of risk estimators and classifiers under model 1. Graphs in the left column present the mean, averaged over all 10,000 sample realizations, of the true risk and all risk estimators considered for LDA, QDA, and OBRC classification. Note for small samples of size $n = 20$ and LDA or QDA classification, surrogate classifiers in the bootstrap risk estimator are occasionally undefined depending on the realized bootstrap sample. These events are thrown out so that only a subset of the original 10,000 sample realizations are used to approximate the mean bootstrap risk estimator. The graphs on the right column of Fig. 1 present the square root of the mean, averaged over all sample realizations, of the square difference between the true risk and each risk estimator, which we call the *empirical RMS*. The square root of the mean, averaged over all sample realizations, of the sample-conditioned MSE of the BRE from (24), which we call the *Bayesian RMS*, is also shown.

The BRE is an unbiased estimator, so the mean true risk and mean BRE curves should be aligned with enough iterations, which is observed. The empirical and Bayesian RMS both approximate the unconditional RMS so that these curves should also be aligned with enough iterations, as observed. Furthermore, the BRE is theoretically optimal in both the sample-conditioned and unconditioned RMS, and as expected, the empirical and Bayesian RMS curves for BRE under each classification rule outperform all other risk estimation rules. Thus, the BRE yields a significant improvement over classical risk estimators in terms of both bias and RMS performance within the Bayesian model. If we compare classification rules, the RMS of BRE is consistently lower for OBRC than LDA and QDA, although there is no theoretical guarantee for this. Similar curves for model 2 are provided in Fig. 2.

To illustrate how the sample-conditioned MSE may be used to assess the accuracy of a risk estimate, suppose that we have observed a sample, trained a classifier, and obtained the BRE and the MSE of the BRE. For this fixed sample, but random parameters in the Bayesian model, the true risk has a mean equal to the BRE and a variance equal to the sample-conditioned MSE so that the random variable $Z = (\widehat{R} - R)/\mathrm{RMS}(\widehat{R}|S)$ must have zero mean and unit variance. This holds for any classification rule. In Fig. 3, we present quantile-quantile (Q-Q) plots of the sample quantiles of Z versus theoretical quantiles from a standard normal distribution. Figure 3a provides Q-Q plots with realizations of Z taken under OBRC classification and BRE risk estimation in model 1 with various sample sizes, along with a 45° reference line, and Fig. 3b provides similar graphs for model 2. Observe that Z appears approximately standard normal, particularly under large sample sizes. Under smaller samples, Z appears more positively skewed but has approximately zero mean and unit variance. Q-Q plots for other classifiers are similar.

In Figs. 4 and 5, we provide examples of decision boundaries for models 3 and 4, respectively, which focus on the effect of multiple classes in two dimensions. Under model 3, where we assume independent covariances, the decision boundaries of OBRC are most similar to QDA, although they are in general of a polynomial order. Under model 4, where we assume homoscedastic covariances,

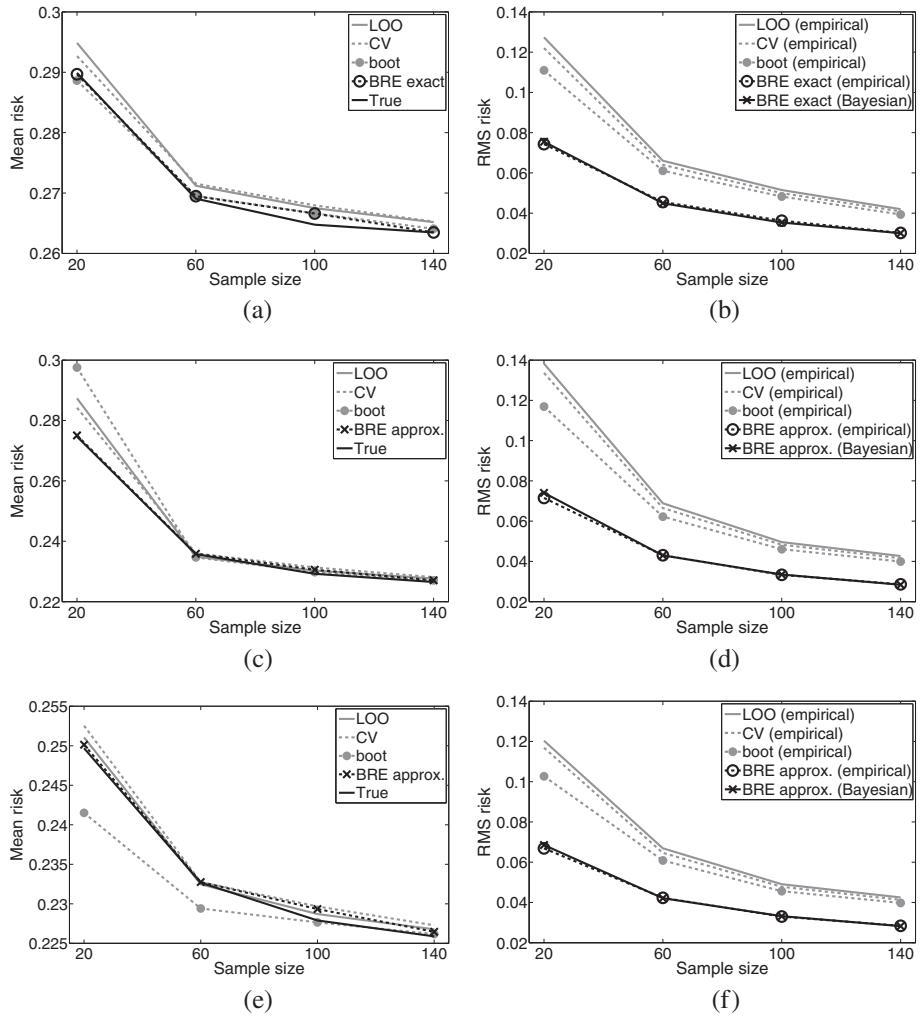

Fig. 1 Mean risks and RMS for model 1, three classification rules (LDA, QDA, and OBRC), and all risk estimators. **a** Mean risk under LDA; **b** RMS risk under LDA; **c** mean risk under QDA; **d** RMS risk under QDA; **e** mean risk under OBRC; **f** RMS risk under OBRC

OBRC is most similar to LDA, although the decision boundaries are not necessarily linear.

In Fig. 6, we present the mean and standard deviation of the true risk with respect to all sample realizations as a function of sample size for models 3 and 4. OBRC outperforms all other classification rules with respect to mean risk, as it must, since the OBRC is defined to minimize mean risk. Although there is no guarantee that OBRC should minimize risk variance, in these examples, the risk variance is lower than in all other classification rules. The performance gain is particularly significant for small samples. Consider Figs. 6a and 6b, where we observe that, at a sample size of 10, the risk of OBRC has a mean of about 0.16 and standard deviation of about 0.065, whereas the risk of the next best classifier, RBF-SVM, has a mean of about 0.22 and standard deviation of about 0.09.

Figure 7 provides the performance of risk estimators under OBRC classification in models 5 and 6, demonstrating performance in 20 dimensions with independent scaled identity covariance priors. Settings in model 5 are designed to produce a low mean risk and model 6 a high mean risk. Graphs in the left column present the mean true risk, averaged over all 10,000 sample realizations; the center column presents empirical and Bayesian RMS curves; and the right column presents Q-Q plots of Z. As in Figs. 1 and 2, the BRE appears unbiased, the empirical and Bayesian RMS curves are aligned, and the RMS curves are optimal. From the Q-Q plots, the distribution of Z appears to be skinny-tailed even under large n, although it is approximately zero mean and unit variance.

In Fig. 8, we present the mean and standard deviation of the true risk of all classifiers as a function of sample size

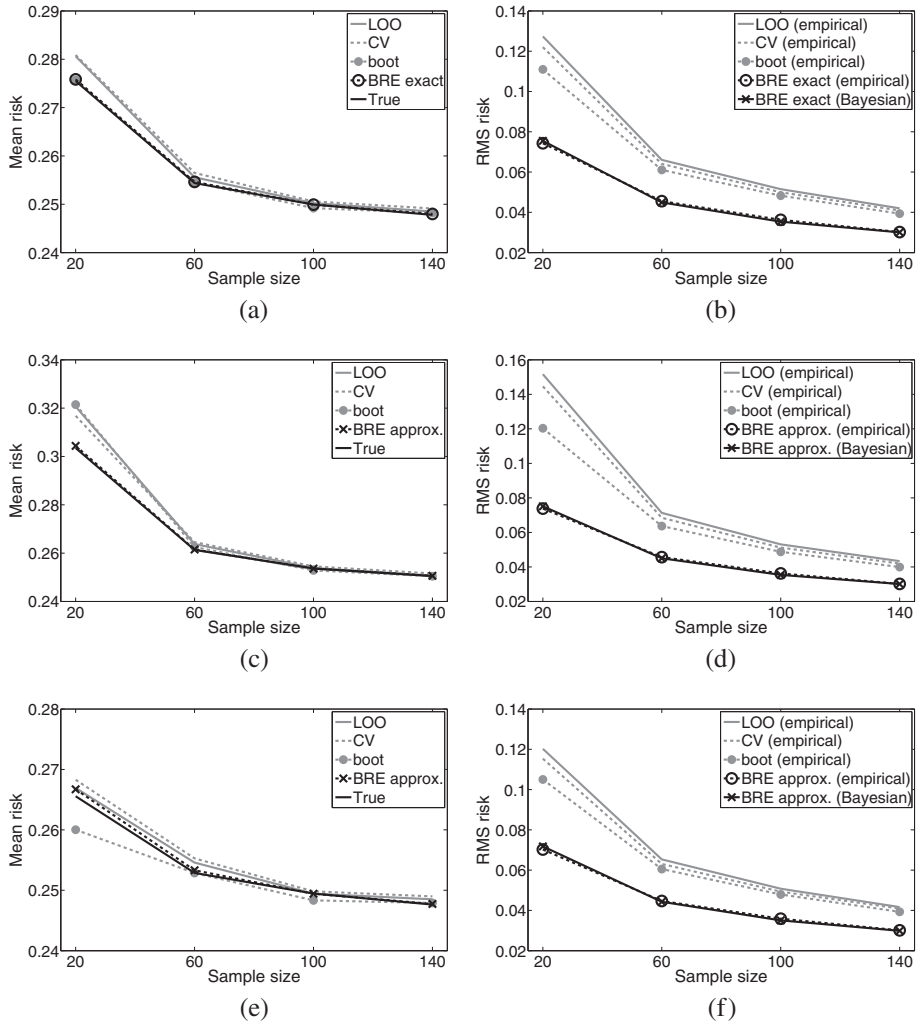

Fig. 2 Mean risks and RMS for model 2, three classification rules (LDA, QDA, and OBRC), and all risk estimators. **a** Mean risk under LDA; **b** RMS risk under LDA; **c** mean risk under QDA; **d** RMS risk under QDA; **e** mean risk under OBRC; **f** RMS risk under OBRC

for models 7 and 8, where model 7 is designed to produce a low mean risk and model 8 a high mean risk. OBRC again outperforms all other classification rules with respect to mean risk, as it should. There is no guarantee that OBRC should minimize risk variance, and although risk variance is lowest for OBRC in Fig. 8b, in Fig. 8d it is actually highest. Performance gain is particularly significant for small samples.

In Figs. 9 and 10, we evaluate the performance of risk estimators and classifiers under the breast cancer dataset

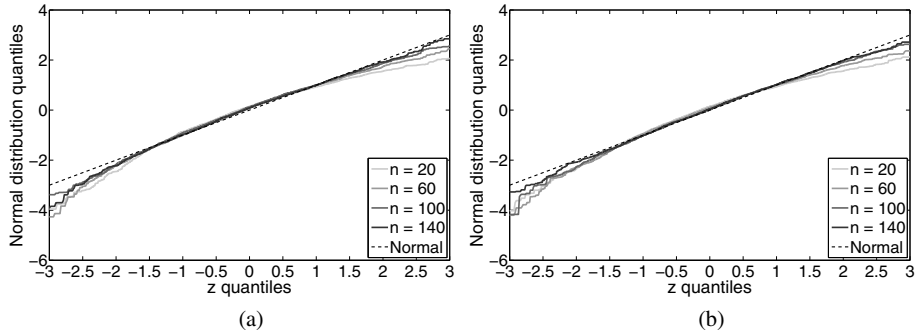

Fig. 3 Q-Q plots of Z under OBRC and BRE. **a** Model 1; **b** model 2

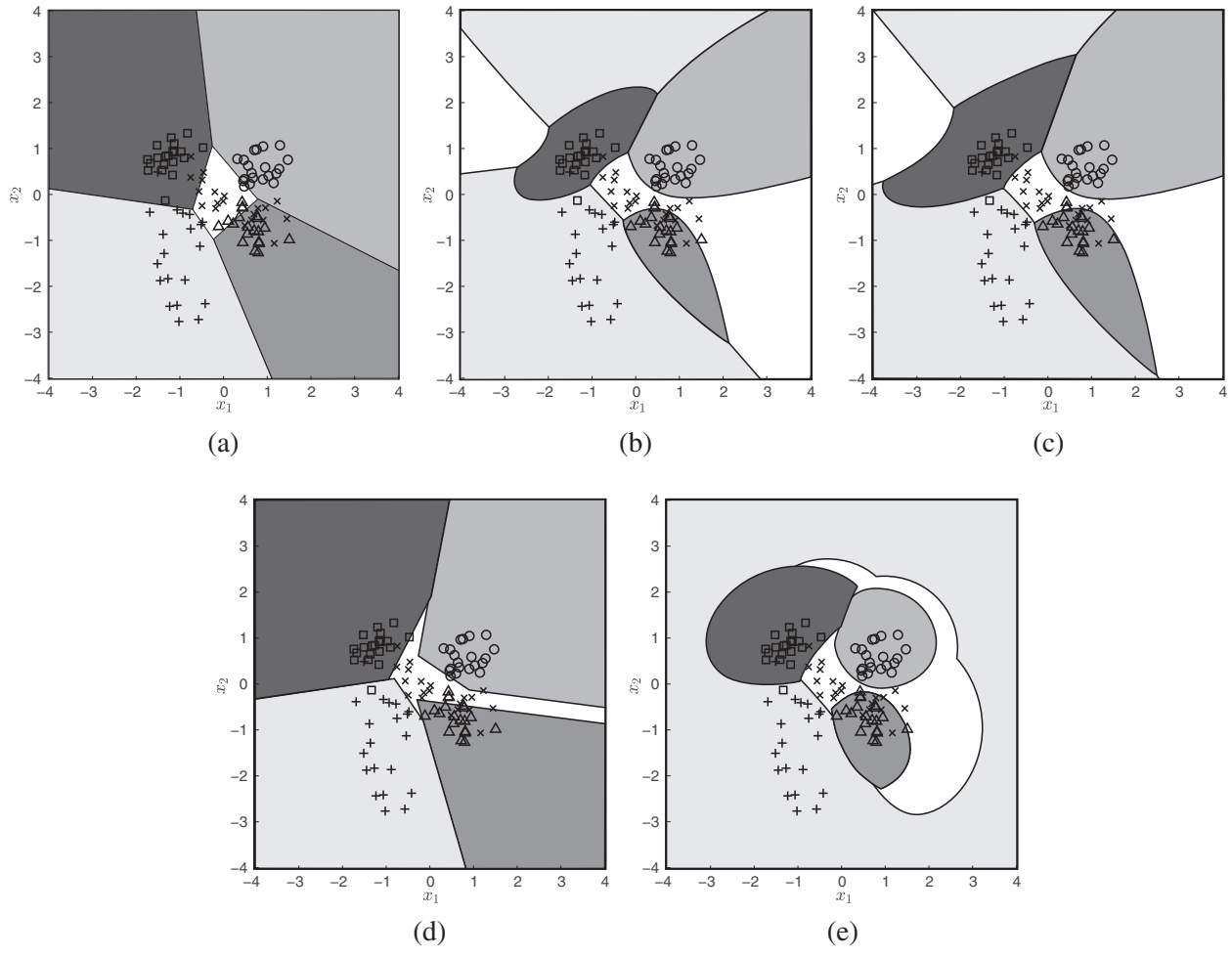

Fig. 4 Example decision boundaries for model 3 with multi-class classification. **a** LDA; **b** QDA; **c** OBRC; **d** L-SVM; **e** RBF-SVM

for $D = 2$ and $D = 5$, respectively. Graphs in the left column present the mean true risk and mean risk estimates, graphs in the center column present the empirical RMS of all risk estimates and the Bayesian RMS for the BRE, and graphs in the right column present the Q-Q plots of Z for various sample sizes. LDA, QDA, and OBRC are presented in the top, center, and bottom rows, respectively. Although the BRE is theoretically unbiased and minimizes RMS when averaged across random distributions in the uncertainty class, when applied to a specific dataset or distribution, we now observe a bias (in the left column) and a discrepancy between the empirical and Bayesian RMS (in the center column). In particular, for all classifiers under $D = 2$ and for LDA under $D = 5$, we observe a high bias, for QDA and OBRC under $D = 5$, we observe a low bias, and in all cases, the Bayesian RMS lies below the empirical RMS. That being said, the empirical RMS still outperforms that of distribution-free resampling error estimators (LOO, CV, and boot). Although resampling estimators are nearly unbiased, they suffer from such large variance under small samples that the BRE, despite imperfections in the Gaussianity assumption and prior construction method, may still outperform in practice thanks to optimization. Turning to classifier performance, in these simulations, LDA appears to outperform QDA and OBRC with independent arbitrary covariances. Keep in mind that Bayesian methods are not guaranteed to be optimal in all datasets and all settings but, rather, are only optimal within the assumed model. In fact, OBRC with homoscedastic arbitrary covariances (not shown in the figures) performs as well as, or significantly better than, LDA, suggesting that covariances in this problem are approximately homoscedastic. From the Q-Q plots, Z deviates from the reference standard normal CDF, with a clear shift in the mean and sometimes variance. For instance, under the LDA classification with $D = 2$ and $n = 70$ (corresponding to Fig. 9c), the mean of Z is 0.76 and the standard deviation is 1.08, and

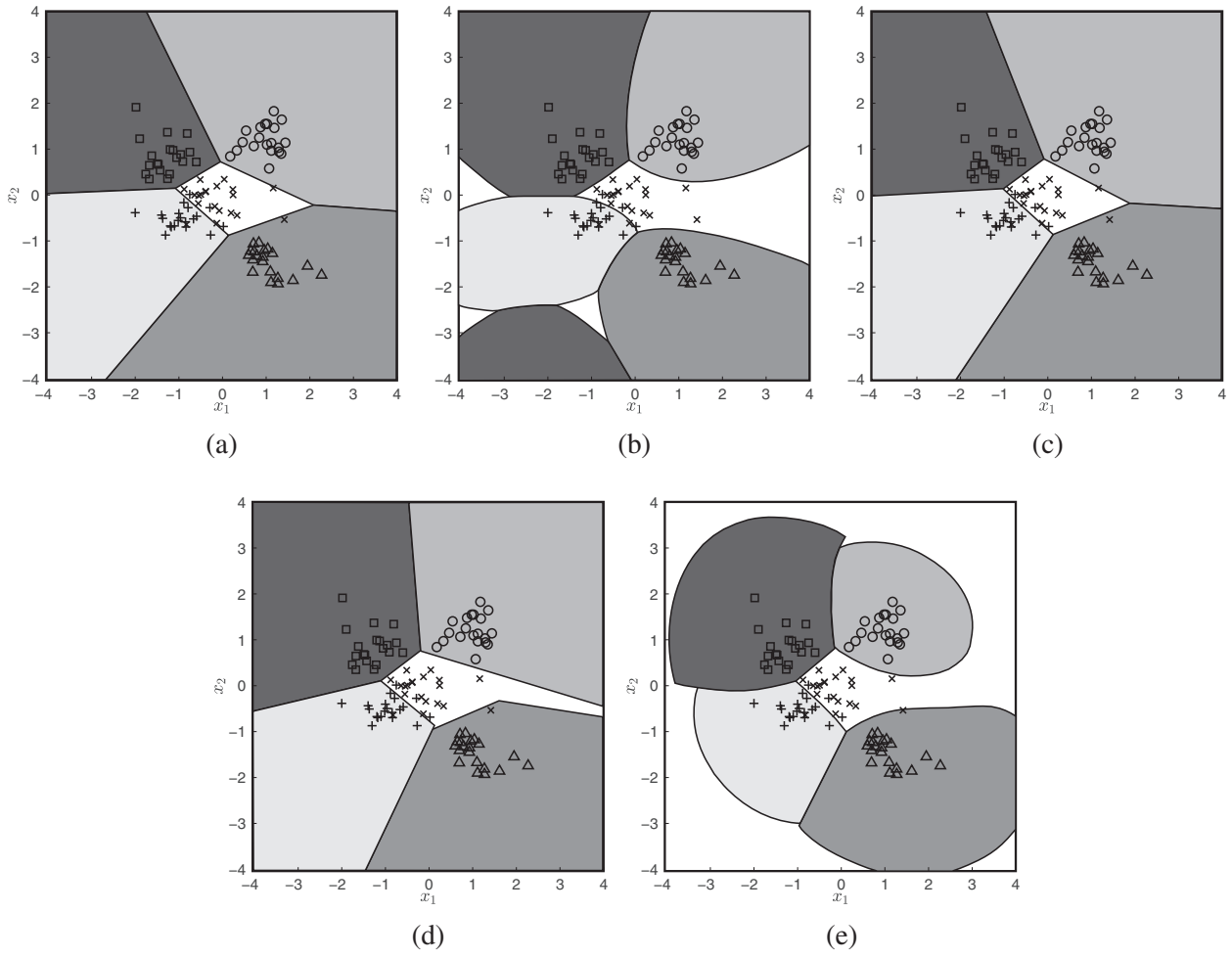

Fig. 5 Example decision boundaries for model 4 with multi-class classification. **a** LDA; **b** QDA; **c** OBRC; **d** L-SVM; **e** RBF-SVM

under the OBRC classification with $D = 5$ and $n = 70$ (corresponding to Fig. 10i), the mean of Z is -1.08 and the standard deviation is 1.49.

In Fig. 11, we present the mean of the true risk with respect to random samples from the TCGA dataset, as a function of sample size, for different feature selection methods and selected feature set sizes. Due to covariance estimation problems, QDA cannot be trained for $D = 20$ in this range of sample sizes. OBRC with calibrated priors consistently outperforms under small samples and performs robustly under large samples. These results depend on the particular features selected and note LDA may have an advantage under FS-2, which minimizes the apparent error of LDA classifiers.

In real applications, data rarely satisfy modeling assumptions, for instance, Gaussianity, and there may be a concern that performance will suffer. Firstly, keep in mind the need to validate assumptions in the Bayesian model. For example, Gaussianity tests and homoscedasticity tests may be used to validate these underlying assumptions. Our real-data simulations demonstrate a few examples of how Gaussianity tests may be used in conjunction with Bayesian methods. Secondly, previous works have shown that Bayesian methods are relatively robust to deviations from a Gaussianity assumption [10, 14]. This is observed, for instance, in Figs. 9 and 10. Thirdly, inference from non-informative priors may serve as a reference. The OBRC under non-informative priors and an arbitrary homoscedastic covariance model behaves similarly to LDA and under an arbitrary independent covariance model behaves similarly to QDA [13, 14]. Thus, the OBRC can be seen as unifying and optimizing these classifiers. This applies in Fig. 11, where OBRC with an appropriate covariance model and non-informative prior performs indistinguishably from LDA. The conditional MSE is also an immensely useful tool to quantify the accuracy of a risk estimator. For instance, one may employ the MSE for censored sampling by collecting batches of sample points

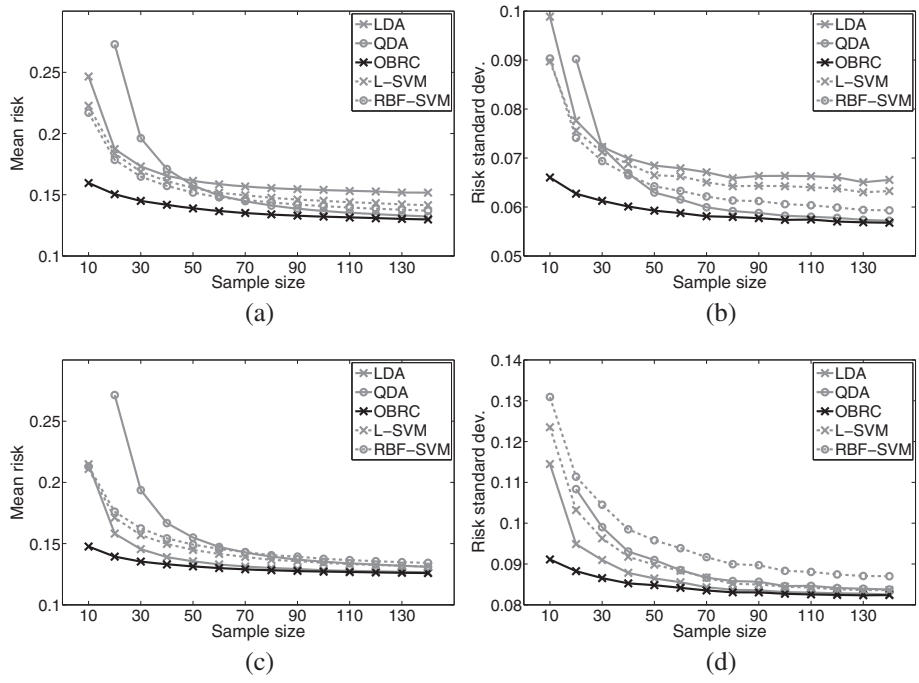

Fig. 6 True risk statistics for models 3 and 4 and five classification rules (LDA, QDA, OBRC, L-SVM, and RBF-SVM). **a** Model 3, mean; **b** model 3, standard deviation; **c** model 4, mean; **d** model 4, standard deviation

until the sample-conditioned MSE reaches an acceptable level, and either an acceptable risk has been achieved or it has been determined that an acceptable risk cannot be achieved. Lastly, although we provide analytic solutions under discrete and Gaussian models, the basic theory for this work does not require these assumptions. For instance, recent work in [30] develops a Bayesian Poisson model for RNA-Seq data, where Bayesian error estimators and optimal Bayesian classifiers are obtained using Markov chain Monte Carlo (MCMC) techniques.

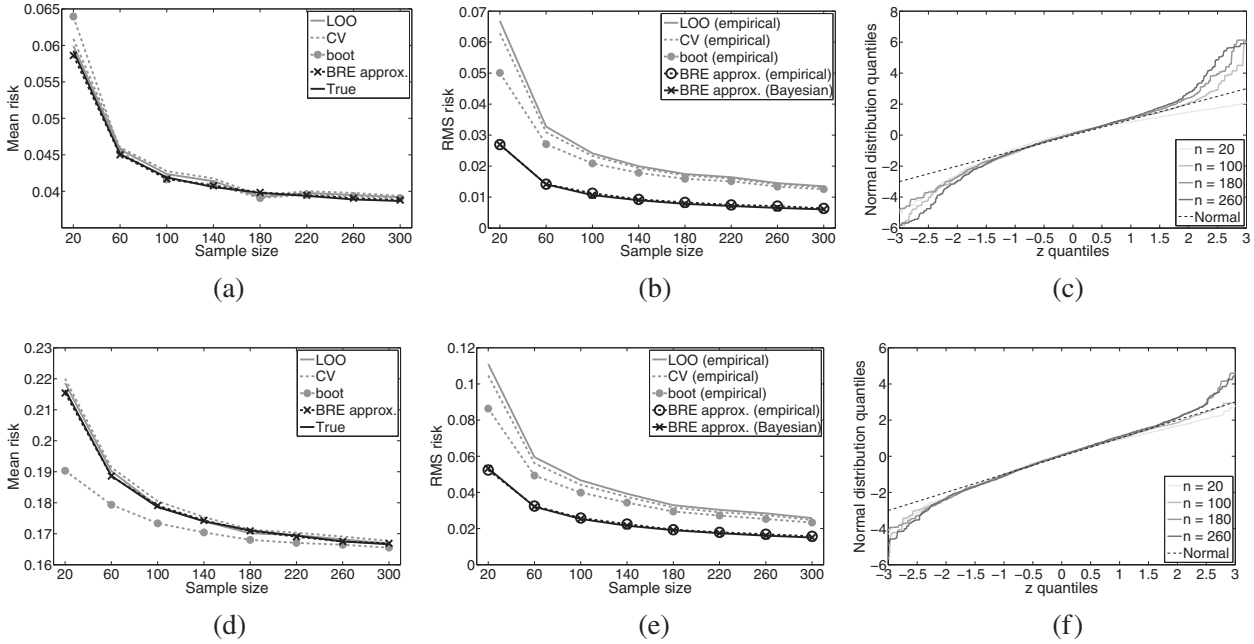

Fig. 7 Mean risks, RMS, and Q-Q plots of Z for models 5 and 6, OBRC classification, and all risk estimators. **a** Mean risk under model 5; **b** RMS risk under model 5; **c** Q-Q plots of Z under model 5; **d** mean risk under model 6; **e** RMS risk under model 6; **f** Q-Q plots of Z under model 6

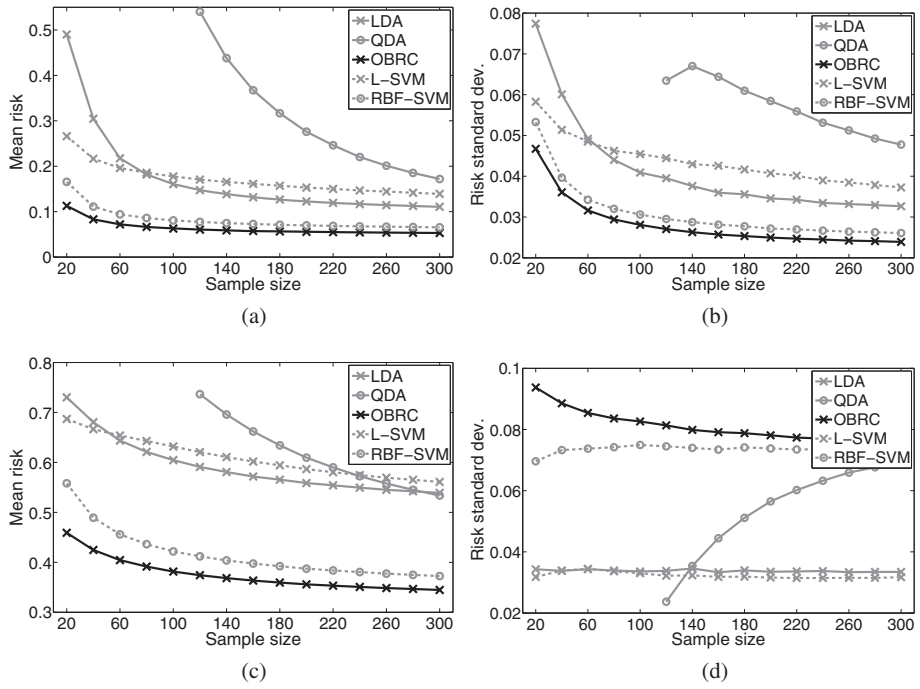

Fig. 8 True risk statistics for models 7 and 8 and five classification rules (LDA, QDA, OBRC, L-SVM, and RBF-SVM). **a** Model 7, mean; **b** model 7, standard deviation; **c** model 8, mean; **d** model 8, standard deviation

6 Conclusion

We have extended optimal Bayesian classification theory to multiple classes and arbitrary loss functions, giving rise to Bayesian risk estimators, the sample-conditioned MSE for arbitrary risk estimators, and optimal Bayesian risk classifiers. We have developed a new interpretation of the conditional MSE based on effective joint densities, which is useful in developing analytic forms and approximations for the conditional MSE. We also provide new analytic solutions for the conditional MSE under homoscedastic covariance models. Simulations based on several synthetic Gaussian models and two real microarray datasets also demonstrate good performance relative to existing methods.

Appendix 1: Discrete models

Consider a discrete sample space, $\mathcal{X} = \{1, 2, \ldots, b\}$. Let p_x^y be the probability that a point from class y is observed in bin $x \in \mathcal{X}$, and let U_x^y be the number of sample points observed from class y in bin x. Note $n_y = \sum_{x=1}^{b} U_x^y$. The discrete Bayesian model defines $\Theta_y = [P_1^y, \ldots, P_b^y]$, with parameter space $\mathcal{T}_y = \Delta^{b-1}$. For each y, we define Dirichlet priors on Θ_y with hyperparameters $\alpha^y = \{\alpha_1^y, \ldots, \alpha_b^y\}$:

$$\pi(\theta_y) \propto \prod_{x=1}^{b} (p_x^y)^{\alpha_x^y - 1}.$$

Assume that Θ_y are mutually independent. Uniform priors are achieved when $\alpha_x^y = 1$ for all x and y. Given data, the posteriors are again Dirichlet with updated hyperparameters, $\alpha_x^{y*} = \alpha_x^y + U_x^y$ for all x and y. For proper posteriors, α_x^{y*} must all be positive for all x and y. The effective density is thus given by:

$$f(x \mid y, S) = \mathrm{E}[P_x^y \mid S] = \frac{\alpha_x^{y*}}{\alpha_+^{y*}}, \tag{31}$$

where $\alpha_+^{y*} = \sum_{x=1}^{b} \alpha_x^{y*}$. Thus, we have

$$\widehat{\varepsilon}^{i,y}(\psi, S) = \sum_{x=1}^{b} \frac{\alpha_x^{y*}}{\alpha_+^{y*}} I_{\psi(x)=i}.$$

The effective joint density, $f(x, w \mid y, z, S)$, for $y = z$, can be found from properties of Dirichlet distributions. We have for any $y \in \{0, \ldots, M-1\}$ and $x, w \in \mathcal{X}$,

$$f(x, w \mid y, y, S) = \mathrm{E}\left[P_x^y P_w^y \mid S\right] = \frac{\alpha_x^{y*}(\alpha_w^{y*} + \delta_{xw})}{\alpha_+^{y*}\left(\alpha_+^{y*} + 1\right)},$$

where δ_{xw} equals 1 if $x = w$ and 0 otherwise. From (28),

$$\begin{aligned}
&\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,y}(\Theta_y) \mid S\right] \\
&\quad = \sum_{x=1}^{b}\sum_{w=1}^{b} \frac{\alpha_x^{y*}\left(\alpha_w^{y*} + \delta_{xw}\right)}{\alpha_+^{y*}(\alpha_+^{y*} + 1)} I_{\psi(x)=i} I_{\psi(w)=j} \\
&\quad = \frac{\widehat{\varepsilon}^{i,y}(\psi, S)\left(\alpha_+^{y*}\widehat{\varepsilon}^{j,y}(\psi, S) + \delta_{ij}\right)}{\alpha_+^{y*} + 1}.
\end{aligned} \tag{32}$$

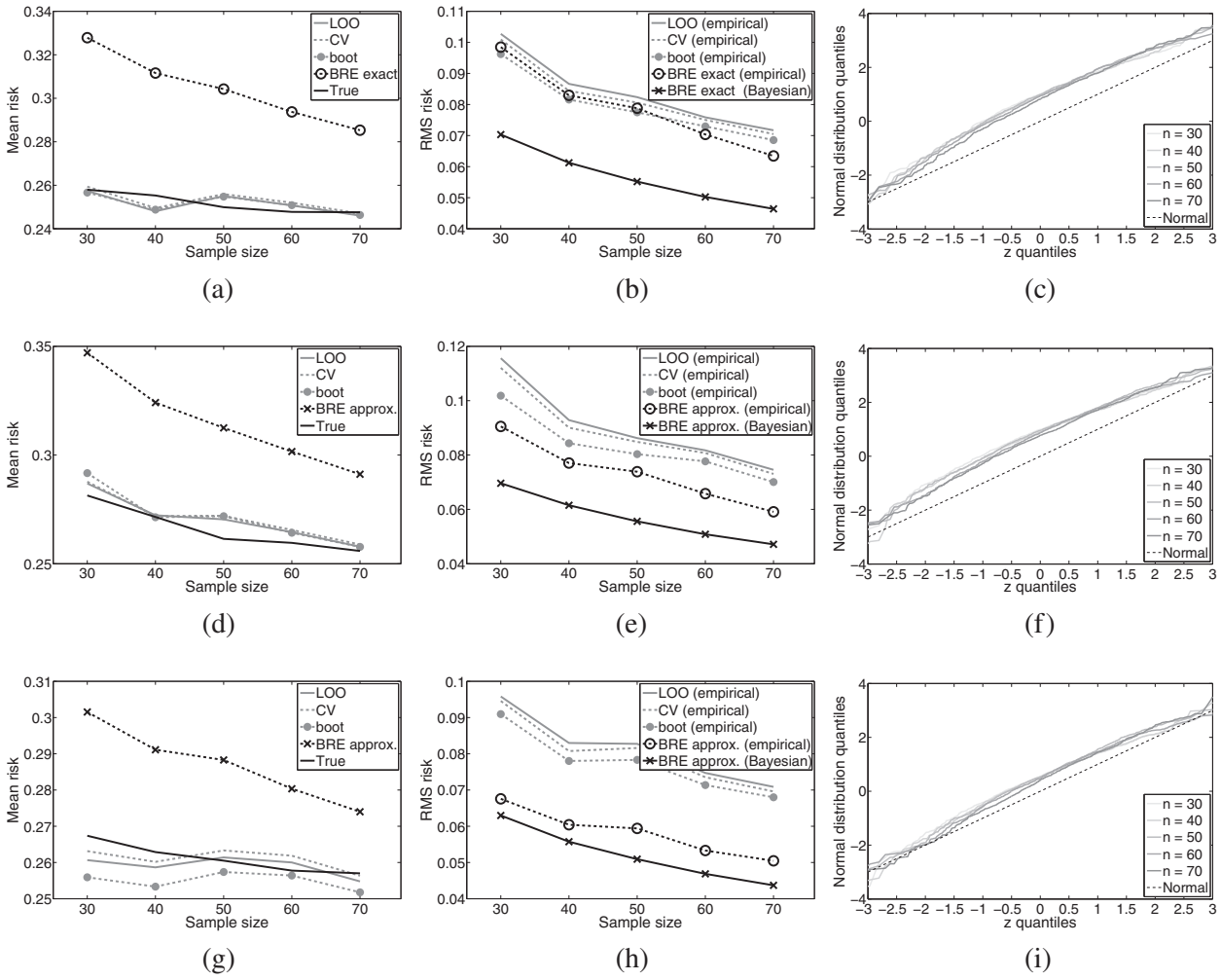

Fig. 9 Mean risks, RMS, and Q–Q plots of Z for the breast cancer dataset, $D = 2$, three classification rules (LDA, QDA, and OBRC), and all risk estimators. **a** Mean risk under LDA; **b** RMS risk under LDA; **c** Q–Q plots of Z under LDA; **d** mean risk under QDA; **e** RMS risk under QDA; **f** Q–Q plots of Z under QDA; **g** mean risk under OBRC; **h** RMS risk under OBRC; **i** Q–Q plots of Z under OBRC

When $y \neq z$, $\mathrm{E}\left[\varepsilon_n^{i,y}(\Theta_y)\varepsilon_n^{j,z}(\Theta_y) \mid S\right]$ may be found from (25).

Appendix 2: Gaussian models

Suppose $\mathcal{X}$ is a D dimensional space in which each point is represented by a column vector and each class-y conditional distribution is Gaussian with mean vector μ_y and covariance matrix Σ_y. We will consider independent covariance models, where the Σ_y are mutually independent prior to observing the data, and homoscedastic covariance models, where Σ_y are identical for all y [13]. We will also consider three structures for the covariance: known, scaled identity, and arbitrary. Throughout, we use μ_y and Σ_y to denote both random quantities and their realizations, and we use $\Sigma_y \succ 0$ to denote a valid covariance matrix, i.e., a symmetric, positive definite matrix. Throughout, we will find analytic forms for the BRE and conditional MSE under binary linear classifiers, ψ, of the form

$$\psi(\mathbf{x}) = \begin{cases} 0 & \text{if } g(\mathbf{x}) \leq 0, \\ 1 & \text{otherwise,} \end{cases} \tag{33}$$

where $g(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ for some vector $\mathbf{a}$ and scalar b, and a superscript T denotes matrix transpose.

Known covariance

Assume that $\Sigma_y \succ 0$ is known so that $\Theta_y = \mu_y$ with parameter space $\mathcal{T}_y = \mathbb{R}^D$. We assume the μ_ys are mutually independent and use the following prior:

$$\pi(\mu_y) \propto \left|\Sigma_y\right|^{-\frac{1}{2}} \exp\left(-\frac{\nu_y}{2}\left(\mu_y - \mathbf{m}_y\right)^T \Sigma_y^{-1}\left(\mu_y - \mathbf{m}_y\right)\right), \tag{34}$$

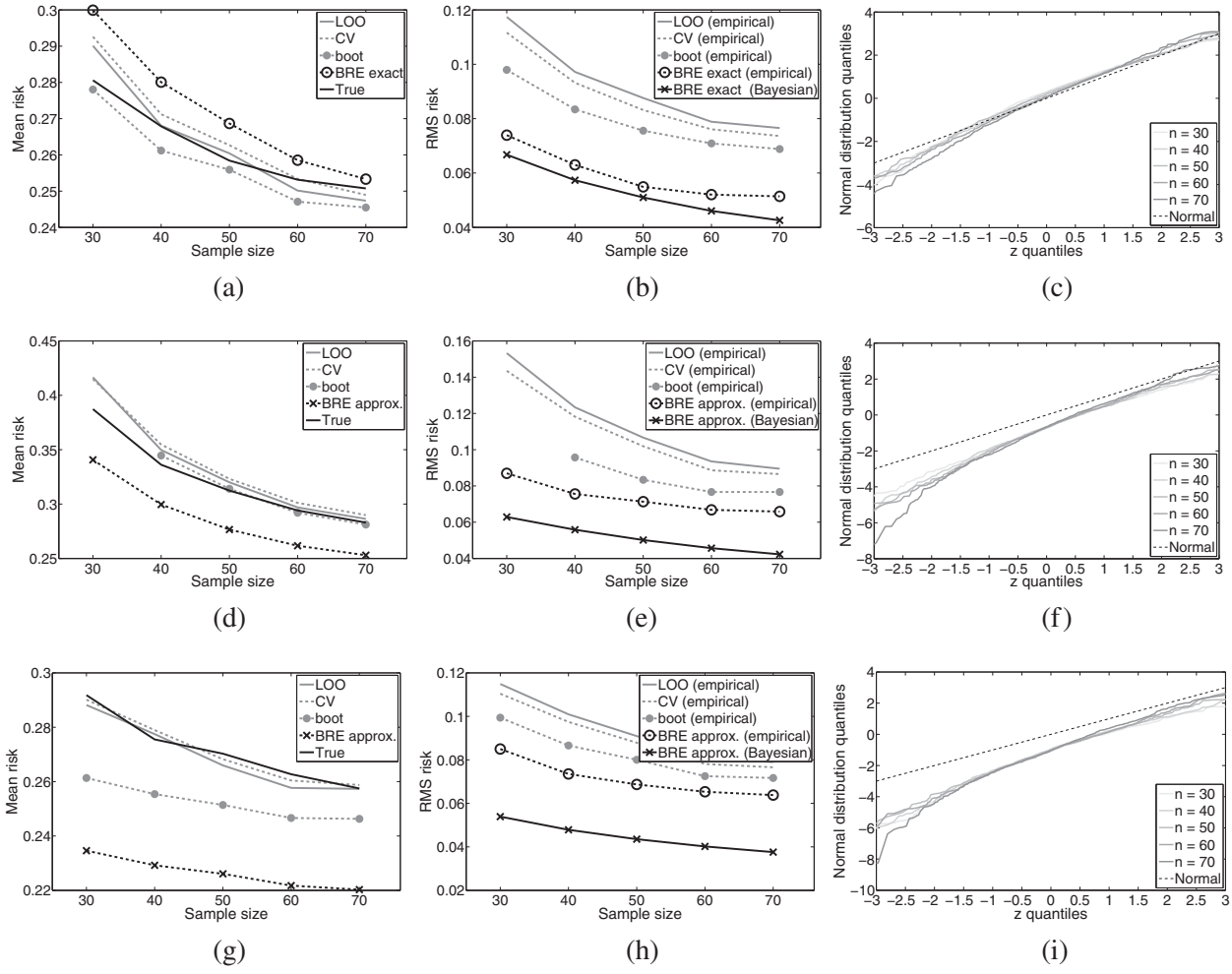

Fig. 10 Mean risks, RMS and Q–Q plots of Z for the breast cancer dataset, $D = 5$, three classification rules (LDA, QDA, and OBRC), and all risk estimators. **a** Mean risk under LDA; **b** RMS risk under LDA; **c** Q-Q plots of Z under LDA; **d** mean risk under QDA; **e** RMS risk under QDA; **f** Q–Q plots of Z under QDA; **g** mean risk under OBRC; **h** RMS risk under OBRC; **i** Q–Q plots of Z under OBRC

with hyperparameters $\nu_y \in \mathbb{R}$ and $\mathbf{m}_y \in \mathbb{R}^D$, where $|\cdot|$ denotes a determinant. When $\nu_y > 0$, this is a Gaussian distribution with mean $\mathbf{m}_y$ and covariance Σ_y/ν_y. Under this model, the posterior is of the same form as the prior, with updated hyperparameters

$$\begin{aligned} \nu_y^* &= \nu_y + n_y, \\ \mathbf{m}_y^* &= \mathbf{m}_y + n_y \frac{\widehat{\mu}_y - \mathbf{m}_y}{\nu_y + n_y}, \end{aligned} \tag{35}$$

where $\widehat{\mu}_y$ is the usual sample mean of training points in class y. We require $\nu_y^* > 0$ for a proper posterior.

The effective density was shown in [13] to be the following Gaussian distribution:

$$f(\mathbf{x} \,|\, y, S) \sim \mathcal{N}\left(\mathbf{m}_y^*, \frac{\nu_y^* + 1}{\nu_y^*}\Sigma_y\right). \tag{36}$$

To find the BRE for a linear classifier, let $P = (-1)^i g(\mathbf{X})$. From the effective density,

$$f(p \,|\, y, S) \sim \mathcal{N}\left((-1)^i g(\mathbf{m}_y^*), \frac{\nu_y^* + 1}{\nu_y^*}\mathbf{a}^T\Sigma_y\mathbf{a}\right). \tag{37}$$

Thus,

$$\begin{aligned} \widehat{\varepsilon}^{i,y}(\psi, S) &= \mathrm{P}((-1)^i g(\mathbf{X}) \le 0 \,|\, y, S) \\ &= \mathrm{P}(P \le 0 \,|\, y, S) \\ &= \Phi\left(-\frac{(-1)^i g(\mathbf{m}_y^*)}{\sqrt{\mathbf{a}^T\Sigma_y\mathbf{a}}}\sqrt{\frac{\nu_y^*}{\nu_y^* + 1}}\right), \end{aligned} \tag{38}$$

where $\Phi(x)$ is the standard normal CDF. This result was also found in [10].

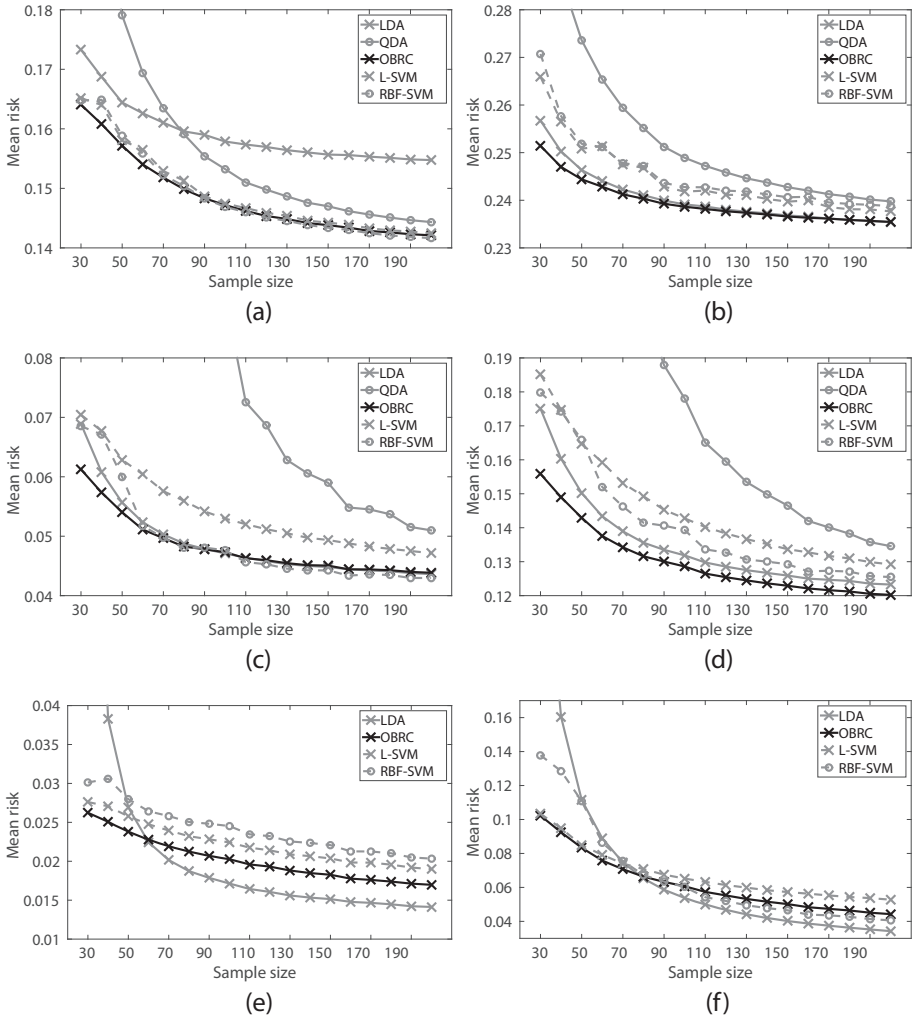

Fig. 11 True risk mean for the TCGA dataset and five classification rules (LDA, QDA, OBRC, L-SVM, and RBF-SVM). **a** FS-1, $D = 2$; **b** FS-2, $D = 2$; **c** FS-1, $D = 5$; **d** FS-2, $D = 5$; **e** FS-1, $D = 20$; **f** FS-2, $D = 20$

To find the MSE under linear classification, note $f(\mathbf{w} \mid \mathbf{x}, y, z, S)$ is of the same form as $f(\mathbf{x} \mid y, S)$ with posterior hyperparameters updated with $\{\mathbf{x}, y\}$ as a new sample point. Hence, for $y = z$,

$$f(\mathbf{w} \mid \mathbf{x}, y, y, S) \sim \mathcal{N}\left(\mathbf{m}_y^* + \frac{\mathbf{x} - \mathbf{m}_y^*}{\nu_y^* + 1}, \frac{\nu_y^* + 2}{\nu_y^* + 1}\Sigma_y\right), \quad (39)$$

and the effective joint density is thus given by

$$f(\mathbf{x}, \mathbf{w} \mid y, y, S) \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}_y^* \\ \mathbf{m}_y^* \end{bmatrix}, \begin{bmatrix} \frac{\nu_y^*+1}{\nu_y^*}\Sigma_y & \frac{1}{\nu_y^*}\Sigma_y \\ \frac{1}{\nu_y^*}\Sigma_y & \frac{\nu_y^*+1}{\nu_y^*}\Sigma_y \end{bmatrix}\right). \quad (40)$$

Now let $Q = (-1)^j g(\mathbf{W})$. Since $\mathbf{X}$ and $\mathbf{W}$ are governed by the effective joint density in (40):

$$f(p, q \mid y, y, S) \sim \mathcal{N}\left(\begin{bmatrix} (-1)^i g(\mathbf{m}_y^*) \\ (-1)^j g(\mathbf{m}_y^*) \end{bmatrix}, \begin{bmatrix} \frac{\nu_y^*+1}{\nu_y^*}\mathbf{a}^T\Sigma_y\mathbf{a} & \frac{(-1)^{i+j}}{\nu_y^*}\mathbf{a}^T\Sigma_y\mathbf{a} \\ \frac{(-1)^{i+j}}{\nu_y^*}\mathbf{a}^T\Sigma_y\mathbf{a} & \frac{\nu_y^*+1}{\nu_y^*}\mathbf{a}^T\Sigma_y\mathbf{a} \end{bmatrix}\right).$$

Hence, from (29), we have

$$\begin{aligned} \mathrm{E}\left[\varepsilon^{i,y}(\psi, \Theta_y)\varepsilon^{j,y}(\psi, \Theta_y) \mid S\right] &= \mathrm{P}(P \le 0 \cap Q \le 0 \mid y, S) \\ &= \Phi\left(-\frac{(-1)^i g(\mathbf{m}_y^*)}{\sqrt{\mathbf{a}^T\Sigma_y\mathbf{a}}}\sqrt{\frac{\nu_y^*}{\nu_y^*+1}}, \right. \\ &\quad \left. -\frac{(-1)^j g(\mathbf{m}_y^*)}{\sqrt{\mathbf{a}^T\Sigma_y\mathbf{a}}}\sqrt{\frac{\nu_y^*}{\nu_y^*+1}}, \frac{(-1)^{i+j}}{\nu_y^*+1}\right), \end{aligned}$$

where $\Phi(x,y,\rho)$ is the joint CDF of two standard normal random variables with correlation ρ. When $y \neq z$, $\mathrm{E}\left[\varepsilon^{i,y}(\psi,\Theta_y)\varepsilon^{j,z}(\psi,\Theta_z) \mid S\right]$ is found from (25).

Homoscedastic arbitrary covariance

Assume $\Theta_y = [\mu_y, \Sigma]$, where the parameter space of μ_y is $\mathbb{R}^D$ and the parameter space of Σ consists of all symmetric positive definite matrices. Further, assume a conjugate prior in which the μ_ys are mutually independent given Σ so that

$$\pi(\theta) = \left(\prod_{y=0}^{M-1} \pi(\mu_y \mid \Sigma)\right)\pi(\Sigma), \tag{41}$$

where $\pi(\mu_y \mid \Sigma)$ is as in (34) with hyperparameters $\nu_y \in \mathbb{R}$ and $\mathbf{m}_y \in \mathbb{R}^D$, and

$$\pi(\Sigma) \propto |\Sigma|^{-\frac{\kappa+D+1}{2}} \exp\left(-\frac{1}{2}\mathrm{trace}\left(\mathbf{S}\Sigma^{-1}\right)\right), \tag{42}$$

with hyperparameters $\kappa \in \mathbb{R}$ and $\mathbf{S}$, a symmetric $D \times D$ matrix. If $\nu_y > 0$, then $\pi(\mu_y \mid \Sigma)$ is Gaussian with mean $\mathbf{m}_y$ and covariance Σ/ν_y. If $\kappa > D-1$ and $\mathbf{S} \succ 0$, then $\pi(\Sigma)$ is an inverse-Wishart distribution with hyperparameters κ and $\mathbf{S}$. If in addition $\kappa > D+1$, the mean of Σ exists and is given by $\mathrm{E}[\Sigma] = \mathbf{S}/(\kappa - D - 1)$; thus, $\mathbf{S}$ determines the shape of the expected covariance. The posterior is of the same form as the prior with the same updated hyperparameters given by (35) and

$$\begin{aligned}\kappa^* &= \kappa + n,\\ \mathbf{S}^* &= \mathbf{S} + \sum_{y=0}^{M-1}(n_y-1)\widehat{\Sigma}_y + \tfrac{\nu_y n_y}{\nu_y+n_y}(\widehat{\mu}_y - \mathbf{m}_y)(\widehat{\mu}_y - \mathbf{m}_y)^T,\end{aligned} \tag{43}$$

where $\widehat{\Sigma}_y$ is the usual sample covariance of training points in class y ($\widehat{\Sigma}_y = 0$ if $n_y \leq 1$). The posteriors are proper if $\nu_y^* > 0$, $\kappa^* > D-1$ and $\mathbf{S}^* \succ 0$.

The effective density for class y is multivariate student t with $k = \kappa^* - D + 1$ degrees of freedom, location vector $\mathbf{m}_y^*$, and scale matrix $\frac{\nu_y^*+1}{k\nu_y^*}\mathbf{S}^*$ [13]. In other words,

$$f(\mathbf{x} \mid y, S) \sim t\left(k, \mathbf{m}_y^*, \frac{\nu_y^*+1}{k\nu_y^*}\mathbf{S}^*\right). \tag{44}$$

To find the BRE under a binary linear classifier of the form (33), let $P = (-1)^i g(\mathbf{X})$. Since P is an affine transformation of a multivariate student t random variable, it has a non-standardized student t distribution [31]:

$$f(p \mid y, S) \sim t\left(k, m_{iy}, \frac{\nu_y^*+1}{k\nu_y^*}\gamma^2\right), \tag{45}$$

where $m_{iy} = (-1)^i g(\mathbf{m}_y^*)$ and $\gamma^2 = \mathbf{a}^T\mathbf{S}^*\mathbf{a}$. The CDF of a non-standardized student t distribution with d degrees of freedom, location parameter m, and scale parameter s^2 is well known, and at zero, it is given by [32],

$$\frac{1}{2} - \frac{\mathrm{sgn}(m)}{2} I\left(\frac{m^2}{m^2+ds^2}; \frac{1}{2}, \frac{d}{2}\right),$$

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

$$\widehat{\varepsilon}^{i,y}(\psi,S) = \frac{1}{2} - \frac{\mathrm{sgn}(m_{iy})}{2} I\left(\frac{m_{iy}^2}{m_{iy}^2 + \frac{\nu_y^*+1}{\nu_y^*}\gamma^2}; \frac{1}{2}, \frac{k}{2}\right). \tag{46}$$

This result was also found in [10].

The effective conditional density for $y = z$ is solved by updating all of the hyperparameters associated with class y with the new sample point, $\{\mathbf{x}, y\}$, resulting in:

$$f(\mathbf{w} \mid \mathbf{x}, y, y, S) \sim t\left(k+1, \mathbf{m}_y^* + \frac{\mathbf{x}-\mathbf{m}_y^*}{\nu_y^*+1}, \frac{\nu_y^*+2}{(k+1)(\nu_y^*+1)}\left(\mathbf{S}^* + \mathbf{S}_y(\mathbf{x})\right)\right), \tag{47}$$

where

$$\mathbf{S}_y(\mathbf{x}) = \frac{\nu_y^*}{\nu_y^*+1}(\mathbf{x}-\mathbf{m}_y^*)(\mathbf{x}-\mathbf{m}_y^*)^T. \tag{48}$$

For $y \neq z$, $f(\mathbf{w} \mid \mathbf{x}, y, z, S)$ is of the same form as the effective density with only hyperparameters associated with the covariance, κ^* and $\mathbf{S}^*$, updated:

$$f(\mathbf{w} \mid \mathbf{x}, y, z, S) \sim t\left(k+1, \mathbf{m}_z^*, \frac{\nu_z^*+1}{(k+1)\nu_z^*}\left(\mathbf{S}^* + \mathbf{S}_y(\mathbf{x})\right)\right). \tag{49}$$

To find the conditional MSE of the BRE, let $Q = (-1)^j g(\mathbf{W})$. For $y = z$,

$$f(q \mid \mathbf{x}, y, y, S) \sim t\left(k+1, m_{iy} + \frac{p-m_{iy}}{\nu_y^*+1}, \frac{\nu_y^*+2}{(k+1)(\nu_y^*+1)}\left(\gamma^2 + \frac{\nu_y^*}{\nu_y^*+1}\left(p-m_{iy}\right)^2\right)\right), \tag{50}$$

where we have used the fact that $(-1)^i\mathbf{a}^T(\mathbf{x}-\mathbf{m}_y^*) = p - m_y$. When $y \neq z$,

$$f(q \mid \mathbf{x}, y, z, S) \sim t\left(k+1, m_{jz}, \frac{\nu_z^*+1}{(k+1)\nu_z^*}\left(\gamma^2 + \frac{\nu_y^*}{\nu_y^*+1}\left(p-m_{iy}\right)^2\right)\right).$$

Since dependency on $\mathbf{X}$ has been reduced to dependency on only P in both of the above distributions, we may write $f(q \mid \mathbf{x}, y, z, S) = f(q \mid p, y, z, S)$ for all y and z. Lemma 1 in Appendix 3 produces an effective joint density given an effective density and an effective conditional density of a specified form. The distributions

$f(p \mid y, S)$ and $f(q \mid p, y, y, S)$ are precisely in the form required by this lemma with $D = 1$. Hence, $[P, Q]^T$ follows a bivariate student t distribution when $y = z$,

$$f(p, q \mid y, y, S) \sim t\left(k, \begin{bmatrix} m_{iy} \\ m_{iy} \end{bmatrix}, \frac{\gamma^2}{k} \begin{bmatrix} \frac{\nu_y^*+1}{\nu_y^*} & \frac{(-1)^{i+j}}{\nu_y^*} \\ \frac{(-1)^{i+j}}{\nu_y^*} & \frac{\nu_y^*+1}{\nu_y^*} \end{bmatrix}\right), \tag{51}$$

and when $y \neq z$,

$$f(p, q \mid y, z, S) \sim t\left(k, \begin{bmatrix} m_{iy} \\ m_{jz} \end{bmatrix}, \frac{\gamma^2}{k} \begin{bmatrix} \frac{\nu_y^*+1}{\nu_y^*} & 0 \\ 0 & \frac{\nu_z^*+1}{\nu_z^*} \end{bmatrix}\right).$$

Thus, $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,y}(\Theta_y) \mid S\right]$ can be found from (29). In particular, when $y = z$,

$$\begin{aligned} \mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,y}(\Theta_y) \mid S\right] &= \mathrm{P}(P \leq 0 \cap Q \leq 0 \mid y, y, S) \\ &= \mathbf{T}\left(-\frac{m_{iy}}{\gamma}\sqrt{\frac{k\nu_y^*}{\nu_y^*+1}}, -\frac{m_{jy}}{\gamma}\right. \\ &\quad \left.\sqrt{\frac{k\nu_y^*}{\nu_y^*+1}}, \frac{(-1)^{i+j}}{\nu_y^*+1}, k\right), \end{aligned} \tag{52}$$

and when $y \neq z$,

$$\begin{aligned} \mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,z}(\Theta_z) \mid S\right] &= \mathrm{P}(P \leq 0 \cap Q \leq 0 \mid y, z, S) \\ &= \mathbf{T}\left(-\frac{m_{iy}}{\gamma}\sqrt{\frac{k\nu_y^*}{\nu_y^*+1}}, -\frac{m_{jz}}{\gamma}\right. \\ &\quad \left.\sqrt{\frac{k\nu_z^*}{\nu_z^*+1}}, 0, k\right), \end{aligned} \tag{53}$$

where $\mathbf{T}(x, y, \rho, d)$ is the joint CDF of two standard multivariate student t random variables with correlation ρ and d degrees of freedom.

Independent arbitrary covariance

Assume $\Theta = [\mu_y, \Sigma_y]$, where the parameter space of μ_y is $\mathbb{R}^D$ and the parameter space of Σ_y consists of all symmetric positive definite matrices. The independent arbitrary covariance model assumes a conjugate prior with independent Θ_ys and

$$\pi(\theta_y) = \pi(\mu_y \mid \Sigma_y)\pi(\Sigma_y), \tag{54}$$

where $\pi(\mu_y \mid \Sigma_y)$ is of the same form as in (34) with hyperparameters $\nu_y \in \mathbb{R}$ and $\mathbf{m}_y \in \mathbb{R}^D$, and $\pi(\Sigma_y)$ is of the same form as in (42) with hyperparameters $\kappa_y \in \mathbb{R}$ and $\mathbf{S}_y$, a symmetric $D \times D$ matrix. The posterior is of the same form as the prior with updated hyperparameters given by (35) and

$$\begin{aligned} \kappa_y^* &= \kappa_y + n_y, \\ \mathbf{S}_y^* &= \mathbf{S}_y + (n_y - 1)\widehat{\Sigma}_y + \tfrac{\nu_y n_y}{\nu_y + n_y}(\widehat{\mu}_y - \mathbf{m}_y)(\widehat{\mu}_y - \mathbf{m}_y)^T. \end{aligned} \tag{55}$$

The posteriors are proper if $\nu_y^* > 0$, $\kappa_y^* > D - 1$ and $\mathbf{S}_y^* \succ 0$.

The effective density for class y is multivariate student t as in (44) with $k_y = \kappa_y^* - D + 1$ and $\mathbf{S}_y^*$ in place of k and $\mathbf{S}^*$, respectively [13]. Further, (45) also holds with $m_{iy} = (-1)^i g(\mathbf{m}_y^*)$ and with k_y and $\gamma_y^2 = \mathbf{a}^T\mathbf{S}_y^*\mathbf{a}$ in place of k and γ^2, respectively. Under binary linear classification, $\widehat{\varepsilon}^{i,y}(\psi, S)$ is given by (46) with k_y and γ_y^2 in place of k and γ^2. The same result was found in [10]. $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,y}(\Theta_y) \mid S\right]$ is solved similarly to before, resulting in (47), (50), (51), and ultimately (52), with k_y, $\mathbf{S}_y^*$ and γ_y^2 in place of k, $\mathbf{S}^*$, and γ^2, respectively. $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,z}(\Theta_z) \mid S\right]$ for $y \neq z$ is found from (25).

Homoscedastic scaled identity covariance

In the homoscedastic scaled identity covariance model, Σ_y is assumed to have a scaled identity structure, i.e., $\Theta_y = [\mu_y, \sigma^2]$ where $\Sigma_y = \sigma^2\mathbf{I}_D$ and $\mathbf{I}_D$ is a $D \times D$ identity matrix. The parameter space of μ_y is $\mathbb{R}^D$ for all y and of σ^2 is $(0, \infty)$. We also assume the μ_ys are mutually independent given σ^2:

$$\pi(\theta) = \left(\prod_{y=0}^{M-1} \pi(\mu_y \mid \sigma^2)\right)\pi(\sigma^2), \tag{56}$$

where $\pi(\mu_y \mid \sigma^2)$ is of the same form as (34) with hyperparameters ν_y and $\mathbf{m}_y$, and

$$\pi(\sigma^2) \propto \left|\sigma^2\right|^{-\frac{(\kappa+D+1)D}{2}} \exp\left(-\frac{\mathrm{trace}(\mathbf{S})}{2\sigma^2}\right), \tag{57}$$

with hyperparameters $\kappa \in \mathbb{R}$ and $\mathbf{S}$, a symmetric $D \times D$ real matrix. When $\nu_y > 0$, $\pi(\mu_y \mid \sigma^2)$ is a univariate Gaussian distribution with mean $\mathbf{m}_y$ and covariance Σ_y/ν_y, and when $(\kappa + D + 1)D > 2$ and $\mathbf{S} \succ 0$, $\pi(\sigma^2)$ is a univariate inverse-Wishart distribution. If in addition $(\kappa + D + 1)D > 4$, then $\mathrm{E}[\sigma^2] = \frac{\mathrm{trace}(\mathbf{S})}{(\kappa+D+1)D-4}$. The form of (57) has been designed so that the posterior is of the same form as the prior with the same hyperparameter update equations given in the arbitrary covariance models, (35) and (43). We require $\nu_y^* > 0$, $(\kappa^* + D + 1)D > 2$, and $\mathbf{S}^* \succ 0$ for a proper posterior.

The effective density for class y is multivariate student t with $k = (\kappa^* + D + 1)D - 2$ degrees of freedom [13]:

$$f(\mathbf{x} \mid y, S) \sim t\left(k, \mathbf{m}_y^*, \frac{\nu_y^* + 1}{k\nu_y^*}\mathrm{trace}(\mathbf{S}^*)\mathbf{I}_D\right). \tag{58}$$

Let $P = (-1)^i g(\mathbf{X})$. Since P is an affine transformation of a multivariate student t random variable, again it has the same form as in (45) with $k = (\kappa^* + D + 1)D - 2$, $m_{iy} = (-1)^i g(\mathbf{m}_y^*)$, and $\gamma^2 = \mathrm{trace}(\mathbf{S}^*)\mathbf{a}^T\mathbf{a}$. Following the same steps as in the homoscedastic arbitrary covariance model, under binary linear classification, $\widehat{\varepsilon}^{i,y}(\psi, S)$ is given

by (46) with the appropriate choice of k, m_{iy}, and γ^2. This was found in [10].

The effective conditional density for $y = z$ is solved by updating all of the hyperparameters associated with class y with the new sample point, $\{\mathbf{x}, y\}$:

$$f(\mathbf{w} \mid \mathbf{x}, y, y, S) \sim t\Big(k + D, \mathbf{m}_y^* + \frac{\mathbf{x} - \mathbf{m}_y^*}{\nu_y^* + 1}, \frac{\nu_y^* + 2}{(k + D)(\nu_y^* + 1)} \text{trace}(\mathbf{S}^* + \mathbf{S}_y(\mathbf{x}))\mathbf{I}_D\Big), \tag{59}$$

where $\mathbf{S}_y(\mathbf{x})$ is given by (48). When $y \neq z$, the effective conditional density is found by updating only hyperparameters associated with the covariance, κ^* and $\mathbf{S}^*$, with the point $\{\mathbf{x}, y\}$. Thus,

$$f(\mathbf{w} \mid \mathbf{x}, y, z, S) \sim t\left(k + D, \mathbf{m}_z^*, \frac{\nu_z^* + 1}{(k + D)\nu_z^*} \text{trace}(\mathbf{S}^* + \mathbf{S}_y(\mathbf{x}))\mathbf{I}_D\right). \tag{60}$$

Lemma 1 in Appendix 3 is used to find an effective joint density. When $y = z$,

$$f(\mathbf{x}, \mathbf{w} \mid y, y, S) \sim t\left(k, \begin{bmatrix} \mathbf{m}_y^* \\ \mathbf{m}_y^* \end{bmatrix}, \frac{\text{trace}(\mathbf{S}^*)}{k} \begin{bmatrix} \frac{\nu_y^* + 1}{\nu_y^*}\mathbf{I}_D & \frac{1}{\nu_y^*}\mathbf{I}_D \\ \frac{1}{\nu_y^*}\mathbf{I}_D & \frac{\nu_y^* + 1}{\nu_y^*}\mathbf{I}_D \end{bmatrix}\right), \tag{61}$$

and when $y \neq z$,

$$f(\mathbf{x}, \mathbf{w} \mid y, z, S) \sim t\left(k, \begin{bmatrix} \mathbf{m}_y^* \\ \mathbf{m}_z^* \end{bmatrix}, \frac{\text{trace}(\mathbf{S}^*)}{k} \begin{bmatrix} \frac{\nu_y^* + 1}{\nu_y^*}\mathbf{I}_D & \mathbf{0}_D \\ \mathbf{0}_D & \frac{\nu_z^* + 1}{\nu_z^*}\mathbf{I}_D \end{bmatrix}\right). \tag{62}$$

$\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,z}(\Theta_y) \mid S\right]$ can be found from (29) by defining $P = (-1)^i g(\mathbf{X})$ and $Q = (-1)^j g(\mathbf{W})$. Following the same steps as in the homoscedastic arbitrary covariance model, one can show that $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,z}(\Theta_y) \mid S\right]$ is equivalent to (52) when $y = z$ and (53) when $y \neq z$, where we plug in appropriate values for k, m_{iy} and γ^2.

Independent scaled identity covariance

Now assume that Σ_y has a scaled identity structure, i.e., $\Theta_y = [\mu_y, \sigma_y^2]$ where $\Sigma_y = \sigma_y^2 \mathbf{I}_D$, and that the parameter space of μ_y is $\mathbb{R}^D$ and of σ_y^2 is $(0, \infty)$ for all y. Also, assume the Θ_ys are mutually independent, with

$$\pi(\theta_y) = \pi(\mu_y \mid \sigma_y^2)\pi(\sigma_y^2), \tag{63}$$

where $\pi(\mu_y \mid \sigma_y^2)$ is of the same form as in (34) with hyperparameters $\nu_y \in \mathbb{R}$ and $\mathbf{m}_y \in \mathbb{R}^D$, and $\pi(\sigma_y^2)$ is of the same form as in (57) with hyperparameters $\kappa_y \in \mathbb{R}$ and $\mathbf{S}_y$, a symmetric $D \times D$ real matrix. The posterior is of the same form as the prior with the same hyperparameter update equations in (35) and (55). We require $\nu_y^* > 0$, $(\kappa_y^* + D + 1)D > 2$ and $\mathbf{S}_y^* \succ 0$ for a proper posterior.

The effective density for class y is multivariate student t, as in (58) with $k_y = (\kappa_y^* + D + 1)D - 2$ and $\mathbf{S}_y^*$ in place of k and $\mathbf{S}^*$, respectively [13]. Under binary linear classification, $\widehat{\varepsilon}^{i,y}(\psi, S)$ is given by (46) with $m_{iy} = (-1)^i g(\mathbf{m}_y^*)$ and with k_y and $\gamma_y^2 = \text{trace}(\mathbf{S}_y^*)\mathbf{a}^T\mathbf{a}$ in place of k and γ^2. The effective joint density, $f(\mathbf{x}, \mathbf{w} \mid y, y, S)$, is solved as before, resulting in (59) and (61) with k_y and $\mathbf{S}_y^*$ in place of k and $\mathbf{S}^*$, respectively. Further, $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,y}(\Theta_y) \mid S\right]$ is solved from (51) resulting in (52), with k_y and γ_y^2 in place of k and γ^2, respectively. $\mathrm{E}\left[\varepsilon^{i,y}(\Theta_y)\varepsilon^{j,z}(\Theta_z) \mid S\right]$ for $y \neq z$ is found from (25).

Appendix 3: Effective joint density lemma

The lemma below is used to derive the effective joint density of Gaussian models in Appendix 2.

Lemma 1. *Suppose* $\mathbf{X}$ *is multivariate student t given by,*

$$f(\mathbf{x}) \sim t\left(k, \mathbf{m}_y^*, \frac{\nu_y^* + 1}{k\nu_y^*}\gamma^2\mathbf{I}_D\right).$$

Further, suppose $\mathbf{W}$ *conditioned on* $\mathbf{X} = \mathbf{x}$ *is multivariate student t given by,*

$$f(\mathbf{w} \mid \mathbf{x}) \sim t\left(k + D, \mathbf{m}_z^* + I\frac{\mathbf{x} - \mathbf{m}_y^*}{\nu_y^* + 1}, \frac{1}{k + D}J\left(\gamma^2 + \frac{\nu_y^*}{\nu_y^* + 1}(\mathbf{x} - \mathbf{m}_y^*)^T(\mathbf{x} - \mathbf{m}_y^*)\right)\mathbf{I}_D\right),$$

where either $I = 0$ *and* $J = \frac{\nu_z^* + 1}{\nu_z^*}$, *or* $I = 1$ *and* $J = \frac{\nu_y^* + 2}{\nu_y^* + 1}$. *Then, the joint density is multivariate student t:*

$$f(\mathbf{x}, \mathbf{w}) \sim t\left(k, \begin{bmatrix} \mathbf{m}_y^* \\ \mathbf{m}_z^* \end{bmatrix}, \frac{\gamma^2}{k} \begin{bmatrix} \frac{\nu_y^* + 1}{\nu_y^*}\mathbf{I}_D & I\frac{1}{\nu_y^*}\mathbf{I}_D \\ I\frac{1}{\nu_y^*}\mathbf{I}_D & K\mathbf{I}_D \end{bmatrix}\right),$$

where $K = \frac{\nu_z^* + 1}{\nu_z^*}$ *when* $I = 0$ *and* $K = \frac{\nu_y^* + 1}{\nu_y^*}$ *when* $I = 1$.

Proof. After some simplification, one can show

$$\begin{aligned} f(\mathbf{x},\mathbf{w}) =& f(\mathbf{x})f(\mathbf{w}\,|\,\mathbf{x}) \\ \propto& \left(1+\frac{\nu_y^*}{\nu_y^*+1}(\mathbf{x}-\mathbf{m}_y^*)^T(\gamma^2\mathbf{I}_D)^{-1}(\mathbf{x}-\mathbf{m}_y^*)\right)^{-\frac{k+D}{2}} \\ &\times \left|\gamma^2\mathbf{I}_D+\frac{\nu_y^*}{\nu_y^*+1}(\mathbf{x}-\mathbf{m}_y^*)^T(\mathbf{x}-\mathbf{m}_y^*)\mathbf{I}_D\right|^{\frac{k+2D-1}{2}} \\ &\times \left|\gamma^2\mathbf{I}_D+\frac{\nu_y^*}{\nu_y^*+1}(\mathbf{x}-\mathbf{m}_y^*)^T(\mathbf{x}-\mathbf{m}_y^*)\mathbf{I}_D\right. \\ &+\frac{1}{J}\left(\mathbf{w}-\mathbf{m}_z^*-I\frac{\mathbf{x}-\mathbf{m}_y^*}{\nu_y^*+1}\right) \\ &\left.\times\left(\mathbf{w}-\mathbf{m}_z^*-I\frac{\mathbf{x}-\mathbf{m}_y^*}{\nu_y^*+1}\right)^T\right|^{-\frac{k+2D}{2}}. \end{aligned}$$

Simplifying further, we obtain

$$\begin{aligned} f(\mathbf{x},\mathbf{w}) \propto& \left(\gamma^2+\frac{\nu_y^*}{\nu_y^*+1}(\mathbf{x}-\mathbf{m}_y^*)^T(\mathbf{x}-\mathbf{m}_y^*)\right. \\ &+\frac{1}{J}\left(\mathbf{w}-\mathbf{m}_z^*-I\frac{\mathbf{x}-\mathbf{m}_y^*}{\nu_y^*+1}\right)^T \\ &\left.\times\left(\mathbf{w}-\mathbf{m}_z^*-I\frac{\mathbf{x}-\mathbf{m}_y^*}{\nu_y^*+1}\right)\right)^{-\frac{k+2D}{2}}. \end{aligned}$$

If $I=0$, then it can be shown that

$$f(\mathbf{x},\mathbf{w}) \propto \left(1+\begin{bmatrix}\mathbf{x}-\mathbf{m}_y^*\\ \mathbf{w}-\mathbf{m}_z^*\end{bmatrix}^T\Lambda^{-1}\begin{bmatrix}\mathbf{x}-\mathbf{m}_y^*\\ \mathbf{w}-\mathbf{m}_z^*\end{bmatrix}\right)^{-\frac{k+2D}{2}},$$

where

$$\Lambda=\begin{bmatrix}\frac{\nu_y^*+1}{\nu_y^*}\gamma^2\mathbf{I}_D & \mathbf{0}_D\\ \mathbf{0}_D & \frac{\nu_z^*+1}{\nu_z^*}\gamma^2\mathbf{I}_D\end{bmatrix}.$$

Similarly, if $I=1$, it can be shown that

$$f(\mathbf{x},\mathbf{w}) \propto \left(1+\begin{bmatrix}\mathbf{x}-\mathbf{m}_y^*\\ \mathbf{w}-\mathbf{m}_z^*\end{bmatrix}^T\Lambda^{-1}\begin{bmatrix}\mathbf{x}-\mathbf{m}_y^*\\ \mathbf{w}-\mathbf{m}_z^*\end{bmatrix}\right)^{-\frac{k+2D}{2}},$$

where

$$\Lambda=\begin{bmatrix}\frac{\nu_y^*+1}{\nu_y^*}\gamma^2\mathbf{I}_D & \frac{1}{\nu_y^*}\gamma^2\mathbf{I}_D\\ \frac{1}{\nu_y^*}\gamma^2\mathbf{I}_D & \frac{\nu_y^*+1}{\nu_y^*}\gamma^2\mathbf{I}_D\end{bmatrix},$$

which completes the proof. □

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
LAD and MRY contributed to the main idea, designed and implemented the algorithms, designed and carried out the simulation, analyzed the results, and drafted the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
The results published here are in part based upon data generated by The Cancer Genome Atlas (TCGA) established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov. The work of LAD is supported by the National Science Foundation (CCF-1422631 and CCF-1453563).

Received: 10 May 2015 Accepted: 21 October 2015
Published online: 24 October 2015

References
1. ER Dougherty, A Zollanvari, UM Braga-Neto, The illusion of distribution-free small-sample classification in genomics. Curr. Genomics. **12**(5), 333–341 (2011)
2. UM Braga-Neto, ER Dougherty, Is cross-validation valid for small-sample microarray classification? Bioinformatics. **20**(3), 374–380 (2004)
3. B Hanczar, J Hua, ER Dougherty, Decorrelation of the true and estimated classifier errors in high-dimensional settings. EURASIP J. Bioinforma. Syst. Biol. **2007**(Article ID 38473), 12 (2007)
4. UM Braga-Neto, ER Dougherty, Exact performance of error estimators for discrete classifiers. Pattern Recogn. **38**(11), 1799–1814 (2005)
5. MR Yousefi, J Hua, C Sima, ER Dougherty, Reporting bias when using real data sets to analyze classification performance. Bioinormatics. **26**(1), 68 (2010)
6. MR Yousefi, J Hua, ER Dougherty, Multiple-rule bias in the comparison of classification rules. Bioinformatics. **27**(12), 1675–1683 (2011)
7. MR Yousefi, ER Dougherty, Performance reproducibility index for classification. Bioinformatics. **28**(21), 2824–2833 (2012)
8. L Devroye, L Gyorfi, G Lugosi, *A probabilistic theory of pattern recognition. Stochastic modelling and applied probability*. (Springer, New York, 1996)
9. LA Dalton, ER 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. **59**(1), 115–129 (2011)
10. LA Dalton, ER Dougherty, 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. **59**(1), 130–144 (2011)
11. LA Dalton, ER Dougherty, Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error–part I: representation. IEEE Trans. Signal Process. **60**(5), 2575–2587 (2012)
12. LA Dalton, ER Dougherty, Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error–part II: consistency and performance analysis. IEEE Trans. Signal Process. **60**(5), 2588–2603 (2012)
13. LA Dalton, ER Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework–part I: discrete and Gaussian models. Pattern Recog. **46**(5), 1301–1314 (2013)
14. LA Dalton, ER Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework–part II: properties and performance analysis. Pattern Recog. **46**(5), 1288–1300 (2013)
15. B Hanczar, J Hua, C Sima, J Weinstein, M Bittner, ER Dougherty, Small-sample precision of ROC-related estimates. Bioinformatics. **26**, 822–830 (2010)
16. H Xu, C Caramanis, S Mannor, S Yun, in *Proceedings of the 48th IEEE Conference on Decision and Control, CDC 2009*. Risk sensitive robust support vector machines (IEEE New York, 2009), pp. 4655–4661
17. H Xu, C Caramanis, S Mannor, Robustness and regularization of support vector machines. J. Mach. Learn. Res. **10**, 1485–1510 (2009)
18. CM Bishop, *Pattern recognition and machine learning vol. 4*. (Springer, New York, NY, 2006)
19. A Gelman, JB Carlin, HS Stern, DB Rubin, *Bayesian data analysis vol. 2*, 3rd edn., (2014)
20. MS Esfahani, ER Dougherty, Incorporation of biological pathway knowledge in the construction of priors for optimal Bayesian classification. IEEE/ACM Trans. Comput. Biol. Bioinform. **11**(1), 202–218 (2014)
21. LA Dalton, ER Dougherty, Application of the Bayesian MMSE estimator for classification error to gene expression microarray data. Bioinformatics. **27**(13), 1822–1831 (2011)

22. BE Boser, IM Guyon, VN Vapnik, in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92*. A training algorithm for optimal margin classifiers (ACM, New York, NY, USA, 1992), pp. 144–152
23. C Cortes, V Vapnik, Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
24. C-C Chang, C-J Lin, LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27–12727 (2011)
25. B Efron, Bootstrap methods: another look at the jackknife. Ann. Stat. **7**(1), 1–26 (1979)
26. B Efron, RJ Tibshirani, *An introduction to the bootstrap*. (CRC Press, Boca Raton, FL, 1994)
27. B Efron, Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc. **78**(382), 316–331 (1983)
28. MJ van de Vijver, YD He, LJ van 't Veer, H Dai, AAM Hart, DW Voskuil, GJ Schreiber, JL Peterse, C Roberts, MJ Marton, M Parrish, D Atsma, A Witteveen, A Glas, L Delahaye, T van der Velde, H Bartelink, S Rodenhuis, ET Rutgers, SH Friend, R Bernards, A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. **347**(25), 1999–2009 (2002)
29. A Zollanvari, UM Braga-Neto, ER Dougherty, On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers. Pattern Recogn. **42**(11), 2705–2723 (2009)
30. JM Knight, I Ivanov, ER Dougherty, MCMC implementation of the optimal Bayesian classifier for non-Gaussian models: model-based RNA-Seq classification. BMC Bioinformatics. **15**(1), 401 (2014)
31. S Kotz, S Nadarajah, *Multivariate T distributions and their applications*. (Cambridge University Press, New York, 2004)
32. NL Johnson, S Kotz, N Balakrishnan, *Continuous univariate distributions vol. 2*, 2nd edn. (John Wiley & Sons, Hoboken, NJ, 1995)