EURASIP Journal on Bioinformatics and Systems Biology
a SpringerOpen Journal

REVIEW Open Access

Scientific knowledge is possible with small-sample classification

Edward R Dougherty[1,2*] and Lori A Dalton[3]

Abstract

A typical small-sample biomarker classification paper discriminates between types of pathology based on, say, 30,000 genes and a small labeled sample of less than 100 points. Some classification rule is used to design the classifier from this data, but we are given no good reason or conditions under which this algorithm should perform well. An error estimation rule is used to estimate the classification error on the population using the same data, but once again we are given no good reason or conditions under which this error estimator should produce a good estimate, and thus we do not know how well the classifier should be expected to perform. In fact, virtually, in all such papers the error estimate is expected to be highly inaccurate. In short, we are given no justification for any claims.

Given the ubiquity of vacuous small-sample classification papers in the literature, one could easily conclude that scientific knowledge is impossible in small-sample settings. It is not that thousands of papers overtly claim that scientific knowledge is impossible in regard to their content; rather, it is that they utilize methods that preclude scientific knowledge. In this paper, we argue to the contrary that scientific knowledge in small-sample classification is possible provided there is sufficient prior knowledge. A natural way to proceed, discussed herein, is via a paradigm for pattern recognition in which we incorporate prior knowledge in the whole classification procedure (classifier design and error estimation), optimize each step of the procedure given available information, and obtain theoretical measures of performance for both classifiers and error estimators, the latter being the critical epistemological issue. In sum, we can achieve scientific validation for a proposed small-sample classifier and its error estimate.

Review

Introduction

It is implicit in the title of this paper that one can entertain the possibility that scientific knowledge is impossible with small-sample classification. In fact, not only might one entertain this impossibility, but perusal of the related literature would most likely lead one to seriously consider that impossibility. It is not that thousands of papers overtly claim that scientific knowledge is impossible with regards to their content; rather, it is that they utilize methods that, *ipso facto*, cannot lead to knowledge. Even though it appears to be almost universally, if tacitly, assumed that scientific knowledge is impossible with small-sample classification - otherwise, why do so many not aspire to such knowledge - we argue to the contrary in this paper that scientific knowledge is possible. But before we make our case, let us examine in more detail why the literature may lead one to believe otherwise.

Consider the following common motif for a small-sample-classification paper, for instance, one proposing a classifier based on gene expression to discriminate types of pathology, stages of a disease, duration of survival, or some other phenotypic difference. Beginning with 30,000 features (genes) and less than 100 labeled sample points (microarrays), some classification rule (algorithm) is selected, perhaps an old one or a new one proposed in the paper. We are given no good reason why this algorithm should perform well. The classification rule is applied to the data and, using the same data, an error estimation rule is used to estimate the classification error on the population, meaning in practice the error rate on future observations. Once again, we are given no good reason why this error estimator should produce a good estimate; in fact, virtually, in all such papers, from what we know about the error estimation rule we would expect

*Correspondence: edward@ece.tamu.edu
[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[2]Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA
Full list of author information is available at the end of the article

the estimate to be inaccurate. At this point, one of two claims is made. If the classification rule is a well-known rule and the purpose of the paper is to produce a classifier for application (say, a biomarker panel), we are told that the authors have achieved their goal of finding such a classifier and its accuracy is validated by the error estimate. If, on the other hand, the purpose is to devise a new classification rule, we are told that the efficacy of the new rule has been validated by its performance, as measured by the error estimate or, by several such error estimates on several different data sets. In either case, we are given no justification for the validation claim. Moreover, in the second case, we are not told the conditions under which the classification rule should be expected to perform well or how well it should be expected to perform.

Amid all of this vacuity, perhaps the reporting of error estimates whose accuracy is a complete mystery is the most puzzling from a scientific perspective. To borrow a metaphor [1], one can imagine Harold Cramér leisurely sailing on the Baltic off the coast of Stockholm, taking in the sights and sounds of the sea, when suddenly a gene-expression classifier to detect prostate cancer pops into his head. No classification rule has been applied, nor is that necessary. All that matters is that Cramér's imagination has produced a classifier that operates on the feature-label distribution of interest with a sufficiently small error rate. Since scientific validity depends on the predictive capacity of a model, while an appropriate classification rule is certainly beneficial to classifier design, epistemologically, the error rate is paramount. Were we to know the feature-label distribution of interest, we could exactly determine the error rate of the proposed classifier. Absent knowledge of the feature-label distribution, the actual error must be estimated from data and the accuracy of the estimate judged from the performance of the error estimation rule employed. Consequently, any paper that applies an error estimation rule without providing a performance characterization relevant to the data at hand is scientifically vacuous. Given the near universality of vacuous small-sample classification papers in the literature, one could easily reach the conclusion that scientific knowledge is impossible in small-sample settings. Of course, this would beg the question of why people are writing vacuous papers and why journals are publishing them. Since the latter are sociological questions, they are outside the domain of the current paper. We will focus on the scientific issues.

Epistemological digression

Before proceeding, we digress momentarily for some very brief comments regarding scientific epistemology (referring to [2] for a comprehensive treatise and to [3] for a discussion aimed at biology and including classifier validity). Our aim is narrow, simply to emphasize the role of prediction in scientific knowledge, not to indulge in broad philosophical issues.

A scientific theory consists of two parts: (1) a *mathematical model* composed of symbols (variables and relations between the variables), and (2) a set of *operational definitions* that relate the symbols to data. A mathematical model alone does not constitute a scientific theory. The formal mathematical structure must yield experimental predictions in accord with experimental observations. As put succinctly by Richard Feynman, "It is whether or not the theory gives predictions that agree with experiment. It is not a question of whether a theory is philosophically delightful, or easy to understand, or perfectly reasonable from the point of view of common sense" [4]. Model validity is characterized by predictive relations, without which the model lacks empirical content. Validation requires that the symbols be tied to observations by some semantic rules that relate not necessarily to the general principles of the mathematical model themselves but to conclusions drawn from the principles. There must be a clearly defined tie between the mathematical model and experimental methodology. Philipp Frank writes, "Reichenbach had explicitly pointed out that what is needed is a bridge between the symbolic system of axioms and the protocols of the laboratory. But the nature of this bridge had been only vaguely described. Bridgman was the first who said precisely that these *relations of coordination* consist in the description of physical operations. He called them, therefore, *operational definitions*" [5]. Elsewhere, we have written, "Operational definitions are required, but their exact formulation in a given circumstance is left open. Their specification constitutes an epistemological issue that must be addressed in mathematical (including logical) statements. Absent such a specification, a purported scientific theory is meaningless" [6].

The validity of a scientific theory depends on the choice of validity criteria and the mathematical properties of those criteria. The observational measurements and the manner in which they are to be compared to the mathematical model must be formally specified. The validity of a theory is relative to this specification, but what is not at issue is the necessity of a set of relations tying the model to operational measurements. Formal specification is mandatory and this necessarily takes the form of mathematical (including logical) statements. Formal specification is especially important in stochastic settings where experimental outcomes reflect the randomness of the stochastic system so that one must carefully define how the outcomes are to be interpreted.

Story telling and intuitive arguments cannot suffice. Not only is complex-system behavior often unintuitive, but stochastic processes and statistics often contradict naïve probabilistic notions gathered from simple experiments like rolling dice. Perhaps even worse is an appeal

to pretty pictures drawn with computer software. The literature abounds with data partitioned according to some clustering algorithm whose partitioning performance is unknown or, even more strangely, justified by some "validation index" that is poorly, if at all, correlated with the error rate of the clustering algorithm [7]. The pretty pictures are usually multi-colored and augmented with all kinds of attractive-looking symbols. They are inevitably followed by some anecdotal commentary. Although all of this may be delightful, it is scientifically meaningless. Putting the artistic touches and enormous calculations aside, all we are presented with is a radical empiricism. Is there any knowledge here? Hans Reichenbach answers, "A mere report of relations observed in the past cannot be called knowledge. If knowledge is to reveal objective relations of physical objects, it must include reliable predictions. A radical empiricism, therefore, denies the possibility of knowledge" [2]. A collection of measurements together with a commentary on the measurements is not scientific knowledge. Indeed, the entire approach "denies the possibility of knowledge," so that its adoption constitutes a declaration of meaninglessness.

Classification error

For two-class classification, the population is characterized by a feature-label distribution F for a random pair $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a vector of D features and Y is the binary label, 0 or 1, of the class containing $\mathbf{X}$. A classifier is a function, ψ, which assigns a binary label, $\psi(\mathbf{X})$, to each feature vector. The error, $\varepsilon[\,\psi]$, of ψ is the probability, $P(\psi(\mathbf{X}) \neq Y)$, that ψ yields an erroneous label. A classifier with minimum error among all classifiers is known as a *Bayes classifier* for the feature-label distribution. The minimum error is called the *Bayes error*. Epistemologically, the error is the key issue since it quantifies the predictive capacity of the classifier.

Abstractly, any pair $\mathcal{M} = (\psi, \varepsilon_\psi)$ composed of a function $\psi : \mathbb{R}^D \to \{0,1\}$ and a real number $\varepsilon_\psi \in [0,1]$ constitutes a *classifier model*, with ε_ψ being simply a number, not necessarily specifying an actual error probability corresponding to ψ. $\mathcal{M}$ becomes a scientific model when it is applied to a feature-label distribution. In practice, the feature-label distribution is unknown and a *classification rule* Ψ_n is used to design a classifier ψ_n from a random sample $S_n = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)\}$ of pairs drawn from the feature-label distribution. Note that a classification rule is a sequence of rules depending on the sample size n. If feature selection is involved, then it is part of the classification rule. A designed classifier produces a classifier model, namely, $(\psi_n, \varepsilon[\,\psi_n])$. Since the true classifier error $\varepsilon[\,\psi_n]$ depends on the feature-label distribution, which is unknown, $\varepsilon[\,\psi_n]$ is unknown. The true error must be estimated by an *estimation rule*, Ξ_n. Thus, the random sample S_n yields a classifier $\psi_n = \Psi_n(S_n)$ and an error estimate $\hat{\varepsilon}[\,\psi_n] = \Xi_n(S_n)$, which together constitute a classifier model $(\psi_n, \hat{\varepsilon}[\,\psi_n])$. Overall, classifier design involves a *rule model* (Ψ_n, Ξ_n) used to determine a sample-dependent classifier model $(\psi_n, \hat{\varepsilon}[\,\psi_n])$. Both $(\psi_n, \varepsilon[\,\psi_n])$ and $(\psi_n, \hat{\varepsilon}[\,\psi_n])$ are random pairs relative to the sampling distribution.

Given a feature-label distribution, error estimation accuracy is commonly measured by the *mean-square error* (*MSE*), defined by $\mathrm{MSE}(\hat{\varepsilon}) = \mathrm{E}[(\hat{\varepsilon} - \varepsilon)^2]$, where for notational ease we denote $\varepsilon[\,\psi_n]$ and $\hat{\varepsilon}[\,\psi_n]$ by ε and $\hat{\varepsilon}$, respectively, or, equivalently, by the square root of the MSE, known as the *root-mean-square* (*RMS*). The expectation used here is relative to the sampling distribution induced by the feature-label distribution. The MSE is decomposed into the bias, $\mathrm{Bias}(\hat{\varepsilon}) = \mathrm{E}[\hat{\varepsilon} - \varepsilon]$, of the error estimator relative to the true error, and the deviation variance, $\mathrm{Var}_{\mathrm{dev}}(\hat{\varepsilon}) = \mathrm{Var}(\hat{\varepsilon} - \varepsilon)$, by

$$\mathrm{MSE}(\hat{\varepsilon}) = \mathrm{Var}_{\mathrm{dev}}(\hat{\varepsilon}) + \mathrm{Bias}(\hat{\varepsilon})^2. \tag{1}$$

When a large amount of data is available, the sample can be split into independent training and test sets, the classifier being designed on the training data and its error being estimated by the proportion of errors on the test data, which is known as the holdout estimator. For holdout, we have the distribution-free bound $\mathrm{RMS}(\widehat{\varepsilon}_{\mathrm{holdout}}|S_{n-m}, F) \leq 1/\sqrt{4m}$, where m is the size of the test sample, S_{n-m} is the training sample and F is any feature-label distribution [8]. $\mathrm{RMS}(\widehat{\varepsilon}|Z)$ indicates that the expectation in the RMS is conditioned on the random vector Z. But when data are limited, the sample cannot be split without leaving too little data to design a good classifier. Hence, training and error estimation must take place on the same data set.

The consequences of training-set error estimation are readily explained by the following formula for the deviation variance:

$$\mathrm{Var}_{\mathrm{dev}}(\hat{\varepsilon}) = \sigma_{\hat{\varepsilon}}^2 + \sigma_\varepsilon^2 - 2\rho\sigma_{\hat{\varepsilon}}\sigma_\varepsilon, \tag{2}$$

where $\sigma_{\hat{\varepsilon}}^2, \sigma_\varepsilon^2$, and ρ are the variance of the error estimate, the variance of the error, and the correlation between the estimated and true errors, respectively. The deviation variance is driven down by small variances or a correlation coefficient near 1.

Consider the popular cross-validation error estimator. For it, the error is estimated on the training data by randomly splitting the training data into k folds (subsets), S_n^i, for $i = 1, 2, ..., k$, training k classifiers on $S_n - S_n^i$, for $i = 1, 2, ..., k$, calculating the proportion of errors of each designed classifier on the appropriate left-out fold, and then averaging these proportions to obtain the cross-validation estimate of the originally designed classifier. Various enhancements are made, such as by repeating the process some number of times and averaging. Letting $k = n$ yields the leave-one-out estimator. The problem with cross-validation is evident from (2): for small samples,

it has large variance and little correlation with the true error. Hence, although with small folds, cross-validation does not suffer too much from bias, it typically has large deviation variance.

To illustrate the matter, we reproduce an example from [9] based on real patient data from a study involving microarrays prepared with RNA from breast tumor specimens from 295 patients, 115 and 180 belonging to the good-prognosis and poor-prognosis classes, respectively. The dataset is reduced to the 2,000 genes with highest variance, these are reduced to 10 via t test feature selection, and a classifier is designed using linear discriminant analysis (LDA). In the simulations, the data are split into two sets. The first set, consisting of 50 examples drawn without replacement from the full dataset, is used for both training and error estimation via leave-one-out cross-validation. The remaining examples are used as a hold-out test set to get an accurate estimate of the true error, which is taken as the true error. There is an assumption that such a hold-out size will give an accurate estimate of the true error. This procedure is repeated 10,000 times. Figure 1 shows the scatter plot for the pairs of true and estimated errors, along with the linear regression of the true error on the estimated error. The means are shown on the axes. What we observe is typical for small samples: large variance and negligible regression between the true and estimated errors [10]. Indeed, one even sees negatively sloping regression lines for cross-validation and bootstrap (another resampling error estimator), and negative correlation between the true and cross-validation estimated errors has been mathematically demonstrated in some basic models [11]. Such error estimates are worthless and can lead to a huge waste of resources in trying to reproduce them [9].

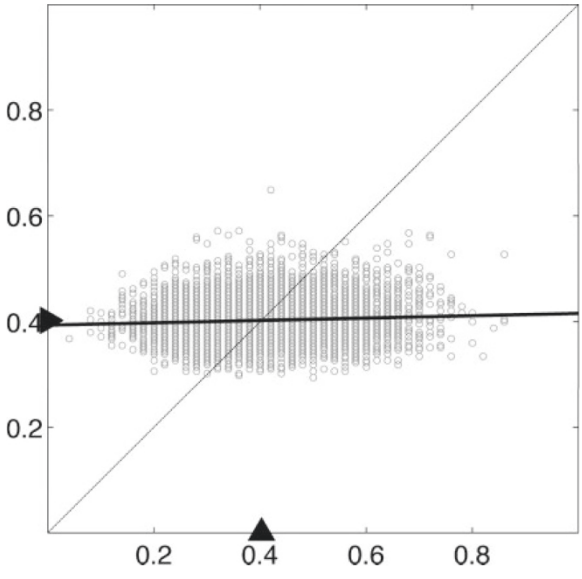

Figure 1 Linear regression between cross-validation and the true error. Scatter plot and linear regression for cross-validation (horizontal axis) and the true error (vertical axis) with sample size 50 for linear discrimination between two classes of breast cancer patients.

RMS bounds

Suppose a sample is collected, a classification rule Ψ_n applied, and the classifier error estimated by an error-estimation rule Ξ_n to arrive at the classifier model $(\psi_n, \hat{\varepsilon}[\,\psi_n])$. If no assumptions are posited regarding the feature-label distribution, then the entire procedure is completely distribution-free. There are three possibilities. First, if no validity criterion is specified, then the classifier model is *ipso facto* epistemologically meaningless. Second, if a validity criterion is specified, say RMS, and no distribution-free results are known about the RMS for Ψ_n and Ξ_n, then again the model is meaningless. Third, if there exist distribution-free RMS bounds concerning Ψ_n and Ξ_n, then these bounds can, in principle, be used to quantify the performance of the error estimator and thereby quantify model validity.

Regarding the third possibility, the following is an example of a distribution-free RMS bound for the leave-one-out error estimator with the discrete histogram rule and tie-breaking in the direction of class 0 [8]:

$$\mathrm{RMS}(\widehat{\varepsilon}_{\mathrm{loo}}|F) \leq \sqrt{\frac{1+6/e}{n} + \frac{6}{\sqrt{\pi\,(n-1)}}}, \tag{3}$$

where F is any feature-label distribution. Although this bound holds for all distributions, it is useless for small samples: for $n = 200$ this bound is 0.506. In general, there are very few cases in which distribution-free bounds are known and, when they are known, they are useless for small samples.

Distribution-based bounds are needed. These require knowledge of the RMS, which means knowledge concerning the second-order moments of the joint distribution between the true and estimated errors. More generally, to fully understand an error estimator we need to know its joint distribution with the true error. Oddly, this problem has historically been ignored in pattern recognition, notwithstanding the fact that error estimation is the epistemological ground for classification. Going back to the 1970s there were some results on the mean and variance of some error estimators for the Gaussian model using LDA. In 1966, Hills obtained the expected value of the resubstitution and plug-in estimators in the univariate model with known common variance [12]. The resubstitution estimate is simply a count of the classification errors on the training data and the plug-in estimate is found by using the data to estimate the feature-label distribution and then finding the error of the designed classifier

on the estimated distribution. In 1972, Foley obtained the expected value of resubstitution in the multivariate model with known common covariance matrix [13]. In 1973, Sorum derived results for the expected value and variance for both resubstitution and leave-one-out in the univariate model with known common variance [14]. In 1973, McLachlan derived an asymptotic representation for the expected value of resubstitution in the multivariate model with unknown common covariance matrix [15]. In 1975, Moran obtained new results for the expected value of resubstitution and plug-in for the multivariate model with known covariance matrix [16]. In 1977, Goldstein and Wolf obtained the expected value of resubstitution for multinomial discrimination [17]. Following the latter, there was a gap of 15 years before Davison and Hall derived asymptotic representations for the expected value and variance of bootstrap and leave-one-out in the univariate Gaussian model with unknown and possibly different covariances [18]. This is the only paper we know of providing analytic results for moments of common error estimators between 1977 and 2005. None of these papers provided representation of the joint distribution or representation of second-order mixed moments, which are needed for the RMS.

This problem has only recently been addressed beginning in 2005, in particular, for the resubstitution and leave-one-out estimators. For the multinomial model, complete enumeration was used to obtain the marginal distributions for the error estimators [11] and then the joint distributions [19]. Exact closed-form representations for second-order moments, including the mixed moments, were obtained, thereby obtaining exact RMS representations for both estimators [11]. For the Gaussian model using LDA in 2009, we obtained the exact marginal distributions for both estimators in the univariate model (known but not necessarily equal class variances) and approximations in the multivariate model (known and equal class covariance matrices) [20]. Subsequently, these were extended to the joint distributions for the true and estimated errors in a Gaussian model [21]. Recently exact closed-form representations for the second-order moments in the univariate model without assuming equal covariances were discovered, thereby providing exact expression of the RMS for both estimators [22]. Moreover, double asymptotic representations for the second-order moments in the multivariate model, sample size and dimension approaching infinity at a fixed rate between the two, were found, thereby providing double asymptotic expressions for the RMS [23]. Finite sample approximations from the double asymptotic method have been shown to possess better accuracy than various simple asymptotic representations (although much more work is needed on this issue) [24,25].

Validity

Let us now consider validity. An obvious way to proceed would be to say that a classifier model (ψ, ε_ψ) is valid for the feature-label distribution F to the extent that ε_ψ approximates the classifier error, $\varepsilon[\psi]$, on F, where the degree of approximation is measured by some distance between ε_ψ and $\varepsilon[\psi]$. For a classifier ψ_n designed from a specific sample, this would mean that we want to measure some distance between $\varepsilon = \varepsilon[\psi_n]$ and $\hat{\varepsilon} = \hat{\varepsilon}[\psi_n]$, say $|\varepsilon - \hat{\varepsilon}|$. To do this, we would have to know the true error and to know that we would need to know F. But if we knew F, we would use the Bayes classifier and would not need to design a classifier from sample data. Since it is the precision of the error estimate that is of consequence, a natural way to proceed would be to characterize validity in terms of the precision of the error estimator $\hat{\varepsilon}[\psi_n] = \Xi_n(S_n)$ as an estimator of $\varepsilon[\psi_n]$, say by $\mathrm{RMS}(\hat{\varepsilon})$. This makes sense because both the true and estimated errors are random functions of the sample and the RMS measures their closeness across the sampling distribution. But again there is a catch: the RMS depends on F, which we do not know. Thus, given the sample without knowledge of F, we cannot compute the RMS.

To proceed, prior knowledge is required, in the sense that we need to assume that the actual (unknown) feature-label distribution belongs to some *uncertainty class*, $\mathcal{U}$, of feature-label distributions. Once RMS representations have been obtained for feature-label distributions in $\mathcal{U}$, distribution-based RMS bounds follow: $\mathrm{RMS}(\hat{\varepsilon}) \leq \max_{G \in \mathcal{U}} \mathrm{RMS}(\hat{\varepsilon}|G)$, where $\mathrm{RMS}(\hat{\varepsilon}|G)$ is the RMS of the error estimator under the assumption that the feature-label distribution is G. We do not know the actual feature-label distribution precisely, but prior knowledge allows us to bound the RMS. For instance, consider using LDA with a feature-label distribution having two equally probable Gaussian class-conditional densities sharing a known covariance matrix. For this model the Bayes error is a one-to-one decreasing function of the distance, m, between the means. Figure 2a shows the RMS to be a one-to-one increasing function of the Bayes error for leave-one-out for dimension $D = 10$ and sample sizes $n = 20, 40, 60$, the RMS and Bayes errors being on the y and x axes, respectively.

Assuming a parameterized model in which the RMS is an increasing function of the Bayes error, $\varepsilon_{\mathrm{bay}}$, we can pose the following question: Given sample size n and $\lambda > 0$, what is the maximum value, $\mathrm{maxBayes}(\lambda)$, of the Bayes error such that $\mathrm{RMS}(\hat{\varepsilon}) \leq \lambda$? If RMS is the measure of validity and λ represents the largest acceptable RMS for the classifier model to be considered meaningful, then the epistemological requirement is characterized by $\mathrm{maxBayes}(\lambda)$. Given the relationship between model parameters and the Bayes error, the inequality $\varepsilon_{\mathrm{bay}} \leq \mathrm{maxBayes}(\lambda)$ can be solved in terms of the parameters to

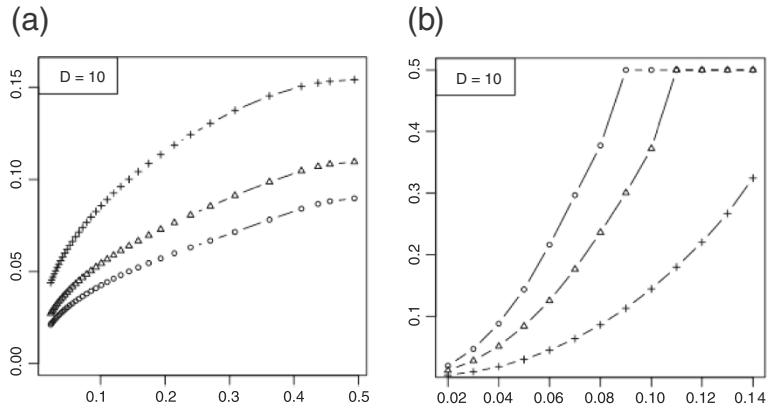

Figure 2 RMS and maxBayes(λ). (a) RMS (*y*-axis) as a function of the Bayes error (*x*-axis) for leave-one-out with dimension $D = 10$ and sample sizes $n = 20$ (plus sign), 40 (triangle), 60 (circle); **(b)** maxBayes(λ) curves corresponding to the RMS curves in part **(a)**.

arrive at a necessary modeling assumption. In the preceding Gaussian example, since ε_{bay} is a decreasing function of m, we obtain an inequality $m \geq m(\lambda)$. Figure 2b shows the maxBayes(λ) curves corresponding to the RMS curves in Figure 2a [26]. These curves show that, assuming Gaussian class-conditional densities and a known common covariance matrix, further assumptions must be made to insure that the RMS is sufficiently small to make the classifier model meaningful.

To have scientific content, small-sample classification requires prior knowledge. Regarding the feature-label distribution, there are two extremes: (1) the feature-label distribution is known, in which case the entire classification problem collapses to finding the Bayes classifier and Bayes error, so there is no classifier design or error estimation issue; and (2) the uncertainty class consists of all feature-label distributions, the distribution-free case, and we typically have no bound, or one that is too loose for practice. In the middle ground, there is a trade-off between the size of the uncertainty class and the size of the sample. The uncertainty class must be sufficiently constrained (equivalently, the prior knowledge must be sufficiently great) that an acceptable bound can be achieved with an acceptable sample size.

MMSE error estimation

Given that one needs a distributional model to achieve useful performance bounds for classifier error estimation, an obvious course of action is to find or define a prior over the uncertainty class of feature-label distributions, and then find an optimal minimum-mean-square-error (MMSE) error estimator relative to that class [27]. This results in a Bayesian approach with the uncertainty class being given a prior distribution and the data being used to construct a posterior distribution, which quantifies everything we know about the feature-label distribution. Benefits of the Bayesian approach are (1) we can incorporate prior knowledge in the whole classification procedure (classifier design and error estimation), which, as we have argued above, is desperately needed in a small-sample setting where the data provide only a meager amount of information; (2) given the mathematical framework, we can optimize each step of the procedure, further addressing the poor performance suffered in small samples; and (3) we can obtain theoretical measures of the performance for both arbitrary classifiers (via the MMSE error estimator) and arbitrary error estimators (via the sample conditioned MSE), perhaps the most important advantage epistemologically. We begin with an overview of optimal MMSE error estimation.

Assume that a sample point has a prior probability c of coming from class 0, and that the class-0 conditional distribution is parameterized by θ_0 and class 1 is parameterized by θ_1. Considering both classes, our model is completely parameterized by $\theta = \{c, \theta_0, \theta_1\}$. Given a random sample, S_n, we design a classifier ψ_n and wish to minimize the MSE between its true error, ε (a function of θ and ψ_n), and an error estimate, $\widehat{\varepsilon}$ (a function of S_n and ψ_n). A key realization is that the expectation in the MSE may now be taken over the uncertainty class conditioned on the observed sample, rather than over the sampling distribution for a fixed (unknown) feature-label distribution. The MMSE error estimator is thus the expected true error, $\widehat{\varepsilon}(\psi_n, S_n) = \mathrm{E}_\theta[\,\varepsilon(\psi_n, \theta)|S_n]$. The expectation given the sample is over the posterior density of θ , denoted by $\pi^*(\theta)$. Thus, we write the Bayesian MMSE error estimator with the shorthand $\widehat{\varepsilon} = \mathrm{E}_{\pi^*}[\,\varepsilon]$.

The Bayesian error estimate is not guaranteed to be the optimal error estimate for any particular feature-label distribution but optimal for a given sample, and assuming the parameterized model and prior probabilities, it is both optimal on average with respect to MSE and unbiased

when averaged over all parameters and samples. These implications apply for any classification rule as long as the classifier is fixed given the sample. To facilitate analytic representations, we assume c, θ_0 and θ_1 are all mutually independent prior to observing the data. Denote the marginal priors of c, θ_0 and θ_1 by $\pi(c)$, $\pi(\theta_0)$ and $\pi(\theta_1)$, respectively, and suppose data are used to find each posterior, $\pi^*(c)$, $\pi^*(\theta_0)$ and $\pi^*(\theta_1)$, respectively. Independence is preserved, i.e., $\pi^*(c,\theta_0,\theta_1) = \pi^*(c)\pi^*(\theta_0)\pi^*(\theta_1)$ [27].

If ψ_n is a trained classifier given by $\psi_n(\mathbf{x}) = 0$ if $\mathbf{x} \in R_0$ and $\psi_n(\mathbf{x}) = 1$ if $\mathbf{x} \in R_1$, where R_0 and R_1 are measurable sets partitioning the sample space, then the true error of ψ_n under the distribution parameterized by θ may be decomposed as

$$\begin{aligned}\varepsilon(\psi_n,\theta) &= c\int_{R_1} f_{\theta_0}(\mathbf{x}|0)\,d\mathbf{x} + (1-c)\int_{R_0} f_{\theta_1}(\mathbf{x}|1)\,d\mathbf{x} \\ &= c\varepsilon^0(\psi_n,\theta_0) + (1-c)\varepsilon^1(\psi_n,\theta_1),\end{aligned} \tag{4}$$

where $f_{\theta_y}(\mathbf{x}|y)$ is the class-y conditional density assuming parameter θ_y is true and ε^y is the error contributed by class y. Owing to the posterior independence between c and θ_0 and between c and θ_1, the Bayesian MMSE error estimator can be expressed as [28]

$$\widehat{\varepsilon}(\psi_n,S_n) = \mathrm{E}_{\pi^*}[c]\,\mathrm{E}_{\pi^*}[\varepsilon^0] + (1-\mathrm{E}_{\pi^*}[c])\mathrm{E}_{\pi^*}[\varepsilon^1]. \tag{5}$$

With a fixed sample and classifier, and given θ_y, the true error, $\varepsilon^y(\psi_n,\theta_y)$, is deterministic. Thus, letting Θ_y be the parameter space of θ_y,

$$\mathrm{E}_{\pi^*}[\varepsilon^y] = \int_{\Theta_y} \varepsilon^y(\psi_n,\theta_y)\pi^*(\theta_y)d\theta_y. \tag{6}$$

Just as the true error for a fixed feature-label distribution is found from the class-conditional densities, $f_{\theta_y}(\mathbf{x}|y)$, the Bayesian MMSE error estimator for an uncertainty class can be found from *effective class-conditional densities*, which are derived by taking the expectations of the individual class-conditional densities with respect to the posterior distribution,

$$f(\mathbf{x}|y) = \int_{\Theta_y} f_{\theta_y}(\mathbf{x}|y)\,\pi^*\left(\theta_y\right)d\theta_y. \tag{7}$$

Specifically, we obtain an equation for the expected true error that parallels that of the true error in (4) [29]:

$$\widehat{\varepsilon}(\psi_n,S_n) = \mathrm{E}_{\pi^*}[c]\int_{R_1} f(\mathbf{x}|0)\,d\mathbf{x} + (1-\mathrm{E}_{\pi^*}[c])\int_{R_0} f(\mathbf{x}|1)\,d\mathbf{x}. \tag{8}$$

Application of Bayesian error estimation to real data, in particular gene-expression microarray data, has been addressed in [30]. This work provides C code implementing the Bayesian error estimator for Gaussian distributions and normal-inverse-Wishart priors for both linear classifiers, with exact closed-form representations, and non-linear classifiers, where closed form-solutions are not available and we instead implement a Monte-Carlo approximation. The code and a toolbox of related utilities are publicly available. In [30] we discuss the suitability of a Gaussian model with normal-inverse-Wishart priors for microarray data and propose a feature selection scheme employing a Shapiro-Wilk Gaussianity test to validate Gaussian modeling assumptions. Furthermore, we propose a methodology for calibrating normal-inverse-Wishart priors for microarray data based on a method-of-moments approach using features discarded by the feature-selection scheme.

Sample-conditioned MSE

The RMS of an error estimator is used to characterize the validity of a classifier model. As we have discussed, if we are in possession of RMS expressions for the feature-label distributions in an uncertainty class, we can bound the RMS, so as to insure a given level of performance. In the case of MMSE error estimation, the priors provide a mathematical framework that can be used for both the analysis of any error estimator and the design of estimators with desirable properties or optimal performance. The posteriors of the distribution parameters imply a (sample-conditioned) distribution on the true classifier error. This randomness in the true error comes from our uncertainty in the underlying feature-label distribution (given the sample). Within the assumed model, this sample-conditioned distribution of the true error contains the full information about error estimator accuracy and we may speak of moments of the true error (for a fixed sample and classifier), in particular the expectation, variance, and sample-conditioned MSE, as opposed to simply the MSE relative to the sampling distribution as in classical error estimation.

Finding the sample-conditioned MSE of MMSE Bayesian error estimators amounts to evaluating the variance of the true error conditioned on the observed sample [28]. The sample-conditioned MSE converges to zero almost surely in both discrete and Gaussian models provided in [31], where closed form expressions for the MSE are available. Further, the exact MSE for arbitrary error estimators falls out naturally in the Bayesian model. That is, if $\widehat{\varepsilon}_\bullet$ is a constant representing an arbitrary error estimate computed from the sample, then the MSE of $\widehat{\varepsilon}_\bullet$ can be evaluated directly from that of the Bayesian error estimator:

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

$\mathrm{MSE}(\widehat{\varepsilon}_\bullet|S_n)$, as well as its square root $\mathrm{RMS}(\widehat{\varepsilon}_\bullet|S_n)$, are minimized when $\widehat{\varepsilon} = \widehat{\varepsilon}_\bullet$.

In a classical approach, nothing is known given a sample, whereas in a Bayesian approach, the sample conditions uncertainty in the RMS and different samples may condition it to different extents. Figure 3 shows probability densities of the sample-conditioned RMS for both the leave-one-out estimator and Bayesian error estimator in a discrete model with $b = 16$ bins. The simulation generates 10,000 distributions drawn from a prior given in [31] and 1,000 samples from each distribution. The unconditional RMS (averaged over both distributions and samples) for both error estimators is also shown, as well as the distribution-free RMS bound on leave-one-out given in (3). In Figure 3, the RMS of the Bayesian error estimator tends to be very close to 0.05 whereas the leave-one-out error estimator has a long tail with substantial mass between 0.05 and 0.2, demonstrating that different samples can condition the RMS to a very significant extent. In addition, the unconditional RMS of the Bayesian error estimator is less than half that of leave-one-out, while Devroye's distribution-free bound on the unconditional RMS is too loose to be useful. Hence, not only does a Bayesian framework permit us to obtain an optimal error estimator and its RMS conditioned on the sample, but performance improvement can be significant.

In [31], a bound on the sample-conditioned RMS of the Bayesian error estimator is provided for the discrete model. With any classifier, beta priors on c and Dirichlet priors on the bin probabilities satisfying mild conditions, and given a sample S_n, $\mathrm{RMS}(\widehat{\varepsilon}_{\mathrm{BEE}}|S_n) \leq 1/\sqrt{4n}$. For comparison, consider the holdout bound $\mathrm{RMS}(\widehat{\varepsilon}_{\mathrm{holdout}}|S_{n-m}, F) \leq 1/\sqrt{4m}$, where m is the size of the test sample. Both bounds still hold if we remove the conditioning, and in this way they become comparable. Since $1/\sqrt{4n} \leq 1/\sqrt{4m}$, under a Bayesian model not only does using the full sample to train the classifier result in a lower true error, but we expect to achieve better RMS performance using training-data error estimation than we would by holding out the entire sample for error estimation. This is a testament to the power of modeling.

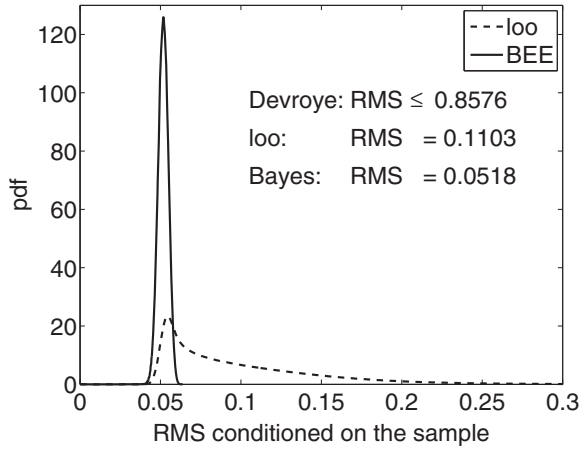

Figure 3 Sample-conditioned RMS probability densities. Probability densities for the sample-conditioned RMS of leave-one-out (dashed line) and the Bayesian error estimator (solid line) in a discrete model with $b = 16$ bins, prior probability $c = 0.5$, $n = 30$ training points, and an average true error of 0.25.

Optimal classification

Since prior knowledge is required to obtain a good error estimate in small-sample settings, an obvious course of action would be to utilize that knowledge for classifier design [29,32]. Whereas ordinary *Bayes classifiers* minimize the misclassification probability when the underlying distributions are known, *optimal Bayesian classification* trains a classifier from data assuming the feature-label distribution is contained in a family parameterized by $\theta \in \Theta$ with some assumed prior density over the states. Formally, we define an optimal Bayesian classifier, ψ_{OBC}, as any classifier satisfying

$$\mathrm{E}_{\pi^*}\left[\varepsilon(\psi_{\mathrm{OBC}}, \theta)\right] \leq \mathrm{E}_{\pi^*}\left[\varepsilon(\psi, \theta)\right] \tag{9}$$

for all $\psi \in \mathcal{C}$, where $\mathcal{C}$ is an arbitrary family of classifiers. Under the Bayesian framework, this is equivalent to minimizing the probability of error as follows:

$$\begin{aligned} \mathrm{P}\left(\psi_n(\mathbf{X}) \neq Y|S_n\right) &= \mathrm{E}_{\pi^*}\left[P\left(\psi_n(\mathbf{X}) \neq Y|\theta, S_n\right)\right] \\ &= \mathrm{E}_{\pi^*}\left[\varepsilon(\psi_n, \theta)\right] \\ &= \widehat{\varepsilon}\left(\psi_n, S_n\right). \end{aligned} \tag{10}$$

An optimal Bayesian classifier can be found by brute force using the closed form solutions for the expected true error (the Bayesian error estimator), when available. However, if $\mathcal{C}$ is the set of all classifiers (with measurable decision regions), then an optimal Bayesian classifier can be found analogously to Bayes classification for a fixed distribution using the effective class-conditional densities. To wit, we can realize an optimal solution without explicitly finding the error for every classifier because the solution can be found pointwise. Specifically, an optimal Bayesian classifier, ψ_{OBC}, satisfying (9) for all $\psi \in \mathcal{C}$, the set of all classifiers with measurable decision regions, exists and is given pointwise by [29]

$$\psi_{\mathrm{OBC}}(\mathbf{x}) = \begin{cases} 0 \text{ if } \mathrm{E}_{\pi^*}[c] f(\mathbf{x}|0) \geq (1 - \mathrm{E}_{\pi^*}[c]) f(\mathbf{x}|1), \\ 1 \text{ otherwise.} \end{cases} \tag{11}$$

If $\mathrm{E}_{\pi^*}[c] = 0$, then this optimal Bayesian classifier is a constant and always assigns class 1, and if $\mathrm{E}_{\pi^*}[c] = 1$ it always assigns class 0. Hence, we will typically assume that $0 < \mathrm{E}_{\pi^*}[c] < 1$.

Essentially, the optimal thing to do is to find the Bayes classifier using $f(\mathbf{x}|y)$ as the true class-conditional distributions. This is like a plug-in rule, only $f(\mathbf{x}|y)$ is not necessarily in the family of distributions $\{f_{\theta_y}(\mathbf{x}|y)\}$, but some other kind of density that happens to result in the optimal classifier. We find the optimal Bayesian classifier without explicitly evaluating the expected true error, $\mathrm{E}_{\pi^*}[\varepsilon(\psi,\theta)]$, for every possible classifier ψ. With regards to both optimal Bayesian classification and Bayesian MMSE error estimation, $f(\mathbf{x}|y)$ contains all of the necessary information in the model about the class-conditional distributions and we do not have to deal with the uncertainty class or priors directly. Upon defining a model, we find $f(\mathbf{x}|y)$ (which depends on the sample because it depends on π^*) and then the whole problem is solved by treating $f(\mathbf{x}|y)$ as the true distribution: optimal classification, the error estimate of the optimal classifier, and the optimal error estimate for arbitrary classifiers. That being said, there is no short-cut to finding the sample-conditioned MSE via the effective density; indeed, there is no notion of variance in the true error of a fixed classifier under the effective class-conditional densities. Moreover, the approach of using the effective class-conditional densities finds an optimal Bayesian classifier over all possible classifiers. On the other hand, there may be advantages to restricting the space of classifiers, for example, in a Gaussian model one may prefer linear classifiers where closed-form Bayesian error estimators have been found [33].

We will present a Bayesian MMSE classifier for the discrete model, which has already been solved. More generally, what we are proposing is not just a few new classifiers, but a new paradigm in classifier design focused on optimization over a concrete mathematical framework. Furthermore, this work ties Bayesian modeling and the Bayesian error estimator together with the old problem of optimal robust filtering; indeed, in the absence of observations, the optimal Bayesian classifier reduces to the Bayesian robust optimal classifier [32,34].

Optimal discrete classification

To illustrate concepts in optimal Bayesian classification, we consider discrete classification, in which the sample space is discrete with b bins. We let p_i and q_i be the class-conditional probabilities in bin $i \in \{1,\ldots,b\}$ for class 0 and 1, respectively, and we define U_j and V_j to be the number of sample points observed in bin $j \in \{1,\ldots,b\}$ from class 0 and 1, respectively. The class sizes are given by $n_0 = \sum_{i=1}^{b} U_i$ and $n_1 = \sum_{i=1}^{b} V_i$. A general discrete classifier assigns each bin to a class, so $\psi_n : \{1,\ldots,b\} \to \{0,1\}$.

The discrete Bayesian model defines $\theta_0 = [p_1,\ldots,p_{b-1}]$ and $\theta_1 = [q_1,\ldots,q_{b-1}]$. The last bin probabilities are not needed since $p_b = 1 - \sum_{i=1}^{b-1} p_i$ and $q_b = 1 - \sum_{i=1}^{b-1} q_i$. The parameter space of θ_0 is defined to be the set of a valid bin probabilities, e.g., $[p_1,\ldots,p_{b-1}] \in \Theta_0$ if and only if $0 \le p_i \le 1$ for $i \in \{1,\ldots,b-1\}$ and $\sum_{i=1}^{b-1} p_i \le 1$. The parameter space Θ_1 is defined similarly. With the parametric model established, we define conjugate Dirichlet priors

$$\pi(\theta_0) \propto \prod_{i=1}^{b} p_i^{\alpha_i^0 - 1} \text{ and } \pi(\theta_1) \propto \prod_{i=1}^{b} q_i^{\alpha_i^1 - 1}. \tag{12}$$

For proper priors, the hyperparameters, α_i^y for $i \in \{1,\ldots,b\}$ and $y \in \{0,1\}$, must be positive, and for uniform priors $\alpha_i^y = 1$ for all i and y. In this setting, the posteriors are again Dirichlet, and when normalized they are given by

$$\pi^*(\theta_0) = \frac{\Gamma\left(n_0 + \sum_{i=1}^{b} \alpha_i^0\right)}{\prod_{k=1}^{b} \Gamma\left(U_k + \alpha_k^0\right)} \prod_{i=1}^{b} p_i^{U_i + \alpha_i^0 - 1}, \tag{13}$$

$$\pi^*(\theta_1) = \frac{\Gamma\left(n_1 + \sum_{i=1}^{b} \alpha_i^1\right)}{\prod_{k=1}^{b} \Gamma\left(V_k + \alpha_k^1\right)} \prod_{i=1}^{b} q_i^{V_i + \alpha_i^1 - 1}, \tag{14}$$

where Γ is the Gamma function.

In the discrete model, for $j \in \{1,\ldots,b\}$ the effective class-conditional densities can be shown to be equal to

$$f(j|0) = \frac{U_j + \alpha_j^0}{n_0 + \sum_{i=1}^{b} \alpha_i^0} \text{ and } f(j|1) = \frac{V_j + \alpha_j^1}{n_1 + \sum_{i=1}^{b} \alpha_i^1}. \tag{15}$$

$f(j|0)$ and $f(j|1)$ may be viewed as effective bin probabilities for each class after combining prior knowledge and observed data. Hence, from (8), the Bayesian MMSE error estimator for an arbitrary classifier ψ_n is

$$\begin{aligned} \widehat{\varepsilon} = &\sum_{j=1}^{b} \mathrm{E}_{\pi^*}[c] \frac{U_j + \alpha_j^0}{n_0 + \sum_{i=1}^{b} \alpha_i^0} \mathbf{I}_{\psi_n(j)=1} \\ &+ (1 - \mathrm{E}_{\pi^*}[c]) \frac{V_j + \alpha_j^1}{n_1 + \sum_{i=1}^{b} \alpha_i^1} \mathbf{I}_{\psi_n(j)=0}, \end{aligned} \tag{16}$$

where $\mathbf{I}_E$ is an indicator function equal to one if E is true and zero otherwise. Exactly the same expression was derived using a brute-force approach in [27]. The optimal Bayesian classifier may now be found directly using (11):

$$\psi_{\mathrm{OBC}}(j) = \begin{cases} 1 \text{ if } \mathrm{E}_{\pi^*}[c] \frac{U_j + \alpha_j^0}{n_0 + \sum_{i=1}^{b} \alpha_i^0} < (1 - \mathrm{E}_{\pi^*}[c]) \frac{V_j + \alpha_j^1}{n_1 + \sum_{i=1}^{b} \alpha_i^1}, \\ 0 \text{ otherwise.} \end{cases} \tag{17}$$

The optimal Bayesian classifier minimizes the Bayesian error estimator by minimizing each term in the sum (16). This is achieved by assigning $\psi_{\mathrm{OBC}}(j)$ the class with

the smaller constant scaling the indicator function. The expected error of the optimal classifier is

$$\widehat{\varepsilon}_{\mathrm{OBC}} = \sum_{j=1}^{b} \min\left\{ \mathrm{E}_{\pi^*}[\,c] \, \frac{U_j + \alpha_j^0}{n_0 + \sum_{i=1}^{b} \alpha_i^0}, \right. \\ \left. (1 - \mathrm{E}_{\pi^*}[\,c]\,) \frac{V_j + \alpha_j^1}{n_1 + \sum_{i=1}^{b} \alpha_i^1} \right\}. \tag{18}$$

In the special case where we have uniform c and uniform priors for the bin probabilities ($\alpha_i^y = 1$ for all i and y), the Bayesian MMSE error estimate is

$$\widehat{\varepsilon} = \sum_{j=1}^{b} \frac{n_0 + 1}{n + 2} \frac{U_j + 1}{n_0 + b} \mathbf{I}_{\psi_n(j)=1} + \frac{n_1 + 1}{n + 2} \frac{V_j + 1}{n_1 + b} \mathbf{I}_{\psi_n(j)=0}, \tag{19}$$

the optimal Bayesian classifier is

$$\psi_{\mathrm{OBC}}(j) = \begin{cases} 1 \text{ if } \frac{n_0+1}{n_0+b}\left(U_j + 1\right) < \frac{n_1+1}{n_1+b}\left(V_j + 1\right), \\ 0 \text{ otherwise,} \end{cases} \tag{20}$$

and the expected error of the optimal classifier is

$$\widehat{\varepsilon}_{\mathrm{OBC}} = \sum_{j=1}^{b} \min\left\{ \frac{n_0 + 1}{n + 2} \frac{U_j + 1}{n_0 + b}, \frac{n_1 + 1}{n + 2} \frac{V_j + 1}{n_1 + b} \right\}. \tag{21}$$

Hence, under uniform priors, when the total number of samples observed in each class is the same ($n_0 = n_1$), the optimal Bayesian classifier is equivalent to the classical discrete histogram rule, which assigns a class to each bin by a majority vote: $\psi_{\mathrm{DHR}}(j) = 1$ if $U_j < V_j$ and $\psi_{\mathrm{DHR}}(j) = 0$ if $U_j \geq V_j$; otherwise, the discrete histogram rule is not necessarily optimal within an arbitrary Bayesian framework.

We take a moment to compare optimal Bayesian classification over an uncertainty class of distributions with Bayes classification for a fixed feature-label distribution. With fixed class-0 probability c and bin probabilities p_i and q_i, the true error of an arbitrary classifier, ψ, is given by

$$\varepsilon = \sum_{j=1}^{b} c p_j \mathbf{I}_{\psi(j)=1} + (1 - c) q_j \mathbf{I}_{\psi(j)=0}. \tag{22}$$

Note a similarity to (16) and (19). The Bayes classifier is given by $\psi_{\mathrm{Bayes}}(j) = 1$ if $cp_j < (1-c)q_j$ and zero otherwise, corresponding to (17) and (20). Finally, the Bayes error is given by

$$\varepsilon_{\mathrm{Bayes}} = \sum_{j=1}^{b} \min\{cp_j, (1 - c)q_j\}, \tag{23}$$

corresponding to (18) and (21). Throughout, c corresponds to $\mathrm{E}_{\pi^*}[\,c]$, p_j corresponds to the effective bin probability $f(j|0) = (U_j + \alpha_j^0)/(n_0 + \sum_{i=1}^{b} \alpha_i^0)$ and similarly q_j corresponds to the effective bin probability $f(j|1)$. In this case, the effective density is a member of our uncertainty class (which contains all possible discrete feature-label distributions), so that the optimal thing to do is simply plug the effective parameters in the fixed-distribution problem.

That being said, the effective density is not always a member of our uncertainty class. Consider an example with $D = 2$ features, an uncertainty class of Gaussian class-conditional distributions with independent arbitrary covariances, and a proper posterior with fixed class-0 probability $c = 0.5$ (hyperparameters are provided in [32]). We consider three classifiers. First is a plug-in classifier, which is the Bayes classifier corresponding to the posterior expected parameters, $c = 0.5$, $\mu_0 = [\,0, 0, \ldots, 0]$, $\mu_1 = [\,1, 1, \ldots, 1]$, and $\Sigma_0 = \Sigma_1 = I_D$. Since the expected covariances are homoscedastic, this classifier is linear. The second is a state-constrained optimal Bayesian classifier, ψ_{SCOBC}, in which we search for a state with corresponding Bayes classifier having smallest expected error over the uncertainty class [34]. Since the Bayes classifier for any particular state in the uncertainty class is quadratic, this classifier is quadratic. Finally, we have the optimal Bayesian classifier, which has been solved analytically in [29], although details are omitted here. In this case, the effective densities are not Gaussian but multivariate student's t distributions, resulting in an optimal Bayesian classifier having a polynomial decision boundary that is higher than quadratic order. Figure 4 shows $\psi_{\mathrm{plug-in}}$ (red), ψ_{SCOBC} (black) and ψ_{OBC} (green). Level curves for the class-conditional distributions corresponding to the expected parameters used in $\psi_{\mathrm{plug-in}}$ are shown in red dashed lines, and level curves for the distributions in the state corresponding to ψ_{SCOBC} are shown in black dashed lines. These were found by setting the Mahalanobis distance to 1. Each classifier is quite distinct, and in particular, the optimal Bayesian classifier is nonquadratic even though all class-conditional distributions in the uncertainty class are Gaussian.

To demonstrate the performance advantage of optimal Bayesian classification via a simulated experiment, we return to the discrete classification problem. Let c and the bin probabilities be generated randomly according to uniform prior distributions. For each fixed feature-label

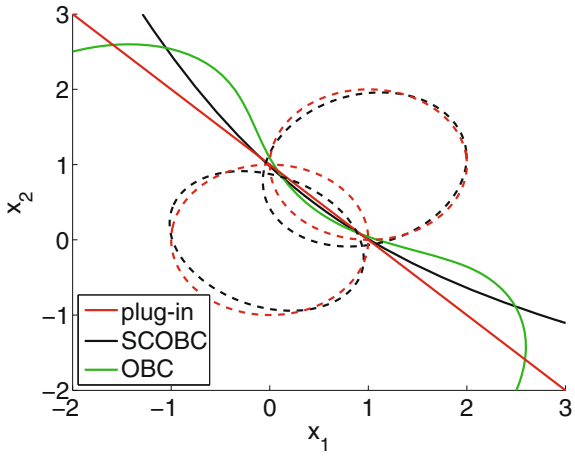

Figure 4 Classifiers for an independent arbitrary covariance Gaussian model. Classifiers for an independent arbitrary covariance Gaussian model with $D = 2$ features and proper posteriors. Whereas the optimal Bayesian classifier (in green) is polynomial with expected true error 0.2007, the state-constrained optimal Bayesian classifier (in black) is quadratic with expected true error 0.2061 and the plug-in classifier (in red) is linear with expected true error 0.2078. These expected true errors are averaged over the posterior on the uncertainty class of states.

distribution, a binomial(n, c) experiment is used to determine the number of sample points in class 0 and the bin for each point is drawn according to the bin probabilities corresponding to its class, thus generating a non-stratified random sample of size n. Both the histogram rule and the new optimal Bayesian classifier from (20), assuming correct priors, are trained from the sample. The true error for each classifier is also calculated exactly via (22) . This is repeated 100,000 times to obtain the average true error for each classification rule, presented in Figure 5 for $b = 2$, 4 and 8 bins. Observe that the average performance of optimal Bayesian classification is indeed superior to that of the discrete histogram rule, especially for larger bin sizes. However, note that optimal Bayesian classifiers are not guaranteed to be optimal for a specific distribution (the optimal classifier is the Bayes classifier), but only optimal when averaged over all distributions in the assumed Bayesian framework.

Conclusions

Scientific knowledge is possible for small-sample classification.

Given the importance of classification throughout science and the crucial epistemological role played by error estimation, it is remarkable that only one paper providing analytic results for moments of common error estimators was published between 1977 and 2005, and that up until 2005, there were no papers providing representation of the joint distribution or of the second-order mixed moments. Today, we are paying the price for this dearth of activity as we are now presented with very large feature sets and small samples across different disciplines, in particular, in high-throughput biology, where the advance of medical science is being hamstrung by a lack of basic knowledge regarding pattern recognition. Moreover, in spite of this obvious crippling lack of knowledge, there is only a minuscule effort to rectify the situation, whereas billions of dollars are wasted on gathering an untold quantity of data that is useless absent the requisite statistical knowledge to make it useful.

No doubt this unfortunate situation would make for a good sociological study. But that is not our field of expertise. Nonetheless, we will put forth a comment made by Thomas Kailath in 1974, about the time that fundamental research in error estimation for small-sample classification came to a halt. He writes, “It was the peculiar atmosphere of the sixties, with its catchwords of ‘building research competence,’ ‘training more scientists,’ etc., that supported the uncritical growth of a literature in which quantity and formal novelty were often prized over significance and attention to scholarship. There was little

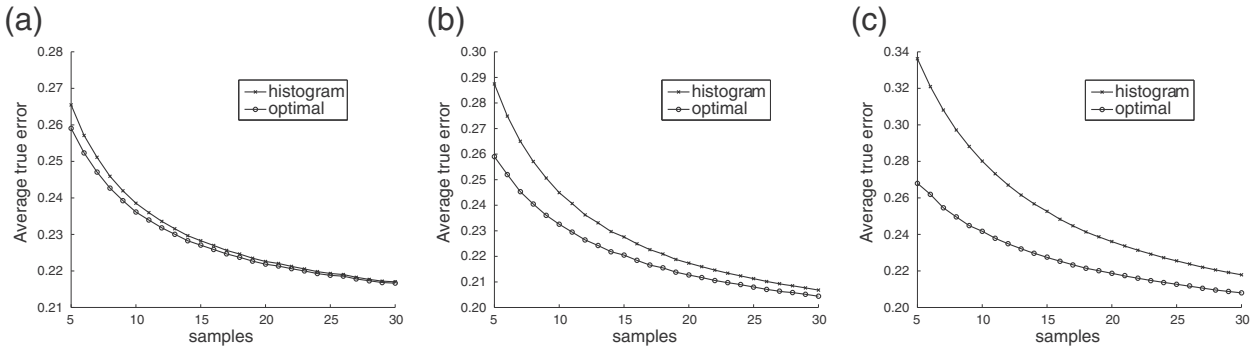

Figure 5 Average true errors for discrete classification. Average true errors on discrete distributions from known priors with uniform c and bin probabilities versus sample size. **(a)** $b = 2$; **(b)** $b = 4$; **(c)** $b = 8$.

concern for fitting new results into the body of old ones; it was important to have 'new' results!" [35]. Although Kailath's observation was aimed at signal processing, the "peculiar atmosphere" of which he speaks is not limited to any particular discipline; rather, he had perceived an "uncritical growth of a literature" lacking "attention to scholarship." One can only wonder what Prof. Kailath's thoughts are today when he surveys a research landscape that produces orders of magnitude more papers but produces less knowledge than that produced by the relative handful of scientists, statisticians, and engineers a half century ago. For those who would question this latter observation in pattern recognition, we suggest a study of the early papers by such pioneers as Theodore Anderson, Albert Bowker, and Rosedith Sitgreaves.

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

Author details
[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA. [2]Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA. [3]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA.

Received: 9 February 2013 Accepted: 6 August 2013
Published: 20 August 2013

References

1. ER Dougherty, U Braga-Neto, Epistemology of computational biology: mathematical models and experimental prediction as the basis of their validity. J. Biol. Syst. **14**, 65–90 (2006)
2. H Reichenbach, *The Rise of Scientific Philosophy* (University of California Press, Berkeley, 1971)
3. ER Dougherty, ML Bittner, *Epistemology of the Cell: A Systems Perspective on Biological Knowledge*. IEEE Press Series on Biomedical Engineering (John Wiley, New York, 2011)
4. R Feynman, *QED: The Strange Theory of Light and Matter* (Princeton University Press, Princeton, 1985)
5. P Frank, *Modern Science and Its Philosophy* (Collier Books, New York, 1961)
6. ER Dougherty, On the epistemological crisis in genomics. Curr. Genomics **9**(2), 69–79 (2008)
7. M Brun, C Sima, J Hua, J Lowey, B Carroll, E Suh, ER Dougherty, Model-based evaluation of clustering validation measures. Pattern Recognit. **40**(3), 807–824 (2007)
8. L Devroye, L Györfi, G Lugosi, *A Probabilistic Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability (Springer, New York, 1996)
9. ER Dougherty, Biomarker development: prudence, risk, and reproducibility. BioEssays **34**(4), 277–279 (2012)
10. B Hanczar, J Hua, ER Dougherty, Decorrelation of the true and estimated classifier errors in high-dimensional settings. EURASIP J. Bioinformatics Syst. Biol. **2007**, 12 (2007). Article ID 38473
11. U Braga-Neto, ER Dougherty, Exact performance of error estimators for discrete classifiers. Pattern Recognit. **38**(11), 1799–1814 (2005)
12. M Hills, Allocation rules and their error rates. J. R. Stat. Soc. Ser. B (Stat. Methodology) **28**, 1–31 (1966)
13. D Foley, Considerations of sample and feature size. IEEE Trans. Inf. Theory **18**(5), 618–626 (1972)
14. MJ Sorum, Estimating the conditional probability of misclassification. Technometrics **13**, 333–343 (1971)
15. GJ McLachlan, An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis. Aust. J. Stat. **15**(3), 210–214 (1973)
16. M Moran, On the expectation of errors of allocation associated with a linear discriminant function. Biometrika **62**, 141–148 (1975)
17. M Goldstein, E Wolf, On the problem of bias in multinomial classification. Biometrics 1977, **33**, 325–331 (1975)
18. A Davison, P Hall, On the bias and variability of bootstrap and cross-validation estimates of error rates in discrimination problems. Biometrica **79**, 274–284 (1992)
19. Q Xu, J Hua, UM Braga-Neto, Z Xiong, E Suh, ER Dougherty, Confidence intervals for the true classification error conditioned on the estimated error. Technol. Cancer Res. Treat. **5**, 579–590 (2006)
20. A Zollanvari, UM Braga-Neto, ER Dougherty, On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers. Pattern Recognit. **42**(11), 2705–2723 (2009)
21. A Zollanvari, UM Braga-Neto, ER Dougherty, On the joint sampling distribution between the actual classification error and the resubstitution and leave-one-out error estimators for linear classifiers. IEEE Trans Inf. Theory **56**(2), 784–804 (2010)
22. A Zollanvari, UM Braga-Neto, ER Dougherty, Exact representation of the second-order moments for resubstitution and leave-one-out error estimation for linear discriminant analysis in the univariate Heteroskedastic Gaussian Model. Pattern Recognit. **45**(2), 908–917 (2012)
23. A Zollanvari, UM Braga-Neto, ER Dougherty, Analytic study of performance of error estimators for linear discriminant analysis. IEEE Trans. Signal Process. **59**(9), 4238–4255 (2011)
24. F Wyman, D Young, D Turner, A comparison of asymptotic error rate expansions for the sample linear discriminant function. Pattern Recognit. **23**, 775–783 (1990)
25. V Pikelis, Comparison of methods of computing the expected classification errors. Automatic Remote Control **5**, 59–63 (1976)
26. ER Dougherty, A Zollanvari, UM Braga-Neto, The illusion of distribution-free small-sample classification in genomics. Curr. Genomics **12**(5), 333–341 (2011)
27. 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**, 115–129 (2011)
28. 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)
29. LA Dalton, ER Dougherty, Optimal classifiers with minimum expected error within a, Bayesian framework–part I: discrete and Gaussian models. Pattern Recognit. **46**(5), 1301–1314 (2013)
30. LA Dalton, ER Dougherty, Application of the Bayesian MMSE estimator for classification error to gene expression microarray data. Bioinformatics **27**(13), 1822–1831 (2011)
31. 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)
32. LA Dalton, ER Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework–part II: properties and performance analysis. Pattern Recognit. **46**(5), 1288–1300 (2013)
33. 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**, 130–144 (2011)
34. ER Dougherty, J Hua, Z Xiong, Y Chen, Optimal robust classifiers. Pattern Recognit. **38**(10), 1520–1532 (2005)
35. T Kailath, A view of three decades of linear filtering theory. IEEE Transact. Inf. Theory **20**(2), 146–181 (1974)

doi:10.1186/1687-4153-2013-10
Cite this article as: Dougherty and Dalton: **Scientific knowledge is possible with small-sample classification.** *EURASIP Journal on Bioinformatics and Systems Biology* 2013 **2013**:10.