



## Corrected small-sample estimation of the Bayes error

Marcel Brun<sup>1</sup>, David L. Sabbagh<sup>1</sup>, Seungchan Kim<sup>2</sup> and Edward R. Dougherty<sup>1, 3,\*</sup>

<sup>1</sup>Department of Electrical Engineering, Texas A&M University, College Station, TX 77840, USA, <sup>2</sup>Cancer Genetics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA and <sup>3</sup>Department of Pathology, University of Texas M. D. Anderson Cancer Center, Houston, Texas, USA

Received on September 6, 2002; revised on November 27, 2002; accepted on December 14, 2002

### ABSTRACT

**Motivation:** A major problem of pattern classification is estimation of the Bayes error when only small samples are available. One way to estimate the Bayes error is to design a classifier based on some classification rule applied to sample data, estimate the error of the designed classifier, and then use this estimate as an estimate of the Bayes error. Relative to the Bayes error, the expected error of the designed classifier is biased high, and this bias can be severe with small samples.

**Results:** This paper provides a correction for the bias by subtracting a term derived from the representation of the estimation error. It does so for Boolean classifiers, these being defined on binary features. Although the general theory applies to any Boolean classifier, a model is introduced to reduce the number of parameters. A key point is that the expected correction is conservative. Properties of the corrected estimate are studied via simulation. The correction applies to binary predictors because they are mathematically identical to Boolean classifiers. In this context the correction is adapted to the coefficient of determination, which has been used to measure non-linear multivariate relations between genes and design genetic regulatory networks. An application using gene-expression data from a microarray experiment is provided on the website <http://gpsnap.tamu.edu/smallsample/> (user:'smallsample', password:'smallsample').

**Contact:** edward@ee.tamu.edu

### INTRODUCTION

Estimation of the Bayes error is a major issue in pattern classification. It is often considered subsidiary to classifier design, in part because one is often primarily interested in constructing a good classifier, and in part because a standard approach to error estimation is to use sample data to estimate the Bayes (optimal) classifier and then

consider the error of the estimated classifier as an estimate of the Bayes (optimal-classifier) error. However, if we are interested in the ability of a feature set to discriminate between classes, the actual measure of interest is the Bayes error, not the error of a designed classifier, nor even the expected error of the designed classifier. And while the approach of classifier design followed by estimation of the error of the designed classifier is generally sound when the sample is large, there are two serious problems for small samples: (1) for a small sample the expected error of the designed classifier can be significantly greater than the Bayes error; and (2) having designed an estimate of the Bayes classifier via some classification rule, the error of that classifier must be estimated using sample data, and this estimation is problematic for small samples. This paper treats the first problem by providing a correction term to subtract from the error of the designed classifier that results in better estimation of the Bayes error. An important aspect of this correction is that it is conservative, meaning that the expected corrected error estimate is lower than the expected uncorrected error estimate but still greater than the Bayes error.

The problem of error estimation relates to three fundamental microarray applications, where small sample issues are ubiquitous (Dougherty, 2002): classification, prediction and network modeling (see website for extensive references).

Prediction and network design are particularly relevant to this paper. If we assume a gene is either ON or OFF, the assumption underlying the Boolean-network model (Kauffman, 1993), then we are in the domain of binary prediction, where the observation vector provides the state of some system of genes and a function on the observation vector provides a predictor of the value of a target gene. This is a view taken in expression-based prediction and in gene regulatory modeling involving probabilistic Boolean networks. If gene-expression data are quantized to 0 and 1, then the prediction is done by

\*To whom correspondence should be addressed.

a binary-valued function defined on binary vectors. In pattern classification and signal processing, the function is called a ‘Boolean classifier’ and a ‘binary filter’, respectively. Since the applications of immediate interest involve genomic signal processing, we will use the latter terminology throughout.

Consider prediction of a binary target random variable  $Y$  based on the observation random variables  $X_1, X_2, \dots, X_N$ . Letting  $\mathbf{X} = (X_1, X_2, \dots, X_N)$ , a random sample of size  $n$  is a set of independent examples,  $(\mathbf{X}^1, Y^1), (\mathbf{X}^2, Y^2), \dots, (\mathbf{X}^n, Y^n)$ , such that  $(\mathbf{X}^k, Y^k)$  is identically distributed to  $(\mathbf{X}, Y)$  for  $k = 1, 2, \dots, n$ . If  $\psi_{opt}$  denotes the optimal filter and  $\psi_n$  denotes an estimate of  $\psi_{opt}$  obtained from the random sample, then the error,  $\epsilon[\psi_n]$ , of the designed filter  $\psi_n$  is the sum of the error of  $\psi_{opt}$  (fixed for the model) and the error owing to estimation:  $\epsilon[\psi_n] = \epsilon[\psi_{opt}] + \Delta(\psi_n, \psi_{opt})$ , where  $\Delta(\psi_n, \psi_{opt})$  is the increase in error owing to using the designed filter instead of the optimal filter. Since the sample is random, we take expectations to obtain the expected error of the designed filter. Letting  $\xi_n = E[\Delta(\psi_n, \psi_{opt})]$ ,

$$E[\epsilon[\psi_n]] = \epsilon[\psi_{opt}] + \xi_n \quad (1)$$

where  $E$  denotes expectation relative to all random samples of size  $n$ . For consistent estimation and sufficiently large samples,  $\xi_n$  is close to 0 and  $E[\epsilon[\psi_n]]$  provides a good estimate of  $\epsilon[\psi_{opt}]$ . For small data sets, this approximation is bad since the estimation error  $\xi_n$  is typically large. Therefore  $E[\epsilon[\psi_n]]$  has strong positive bias relative to estimation of  $\epsilon[\psi_{opt}]$ .

In experimental situations, we have to estimate  $E[\epsilon[\psi_n]]$  by an estimator  $\bar{E}[\epsilon[\psi_n]]$ , and the experimental estimation error is given by the difference  $\bar{\xi}_n = \bar{E}[\epsilon[\psi_n]] - \epsilon[\psi_{opt}]$ . If  $\bar{E}[\epsilon[\psi_n]]$  is unbiased then  $E[\bar{E}[\epsilon[\psi_n]]] = E[\epsilon[\psi_n]]$  and  $E[\bar{\xi}_n] = \xi_n$ .

Equation (1) motivates us to find a correction  $\tau_n$  and to estimate  $\epsilon[\psi_{opt}]$  by

$$\lambda_n = E[\epsilon[\psi_n]] - \tau_n. \quad (2)$$

If  $\tau_n > 0$ , then  $\lambda_n < E[\epsilon[\psi_n]]$  and the positive bias is reduced relative to estimating the error by  $E[\epsilon[\psi_n]]$ . It is important that the correction be conservative, meaning that  $\lambda_n \geq \epsilon[\psi_{opt}]$ . Putting together equations (1) and (2) yields

$$\lambda_n = \epsilon[\psi_{opt}] + \xi_n - \tau_n \quad (3)$$

Our task is to find a correction factor  $\tau_n$  such that  $0 \leq \tau_n \leq \xi_n$ . In experimental situations, if  $E[\epsilon[\psi_n]]$  is estimated by  $\bar{E}[\epsilon[\psi_n]]$ , then  $\lambda_n$  is estimated by  $\bar{\lambda}_n = \bar{E}[\epsilon[\psi_n]] - \tau_n$

## LOWER BOUND ON THE ESTIMATION ERROR

Consider a Boolean function operating on binary  $N$ -vectors,  $\psi : \{0, 1\}^N \rightarrow \{0, 1\}$ . As an estimator of

a binary random variable  $Y$ , the *mean-absolute error* (MAE) of  $\psi$  is defined by  $\epsilon[\psi] = E[|Y - \psi(\mathbf{X})|]$ . An optimal filter minimizes  $\epsilon[\psi]$ . We adopt the following notations:  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$  is the set of all binary  $N$ -vectors (configurations), where  $m = 2^N$ ; for  $i = 1, 2, \dots, m$ ,  $r_i$  is the probability of observing  $\mathbf{x}_i$  and  $p_i = P(Y = 1|\mathbf{x}_i)$  is the conditional probability for  $Y = 1$ , given  $\mathbf{x}_i$  (both with respect to the model).  $\mathbf{p} = (r_1, \dots, r_m, p_1, \dots, p_m)$  is the probability vector determining the model.

For each configuration  $\mathbf{x}_i$ , its contribution to  $\epsilon[\psi_{opt}]$ , the error of the optimal filter, is given by  $r_i \rho_i$ , where  $\rho_i = \min(p_i, 1 - p_i)$ , and

$$\epsilon[\psi_{opt}] = \sum_{i=1}^m r_i \rho_i \quad (4)$$

Equation 1 applies directly, with  $\xi_n$  being a function of  $\mathbf{p}$ , namely,  $\xi_n(\mathbf{p})$ .

Given a sample  $S$  of  $n$  pairs, the filter  $\psi_n$  is designed on the configuration  $\mathbf{x}_i$  by the relative frequency  $\eta_i$ , which is the ratio of the number of times the pair  $(1, \mathbf{x}_i)$  is observed and the number of times  $\mathbf{x}_i$  is observed in  $S$ , respectively. If  $\eta_i > 0.5$ , then  $\psi_n(\mathbf{x}_i) = 1$ ; otherwise,  $\psi_n(\mathbf{x}_i) = 0$ . If  $\mathbf{x}_i$  is not observed, then various rules can be adopted by which to define  $\psi_n(\mathbf{x}_i)$ . For the purposes of our analysis, if  $\mathbf{x}_i$  is not observed, then  $\psi_n(\mathbf{x}_i)$  is defined randomly in  $\{0, 1\}$  with the same probability  $P(Y = 1|\mathbf{x}_i) = P(Y = 0|\mathbf{x}_i) = 0.5$ .

The following theorem provides an explicit representation for the expected error increase  $\xi_n(\mathbf{p})$  for a model  $\mathbf{p}$  and a sample size  $n$ . Its proof is given at the website.

**THEOREM 1.** *The expected error increase  $\xi_n(\mathbf{p})$  is given by*

$$\begin{aligned} \xi_n(\mathbf{p}) = & \sum_{\{i:p_i < 0.5\}} (0.5 - \rho_i) c_1(\rho_i, r_i, n) \\ & + \sum_{\{i:p_i \geq 0.5\}} (0.5 - \rho_i) c_2(\rho_i, r_i, n) \end{aligned} \quad (5)$$

with

$$\begin{aligned} c_1(\rho_i, r_i, n) = & 2r_i \left\{ \frac{(1 - r_i)^n}{2} \right. \\ & \left. + (1 - r_i)^n \sum_{k=1}^n C_n^k \left( \frac{r_i}{1 - r_i} \right)^k \left( 1 - B_{k, \rho_i} \left( \left\lfloor \frac{k}{2} \right\rfloor \right) \right) \right\} \end{aligned} \quad (6)$$

$$\begin{aligned} c_2(\rho_i, r_i, n) = & 2r_i \left\{ \frac{(1 - r_i)^n}{2} \right. \\ & \left. + (1 - r_i)^n \sum_{k=1}^n C_n^k \left( \frac{r_i}{1 - r_i} \right)^k \left( B_{k, 1 - \rho_i} \left( \left\lfloor \frac{k}{2} \right\rfloor \right) \right) \right\}. \end{aligned} \quad (7)$$

The coefficients  $c_1$  and  $c_2$  depend only on the probabilities  $r_i$  and the error contributions  $\rho_i$ . Whether  $p_i$  is less

than or greater than 0.5 determines the use of  $c_1$  or  $c_2$  in the sum. This allows us to find a simple lower bound  $\tau_n(\mathbf{p})$  of  $\xi_n(\mathbf{p})$  that depends only on the probabilities  $r_i$  and the error contributions  $\rho_i$ , and therefore does not depend on whether  $\psi_{opt}(\mathbf{x}_i) = 0$  or  $\psi_{opt}(\mathbf{x}_i) = 1$ . Since we want a lower bound for  $\xi_n(\mathbf{p})$ , we can use the minimum value between  $c_1$  and  $c_2$  for each configuration. For each configuration,  $c_1(\rho_i, r_i, n) \leq c_2(\rho_i, r_i, n)$ . Let

$$c(\rho_i, r_i, n) = \min[c_1(\rho_i, r_i, n), c_2(\rho_i, r_i, n)] = c_1(\rho_i, r_i, n) \quad (8)$$

and define the correction factor for the model  $\mathbf{p}$  by

$$\tau_n(\mathbf{p}) = \sum_{i=1}^m c(\rho_i, r_i, n)(0.5 - \rho_i). \quad (9)$$

Then  $\xi_n(\mathbf{p}) \geq \tau_n(\mathbf{p})$  and  $\lambda_n(\mathbf{p}) \geq \epsilon[\psi_{opt}]$ . Thus the corrected estimate is conservative.

The tightness of the lower bound depends on the differences between  $c_1$  and  $c_2$ . The larger these differences, the looser the bound. To compare these values, we let  $N = 3$  and compute the values of  $c_1(\rho, r, n)$  and  $c_2(\rho, r, n)$  for  $\rho = 0.01, 0.02, \dots, 0.49, 0.50$  and  $n = 10, 20, \dots, 400$ . The results for  $c_1$  and  $c_2$  and the difference  $c_2 - c_1$  are plotted in Figures A1 and A2, provided in the website, respectively. The difference between  $c_1$  and  $c_2$  is small for small  $n$  and small  $\rho$ . In these cases, the bound will be near the real value for  $\xi_n(\mathbf{p})$ .

### MODEL TO COMPUTE THE CORRECTION

To compute  $\tau_n(\mathbf{p})$  in (9) we need to know  $\rho_i$  and  $r_i$  for all the configurations  $\mathbf{x}_i$ . This means there are  $2 \times m$  parameters to be estimated (or to be assumed as prior knowledge). Since we have an interest in small samples, we introduce a model with fewer parameters.

#### Dirichlet model

The model assumes the probabilities  $r_i$  are normalized random variables arising from a gamma distribution with shape parameter  $\kappa$  (varying) and scale parameter  $\beta$  fixed at 1, and the configurations possess equal error contributions,  $\rho_i = \rho$ . A normalization is used to satisfy the probability requirement  $\sum_{i=1}^m r_i = 1$ , and the resulting distribution is a multivariate Dirichlet distribution  $D(\mu_1, \dots, \mu_{m-1}; \mu_m)$  with  $\mu_i = \kappa$ , for  $i = 1, \dots, m$  (Wilks, 1962). The model parameters are the error contribution  $\rho$  and the shape parameter  $\kappa$ . The model, denoted by  $(\rho, \kappa)$ , is summarized by  $\rho_i = \rho$ ,  $i = 1, \dots, m$ , and

$$r_i = \frac{r'_i}{\sum_{i=1}^m r'_i}; \text{ with } r'_i \sim \text{gamma}(\kappa). \quad (10)$$

The model  $(\rho, \kappa)$  corresponds to a class of distributions  $\mathbf{p} = (r_1, \dots, r_m, p_1, \dots, p_m)$ . The probabilities  $r_1, r_2, \dots, r_m$  are only specified up to selection via (10), and the conditional probabilities  $p_1, p_2, \dots, p_m$  are specified only up to the requirement that  $p_i = \rho$  or  $p_i = 1 - \rho$  (that is,  $\rho_i = \rho$ ).

For any  $\kappa$ , the probabilities  $r_i$  can take different values. Hence  $\tau_n(\mathbf{p})$  and  $c(\rho, r_i, n)$  are random variables. We define  $\tau_n(\rho, \kappa)$  to be the expected value of  $\tau_n(\mathbf{p})$  relative to the distribution of  $r_1, r_2, \dots, r_m$ . From (9),

$$\tau_n(\rho, \kappa) = \sum_{i=1}^m E[c(\rho, r_i, n)](0.5 - \rho). \quad (11)$$

Since the correction is a function of the model, not the particular distribution, the corrected estimate for the error of the optimal filter (2) becomes

$$\lambda_n(\mathbf{p}) = E[\epsilon[\psi_n]](\mathbf{p}) - \tau_n(\rho, \kappa). \quad (12)$$

We can use Monte Carlo simulation, via (9), to obtain  $\tau_n(\rho, \kappa)$  to a desired degree of precision. Since it requires computation of the correction many times, it is useful to obtain a faster approximate solution. Since the order of the probabilities does not affect the correction, we can re-order  $r_1, r_2, \dots, r_m$  from smallest to largest without changing the correction  $\tau_n(\mathbf{p})$  for a particular distribution. If we denote this re-ordering by  $r_{(1)}, r_{(2)}, \dots, r_{(m)}$ , then (9) can be rewritten with  $r_{(i)}$  in place of  $r_i$ . Taking expectations in the new equation yields (11) with  $E[c(\rho, r_{(i)}, n)]$  instead of  $E[c(\rho, r_i, n)]$ . While  $E[c(\rho, r_{(i)}, n)] \neq c(\rho, E[r_{(i)}], n)$ , experimentation shows they tend to be close. With this in mind, we have the approximation

$$\tau_n(\rho, \kappa) \approx \sum_{i=1}^m c(\rho, E[r_{(i)}], n)(0.5 - \rho). \quad (13)$$

To estimate  $\tau_n(\rho, \kappa)$ , we can obtain empirical means  $\bar{r}_{(1)}, \bar{r}_{(2)}, \dots, \bar{r}_{(m)}$  for  $r_{(1)}, r_{(2)}, \dots, r_{(m)}$  using the parameter  $\kappa$  of the model and Monte Carlo simulation. These yield

$$\tau_n(\rho, \kappa) \approx \sum_{i=1}^m c(\rho, \bar{r}_{(i)}, n)(0.5 - \rho). \quad (14)$$

This estimation is very fast. It needs only some sample points from the gamma distribution.

#### Estimation of the model

Typically the model is not known beforehand. Therefore, to compute the correction  $\tau_n(\rho, \kappa)$ , we need to estimate  $\rho$  and  $\kappa$  from the data, which is a set  $S$  of  $n$  pairs,  $(\mathbf{x}, y)$ .

We can compute  $\kappa$  as a function of the variance  $\sigma^2$  of the Dirichlet distribution,

$$\hat{\kappa} = \frac{1}{N^3} \left( \frac{N-1}{S^2} - N^2 \right) \quad (15)$$

where  $S^2$  is the estimator, from the data, of  $\sigma^2$  (Wilks, 1962).

To estimate  $\rho$ , we use the fact that the model assumes that all conditional probabilities are equal. This implies that  $\rho = \epsilon[\psi_{opt}]$ . Hence, an estimator of the error of the optimal filter estimates  $\rho$ . First, we obtain a cross-validation high-biased estimate  $\tilde{\rho}_n$  of  $\rho$ . Next, we obtain the resubstitution low-biased estimate  $\check{\rho}_n$ . A rough estimator of  $\rho$  is given by averaging the high- and low-biased estimators,

$$\hat{\rho} = \frac{\tilde{\rho}_n + \check{\rho}_n}{2}. \quad (16)$$

Our ultimate goal is to obtain a conservative estimate of  $\rho = \epsilon[\psi_{opt}]$  less biased than the cross-validation estimate. The estimate  $\hat{\rho}$  provides the starting point for obtaining the desired estimate. The estimates  $\hat{\rho}$  and  $\hat{\kappa}$  are used with (9) to compute the correction  $\tau_n(\rho, \kappa)$ .

### Confidence

The correction  $\tau_n(\rho, \kappa)$  depends on its parameters  $\rho$  and  $\kappa$ . Poor estimation of the parameters can cause too high a correction, which would then underestimate the error of the optimal filter. To avoid this problem, we can define our confidence in  $\rho$  via an interval containing  $\rho$ . The correction can be computed conservatively by taking the minimum value of  $\tau_n(\rho', \kappa)$  over all  $\rho'$  in the interval. A practical way to define the interval is to postulate a Gaussian distribution with mean  $\rho$  and standard deviation  $\sigma_\rho$ . A 99% confidence interval relative to this distribution is given by  $[\rho - 2.58\sigma_\rho, \rho + 2.58\sigma_\rho]$ . Relative to  $\sigma_\rho$ , the correction is

$$\tau_n^{\sigma_\rho}(\rho, \kappa) = \min_{\rho' \in [\rho - 2.58\sigma_\rho, \rho + 2.58\sigma_\rho]} \tau_n(\rho', \kappa). \quad (17)$$

The larger the value of  $\sigma_\rho$ , the lower our confidence in the parameter  $\rho$ , and the smaller the correction. The probability model is still determined by  $(\rho, \kappa)$ ; however, the correction is determined by both  $(\rho, \kappa)$  and  $\sigma_\rho$ .

Figure 1 shows the correction  $\tau_n^{\sigma_\rho}(\rho, \kappa)$  computed for fixed  $\rho = 0.2$  and  $\kappa = 1$ . Curves are shown for confidences  $\sigma_\rho = 0.01, 0.03$  and  $0.05$ , each as a function of the sample size (for  $n = 10, 20, 50, 70, 100, 150$ ). For large training sets (large  $n$ ) the correction decreases toward zero, owing to the better precision of the estimator, and it decreases also for larger  $\sigma_\rho$  (or smaller confidence in the parameter  $\rho$ ).

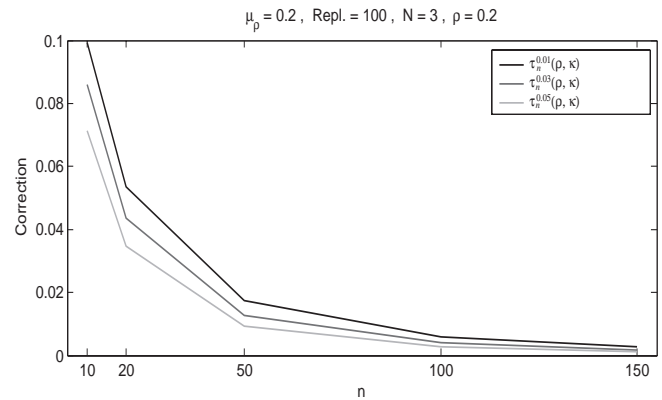


Fig. 1. Corrections using the model parameters.

### SIMULATION STUDIES OF CORRECTION PROPERTIES

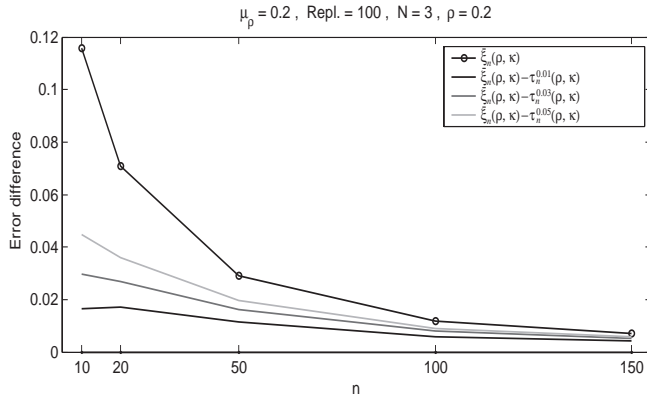
The most basic question regarding the correction is how the corrected and uncorrected error estimations compare. Another concern is robustness relative to using the wrong model parameters, especially with our desire not to underestimate the true error of the optimal filter? A related issue concerns what confidence values  $\sigma_\rho$  provide acceptable robustness and still give useful correction. In practice, we do not know the model. What is the effect of using estimated errors of designed filters rather than their true errors? To address these questions, we employ simulations to analyze the behavior of four models,  $\rho = 0.1, 0.2, 0.3$  and  $0.4$ , and  $\kappa = 1$ , for sample size  $n = 20$  and three variables ( $N = 3$ ).

To compare the correction  $\tau_n(\rho, \kappa)$  with the expected error increase  $\xi_n(\rho, \kappa)$  for a given model  $(\rho, \kappa)$ , we generate 100 joint distributions  $\Pi_1, \Pi_2, \dots, \Pi_{100}$  of  $(Y, \mathbf{X})$ . For  $h = 1, 2, \dots, 100$ , the distribution  $\Pi_h$  is defined by its probabilities  $r_1^h, r_2^h, \dots, r_m^h$  and conditional probabilities  $p_1^h, p_2^h, \dots, p_m^h$ . The probabilities  $r_1^h, r_2^h, \dots, r_m^h$  are generated randomly using (10). The conditional probabilities  $p_1^h, p_2^h, \dots, p_m^h$  are defined by  $p_i^h = 1 - \rho$  for  $i = 1, 2, \dots, \frac{m}{2}$  and  $p_i^h = \rho$  for  $i = \frac{m}{2} + 1, \dots, m$ . This puts half of the configurations in the 1-set of the optimal filter  $\psi_{opt}^h$  and half in the 0-set. Because in practical situations it is very unlikely to have constant error contribution,  $\min(p_i, 1 - p_i) = \rho$ , for all configurations, Gaussian noise, with  $\mu = 0$  and  $\sigma^2 = 0.01$ , is added to the conditional probabilities, thresholding the values to stay in the interval  $[0, 1]$ . For each distribution,  $\psi_{opt}^h$  and its error  $\epsilon[\psi_{opt}^h]$  are found directly from the parameters of the distribution.

From each of the 100 distributions corresponding to the model  $(\rho, \kappa)$ , we generate 100 random

**Table 1.** Corrections computed for different models, for  $n = 20$

$\rho$	$\bar{\xi}_n(\rho, \kappa)$	$\tau_n^{0.05}(\rho, \kappa)$	$\tau_n^{0.03}(\rho, \kappa)$	$\tau_n^{0.01}(\rho, \kappa)$
0.1	0.056	0.035	0.038	0.043
0.2	0.071	0.035	0.044	0.054
0.3	0.076	0.034	0.049	0.062
0.4	0.056	0.000	0.020	0.047



**Fig. 2.** MSE differences.

samples  $S_1^h, S_2^h, \dots, S_{100}^h$  of size  $n$  and design estimators  $\psi_{n,1}^h, \psi_{n,2}^h, \dots, \psi_{n,100}^h$  of the optimal filter  $\psi_{opt}^h$ . Since the model distribution is known, the errors of  $\psi_{n,1}^h, \psi_{n,2}^h, \dots, \psi_{n,100}^h$  can be directly computed and averaged to yield an estimate  $\bar{E}[\epsilon[\psi_n^h]]$  of  $E[\epsilon[\psi_n^h]]$ . This in turn provides an estimate of estimation error for distribution  $\Pi_h$  by

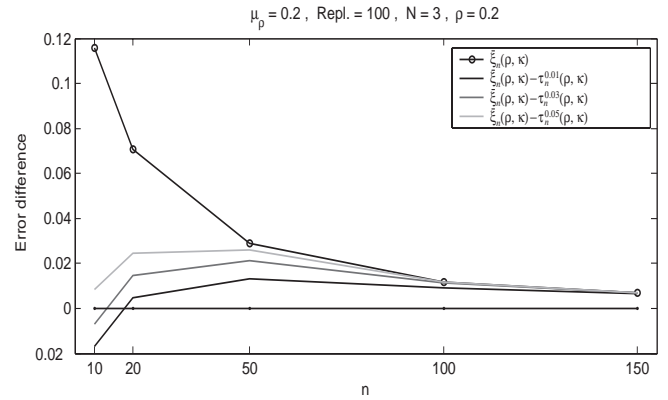
$$\bar{\xi}_n^h(\rho, \kappa) = \bar{E}[\epsilon[\psi_n^h]] - \epsilon[\psi_{opt}^h]. \quad (18)$$

Averaging this estimate over the 100 distributions provides an estimate of the expected design error,  $\bar{\xi}_n(\rho, \kappa)$ , for the model.

Table 1 shows a comparison between the estimation error and the correction, averaged over the 100 distributions, for models  $\kappa = 1$  and  $\rho = 0.1, 0.2, 0.3$  and  $0.4$ , and for  $n = 20$  samples. The correction has been computed using the model parameters  $\rho$  and  $\kappa$ .

Figure 2 shows  $\bar{\xi}_n(\rho, \kappa)$  and the corrected estimated design error,  $\bar{\xi}_n(\rho, \kappa) - \tau_n^{\sigma_\rho}(\rho, \kappa)$  for different  $n$  and  $\sigma_\rho$ , for the model ( $\rho = 0.2, \kappa = 1$ ). The difference is nonnegative.

In considering the difference,  $\bar{\xi}_n(\rho, \kappa) - \tau_n^{\sigma_\rho}(\rho, \kappa)$ , the expected error increase  $\bar{\xi}_n(\rho, \kappa)$  is obtained by averaging estimation errors over distributions covered by the model ( $\rho, \kappa$ ); however,  $\tau_n^{\sigma_\rho}(\rho, \kappa)$  is calculated solely upon the values of  $\rho$  and  $\kappa$ . There is no anomaly here



**Fig. 3.** MSE differences with wrong  $\kappa$ .

because our intention is to utilize the same correction for all distributions covered by the model. The question arises as to the comparison between the model-based correction and the distributional corrections according to (9). Figure A3 on the website shows, as a function of  $\rho$  and  $\kappa$ : the expected design error,  $\bar{\xi}_n(\rho, \kappa)$ , for the model; the expected correction,  $\bar{\tau}_n^{\sigma_\rho}(\rho, \kappa)$ , for the model, based on the distributional corrections of (9), not the model-based correction; and the model-based correction,  $\tau_n^{\sigma_\rho}(\rho, \kappa)$ , for  $n = 20$  and  $\sigma_\rho = 0.01$ . The figure shows  $\tau_n^{\sigma_\rho}(\rho, \kappa)$  and  $\bar{\tau}_n^{\sigma_\rho}(\rho, \kappa)$  to be very close, and  $\bar{\tau}_n^{\sigma_\rho}(\rho, \kappa) \leq \bar{\xi}_n(\rho, \kappa)$  throughout. Moreover, for the most part, the model correction is more conservative than the average of the distributional corrections for the model. The exception is when  $\rho$  is very small and  $\kappa$  is simultaneously large. Even in these extreme cases, the model-based correction only slightly exceeds the expected design error, and this is for the high confidence setting  $\sigma_\rho = 0.01$ .

To illustrate the importance of  $\kappa$ , Figure 3 shows the result of the same simulations as in Figure 2 ( $\rho = 0.2, \kappa = 1$ ); however, applying the correction  $\tau_n^{\sigma_\rho}(\rho, 0)$  instead of  $\tau_n^{\sigma_\rho}(\rho, 1)$ . The figure shows that the difference  $\bar{\xi}_n(\rho, 1) - \tau_n^{\sigma_\rho}(\rho, 0)$  can be negative in some cases ( $n = 10$  and  $\sigma_\rho = 0.001$ ). In such a case,  $\epsilon[\psi_{opt}]$  can be underestimated by  $\bar{E}[\epsilon[\psi_n]] - \tau_n^{\sigma_\rho}(\rho, \kappa)$ . Selecting the wrong model can result in a nonconservative estimate of  $\epsilon[\psi_{opt}]$ , which is not desirable.

We have seen that  $\bar{\xi}_n(0.2, 1) - \tau_n^{\sigma_\rho}(0.2, 1) \geq 0$ , which means the corrected error estimate is conservative; however, this is judged relative to expectation for the model, which is the average of the sample averages  $\bar{E}[\epsilon[\psi_n^h]] - \tau_n^{\sigma_\rho}(0.2, 1)$  over the 100 selected distributions. But what of the individual corrections? For the 100 distributions, Figure A4 on the website shows the scatter plot between the error  $\epsilon[\psi_{opt}]$  and the estimates  $\bar{E}[\epsilon[\psi_n^h]]$  compared with the corrected estimates,  $\bar{E}[\epsilon[\psi_n^h]] -$

**Table 2.** Coefficients of determination computed for different models, for  $n = 20$ 

$\rho$	$\bar{E}[\theta(\mathbf{p}) - \bar{E}[\theta_n](\mathbf{p})]$	$\bar{E}[\theta(\mathbf{p}) - \omega_n(\mathbf{p})]$ $\sigma_\rho = 0.05$	$\bar{E}[\theta(\mathbf{p}) - \omega_n(\mathbf{p})]$ $\sigma_\rho = 0.03$	$\bar{E}[\theta(\mathbf{p}) - \omega_n(\mathbf{p})]$ $\sigma_\rho = 0.01$
0.1	0.135	0.045	0.037	0.023
0.2	0.145	0.063	0.041	0.017
0.3	0.129	0.057	0.024	-0.005
0.4	0.074	0.074	0.035	0.021

$\tau_n^{0.05}(0.2, 1)$ . In all cases, the corrected estimate exceeds  $\epsilon[\psi_{opt}] = 0.2$ .

### COEFFICIENT OF DETERMINATION

The coefficient of determination (COD) measures the degree to which a target random variable can be better predicted based on a set of observation random variables than in the absence of any observations. Whereas the correlation coefficient measures a linear relation between the target and a single variable, the COD can be used for any functional relation between the target and a set of variables. Estimation of the COD depends on estimation of the optimal-filter error. Obtaining a corrected estimate of this error will lead to better estimation of the coefficient. The COD is used to measure multivariate gene interaction (Kim *et al.*, 2000) and to construct gene regulatory networks (Shmulevich *et al.*, 2002).

If the variable  $Y$  is to be predicted from the vector  $\mathbf{X} = (X_1, X_2, \dots, X_N)$ , then the coefficient of determination  $\theta$  is defined by

$$\theta = \frac{\epsilon[\psi_0] - \epsilon[\psi_{opt}]}{\epsilon[\psi_0]} \quad (19)$$

where  $\psi_0$  is the best constant predictor of  $Y$  in the absence of observations. For a sample of size  $n$ , we obtain an estimator of  $\theta$  by

$$\theta_n = \frac{\epsilon[\psi_{n,0}] - \epsilon[\psi_n]}{\epsilon[\psi_{n,0}]} \quad (20)$$

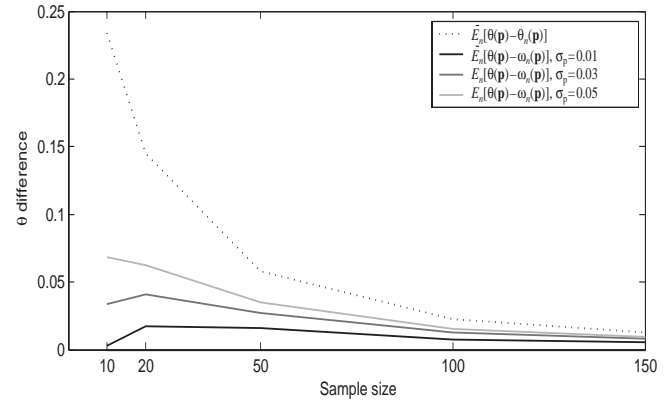
where  $\psi_{n,0}$  is an estimate of the optimal constant filter  $\psi_0$  based on the sample. Since  $\epsilon[\psi_0]$  can be precisely estimated from relatively small samples, a good approximation of  $\theta_n$  is obtained with  $\epsilon[\psi_{n,0}] \approx \epsilon[\psi_0]$ . Taking the expectation over all the samples of size  $n$  yields

$$E[\theta_n] \approx \frac{\epsilon[\psi_0] - E[\epsilon[\psi_n]]}{\epsilon[\psi_0]}. \quad (21)$$

Rewriting the right-hand side gives

$$E[\theta_n] \approx \theta - \frac{\xi_n(\mathbf{p})}{\epsilon[\psi_0]}. \quad (22)$$

Hence,  $E[\theta_n]$  provides a low biased estimate of  $\theta$ .


**Fig. 4.** Expected differences between true and estimated  $\theta$ .

The bias in  $E[\theta_n]$  is caused by the estimation error  $\xi_n(\mathbf{p})$ . A *corrected estimate*,  $\omega_n(\mathbf{p})$ , is defined by replacing  $E[\epsilon[\psi_n]]$  in (21) by its corrected value, to yield

$$\omega_n(\mathbf{p}) = \frac{\epsilon[\psi_0] - \lambda_n(\mathbf{p})}{\epsilon[\psi_0]} \quad (23)$$

According to the definition of  $\lambda_n(\mathbf{p})$ ,

$$\omega_n(\mathbf{p}) = \theta(\mathbf{p}) - \frac{\xi_n(\mathbf{p}) - \tau_n(\mathbf{p})}{\epsilon[\psi_0]} \quad (24)$$

$\omega_n(\mathbf{p})$  is low biased, but less low-biased than  $E[\theta_n]$ . When the distributional model  $(\rho, \kappa)$  is employed, the definition of  $\lambda_n(\mathbf{p})$  takes the model into account according (12). Note that  $\epsilon[\psi_0]$  is a function of  $\mathbf{p}$  in the preceding equations.

We can use simulation to investigate properties of the COD correction using the same approach as for the properties of the error correction. A key measure is the expected value of the difference  $\theta(\mathbf{p}) - \omega_n(\mathbf{p})$  across the model  $(\rho, \kappa)$ , namely,

$$E_{(\rho, \kappa)}[\theta(\mathbf{p}) - \omega_n(\mathbf{p})] = E_{(\rho, \kappa)}\left[\frac{\xi_n(\mathbf{p}) - \tau_n(\mathbf{p})}{\epsilon[\psi_0](\mathbf{p})}\right]. \quad (25)$$

Table 2 gives values of the expectation for various values of  $\rho$ ,  $\sigma_\rho = 0.01, 0.03, 0.05$ ,  $\kappa = 1$ , and  $n = 20$ . For

**Table 3.** Corrections computed for different models, for  $n = 20$

$\rho$	$\tilde{E}[\epsilon[\psi_n]] - \epsilon[\psi_{opt}]$	$\tau_n^{0.05}(\rho, \kappa)$	$\tau_n^{0.03}(\rho, \kappa)$	$\tau_n^{0.01}(\rho, \kappa)$
0.1	0.056	0.035	0.038	0.043
0.2	0.071	0.035	0.044	0.054
0.3	0.076	0.034	0.049	0.062
0.4	0.054	0.000	0.020	0.047

**Table 4.** Coefficients of determination computed for different models, for  $n = 20$

$\rho$	$\tilde{E}[\theta(\mathbf{p})] - \tilde{E}[\theta_n(\mathbf{p})]$	$E[\theta(\mathbf{p}) - \tilde{\omega}_n(\mathbf{p})]$ $\sigma_\rho = 0.05$	$E[\theta(\mathbf{p}) - \tilde{\omega}_n(\mathbf{p})]$ $\sigma_\rho = 0.03$	$E[\theta(\mathbf{p}) - \tilde{\omega}_n(\mathbf{p})]$ $\sigma_\rho = 0.01$
0.1	0.152	0.056	0.047	0.032
0.2	0.168	0.080	0.057	0.031
0.3	0.159	0.082	0.046	0.015
0.4	0.105	0.105	0.061	0.001

comparison, the table also shows the expected difference between  $E[\theta_n(\mathbf{p})]$  and  $\theta(\mathbf{p})$ . Note that for  $\rho = 0.3$  and  $\rho = 0.4$ ,  $\sigma_\rho = 0.01$  yields optimistic estimates. This means that one should be prudent when using such a strong confidence for high  $\rho$  values. For fixed  $\rho = 0.2$ , Figure 4 shows the expectations as a function of  $n$ . A scatter plot corresponding to the error scatter plot of Figure A4 is shown in Figure A5 on the website for  $\rho = 0.2$  and  $n = 20$ .

It is interesting to note that there is strong rank correlation between the true values  $\theta(\mathbf{p})$  and both the uncorrected and corrected estimates for all values of  $\rho \leq 0.3$  (for  $\rho \geq 0.3$ , the COD is very low). The rank correlations tend to be slightly better for the corrected estimates. For instance, for  $\rho = 0.2$ ,  $\kappa = 1$ , and  $n = 20$ , the rank correlations for  $\tilde{E}[\theta_n(\mathbf{p})]$ ,  $\omega_n^{0.001}(\mathbf{p})$ ,  $\omega_n^{0.01}(\mathbf{p})$ ,  $\omega_n^{0.03}(\mathbf{p})$ , and  $\omega_n^{0.05}(\mathbf{p})$  are 0.946, 0.949, 0.951, 0.954 and 0.953, respectively.

### ERROR ESTIMATION FROM SAMPLE DATA

Thus far, we have considered correction of  $E[\epsilon[\psi_n]]$  as an estimate of  $\epsilon[\psi_{opt}]$ , along with the corresponding COD correction. In practice, the model is unknown and we now estimate  $E[\epsilon[\psi_n]]$ . Here we consider cross-validation estimation by randomly splitting the sample into training and test data to obtain error estimates, which are averaged to obtain a close-to-unbiased estimate,  $\tilde{E}[\epsilon[\psi_n]]$ , of  $E[\epsilon[\psi_n]]$ . Although  $\tilde{E}[\epsilon[\psi_n]]$  is essentially unbiased, it possesses a large variance (Devroye *et al.*, 1996), so that individual estimates can vary widely. This yields the cross-validation estimate of the corrected error, where  $\hat{\rho}$  and  $\hat{\kappa}$

are estimated in accordance with (15) and (16):

$$\tilde{\lambda}_n(\mathbf{p}) = \tilde{E}[\epsilon[\psi_n]] - \tau_n(\hat{\rho}, \hat{\kappa}). \quad (26)$$

We obtain an estimate  $\tilde{\omega}_n(\mathbf{p})$  of  $\omega_n(\mathbf{p})$  by replacing  $\lambda_n(\mathbf{p})$  by  $\tilde{\lambda}_n(\mathbf{p})$  in (23).

In analogy to Table 1, Table 3 shows a comparison between the estimation error and the correction, averaged over the 100 distributions, for models  $\kappa = 1$  and  $\rho = 0.1, 0.2, 0.3$  and  $0.4$ , and for  $n = 20$  samples, except here cross-validation is used for estimations. The correction has been computed using the model parameters  $\rho$  and  $\kappa$ . A scatter plot analogous to Figure A4 is shown in Figure A6 on the website with cross-validation estimation having been used. The key point is that, for each of the 100 distributions, the average distributional correction is conservative. Finally, Table 4 corresponds to Table 2 for the COD, here cross-validation being used. An application using  $\hat{\rho}$  and  $\hat{\kappa}$  is given on the website.

### CONCLUSION

A major problem of binary filters and pattern classification is estimation of the error of a designed filter when only small samples are available. This paper provides a correction for the high bias of the expected error of the designed filter as an estimate of the error of the optimal filter. Although the general theory applies to any binary problem, a model-based approach is introduced to greatly reduce the required number of parameters. A key point is that the expected correction is conservative. The correction has been adapted to the COD. An application to genotoxic stress analysis is provided on the website.

---

**ACKNOWLEDGEMENT**

This research was funded in part by the National Human Genome Research Institute.

**REFERENCES**

- Devroye, L., Györfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*. Springer, New York.
- Dougherty, E.R. (2002) Small sample issues for microarray-based classification. *Comp. Func. Genomics*, **2**, 28–34.
- Kauffman, S. (1993) *The Origins of Order: Self-organization and Selection in Evolution*. Oxford University Press, New York.
- Kim, S., Dougherty, E.R., Bittner, M., Chen, Y., Sivakumar, K., Meltzer, P. and Trent, J.M. (2000) A general nonlinear framework for the analysis of gene interaction via multivariate expression arrays. *Biomedical Optics*, **5**, 411–424.
- Shmulevich, I., Dougherty, E.R., Kim, S. and Zhang, W. (2002) Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory network. *Bioinformatics*, **18**, 261–274.
- Wilks, S.S. (1962) *Mathematical Statistics*. John Wiley, New York.

---

† A more complete bibliography is on the website