Supplementary Material for “Application of the Bayesian MMSE Estimator for Classification Error to Gene-Expression Microarray Data”

Lori A. Dalton\textsuperscript{1} and Edward R. Dougherty\textsuperscript{1,2,3}

\textsuperscript{1}Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA
\textsuperscript{2}Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004 USA
\textsuperscript{3}Department of Bioinformatics and Computational Biology, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030 USA

\section{I. INTRODUCTION}

This supplementary material contains a review of classical error estimation schemes, and provides several performance results that were left out of the paper. These include results for both our synthetic high-dimensional model and real empirical breast cancer data, with and without feature selection.

\section{II. REVIEW OF CLASSICAL ERROR ESTIMATORS}

The introduction of our paper discusses the leave-one-out and cross-validation error estimators. In our implementation of cross-validation, we use $k = 5$ folds and 5 repetitions, each with different partitions.

The basic bootstrap zero estimator, $\hat{\varepsilon}_{\text{boot}}$, \cite{3, 4} generates $B$ bootstrap samples, each consisting of $n$ equally-likely draws with replacement from the original sample of size $n$. Each bootstrap sample is then used to design a surrogate classifier, and the points left out of the bootstrap sample are used as holdout to estimate the error of the surrogate classifier. The bootstrap zero estimator is the average of these errors. Like cross-validation, this error estimator is randomized because of the randomly selected bootstrap samples and tends to be pessimistic because the expected bootstrap sample size is only $0.632 n$. In our simulations, we use the popular .632 bootstrap error estimator with $B = 100$, which attempts to correct the pessimistic bias of the bootstrap zero estimator with optimistically biased resubstitution. In particular,

$$\hat{\varepsilon}_{\text{boot}} = (1 - 0.632) \hat{\varepsilon}_{\text{resub}} + 0.632 \hat{\varepsilon}_{\text{b0}}.$$ 

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

$$\hat{\varepsilon}_{\text{bol}} = \frac{1}{n} \sum_{i=1}^{n} \left( I_{y_i=0} \int_{\phi_{x} (x) = 1} f_i(x - x_i) dx + I_{y_i=1} \int_{\phi_{x} (x) = 0} f_i(x - x_i) dx \right).$$

We use spherical Gaussian kernels with the same variance used for all points in a class. The kernel variances are determined by the method proposed in the original paper \cite{1}.

\section{III. SYNTHETIC DATA WITH FEATURE SELECTION}

In this section, we provide a simulation study on performance versus sample size using nearly the same model from Section 3.1 in the paper. The model parameters for the simulation are the same as the equal variance model chosen in \cite{6}, namely $\sigma_0 = \sigma_1 = 0.6$, $m_0 = 0.23$ and $m_1 = 0.8$. These parameters were selected for an LDA classifier designed from 120 sample points and 5 features selected with a $t$-test to give an average true error of 0.25. A summary of our synthetic model parameters is shown in Table I.

As in the paper, we first analyze the quality of features selected by the three-stage feature selection algorithm. In Figure 1(a) we graph the percentage of selected features that are global features. Recall that only 20 of the 20000 features in our model are global features. We see the same themes that were observed in Figure 3 of the paper, that is, as long as the feature size is reasonable for the sample size the percentage of selected features that are global features is large as long as the feature size is reasonable for the sample size. For instance, with 60 sample points we see in Figure 1(a) that the percentage of global features selected drops from 90% with one feature to 75% for 10 features.

Figure 1(b) shows the percentage of selected feature sets that are not rejected on either class at a 95% significance level by the same multivariate Shapiro-Wilk test used in the paper \cite{5}. This graph corroborates the results from Figure 4 in the paper, namely that our three-stage feature selection algorithm with a univariate Gaussianity test still tends to produce features passing a multivariate Gaussianity test. The dip in Figure 1(b) is likely due to two opposing phenomena: as we increase the number of sample points the multivariate Gaussianity test tends to reject feature sets more often, but
at the same time the percentage of global features shown in Figure 1(a) increases so that the probability that the selected features are truly jointly Gaussian increases. For our performance study versus sample size, a summary of all synthetic data simulation settings is shown in Table II. Figure 2 shows RMS deviation from true error for all error estimators implemented on the model with LDA on 1, 5 and 10 selected features with respect to sample size, as well as the expected true error for all of these simulations in Figure 2(d). Lines marked with ‘o’ represent the Bayesian error estimator with flat priors, and lines marked with ‘x’ represent the Bayesian error estimator with the calibrated priors.

For small feature sets, the calibrated Bayesian error estimator shows very good results. For larger feature sets, the Bayesian error estimator with flat priors can have good performance with a small sample size, for example see Figure 2(c). One reason for this flat-prior performance is that when the sample is small and feature size is large, the posterior is likely to be improper for at least one of the classes. As explained in the paper, our Bayesian error estimation code outputs the error 0.5 for that class, which is close to the true error for small sample sizes, as seen in Figure 2(d). Meanwhile, calibrated priors require a larger sample size to beat the flat prior, probably for accurate hyperparameter estimation.

IV. SYNTHETIC DATA WITHOUT FEATURE SELECTION

In this section, we provide several Monte Carlo simulations on synthetic Gaussian distributions without feature selection. These results demonstrate the accuracy of the Bayesian error estimator approximation method described in Section 2.3 of the paper, and also provide results on synthetic data with the QDA and 3NN classification schemes.

The variable parameters of the experiment are the training sample size, $n$, the number of features, $D$, and the Bayes error, used to control the difficulty of the classification problem. Since we are not using feature selection, we compute this Bayes error exactly.

The a priori probabilities for each class are 0.5, and the samples are non-stratified. In each iteration, the sample size in each class is determined by a binomial$(0.5, n)$ experiment and the corresponding number of sample points is randomly generated according to the distributions defined for each class. We use a Gaussian model with fixed means (class 0 is centered at $[0, 0, \ldots, 0]$ and class 1 at $[1, 0, \ldots, 0]$), and the same covariance matrix for both classes. This covariance matrix has the same variance for each feature (with magnitude determined by the desired Bayes error) and no correlation between features.

Once the sample has been generated, it is used to train either an LDA, QDA or 3NN classifier. With the classifier fixed, the true error is deterministic within the fixed model and either computed exactly (for LDA) or approximated using 100,000 testing points (for QDA and 3NN). We use the same error estimators as in the paper, with regular bolstered resubstitution for LDA and QDA (bol), and semi-bolstered resubstitution for 3NN (semi-bol). When LDA is used, the exact Bayesian error estimator is applied with a flat prior. For all classification rules, the approximate Bayesian error estimator is also calculated with a flat prior. This entire process is repeated either 100,000 times (for LDA) or 10,000 times (for QDA and 3NN) to obtain approximate root-mean-square (RMS) deviations from the true error for each error estimator. Note that the calibrated BEE is not applied because there is no calibration data from feature selection.

Figures 4 through 6 show RMS deviation from true error for all error estimators with LDA on 1, 2 and 3 features, respectively. Both the approximate and exact Bayesian error estimators assuming Gaussian distributions and flat priors are applied, where lines marked with ‘o’ represent the exact Bayesian error estimator, and lines marked with ‘x’ represent the approximate result, which is very accurate.

Figures 7 through 9 show performance with QDA classification on 1, 2 and 3 features, respectively, and Figs. 10 through 12 show performance with 3NN on 1, 2 and 3 features, respectively. The Bayesian error estimator generally has very good performance, especially for moderate Bayes errors. It also tends to have excellent performance over a much
wider range of Bayes errors than the other error estimators. This is clearly illustrated in Figure 10(c) with the 3NN classification rule and 1 feature. In this example, the leave-one-out, cross-validation and bootstrap estimators may have very poor performance due to the instability of the surrogate classifiers generated by the nearest neighbor classification rule. The Bayesian error estimator does not suffer from these problems because it is always calculated from the actual designed classifier [2].

V. EMPIRICAL BREAST CANCER DATA WITH FEATURE SELECTION

Simulations in this section use exactly the same model and methodology described in Section 3.2 of the paper.

Figure 3 shows the percentage of selected feature sets that are not rejected by a multivariate Shapiro-Wilk test on either class at a 95% significance level. We use the same multivariate Gaussianity test as before, and the results are slightly less optimistic than seen in Figure 1(b) for the synthetic data model, which is to be expected since the synthetic data model is somewhat idealized with a handful of perfect multivariate Gaussian global features. However, in all simulations the probability of rejecting the multivariate Gaussianity hypothesis does not drop below 75%. Also, as we increase feature size at some point the curves seem to stop dropping, which may be due to the same phenomena described regarding Figure 1(b) in Section III.

Finally, we have provided the complete set of simulations for QDA classification from Section 3.2 of the paper in Figure 13.

VI. EMPIRICAL BREAST CANCER DATA WITHOUT FEATURE SELECTION

The simulations in this section use the same empirical data and methodology as in Section 3.2 of the paper, except we use fixed feature sets before we begin the experiments. Hence, there is no feature selection. The purpose of these simulations is to observe performance when we remove feature selection from the analysis.

All feature sets consist of genes listed in [7] (CENPA, BBC3, CFFM4, TGFβ3 and DKFZP564D0462), and are selected so that a multivariate Shapiro-Wilk test applied to the full data set does not reject Gaussianity over either of the classes at a 95% significance level.

RMS deviation from the holdout estimated true error for several error estimators with LDA, QDA and 3NN are shown in Figs. 14, 15, and 16, respectively. In each figure, 1 to 5 fixed features are used and sample size varies between 20 and 70. Again, the same battery of popular error estimators used before are included, as well as the flat BEE marked with ‘o’, and the calibrated BEE marked with ‘x’. Note that the prior of the calibrated BEE is evaluated using the fixed left out features. Although we do not employ feature selection, the quality of these graphs is similar to those provided in Section 3.2 of the paper.

REFERENCES


TABLE II

CLASSIFICATION SCHEMES AND SETTINGS FOR SIMULATION WITH SYNTHETIC HIGH-DIMENSIONAL DATA

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Size of data</th>
<th>Feature selection</th>
<th>BEE calibration</th>
<th>Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Original</td>
<td>1st t-test</td>
<td>2nd t-test</td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td>30 to 200</td>
<td>5000</td>
<td>20000</td>
<td>1000</td>
</tr>
<tr>
<td>LDA</td>
<td>30 to 200</td>
<td>5000</td>
<td>20000</td>
<td>1000</td>
</tr>
<tr>
<td>LDA</td>
<td>n = 60</td>
<td>5000</td>
<td>20000</td>
<td>1000</td>
</tr>
<tr>
<td>LDA</td>
<td>n = 120</td>
<td>5000</td>
<td>20000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Fig. 3. Percentage of three-stage selected features that are not rejected by a multivariate Shapiro-Wilk test on either class at a 95% significance level with real breast cancer data.
Fig. 4. RMS deviation from true error for the synthetic 1 feature Gaussian model with LDA classification versus Bayes error and sample size.

Fig. 5. RMS deviation from true error for the synthetic 2 feature Gaussian model with LDA classification versus Bayes error and sample size.
Fig. 6. RMS deviation from true error for the synthetic 3 feature Gaussian model with LDA classification versus Bayes error and sample size.

Fig. 7. RMS deviation from true error for the synthetic 1 feature Gaussian model with QDA classification versus Bayes error and sample size.
Fig. 8. RMS deviation from true error for the synthetic 2 feature Gaussian model with QDA classification versus Bayes error and sample size.

Fig. 9. RMS deviation from true error for the synthetic 3 feature Gaussian model with QDA classification versus Bayes error and sample size.
Fig. 10. RMS deviation from true error for the synthetic 1 feature Gaussian model with 3NN classification versus Bayes error and sample size.

Fig. 11. RMS deviation from true error for the synthetic 2 feature Gaussian model with 3NN classification versus Bayes error and sample size.
Fig. 12. RMS deviation from true error for the synthetic 3 feature Gaussian model with 3NN classification versus Bayes error and sample size.

(a) Bayes error = 0.1
(b) Bayes error = 0.2
(c) sample size 30
(d) Bayes error = 0.3
(e) Bayes error = 0.4
(f) sample size 100

Fig. 13. RMS deviation from true error and the expected true error for empirical measurements from a breast cancer study with two stage feature selection and QDA classification versus sample size.

(a) 1 feature
(b) 2 features
(c) 3 features
(d) 4 features
(e) 5 features
(f) expected true error
Fig. 14. RMS deviation from true error for empirical measurements from a breast cancer study with fixed features and LDA classification versus sample size.

Fig. 15. RMS deviation from true error for empirical measurements from a breast cancer study with fixed features and QDA classification versus sample size.
Fig. 16. RMS deviation from true error for empirical measurements from a breast cancer study with fixed features and 3NN classification versus sample size.