Tracking cross-validated estimates of prediction error as studies accumulate *

Lo-Bin Chang
Department of Applied Mathematics and Statistics, Johns Hopkins University

Donald Geman
Department of Applied Mathematics and Statistics, Johns Hopkins University

Abstract. In recent years "reproducibility" has emerged as a key factor in evaluating applications of statistics to the biomedical sciences, for example learning predictors of disease phenotypes from high-throughput "omics" data. In particular, "validation" is undermined when error rates on newly acquired data are sharply higher than those originally reported. More precisely, when data are collected from m "studies" representing possibly different sub-phenotypes, more generally different mixtures of sub-phenotypes, the error rates in cross-study validation (CSV) are observed to be larger than those obtained in ordinary randomized cross-validation (RCV), although the "gap" seems to close as m increases. Whereas these findings are hardly surprising for a heterogeneous underlying population, this discrepancy is then seen as a barrier to translational research. We provide a statistical formulation in the large sample limit: studies themselves are modeled as components of a mixture and all error rates are optimal (Bayes) for a two-class problem. Our results cohere with the trends observed in practice and suggest what is likely to be observed with large samples and consistent density estimators, namely that the CSV error rate exceeds the RCV error rates for any m, the latter (appropriately averaged) increases with m, and both converge to the optimal rate for the whole population.

*The authors gratefully acknowledge the support of the Defense Advanced Research Projects Agency under contract FA8650-11-1-7151, and the National Science Council under grant 100-2115-M-009-007-MY2, partial support of the Center of Mathematical Modeling & Scientific Computing and the National Center for Theoretical Science, Hsinchu, Taiwan.
Key words: Classification, prediction error, cross-validation, Bayes rule, mixture model.

1 Introduction

We provide a statistical analysis of certain empirical observations encountered in attempting to validate applications of statistical learning to prediction. Suppose a classifier is constructed from data in order to distinguish between two classes or hypotheses, call them class 1 and class 2. Suppose further that data used to learn the classifier have been assembled from m sources or "studies," each consisting of samples from each of the two classes. The studies are often associated with "sub-classes," which can exhibit considerable inter-study diversity but still represent the same general class. In fact, each study population may not represent a pure sub-class but rather a *mixture of sub-classes*, and several study populations may be drawn from the same mixture. The standard approach to quantifying the performance of the learning algorithm is to estimate the generalization error using some form of cross-validation. One way is to randomly and repeatedly select some fraction (say ninety percent) of the pooled data for training and test the classifier on the remaining (say ten percent) of the samples, averaging the results. There are other plausible variations, such as "within-study cross validation," in which the error rate for each study is estimated by cross-validation and then the results are averaged or summarized. We shall focus on the pooled one, call it *RCV* for *randomized cross-validation* and denote the estimated error rate by $\hat{e}_{RCV}(m)$. Notice that the identity of the individual studies is lost in testing (although possibly maintained during training). Another possibility, call it *CSV* for *cross-study validation*, is to maintain the identity of the individual studies by leaving each study out in turn, training on the other $m-1$ studies, and testing on the left out study. Let $\hat{e}_{CSV}$ denote the average error rate.

Aggregating data from multiple sources to learn and test prediction rules is common in the analysis high-dimensional biomolecular data, where the motivation is usually to increase consistency by accounting for population heterogeneity [Ma et al (2013), Xu et al (2008), Xu et al (2005), Shen, Ghosh and Chinnaiyan(2004), Teschendorff et al (2006), Michiels, Koscielny and Hill (2005), Yang et al (2008)]. In particular, there has been a substantial effort to develop predictors of disease based on "omics" data generated with high-throughput technologies. A notable example is distinguishing between two cellular phenotypes from mRNA transcript levels ("gene expression") collected from cells in assayed tissue, for instance detecting the presence of disease (e.g., "tumor" vs "normal"), discriminating among cancer sub-types (e.g., "GIST" vs "LMS") and predicting clinical outcomes (e.g., "poor prognosis" vs "good prognosis"). Generally, each individual has a (usually hidden) sub-phenotype label, and a "study population" is a collection of samples which is a particular mixture of sub-phenotypes. Validation methodology has become a core issue in this setting. Whereas many papers have reported high accuracies (usually estimated by cross-validation), finding prediction rules and "signatures" (genes or gene products supporting the rules) that give consistent results across multiple trials remains elusive [Ioannidis et al (2009), Sung et al (2012), Ma, Funk and Price (2010)]. Sometimes a promisingly low rate $\hat{e}_{RCV}(m)$ is reported in one paper but is seen to fail "independent validation" when afterwards either the same classifier is applied to a new study $m+1$, or $\hat{e}_{CSV}(m+1)$ yields a substantially higher error rate. Such observations are then interpreted as calling into question the "reproducibility" of the results.

The "stability" of signatures and feature selection has recently been analyzed in Meinshausen and Buhlmann (2010) and Kirk (2013), and reviewed in He and Yu (2010); as for the role of sample size in reproducibility, see Ein-Dor, Zuk and Domany (2006) for the effect on stability and Braga-Neto and Dougherty (2004) for the effect on *RCV*. For these and other reasons, the implications for clinical practice are widely acknowledged to remain limited; see Altman et al (2011), Marshall (2011), Evan et al (2011) and the discussion in Winslow et al (2012). The discrepancy between the efficacy of biomarkers reported in development and the scarcity of robust omics-based tests delivered to the clinic was the subject of a recent in-depth study by the United States Institute of

Medicine [Micheel, Nass and Omenn (2012)].

The direct origin of this work is a series of experiments conducted in Ma et al (2013) to quantify the impact of "study-effects" on predictive performance by comparing variations on RCV and CSV, referred to as "comparative cross-validation analysis." The authors gathered publicly available gene expression data from 1,470 microarray samples of 6 lung phenotypes from 26 independent experimental studies. For several binary scenarios (e.g., one lung disease vs normal), and using SVM learning, the authors plotted two curves, roughly corresponding to $\hat{e}_{RCV}(m)$ and $\hat{e}_{CSV}(m)$; see Ma et al(2013) for details about sample sizes, etc. One general observation was that the CSV yielded systematically larger error rates; two others were that $\hat{e}_{RCV}(m)$ tended to increase with m and that the "gap" between $\hat{e}_{RCV}(m)$ and $\hat{e}_{CSV}(m)$ seemed to decrease. In view of these observations one conclusion was that by examining how fast CSV "catches up" with RCV as the number of studies is increased, one can estimate when "sufficient" diversity has been achieved for learning a robust molecular predictor likely to translate effectively to new clinical settings.

The observed difference between RCV and CSV makes good sense from a statistical perspective, particularly when m is small and the population is very heterogeneous for at least one of the two classes. As usual, we assume all the samples are independently drawn. Clearly, the key distinction between RCV and CSV is that in RCV the training and testing data follow the same distribution, which is necessary to have unbiased error estimates (for the training sample size). In contrast, in CSV, this condition is often violated in practice, i.e., the data for at least one of the two classes are not identically distributed across studies. This can occur for a variety of reasons, often domain-dependent, but usually associated with either differences in measurement technologies and experimental protocols or with inherent heterogeneity in the underlying population. For example, in the case of gene expression, one source of variation is "technical" and refers to "lab effects" and "batch-effects" [see e.g. Shi et al (2006), Hoen et al (2013)]; expression values derived from the same tissue may vary considerably from platform to platform and/or from day to day. Depending on the enrollment protocol, a study population may also be a mixture over ethnic groups and other factors. The main source of interest here is the inherent diversity within the same general phenotype as discussed above, for example due to environmental or geographic differences. Clearly randomized sampling obscures systematic differences associated with "study-effects" whereas CSV preserves them. Finally, whereas the issue of bias in RCV has been considered from the viewpoint of density estimation Scott and Terrell (1987), model selection Varma and Simon (2006), tuning parameters in supervised learning Tibshirani and Tibshirani (2009) and specifically for microarray data Braga-Neto (2007), our concern is not RCV itself but rather the comparison with CSV.

Our objective is to explain the empirical behavior of the two curves $m \longrightarrow \hat{e}_{RCV}(m)$ and $m \longrightarrow \hat{e}_{CSV}(m)$. In order to provide a theoretical analysis we make several simplifying assumptions. First, we remove the effects of sample size and the choice of the classification method by assuming there is sufficient data to estimate the true distributions and therefore use the Bayes classifier, a weighted likelihood ratio test. In addition, we assume an ideal situation in which the observation vector has already been corrected for all the study-to-study differences described above except for phenotype diversity, i.e., the existence of sub-phenotypes; this then is the source of the difference in the distribution of the data from study to study. Overall population heterogeneity is then modeled as a mixture of possible studies. (Consequently, each element of this mixture is itself a mixture over sub-phenotypes; a finer, multi-scale analysis might make this explicit by identifying studies with the parameters of a multinomial model for the sub-phenotype mixture. However, this is beyond the scope of this paper.) The selection of m studies then corresponds to drawing m iid samples from the mixture variable. Since studies are sampled with replacement, and the same one may appear multiple times. In the next section we will characterize this as the limiting case of a finite sample

scenario using the estimated Bayes classifier for predictions.

In this large sample limit $e_{RCV}(m)$ and $e_{CSV}(m)$ represent optimal error rates in the two corresponding scenarios. The former is the error rate when the data under each class follow the same (mixture) distribution that appears in the likelihood ratio for that class. The latter is the error rate when the data for each class follow one of the distributions in the mixture and this term is left out of each mixture in the likelihood ratio. Both error rates are random variables due to the randomized selection of studies. We prove three results:

1. *With probability one,*
$$e_{CSV}(m) \geq e_{RCV}(m), \;\; m \geq 2.$$
The interpretation is that with CSV the distributions followed by the data may differ from those appearing in the likelihood ratio.

2. $m \to \mathbb{E}[e_{RCV}(m)]$ *is increasing, where the expectation is with respect to choices for the* m *studies. Thus, on average, the larger the mixture the harder the problem, which is not surprising.*

3. *With probability one* $e_{RCV}(m) \to e_{opt}$*, and* $e_{CSV}(m) \to e_{opt}$ *in probability where* e_{opt} *is the Bayes rate for the whole population. Therefore CSV eventually "catches-up".*

Following a more precise formulation in the next section, three theorems are stated in §3, followed by experiments with both simulated and real data in §4 and a Discussion in §5. The proofs are in the Appendix.

2 Mathematical Formulation

We begin with the basic notation. *As explainded above, a study is identified with a mixture of subclasses.* Assume the observed data are continuous with a possibly different d-dimensional density for each class for each study; everything goes through the same way with discrete distributions. We also allow the class likelihoods to depend on the studies. As for the mixture model over studies, the distribution is unrestricted.

$$\begin{aligned}
X &: \text{ a random vector in } \mathbf{R}^d \text{ representing the observation,} \\
Y &: \text{ a binary random variable representing the class,} \\
Z &: \text{ a random variable representing a study,} \\
p_z(k) &: \text{ prior class probabilities given } Z = z \text{ for } k = 1, 2, \\
f_z(x|k) &: \text{ class-conditional densities of } X \text{ given } Z = z, Y = k.
\end{aligned}$$

First consider the classification problem at the population level. Following the notation above, the joint distribution of the observation $X = x$ and the class $Y = k$ is $f(x, k) \doteq \mathbb{E}(p_Z(k) f_Z(x|k))$. By integrating over x and writing $f(x, k) = f(x|k)p(k)$, the probabilities of two classes are

$$p(1) = \mathbb{E}(p_Z(1)), \;\; p(2) = \mathbb{E}(p_Z(2)),$$

and the densities of two classes are

$$f(x|1) = \frac{1}{p(1)} \mathbb{E}(p_Z(1) f_Z(x|1)), \;\; f(x|2) = \frac{1}{p(2)} \mathbb{E}(p_Z(2) f_Z(x|2)).$$

The Bayes classifier compares $p(1)f(X|1)$ with $p(2)f(X|2)$ and the Bayes error rate is

$$e_{opt} = \mathbb{P}\left(\frac{p(1)f(X|1)}{p(2)f(X|2)} > 1 \middle| X \sim f(\cdot|2)\right) \cdot p(2) + \mathbb{P}\left(\frac{p(1)f(X|1)}{p(2)f(X|2)} \leq 1 \middle| X \sim f(\cdot|1)\right) \cdot p(1).$$

In applications, we have a limited number of samples collected by aggregating study populations. Each sample is an observation together with a class label. Our analysis is carried out in the large sample limit of the following finite-sample scenario for generating data. First, m studies are selected from m i.i.d. realizations from Z; call these $z_1, ..., z_m$. There is a study population S_j associated with each $j = 1, ..., m$ which consists of $n_{1,j}$ labeled samples from $f_{z_j}(\cdot|1)$ and $n_{2,j}$ labeled samples from $f_{z_j}(\cdot|2)$; the total sample size of S_j is $n_j = n_{1,j} + n_{2,j}$. Again, several S_j may represent the same study in the sense of a sub-phenotype mixture. However, the sample sizes n_j are not meaningful and depend on various extraneous factors. The study data $S_1, ..., S_m$ are used to learn the classifiers appearing in RCV and CSV, leading to error estimates $\hat{e}_{RCV}(m)$ and $\hat{e}_{CSV}(m)$, which of course also depend on $z_1, ..., z_m$.

In order to minimize the dependence of our results on the specific choice of classifier methodology (and the inherent complications in selecting its parameters) and to remove the effect of sample size, we assume the underlying densities and class probabilities can be perfectly estimated from the data and hence the Bayes classifier is available, i.e., the class with highest posterior mass is selected. This also allows us to investigate the optimal performance possible under each form of cross-validation. In other words, we perform our analysis in the large sample limit where $n_{1,j}, n_{2,j} \to \infty$ for each j, assuming $\frac{n_{k,j}}{n_j} \approx p_{z_j}(k)$ for $k = 1, 2$ and that our estimates of the class-conditional densities $f_{z_j}(x|k), k = 1, 2$, are consistent. Finally, the distributions over aggregated studies are uniform mixtures because we are weighting the studies identically; imbalances among study populations have already been accounted for in sampling $z_1, ..., z_m$. This is consistent with what is done in practice. (If the mixture densities are learned directly from the pooled data, then adjustments must be made for varying the sample sizes.)

For RCV, the error rate is then the Bayes error rate associated with two mixture densities

$$\frac{1}{\sum_{j=1}^m p_{z_j}(1)} \sum_{j=1}^m p_{z_j}(1) f_{z_j}(x|1) \quad \text{and} \quad \frac{1}{\sum_{j=1}^m p_{z_j}(2)} \sum_{j=1}^m p_{z_j}(2) f_{z_j}(x|2),$$

which is

$$\begin{aligned} & e_{RCV}(z_1, ..., z_m) \\ = \; & \frac{\sum_{j=1}^m p_{z_j}(2)}{m} \cdot \mathbb{P}\left(\frac{\sum_{j=1}^m p_{z_j}(1) f_{z_j}(X|1)}{\sum_{j=1}^m p_{z_j}(2) f_{z_j}(X|2)} > 1 \middle| X \sim \frac{1}{\sum_{j=1}^m p_{z_j}(2)} \sum_{j=1}^m p_{z_j}(2) f_{z_j}(\cdot|2)\right) \\ & + \frac{\sum_{j=1}^m p_{z_j}(1)}{m} \cdot \mathbb{P}\left(\frac{\sum_{j=1}^m p_{z_j}(1) f_{z_j}(X|1)}{\sum_{j=1}^m p_{z_j}(2) f_{z_j}(X|2)} \leq 1 \middle| X \sim \frac{1}{\sum_{j=1}^m p_{z_j}(1)} \sum_{j=1}^m p_{z_j}(1) f_{z_j}(\cdot|1)\right). \end{aligned}$$

For CSV, the error rate is the average of m cross study error rates

$$\begin{aligned} e_{CSV}(z_1, ..., z_m) \;=\; & \frac{1}{m} \sum_{j=1}^m \left\{ p_{z_j}(2) \cdot \mathbb{P}\left(\frac{\sum_{i \neq j} p_{z_i}(1) f_{z_i}(X|1)}{\sum_{i \neq j} p_{z_i}(2) f_{z_i}(X|2)} > 1 \middle| X \sim f_{z_j}(\cdot|2)\right) \right. \\ & \left. + p_{z_j}(1) \cdot \mathbb{P}\left(\frac{\sum_{i \neq j} p_{z_i}(1) f_{z_i}(X|1)}{\sum_{i \neq j} p_{z_i}(2) f_{z_i}(X|2)} \leq 1 \middle| X \sim f_{z_j}(\cdot|1)\right) \right\}. \end{aligned}$$

3 Results

In the previous section we considered the joint distribution of X, the observation, Y, the class variable, and Z, the study. The m studies are associated with m i.i.d. realizations of Z, which then determine the RCV and CSV error rates. For simplicity, these rates be denoted by

$$e_{RCV}(m) = e_{RCV}(Z_1, \cdots, Z_m) \text{ and } e_{CSV}(m) = e_{CSV}(Z_1, \cdots, Z_m).$$

These are of course random variables and the probability in Theorem 1 and expectations in Theorem 2 are with respect to $Z_1, \cdots, Z_m$. The only assumption in the theorems below is representing the studies as i.i.d. realizations of Z.

As discussed above, we might expect that the error rate in CSV validation would exceed that in RCV due to the fact that in CSV the distributions under which the error rates are computed are not necessarily the same as the distributions in the likelihood ratio. This of course is the large-sample version of "testing" on a left-out study and "training" on the others combined. The following result states that this is indeed true in the pointwise sense, i.e., with probability one with respect to choosing the studies.

Theorem 1. *For $m \geq 2$,*

$$\mathbb{P}[e_{CSV}(m) \geq e_{RCV}(m)] = 1.$$

Next, we consider the monotonicity of $e_{RCV}(m)$ and $e_{CSV}(m)$. First, we would not expect either curve to be monotone a.s. since adding one more study could make the classification problem either easier or harder. However, what about the functions $m \to \mathbb{E}[e_{RCV}(m)]$ and $m \to \mathbb{E}[e_{CSV}(m)]$? For RCV it would appear that the prediction problem becomes increasingly difficult as the number of studies increases, and this is indeed the case:

Theorem 2. *For $m \geq 2$,*

$$\mathbb{E}[e_{RCV}(m+1)] \geq \mathbb{E}[e_{RCV}(m)].$$

As for $\mathbb{E}[e_{CSV}(m)]$, we might expect this to *decrease* as m increases based on the argument that we are seeing more and more of the population as m increases and hence should be less and less "surprised" when testing on a left out study. This, however, is not necessarily the case. In the following example, $\mathbb{E}[e_{CSV}(m)]$ is not decreasing, nor even is the expected gap decreasing.

Example. Let the hidden variable Z assume only two values $z = 1$ and $z = 2$ with probabilities ϵ and $1 - \epsilon$. Let $f_1(x|2) = f_2(x|2) = \mathbb{1}_{[0,1]}(x)$ and

$$f_1(x|1) = 2 \cdot \mathbb{1}_{[0,\frac{1}{8}]}(x) + \frac{6}{31} \cdot \mathbb{1}_{[-\frac{31}{8},0)}(x), \quad f_2(x|1) = \frac{1}{4}\mathbb{1}_{[-\frac{31}{8},\frac{1}{8}]}(x).$$

Then when ϵ is sufficiently small,

$$\mathbb{E}[e_{CSV}(2) - e_{RCV}(2)] < \mathbb{E}[e_{CSV}(3) - e_{RCV}(3)].$$

(We omit the messy computation.) Since we know from Theorem 2 that $\mathbb{E}[e_{RCV}(m)]$ is nondecreasing, we obtain

$$\mathbb{E}[e_{CSV}(2)] < \mathbb{E}[e_{CSV}(3)].$$

Finally, we consider the asymptotic behavior of the two curves as $m \to \infty$. Recall that $f(x|1) = \frac{1}{p(1)}\mathbb{E}(p_Z(1)f_Z(x|1))$ and $f(x|2) = \frac{1}{p(2)}\mathbb{E}(p_Z(2)f_Z(x|2))$.

Theorem 3. *As $m \to \infty$, $e_{RCV}(m) \to e_{opt}$ almost surely and $e_{CSV}(m) \to e_{opt}$ in probability. Moreover, if the set $\{x : p(1)f(x|1) = p(2)f(x|2) \neq 0\}$ has measure zero, then $e_{CSV}(m) \to e_{opt}$ almost surely.*

4 Numerical Experiments

Based on the mathematical formulation in Section 2, we provide two experiments with specific mixture models in order to illustrate the various properties of the two curves stated in the three theorems. These experiments also reveal some of the subtleties. The models for the first experiment is hand-crafted whereas the one for second experiment is learned from real data, in fact from the data employed in study of lung diseases Ma et al (2013) which motivated this paper.

Experiment 1.

The intent is to model a very simple scenario in which there is a one-dimensional observation with a Gaussian distribution under each class for each study. For instance, X might represent blood pressure or heart rate, the two classes might represent "normal" and "elevated" and Z might corresponds to a geographic location or some other stratification of a large and diverse population.

Let $Z = (Z(1), Z(2))$, where $Z(1), Z(2)$ are i.i.d. $U(0,1)$ random variables. Let $p_z(1) = p_z(2) = 0.5$ for all $z \in [0,1]$. In class 1, X is normally distributed with mean $Z(1)$ and variance σ^2; in class 2, X is also normally distributed with mean $0.5 + Z(2)$ and variance σ^2. Then, for any given $Z_1, \cdots, Z_m$ i.i.d.$\sim Z$, we can compute the error rate $e_{RCV}(m)$ of the randomized cross-validation and the error rate $e_{RCV}(m)$ of the cross-study-validation using Monte Carlo integrations. The difficulty of the problem is determined by σ.

Figure 1 depicts the results for $\sigma = 0.8, 1.0, 1.2$. As expected, larger variances correspond to larger error rates. The left panels show $e_{RCV}(m)$ and $e_{RCV}(m)$ for $m = 2, ..., 7$ for one specific sequence $Z_1, ..., Z_7$. First, we see that $e_{RCV}(m) \leq e_{RCV}(m)$ for each m as stated in Theorem 1. Second, neither curve of the curves $e_{RCV}(m), m = 2, ..., 7$ nor $e_{RCV}(m), m = 2, ..., 7$ is monotone. Apparently, in this particular sample, the two means $Z(1)$ and $0.5 + Z(2)$ are very close in the fifth study for $\sigma = 0.8$ and in the third study for $\sigma = 1.2$.

The right panels show the expected values $\mathbb{E}[e_{RCV}(m)]$ and $\mathbb{E}[e_{CSV}(m)]$ for $m = 2, ..., 7$. These expectations are also calculated also by Monte Carlo integration. Notice that in this example both curves are monotone, although in general only $m \to \mathbb{E}[e_{RCV}(m)]$ is guaranteed to be monotone (Theorem 2). In this experiment, around seven studies seems to be sufficient to capture the diversity in the population in the sense that CSV has "caught-up" with RCV.

Finally, we repeated the experiment for different choices of $p_z(1)$, including $p_z(1) = .1$, but the results were qualitatively quite similar.

Experiment 2.

In the previous example the distributions were designed to have considerable study-to-study variability, which is then reflected in the behavior of the two curves, for example the irregularity and large initial gap. As the next experiment shows, this level of variability can be found in real data as well. In this experiment, the models are learned from gene expression data.

We obtained the GCRMA lung disease microarray data set from Price Lab at institute for systems biology (price.systemsbiology.net), which consists of RNA counts derived from the tissue of patients with either "normal" lungs or diagnosed with one of several lung diseases Ma et al (2013). In order to have two classes, we focused on the patient profiles for Adenocarcinoma (called ADC) and the non-disease phenotype (called NORM). There are six studies (or labs) containing both ADC and NORM data. Each study has includes expression values for order 10^4 genes with the numbers of patients ranging from several tens to several hundreds.

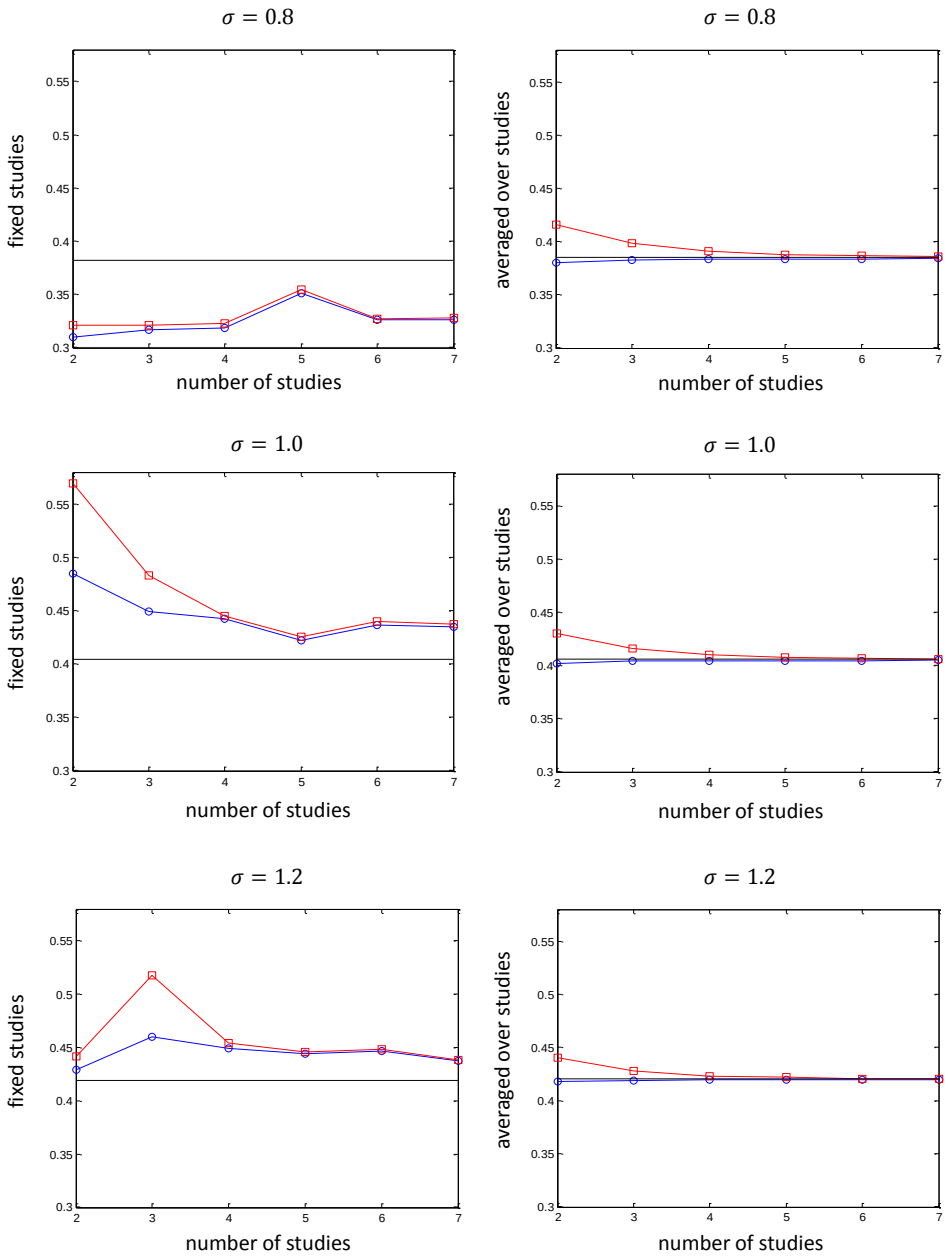

Figure 1: **Designed Model.** $Z = (Z(1), Z(2))$ i.i.d.$\sim U(0,1)$; $f_z(x), g_z(x)$ are respectively $N(z(1), \sigma^2)$ and $N(0.5+z(2), \sigma^2)$. The three rows correspond to $\sigma = 0.8, 1.0, 1.2$. The two columns correspond to to a fixed sequence $Z_1, ..., Z_7$ (left) and the average over $Z_1, ..., Z_7$. The error rates for RCV are in blue and for CSV in red. The black line is the optimal error rate for the whole population.

We selected three differentially-expressed genes for ADC vs. NORM for each study using the Wilcoxon rank-sum test. However, it is well-known that training a classifier on the same data used

for filtering (extracting a reduced feature set) can lead to rampant over-fitting and very biased error estimates, especially when the number of features far exceeds the sample sizes. As a result, we used one-third of data determine the three genes and the remaining two-thirds the data for learning the distributions. The three genes, say g_1, g_2, g_3, are not simply the ones with the smallest p-values. Whereas g_1 has the smallest p-value, g_2 has the smallest p-value among those genes whose absolute correlation with g_1 is smaller than 0.2, and g_3 has the smallest p-value among those whose absolute correlations with both g_1 and g_2 are smaller than 0.2. Then, for each of the six studies, we learned a trivariate normal distribution for $X = (X_1, X_2, X_3)$, the expressions of g_1, g_2, g_3, for the ADC and NORM classes. Finally, Z is assumed to be uniformly distributed on $\{1, 2, 3, 4, 5, 6\}$.

Similar to the first experiment, the left panel of Figure 2 shows the two curves for a particular realization $Z_1, \cdots, Z_7$ generated i.i.d. from $U\{1, 2, 3, 4, 5, 6\}$. (The choice of $m_{max} = 7$ was dictated by computational issues.) For this particular choice of studies, adding the third one results in a relatively large jump from $e_{CSV}(2)$ to $e_{CSV}(3)$ but has less influence on e_{RCV}. The trivariate normal densities for the two classes of the third study are in fact quite different from the mixture of the first two studies, but the mixture density for the three studies is still similar to the mixture of the first two. The right panel shows that, for this model, the expectations of both $e_{RCV}(m)$ and $e_{RCV}(m)$ are monotone. All error rates and expectations are calculated by Monte Carlo integrations.

Clearly cross-study error rates are quite pessimistic relative to RCV, where the (expected) error rate increases slowly and approaches the Bayes error rate $e_{opt} = 0.0667$. Even though the study variable Z assumes only six possible values, it appears that considerably more than $m = 7$ samples is necessary for the expected CSV rate to get near e_{opt}.

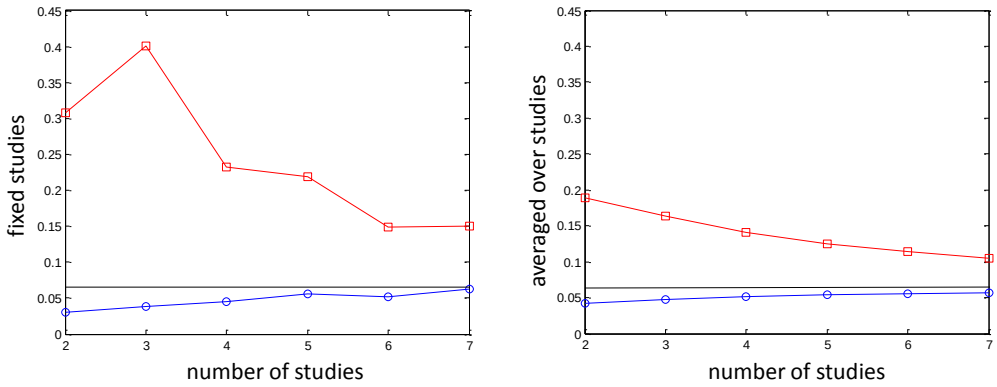

Figure 2: **Models learned from real data.** The population is a mixture of six possible subphenotype mixtures, with $Z \sim U\{1, 2, 3, 4, 5, 6\}$. The observation $X = (X_1, X_2, X_3)$ is the expression of three genes, and for each of the two classes ADC (a lung disease) and NORM, the learned model is trivariate Gaussian. The left panel shows the error rate of the randomized cross-validation in blue and of cross-study-validation (red) for one particular choice of seven studies. The right panel shows the expected error rates, $\mathbb{E}e_{RCV}(m)$ (blue) and $\mathbb{E}e_{CSV}(m)$ (red). The black line is the optimal error rate for the whole population.

5 Discussion

Many questions remain about the behavior of the two curves. We have only considered the large sample limit, but even here several issues are unresolved. We know that $\mathbb{E}(e_{RCV}(m)) \nearrow e_{opt}$ (even a.s.), and that $e_{CSV}(m)$ converges in probability to e_{opt}. Under what conditions is $\mathbb{E}(e_{CSV}(m))$ decreasing in m, and hence $\mathbb{E}(e_{CSV}(m)) \searrow e_{opt}$? More generally, is $\mathbb{E}(e_{CSV}(m)) \geq e_{opt}$?. Can e_{opt} be estimated from the history $\{e_{RCV}(k), e_{CSV}(k), k = 1, ..., m\}$?.

Turning to finite samples, it might be possible to obtain results under some assumption on the distribution of the data and classification methodology, for example assuming Gaussian data and the estimated Bayes rule (i.e., *LDA* etc.), or some variation on logistic regression or large margin classifiers. Such results could be useful in practice. In particular, in many applications, notably the biomedical ones highlighted here, a key issue is how to make rational decisions about how data is collected. (The importance of "original data" in computational biology cannot be over-emphasized, partially due to the effort required to collect them.) What is "enough" data? More specifically, for prediction, when is the data collected from m studies sufficient to capture the diversity of the underlying population and be confident that error estimates will stand up to "independent validation"? Informally, the answer should be "when the two curves converge." How does one quantify this and provide a practical recipe? Finally, as suggested above, with additional regularity it might be possible to construct a interval estimate for e_{opt}, which could be an indirect but effective way to answer the question about sufficient data.

6 Proofs

Theorem 1 utilizes the following lemma. Let X be a random variable generated from either density $f_0(x)$ under class 1 or from density $g_0(x)$ under class 2, with

$$P_1(\cdot) = \mathbb{P}(\cdot|X \sim f_0), \ P_2(\cdot) = \mathbb{P}(\cdot|X \sim g_0).$$

Further, let $\tilde{f}(x)$ and $\tilde{g}(x)$ be any two densities and consider the two mixtures:

$$\bar{f}(x) = (1-\alpha)\tilde{p}(1)\tilde{f}(x) + \alpha p(1) f_0(x), \ \bar{g}(x) = (1-\alpha)\tilde{p}(2)\tilde{g}(x) + \alpha p(2) g_0(x),$$

where $\tilde{p}(1) + \tilde{p}(2) = p(1) + p(2) = 1$ and $0 \leq \tilde{p}(1), \tilde{p}(2), p(1), p(2), \alpha \leq 1$. Notice that the $\bar{f}$ and $\bar{g}$ defined above are nonnegative functions but not density functions.

Lemma: The error rate of the likelihood ratio classifier with $\tilde{f}(x)$ and $\tilde{g}(x)$ is greater than or equal to the error rate of the likelihood ratio classifier using any mixture of these with the true densities. That is

$$\begin{aligned} p(1) \cdot P_1\left(\frac{\tilde{p}(1)\tilde{f}(X)}{\tilde{p}(2)\tilde{g}(X)} \leq 1\right) + p(2) \cdot P_2\left(\frac{\tilde{p}(1)\tilde{f}(X)}{\tilde{p}(2)\tilde{g}(X)} > 1\right) \\ \geq p(1) \cdot P_1\left(\frac{\bar{f}(X)}{\bar{g}(X)} \leq 1\right) + p(2) \cdot P_2\left(\frac{\bar{f}(X)}{\bar{g}(X)} > 1\right). \end{aligned}$$

Proof of the Lemma. Let

$$\begin{aligned} A &= \{x|p(1)f_0(x) \leq p(2)g_0(x) \text{ and } \tilde{p}(1)\tilde{f}(x) > \tilde{p}(2)\tilde{g}(x)\} = A_1 \cup A_2 \\ B &= \{x|p(1)f_0(x) > p(2)g_0(x) \text{ and } \tilde{p}(1)\tilde{f}(x) \leq \tilde{p}(2)\tilde{g}(x)\} = B_1 \cup B_2 \\ C &= \{x|p(1)f_0(x) > p(2)g_0(x) \text{ and } \tilde{p}(1)\tilde{f}(x) > \tilde{p}(2)\tilde{g}(x)\} \subseteq \{x|\bar{f}(x) > \bar{g}(x)\} \\ D &= \{x|p(1)f_0(x) \leq p(2)g_0(x) \text{ and } \tilde{p}(1)\tilde{f}(x) \leq \tilde{p}(2)\tilde{g}(x)\} \subseteq \{x|\bar{f}(x) \leq \bar{g}(x)\}, \end{aligned}$$

where

$$A_1 = \{x | x \in A, \bar{f}(x) > \bar{g}(x)\},\ A_2 = \{x | x \in A, \bar{f}(x) \leq \bar{g}(x)\},$$
$$B_1 = \{x | x \in B, \bar{f}(x) > \bar{g}(x)\},\ B_2 = \{x | x \in B, \bar{f}(x) \leq \bar{g}(x)\}.$$

Now, consider any pair of density functions F, G satisfying

$$\begin{aligned} F(x) &= f_0(x) \quad \text{if } x \in C \cup D \\ G(x) &= g_0(x) \quad \text{if } x \in C \cup D \\ F(x) &= 0 \quad \text{if } x \in A_1 \cup B_1 \\ G(x) &= 0 \quad \text{if } x \in A_2 \cup B_2 \\ F(x) &> 0 \quad \text{if } x \in A_2 \cup B_2 \\ G(x) &> 0 \quad \text{if } x \in A_1 \cup B_1. \end{aligned}$$

The Bayes error rate is

$$\begin{aligned} & p(1) \cdot P_1(A \cup D) + p(2) \cdot P_2(B \cup C) \\ = \ & p(1) \cdot [P_1(A_1)) + P_1(A_2)) + P_1(D))] + p(2) \cdot [P_2(B_1) + P_2(B_2) + P_2(C)]. \end{aligned}$$

Since this is optimal, it must be less than or equal to

$$\begin{aligned} & p(1) \cdot P_1(p(1)F(X) \leq p(2)G(X))) + p(2) \cdot P_2(p(1)F(X) > p(2)G(X))) \\ = \ & p(1) \cdot [P_1(A_1) + P_1(B_1) + P_1(D)] + p(2) \cdot [P_2(A_2) + P_2(B_2) + P_2(C)], \end{aligned}$$

and we have

$$p(1) \cdot P_1(A_2) + p(2) \cdot P_2(B_1) \leq p(1) \cdot P_1(B_1) + p(2) \cdot P_2(A_2),$$

and therefore

$$\begin{aligned} & p(1) \cdot [P_1(A_2) + P_1(B_2) + P_1(D)] + p(2) \cdot [P_2(A_1) + P_2(B_1) + P_2(C)] \\ \leq \ & p(1) \cdot [P_1(B_1) + P_1(B_2) + P_1(D)] + p(2) \cdot [P_2(A_1) + P_2(A_2) + P_2(C)]. \end{aligned}$$

Now notice that

$$A_2 \cup B_2 \cup D = \{x | \bar{f}(x) \leq \bar{g}(x)\},\ A_1 \cup B_1 \cup C = \{x | \bar{f}(x) > \bar{g}(x)\},$$

and

$$B_1 \cup B_2 \cup D = \{x | \tilde{p}(1)\tilde{f}(x) \leq \tilde{p}(2)\tilde{g}(x)\},\ A_1 \cup A_2 \cup C = \{x | \tilde{p}(1)\tilde{f}(x) > \tilde{p}(2)\tilde{g}(x)\}.$$

Hence, we obtain

$$\begin{aligned} & p(1) \cdot P_1(\bar{f}(X) \leq \bar{g}(X)) + p(2) \cdot P_2(\bar{f}(X) > \bar{g}(X)) \\ \leq \ & p(1) \cdot P_1(\tilde{p}(1)\tilde{f}(X) \leq \tilde{p}(2)\tilde{g}(X)) + p(2) \cdot P_2(\tilde{p}(1)\tilde{f}(X) > \tilde{p}(2)\tilde{g}(X)), \end{aligned}$$

and the proof is completed. □

Remark: From the above lemma, we can show that the error rate associated with the ratio of two functions

$$\bar{f}(x) = (1-\alpha)\tilde{p}(1)\tilde{f}(x) + \alpha p(1) f_0(x),\ \bar{g}(x) = (1-\alpha)\tilde{p}(2)\tilde{g}(x) + \alpha p(2) g_0(x)$$

is nonincreasing in α.

For simplicity, we introduce the following notation:

$$R_1 = R_1(z_1, \cdots, z_m) = \left\{ \frac{\sum_{i=1}^m p_{z_i}(1) f_{z_i}(X|1)}{\sum_{i=1}^m p_{z_i}(2) f_{z_i}(X|2)} \leq 1 \right\}$$
$$R_2 = R_2(z_1, \cdots, z_m) = \left\{ \frac{\sum_{i=1}^m p_{z_i}(1) f_{z_i}(X|1)}{\sum_{i=1}^m p_{z_i}(2) f_{z_i}(X|2)} > 1 \right\}$$
$$c(k) = \sum_{j=1}^m p_{z_j}(k) \text{ for } k = 1, 2.$$

Therefore,

$$\begin{aligned} e_{RCV}(z_1, ..., z_m) &= \frac{c(2)}{m} \mathbb{P}\left(R_2 \left| X \sim \frac{1}{c(2)} \sum_{j=1}^m p_{z_j}(2) f_{z_j}(\cdot|2) \right. \right) \\ &+ \frac{c(1)}{m} \mathbb{P}\left(R_1 \left| X \sim \frac{1}{c(1)} \sum_{j=1}^m p_{z_j}(1) f_{z_j}(\cdot|1) \right. \right). \end{aligned}$$

Now, since for $k = 1, 2$,

$$\begin{aligned} & \mathbb{P}\left(R_k \left| X \sim \frac{1}{c(k)} \sum_{j=1}^m p_{z_j}(k) f_{z_j}(\cdot|k) \right. \right) \\ = & \int_{R_k} \frac{1}{c(k)} \sum_{j=1}^m p_{z_j}(k) f_{z_j}(x|k) dx = \frac{1}{c(k)} \sum_{j=1}^m p_{z_j}(k) \int_{R_k} f_{z_j}(x|k) dx \\ = & \frac{1}{c(k)} \sum_{j=1}^m p_{z_j}(k) \mathbb{P}(R_k | X \sim f_{z_j}(\cdot|k)), \end{aligned} \tag{6.1}$$

we obtain

$$e_{RCV}(z_1, ..., z_m) = \frac{1}{m} \sum_{j=1}^m p_{z_j}(2) \mathbb{P}(R_2 | X \sim f_{z_j}(\cdot|2)) + \frac{1}{m} \sum_{j=1}^m p_{z_j}(1) \mathbb{P}(R_1 | X \sim f_{z_j}(\cdot|1)) \tag{6.2}$$

Proof of Theorem 1. Assume that we have m studies $z_1, z_2, \cdots, z_m$. From the above lemma, with $f_{z_j}(\cdot|1)$ and $f_{z_j}(\cdot|2)$ playing the roles of f_0 and g_0, we have for each $j \in \{1, 2, \cdots, m\}$,

$$\begin{aligned} & p_{z_j}(2) \cdot \mathbb{P}\left(\frac{\sum_{i \neq j} p_{z_i}(1) f_{z_i}(X|1)}{\sum_{i \neq j} p_{z_i}(2) f_{z_i}(X|2)} > 1 \left| X \sim f_{z_j}(\cdot|2) \right. \right) \\ & + p_{z_j}(1) \cdot \mathbb{P}\left(\frac{\sum_{i \neq j} p_{z_i}(1) f_{z_i}(X|1)}{\sum_{i \neq j} p_{z_i}(2) f_{z_i}(X|2)} \leq 1 \left| X \sim f_{z_j}(\cdot|1) \right. \right) \\ \geq & p_{z_j}(2) \cdot \mathbb{P}(R_2 | X \sim f_{z_j}(\cdot|2)) + p_{z_j}(1) \cdot \mathbb{P}(R_1 | X \sim f_{z_j}(\cdot|1)). \end{aligned}$$

This implies

$$\begin{aligned}
&\frac{1}{m}\sum_{j=1}^{m} p_{z_j}(2)\cdot\mathbb{P}\left(\frac{\sum_{i\neq j} p_{z_i}(1)f_{z_i}(X|1)}{\sum_{i\neq j} p_{z_i}(2)f_{z_i}(X|2)}>1\middle| X\sim f_{z_j}(\cdot|2)\right)\\
&\qquad+\frac{1}{m}\sum_{j=1}^{m} p_{z_j}(1)\cdot\mathbb{P}\left(\frac{\sum_{i\neq j} p_{z_i}(1)f_{z_i}(X|1)}{\sum_{i\neq j} p_{z_i}(2)f_{z_i}(X|2)}\leq 1\middle| X\sim f_{z_j}(\cdot|1)\right)\\
\geq\ &\frac{1}{m}\sum_{j=1}^{m} p_{z_j}(2)\mathbb{P}(R_2|X\sim f_{z_j}(\cdot|2))+\frac{1}{m}\sum_{j=1}^{m} p_{z_j}(1)\mathbb{P}(R_1|X\sim f_{z_j}(\cdot|1))
\end{aligned}$$

Hence, using equation (6.2) we have

$$e_{CSV}(m)\geq e_{RCV}(m)\qquad \text{a.s.}$$

□

Proof of Theorem 2. Using the optimality of the Bayes error rate, we have

$$\begin{aligned}
&e_{RCV}(m)\\
&\leq \frac{\sum_{j=1}^{m} p_{Z_j}(2)}{m}\cdot\mathbb{P}\left(R_2(Z_1,\cdots,Z_{m+1})\middle| X\sim\frac{1}{\sum_{j=1}^{m} p_{Z_j}(2)}\sum_{i=1}^{m} p_{Z_i}(2)f_{Z_j}(\cdot|2)\right)\\
&\quad+\frac{\sum_{j=1}^{m} p_{Z_j}(1)}{m}\cdot\mathbb{P}\left(R_1(Z_1,\cdots,Z_{m+1})\middle| X\sim\frac{1}{\sum_{j=1}^{m} p_{Z_j}(1)}\sum_{i=1}^{m} p_{Z_i}(1)f_{Z_i}(\cdot|1)\right)\\
&=\frac{1}{m}\sum_{j=1}^{m} p_{Z_j}(2)\mathbb{P}(R_2(Z_1,\cdots,Z_{m+1})|X\sim f_{Z_j}(\cdot|2))\\
&\qquad+\frac{1}{m}\sum_{j=1}^{m} p_{Z_j}(1)\mathbb{P}(R_1(Z_1,\cdots,Z_{m+1})|X\sim f_{Z_j}(\cdot|1))
\end{aligned}$$

where the last equality is obtained using the same argument in equation (6.1). Notice that in $\mathbb{P}(\cdot)$, the probability is with respect to X only, i.e., $Z_1,...,Z_m$ are fixed.

Thus,

$$\begin{aligned}
\mathbb{E}[e_{RCV}(m)]\ &\leq\ \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\ p_{Z_j}(2)\cdot\mathbb{P}(R_2(Z_1,\cdots,Z_{m+1})|X\sim f_{Z_j}(\cdot|2))\\
&\qquad\qquad+p_{Z_j}(1)\cdot\mathbb{P}(R_1(Z_1,\cdots,Z_{m+1})|X\sim f_{Z_j}(\cdot|1))]\\
&=\ \mathbb{E}[\ p_{Z_1}(2)\cdot\mathbb{P}(R_2(Z_1,\cdots,Z_{m+1})|X\sim f_{Z_1}(\cdot|2))\\
&\qquad+p_{Z_1}(1)\cdot\mathbb{P}(R_1(Z_1,\cdots,Z_{m+1})|X\sim f_{Z_1}(\cdot|1))]\\
&=\ \frac{1}{m+1}\sum_{j=1}^{m+1}\mathbb{E}[\ p_{Z_j}(2)\cdot\mathbb{P}(R_2(Z_1,\cdots,Z_{m+1})|X\sim f_{Z_j}(\cdot|2))\\
&\qquad\qquad+p_{Z_j}(1)\cdot\mathbb{P}(R_1(Z_1,\cdots,Z_{m+1})|X\sim f_{Z_j}(\cdot|1))].
\end{aligned}$$

Again, by equation (6.2), we obtain

$$\mathbb{E}[e_{RCV}(m)]\leq\mathbb{E}[e_{RCV}(m+1)].$$

□

Proof of Theorem 3. In this proof, in order to make it more clear what is fixed and random, we replace $z_1, z_2, \cdots$ by $Z_1, Z_2, \cdots$, an i.i.d. sequence. Then, for almost any *fixed* x, we have

$$\begin{aligned} &\tfrac{1}{m}\textstyle\sum_{i=1}^m p_{Z_i}(1) f_{Z_i}(x|1) \to \mathbb{E} p_Z(1) f_Z(x|1) = p(1)\cdot f(x|1) \\ \text{and} \quad &\tfrac{1}{m}\textstyle\sum_{i=1}^m p_{Z_i}(2) f_{Z_i}(x|2) \to \mathbb{E} p_Z(2) f_Z(x|2) = p(2)\cdot f(x|2) \end{aligned} \tag{6.3}$$

with probability one. By a standard argument in measure theory, with probability one, (6.3) holds for *almost every* x (with respect to Lebesgue measure). Let

$$E = \{x : p(1)f(x|1) > p(2)f(x|2)\},$$

$$F = \{x : p(1)f(x|1) \le p(2)f(x|2)\},$$

$$G = \{x : p(1)f(x|1) = p(2)f(x|2)\},$$

$$E_m = \left\{ x : \frac{1}{m}\sum_{i=1}^m p_{Z_i}(1) f_{Z_i}(x|1) > \frac{1}{m}\sum_{i=1}^m p_{Z_i}(2) f_{Z_i}(x|2) \right\}$$

and

$$F_m = \left\{ x : \frac{1}{m}\sum_{i=1}^m p_{Z_i}(1) f_{Z_i}(x|1) \le \frac{1}{m}\sum_{i=1}^m p_{Z_i}(2) f_{Z_i}(x|2) \right\}.$$

Then, we know that almost surely $\lim_{m\to\infty} \mathbb{1}_{E_m}(x) = 1$ for almost every $x \in E$, and almost surely $\lim_{m\to\infty} \mathbb{1}_{F_m}(x) = 1$ for almost every $x \in F \setminus G$. Therefore, almost surely, for almost every $x \in G$,

$$\begin{aligned} & \lim_{m\to\infty} \left(\mathbb{1}_{E_m}(x) \frac{1}{m}\sum_{i=1}^m p_{Z_i}(2) f_{Z_i}(x|2) + \mathbb{1}_{F_m}(x) \frac{1}{m}\sum_{i=1}^m p_{Z_i}(1) f_{Z_i}(x|1) \right) \\ = & \lim_{m\to\infty} (p(2)\mathbb{1}_{E_m}(x) f(x|2) + p(1)\mathbb{1}_{F_m}(x) f(x|1)) \\ = & \lim_{m\to\infty} (p(1)\mathbb{1}_{E_m}(x) f(x|1) + p(1)\mathbb{1}_{F_m}(x) f(x|1)) \\ = & \; p(1) f(x|1) = p(2)\mathbb{1}_E(x) f(x|2) + p(1)\mathbb{1}_F(x) f(x|1). \end{aligned}$$

Thus, almost surely for almost all x,

$$\begin{aligned} &\lim_{m\to\infty} \left(\mathbb{1}_{E_m}(x) \tfrac{1}{m}\textstyle\sum_{i=1}^m p_{Z_i}(2) f_{Z_i}(x|2) + \mathbb{1}_{F_m}(x) \tfrac{1}{m}\sum_{i=1}^m p_{Z_i}(1) f_{Z_i}(x|1) \right) \\ &= p(2)\mathbb{1}_E(x) f(x|2) + p(1)\mathbb{1}_F(x) f(x|1). \end{aligned} \tag{6.4}$$

Now, since

$$e_{RCV}(m) = \int \left(\mathbb{1}_{E_m}(x) \frac{1}{m}\sum_{i=1}^m p_{Z_i}(2) f_{Z_i}(x|2) + \mathbb{1}_{F_m}(x) \frac{1}{m}\sum_{i=1}^m p_{Z_i}(1) f_{Z_i}(x|1) \right) dx,$$

by the Generalized Lebesgue Dominated Convergence theorem, almost surely

$$\lim_{m\to\infty} e_{RCV}(m) = \int \left(p(2)\mathbb{1}_E(x) f(x|2) + p(1)\mathbb{1}_F(x) f(x|1) \right) dx = e_{opt}.$$

Next, write

$$e_{CSV}(m) = \frac{1}{m}\sum_{i=1}^m \int (p_{Z_i}(2)\mathbb{1}_{{}_i E_m}(x) f_{Z_i}(x|2) + p_{Z_i}(1)\mathbb{1}_{{}_i F_m}(x) f_{Z_i}(x|1)) dx$$

where ${}_iE_m = \{x : \frac{1}{m}\sum_{j\neq i} p_{Z_j}(1) f_{Z_j}(x|1) > \frac{1}{m}\sum_{j\neq i} p_{Z_j}(2) f_{Z_j}(x|2)\}$ and ${}_iF_m = \{x : \frac{1}{m}\sum_{j\neq i} p_{Z_j}(1) f_{Z_j}(x|1) \leq \frac{1}{m}\sum_{j\neq i} p_{Z_j}(2) f_{Z_j}(x|2)\}$. Then, since $Z_1, \cdots, Z_m$ are i.i.d., we have

$$\begin{aligned}\mathbb{E}(e_{CSV}(m)) &= \mathbb{E}\int (p_{Z_1}(2)\mathbb{1}_{{}_1E_m}(x) f_{Z_1}(x|2) + p_{Z_1}(1)\mathbb{1}_{{}_1F_m}(x) f_{Z_1}(x|1))dx \\ &= \mathbb{E}\int (p(2)\mathbb{1}_{{}_1E_m}(x) f(x|2) + p(1)\mathbb{1}_{{}_1F_m}(x) f(x|1))dx.\end{aligned}$$

Similar to the proof of equation (6.4), almost surely for almost every x

$$\lim_{m\to\infty}\Big(p(2)\mathbb{1}_{{}_1E_m}(x) f(x|2) + p(1)\mathbb{1}_{{}_1F_m}(x) f(x|1)\Big) = p(2)\mathbb{1}_E(x) f(x|2) + p(1)\mathbb{1}_F(x) f(x|1).$$

By dominated convergence theorem, we obtain $\lim_{m\to\infty}\mathbb{E}(e_{CSV}(m)) = e_{opt}$. Because $e_{CSV}(m) \geq e_{RCV}(m)$ by Theorem 1, we have

$$\lim_{m\to\infty}\mathbb{E}|e_{CSV}(m) - e_{RCV}(m)| = \lim_{m\to\infty}\mathbb{E}(e_{CSV}(m) - e_{RCV}(m)) = e_{opt} - \lim_{m\to\infty}\mathbb{E}e_{RCV}(m) = 0,$$

by the dominated convergence theorem again. Therefore, $e_{CSV}(m) - e_{RCV}(m)$ converges to zero in probability. Hence, $e_{CSV}(m)$ converges to e_{opt} in probability. Now if the set $\{x : p(1)f(x|1) = p(2)f(x|2) \neq 0\}$ is of measure zero, then, almost surely,

$$e_{CSV}(m) = \int_{E\cup(F\setminus G)}\Big(\frac{1}{m}\sum_{i=1}^{m} p_{Z_i}(2)\mathbb{1}_{{}_iE_m}(x) f_{Z_i}(x|2) + \frac{1}{m}\sum_{i=1}^{m} p_{Z_i}(1)\mathbb{1}_{{}_iF_m}(x) f_{Z_i}(x|1)\Big)dx.$$

Here we have used the fact that $\mathbb{E}I_G = 0$ by Fubini's theorem where

$$I_G = \int_G\Big(\frac{1}{m}\sum_{i=1}^{m} p_{Z_i}(2)\mathbb{1}_{{}_iE_m}(x) f_{Z_i}(x|2) + \frac{1}{m}\sum_{i=1}^{m} p_{Z_i}(1)\mathbb{1}_{{}_iF_m}(x) f_{Z_i}(x|1)\Big)dx$$

which implies $I_G = 0$ almost surely. Since almost surely $\lim_{m\to\infty}\min_{1\leq i\leq m}\mathbb{1}_{{}_iE_m}(x) = 1$ for almost every $x \in E$ and almost surely $\lim_{m\to\infty}\min_{1\leq i\leq m}\mathbb{1}_{{}_iF_m}(x) = 1$ for almost every $x \in F\setminus G$, following the same argument in the proof of $\lim_{m\to\infty} e_{RCV}(m) = e_{opt}$ above, we can obtain that almost surely

$$\begin{aligned}&\lim_{m\to\infty}(\frac{1}{m}\sum_{i=1}^{m} p_{Z_i}(2)\mathbb{1}_{{}_iE_m}(x) f_{Z_i}(x|2) + \frac{1}{m}\sum_{i=1}^{m} p_{Z_i}(1)\mathbb{1}_{{}_iF_m}(x) f_{Z_i}(x|1)) \\ &= p(2)\mathbb{1}_E(x) f(x|2) + p(1)\mathbb{1}_F(x) f(x|1)\end{aligned}$$

for almost every $x \in E \cup (F \setminus G)$. Hence, by dominated convergence theorem, $e_{CSV}(m)$ converges to e_{opt} almost surely.

□

References

[1] R.B. Altman, H.K. Kroemer, C.A. McCarty, M.J. Ratain, and D. Roden. Pharmacogenomics: will the promise be fulfilled? *Nature Reviews Genetics*, 12:69–73, 2011.

[2] U.M. Braga-Neto. Fads and fallacies in the name of small-sample microarray classification. *Signal Processing Magazine, IEEE*, 24:91–99, 2007.

[3] U.M. Braga-Neto and E. R. Dougherty. Is cross-validation valid for small-sample microarray classification. *Bioinformatics*, 20:374–380, 2004.

[4] L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA*, 103:5923–5928, 2006.

[5] L. Shi et al. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24:1151–1161, 2006.

[6] P.A. Hoen et al. Reproducibility of high-throughput mrna and small rna sequencing across laboratories. *Nature Biotechnology*, 31:1015–1022, 2013.

[7] J.P. Evans, E.M. Meslin, T.M. Marteau, and T. Caulfield. Deflating the genomic bubble. *Science*, 331:861–862, 2011.

[8] Z. He and W. Yu. Stable feature selection for biomarker discovery. *Comput. Biol. Chem.*, 34:215–225, 2010.

[9] J.P.A. Ioannidis, D.B. Allison, C.A. Ball, I. Coulibaly, X. Cui, A.C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G.P. Page, E. Petretto, and V. Noort. Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41:149–155, 2009.

[10] P. Kirk, A. Witkover, C. Bangham, S. Richardson, A. Lewin, and M. Stumpf. Balancing the robustness and predictive performance of biomarkers. *Journal of Computational Biology*, 20, 2013.

[11] S. Ma, C.C. Funk, and N.D. Price. Systems approaches to molecular cancer diagnostics. *Discovery Medicine*, 10:531–542, 2010.

[12] S. Ma, J. Sung, A. Magis, Y. Wang, D. Geman, and N. Price. Measuring the effect of inter-study variability on learning molecular signatures. *preprint*, 2013.

[13] E. Marshall. Waiting for the revolution. *Science*, 331:526–529, 2011.

[14] N. Meinshausen and P. Buhlmann. Stability selection. *J. Roy. Stat. Soc. B*, 72, 2010.

[15] C.M. Micheel, S.J. Nass, and G.S. Omenn. *Evolution of translational omics: lessons learned and the path forward*. The National Academies Press, 2012.

[16] S. Michiels, S. Koscielny, and C. Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365:488–492, 2005.

[17] D.W. Scott and G.R. Terrell. Biased and unbaised cross-validation in density estimation. *Journal of the American Statistical Association*, 82:1131–1146, 1987.

[18] R. Shen, D. Ghosh, and A. Chinnaiyan. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics*, 5:94, 2004.

[19] J. Sung, Y. Wang, S. Chandrasekaran, D.M. Witten, and N.D. Price. Molecular signatures from omics data: From chaos to consensus. *Biotechnology Journal*, 7:946–957, 2012.

[20] A. Teschendorff, A. Naderi, N. Barbosa-Morais, S. Pinder, I. Ellis, S. Aparicio, J. Brenton, and C. Caldas. A consensus prognostic gene expression classifier for er positive breast cancer. *Genome Biology*, 7:r101, 2006.

[21] R.J. Tibshirani and R. Tibshirani. A bias correction for the minimum error rate in cross-validation. *Annals of Applied Statistics*, 3:822–829, 2009.

[22] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:1397873, 2006.

[23] R.L. Winslow, N. Trayanova, D. Geman, and M. Miller. Computational medicine: translating models to clinical care. *Science Translational Medicine*, 4:158rv11, 2012.

[24] L. Xu, A-C Tan, D. Naiman, D. Geman, and R. Winslow. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, 21:3905–3911, 2005.

[25] L. Xu, A.C. Tan, R. L. Winslow, and D. Geman. Merging microarray data from separate breast cancer studies provides a robust prognostic signature. *BMC Bioinformatics*, 9:125, 2008.

[26] H. Yang, C.A. Harrington, K. Vartanian, C.D. Coldren, R. Hall, and G.A. Churchill. Randomisation in laboratory procedure is key to obtaining reproducible microarray results. *PLos One*, 3(11):e3724, 2008.