LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Anne-Laure Boulesteix, Robert Hable, Sabine Lauer, Manuel Eugster

A Statistical Framework for Hypothesis Testing in Real Data Comparison Studies

Technical Report Number 136, 2013
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

A Statistical Framework for Hypothesis Testing in Real Data Comparison Studies

Anne-Laure Boulesteix[1], Robert Hable[2,3],

Sabine Lauer[1], Manuel J. A. Eugster[2,4]

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich (LMU), Marchioninistr. 15, D-81377 Munich, Germany. boulesteix@ibe.med.uni-muenchen.de

[2] Department of Statistics, University of Munich (LMU), Munich, Germany

[3] Chair of Stochastics, Department of Mathematics, University of Bayreuth, Bayreuth, Germany

[4] Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Espoo, Finland

Anne-Laure Boulesteix is a statistician and biometrician, Department of Medical Informatics, Biometry and Epidemiology, University of Munich (LMU), Munich, Germany. Robert Hable is a mathematician and statistician, Department of Mathematics, University of Bayreuth, Bayreuth, Germany. Sabine Lauer is a statistician and sociologist, Department of Medical Informatics, Biometry and Epidemiology, University of Munich (LMU), Munich, Germany. Manuel Eugster is a computer scientist and statistician, Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Espoo, Finland.

ABSTRACT

In computational sciences, including computational statistics, machine learning, and bioinformatics, most abstracts of articles presenting new supervised learning methods end with a sentence like "our method performed better than existing methods on real data sets", e.g. in terms of error rate. However, these claims are often not based on proper statistical tests and, if such tests are performed (as usual in the machine learning literature), the tested hypothesis is not clearly defined and poor attention is devoted to the type I and type II error. In the present paper we aim to fill this gap by providing a proper statistical framework for hypothesis tests comparing the performance of supervised learning methods based on several real data sets with unknown underlying distribution. After giving a statistical interpretation of ad-hoc tests commonly performed by machine learning scientists, we devote special attention to power issues and suggest a simple method to determine the number of data sets to be included in a comparison study to reach an adequate power. These methods are illustrated through three comparison studies from the literature and an exemplary benchmarking study using gene expression microarray data. All our results can be reproduced using R-codes and data sets available from the companion website `http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/compstud2013`.

KEYWORDS

Benchmarking, supervised learning, comparison, testing, performance

1. INTRODUCTION

Almost all machine learning or computational statistics articles on supervised learning methods include a more or less extensive comparison study based on real data sets of moderate size assessing the respective performance (typically in terms of prediction error) of a few algorithms, either new or already described in previous literature. These comparisons are often termed as "benchmark experiments" (e.g. Hothorn et al., 2005) in the statistical and machine learning communities.

However, the statistical foundations of these comparisons are usually given surprisingly poor attention in concrete comparison studies, although they are addressed in a large body of methodological literature. For simplicity we will from now on talk about the comparison of two supervised classification methods, but all the ideas discussed here are also relevant to the comparison of more than two methods or to other supervised learning problems. These two methods are used to derive a "classification rule" from an available training sample.

In a seminal paper Dietterich (1998) proposes a general taxonomy of the problems related to performance evaluation and comparison in supervised learning. The comparison of two methods for a particular distribution in the absence of large sample is termed "Question 8" while Questions 1 to 7 consider error estimation (as opposed to comparison of methods), situations with large samples, and/or the assessment of specific classification rules (as opposed to the problem – considered here – of the assessment of the methods used to derive classification rules). The paper considers a few simple testing procedures based on resampling-based estimates of the prediction error and compares them empirically via simulations in terms of type I error and power. The conclusion is that none of these procedures is completely satisfactory. The essential problem is that they do not use adequate estimates of the unconditional variance of the error estimates. Estimating the unconditional variance based on the empirical variance over resampling iterations implies a violation of independence assumptions and thus a reduction of the effective degrees of freedom

(Bouckaert, 2003).

Nadeau and Bengio (2003) provide an overview of estimators of the unconditional variance of resampling-based error estimates used in the machine learning community and suggest two additional estimators that can be applied to error estimators obtained through repeated splitting into training and test data. They present benchmark experiments as a natural application of their new estimators and stress that naive estimators of the variance (e.g. the empirical variance of the estimated error over resampling iterations) often lead to false positives in the sense that researchers see a difference in performance between the considered methods although there is no such difference. Hothorn et al. (2005) give a statistical interpretation of these issues and recommend to estimate the variance of resampling-based error estimates through bootstrapping.

Dietterich (1998) also mentions a further question termed as "Question 9" in his taxonomy and referring to several domains, where the term "domain" here denotes a data set with its own underlying distribution:

> *"Question 9: Given two learning algorithms A and B and data sets from several domains, which algorithm will produce more accurate classifiers when trained on examples from new domains? This is perhaps the most fundamental and difficult question in machine learning."*

Indeed, almost all comparison studies based on real data consider not one but several data sets, thus implicitly involving several different underlying distributions. While it is usual to perform hypothesis tests to compare the performance of different methods in the context of benchmarkring studies (Demšar, 2006), the literature on the *statistical* interpretation of such tests based on multiple data sets is surprisingly sparse. One of the few articles on this topic suggests to use a mixed model approach with the data sets as *subjects* with random intercept and the methods as fixed effects (Eugster et al., 2012).

In this paper, our aim is to give a statistical formulation of tests performed in the context of comparison studies and to correspondingly interpret results of published comparison studies, with focus on classification based on high-dimensional gene expression data as an highlighting illustration. The paper is structured as follows. In Section 2 we outline the epistemological background of comparisons of prediction errors from a statistical perspective contrasting with the machine learning perspective taken by most papers handling this topic. Section 3 is especially devoted to tests in the context of comparison studies based on several real data sets and gives an original formulation of these tests in a strict statistical framework. In particular, Section 3 suggests a power calculation approach in this context. Section 4 presents an application of these methods to three comparison studies of classification methods for microarray data from the literature, while Section 5 describes an exemplary benchmark study based on 50 data sets. The appendix contains some technicalities which are needed for the mathematically rigorous formulation of the testing problem given in Section 3.

2. EPISTEMOLOGICAL BACKGROUND

Settings

From a statistical point of view, binary supervised classification can be described in the following way. On the one hand, we have a response variable taking values in $\mathcal{Y}$ $= \{0, 1\}$. On the other hand, we have predictors taking values in $\mathcal{X} \subset \mathbb{R}^p$ that will be used for constructing a classification rule. Predictors $\boldsymbol{X}$ and response Y follow an unknown joint distribution on $\mathcal{X} \times \mathcal{Y}$ denoted by P. The observed *i.i.d.* sample of size n is denoted by $s_0 = \{(\boldsymbol{x}_1, y_1)...(\boldsymbol{x}_n, y_n)\}$. The classification task consists in building a decision function $\hat{f}$ that maps elements of the predictor space $\mathcal{X}$ into the response space $\mathcal{Y}$:

$$\begin{array}{rccc} \hat{f}^{s_0}: & \mathcal{X} & \mapsto & \mathcal{Y}, \\ & \boldsymbol{x} & \mapsto & \hat{f}^{s_0}(\boldsymbol{x}), \end{array}$$

where the superscript s_0 indicates that the decision function is built using the sample s_0. For simplicity we assume that the classification method is deterministic, i.e. that $\hat{f}^{s_0}$ is uniquely defined given the sample s_0. There are many possible methods to fit a function $\hat{f}^{s_0}$ based on a sample s_0, which we denote as $M_1, \dots, M_K$. The decision function obtained by fitting method M_k to the sample s_0 is denoted as $\hat{f}^{s_0}_{M_k}$.

The true error of this classification rule $\hat{f}^{s_0}_{M_k}$ can be written as

$$\varepsilon(\hat{f}^{s_0}_{M_k}, P) = \boldsymbol{E}_P \left[L\left(\hat{f}^{s_0}_{M_k}(\boldsymbol{X}), Y \right) \right] \tag{1}$$

where $\boldsymbol{E}_P$ stands for the mean over the joint distribution P of $\boldsymbol{X}$ and Y, and $L(.,.)$ is an adequate loss function, e.g. the indicator loss yielding the classification error rate considered in this paper.

The true error $\varepsilon(\hat{f}^{s_0}_{M_k}, P)$ of method M_k constructed using sample s_0 is commonly referred to as *conditional* error since it corresponds to the decision function constructed on the specific sample s_0. The notation $\varepsilon(\hat{f}^{s_0}_{M_k}, P)$ stresses that this error depends on the distribution P as well as on the method M_k and sample s_0 used to fit the classification rule.

In this perspective, the error $\varepsilon(\hat{f}^{S}_{M_k}, P)$ (corresponding to Eq. (1) with s_0 replaced by S) should be seen as a random variable, where S stands for a random i.i.d. sample that follows the distribution P^n. The mean of this random variable over P^n is commonly referred to as the *unconditional* true error rate of method M_k. In this paper it is denoted as

$$\varepsilon(n, M_k, P) \;=\; \boldsymbol{E}_{P^n}[\varepsilon(\hat{f}^{S}_{M_k}, P)].$$

It depends only on the method M_k, on the size n of the sample and on the joint distribution P of $\boldsymbol{X}$ and Y. Note that the joint distribution P is involved twice in this formula.

Method 2 is "better" than method 1: what does this mean?

Now suppose that we are interested in the relative performance of two methods M_1 and M_2. What does it mean when we ask whether method M_2 is "better" than method M_1 or vice-versa? An applicant (for instance a biologist) who collected a specific data set s_0 in his lab is primarily interested in whether the classification rule $\hat{f}_{M_2}^{s_0}$ fitted with method M_2 has a smaller error on future independent data than the classification rule $\hat{f}_{M_1}^{s_0}$ fitted with method M_1 or vice-versa, i.e. whether $\varepsilon(\hat{f}_{M_2}^{s_0}, P) < \varepsilon(\hat{f}_{M_1}^{s_0}, P)$. We define the null- and alternative hypotheses of the applicant correspondingly as

$$\begin{array}{rlcll} & H_0^{(cond)}: & \varepsilon(\hat{f}_{M_2}^{s_0}, P) & - & \varepsilon(\hat{f}_{M_1}^{s_0}, P) & \geq 0 \\ \text{vs.} & H_1^{(cond)}: & \varepsilon(\hat{f}_{M_2}^{s_0}, P) & - & \varepsilon(\hat{f}_{M_1}^{s_0}, P) & < 0. \end{array}$$

These hypotheses can be seen as conditional (hence the exponent "(*cond*)") in the sense that they are conditional on a fixed sample s_0.

In contrast, statisticians or machine learners doing methodological research are not primarily interested in the performance of the classification rule fitted on a specific sample s_0 but rather on the mean performance over different samples. In mathematical words, they are interested in the comparison of the unconditional errors $\varepsilon(n, M_1, P)$ and $\varepsilon(n, M_2, P)$, yielding the corresponding hypotheses:

$$\begin{array}{rlcll} & H_0^{(uncond)}: & \varepsilon(n, M_2, P) & - & \varepsilon(n, M_1, P) & \geq 0 \\ \text{vs.} & H_1^{(uncond)}: & \varepsilon(n, M_2, P) & - & \varepsilon(n, M_1, P) & < 0, \end{array}$$

with the exponent "(*uncond*)" standing for "unconditional". When methodological researchers write that method M_2 is better than the standard method M_1, they implicitly mean that the unconditional error is smaller for M_2 than for M_1 and that $H_0^{(uncond)}$ can be rejected.

In practical data analysis, when the distribution P is unknown, it is not easy to test $H_0^{(uncond)}$. A natural estimator of the difference $\varepsilon(n, M_2, P) - \varepsilon(n, M_1, P)$ is the difference between resampling-based error estimates obtained with the two considered methods, e.g. cross-validation error estimates. The problem is that the true *unconditional* variance of this difference under $H_0^{(uncond)}$ is unknown and difficult to estimate. Many naive or more complex estimates of this variance have been considered in the literature (Dietterich, 1998; Nadeau and Bengio, 2003; Hanczar and Dougherty, 2010), but they rely on uncertain assumptions that are most often not met in practical cases. The essential problem is that they are all based on the available sample s_0, while their target is actually the variance over different samples drawn from P^n. Hence, they are all conditional in some way. To date, there exists no widely accepted test for testing $H_0^{(uncond)}$ based on a real data set with unknown underlying distribution.

Epistemological background: theory and simulations

For a given distribution P and a specific n, however, it is easy to empirically approximate the unconditional error $\varepsilon(n, M_1, P)$ of a given classification method M_1 via simulations. A straightforward procedure is as follows:

1. Randomly draw a huge number n_{test} of independent realizations of P, yielding a so-called "test sample" $s^{(T)} = \{(\boldsymbol{x}_1^{(T)}, y_1^{(T)}), \ldots, (\boldsymbol{x}_{n_{\text{test}}}^{(T)}, y_{n_{\text{test}}}^{(T)})\}$. For instance, $n_{\text{test}} = 10000$ may be appropriate.

2. For $b = 1, \ldots, B$, with B large (typically $B \geq 1000$):

   2.a. Draw n i.i.d. realizations from the distribution P, yielding the training sample $s^{(b)} = ((\boldsymbol{x}_1^{(b)}, y_1^{(b)}), \ldots, (\boldsymbol{x}_n^{(b)}, y_n^{(b)}))$.

   2.b. Fit method M_1 to $s^{(b)}$, yielding the classification rule $\hat{f}_{M_1}^{s^{(b)}}$.

2.c. Estimate the true error of $\hat{f}^{s^{(b)}}_{M_1}$ based on the test sample drawn in step 1 as the proportion of misclassified realizations in the test sample

$$\hat{\varepsilon}(\hat{f}^{s^{(b)}}_{M_1}, P) \;=\; \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} L(\hat{f}^{s^{(b)}}_{M_1}(\boldsymbol{x}_i^{(T)}), y_i^{(T)}).$$

3. Estimate the true unconditional error $\varepsilon(n, M_1, P)$ as

$$\hat{\varepsilon}(n, M_1, P) \;=\; \frac{1}{B} \sum_{b=1}^{B} \hat{\varepsilon}(\hat{f}^{s^{(b)}}_{M_1}, P).$$

Note that this approach implies two embedded approximation procedures: approximation of the true error of $\hat{f}^{s^{(b)}}_{M_1}$ using a huge test sample $s^{(T)}$, and approximation of the unconditional error over P^n by averaging over a large number B of random training samples $s^{(b)}$.

Note that in specific cases, the unconditional error might even be derived analytically. No matter whether one derives this error analytically or via simulations, two parameters have to be chosen: the sample size n and the joint distribution P. Except for trivial examples (e.g. an algorithm that randomly generates a rule f without looking at the data), it is not to be expected that the new method M_2 performs better than the standard method M_1 for all sample sizes and all imaginable joint distributions – the so-called "no free lunch"-theorem (Wolpert, 2001).

Epistemological background: real data study

The essential limitation of simulations and analytical results is that the chosen distribution P often does not reflect the complexity of "real life distributions". Therefore, performance estimation on real data is usually considered as extremely important. In essence, the goal of such studies is to evaluate the considered classification methods M_1 and M_2 on "real life distributions" P, i.e. to compare $\varepsilon(n, M_k, P)$ for $k = 1, 2$. There are however two important problems related to real data studies.

The first problem ("variability of error estimation") is related to the fact that for a specific data set the underlying distribution P is essentially unknown in practice. It is thus impossible to derive the prediction error analytically. If a huge sample is given, we can obtain a good approximation by drawing numerous non-overlapping samples of the considered size n out of the huge sample, and estimating the error on the rest of the sample. However, in most practical situations no huge samples are given and resampling methods are then used to address this estimation issue. Such methods, however, are known to poorly estimate the unconditional error because they suffer a very high variance and/or a high bias resulting in a large mean squared error (Braga-Neto and Dougherty, 2005; Zollanvari et al., 2009; Dougherty et al., 2011; Zollanvari et al., 2011; Dalton and Dougherty, 2012a,b).

The second problem ("variability across data sets") is that each real life data set follows its own distribution. In the example of microarray gene expression data, the leukemia data set of size $n = 38$ by Golub et al. (1999) follows a certain distribution P_1 while the breast cancer data set of size $n = 76$ by Van't Veer et al. (2002) follows another distribution P_2. Even if we had a good estimate of the unconditional error $\varepsilon(n = 38, M_1, P_1)$ for method M_1, it would tell us nothing about $\varepsilon(n = 38, M_1, P_2)$ and $\varepsilon(n = 76, M_1, P_2)$, except if we assume that P_2 is somehow "similar" to P_1, an assumption that may make sense in some exceptional cases, but not in general. When researchers perform a real data study based on several data sets, they implicitly aim to capture the variability across data sets, i.e. across distributions.

A problem related to the variability across data sets is the definition of the aimed area of application. It can be defined in a very general way without restrictions, or the authors may choose to focus on a very particular area of application with restrictions regarding the structure of the data and/or the substantive context of the data sets. No matter how the area of application is defined and how example data sets are selected, the two sources of variability (variability of error estimation and variability across data sets) imply a high variance. The variance of error estimation can be addressed by using an estimation method known to have smaller variance: for instance, we know that repeated subsampling with a training/test splitting ratio of, say, 2:1 has smaller (unconditional) variance than leave-one-out cross-validation (LOOCV). However, the variance remains high even for the less variable methods. As to the variability across data sets, it can only be addressed by increasing the number of candidate data sets.

Therefore – after motivating the issues with an illustrative exemplary scenario – we formally address the problem of testing the difference between the unconditional errors of two methods M_1 and M_2 in a real data study with several data sets and suggest a statistical framework, including power considerations.

Examples

Suppose that authors want to compare two simple statistical procedures for supervised classification based on high-dimensional microarray data to be used after a variable selection step: linear discriminant analysis (method A, considered as the reference), and diagonal linear discriminant analysis (method B, considered as a new method suggested by the authors).

Method B is expected to work well if the covariance matrix is indeed diagonal (which depends on P). Method A may have problems if the sample size n is not large compared to the number of predictors (which also depends on P via the dimension of $\boldsymbol{X}$), because the estimated covariance matrix is then ill-conditioned and it inversion is problematic. But it may work well if the true model implied by the distribution P involves correlations between the variables and if n is large enough. Quite generally, for learning algorithms based on the estimation of parameters of a stochastic model, the closer the assumed model to the true distribution P, the better the prediction accuracy in infinite sample settings. Things are more complicated for learning algorithms that are not based on an underlying stochastic model and in finite sample settings, but we can again say that the error essentially depends on n

and P.

When evaluating methods based on simulated data, the choice of the distribution P and the sample size n is thus crucial. A particular distribution and a particular sample size n can be realistic in a considered area of application, but not realistic for another one. In practice, it is recommended that researchers consider distributions and sample sizes that are typical for the area of application they are aiming at. For example, a sample size of $n = 100$ and a distribution P involving $p = 10000$ multivariate Gaussian covariates with block diagonal structure, of which $p^* = 20$ are related to the response class Y through a logistic model may more or less reflect the reality of microarray gene-expression data, but not of large epidemiological studies involving only a handful of covariates, or of imaging data with complex spatial structure.

In practice, methodological researchers often develop methods addressing a particular type of distribution P (sometimes combined with a particular sample size range). They correspondingly design a simulation study based on such a distribution and/or sample size. For instance, the authors comparing methods A and B would probably consider simulation settings with diagonal covariance structure. While it is obviously more than recommended to evaluate the new classification method in the data setting it was designed for, it is also recommended to i) clearly state that the superiority of the new method is specific to the investigated distributions (here: diagonal covariance matrix), ii) investigate other distributions as well to examine the robustness of the method to changes in the distribution (e.g. distributions with block-diagonal covariance matrix).

In real life, the data do not stem from simple joint distributions like those commonly used in simulations. That is why real life comparison studies are considered as very important in computational science. If we refer again to the example of methods A and B, the success of classification method B in the real data study with microarray data then depends on three essential factors: i) whether the considered data sets have a diagonal covariance structure, ii) how the new method performs if the structure is indeed diagonal, iii) whether the method is robust against deviations from the diagonal covariance structure. Obviously, method B will comparatively perform better in the real data study if condition i) holds. And it will have more impact in the future if diagonal covariance structures often occur not only in the selected data sets but in the whole considered area of application.

More generally, we can say that the data sets selected for the real data study should reflect the aimed area of application. For example, if one consciously selects microarray gene expression data with diagonal covariance structure, the real data analysis section should not be written as if the area of application were "microarray gene expression data" in general. Conversely, one could say that data sets should be selected randomly within the stated field of application. Most importantly, data sets should not be selected a posteriori based on the obtained results, because this procedure generates a substantial bias (Yousefi et al., 2010).

3. TESTING FRAMEWORK FOR REAL DATA STUDIES

Settings

Webb (2000) states that "it is debatable whether error rates in different domains [where the term domain here refers to the underlying joint distribution] are commensurable, and hence whether averaging error rates across domains is very meaningful. Nonetheless, a low average error rate is indicative of a tendency toward low error rates for individual domains." In this section, we try to address this issue in terms of statistical testing. More precisely, we consider the problem of "testing the error difference" between two methods M_1 and M_2 based on several real-life data sets. The first task we have to address is thus to properly define the testing problem at hand.

Let us consider a given area of application. We independently and randomly draw J data sets belonging to this area. It is questionable whether the selection of data sets would really be random over the area in practice. For example, researchers may preferably look for data sets in their favorite database. This database is not necessarily representative for the whole area, for instance because it includes, say, many small data sets (the distribution of n is then not the same as in the whole area) or more data sets for a particular disease (possibly implying specific forms for P). For simplicity, however, we assume in this paper that the data sets are randomly drawn from the considered area.

These data sets are denoted as $D_1, \ldots, D_J$. We intentionally do not use the notation S from the previous section to stress that the situation is now different: the data sets $D_1, \ldots, D_J$ are not drawn from the same distribution (as was the case for $s^{(1)}, \ldots, s^{(b)}$ in the previous section). Each data set D_j is as a realization of $P_j^{n_j}$, where n_j is its size and P_j is the distribution of the underlying population. As we assume that data sets are randomly drawn from the considered area, the distribution P_j is the outcome of a random variable $\Phi_j : \Omega \rightarrow \mathcal{V}$ where $\mathcal{V}$ is the set of all possible distributions (in the area of application). Furthermore, the size n_j of the data set D_j is the outcome of a random variable $N_j : \Omega \rightarrow \mathbb{N}$. The random variables $(\Phi_1, N_1), \ldots, (\Phi_J, N_J)$ are i.i.d. but, of course, we only observe $N_j = n_j$ and cannot observe $\Phi_j = P_j$ for every $j \in \{1, \ldots, J\}$.

Test hypothesis

When randomly drawing a data set D, one thus implicitly randomly draws simultaneously a distribution P and n realizations of this distribution. Both imply a certain variability. Importantly, P can now be seen as the outcome of a random variable Φ and n as the outcome of a random variable N. Note that Φ and N are not necessarily independent. The test hypotheses of interest that are implicitly considered when comparing the performance of methods based on different real data sets can be stated as

$$\begin{array}{rrcrcl} & H_0^{(real)}: & \boldsymbol{E}(\varepsilon(N, M_2, \Phi)) & - & \boldsymbol{E}(\varepsilon(N, M_1, \Phi)) & \geq 0 \\ \text{vs.} & H_1^{(real)}: & \boldsymbol{E}(\varepsilon(N, M_2, \Phi)) & - & \boldsymbol{E}(\varepsilon(N, M_1, \Phi)) & < 0, \end{array}$$

where $\boldsymbol{E}$ denotes the expectation over the random variables Φ and N here. Accordingly, $\varepsilon(n, M_k, P)$ is now the outcome of the random variable $\varepsilon(N, M_k, \Phi)$, $k \in \{1, 2\}$.

In comparison studies based on real data, researchers typically estimate the error for each data set using a resampling procedure such as, e.g., repeated splitting into training and test set. Let $e(n, M_k, D)$ denote the error of method M_k estimated for data set D with the chosen resampling procedure. The estimated error $e(n, M_k, D)$ can be seen as an estimator of the unknown parameter $\varepsilon(n, M_k, P)$. In resampling procedures, however, the training data set used at each resampling iteration is smaller than n, hence leading to an error $e(n, M_k, D)$ larger than $\varepsilon(n, M_k, P)$ on average. Nevertheless, if we assume that this bias is equal for both considered methods M_1 and M_2, i.e. that

$$\boldsymbol{E}_{P^n}(e(n, M_1, D)) - \varepsilon(n, M_1, P) \;=\; \boldsymbol{E}_{P^n}(e(n, M_2, D)) - \varepsilon(n, M_2, P),$$

we have

$$\varepsilon(n, M_2, P) - \varepsilon(n, M_1, P) \;=\; \boldsymbol{E}_{P^n}(e(n, M_2, D) - e(n, M_1, D)) \qquad (2)$$

where $\boldsymbol{E}_{P^n}$ denotes the expectation over the data set D drawn i.i.d. from P. The data set D can be seen as the outcome of a random variable $\mathbf{D}$ of which the conditional distribution given $\Phi = P$ and $N = n$ is equal to P^n. Then, using the relation

$$\boldsymbol{E}\big(\boldsymbol{E}(e(N, M_2, \mathbf{D}) - e(N, M_1, \mathbf{D})|\Phi,\ N)\big) = \boldsymbol{E}(e(N, M_2, \mathbf{D}) - e(N, M_1, \mathbf{D})), \qquad (3)$$

we can formulate the null-hypothesis $H_0^{(real)}$ as

$$\begin{array}{rrcl} & H_0^{(real)}: & \boldsymbol{E}(e(N, M_2, \mathbf{D}) \quad - \quad \boldsymbol{E}(e(N, M_1, \mathbf{D}) & \geq 0 \\ \text{vs.} & H_1^{(real)}: & \boldsymbol{E}(e(N, M_2, \mathbf{D}) \quad - \quad \boldsymbol{E}(e(N, M_1, \mathbf{D}) & < 0. \end{array}$$

This formulation is advantageous because we have access to independent and identically distributed realizations $e(n_j, M_2, D_j) - e(n_j, M_1, D_j)$ (with $j = 1, \ldots, J$) of $e(N, M_2, \mathbf{D}) - e(N, M_1, \mathbf{D})$, thus yielding estimates of its mean and its variance. Note that the exact measure theoretic formulation of (3) needs some more care because: (i) Φ is a random variable which takes its values in a set $\mathcal{V}$ of probability measures P so that we need a suitable σ-algebra on $\mathcal{V}$, (ii) the size $n = N$ of the data set $D = \mathbf{D}$ is random so that a naive formalization would yield that $\mathbf{D} \in (\mathcal{X} \times \mathcal{Y})^N$ where the dimension N is random, and (iii) the existence of a suitable random variable $\mathbf{D}$ (of which the conditional distribution given $\Phi = P$ and $N = n$ is equal to P^n) is not obvious. The mathematically correct derivation of the above testing problem which takes care of these technicalities is deferred to the appendix.

Let $\Delta e(n_j, D_j) = e(n_j, M_2, D_j) - e(n_j, M_1, D_j)$ (with $j = 1, \ldots, J$) be independent and identically distributed realizations of $e(N, M_2, \mathbf{D}) - e(N, M_1, \mathbf{D})$ and define $\overline{\Delta e} = \frac{1}{J} \sum_{j=1}^{J} \Delta e(n_j, D_j)$. Under normality assumption or for very large J it is now possible to perform a paired sample t-test to test $H_0^{(real)}$. The test statistic

$$T = \frac{\overline{\Delta e}}{\sqrt{\frac{1}{J} \frac{1}{J-1} \sum_{j=1}^{J} (\Delta e(n_j, D_j) - \overline{\Delta e})^2}}$$

follows a Student distribution with $J - 1$ degrees of freedom under $H_0^{(real)}$. This type of test is performed in many machine learning studies, where it is usual to apply new and existing methods to a large number of data sets (Demšar, 2006), for instance from databases especially designed for this purpose. Non-parametric tests such as the Wilcoxon signed-rank test are also commonly applied in this context (Demšar, 2006) including adjustment procedures for multiple comparisons or the use of global test statistics for the comparison of more than two methods (Garcia and Herrera, 2008; Garcia et al., 2010). When authors state in their abstracts that “the new method performs better than existing methods on real data sets”, they implicitly say that $H_0^{(real)}$ can be rejected, although most of them do not perform the test and give almost no attention to the theoretical background of these tests.

Decomposition of the variance

The variance term $\mathrm{Var}(\Delta e(N, \mathbf{D}))$ consists of two parts: firstly, the variance of $\Delta e(N, \mathbf{D})$ conditional on the distribution $\Phi = P$ and the sample size $N = n$; secondly, the variance of $\varepsilon(N, M_2, \Phi) - \varepsilon(N, M_1, \Phi)$:

$$\mathrm{Var}(\Delta e(N, \mathbf{D})) \quad = \quad \mathrm{Var}\big(\varepsilon(N, M_2, \Phi) - \varepsilon(N, M_1, \Phi)\big) + \boldsymbol{E}\big(\mathrm{Var}(\Delta e(N, \mathbf{D})|\Phi,\, N)\big) \tag{4}$$

Essentially, this follows from the law of total variance, Equation (2), and the fact that the conditional distribution of $\mathbf{D}$ given $\Phi = P$ and $N = n$ is equal to P^n; the exact derivation needs some more care again and is deferred to the appendix.

The literature on error estimation and comparison usually focuses on the second part of the variance in (4): several estimators have been proposed (Nadeau and Bengio, 2003). The first part of the variance is considered as disturbing factor by many authors. Data sets that behave differently from the other are generally considered as cumbersome outliers and sometimes even excluded from the comparison study, possibly leading to a substantial bias (Yousefi et al., 2010).

This part of the variance is also tightly related to the optimistic bias commonly observed in studies assessing new methods in comparison studies (Jelizarow et al., 2010). Researchers tend to overfit their new method to specific example data sets while developing them. The variance across data sets being high, this new method that has been optimized to these particular data sets is likely to perform much worse on other data sets.

Power considerations

Considering the one-sided one-sample t-test outlined above, the number J of data sets necessary to detect a given effect size $\frac{\Delta}{\sigma}$ at a certain power $1-\beta$ can be derived from the formula

$$J \;\approx\; \frac{[t_{1-\alpha,J-1} + t_{1-\beta,J-1}]^2}{(\frac{\Delta}{\sigma})^2},$$

where $t_{\alpha,df}$ denotes the α-quantile of the Student distribution with df degrees of freedom, α denotes the type I error (typically $\alpha = 0.05$), β denotes the type II error, Δ denotes the difference that we want to be able to detect, and σ denotes the standard deviation (Bock, 1998).

Conversely, for a given J, a given Δ and a given σ, one can also compute the power $1-\beta$ of the test as

$$1-\beta \;=\; \Phi_{t_{df=J-1}}\left(\sqrt{J\cdot\frac{\Delta^2}{\sigma^2}} - t_{1-\alpha,df=J-1}\right),$$

where $\Phi_{t_{df=J-1}}$ denotes the cumulative distribution function of the Student distribution with $J-1$ degrees of freedom. It is common practice to use such formulas to derive the adequate sample size in various types of experiments including, e.g. animal trials or clinical trials. Such a statistical planning, however, is never considered when performing benchmark experiments, even if the researchers want to eventually perform statistical tests. In this paper, we suggest to also give attention to such power considerations when planning or performing a comparison study involving several data sets. These issues are illustrated in the next section through an application to comparison studies of microarray-based classification methods.

4. ILLUSTRATION: POWER OF PUBLISHED STUDIES

In this section we examine comparison studies from the literature on microarray-based supervised classification with respect to the issue discussed above. In particular, we compute the empirical variance of the difference between methods and estimate the power of the comparisons.

Considered studies

We selected comparison studies (i) whose aim was *not* to establish the superiority of a “new method”, hence warranting a certain level of neutrality (Boulesteix et al., 2008; Boulesteix and Eugster, 2012), (ii) focusing on diagnosis or prognosis based on high-dimensional gene expression microarray data, (iii) representing the estimated error rates in form of a table (thus providing access to the exact figures) rather than in form of graphics, (iv) examining at least five data sets $N \geq 5$ and at least five classification methods. Three of the considered comparison studies fulfilled these four criteria: the study by Lee et al. (2005) comparing six methods on six cancer sdata sets, the study by Lai et al. (2006) comparing 13 methods on seven cancer data sets, and the study by Statnikov et al. (2005) comparing eight methods on 11 cancer data sets.

Standard deviation

We first derived the standard deviation of each pairwise difference between methods for each study. There were $6 \cdot 5/2 = 15$ pairwise differences for the Lee study, $13 \cdot 12/2 = 78$ pairwise differences for the Lai study, and $8 \cdot 7/2 = 28$ pairwise differences for the Statnikov study. The boxplots representing the standard deviations of these pairwise differences are displayed in Figure 1.

It can be seen from these boxplots that the range of the standard deviations is surprisingly the same for the three studies, except for a few more extreme values for the Statnikov study. Furthermore, the standard deviations are very different within the studies: some pairs of methods have a very low standard deviation (indicating that the data sets are similar with respect to the difference in performance), while other show large standard deviations.

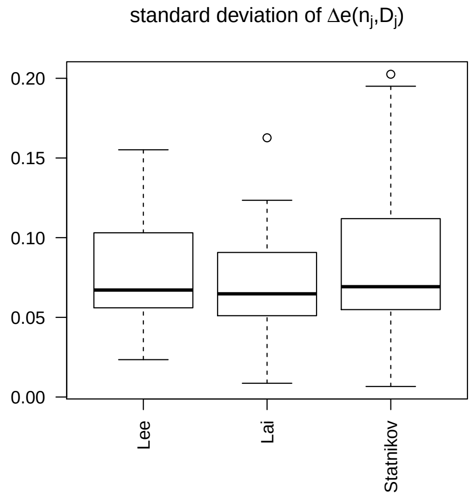

Figure 1: Estimated standard deviation of $\Delta e(n_j, D_j)$ in the three studies by Lee et al. (2005), Lai et al. (2006) and Statnikov et al. (2005).

Power considerations

Assuming values of the standard deviation σ in the range of those observed in the three investigated studies, Figure 2 (left panel) represents the number J of data sets required to detect a difference of error rate of Δ with power 80%. The right panel of Figure 2 displays the reached power against the number of data sets J for $\Delta = 0.05$ and different values of the standard deviation σ in the range of those observed in the three investigated studies. This figure suggests that most published comparison studies (either neutral comparison studies or comparison studies included in an original article) are substantially underpowered. In this perspective, we propose and recommend that researchers conducting comparison studies explicitly address such power issues for the design and/or interpretation of their benchmark experiments.

5. AN EXEMPLARY BENCHMARK STUDY

In order to explicitly illustrate these problems in real benchmark experiments we perform an exemplary benchmark study based on $J = 50$ microarray data sets with binary response as prepared by de Souza et al.

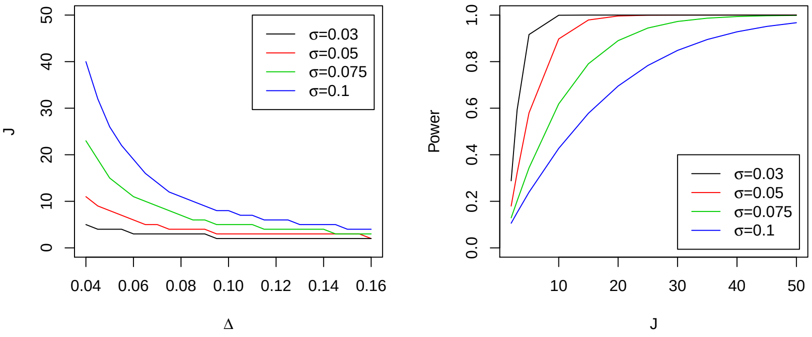

Figure 2: **Left**: Number J of data sets requested to detect Δ for different values of σ (black: $\sigma = 0.03$, red: $\sigma = 0.05$, green: $\sigma = 0.075$, blue: $\sigma = 0.1$) and power=80% and **Right**: Power to detect a difference of $\Delta = 0.05$ for different values of σ (black: $\sigma = 0.03$, red: $\sigma = 0.05$, green: $\sigma = 0.075$, blue: $\sigma = 0.1$).

(2010). These data sets and corresponding R Code are available from our companion website `http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/compstud2013`. For each data set, repeated subsampling with 4/5 of the observations in the training sets and 300 resampling iterations is used for error estimation.

The considered classification methods are: i) diagonal linear discriminant analysis (DLDA) without variable selection (DLDA-all), ii) DLDA with 500 selected variables (DLDA-500), iii) DLDA with 20 selected variables (DLDA-20), and iv) DLDA with 10 selected variables (DLDA-10). Variable selection is performed by selecting the variables yielding the smallest p-values when testing the equality of the means in the two groups $Y = 0/1$ with a classical t-test. All analyses are performed using the Bioconductor package `CMA` (Slawski et al., 2008).

Figure 3 displays the boxplots of the estimated errors over the 50 data sets using the four considered classification methods (top) and the boxplots of the six pairwise differences (bottom) also showing their respective standard deviation. It can be clearly seen from this figure that the standard deviations of the differences vary a lot depending on the considered pair of methods and that their range is the similar

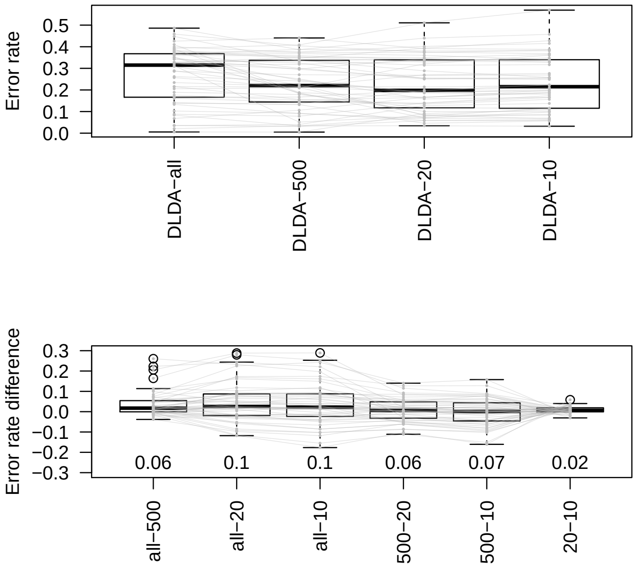

Figure 3: **Top:** Boxplots of the estimated errors over the 50 data sets using the four considered classification methods. **Bottom:** Boxplots of the six pairwise differences also showing their respective standard deviation. The gray lines connect points corresponding to the same data set.

to the values observed in the comparison studies from the literature discussed in the previous section. The results of the power considerations presented in Figure 2 are thus also relevant to our exemplary benchmark study.

		Matched-pair t-test		**Wilcoxon test**	
Comparison	Difference	t	p-value	W	p-value
DLDA-all vs. DLDA-500	0.038	4.22	5e-05	1065	2e-05
DLDA-all vs. DLDA-20	0.045	3.252	0.00104	926	0.00272
DLDA-all vs. DLDA-10	0.036	2.466	0.0086	866	0.01387
DLDA-500 vs. DLDA-20	0.007	0.875	0.19298	725	0.2005
DLDA-10 vs. DLDA-500	0.002	0.198	0.4221	622	0.46433
DLDA-10 vs. DLDA-20	0.009	3.823	0.00019	999	0.00025

Table 1: Results of the one-sided matched-pair t-test and one-sided Wilcoxon-ranked-sum test for all six pairwise comparisons of differences in the error rates over the 50 data sets when testing hypothesis $H_0^{(real)}$ vs. $H_1^{(real)}$. For each comparison, the first method plays the role of M_1 and the second method plays the role of M_2.

The results of the one-sided matched-pair t-test and one-sided Wilcoxon-signed-rank-test for the six considered pairs of methods are displayed in Table 1. Four of the six pairwise differences between error rates are statistically significant when tested at an α-level of 0.05 using the parametric matched-pair t-test and the non-parametric Wilcoxon-signed-rank-test. The largest difference of $\Delta = 0.045$ can be found between DLDA-all and DLDA-20, where the standard deviation is as large as 0.1. The results from Table outline the importance of the standard deviation of the difference: the comparisons DLDA-all vs. DLDA-500 and DLDA-all vs. DLDA-10 yield almost equal means differences (0.038 and 0.036, respectively) but the first comparison leads to a much lower p-value due to the smaller standard deviation of 0.06 (see Figure 3). Similarly, the mean difference for the comparison DLDA-10 vs. DLDA-20 is moderate (0.009) but yields small p-values due to the very small standard deviation of 0.02.

On the whole, these results again stress the importance of a sufficient number of data sets in comparison studies, especially in the context of high-dimensional data investigated in our example. Differences between methods are often moderate and their standard deviations may be comparatively large, thus making comparison studies based on the usual number of, say, 5 to 10 data sets substantially underpowered.

6. CONCLUDING REMARKS

In this paper we proposed a statistical formulation and interpretation of hypothesis tests performed in the context of comparison studies comparing the performance of supervised classification methods based on several data sets with unknown underlying distribution. Although we focused on classification problems, the developed ideas may be easily extended to other problems through the use of a different loss function.

At the light of this framework, we examined published comparison studies assessing classification methods for high-dimensional microarray data. Considering the large variance of the difference in performance across the data sets, we found that a very large number of data sets would be necessary to reach an adequate power, i.e. to have a large probability to detect relevant differences in error rates as statistically significant in the testing framework. We conclude that most published comparison studies (either neutral or part of an original article) are substantially underpowered.

As an outlook, we point to the parallels that can be made between comparison studies in the context of supervised learning and experiments from application fields of statistics (e.g. biomedicine). Roughly speaking, a comparison study comparing supervised learning methods shows some similarities with a clinical trial. For example, it might be helpful to first perform a pilot study with a limited number of data sets to provide a first raw estimate of the variance in order to evaluate the necessary number of data sets. Subgroup analyses may also make sense, e.g. focusing on data sets of particular sizes, with particular numbers of predictors, etc. The same rules should be observed as when performing subgroup analyses in biomedical sciences: relevant subgroups should be defined prior to the analysis and all the results should be reported – in order to avoid fishing for significance. Alternatively, it may make sense to model Δe as a dependent variable with data set characteristics as independent variables (for instance size, number of predictors, ratio between size and number of predictors, signal strength, balance between the class, etc). This would be an alternative to the previously suggested Bradley-Terry model approach (Eugster et al., 2010). Going one step further, meta-analyses of comparison studies would also be conceivable.

To conclude, we believe that statisticians working on methodological research projects such as the development of new supervised learning methods (including ourselves) should probably take more care of rules that they themselves (rightly) impose on their statistical consulting clients: take care of sample size and power issues, pay attention to the underlying hypothesis when performing a test, and not over-interpret results that are not statistically significant.

References

Billingsley, P., 1986. Probability and measure, 2nd Edition. John Wiley & Sons Inc., New York.

Bock, J., 1998. Bestimmung des stichprobenumfangs. München/Wien: Oldenbourg.

Bouckaert, R., 2003. Choosing between two learning algorithms based on calibrated tests. In: MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-. Vol. 20. p. 51.

Boulesteix, A., Eugster, M., 2012. A plea for neutral comparison studies in computational sciences. Technical Report: http://arxiv.org/abs/1208.2651.pdf.

Boulesteix, A., Strobl, C., Augustin, T., Daumer, M., 2008. Evaluating microarray-based classifiers: an overview. Cancer Informatics 6, 77–97.

Braga-Neto, U., Dougherty, E., 2005. Exact performance of error estimators for discrete classifiers. Pattern Recognition 38, 1799–1814.

Dalton, L., Dougherty, E., 2012a. Exact sample conditioned mse performance of the bayesian mmse estimator for classification error – part I: Representation. IEEE Transactions on Signal Processing 60, 2575–2587.

Dalton, L., Dougherty, E., 2012b. Exact sample conditioned mse performance of the bayesian mmse estimator for classification error – part ii: Consistency and performance analysis. IEEE Transactions on Signal Processing 60, 2588–2603.

de Souza, B., de Carvalho, A., Soares, C., 2010. A comprehensive comparison of ml algorithms for gene expression data classification. In: The 2010 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30.

Dietterich, T., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation 10, 1895–1923.

Dougherty, E., Zollanvari, A., Braga-Neto, U., 2011. The illusion of distribution-free small-sample classification in genomics. Current Genomics 12, 333–341.

Eugster, M., Hothorn, T., Leisch, F., 2012. Domain-based benchmark experiments: exploratory and inferential analysis. Austrian Journal of Statistics 41, 5–26.

Eugster, M., Leisch, F., Strobl, C., 2010. (Psycho-)analysis of benchmark experiments. Technical Report 78, Department of Statistics, LMU.

Garcia, S., Fernandez, A., Luengo, J., Herrera, F., 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences 180 (10), 2044–2064.

Garcia, S., Herrera, F., 2008. An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. Journal of Machine Learning Research 9, 2677–2694.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., et al., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

Hanczar, B., Dougherty, E., 2010. On the comparison of classifiers for microarray data. Current Bioinformatics 5, 29–39.

Hoeffding, W., Wolfowitz, J., 1958. Distinguishability of sets of distributions. (The case of independent and identically distributed chance variables). Annals of Mathematical Statistics 29, 700–718.

Hothorn, T., Leisch, F., Zeileis, A., Hornik, K., 2005. The design and analysis of benchmark experiments. Journal of Computational and Graphical Statistics 14, 675–699.

Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., Boulesteix, A., 2010. Over-optimism in bioinformatics: an illustration. Bioinformatics 26, 1990–1998.

Lai, C., Reinders, M., Van't Veer, L., Wessels, L., 2006. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics 7, 235.

Lee, J., Lee, J., Park, M., Song, S., 2005. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis 48, 869–885.

Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. Machine Learning 52, 239–281.

Slawski, M., Daumer, M., Boulesteix, A., 2008. CMA–a comprehensive bioconductor package for supervised classification with high dimensional data. BMC Bioinformatics 9, 439.

Statnikov, A., Aliferis, C., Tsamardinos, I., Hardin, D., Levy, S., 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21, 631–643.

Van't Veer, L., Dai, H., Van De Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., Van Der Kooy, K., Marton, M., Witteveen, A., et al., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536.

Webb, G., 2000. Multiboosting: A technique for combining boosting and wagging. Machine Learning 40, 159–196.

Wolpert, D., 2001. The supervised learning no-free-lunch theorems. In: Proceedings of the 6th Online World Conference on Soft Computing in Industrial Applications. pp. 10–24.

Yousefi, M., Hua, J., Sima, C., Dougherty, E., 2010. Reporting bias when using real data sets to analyze classification performance. Bioinformatics 26, 68–76.

Zollanvari, A., Braga-Neto, U., Dougherty, E., 2009. On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers. Pattern Recognition 42, 2705–2723.

Zollanvari, A., Braga-Neto, U., Dougherty, E., 2011. Analytic study of performance of error estimators for linear discriminant analysis. IEEE Transactions on Signal Processing 59, 4238–4255.

Appendix

The appendix contains some technicalities which are needed for a mathematically rigorous formulation of the testing problem in Section 3. In particular, it is shown that there are suitable random variables N and $\mathbf{D}$ such that $e(n, M_2, D) - e(n, M_1, D)$ can be seen as the observed realization of the random variable $e(N, M_2, \mathbf{D}) - e(N, M_1, \mathbf{D})$ where also the distribution P (which generates the data set D) is randomly chosen. This is crucial in order to apply the central limit theorem and, otherwise, using the t-rest to test $H_0^{(real)}$ in Section 3 would not be justified.

Preliminaries

Let $\big((\mathcal{X} \times \mathcal{Y})^\infty, \mathcal{B}^\infty\big)$ be the countably-infinite-product space of the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B})$. Let $\mathcal{V}$ be the set of all possible distributions (in the area of application) endowed with the Borel-σ-algebra $\mathfrak{B}_\mathcal{V}$ with respect to the total variation norm.

We arbitrarily fix any $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}$ and define the function

$$\gamma \,:\, (\mathcal{X} \times \mathcal{Y})^\infty \times \mathbb{N} \to (\mathcal{X} \times \mathcal{Y})^\infty \,, \quad (D, n) \;\mapsto\; \gamma(D, n) \;=:\; D^{(n)}$$

by

$$D^{(n)} \;=\; \big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), (x_0, y_0), (x_0, y_0), \ldots\big) \in (\mathcal{X} \times \mathcal{Y})^\infty$$

for $D = \big((x_1, y_1), (x_2, y_2), \ldots\big) \in (\mathcal{X} \times \mathcal{Y})^\infty$ and $n \in \mathbb{N}$. Note that γ is measurable with respect to $\mathcal{B}^\infty \otimes 2^\mathbb{N}$ and $\mathcal{B}^\infty$. Similarly, define

$$P^{(n)} \;=\; P^n \otimes \delta^\infty_{(x_0, y_0)}$$

on $\big((\mathcal{X} \times \mathcal{Y})^\infty, \mathcal{B}^\infty\big)$ for every $P \in \mathcal{V}$.

Lemma *The mapping*

$$\tau \;:\; \big(\mathcal{V}\times\mathbb{N}\big)\times\mathcal{B}^{\infty} \;\to\; [0,1]\,, \qquad \big((P,n),B\big) \;\mapsto\; \tau_{P,n}(B) \;=\; P^{(n)}(B)$$

is a Markov kernel from $\big(\mathcal{V}\times\mathbb{N},\mathfrak{B}_{\mathcal{V}}\otimes 2^{\mathbb{N}}\big)$ *to* $\big((\mathcal{X}\times\mathcal{Y})^{\infty},\mathcal{B}^{\infty}\big)$*.*

Proof: Since $\tau_{P,n} = P^{(n)}$ is a probability measure on $\big((\mathcal{X}\times\mathcal{Y})^{\infty},\mathcal{B}^{\infty}\big)$, it remains to prove that $(P,n)\mapsto P^{(n)}(B)$ is measurable for every $B\in\mathcal{B}^{\infty}$. Since $\mathbb{N}$ is countable, it is enough to show that $P\mapsto P^{(n)}(B)$ is measurable for every $n\in\mathbb{N}$ and $B\in\mathcal{B}^{\infty}$. To this end, it is shown in the following that the mapping $P\mapsto P^{(n)}(B)$ is even continuous (with respect to the total variation norm): Let $(P_k)_{k\in\mathbb{N}_0}\subset\mathcal{V}$ such that

$$\lim_{k\to\infty}\big\|P_k-P_0\big\|_{\mathrm{TV}} \;=\; 0\,. \tag{5}$$

According to (Hoeffding and Wolfowitz, 1958, Assertions (4.4) and (4.5)), the product measures P_k^n and P_0^n on $(\mathcal{X}\times\mathcal{Y})^n$ fulfill

$$\big\|P_k^n-P_0^n\big\|_{\mathrm{TV}} \;\leq\; n\cdot\big\|P_k-P_0\big\|_{\mathrm{TV}}\,. \tag{6}$$

For $D=\big((x_1,y_1),(x_2,y_2),\dots\big)\in(\mathcal{X}\times\mathcal{Y})^{\infty}$, put $D_n=\big((x_1,y_1),\dots,(x_n,y_n)\big)$ and $D_\infty=\big((x_{n+1},y_{n+1}),(x_{n+2},y_{n+2}),\dots\big)$; that is, $D=(D_n,D_\infty)$. In addition, define

$g_{D_\infty}(D_n) := I_B(D_n, D_\infty) = I_B(D)$ for every $D = (D_n, D_\infty)$. Then,

$$
\begin{aligned}
\Big| P_k^{(n)}(B) - P_0^{(n)}(B) \Big| \;&= \\
&= \left| \iint I_B(D_n, D_\infty)\, P_k^n\big(dD_n\big)\, \delta^\infty_{(x_0,y_0)}(dD_\infty) - \right. \\
&\qquad \left. - \iint I_B(D_n, D_\infty)\, P_0^n\big(dD_n\big)\, \delta^\infty_{(x_0,y_0)}(dD_\infty) \right| \\
&\leq \int \left| \int g_{D_\infty}\, dP_k^n - \int g_{D_\infty}\, dP_0^n \right| \delta^\infty_{(x_0,y_0)}(dD_\infty) \;\leq \\
&\leq \int \big\| P_k^n - P_0^n \big\|_{\mathrm{TV}}\, \delta^\infty_{(x_0,y_0)}(dD_\infty) \;\overset{(6)}{\leq}\; n \cdot \big\| P_k - P_0 \big\|_{\mathrm{TV}}.
\end{aligned}
$$

Therefore, $\lim_{k\to\infty} P_k^{(n)}(B) = P_0^{(n)}(B)$ follows from (6). □

Now, we can prove existence of suitable random variables $\mathbf{D}$, Φ, and N such that the conditional distribution of $\gamma(\mathbf{D}, N) = \mathbf{D}^{(N)}$ given $\Phi = P$ and $N = n$ is equal to $P^{(n)}$.

Theorem *Let $Q_{\Phi,N}$ be any distribution on $\big(\mathcal{V} \times \mathbb{N}, \mathfrak{B}_{\mathcal{V}} \otimes 2^{\mathbb{N}}\big)$. Then, there are a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and random variables*

$$
\mathbf{D} \;=\; \big((X_1, Y_1), (X_2, Y_2), \ldots\big) \;:\; (\Omega, \mathcal{A}) \;\longrightarrow\; \big((\mathcal{X} \times \mathcal{Y})^\infty, \mathcal{B}^\infty\big),
$$

$$
\Phi \;:\; (\Omega, \mathcal{A}) \;\longrightarrow\; \big(\mathcal{V}, \mathfrak{B}_{\mathcal{V}}\big), \qquad \textit{and} \qquad N \;:\; (\Omega, \mathcal{A}) \;\longrightarrow\; \big(\mathbb{N}, 2^{\mathbb{N}}\big)
$$

such that the joint distribution of Φ and N is equal to $Q_{\Phi,N}$ and the conditional distribution of $\mathbf{D}^{(N)} = \gamma(\mathbf{D}, N)$ given $\Phi = P$ and $N = n$ is equal to $P^{(n)}$, i.e.,

$$
\mathscr{L}\Big(\gamma(\mathbf{D}, N) \,\Big|\, (\Phi, N) = (P, n)\Big) \;=\; P^{(n)}\,. \tag{7}
$$

Proof: According to the above lemma, τ is a Markov kernel so that we can define a probability measure Q on $\big((\mathcal{X}\times\mathcal{Y})^\infty \times \mathcal{V} \times \mathbb{N}\,,\,\mathcal{B}^\infty \otimes \mathfrak{B}_\mathcal{V} \otimes 2^\mathbb{N}\big)$ via

$$Q(C) \;=\; \iint I_C(D,P,n)\,P^{(n)}(dD)\,Q_{\Phi,N}\big(d(P,n)\big) \qquad \forall\, C \in \mathcal{B}^\infty {\otimes} \mathfrak{B}_\mathcal{V} \otimes 2^\mathbb{N}.$$

Then, there are a probability space $(\Omega,\mathcal{A},\mathbb{P})$ and random variables

$$\mathbf{D} \;=\; \big((X_1,Y_1),(X_2,Y_2),\dots\big) \;:\; (\Omega,\mathcal{A}) \;\longrightarrow\; \big((\mathcal{X}\times\mathcal{Y})^\infty,\mathcal{B}^\infty\big),$$

$$\Phi \;:\; (\Omega,\mathcal{A}) \;\longrightarrow\; \big(\mathcal{V},\mathfrak{B}_\mathcal{V}\big), \qquad \text{and} \qquad N \;:\; (\Omega,\mathcal{A}) \;\longrightarrow\; \big(\mathbb{N},2^\mathbb{N}\big)$$

such that the joint distribution of $\mathbf{D}$, Φ, and N is equal to Q; in particular, this means that $Q_{\Phi,N}$ is the joint distribution of Φ and N. In order to show (7), fix any $B \in \mathcal{B}^\infty$ and $C \in \mathfrak{B}_\mathcal{M} \otimes 2^\mathbb{N}$. Then, it follows from the definition of Q that

$$\begin{aligned} \mathbb{P}\big(\gamma(\mathbf{D},N)\in B,\ (\Phi,N)\in C\big) \;&=\; \int I_B\big(\gamma(\mathbf{D},N)\big) I_C(\Phi,N)\,d\mathbb{P} \;= \\ =\; &\int_C \int I_B\big(\gamma(D,n)\big)\,P^{(n)}\big(dD\big)\;Q_{\Phi,N}\big(d(P,n)\big) \;= \\ \overset{(*)}{=}\; &\int_C \int I_B(D)\,P^{(n)}\big(dD\big)\,Q_{\Phi,N}\big(d(P,n)\big) \;=\; \int_C P^{(n)}(B)\,Q_{\Phi,N}\big(d(P,n)\big) \end{aligned}$$

where $(*)$ follows from the definition of γ and $P^{(n)}$. □

Exact formulation of the testing problem

Again, let $e(n,M_k,D_n)$ denote the estimated error of method M_k by use of the data set $D_n \in (\mathcal{X}\times\mathcal{Y})^n$. For $D=(D_n,D_\infty)\in(\mathcal{X}\times\mathcal{Y})^\infty$, we also write

$$e\big(n,M_k,D_n\big) \;=\; e\big(n,M_k,D\big) \;=\; e\big(n,M_k,\gamma(D,n)\big) \tag{8}$$

but note that $e\big(n,M_k,D\big)$ and $e\big(n,M_k,\gamma(D,n)\big)$ only depend on D_n, that is, the first n data points (x_i,y_i) of D. As in the previous subsection, let $\mathbf{D}$ be a random variable such that the conditional distribution of $\mathbf{D}^{(N)}=\gamma(\mathbf{D},N)$ given $\Phi=P$ and

$N = n$ is equal to $P^{(n)}$. Then, $e\big(n, M_k, D\Big) = e\big(n, M_k, D_n\big)$ can be seen as the observed realization of the random variable $e\big(N, M_k, \mathbf{D}\big)$.

Recall the assumption

$$\boldsymbol{E}_{P^n}\Big[e\big(n, M_1, D_n\big)\Big] - \varepsilon(n, M_1, P) = \boldsymbol{E}_{P^n}\Big[e\big(n, M_2, D_n\big)\Big] - \varepsilon(n, M_2, P),$$

from Section 3 where $\boldsymbol{E}_{P^n}$ denotes the expectation over the data set D_n drawn i.i.d. from P. Then, it follows that

$$\begin{aligned}
\varepsilon(n, M_2, P) - \varepsilon(n, M_1, P) \; &= \; \boldsymbol{E}_{P^n}\Big[e\big(n, M_2, D_n\big) - e\big(n, M_1, D_n\big)\Big] \; = \\
&= \; \int e\big(n, M_2, D_n\big) - e\big(n, M_1, D_n\big)\, P^n(dD_n) \; = \\
&\overset{(8)}{=} \; \int e\big(n, M_2, D\big) - e\big(n, M_1, D\big)\, P^{(n)}\big(dD\big) \; = \\
&\overset{(7)}{=} \; \boldsymbol{E}\Big[e\big(n, M_2, \gamma(\mathbf{D}, N)\big) - e\big(n, M_1, \gamma(\mathbf{D}, N)\big)\,\Big|\,(\Phi, N) = (P, n)\Big] \; = \\
&\overset{(8)}{=} \; \boldsymbol{E}\Big[e\big(N, M_2, \mathbf{D}\big) - e\big(N, M_1, \mathbf{D}\big)\,\Big|\,(\Phi, N) = (P, n)\Big] \, . \qquad (9)
\end{aligned}$$

Hence,

$$\boldsymbol{E}\big[\varepsilon(N, M_2, \Phi) - \varepsilon(N, M_1, \Phi)\big] \; = \; \boldsymbol{E}\Big[e\big(N, M_2, \mathbf{D}\big) - e\big(N, M_1, \mathbf{D}\big)\Big] \, .$$

In order to support the claim that M_2 is better than M_1, we may therefore consider the testing problem

$$\begin{aligned}
H_0^{(\text{real})} \; &: \; \boldsymbol{E}e\big(N, M_2, \mathbf{D}\big) \; - \; \boldsymbol{E}e\big(N, M_1, \mathbf{D}\big) \; \geq \; 0 \\
\text{vs. } H_1^{(\text{real})} \; &: \; \boldsymbol{E}e\big(N, M_2, \mathbf{D}\big) \; - \; \boldsymbol{E}e\big(N, M_1, \mathbf{D}\big) \; < \; 0 \, .
\end{aligned}$$

This testing problem is feasible because we observe i.i.d. realizations of $e\big(N, M_2, \mathbf{D}\big) - e\big(N, M_1, \mathbf{D}\big)$.

Now, we are also able to rigorously show the variance decomposition (4): As in

Section 3, define $\Delta e(N, \mathbf{D}) = e(N, M_2, \mathbf{D}) - e(N, M_1, \mathbf{D})$. Then, (4) follows from

$$\begin{aligned} \mathrm{Var}(\Delta e(N, \mathbf{D})) &\overset{(*)}{=} \mathrm{Var}\Big(\boldsymbol{E}\big(\Delta e(N, \mathbf{D})\big|\Phi,\, N\big)\Big) + \boldsymbol{E}\Big(\mathrm{Var}\big(\Delta e(N, \mathbf{D})\big|\Phi,\, N\big)\Big) \ = \\ &\overset{(9)}{=} \ \mathrm{Var}\big(\varepsilon(N, M_2, \Phi) - \varepsilon(N, M_1, \Phi)\big) + \boldsymbol{E}\big(\mathrm{Var}(\Delta e(N, \mathbf{D})|\Phi,\, N)\big) \end{aligned}$$

where $(*)$ is the well-known law of total variance, see e.g. (Billingsley, 1986, Problem 34.10 b).