Mach Learn (2015) 100:285–304
DOI 10.1007/s10994-015-5486-z

A Bayesian approach for comparing cross-validated algorithms on multiple data sets

Giorgio Corani[1] · **Alessio Benavoli**[1]

Received: 14 November 2014 / Accepted: 3 March 2015 / Published online: 24 March 2015

Abstract We present a Bayesian approach for making statistical inference about the accuracy (or any other score) of two competing algorithms which have been assessed via cross-validation on multiple data sets. The approach is constituted by two pieces. The first is a novel *correlated* Bayesian t test for the analysis of the cross-validation results on a single data set which accounts for the correlation due to the overlapping training sets. The second piece merges the posterior probabilities computed by the Bayesian correlated t test on the different data sets to make inference on multiple data sets. It does so by adopting a Poisson-binomial model. The inferences on multiple data sets account for the different uncertainty of the cross-validation results on the different data sets. It is the first test able to achieve this goal. It is generally more powerful than the signed-rank test if ten runs of cross-validation are performed, as it is anyway generally recommended.

Keywords Bayesian hypothesis tests · Signed-rank test · Cross-validation · Poisson-binomial · Hypothesis test · Evaluation of classifiers

1 Introduction

A typical problem in machine learning is to compare the accuracy of two competing classifiers on a data set D. Usually one measures the accuracy of both classifiers via k-folds cross-validation. After having performed cross-validation, one has to decide if the accuracy of the two classifiers on data set D is significantly different. The decision is made using a statistical

Editors: João Gama, Indrė Žliobaitė, Alípio M. Jorge, and Concha Bielza.

✉ Giorgio Corani
giorgio@idsia.ch

Alessio Benavoli
alessio@idsia.ch

[1] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Scuola Universitaria Professionale della Svizzera Italiana (SUPSI), Università della Svizzera Italiana (USI), Manno, Switzerland

hypothesis test which analyzes the measures of accuracy yielded by cross-validation on the different folds. Using a t test is however a naive choice. The t test assumes the measures of accuracy taken on the different folds to be independent. Such measures are instead correlated because of the overlap of the training sets built during cross-validation. As a result the t test is *not* calibrated, namely its rate of Type I errors is much larger than the nominal size[1] α of the test. Thus the t test is not suitable for analyzing the cross-validation results (Dietterich 1998; Nadeau and Bengio 2003).

A suitable approach is instead the correlated[2] t test (Nadeau and Bengio 2003), which adjusts the t test accounting for correlation. The statistic of the correlated t test is composed by two pieces of information: the mean difference of accuracy between the two classifiers (computed averaging over the different folds) and the uncertainty of such estimate, known as the *standard error*. The standard error of the correlated t test accounts for correlation, differently from the t test. The correlated t test is the recommended approach for the analysis of cross-validation results on a single data set (Nadeau and Bengio 2003; Bouckaert 2003).

Assume now that the two classifiers have assessed via cross-validation on a collection of data sets $\boldsymbol{D} = \{D_1, D_2, \ldots, D_q\}$. One has to decide if the difference of accuracy between the two classifiers on the multiple data sets of $\boldsymbol{D}$ is significant. The recommended approach is the signed-rank test (Demšar 2006). It is a non-parametric test. As such it is derived under mild assumptions and is robust to outliers. A Bayesian counterpart of the signed-rank test (Benavoli et al. 2014) has been also recently proposed. However the signed-rank test considers only the mean difference of accuracy measured on each data set, ignoring the associated uncertainty.

Dietterich (1998) pointed out the need for a test able to compare two classifier on multiple data sets accounting for the uncertainty of the results on each data set. Tests dealing with this issue have been devised only recently. Otero et al. (2014) proposes an interval-valued approach to considers the uncertainty of the cross-validation results on each data set. When working with multiple data sets, the interval uncertainty is propagated. In some cases the interval becomes wide, preventing to achieve a conclusion.

The Poisson-binomial test (Lacoste et al. 2012) performs inference on multiple data sets accounting for the uncertainty of the result on each data set. First it computes on each data set the posterior probability of the difference of accuracy being significant; then it merges such probabilities through a Poisson-binomial distribution to make inference on $\boldsymbol{D}$. Its limit is that the posterior probabilities computed on the individual data sets assume that the two classifiers have been compared on a *single* test set. It does not manage the multiple correlated test sets produced by cross-validation. This limits its applicability, since classifiers are typically assessed by cross-validation.

To design a test able to perform inference on multiple data sets accounting for the uncertainty of the estimates yielded by cross-validation is a challenging task.

In this paper we solve this problem. Our solution is based on two main steps. First we develop a Bayesian counterpart of the correlated t test (its posterior probabilities are later exploited to build a Poisson-binomial distribution). We design a generative model for the correlated results of cross-validation and we analytically derive the posterior distribution of the mean difference of accuracy between the two classifiers. Moreover, we show that for a particular choice of the prior over the parameters, the posterior distribution coincides with the sampling distribution of the correlated t test by Nadeau and Bengio (2003). Under the

[1] Consider performing many experiments in which the data are generated under the null hypothesis. A test executed with size α is correctly calibrated if its rate of rejection of the null hypothesis is not $> \alpha$.

[2] Nadeau and Bengio (2003) refer to this test as the *corrected* t test. We adopt in this paper the more informative terminology of *correlated* t test.

matching prior the inferences of the Bayesian correlated t test and of the frequentist correlated t test are numerically equivalent. The meaning of the inferences is however different. The inference of the frequentist test is a p value; the inference of the Bayesian test is a posterior probability. The posterior probabilities computed on the individual data sets can be combined to make further Bayesian inference on multiple data sets.

After having computed the posterior probabilities on each individual data set through the correlated Bayesian t test, we merge them to make inference on $\boldsymbol{D}$, borrowing the intuition of the Poisson-binomial test (Lacoste et al. 2012). This is the second piece of the solution. We model each data set as a Bernoulli trial, whose possible outcomes are the win of the first or the second classifier. The probability of success of the Bernoulli trial corresponds to the posterior probability computed by the Bayesian correlated t test on that data set. The number of data sets on which the first classifier is more accurate than the second is a random variable which follows a Poisson-binomial distribution. We use this distribution to make inference about the difference of accuracy of the two classifiers on $\boldsymbol{D}$. The resulting approach couples the Bayesian correlated t test and the Poisson-binomial approach; we call it the *Poisson test*.

It is worth discussing an important difference between the signed-rank and the Poisson test. The signed rank test assumes the results on the individual data sets to be i.i.d. The Poisson test assumes them to be independent but *not* identically distributed, which can be advocated as follows. The different data sets $D_1, \ldots, D_q$ have different size and complexity. The uncertainty of the cross-validation result is thus different on each data set, breaking the assumption of the results on different data sets to be identically distributed.

We compare the Poisson and the signed-rank test through extensive simulations, performing either one run or ten runs of cross-validation. When we perform one run of cross-validation, the estimates are affected by important uncertainty. In this case the Poisson behaves cautiously and it is less powerful than the signed-rank test. When we perform ten runs of cross-validation, the uncertainty of the cross-validation estimate decreases. In this case the Poisson test is generally *more* powerful than the signed-rank test. To perform ten runs rather than a single one run of cross-validation is anyway recommended to obtain robust cross-validation estimates (Bouckaert 2003). The signed-rank test does not account for the uncertainty of the estimates and thus its power is roughly the same whether one or ten runs of cross-validation are performed.

Under the null hypothesis, the Type I errors of both test are correctly calibrated in all the investigated settings.

The paper is organized as follows: Sect. 2 presents the methods for inference on a single data set; Sect. 3 presents the methods for inference on multiple data set; Sect. 4 presents the experimental results.

2 Inference from cross-validation results on a single data set

2.1 Problem statement and frequentist tests

We want to statistically compare the accuracy of two classifiers which have been assessed via m runs of k-folds cross-validation. We provide both classifiers with the same training and test sets and we compute the difference of accuracy between the two classifiers on each test set. This yields the *differences of accuracy* $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, where $n = mk$. We denote the sample mean and the sample variance of the differences as $\overline{x}$ and $\hat{\sigma}^2$.

A statistical test has to establish whether the mean difference between the two classifier is significantly different from zero, analyzing the vector of results $\boldsymbol{x}$. Such results are correlated

because of the overlapping training sets. Nadeau and Bengio (2003) prove that there is no unbiased estimator of such correlation. They assume the correlation to be $\rho = \frac{n_{te}}{n}$, where n_{te}, n_{tr} and n_{tot} denote the size of the training set, of the test set and of the whole available data set. Thus $n_{tot} = n_{tr} + n_{te}$. The statistic of the correlated t test is:

$$t = \frac{\overline{x}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\rho}{1-\rho}\right)}} = \frac{\overline{x}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{n_{te}}{n_{tr}}\right)}}. \tag{1}$$

Its sampling distribution is a Student with $n-1$ degrees of freedom. The correlation heuristic has proven to be effective and the correlated t test is much closer to a correct calibration than the standard t test (Nadeau and Bengio 2003). The correlation heuristic of Nadeau and Bengio (2003) is derived assuming *random selection* of the instances which compose the different training and test sets used in cross-validation. Under random selection the different test sets overlap. The standard cross-validation yields non-overlapping test sets. This is also the setup we consider in this paper. The correlation heuristic of Nadeau and Bengio (2003) is anyway effective also with the standard cross-validation (Bouckaert 2003).

The denominator of the statistics is the *standard error*, namely the standard deviation of the estimate of $\overline{x}$. The standard error increases with $\hat{\sigma}^2$, which typically increases on smaller data sets. On the other hand the standard error decreases with $n = mk$. Previous studies (Kohavi 1995) recommend to set the number of folds to $k = 10$ to obtain a reliable estimate from cross-validation. This has become a standard choice. Having set $k = 10$, one can further decrease the standard error of the test by increasing the number or runs m. Indeed Bouckaert (2003) and (Witten et al. 2011, Sec. 5.3) recommend to perform $m = 10$ runs of ten-folds cross-validation.

The correlated t test has been originally designed to analyze the results of a single run of cross-validation. Indeed its correlation heuristic models the correlation due to overlapping training sets. When multiple runs of cross-validation are performed, there is an additional correlation due to overlapping test sets. We are unaware of approaches able to represent also this second type of correlation, which is usually ignored.

2.2 Bayesian t test for uncorrelated observations

Before introducing the Bayesian t test for correlated observations, we briefly discuss the Bayesian inference in the uncorrelated case. Assume we have a vector of independent and identically distributed observations of a variable X, i.e., $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, and that we aim to test if the mean of X is positive. In the Bayesian t test we assume that the likelihood of the observations is Normal with unknown mean μ and and unknown precision ν (the precision is the inverse of variance $\nu = 1/\sigma^2$):

$$p(\boldsymbol{x}|\mu, \nu) = \prod_{i=1}^{n} N(x_i; \mu, 1/\nu). \tag{2}$$

Our aim is to compute the posterior of μ (here ν is a nuisance parameter). A natural prior for μ, ν is the Normal-Gamma distribution (Bernardo and Smith 2009, Chap. 5), which is conjugate with the likelihood model:

$$p(\mu, \nu|\mu_0, k_0, a, b) = N\left(\mu; \mu_0, \frac{k_0}{\nu}\right) G(\nu; a, b) = NG(\mu, \nu; \mu_0, k_0, a, b).$$

It is the product of a Normal distribution over μ (with precision ν/k_0 proportional to ν) and a Gamma distribution over ν and depends on four parameters μ_0, k_0, a, b. Updating

Table 1 Posterior parameters for the uncorrelated case

Parameter	Analytical expression	Under matching prior
μ_n	$\frac{\mu_0/k_0+n\overline{x}}{\frac{1}{k_0}+n}$	$\overline{x}$
k_n	$\frac{1}{\frac{1}{k_0}+n}$	$\frac{1}{n}$
a_n	$a+\frac{n}{2}$	$\frac{n-1}{2}$
b_n	$b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{\frac{1}{k_0}n(\overline{x}-\mu_0)^2}{2(\frac{1}{k_0}+n)}$	$\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2$

the prior-normal gamma with the normal likelihood, one obtains a posterior normal-gamma joint distribution with updated parameters (μ_n, k_n, a_n, b_n), whose values are reported in first column of Table 1 (see also Murphy 2012, Chap. 4). Marginalizing out the precision from the Normal-Gamma posterior one obtains the posterior marginal distribution of the mean, which follows a Student distribution:

$$p(\mu|\boldsymbol{x}, \mu_0, k_0, a, b) = \mathrm{St}\left(\mu; 2a_n, \mu_n, \frac{b_n k_n}{a_n}\right).$$

Then, the Bayesian t test for the positiveness of μ is:

$$P(\mu > 0|\boldsymbol{x}, \mu_0, k_0, a, b) = \int_0^{\infty} \mathrm{St}\left(\mu; 2a_n, \mu_n, \frac{b_n k_n}{a_n}\right) d\mu = \mathcal{T}_{2a_n}\left(\frac{\mu_n}{\sqrt{\frac{b_n k_n}{a_n}}}\right) > 1-\alpha, \tag{3}$$

where $\mathcal{T}_{2a_n}(z)$ denotes the cumulative distribution of the standardized Student distribution with $2a_n$ degrees of freedom computed at z. By choosing $\alpha = 0.05$, we can assess the positivity of μ with posterior probability 0.95. If the prior parameters are set as follows: $\{\mu_0 = 0, k_0 \to \infty, a = -1/2, b = 0\}$, from Eqn.(3) it follows that $P(\mu > 0|\boldsymbol{x}, \mu_0, k_0, a, b) = 1 - p$, where p is the p value of the frequentist t test. See Murphy 2012, Chap. 4 for further details on the correspondence between frequentist and Bayesian t tests. In fact, for these values, the posterior reduces to $\mathrm{St}\left(\mu; n-1, \overline{x}, \sigma^2/n\right)$, as shown also in the second column in Table 1. Therefore, if we consider this matching (improper) prior, the Bayesian and frequentist t test coincide.

2.3 A novel Bayesian t test for correlated observations

Assume now that the observations of the variable X, $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, are identically distributed but dependent. In particular, consider the case in which the observations have the same mean μ, the same precision ν and are equally correlated with each other with correlation $\rho > 0$. This is for instance the case in which the n observations are the n differences of accuracy among two classifiers yielded by cross-validation. The data generating process can be modelled as follows:

$$\mathbf{x} = \mathbf{H}\mu + \mathbf{v} \tag{4}$$

where $\mathbf{H}_{n\times 1}$ is a vector of ones ($\mathbf{H}_{n\times 1} = \mathbf{1}_{n\times 1}$) and $\mathbf{v}$ is a noise vector with zero mean and covariance matrix $\mathbf{\Sigma}_{n\times n}$ patterned as follows: each diagonal elements equals $\sigma^2 = 1/\nu$; each non-diagonal element equals $\rho\sigma^2$. This is the so-called *intraclass covariance matrix* (Press

2012). We define $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{M}$, where $\boldsymbol{M}$ is the (n × n) correlation matrix. As an example, with $n = 3$ we have:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix} \qquad \boldsymbol{M} = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \tag{5}$$

To allow for $\boldsymbol{\Sigma}$ to be invertible and positive definite, we require $\sigma^2 > 0$ and $0 \leq \rho < 1$. The correlation among the cross-validation results is positive anyway (Nadeau and Bengio 2003). These two conditions define the *admissibility region* of the parameters.

In the Bayesian t test for correlated observations, we assume the noise vector $\mathbf{v}$ to be follow a multivariate Normal distribution: $\mathbf{v} \sim \text{MVN}(0, \boldsymbol{\Sigma})$. The likelihood corresponding to (4) is:

$$p\left(\mathbf{x}|\mu, \boldsymbol{\Sigma}\right) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{H}\mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{H}\mu)\right)}{(2\pi)^{n/2}\sqrt{|\boldsymbol{\Sigma}|}}. \tag{6}$$

Equation (6) reduces to Eq. (2) in the uncorrelated case ($\rho = 0$). As in the previous section, our aim is to test the positivity of μ. To this end, we need to estimate the model parameters: μ, σ^2 and ρ.

Theorem 1 *The maximum likelihood estimator of (μ,σ^2,ρ) from the model (6) is not asymptotically consistent: it does not converge to the true value of the parameters as $n \to \infty$.*

The proof is given in "Appendix". By computing the derivatives of the likelihood w.r.t. the parameters, it shows that the maximum likelihood estimate of μ, σ^2 is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and, respectively, $\hat{\sigma}^2 = tr(\boldsymbol{M}^{-1}\mathbf{Z})$, where $\mathbf{Z} = (\mathbf{x} - \mathbf{H}\hat{\mu})(\mathbf{x} - \mathbf{H}\hat{\mu})^T$. Thus $\hat{\sigma}^2$ depends on ρ through M. By plugging these estimates into the likelihood and computing the derivative w.r.t. ρ, we show that the derivative is never zero in the admissibility region. The derivative decreases with ρ and does not depend on the data. Hence, the maximum likelihood estimate of ρ is $\hat{\rho} = 0$ regardless the observations. When the number of observations n increases, the likelihood gets more concentrated around the maximum likelihood estimate. Thus the maximum likelihood estimate is not asymptotically consistent whenever $\rho \neq 0$. This will also be true for the Bayesian estimate, since the likelihood dominates the conjugate prior for large n. This means that we cannot consistently estimate all the three parameters (μ, σ^2, ρ) from data.

2.4 Introducing the correlation heuristic

To enable inferences from correlated samples we renounce estimating ρ from data. We adopt instead the correlation heuristic of (Nadeau and Bengio 2003), setting $\rho = \frac{n_{te}}{n_{tot}}$, where n_{te} and n_{tot} are the size of test set and of the entire data set. Having fixed the value of ρ, we can derive the posterior marginal distribution of μ.

Theorem 2 *Choose $p(\mu, \nu|\mu_0, k_0, a, b) = NG(\mu, \nu; \mu_0, k_0, a, b)$ as joint prior over μ, ν. Update it with the likelihood of Eq. (6). The posterior distribution of the parameters is $p(\mu, \nu|\mathbf{x}, \mu_0, k_0, a, b, \rho) = NG(\mu, \nu; \tilde{\mu}_n, \tilde{k}_n, \tilde{a}_n, \tilde{b}_n)$ and the posterior marginal over μ is a Student distribution:*

$$p(\mu|\mathbf{x}, \mu_0, k_0, a, b, \rho) = St\left(\mu; 2\tilde{a}_n, \tilde{\mu}_n, \frac{\tilde{b}_n \tilde{k}_n}{\tilde{a}_n}\right). \tag{7}$$

The expression of the parameters and their values are reported in Table 2.

Table 2 Posterior parameters for the correlated case

Parameter	Analytical expression	Under matching prior
$\tilde{\mu}_n$	$\frac{\mathbf{H}^T \boldsymbol{M}^{-1}\mathbf{x} + \frac{\mu_0}{k_0}}{\mathbf{H}^T \boldsymbol{M}^{-1}\mathbf{H} + \frac{1}{k_0}}$	$\frac{\sum_{i=1}^n x_i}{n}$
$\tilde{k}_n$	$\frac{1}{\mathbf{H}^T \boldsymbol{M}^{-1}\mathbf{H} + \frac{1}{k_0}}$	$\frac{1}{\mathbf{H}^T \boldsymbol{M}^{-1}\mathbf{H}}$
$\tilde{a}_n$	$a + \frac{n}{2}$	$\frac{n-1}{2}$
$\tilde{b}_n$	$\frac{1}{2}\Big((\mathbf{x} - \mathbf{H}\hat{\mu})^T \boldsymbol{M}^{-1}(\mathbf{x} - \mathbf{H}\hat{\mu}) + 2b - \frac{\mu_0^2}{k_0}$ $-\hat{\mu}^2 \mathbf{H}^T \boldsymbol{M}^{-1}\mathbf{H} + \tilde{\mu}^2 \left(\mathbf{H}^T \boldsymbol{M}^{-1}\mathbf{H} + \frac{1}{k_0}\right)\Big)$	$\frac{1}{2}(\mathbf{x} - \mathbf{H}\hat{\mu})^T \boldsymbol{M}^{-1}(\mathbf{x} - \mathbf{H}\hat{\mu})$

Corollary 1 *Under the matching prior ($\mu_0 = 0, k_0 \to \infty, a = -1/2, b = 0$), the posterior marginal distribution of μ simplifies as:*

$$St\left(\mu; n-1, \bar{x}, \left(\frac{1}{n} + \frac{\rho}{1-\rho}\right)\hat{\sigma}^2\right) \tag{8}$$

where $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ *and* $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ *and, therefore,*

$$P[\mu > 0 | \boldsymbol{x}, \mu_0, k_0, a, b, \rho] = \mathcal{T}_{n-1}\left(\frac{\bar{x}}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\rho}{1-\rho}}}\right) \tag{9}$$

The proof of both the theorem and corollary are given in "Appendix". Under the matching prior the posterior Student distribution (9) coincides with the sampling distribution of the statistic of the correlated t test by Nadeau and Bengio (2003). This implies that given the same test size α, the Bayesian correlated t test and the frequentist correlated t test take the same decisions. In other words, the posterior probability $P(\mu > 0 | \boldsymbol{x}, \mu_0, k_0, a, b, \rho)$ equals $1 - p$ where p is the p value of the correlated t test.

3 Inference on multiple data sets

Consider now the problem of comparing two classifiers on q different data sets, after having assessed both classifiers via cross-validation on each data set. The mean *difference* of accuracy on each data set are stored in vector $\overline{\boldsymbol{x}} = \{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_q\}$. The recommended test to compare two classifiers on multiple data sets is the signed-rank test (Demšar 2006).

The signed-rank test assumes the $\overline{x}_i$'s to be i.i.d. and generated from a symmetric distribution. The null hypothesis is that the median of the distribution is M. When the test accept the alternative hypothesis it claims that the median of the distribution is significantly different from M.

The test ranks the $\overline{x}_i$'s according to their absolute value and then compares the ranks of the positive and negative differences. The test statistic is:

$$T^+ = \sum_{\{i:\, \overline{x}_i \geq 0\}} r_i(|\overline{x}_i|) = \sum_{1 \leq i \leq j \leq n} T_{ij}^+,$$

where $r_i(|\overline{x}_i|)$ is the rank of $|\overline{x}_i|$ and

$$T_{ij}^{+} = \begin{cases} 1 & \text{if } \overline{x}_i \geq \overline{x}_j, \\ 0 & \text{otherwise.} \end{cases}$$

For a large enough number of samples (e.g., $q > 10$), the sampling distribution of the statistic under the null hypothesis is approximately normal with mean $1/2$. Being non-parametric, the signed-rank test does *not* average the results across data sets. This is a sensible approach since the average of results referring to different domains is in general meaningless. The test is moreover robust to outliers.

A limit of the signed-rank test is that does not consider the standard error of the $\overline{x}_i$'s. It assumes the samples to be i.i.d and thus all the $\overline{x}_i$'s to have equal uncertainty. This is a questionable assumptions. The data sets typically have different size and complexity. Moreover one could have performed a different number of cross-validation runs on different data sets. For these reasons the $\overline{x}_i$'s typically have different uncertainties; thus they are *not* identically distributed.

3.1 Poisson-binomial inference on multiple data sets

Our approach to make inference on multiple data sets is inspired to the Poisson-binomial test (Lacoste et al. 2012). As a preliminary step we perform cross-validation on each data set and we analyze the results through the Bayesian correlated t test. We denote by p_i the posterior probability that the second classifier is more accurate than the first on the ith data set. This is computed according to Eq.(9): $p_i = p(\mu_i > 0|\mathbf{x}_i, \mu_0, k_0, a, b, \rho)$. We consider each data set as an independent Bernoulli trial, whose possible outcome are the win of the first or of the second classifier. The probability of success (win of the second classifier) of the ith Bernoulli trial is p_i.

The number of data sets in which the second classifier is more accurate than the first classifier is a random variable X which follows a Poisson-binomial distribution (Lacoste et al. 2012). The Poisson-binomial distribution is a generalization of the binomial distribution in which the Bernoulli trials are allowed to have different probability of success. This probabilities are computed by Bayesian correlated t test and thus account both for the mean and the standard error of the cross-validation estimates. The probability of success is different on each data set, and thus the test does not assume the results on the different data sets to be identically distributed.

The cumulative distribution function of X is:

$$P(X \leq k) = \sum_{i=0}^{k} \xi(i) = \sum_{i=0}^{k} \left(\sum_{A \in \mathcal{F}_i} \prod_{i \in A} p_i \prod_{i \in A^c} (1 - p_i) \right) \tag{10}$$

where $\xi(i) = P(X = i)$, $\mathcal{F}_i$ is the set of all subsets of i integers that can be drawn from $\{1, 2, 3, \ldots, q\}$ and A^c is the complement of A: $A^c = \{1, 2, 3, , q\} \backslash A$. Hong (2013) discusses several methods to exactly compute the Poisson-binomial distribution. We adopt a sampling approach. We simulate q biased coin, one for each data set. The bias of the ith coin is p_i. We simulate the q coins 100,000 times. We count the proportion of times in which $x = 1, x = 2, \ldots, X = q$ out of the 100,000 trials. This yields a numerical approximation of the Poisson-binomial distribution.

The Poisson binomial test declares the second classifier significantly more accurate than the first classifier if $P(X > q/2) > 1 - \alpha$, namely if the probability of the second classifier being more accurate than the first on more than half the data sets is larger than $1 - \alpha$.

Table 3 Example of comparison of two classifiers in multiple datasets

	Datasets	μ_i	σ_i
Case 1	$D_1, \ldots, D_5$	0.1	0.05
	$D_6, \ldots, D_{10}$	−0.1	0.05
Case 2	$D_1, \ldots, D_5$	0.1	0.05
	$D_6, \ldots, D_{10}$	−0.1	0.15

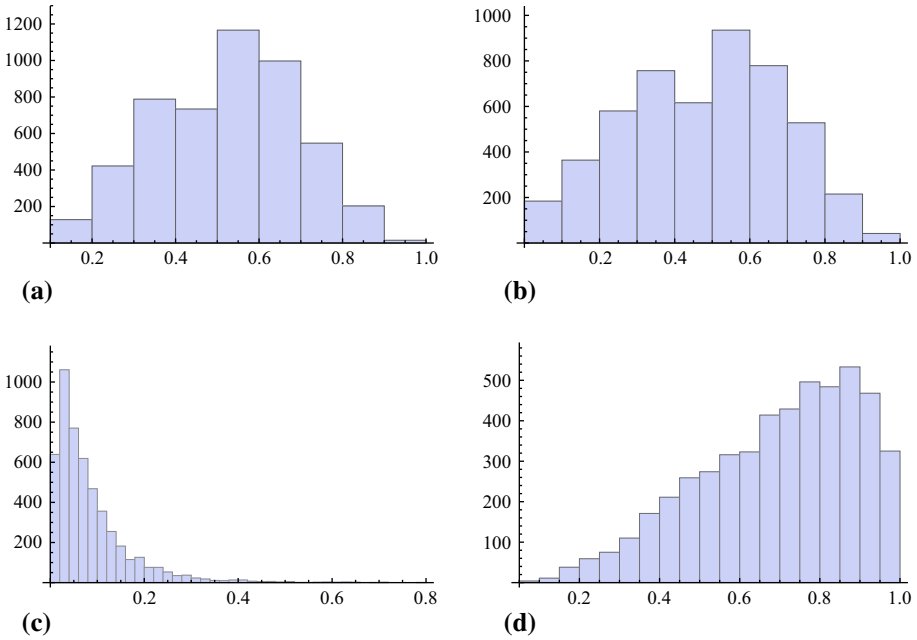

Fig. 1 Distribution of $P(X > q/2)$ for the Poisson test and distribution of the p values for the Wilcoxon signed-rank test in the two cases. **a** Wilcoxon case 1. **b** Wilcoxon case 2. **c** Poisson case 1. **d** Poisson case 2

3.2 Example

In order to understand the differences between the Poisson test and the Wilcoxon signed-rank test, consider the artificial example of Table 3.

In case 1, classifier A is more accurate than classifier B on five data sets. Classifier B is more accurate than classifier A on the remaining five data sets. Parameter μ_i and σ_i represent the mean and the standard deviation of the actual difference of accuracy among the two classifiers on each data set. The absolute value of μ_i is equal on all data sets and σ_i is equal on all data sets.

In case 2, the mean differences μ_i are the same as in case 1, but the standard deviation in $D_6, \ldots, D_{10}$ is three times larger. We have generated observations as follows

$$x_{ji} \sim N\left(\mu_i, \sigma_i^2\right),$$

for $i = 1, \ldots, 10$ (ten-folds cross-validation) and for the $j = 1, \ldots, 10$ datasets (here $\rho = 0$ but the results are similar if we consider a correlated model). Figure 1 shows the distribution of $P(X > q/2)$ (classifier A is better than B) for the Poisson test and the distribution of the

p values for the Wilcoxon signed-rank test in the two cases (computed for 5000 Monte Carlo runs). It can be observed that the distribution for Wilcoxon signed-rank test is practically unchanged in the two cases, while the distribution of the Poisson test is very different. The Poisson test is thus able to distinguish the two cases: it takes into account the variance of the mean accuracy in the ten-folds cross-validation of each dataset, while the Wilcoxon signed-rank test does not.

4 Experiments

The calibration and the power of the correlated t test have been already extensively studied by (Nadeau and Bengio 2003; Bouckaert 2003) and we refrain from doing it here. The same results apply to the Bayesian correlated t test, since the frequentist and the Bayesian correlated t test take the same decisions. The main result of such studies is that the rate of Type I errors of the correlated t test is considerably closer to the nominal test size α than the rate of Type I error of the standard t test. In the following we thus present results dealing with the inference on multiple data sets.

4.1 Two classifiers with known difference of accuracy

We generate the data sets sampling the instances from the Bayesian network $C \rightarrow F$, where C is the binary class with states $\{c_0, c_1\}$ and F is a binary feature with states $\{f_0, f_1\}$. The parameters are: $P(c_0) = 0.5$; $P(f_0|c_0) = \theta$; $P(f_0|c_1) = 1 - \theta$ with $\theta > 0.5$. We refer to this model with exactly these parameters as BN.

Notice that if the BN model is used both the generate the instances and to issue the prediction, its expected accuracy is[3] θ.

Once a data set is generated, we assess via cross-validation the accuracy of two classifiers. The first classifier is the majority predictor also known as *zeroR*. It predicts the most frequent class observed in the training set. If the two classes are equally frequent in the training set, it randomly draws the prediction. Its expected accuracy is thus 0.5.

The second classifier is $\hat{BN}$, namely the Bayesian network $C \rightarrow F$ with parameters learned from the training data. The actual difference of accuracy between the two classifiers is thus approximately $\delta_{acc} = \theta - 0.5$. To simulate the difference of accuracy δ_{acc} between the two classifiers we set $\theta = 0.5 + \delta_{acc}$ in the parameters of the BN model. We repeat experiments using different values of δ_{acc}.

We perform the tests in a one-sided fashion: the null hypothesis is that zeroR is less or equally accurate than $\hat{BN}$. The alternative hypothesis is that $\hat{BN}$ is more powerful than zeroR. We set the size of both the signed rank and the Poisson tests to $\alpha = 005$. We measure the power of a test as the rate of rejection of the null hypothesis when $\delta_{acc} > 0$.

We present results obtained with $m = 1$ and $m = 10$ runs of cross-validation.

4.2 Fixed difference of accuracy on all data sets

As a first experiment, we set the actual difference of accuracy δ_{acc} among the two classifiers as identical on all the q data sets. We assume the availability of $q = 50$ data sets. This is a

[3] The proof is as follows. Consider the instances with $F = f_0$. We have that $P(c_0|f_0) = \theta > 0.5$, so the model always predicts c_0 if $F = f_0$. This prediction is accurate with probability θ. Regarding the instances with $F = f_1$, the most probable class is c_1. Also this prediction is accurate with probability θ. Overall the classifier has probability θ of being correct.

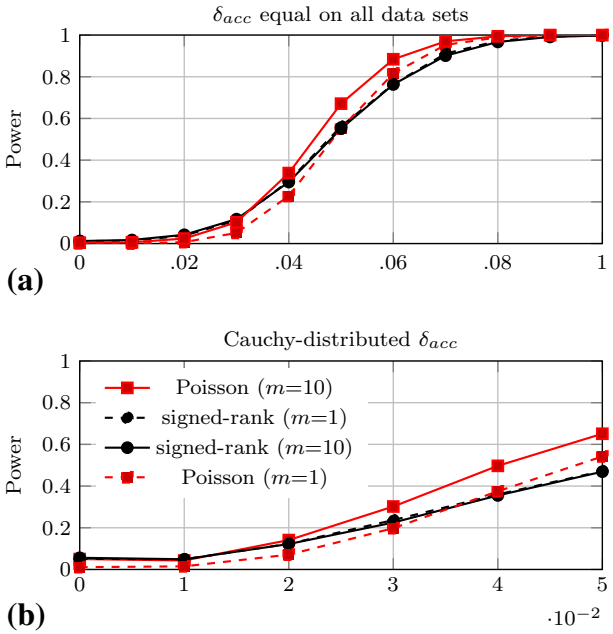

Fig. 2 Power and calibration of the tests over multiple data sets. The plots share the same legend. The Poisson test has squared marks. The signed-rank test has *circle marks*. *Dashed lines* refer to one run of cross-validation, *solid lines* refer to ten runs of cross-validation. The plots refer to the case $q = 50$. **a** Difference of accuracy (δ_{acc}). **b** Mean difference of accuracy ($\bar{\delta}_{acc}$)

common size for a comparison of classifiers. We consider the following different values of δ_{acc}: $\{0, 0.01, 0.02, \ldots, 0.1\}$.

For each value of δ_{acc} we repeat 5000 experiments as follows. We allow the various data sets to have different size $\boldsymbol{s} = s_1, s_2, \ldots, s_q$. We draw the sample size of each data set uniformly from $\mathcal{S} = \{25, 50, 100, 250, 500, 1000\}$. We generate each data set using the BN model; then we assess via cross-validation the accuracy of both zero R and $\hat{BN}$. We then compare the two classifiers via the Poisson and the signed-rank test.

The results are shown in Fig. 2a. Both tests yield Type I error rate lower than 0.05 when $\delta_{acc} = 0$; thus they are correctly calibrated. The power of the tests can be assessed looking at the results for strictly positive values of δ_{acc}. If one run of cross-validation is performed, the Poisson test is generally *less* powerful than the signed-rank test. However if ten runs of cross-validation are performed, the Poisson is generally *more* powerful than the signed rank. The signed-rank does not account for the uncertainty of the estimates and thus its power is roughly the same regardless whether one or ten runs of cross-validation have been performed.

The same conclusions are confirmed in the case $q = 25$.

4.3 Difference of accuracy sampled from the Cauchy distributions

We remove the assumption of δ_{acc} being equal for all data sets. Instead for each data set we sample δ_{acc} from a Cauchy distribution. We set the median and the scale parameter of the Cauchy to a value $\overline{\delta_{acc}} > 0$. A different value of $\overline{\delta_{acc}}$ defines a different experimental

Table 4 Comparison of the decision of the Poisson and the signed-rank test on real data sets

	Naive Bayes	J48	J48-gr	AODE	HNB
Data sets 1–27					
Naive Bayes	–	1/0	1/0	1/1	1/1
J48	–	–	1/1	1/0	0/0
J48-gr	–	–	–	1/0	0/0
AODE	–	–	–	–	0/0
Data sets 28–54					
Naive Bayes	–	1/0	1/0	1/1	1/1
J48	–	–	1/1	1/0	0/0
J48-gr	–	–	–	1/0	0/0
AODE	–	–	–	–	0/0
Data sets 1–54					
Naive Bayes	–	1/0	1/0	1/1	1/1
J48	–	–	1/1	1/1	0/1
J48-gr	–	–	–	1/0	0/1
AODE	–	–	–	–	0/0

The entries of the table have the following meaning: < Poisson decision > / < signed-rank decision >. The decision is about the classifier of the current column being significantly more accurate than the classifier of the current row. For instance the entry 1/0 means that only the Poisson test claims the difference to be significant

setting. We consider the following values of $\overline{\delta}_{acc}$: $\{0, 0.01, 0.02, \ldots, 0.05\}$. We run 5,000 experiments for each value of $\overline{\delta}_{acc}$. We assume the availability of $q = 50$ data sets.

Sampling from the Cauchy one sometimes obtains values of δ_{acc} whose absolute value is larger than 0.5. It is not possible to simulate difference of accuracy that large. Thus sampled values of δ_{acc} larger than 0.5 or smaller than -0.5 are capped to 0.5 and -0.5 respectively.

The results are given in Fig. 2b. Both tests are correctly calibrated for $\delta_{acc} = 0$. This is noteworthy since values sampled from the Cauchy are often aberrant and can easily affect the inference of parametric tests.

Let us analyze the power of the tests for $\delta_{acc} > 0$. If one run of cross-validation is performed, the Poisson test is slightly *less* powerful than the signed-rank test. If ten runs of cross-validation are performed, the Poisson test is *more* powerful than the signed-rank test.

Such findings are confirmed by repeating the simulation with a number of data sets $q = 25$.

4.4 Application to real data sets

We consider 54 data sets[4] from the UCI repository. We consider five different classifiers: naive Bayes, averaged one-dependence estimator (AODE), hidden naive Bayes (HNB), J48 decision tree and J48 grafted (J48-gr). All the algorithms are described in (Witten et al. 2011). On each data set we run ten runs of ten-folds cross-validation using the WEKA[5] software.

We then compare each couple of classifiers via the signed-rank and the Poisson test.

We sort the data sets alphabetically and we repeat the analysis three times. The first time we compare the classifiers on data sets 1–27; the second time we compare the classifiers on data sets 28–54; the third time we compare the classifiers on all data sets. The results are given in Table 4. The zeros and the ones in Table 4 indicate respectively that the null or the alternative hypothesis has been accepted.

[4] Available from http://www.cs.waikato.ac.nz/ml/weka/datasets.html.

[5] Available from http://www.cs.waikato.ac.nz/ml/weka.

The Poisson test detects seven significant differences out of the ten comparison in all the three experiments. It consistently detects the same seven significances in all the three experiments. The signed-rank test is less powerful. It detects only three significances in the first and in the second experiment. When all data sets are available its power increases and it detects three further differences, arriving to six detected differences. Overall the Poisson test is both more powerful and more replicable.

The detected differences are in agreement with what is known in literature: both AODE and HNB are recognized as significantly more accurate than naive Bayes, J48-gr is recognized as significantly more accurate than both naive Bayes and J48. The two tests take different decisions when comparing couples of high-performance classifiers such as HNB, AODE and J48-gr.

4.5 Software

At the link www.idsia.ch/~giorgio/poisson/test-package.zip we provide both the Matlab and the R implementation of our test. They can be used by a researcher who wants to compare any two algorithms assessed via cross-validation on multiple data sets. The package also allows reproducing the experiments of this paper.

The procedure can be easily implemented also in other computational environments. The standard t test is available within every computational package. The frequentist correlated t test can be implemented by simply changing the statistic of the standard t test, according to Eq. (1). Under the matching prior, the posterior probability of the null computed by the Bayesian correlated t test correspond to the p value computed by the one-sided frequentist correlated t test. Once the posterior probabilities are computed on each data set, it remains to compute the Poisson-binomial probability distribution. The Poisson-binomial distribution can be straightforwardly computed via sampling, while exact approaches (Hong 2013) are more difficult to implement.

5 Conclusions

To our knowledge, the Poisson test is the first test which compares two classifiers on multiple data sets accounting for the correlation and the uncertainty of the results generated by cross-validation on each individual data set. The test is usually more powerful than the signed-rank if ten runs of cross-validation are performed, which is anyway common practice. A limit of the approach based on the Poisson-binomial is that its inferences refer to the sample of provided data sets rather than to the population from which the data sets have been drawn. A way to overcome this limit could be the development a hierarchical test able to make inference on the population of data sets.

Appendix: Proof of Theorem 1

Preliminaries

The symmetry of the correlation matrix $\boldsymbol{M}$ implies that $\boldsymbol{M}^{-1}$ is symmetric too. Assuming $n = 3$ as an example, its structure is:

$$\boldsymbol{M} = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \qquad \boldsymbol{M}^{-1} = \frac{1}{|\boldsymbol{M}|}\text{Adj}(\mathbf{M}) = \frac{1}{|\boldsymbol{M}|}\begin{bmatrix} \alpha & \beta & \beta \\ \beta & \alpha & \beta \\ \beta & \beta & \alpha \end{bmatrix}$$

where α, β are the entries of the adjugate matrix.

We get:

$$\mathbf{H}^T\mathbf{M}^{-1}\mathbf{H} = \frac{n\alpha + n(n-1)\beta}{|\boldsymbol{M}|} = \frac{n(\alpha + (n-1)\beta)}{|\boldsymbol{M}|} \tag{11}$$

$$\mathbf{H}^T\mathbf{M}^{-1}\mathbf{x} = \frac{\sum_{i=1}^{n}(\alpha + (n-1)\beta)x_i}{|\boldsymbol{M}|} \tag{12}$$

Moreover,

$$\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{H} = \mathbf{H}^T\boldsymbol{\Sigma}^{-1}x = \frac{1}{\sigma^2|\boldsymbol{M}|}[\alpha + (n-1)\beta]\sum_i x_i. \tag{13}$$

Estimating $\boldsymbol{\mu}$

From (6), the log-likelihood is:

$$L(\mu, \sigma^2, \rho) = -\frac{1}{2}(\mathbf{x} - \mathbf{H}\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{H}\mu) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|).$$

Its derivative w.r.t. μ is:

$$\begin{aligned} &\frac{\partial}{\partial\mu}\left(-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{H}\mu) + \frac{1}{2}\mathbf{H}^T\mu\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{H}\mu)\right) \\ &= \frac{\partial}{\partial\mu}\left(\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{H}\mu + \frac{1}{2}\mathbf{H}^T\mu\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{H}^T\mu\boldsymbol{\Sigma}^{-1}\mathbf{H}\mu\right) \\ &= \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{H} + \frac{1}{2}\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \mu\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H} = \mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \mu\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H} \end{aligned}$$

where in the last passage we have used the first equality in (13).

Substituting $\boldsymbol{\Sigma}$ with $\sigma^2\boldsymbol{M}$, equating the derivative to 0 and using equations (11) and (12) we get:.

$$\mu = \frac{\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}{\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H}} = \frac{\sum_{i=1}^{n} x_i}{n},$$

which is the traditional maximum likelihood estimator of the mean. It does not depend on σ^2 or ρ.

Estimating $\boldsymbol{\sigma^2}$

Recalling that $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{M}$ and thus $|\boldsymbol{\Sigma}| = \left(\sigma^2\right)^n|\boldsymbol{M}|$, the log-likelihood is:

$$\begin{aligned} L(\mu, \sigma^2, \rho) &= -\frac{1}{2\sigma^2}(\mathbf{x} - \mathbf{H}\mu)^T\boldsymbol{M}^{-1}(\mathbf{x} - \mathbf{H}\mu) - \frac{1}{2}\log\left((\sigma^2)^n|\boldsymbol{M}|\right) - \frac{n}{2}\log(2\pi) \\ &= -\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{H}\mu)^T\boldsymbol{M}^{-1}(\mathbf{x}-\mathbf{H}\mu) - \frac{1}{2}\log\left((\sigma^2)^n\right) - \underbrace{\frac{1}{2}\log(|\boldsymbol{M}|) - \frac{n}{2}\log(2\pi)}_{\text{not depending on } \sigma^2}. \end{aligned}$$

Only the first two terms of the above expression are relevant for the derivative. Thus, by replacing μ with $\hat{\mu}$ and by equating to zero the derivative. we obtain

$$\frac{\partial}{\partial \sigma^2} L(\hat{\mu}, \sigma^2, \rho) = \frac{1}{2(\sigma^2)^2} (\mathbf{x} - \mathbf{H}\hat{\mu})^T \boldsymbol{M}^{-1} (\mathbf{x} - \mathbf{H}\hat{\mu}) - \frac{1}{2} \frac{1}{(\sigma^2)^n} n(\sigma^2)^{n-1} = 0$$

Finally, we get:

$$\bar{\sigma}^2 = \frac{(\mathbf{x} - \mathbf{H}\hat{\mu})^T \boldsymbol{M}^{-1} (\mathbf{x} - \mathbf{H}\hat{\mu})}{n}$$

The product $(\mathbf{x} - \mathbf{H}\hat{\mu})^T_{1\times n} \boldsymbol{M}^{-1}_{n\times n} (\mathbf{x} - \mathbf{H}\hat{\mu})_{n\times 1}$ yields a scalar. The trace of a scalar is the scalar itself. The trace is invariant under cyclic permutations: $tr(ABC) = tr(BCA)$. We thus have:

$$(\mathbf{x} - \mathbf{H}\hat{\mu})^T \boldsymbol{M}^{-1} (\mathbf{x} - \mathbf{H}\hat{\mu}) = tr[(\mathbf{x} - \mathbf{H}\hat{\mu})^T \boldsymbol{M}^{-1} (\mathbf{x} - \mathbf{H}\hat{\mu})] =$$
$$tr[\boldsymbol{M}^{-1} (\mathbf{x} - \mathbf{H}\hat{\mu})(\mathbf{x} - \mathbf{H}\hat{\mu})^T] = tr(\boldsymbol{M}^{-1}\mathbf{Z})$$

where $\mathbf{Z} = (\mathbf{x} - \mathbf{H}\hat{\mu})(\mathbf{x} - \mathbf{H}\hat{\mu})^T$ and so

$$\bar{\sigma}^2 = \frac{tr(\boldsymbol{M}^{-1}\mathbf{Z})}{n} \tag{14}$$

Thus $\bar{\sigma}^2$ depends on the correlation ρ through $\boldsymbol{M}$.

Useful lemmas for estimating ρ

Lemma 1 *The determinant of $\boldsymbol{M}$ is:* $(1 + (n-1)\rho)(1-\rho)^{n-1}$.

Proof Consider the ith and the jth ($i \neq j$) row of matrix $\boldsymbol{M}_{n\times n}$, containing elements $\{m_{i1}, m_{i2}, \ldots, m_{in}\}$ and $\{m_{j1}, m_{j2}, \ldots, m_{jn}\}$ respectively. The value of $|\boldsymbol{M}|$ does not change if we substitute each element of the ith row as follows:

$$m_{ik} \leftarrow m_{ik} + b \cdot m_{jk} \quad \forall k \in \{1, 2, \ldots, n\}$$

where b is any scalar weight and in particular for $b = 1$. Then, consider the matrix $\mathbf{N}$ obtained by adding the second row to the first row ($b = 1$), then the third row to the first row, ... then the nth row to the first row:

$$\boldsymbol{M} = \begin{vmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho & \rho & \rho & \ldots & 1 \end{vmatrix}$$

$$\boldsymbol{N} = \begin{vmatrix} 1+(n-1)\rho & 1+(n-1)\rho & 1+(n-1)\rho & \ldots & 1+(n-1)\rho \\ \rho & 1 & \rho & \ldots & \rho \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho & \rho & \rho & \ldots & 1 \end{vmatrix}$$

Consider now matrix $\boldsymbol{O}$ defined as follows:

$$\boldsymbol{O} = \begin{vmatrix} 1 & 1 & 1 & \ldots & 1 \\ \rho & 1 & \rho & \ldots & \rho \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho & \rho & \rho & \ldots & 1 \end{vmatrix}$$

Then, $|\boldsymbol{M}| = |\boldsymbol{N}| = (1 + (n-1)\rho) \cdot |\boldsymbol{O}|$. Consider now adding the elements of the first row of $|\boldsymbol{O}|$ to the second row of $|\boldsymbol{O}|$, using the scalar weight $b = -\rho$. Then add $-\rho$ times

the first row to the third, to the fourth, ... to the nth row of $|\boldsymbol{O}|$. This yields matrix $\boldsymbol{P}$, with $|\boldsymbol{P}| = |\boldsymbol{O}|$:

$$\boldsymbol{P} = \begin{vmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & 1-\rho & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1-\rho \end{vmatrix}$$

We have $|\boldsymbol{P}| = (1-\rho)^{n-1}$ and thus:

$$|\boldsymbol{M}| = |\boldsymbol{N}| = (1+(n-1)\rho)|\boldsymbol{O}| = (1+(n-1)\rho)|\boldsymbol{P}| = (1+(n-1)\rho)(1-\rho)^{n-1}$$

□

Lemma 2 *The entries of* $\boldsymbol{M}^{-1}$ *are* $\alpha = (1+(n-2)\rho)(1-\rho)^{n-2}$ *and* $\beta = -\rho(1-\rho)^{n-2}$.

Proof By definition of adjugate matrix, α is the determinant of each principal minor of $\boldsymbol{M}$. Consider the principal minor obtained by removing the first row and the first column from $\boldsymbol{M}$. This sub-matrix has the same structure of $\boldsymbol{M}$, but with dimension $(n-1)\times(n-1)$. Its determinant is thus $(1+(n-2)\rho)(1-\rho)^{n-2}$, which gives the value of α. The same result is obtained considering any other principal minor.

Parameter β corresponds instead to the determinant of any non-principal minor of $\boldsymbol{M}$, multiplied by -1^{i+j}, where i and j are respectively the index of the row and the column removed from $\boldsymbol{M}$ to obtain the minor. Consider the minor obtained by removing the first row and the second column:

$$\boldsymbol{Q} = \begin{vmatrix} \rho & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \rho & \dots & \dots & \rho & 1 \end{vmatrix}$$

By subtracting the first row ($i=1$) from the second row ($j=2$), the first row from the third row, the first row from the nth row we get:

$$\boldsymbol{R} = \begin{vmatrix} \rho & \rho & \rho & \dots & \rho \\ 0 & 1-\rho & 0 & \dots & 0 \\ 0 & 0 & 1-\rho & \dots & 0 \\ 0 & \dots & \dots & 0 & 1-\rho \end{vmatrix}$$

whose determinant is $(1-\rho)^{n-2}\rho$. The value of β is thus $-\rho(1-\rho)^{n-2}$, the minus sign being due to the sum of i and j being an odd number. The same result is obtained considering any other principal minor. □

Estimating ρ

The log-likelihood evaluated in $\hat{\mu}, \bar{\sigma}^2$ is:

$$L(\rho,\mu,\sigma^2)\big|_{\hat{\mu},\bar{\sigma}^2} = -\frac{1}{2\bar{\sigma}^2}(\mathbf{x}-\mathbf{H}\hat{\mu})^T\boldsymbol{M}^{-1}(\mathbf{x}-\mathbf{H}\hat{\mu}) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left((\bar{\sigma}^2)^n|\boldsymbol{M}|\right)$$

$$= -\frac{\hat{\mu}^2}{2\bar{\sigma}^2} Tr(\boldsymbol{M}^{-1}\mathbf{Z}) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left((\bar{\sigma}^2)^n|\boldsymbol{M}|\right)$$
$$= \underbrace{-\frac{n\hat{\mu}^2}{2} - \frac{n}{2}\log(2\pi)}_{\text{not depending on } \rho} - \frac{1}{2}\log\left((\bar{\sigma}^2)^n|\boldsymbol{M}|\right)$$

where in the last passage we have exploited that $\bar{\sigma}^2 = Tr(\boldsymbol{M}^{-1}\mathbf{Z})/n$ as shown in (14).

The derivative w.r.t. ρ is:

$$\frac{\partial}{\partial\rho} L(\mu, \sigma^2, \rho)\Big|_{\hat{\mu},\bar{\sigma}^2} = \frac{\partial}{\partial\rho}\left(-\frac{1}{2}\log((\bar{\sigma}^2)^n|\boldsymbol{M}|)\right) = \frac{\partial}{\partial\rho}\left(-\frac{1}{2}\log\left[\left(\frac{Tr(\boldsymbol{M}^{-1}\mathbf{Z})}{n}\right)^n|\boldsymbol{M}|\right]\right)$$
$$= \frac{\partial}{\partial\rho}\left(-\frac{1}{2}n\log\left(\frac{Tr(\boldsymbol{M}^{-1}\mathbf{Z})}{n}\right) - \frac{1}{2}log|\boldsymbol{M}|\right)$$
$$= \frac{\partial}{\partial\rho}\left(-\frac{1}{2}n\log\left(Tr(\boldsymbol{M}^{-1}\mathbf{Z})\right) + \frac{1}{2}n\log(n) - \frac{1}{2}log|\boldsymbol{M}|\right)$$
$$= -\frac{1}{2}\frac{\partial}{\partial\rho}\left(n\log\left(Tr(\boldsymbol{M}^{-1}\mathbf{Z})\right)\right) - \frac{1}{2}\frac{\partial}{\partial\rho}(log|\boldsymbol{M}|)$$
$$= -\frac{1}{2}\frac{n}{Tr(\boldsymbol{M}^{-1}\mathbf{Z})}\frac{\partial}{\partial\rho}\left(Tr(\boldsymbol{M}^{-1}\mathbf{Z})\right) - \frac{1}{2}\frac{1}{|\boldsymbol{M}|}\frac{\partial}{\partial\rho}(|\boldsymbol{M}|)$$

Let us now consider $Tr\left(\boldsymbol{M}^{-1}\mathbf{Z}\right) = \frac{1}{|\boldsymbol{M}|}Tr\left(\text{Adj}(\mathbf{M})\,\mathbf{Z}\right)$ and define $\boldsymbol{S} = \text{Adj}(\mathbf{M})\,\mathbf{Z}$. Let us denote the difference between an observation and the maximum likelihood mean as $\delta_i = x_i - \hat{\mu}_i$. The ith diagonal element of $\boldsymbol{S}$ is

$$s_{ii} = \alpha\delta_i^2 + \beta\left(\sum_{i\neq j}\delta_i\delta_j\right)$$

Notice that $\delta_i^2 + \left(\sum_{i\neq j}\delta_i\delta_j\right) = 0$, due to the following relation:

$$\delta_i^2 + \left(\sum_{i\neq j}\delta_i\delta_j\right) = \delta_i\left(\delta_i + \sum_{i\neq j}\delta_j\right) = \delta_i\left(\sum_i x_i - n\hat{\mu}\right) = 0$$

We can then rewrite $s_{ii} = (\alpha - \beta)\delta_i^2$. Summing over all the elements of the diagonal, we get:

$$Tr\left(\boldsymbol{M}^{-1}\mathbf{Z}\right) = \frac{(\alpha-\beta)\sum_{i=1}^n\delta_i^2}{|\boldsymbol{M}|} = \frac{(\alpha-\beta)f(\mathbf{x})}{|\boldsymbol{M}|}$$

where $f(\mathbf{x}) = \sum_{i=1}^n\delta_i^2$ depends only on the data.

By equating to zero the derivative of the log-likelihood w.r.t. ρ, we obtain:

$$0 = -\frac{1}{2}\frac{n}{Tr(\boldsymbol{M}^{-1}\mathbf{Z})}\frac{\partial}{\partial\rho}\left(Tr(\boldsymbol{M}^{-1}\mathbf{Z})\right) - \frac{1}{2}\frac{1}{|\boldsymbol{M}|}\frac{\partial}{\partial\rho}(|\boldsymbol{M}|)$$
$$= \frac{n|\boldsymbol{M}|}{(\alpha-\beta)f(\mathbf{x})}\left(f(\mathbf{x})\frac{\partial}{\partial\rho}(\alpha-\beta)\frac{1}{|\boldsymbol{M}|}\right) + \frac{1}{|\boldsymbol{M}|}\frac{\partial}{\partial\rho}(|\boldsymbol{M}|)$$
$$= n(1-\rho)\frac{\partial}{\partial\rho}\left(\frac{1}{1-\rho}\right) + \frac{1}{|\boldsymbol{M}|}\frac{\partial}{\partial\rho}(|\boldsymbol{M}|)$$

where we have exploited that $\alpha - \beta = |\boldsymbol{M}|/(1-\rho)$. Since

$$\frac{1}{|\boldsymbol{M}|}\frac{\partial}{\partial\rho}(|\boldsymbol{M}|) = -\frac{n(n-1)\rho}{(1+(n-1)\rho)(1-\rho)}$$

it can easily be shown that

$$0 = n(1-\rho)\frac{\partial}{\partial\rho}\left(\frac{1}{1-\rho}\right) + \frac{1}{|\boldsymbol{M}|}\frac{\partial}{\partial\rho}(|\boldsymbol{M}|) = \frac{n}{(\rho-1)(1+(n-1)\rho)}$$

Thus, there is no value $\rho \in [0,1)$ which can make the derivative equal to zero and the derivative is always decreasing in ρ. Thus the maximum likelihood estimate of ρ is $\hat{\rho} = 0$. For any fixed ρ, it can easily be shown that the Hessian of the likelihood w.r.t. μ, σ^2 computed at $\hat{\mu}, \bar{\sigma}^2$ is negative definite. In fact, we have that

$$\frac{\partial^2}{\partial\mu^2}L(\mu,\sigma^2,\rho)\Big|_{\hat{\mu},\bar{\sigma}^2} = -\boldsymbol{H}^T\boldsymbol{M}^{-1}\boldsymbol{H}^T, \quad \frac{\partial^2}{\partial(\sigma^2)^2}L(\mu,\sigma^2,\rho)\Big|_{\hat{\mu},\bar{\sigma}^2} = -\frac{1}{2}\frac{1}{(\bar{\sigma}^2)^2}$$

and $\frac{\partial^2}{\partial\sigma^2\partial\mu}L(\mu,\sigma^2,\rho)|_{\hat{\mu},\bar{\sigma}^2} = 0$. Thus, $\hat{\mu}, \bar{\sigma}^2, \hat{\rho}$ is the maximum likelihood estimator. Since $\hat{\rho} = 0$, this estimator is not consistent whenever the true correlation is not zero (strictly positive).

Proof of Theorem 2

Let us define $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then:

$$\begin{aligned}(\mathbf{x}-\mathbf{H}\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{H}\mu) &= (\mathbf{x}-\mathbf{H}(\mu-\hat{\mu}+\hat{\mu}))^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{H}(\mu-\hat{\mu}+\hat{\mu}))\\ &= (\mathbf{x}-\mathbf{H}\hat{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{H}\hat{\mu}) + (\mu-\hat{\mu})\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{H}(\mu-\hat{\mu}).\end{aligned}$$

Let us define $\nu = 1/\sigma^2$, then we can rewrite the likelihood as:

$$\begin{aligned}p(\mathbf{x}|\mu,\nu,\rho) = &\frac{\nu^{n/2-1/2}}{(2\pi)^{n/2}\sqrt{|\boldsymbol{M}|}}\exp\left(-\frac{\nu}{2}(\mathbf{x}-\mathbf{H}\hat{\mu})^T\boldsymbol{M}^{-1}(\mathbf{x}-\mathbf{H}\hat{\mu})\right)\\ &\times\nu^{1/2}\exp\left(-\frac{\nu}{2}(\mu-\hat{\mu})\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}(\mu-\hat{\mu})\right)\end{aligned} \tag{15}$$

Given ρ, the likelihood (15) has the structure of a Normal-Gamma distribution. Therefore, for the unknown parameters μ, ν, we consider the conjugate prior:

$$p(\mu|\nu,\rho) = N(\mu;\mu_0,k_0/\nu), \quad p(\nu|\rho) = G(\nu;a,b), \tag{16}$$

with parameters μ_0, k_0, a, b. By combining the likelihood and the prior, we obtain the joint:

$$\begin{aligned}p(\mu,\nu,\mathbf{x}|\rho) &= p(\mathbf{x}|\mu,\nu,\rho)p(\mu|\nu,\rho)p(\nu|\rho)\\ &\propto \frac{\nu^{\frac{n+2a}{2}-1}}{(2\pi)^{n/2}\sqrt{|\boldsymbol{M}|}}\exp\left(-\frac{\nu}{2}(\mathbf{x}-\mathbf{H}\hat{\mu})^T\boldsymbol{M}^{-1}(\mathbf{x}-\mathbf{H}\hat{\mu}) - b\nu\right)\\ &\quad\times\nu^{\frac{1}{2}}\exp\left(-\frac{\nu}{2}(\mu-\hat{\mu})^2\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H} - \frac{\nu}{2k_0}(\mu-\mu_0)^2\right).\end{aligned}$$

Let us define the posterior mean

$$\tilde{\mu} = \left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H} + \frac{1}{k_0}\right)^{-1}\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{x} + \frac{\mu_0}{k_0}\right),$$

then

$$
\begin{aligned}
&(\mu-\hat{\mu})^2\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}(\mu-\mu_0)^2\\
&=\mu^2\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}\right)-2\mu\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}\hat{\mu}+\frac{\mu_0}{k_0}\right)+\hat{\mu}^2\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{\mu_0^2}{k_0}\\
&=\mu^2\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}\right)-2\mu\left(\mathbf{H}^T\boldsymbol{M}^{-1}\boldsymbol{x}+\frac{\mu_0}{k_0}\right)+\hat{\mu}^2\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{\mu_0^2}{k_0}\\
&=(\mu-\tilde{\mu})^2\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}\right)+\hat{\mu}^2\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{\mu_0^2}{k_0}-\tilde{\mu}^2\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}\right).
\end{aligned}
$$

Thus, we can rewrite the joint as $p(\mu,\nu,\mathbf{x}|\rho)\propto\ell_1\ell_2$ with

$$
\ell_1=\nu^{1/2}\exp\left(-\frac{\nu}{2}(\mu-\tilde{\mu})^2\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}\right)\right)\propto N\left(\mu;\tilde{\mu},\frac{1}{\nu}\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}\right)^{-1}\right)
$$

and

$$
\begin{aligned}
\ell_2=&\frac{\nu^{\frac{n+2a}{2}-1}}{(2\pi)^{n/2}\sqrt{|\boldsymbol{M}|}}\exp\Bigg(-\frac{\nu}{2}\Bigg((\mathbf{x}-\mathbf{H}\hat{\mu})^T\boldsymbol{M}^{-1}(\mathbf{x}-\mathbf{H}\hat{\mu})+2b\\
&-\hat{\mu}^2\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}-\frac{\mu_0^2}{k_0}+\tilde{\mu}^2\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}\right)\Bigg)\Bigg)\propto\frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{M}|}}G(\nu;\tilde{a},\tilde{b})\frac{\Gamma(\tilde{a})}{\tilde{\beta}^{\tilde{a}}}
\end{aligned}
$$

with $\tilde{a}=a+\frac{n}{2}$ and

$$
\tilde{b}=\frac{1}{2}\left((\mathbf{x}-\mathbf{H}\hat{\mu})^T\boldsymbol{M}^{-1}(\mathbf{x}-\mathbf{H}\hat{\mu})+2b-\hat{\mu}^2\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}-\frac{\mu_0^2}{k_0}+\tilde{\mu}^2\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}\right)\right).
$$

Hence, it follows that the posterior is

$$
p(\mu,\nu|\boldsymbol{x},\rho)=N\left(\mu;\tilde{\mu},\frac{1}{\nu}\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}\right)^{-1}\right)G(\nu;\tilde{a},\tilde{b}).
$$

The marginal posterior of μ can be obtained by marginalizing out ν:

$$
\begin{aligned}
p(\mu|\boldsymbol{x},\rho)&\propto\frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{M}|}}\left((\mu-\tilde{\mu})^2\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}\right)+2\tilde{b}\right)^{-(2\tilde{a}+1)/2}\\
&\propto St\left(\mu;2\tilde{a},\tilde{\mu},\frac{\tilde{b}}{\tilde{a}}\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}+\frac{1}{k_0}\right)^{-1}\right).
\end{aligned}
\tag{17}
$$

Proof of Corollary 1

Let us consider the matching prior $\mu_0=0$, $k_0\to\infty$, $a=-1/2$, $b=0$, then (17) becomes

$$
p(\mu|\boldsymbol{x},\rho)\propto St\left(\mu;n-1,\hat{\mu},\frac{(\mathbf{x}-\mathbf{H}\hat{\mu})^T\boldsymbol{M}^{-1}(\mathbf{x}-\mathbf{H}\hat{\mu})}{(n-1)\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}\right)}\right).
\tag{18}
$$

By exploiting $(\mathbf{x}-\mathbf{H}\hat{\mu})^T\boldsymbol{M}^{-1}(\mathbf{x}-\mathbf{H}\hat{\mu}) = Tr(\boldsymbol{M}^{-1}\boldsymbol{Z}) = \frac{\alpha-\beta}{|\boldsymbol{M}|}\sum_{i=1}^{n}\delta_i$ and $\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H} = (\alpha n+\beta n(n-1))/|\boldsymbol{M}|$, then we have that

$$\frac{(\mathbf{x}-\mathbf{H}\hat{\mu})^T\boldsymbol{M}^{-1}(\mathbf{x}-\mathbf{H}\hat{\mu})}{(n-1)\left(\mathbf{H}^T\boldsymbol{M}^{-1}\mathbf{H}\right)} = \frac{(\alpha-\beta)\sum_{i=1}^{n}\delta_i}{(n-1)(\alpha n+\beta n(n-1))} = \frac{1}{n}\frac{(\alpha-\beta)}{(\alpha+\beta(n-1))}\hat{\sigma}^2,$$

where $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\delta_i$. Hence, one gets

$$\frac{1}{n}\frac{(\alpha-\beta)}{(\alpha+\beta(n-1))}\hat{\sigma}^2 = \frac{1}{n}\frac{1+(n-1)\rho}{1-\rho}\hat{\sigma}^2 = \left(\frac{1}{n}+\frac{\rho}{1-\rho}\right)\hat{\sigma}^2,$$

which ends the proof.

References

Benavoli, A., Mangili, F., Corani, G., Zaffalon, M., & Ruggeri, F. (2014). A Bayesian Wilcoxon signed-rank test based on the Dirichlet process. In Proceedings of the 31st International Conference on Machine Learning (ICML 2014) (pp. 1026–1034).

Bernardo, J. M., & Smith, A. F. M. (2009). *Bayesian theory* (Vol. 405). Chichester: Wiley.

Bouckaert, R. R. (2003). Choosing between two learning algorithms based on calibrated tests. In Proceedings of the 20th International Conference on Machine Learning (ICML-03) (pp. 51–58).

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

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

Hong, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics and Data Analysis*, *59*, 41–51.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial intelligence-Volume 2 (pp. 1137–1143). Morgan Kaufmann Publishers Inc.

Lacoste, A., Laviolette, F., & Marchand, M. (2012). Bayesian comparison of machine learning algorithms on single and multiple datasets. In Proceeding of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12) (pp. 665–675).

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge: MIT press.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281.

Otero, J., Sánchez, L., Couso, I., & Palacios, A. (2014). Bootstrap analysis of multiple repetitions of experiments using an interval-valued multiple comparison procedure. *Journal of Computer and System Sciences*, *80*(1), 88–100.

Press, S. J. (2012). *Applied multivariate analysis: Using Bayesian and frequentist methods of inference*. Mineola: Courier Dover Publications.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, USA: Morgan Kaufmann.