

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

Giorgio Corani and Alessio Benavoli

Received: date / Accepted: date

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.

1 Introduction

2 A typical problem in machine learning is to compare the accuracy of two competing
3 classifiers on a data set D . Usually one measures the accuracy of both classifiers via
4 k -folds cross-validation. After having performed cross-validation, one has to decide
5 if the accuracy of the two classifiers on data set D is significantly different. The
6 decision is made using a statistical hypothesis test which analyzes the measures
7 of accuracy yielded by cross-validation on the different folds. Using a *t*-test is
8 however a naive choice. The *t*-test assumes the measures of accuracy taken on the
9 different folds to be independent. Such measures are instead correlated because of
10 the overlap of the training sets built during cross-validation. As a result the *t* test
11 is *not* calibrated, namely its rate of Type I errors is much larger than the nominal

G. Corani and A. Benavoli are with
Istituto Dalle Molle di studi sull'Intelligenza Artificiale (IDSIA), Scuola universitaria profes-
sionale della Svizzera italiana (SUPSI), Università della Svizzera italiana (USI)
Manno, Switzerland
E-mail:
giorgio@idsia.ch
alessio@idsia.ch

size¹ α of the test. Thus the t -test is not suitable for analyzing the cross-validation results (Dietterich, 1998; Nadeau & Bengio, 2003).

A suitable approach is instead the correlated² t -test (Nadeau & 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 & Bengio, 2003; Bouckaert, 2003).

Assume now that the two classifiers have assessed via cross-validation on a collection of data sets $\mathbf{D} = \{D_1, D_2, \dots, D_q\}$. One has to decide if the difference of accuracy between the two classifiers on the multiple data sets of \mathbf{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 \mathbf{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

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 greater than α .

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

56 the correlated t -test by Nadeau & Bengio (2003). Under the matching prior the
 57 inferences of the Bayesian correlated t -test and of the frequentist correlated t -test
 58 are numerically equivalent. The meaning of the inferences is however different. The
 59 inference of the frequentist test is a p -value; the inference of the Bayesian test is a
 60 posterior probability. The posterior probabilities computed on the individual data
 61 sets can be combined to make further Bayesian inference on multiple data sets.

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

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

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

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

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

97 **2 Inference from cross-validation results on a single data set**

98 2.1 Problem statement and frequentist tests

99 We want to statistically compare the accuracy of two classifiers which have been
 100 assessed via m runs of k -folds cross-validation. We provide both classifiers with

101 the same training and test sets and we compute the difference of accuracy be-
 102 tween the two classifiers on each test set. This yields the *differences of accuracy*
 103 $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, where $n = mk$. We denote the sample mean and the sample
 104 variance of the differences as \bar{x} and $\hat{\sigma}^2$.

105 A statistical test has to establish whether the mean difference between the two
 106 classifier is significantly different from zero, analyzing the vector of results \mathbf{x} . Such
 107 results are correlated because of the overlapping training sets. Nadeau & Bengio
 108 (2003) prove that there is no unbiased estimator of such correlation. They assume
 109 the correlation to be $\rho = \frac{n_{te}}{n}$, where n_{te} , n_{tr} and n_{tot} denote the size of the training
 110 set, of the test set and of the whole available data set. Thus $n_{tot} = n_{tr} + n_{te}$. The
 111 statistic of the correlated t -test is:

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

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

121 The denominator of the statistics is the *standard error*, namely the standard
 122 deviation of the estimate of \bar{x} . The standard error increases with $\hat{\sigma}^2$, which typically
 123 increases on smaller data sets. On the other hand the standard error decreases with
 124 $n = mk$. Previous studies (Kohavi, 1995) recommend to set the number of folds
 125 to $k=10$ to obtain a reliable estimate from cross-validation. This has become a
 126 standard choice. Having set $k=10$, one can further decrease the standard error
 127 of the test by increasing the number of runs m . Indeed Bouckaert (2003) and
 128 (Witten et al., 2011, Sec.5.3) recommend to perform $m=10$ runs of ten folds cross-
 129 validation.

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

136 2.2 Bayesian t-test for uncorrelated observations

Before introducing the Bayesian t-test for correlated observations, we briefly dis-
 cuss the Bayesian inference in the uncorrelated case. Assume we have a vec-
 tor of independent and identically distributed observations of a variable X , i.e.,
 $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, 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 unknown precision ν (the precision is the inverse of variance $\nu = 1/\sigma^2$):

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

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 & 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 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:

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

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

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

138 where $\mathcal{T}_{2a_n}(z)$ denotes the cumulative distribution of the standardized Student
 139 distribution with $2a_n$ degrees of freedom computed at z . By choosing $\alpha = 0.05$, we
 140 can assess the positivity of μ with posterior probability 0.95. If the prior parameters
 141 are set as follows: $\{\mu_0 = 0, k_0 \rightarrow \infty, a = -1/2, b = 0\}$, from Eqn.(3) it follows
 142 that $P(\mu > 0|\mathbf{x}, \mu_0, k_0, a, b) = 1 - p$, where p is the p -value of the frequentist t-test.
 143 See (Murphy, 2012, Chap.4) for further details on the correspondence between
 144 frequentist and Bayesian t -tests. In fact, for these values, the posterior reduces to
 145 $\text{St}(\mu; n - 1, \bar{x}, \sigma^2/n)$, as shown also in the second column in Table 1. Therefore,
 146 if we consider this matching (improper) prior, the Bayesian and frequentist t-test
 147 coincide.

148 **2.3 A novel Bayesian t-test for correlated observations**

149 Assume now that the observations of the variable X , $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, are
 150 identically distributed but dependent. In particular, consider the case in which the
 151 observations have the same mean μ , the same precision ν and are equally correlated
 152 with each other with correlation $\rho > 0$. This is for instance the case in which the

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

Table 1 Posterior parameters for the uncorrelated case.

153 n observations are the n differences of accuracy among two classifiers yielded by
 154 cross-validation. The data generating process can be modelled as follows:

$$\mathbf{x} = \mathbf{H}\mu + \mathbf{v} \quad (4)$$

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 $\Sigma_{n \times n}$ patterned as follows: each diagonal elements equals $\sigma^2 = 1/\nu$; each non-diagonal element equals $\rho\sigma^2$. This is the so-called *intra-class covariance matrix* (Press, 2012). We define $\Sigma = \sigma^2 M$, where M is the $(n \times n)$ correlation matrix. As an example, with $n = 3$ we have:

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix} \quad M = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \quad (5)$$

155 To allow for Σ to be invertible and positive definite, we require $\sigma^2 > 0$ and
 156 $0 \leq \rho < 1$. The correlation among the cross-validation results is positive anyway
 157 (Nadeau & Bengio, 2003). These two conditions define the *admissibility region* of
 158 the parameters.

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

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

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

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

168 The proof is given in Appendix. By computing the derivatives of the likelihood
 169 w.r.t. the parameters, it shows that the maximum likelihood estimate of μ, σ^2 is
 170 $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$ and, respectively, $\hat{\sigma}^2 = \text{tr}(M^{-1}\mathbf{Z})$, where $\mathbf{Z} = (\mathbf{x} - \mathbf{H}\hat{\mu})(\mathbf{x} - \mathbf{H}\hat{\mu})^T$.
 171 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.

Introducing the correlation heuristic

To enable inferences from correlated samples we renounce estimating ρ from data. We adopt instead the correlation heuristic of (Nadeau & 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 Eqn. (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:

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

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

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

Table 2 Posterior parameters for the correlated case

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

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

194 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,

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

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

202 3 Inference on multiple data sets

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

208 The signed-rank test assumes the \bar{x}_i 's to be i.i.d. and generated from a sym-
 209 metric distribution. The null hypothesis is that the median of the distribution is
 210 M . When the test accept the alternative hypothesis it claims that the median of
 211 the distribution is significantly different from M .

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

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

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

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

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

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

223 3.1 Poisson-binomial inference on multiple data sets

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

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

242 The cumulative distribution function of X is:

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

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

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

255 3.2 Example

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

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

	Datasets	μ_i	σ_i
Case 1	D_1, \dots, D_5	0.1	0.05
	D_6, \dots, D_{10}	-0.1	0.05
Case 2	D_1, \dots, D_5	0.1	0.05
	D_6, \dots, D_{10}	-0.1	0.15

Table 3 Example of comparison of two classifiers in multiple datasets

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

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

for $i = 1, \dots, 10$ (10-fold cross-validation) and for the $j = 1, \dots, 10$ 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 10-fold cross-validation of each dataset, while the Wilcoxon signed-rank test does not.

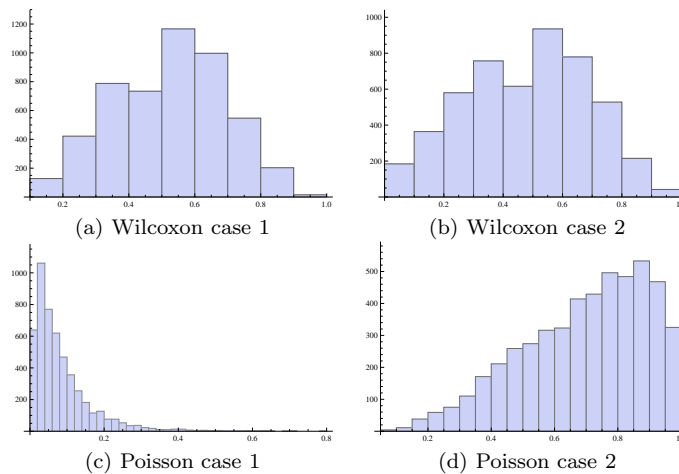


Fig. 1 Distribution of $P(X > q/2)$ for the Poisson test and distribution of the p -values for the Wilcoxon signed-rank test in the two cases

273 4 Experiments

274 The calibration and the power of the correlated t -test have been already extensively
 275 studied by (Nadeau & Bengio, 2003; Bouckaert, 2003) and we refrain from doing it

276 here. The same results apply to the Bayesian correlated t -test, since the frequentist
 277 and the Bayesian correlated t -test take the same decisions. The main result of such
 278 studies is that the rate of Type I errors of the correlated t -test is considerably closer
 279 to the nominal test size α than the rate of Type I error of the standard t -test. In
 280 the following we thus present results dealing with the inference on multiple data
 281 sets.

282 4.1 Two classifiers with known difference of accuracy

283 We generate the data sets sampling the instances from the Bayesian network $C \rightarrow$
 284 F , where C is the binary class with states $\{c_0, c_1\}$ and F is a binary feature with
 285 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
 286 $\theta > 0.5$. We refer to this model with exactly these parameters as BN.

287 Notice that if the BN model is used both to generate the instances and to
 288 issue the prediction, its expected accuracy is³ θ .

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

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

299 We perform the tests in a one-sided fashion: the null hypothesis is that zeroR is
 300 less or equally accurate than \hat{BN} . The alternative hypothesis is that \hat{BN} is more
 301 powerful than zeroR. We set the size of both the signed rank and the Poisson
 302 tests to $\alpha=0.05$. We measure the power of a test as the rate of rejection of the null
 303 hypothesis when $\delta_{acc} > 0$.

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

305 4.2 Fixed difference of accuracy on all data sets

306 As a first experiment, we set the actual difference of accuracy δ_{acc} among the two
 307 classifiers as identical on all the q data sets. We assume the availability of $q=50$
 308 data sets. This is a common size for a comparison of classifiers. We consider the
 309 following different values of δ_{acc} : $\{0, 0.01, 0.02, \dots, 0.1\}$.

310 For each value of δ_{acc} we repeat 5,000 experiments as follows. We allow the
 311 various data sets to have different size $\mathbf{s} = s_1, s_2, \dots, s_q$. We draw the sample size
 312 of each data set uniformly from $\mathcal{S} = \{25, 50, 100, 250, 500, 1000\}$. We generate each
 313 data set using the BN model; then we assess via cross-validation the accuracy of

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.

314 both zeroR and $\hat{B}N$. We then compare the two classifiers via the Poisson and the
 315 signed-rank test.

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

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

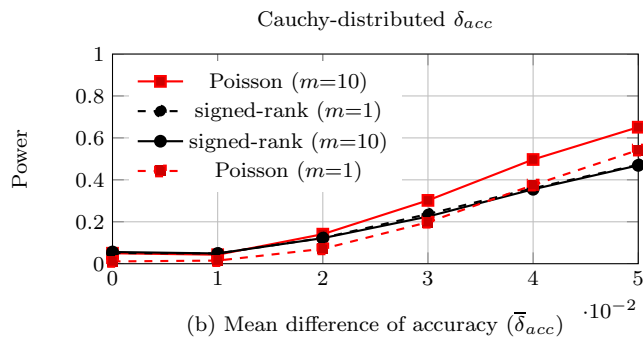
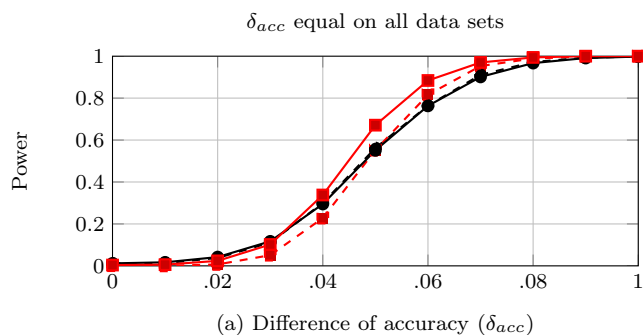


Fig. 2 Power and calibration of the tests over multiple data sets. The plots share the same legend. The Poisson test has squared marks. The signed-rank test has circle marks. Dashed lines refer to one run of cross-validation, solid lines refer to ten runs of cross-validation. The plots refer to the case $q=50$.

325 4.3 Difference of accuracy sampled from the Cauchy distributions

326 We remove the assumption of δ_{acc} being equal for all data sets. Instead for each
 327 data set we sample δ_{acc} from a Cauchy distribution. We set the median and the

328 scale parameter of the Cauchy to a value $\overline{\delta_{acc}} > 0$. A different value of $\overline{\delta_{acc}}$ de-
 329 fines a different experimental setting. We consider the following values of $\overline{\delta_{acc}}$:
 330 $\{0, 0.01, 0.02, \dots, 0.05\}$. We run 5,000 experiments for each value of $\overline{\delta_{acc}}$. We as-
 331 sume the availability of $q=50$ data sets.

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

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

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

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

345 4.4 Application to real data sets

346 We consider 54 data sets⁴ from the UCI repository. We consider five different
 347 classifiers: naive Bayes, averaged one-dependence estimator (AODE), hidden naive
 348 Bayes (HNB), J48 decision tree and J48 grafted (J48-gr). All the algorithms are
 349 described in (Witten et al., 2011). On each data set we run ten runs of ten-folds
 350 cross-validation using the WEKA⁵ software.

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

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

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

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

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

⁵ Available from <http://www.cs.waikato.ac.nz/ml/weka>.

<i>Data sets 1-27</i>					
	naive Bayes	J48	J48-gr	AODE	HNB
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
<i>Data sets 28-54</i>					
	naive Bayes	J48	J48-gr	AODE	HNB
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
<i>Data sets 1-54</i>					
	naive Bayes	J48	J48-gr	AODE	HNB
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

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

370 4.5 Software

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

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

385 5 Conclusions

386 To our knowledge, the Poisson test is the first test which compares two classifiers on
 387 multiple data sets accounting for the correlation and the uncertainty of the results
 388 generated by cross-validation on each individual data set. The test is usually more
 389 powerful than the signed-rank if ten runs of cross-validation are performed, which

390 is anyway common practice. A limit of the approach based on the Poisson-binomial
391 is that its inferences refer to the sample of provided data sets rather than to the
392 population from which the data sets have been drawn. A way to overcome this
393 limit could be the development a hierarchical test able to make inference on the
394 population of data sets.

395 References

- 396 Benavoli, A., Mangili, F., Corani, G., Zaffalon, M., and Ruggeri, F. A Bayesian
397 Wilcoxon signed-rank test based on the Dirichlet process. In *Proceedings of the*
398 *31st International Conference on Machine Learning (ICML 2014)*, pp. 1026–1034,
399 2014.
- 400 Bernardo, José M and Smith, Adrian FM. *Bayesian theory*, volume 405. Wiley
401 Chichester, 2009.
- 402 Bouckaert, Remco R. Choosing between two learning algorithms based on cal-
403 ibrated tests. In *Proceedings of the 20th International Conference on Machine*
404 *Learning (ICML-03)*, pp. 51–58, 2003.
- 405 Demšar, Janez. Statistical comparisons of classifiers over multiple data sets. *The*
406 *Journal of Machine Learning Research*, 7:1–30, 2006.
- 407 Dietterich, Thomas G. Approximate statistical tests for comparing supervised
408 classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.
- 409 Hong, Yili. On computing the distribution function for the Poisson binomial
410 distribution. *Computational Statistics & Data Analysis*, 59:41–51, 2013.
- 411 Kohavi, Ron. A study of cross-validation and bootstrap for accuracy estimation
412 and model selection. In *Proceedings of the 14th International Joint Conference*
413 *on Artificial intelligence-Volume 2*, pp. 1137–1143. Morgan Kaufmann Publishers
414 Inc., 1995.
- 415 Lacoste, Alexandre, Laviolette, François, and Marchand, Mario. Bayesian com-
416 parison of machine learning algorithms on single and multiple datasets. In
417 *Proc.of the Fifteenth International Conference on Artificial Intelligence and Statis-*
418 *tics (AISTATS-12)*, pp. 665–675, 2012.
- 419 Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- 420 Nadeau, Claude and Bengio, Yoshua. Inference for the generalization error. *Ma-*
421 *chine Learning*, 52(3):239–281, 2003.
- 422 Otero, José, Sánchez, Luciano, Couso, Inés, and Palacios, Ana. Bootstrap analysis
423 of multiple repetitions of experiments using an interval-valued multiple compar-
424 ison procedure. *Journal of Computer and System Sciences*, 80(1):88–100, 2014.
- 425 Press, S James. *Applied multivariate analysis: using Bayesian and frequentist methods*
426 *of inference*. Courier Dover Publications, 2012.
- 427 Witten, Ian H, Frank, Eibe, and Hall, Mark A. *Data mining: Practical machine*
428 *learning tools and techniques*. 2011.

429 **Appendix**430 **Proof of Theorem 1**

431 Preliminaries

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

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

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

We get:

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

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

433 Moreover,

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

434 Estimating μ

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:

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

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

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

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

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

438 Estimating σ^2

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

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

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

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

Finally, we get:

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

The product $(\mathbf{x} - \mathbf{H}\hat{\mu})_{1 \times n}^T \mathbf{M}_{n \times n}^{-1}(\mathbf{x} - \mathbf{H}\hat{\mu})_{n \times 1}$ 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:

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

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

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

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

443 Useful lemmas for estimating ρ

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

445 *Proof* Consider the i -th and the j -th ($i \neq j$) row of matrix $\mathbf{M}_{n \times n}$, containing elements
446 $\{m_{i1}, m_{i2}, \dots, m_{in}\}$ and $\{m_{j1}, m_{j2}, \dots, m_{jn}\}$ respectively. The value of $|\mathbf{M}|$ does not change
447 if we substitute each element of the i -th row as follows:

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

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 n -th row to the first row:

$$\mathbf{M} = \begin{vmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \dots & \dots & \dots & \dots & \dots \\ \rho & \rho & \rho & \dots & 1 \end{vmatrix}$$

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

Consider now matrix \mathbf{O} defined as follows:

$$\mathbf{O} = \begin{vmatrix} 1 & 1 & 1 & \dots & 1 \\ \rho & 1 & \rho & \dots & \rho \\ \dots & \dots & \dots & \dots & \dots \\ \rho & \rho & \rho & \dots & 1 \end{vmatrix}$$

Then, $|\mathbf{M}| = |\mathbf{N}| = (1 + (n-1)\rho) \cdot |\mathbf{O}|$. Consider now adding the elements of the first row of $|\mathbf{O}|$ to the second row of $|\mathbf{O}|$, using the scalar weight $b=-\rho$. Then add $-\rho$ times the first row to the third, to the fourth, ... to the n -th row of $|\mathbf{O}|$. This yields matrix \mathbf{P} , with $|\mathbf{P}| = |\mathbf{O}|$:

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

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

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

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

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

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

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

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 n -th row we get:

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

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

458 Estimating ρ

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

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

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

The derivative w.r.t. ρ is:

$$\begin{aligned} \left. \frac{\partial}{\partial \rho} L(\mu, \sigma^2, \rho) \right|_{\hat{\mu}, \hat{\sigma}^2} &= \frac{\partial}{\partial \rho} \left(-\frac{1}{2} \log((\hat{\sigma}^2)^n |\mathbf{M}|) \right) = \frac{\partial}{\partial \rho} \left(-\frac{1}{2} \log \left[\left(\frac{\text{Tr}(\mathbf{M}^{-1} \mathbf{Z})}{n} \right)^n |\mathbf{M}| \right] \right) \\ &= \frac{\partial}{\partial \rho} \left(-\frac{1}{2} n \log \left(\frac{\text{Tr}(\mathbf{M}^{-1} \mathbf{Z})}{n} \right) - \frac{1}{2} \log |\mathbf{M}| \right) \\ &= \frac{\partial}{\partial \rho} \left(-\frac{1}{2} n \log (\text{Tr}(\mathbf{M}^{-1} \mathbf{Z})) + \frac{1}{2} n \log(n) - \frac{1}{2} \log |\mathbf{M}| \right) \\ &= -\frac{1}{2} \frac{\partial}{\partial \rho} (n \log (\text{Tr}(\mathbf{M}^{-1} \mathbf{Z}))) - \frac{1}{2} \frac{\partial}{\partial \rho} (\log |\mathbf{M}|) \\ &= -\frac{1}{2} \frac{n}{\text{Tr}(\mathbf{M}^{-1} \mathbf{Z})} \frac{\partial}{\partial \rho} (\text{Tr}(\mathbf{M}^{-1} \mathbf{Z})) - \frac{1}{2} \frac{1}{|\mathbf{M}|} \frac{\partial}{\partial \rho} (|\mathbf{M}|) \end{aligned}$$

460 Let us now consider $\text{Tr}(\mathbf{M}^{-1} \mathbf{Z}) = \frac{1}{|\mathbf{M}|} \text{Tr}(\text{Adj}(\mathbf{M}) \mathbf{Z})$ and define $\mathbf{S} = \text{Adj}(\mathbf{M}) \mathbf{Z}$. Let us
461 denote the difference between an observation and the maximum likelihood mean as $\delta_i = x_i - \hat{\mu}_i$.
462 The i -th diagonal element of \mathbf{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:

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

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

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

464 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:

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

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

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

it can easily be shown that

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

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 $\hat{\mu}, \hat{\sigma}^2$ is negative definite. In fact, we have that

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

465 and $\frac{\partial^2}{\partial \sigma^2 \partial \mu} L(\mu, \sigma^2, \rho) \Big|_{\hat{\mu}, \hat{\sigma}^2} = 0$. Thus, $\hat{\mu}, \hat{\sigma}^2, \hat{\rho}$ is the maximum likelihood estimator. Since
 466 $\hat{\rho} = 0$, this estimator is not consistent whenever the true correlation is not zero (strictly
 467 positive).

468 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:

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

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

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

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

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

Let us define the posterior mean

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

then

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

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

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

and

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

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

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

Hence, it follows that the posterior is

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

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

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

471 Proof of Corollary 1

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

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

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

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

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

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

472 which ends the proof.