Inference for the Generalization Error

Claude Nadeau
Health Canada, AL0900B1
Ottawa ON, Canada K1A 0L2
jcnadeau@altavista.net
Yoshua Bengio
CIRANO and *Dept. IRO, Université de Montréal*
C.P. 6128 Succ. Centre-Ville, Montréal, Quebec, Canada H3C 3J7
bengioy@iro.umontreal.ca

April 5, 2001

Abstract

In order to compare learning algorithms, experimental results reported in the machine learning literature often use statistical tests of significance to support the claim that a new learning algorithm generalizes better. Such tests should take into account the **variability due to the choice of training set** and not only that due to the test examples, as is often the case. This could lead to gross underestimation of the variance of the cross-validation estimator, and to the wrong conclusion that the new algorithm is significantly better when it is not. We perform a theoretical investigation of the variance of a cross-validation estimator of the generalization error that takes into account the variability due to the randomness of the *training set* as well as *test examples*. Our analysis shows that all the variance estimators that are based only on the results of the cross-validation experiment must be biased. This analysis allows us to propose new estimators of this variance. We show, via simulations, that tests of hypothesis about the generalization error using those new variance estimators have better properties than tests involving variance estimators currently in use and listed in (Dietterich, 1998). In particular, the new tests have correct size and good power. That is, the new tests do not reject the null hypothesis too often when the hypothesis is true, but they tend to frequently reject the null hypothesis when the latter is false.

1 Generalization Error and its Estimation

In order to compare learning algorithms, experimental results reported in the machine learning literature often use statistical tests of significance. Unfortunately, these tests often rely solely on the variability due to the test examples and do not take into account the **variability due to the randomness of the training set**. We perform a theoretical investigation of the variance of a cross-validation estimator of the generalization error that takes into account the variability due to the choice of training sets as well as of test examples (hold-out set). When applying a learning algorithm (or comparing several algorithms), one is typically interested in estimating its generalization error. Its estimation is rather trivial through cross-validation or the bootstrap. Providing a variance estimate of the cross-validation estimator, so that hypothesis testing and/or confidence intervals are possible, is more difficult, especially, as pointed out in (Hinton, Neal, Tibshirani, & DELVE team members, 1995), if one wants to take into account various sources of variability such as the choice of the training set (Breiman, 1996) or initial conditions of a learning algorithm (Kolen & Pollack, 1991). A notable effort in that direction is Dietterich's work (Dietterich, 1998). See also the review of bounds of the accuracy of various cross-validation estimates in (Devroye, Gyrfi, & Lugosi, 1996). Building upon (Dietterich, 1998), in this paper we take into account the variability due to the choice of training sets and test examples. Specifically, an investigation of the variance to be estimated allows us to provide two new variance estimators, one of which is conservative by construction.

The choice of estimator for the variance of an average test error (or of the difference between the average error made by two algorithms) is very important: a poor choice of estimator (especially if it is **liberal**, i.e. underestimates the variance) could lead to a profusion of publications in which method A is incorrectly claimed to be better than a previously proposed method B. Because of the approximately normal behavior of average error, an underestimation of the standard deviation by a factor 2, say, can yield to about 6 times more "false claims" than would have been expected for a test level of 5%. If the habit of making rigorous statistical comparisons and in particular avoiding the use of a liberal estimator is not ingrained in the machine learning community, it could be tempting for many researchers to use a liberal estimate of variance when comparing their preferred algorithm against the competition. For this reason, it is very important that reviewers insist on analyses of results that avoid liberal estimators of variance (for confidence intervals or to test the null hypothesis of method A being not better than method B).

Let us define what we mean by "generalization error" and say how it will be estimated in this paper. We assume that data is available in the form $Z_1^n = \{Z_1, \ldots, Z_n\}$. For example, in the case of supervised learning, $Z_i = (X_i, Y_i) \in \mathcal{Z} \subseteq \mathbb{R}^{p+q}$, where p and q denote the dimensions of the X_i's (inputs) and the Y_i's (outputs). We also assume that the Z_i's are independent with $Z_i \sim P(Z)$, where the generating distribution P is unknown. Let $\mathcal{L}(D; Z)$, where D represents a subset of size $n_1 \leq n$ taken from Z_1^n, be a function from $\mathcal{Z}^{n_1} \times \mathcal{Z}$ to $\mathbb{R}$. For instance, this function could be the loss incurred by the decision that a learning algorithm trained on D makes on a new example Z.

We are interested in estimating ${}_n\mu \equiv E[\mathcal{L}(Z_1^n; Z_{n+1})]$ where $Z_{n+1} \sim P(Z)$ is independent of Z_1^n. The subscript $_n$ stands for the size of the training set (Z_1^n here). Note that the above expectation is taken over Z_1^n and Z_{n+1}, meaning that we are interested in the performance of an algorithm rather than the performance of the specific decision function it yields on the data at hand. Dietterich (Dietterich, 1998) has introduced a taxonomy of statistical questions in Machine Learning, which we briefly summarize here. At the top level is whether the questions refer to single or multiple domains (type 9). For single domains, Dietterich distinguishes between the analysis of a single (given or already trained) predictor (types 1 through 4) and the analysis of a learning algorithm that can yield such predictors given a training set (types 5 through 8). For the former, the training set is considered fixed, whereas for the latter the training set is considered to be random. In this paper, we are concerned with the analysis of the performance of *learning algorithms* (types 5 through 8), not of particular *trained predictors*. Dietterich further splits types 5 through 8 according to whether the sample size is large or not and whether one is interested in the generalization error of a single algorithm or wants to compare the generalization errors of various algorithms.

Let us now introduce some notation and definitions. We shall call $_n\mu$ the generalization error even though it can go beyond that as we now illustrate. Here are two examples.

- **Generalization error**

  We may take as our basic measurement

  $$\mathcal{L}(D;Z) = \mathcal{L}(D;(X,Y)) = Q(F(D)(X),Y), \tag{1}$$

  where F represents a learning algorithm that yields a function $f = F(D)$ ($F(D) : \mathbb{R}^p \to \mathbb{R}^q$), when training the algorithm on D, and Q is a loss function measuring the inaccuracy of a decision $f(X)$ when Y is observed. For instance, for classification problems, we could have

  $$Q(\hat{y},y) = I[\hat{y} \neq y], \tag{2}$$

  where $I[\ ]$ is the indicator function, and in the case of regression,

  $$Q(\hat{y},y) = \parallel \hat{y} - y \parallel^2, \tag{3}$$

  where $\parallel \cdot \parallel$ is the Euclidean norm. In that case $_n\mu = E[\mathcal{L}(Z_1^n, Z_{n+1})]$ is the generalization error of algorithm F on data sampled from P.

- **Comparison of generalization errors**

  Sometimes, what we are interested in is not the performance of algorithms *per se*, but how two algorithms compare with each other. In that case we may want to consider

  $$\mathcal{L}(D;Z) = \mathcal{L}(D;(X,Y)) = Q(F_A(D)(X),Y) - Q(F_B(D)(X),Y), \tag{4}$$

  where $F_A(D)$ and $F_B(D)$ are decision functions obtained when training two algorithms (respectively A and B) on D, and Q is a loss function. In this case $_n\mu$ would be a difference of generalization errors.

The generalization error is often estimated via some form of cross-validation. Since there are various versions of the latter, we lay out the specific form we use in this paper.

- Let S_j be a random set of n_1 distinct integers from $\{1, \ldots, n\}(n_1 < n)$. Here n_1 is not random and represents the size of a training set. We shall let $n_2 = n - n_1$ be the size of the corresponding test set (or hold-out set).
- Let $S_1, \ldots S_J$ be such random index sets (of size n_1), sampled independently of each other, and let $S_j^c = \{1, \ldots, n\} \setminus S_j$ denote the complement of S_j.
- Let $Z_{S_j} = \{Z_i | i \in S_j\}$ be the training set obtained by subsampling Z_1^n according to the random index set S_j. The corresponding test set (or hold-out set) is $Z_{S_j^c} = \{Z_i | i \in S_j^c\}$.
- Let $L(j, i) = \mathcal{L}(Z_{S_j}; Z_i)$. According to (1), this could be the error an algorithm trained on the training set Z_{S_j} makes on example Z_i. According to (4), this could be the difference of such errors for two different algorithms.
- Let $\hat{\mu}_j = \frac{1}{n_2} \sum_{i \in S_j^c} L(j, i)$ denote the usual "average test error" measured on the test set $Z_{S_j^c}$.

Then the cross-validation estimate of the generalization error considered in this paper is

$$ {}^{n_2}_{n_1}\hat{\mu}_J = \frac{1}{J} \sum_{j=1}^{J} \hat{\mu}_j. \tag{5} $$

Note that this an unbiased estimator of ${}_{n_1}\mu = E[\mathcal{L}(Z_1^{n_1}, Z_{n+1})]$, which is not quite the same as ${}_{n}\mu$.

This paper is about the estimation of the variance of the above cross-validation estimator. There are many variants of cross-validation, and the above variant is close to the popular K-fold cross-validation estimator, which has been found more reliable than the leave-one-out estimator (Kohavi, 1995). It should be noted that our goal in this paper is not to compare algorithms in order to perform *model selection* (i.e. to choose exactly one among several learning algorithms for a particular task, given a data set on which to train them). The use of cross-validation estimators for model selection has sparked a debate in the last few years (Zhu & Rohwer, 1996; Goutte, 1997) related to the "no free lunch theorem" (Wolpert & Macready, 1995), since cross-validation model selection often works well in practice but it is probably not a universally good procedure.

This paper does not address the issue of model selection but rather that of estimating the uncertainty in a cross-validation type of estimator for generalization error, namely ${}^{n_2}_{n_1}\hat{\mu}_J$. For this purpose, this paper studies estimators of the *variance* of ${}^{n_2}_{n_1}\hat{\mu}_J$ (in the sense that different values of ${}^{n_2}_{n_1}\hat{\mu}_J$ would have been obtained if a different data set Z_1^n had been sampled from the same unknown underlying distribution P and different random index sets S_j's had been generated). The application of the estimators studied in this paper may be for example (1) to provide confidence intervals around estimated generalization error, or (2) to perform a hypothesis test in order to determine whether an algorithm's estimated performance is significantly above or below the performance of another algorithm. The latter is very important when researchers introduce a new learning algorithm and they want to show that it brings a significant improvement with respect to previously known algorithms.

We first study theoretically the variance of ${}^{n_2}_{n_1}\hat{\mu}_J$ in Section 2. This will lead us to two new variance estimators we develop in Section 3. Section 4 shows how to test hypotheses or construct confidence intervals. Section 5 describes a simulation study we performed to see how the proposed statistics behave compared to statistics already in use. Section 6 concludes the paper. Before proceeding with the rest of the paper, some readers may prefer to read Appendix A.0 that presents some statistical prerequisites relevant to the rest of the paper.

2 Analysis of $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$

In this section, we study $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ and discuss the difficulty of estimating it. This section is important as it enables us to understand why some inference procedures about ${}_{n_1}\mu$ presently in use are inadequate, as we shall underline in Section 4. This investigation also enables us to develop estimators of $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ in Section 3. Before we proceed, we state a lemma that will prove useful in this section, and later ones as well.

Lemma 1 *Let $U_1, \ldots, U_K$ be random variables with common mean β and the following covariance structure*

$$Var[U_k] = \delta,\ \forall k \qquad Cov[U_k, U_{k'}] = \gamma,\ \forall k \neq k'.$$

Let $\pi = \frac{\gamma}{\delta}$ be the correlation between U_k and $U_{k'}$ ($k \neq k'$). Let $\bar{U} = k^{-1}\sum_{k=1}^{K} U_k$ and $S_U^2 = \frac{1}{K-1}\sum_{k=1}^{K}(U_k - \bar{U})^2$ be the sample mean and sample variance respectively. Then

1. $Var[\bar{U}] = \gamma + \frac{(\delta-\gamma)}{K} = \delta\left(\pi + \frac{1-\pi}{K}\right)$.
2. *If the stated covariance structure holds for all K (with γ and δ not depending on K), then*
    - $\gamma \geq 0$,
    - $lim_{K\rightarrow\infty} Var[\bar{U}] = 0 \Leftrightarrow \gamma = 0$.
3. $E[S_U^2] = \delta - \gamma$.

Proof

1. *This result is obtained from a standard development of $Var[\bar{U}]$.*

2. *If $\gamma < 0$, then $Var[\bar{U}]$ would eventually become negative as K is increased. We thus conclude that $\gamma \geq 0$. From item 1, it is obvious that $Var[\bar{U}]$ goes to zero as K goes to infinity if and only if $\gamma = 0$.*

3. *Again, this only requires careful development of the expectation. The task is somewhat easier if one uses the identity*

$$S_U^2 = \frac{1}{K-1}\sum_{k=1}^{K}(U_k^2 - \bar{U}^2) = \frac{1}{2K(K-1)}\sum_{k=1}^{K}\sum_{k'=1}^{K}(U_k - U_{k'})^2.$$

Although we only need it in Section 4, it is natural to introduce a second lemma here as it is a continuation of Lemma 1.

Lemma 2 *Let $U_1, \ldots, U_K, U_{K+1}$ be random variables with mean, variance and covariance as described in Lemma 1. In addition, assume that the vector $(U_1, \ldots, U_K, U_{K+1})$ follows the multivariate Gaussian distribution. Again, let $\bar{U} = K^{-1}\sum_{k=1}^{K} U_k$ and $S_U^2 = \frac{1}{K-1}\sum_{k=1}^{K}(U_k - \bar{U})^2$ be respectively the sample mean and sample variance of $U_1, \ldots, U_K$. Then*

1. $\sqrt{1-\pi}\ \frac{U_{K+1}-\beta}{\sqrt{S_U^2}} \sim t_{K-1}$,

2. $\sqrt{\frac{1-\pi}{1+(K-1)\pi}}\ \frac{\sqrt{K}(\bar{U}-\beta)}{\sqrt{S_U^2}} \sim t_{K-1}$,

where $\pi = \frac{\gamma}{\delta}$ as in Lemma 1, and t_{K-1} refers to Student's t distribution with $(K-1)$ degrees of freedom.
Proof *See Appendix A.1.*

To study $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ we need to define the following covariances. In the following, S_j and $S_{j'}$ are independent random index sets, each consisting of n_1 distinct integers from $\{1, \ldots, n\}$. Also, expectations are totally unconditional, that is expectations (as well as variances and covariances) are taken over Z_1^n, S_j, $S_{j'}$, i and i'.

- Let $\sigma_0 = \sigma_0(n_1) = Var[L(j,i)]$ when i is randomly drawn from S_j^c. To establish that σ_0 does not depend on n_2 we note that $Var[L(j,i)] = E_{S_j,i}[Var_{Z_1^n}[\mathcal{L}(Z_{S_j}; Z_i)|S_j, i]] + Var_{S_j,i}[E_{Z_1^n}[\mathcal{L}(Z_{S_j}; Z_i)|S_j, i]]$. Now the distribution of $\mathcal{L}(Z_{S_j}; Z_i)$ does not depend on the particular realization of S_j and i, it is just the distribution of $\mathcal{L}(Z_1^{n_1}; Z_{n_1+1})$. Thus $\sigma_0 = E_{S_j,i}[Var[\mathcal{L}(Z_1^{n_1}; Z_{n_1+1})]] + Var_{S_j,i}[\,{}_{n_1}\mu] = Var[\mathcal{L}(Z_1^{n_1}; Z_{n_1+1})]$ depends only on n_1, not on n_2.

- Let $\sigma_1 = \sigma_1(n_1, n_2) = Var[\hat{\mu}_j]$.

- Let $\sigma_2 = \sigma_2(n_1, n_2) = Cov[L(j,i), L(j',i')]$, with $j \neq j'$, i and i' randomly and independently drawn from S_j^c and $S_{j'}^c$ respectively.

- Let $\sigma_3 = \sigma_3(n_1) = Cov[L(j,i), L(j,i')]$ for $i, i' \in S_j^c$ and $i \neq i'$, that is i and i' are sampled without replacement from S_j^c. Using a similar argument as for σ_0 allows one to show that σ_3 does not depend on n_2.

Let us look at the mean and variance of $\hat{\mu}_j$ (i.e., over one set) and ${}^{n_2}_{n_1}\hat{\mu}_J$ (i.e. over J sets). Concerning expectations, we obviously have $E[\hat{\mu}_j] = {}_{n_1}\mu$ and thus $E[{}^{n_2}_{n_1}\hat{\mu}_J] = {}_{n_1}\mu$. From Lemma 1, we have

$$\sigma_1 = \sigma_1(n_1, n_2) = Var[\hat{\mu}_j] = \sigma_3 + \frac{\sigma_0 - \sigma_3}{n_2} = \frac{(n_2 - 1)\sigma_3 + \sigma_0}{n_2}. \tag{6}$$

For $j \neq j'$, we have

$$Cov[\hat{\mu}_j, \hat{\mu}_{j'}] = \frac{1}{n_2^2} \sum_{i \in S_j^c} \sum_{i' \in S_{j'}^c} Cov[L(j,i), L(j',i')] = \sigma_2, \tag{7}$$

and therefore (using Lemma 1 again)

$$Var[{}^{n_2}_{n_1}\hat{\mu}_J] = \sigma_2 + \frac{\sigma_1 - \sigma_2}{J} = \sigma_1 \left(\rho + \frac{1-\rho}{J}\right) = \sigma_2 + \frac{\sigma_3 - \sigma_2}{J} + \frac{\sigma_0 - \sigma_3}{n_2 J}, \tag{8}$$

where $\rho = \frac{\sigma_2}{\sigma_1} = corr[\hat{\mu}_j, \hat{\mu}_{j'}]$. Asking how to choose J amounts to asking how large is ρ. If it is large, then taking $J > 1$ (rather than $J = 1$) does not provide much improvement in the estimation of ${}_{n_1}\mu$. We provide some guidance on the choice of J in Section 5.

Equation (8) lends itself to an interesting interpretation. First we get that $\sigma_2 = Var[{}^{n_2}_{n_1}\hat{\mu}_\infty]$ with

$${}^{n_2}_{n_1}\hat{\mu}_\infty = \lim_{J \to \infty} {}^{n_2}_{n_1}\hat{\mu}_J = \lim_{J \to \infty} \frac{1}{J} \sum_{j=1}^{J} \hat{\mu}_j = \frac{1}{\binom{n}{n_1} n_2} \sum_{s \in C(\{1,\ldots,n\}, n_1)} \sum_{i \in \{1,\ldots,n\} \setminus s} \mathcal{L}(Z_s; Z_i),$$

where $C(\{1, \ldots, n\}, n_1)$, as defined in Appendix A.2, is the set of all possible subsets of n_1 distinct integers from $\{1, \ldots, n\}$. We justify the last equality as follows. What happens when J goes to infinity is that all possible errors (there are $\binom{n}{n_1} n_2$ different ways to choose a training set and a test example) appear with relative frequency $\frac{1}{\binom{n}{n_1} n_2}$. In other words, ${}^{n_2}_{n_1}\hat{\mu}_\infty$ is like ${}^{n_2}_{n_1}\hat{\mu}_J$ except that all $\binom{n}{n_1}$ possible training sets are chosen exactly once. Briefly, sampling infinitely often with replacement is equivalent to sampling exhaustively without replacement (i.e. a census). We also have ${}^{n_2}_{n_1}\hat{\mu}_\infty = E_{S_j}[\hat{\mu}_j | Z_1^n]$ $\forall j$ and therefore ${}^{n_2}_{n_1}\hat{\mu}_\infty = E[{}^{n_2}_{n_1}\hat{\mu}_J | Z_1^n]$. Thus $\sigma_2 = Var[E[{}^{n_2}_{n_1}\hat{\mu}_J | Z_1^n]]$ so that we must have $E[Var[{}^{n_2}_{n_1}\hat{\mu}_J | Z_1^n]] = \frac{\sigma_1 - \sigma_2}{J}$.

We shall often encounter $\sigma_0, \sigma_1, \sigma_2$ and σ_3 in the future, so some knowledge about those quantities is valuable. Here's what we can say about them.

Proposition 1 *For given n_1 and n_2, we have $0 \leq \sigma_2 \leq \sigma_1 \leq \sigma_0$ and $0 \leq \sigma_3 \leq \sigma_1$.*

Proof *For $j \neq j'$ we have*

$$\sigma_2 = Cov[\hat{\mu}_j, \hat{\mu}_{j'}] \leq \sqrt{Var[\hat{\mu}_j]\, Var[\hat{\mu}_{j'}]} = \sigma_1.$$

Since $\sigma_0 = Var[L(j,i)], i \in S_j^c$ and $\hat{\mu}_j$ is the mean of the $L(j,i)$'s, then $\sigma_1 = Var[\hat{\mu}_j] \leq Var[L(j,i)] = \sigma_0$. The fact that $\lim_{J\to\infty} Var[\,{}_{n_1}^{n_2}\hat{\mu}_J] = \sigma_2$ provides the inequality $0 \leq \sigma_2$.

Regarding σ_3, we deduce $\sigma_3 \leq \sigma_1$ from (6) while $0 \leq \sigma_3$ is derived from the fact that $\lim_{n_2\to\infty} Var[\hat{\mu}_j] = \sigma_3$.

Naturally the inequalities are strict provided $L(j,i)$ is not perfectly correlated with $L(j,i')$, $\hat{\mu}_j$ is not perfectly correlated with $\hat{\mu}_{j'}$, and the variances used in the proof are positive.

A natural question about the estimator ${}_{n_1}^{n_2}\hat{\mu}_J$ is how n_1, n_2 and J affect its variance.

Proposition 2 *The variance of ${}_{n_1}^{n_2}\hat{\mu}_J$ is non-increasing in J and n_2.*
Proof

- *$Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$ is non-increasing (decreasing actually, unless $\sigma_1 = \sigma_2$) in J as obviously seen from (8). This means that averaging over many train/test improves the estimation of ${}_{n_1}\mu$.*
- *From (8), we see that to show that $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$ is non-increasing in n_2, it is sufficient to show that σ_1 and σ_2 are non-increasing in n_2. For σ_1, this follows from (6) and Proposition 3. Regarding σ_2, we show in Appendix A.2 that it is non-increasing in n_2. All this to say that for a given n_1, the larger the test set size, the better the estimation of ${}_{n_1}\mu$.*

The behavior of $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$ with respect to n_1 is unclear, but we conjecture as follows.

Conjecture 1 *In most situations, $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$ should decrease in n_1.*
Argument [1] *The variability in ${}_{n_1}^{n_2}\hat{\mu}_J$ comes from two sources: sampling decision rules (training process) and sampling testing examples. Holding n_2 and J fixed freezes the second source of variation as it solely depends on those two quantities, not n_1. The problem to solve becomes: how does n_1 affect the first source of variation? It is not unreasonable to expect that the decision function yielded by a "stable" learning algorithm is less variable when the training set is larger. See (Kearns & Ron, 1997) showing that for a large class of algorithms including those minimizing training error, cross-validation estimators are not much worse than the training error estimator (which itself improves in $O(VCdim/n_1)$ as the size of the training set increases (Vapnik, 1982)). Therefore we conclude that, for many cases of interest, the first source of variation, and thus the total variation (that is $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$) is decreasing in n_1.*

Regarding the estimation of $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$, we show below that we can easily estimate unbiasedly $(\sigma_1 - \sigma_2)$, $(\sigma_0 - \sigma_3)$ and $(\sigma_2 + ({}_{n_1}\mu)^2)$.

[1] Here we are not trying to prove the conjecture but to justify our intuition that it is correct.

- From Lemma 1, we obtain readily that the sample variance of the $\hat{\mu}_j$'s (call it $S^2_{\hat{\mu}_j}$ as in equation (9)) is an unbiased estimate of $\sigma_1 - \sigma_2 = \sigma_3 - \sigma_2 + \frac{\sigma_0 - \sigma_3}{n_2}$. Let us interpret this result. Given Z_1^n, the $\hat{\mu}_j$'s are J independent draws (with replacement) from a hat containing all $\binom{n}{n_1}$ possible values of the $\hat{\mu}_j$'s. The sample variance of those J observations ($S^2_{\hat{\mu}_j}$) is therefore an unbiased estimator of the variance of $\hat{\mu}_j$, given Z_1^n, i.e. an unbiased estimator of $Var[\hat{\mu}_j|Z_1^n]$, not $Var[\hat{\mu}_j]$. This permits an alternative derivation of the expectation of the sample variance. Indeed, we have

$$\begin{aligned} E_{Z_1^n, S_\cdot}[S^2_{\hat{\mu}_j}] &= E_{Z_1^n}[E_{S_\cdot}[S^2_{\hat{\mu}_j}|Z_1^n]] = E_{Z_1^n}[Var_{S_j}[\hat{\mu}_j|Z_1^n]] \\ &= Var_{Z_1^n, S_j}[\hat{\mu}_j] - Var_{Z_1^n}[E_{S_j}[\hat{\mu}_j|Z_1^n]] = \sigma_1 - Var_{Z_1^n}[\,{}^{n_2}_{n_1}\hat{\mu}_\infty] = \sigma_1 - \sigma_2, \end{aligned}$$

  where $S_\cdot$ denotes the random index sets $(S_1, \ldots, S_J)$. Note that $E[\hat{\mu}_j|Z_1^n] = {}^{n_2}_{n_1}\hat{\mu}_\infty$ and $Var[\,{}^{n_2}_{n_1}\hat{\mu}_\infty] = \sigma_2$ both come from the discussion after equation (8).

- For a given j, the sample variance of the $L(j,i)$'s ($i \in S_j^c$) is unbiased for $\sigma_0 - \sigma_3$ according to Lemma 1 again. We may average these sample variances over j to obtain a more accurate estimate of $\sigma_0 - \sigma_3$.

- From equation (7) we have $E[\hat{\mu}_j\hat{\mu}_{j'}] = \sigma_2 + (\,_{n_1}\mu)^2$ for $j \neq j'$, so the sample average of the $\hat{\mu}_j\hat{\mu}_{j'}$ will be unbiased for $(\sigma_2 + (\,_{n_1}\mu)^2)$.

Since we can estimate $(\sigma_1 - \sigma_2)$, $(\sigma_0 - \sigma_3)$ and $(\sigma_2 + (\,_{n_1}\mu)^2)$ without bias, we are thus able to estimate unbiasedly any linear combination of $(\sigma_0 - \sigma_3)$, $(\sigma_3 - \sigma_2)$ and $(\sigma_2 + (\,_{n_1}\mu)^2)$. This is not sufficient to estimate $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ shown in (8) unbiasedly. We now tackle the question of whether or not there exists an unbiased estimator of $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$. Potential estimators may be put in two classes: (i) those that are linear and/or quadratic in the $L(j,i)$'s, (ii) those that are not. Because of the general framework of the paper, it is impossible to say anything about the distribution of the $L(j,i)$'s beyond their means and covariances (to say anything more requires assumptions about the distribution of Z_1^n, the learning algorithms and the loss function $\mathcal{L}$). Hence we are only able to derive mathematical expectations for estimators within class (i). We obviously cannot identify an unbiased estimator of $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ in class (ii) since we cannot derive expectations in this class. The following proposition shows that there is no unbiased estimator of $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ in class (i).

Proposition 3 *There is no general unbiased estimator of $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ that involves the $L(j,i)$'s in a quadratic and/or linear way.*

Proof *Let $\vec{L}_j$ be the vector of the $L(j,i)$'s involved in $\hat{\mu}_j$ and $\vec{L}$ be the vector obtained by concatenating the $\vec{L}_j$'s; $\vec{L}$ is thus a vector of length $n_2 J$. We know that $\vec{L}$ has expectation $_{n_1}\mu \mathbf{1}_{n_2 J}$ and variance*

$$Var[\vec{L}] = \sigma_2 \mathbf{1}_{n_2 J}\mathbf{1}'_{n_2 J} + (\sigma_3 - \sigma_2) I_J \otimes (\mathbf{1}_{n_2}\mathbf{1}'_{n_2}) + (\sigma_0 - \sigma_3) I_{n_2 J},$$

where I_k is the identity matrix of order k, $\mathbf{1}_k$ is the $k \times 1$ vector filled with 1's and $\otimes$ denotes

Kronecker's product. We consider estimators of $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$ *of the following form*

$$\hat{V}[\,{}_{n_1}^{n_2}\hat{\mu}_J] = \vec{L}'A\vec{L} + b'\vec{L}$$

Using the fact that $\mathbf{1}_{n_2J}'A\mathbf{1}_{n_2J} = trace(A\mathbf{1}_{n_2J}\mathbf{1}_{n_2J}')$*, we have*

$$\begin{aligned} E[\hat{V}[\,{}_{n_1}^{n_2}\hat{\mu}_J]] &= trace(A\,Var[\vec{L}]) + E[\vec{L}]'AE[\vec{L}] + b'E[\vec{L}] \\ &= (\sigma_3 - \sigma_2)trace(A(I_J \otimes (\mathbf{1}_{n_2}\mathbf{1}_{n_2}'))) + (\sigma_0 - \sigma_3)trace(A) \\ &+ (\sigma_2 + {}_{n_1}\mu^2)\mathbf{1}_{n_2J}'A\mathbf{1}_{n_2J} + {}_{n_1}\mu b'\mathbf{1}_{n_2J}. \end{aligned}$$

Since we wish $\hat{V}[\,{}_{n_1}^{n_2}\hat{\mu}_J]$ *to be unbiased for* $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$*, we want* $0 = b'\mathbf{1}_{n_2J} = \mathbf{1}_{n_2J}'A\mathbf{1}_{n_2J}$ *to get rid of* ${}_{n_1}\mu$ *in the above expectation. Once those restrictions are incorporated into* $\hat{V}[\,{}_{n_1}^{n_2}\hat{\mu}_J]$*, we have*

$$E[\hat{V}[\,{}_{n_1}^{n_2}\hat{\mu}_J]] = (\sigma_3 - \sigma_2)trace(A(I_J \otimes (\mathbf{1}_{n_2}\mathbf{1}_{n_2}'))) + (\sigma_0 - \sigma_3)trace(A).$$

Since $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$ *is not a linear combination of* $(\sigma_3 - \sigma_2)$ *and* $(\sigma_0 - \sigma_3)$ *alone, we conclude that* $\hat{V}[\,{}_{n_1}^{n_2}\hat{\mu}_J]$ *cannot be unbiased for* $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$ *in general.*

3 Estimation of $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$

We are interested in estimating ${}_{n_1}^{n_2}\sigma_J^2 = Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$ where ${}_{n_1}^{n_2}\hat{\mu}_J$ is as defined in (5). We provide two new estimators of $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$ that shall be compared, in Section 5, to estimators currently in use and presented in Section 4. The first estimator is simple but may have a positive or negative bias for the actual variance $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$. The second estimator is meant to lead to conservative inference (see Appendix A.0), that is, if our Conjecture 1 is correct, its expected value exceeds the actual variance $Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]$.

3.1 First Method: Approximating ρ

Let us recall from (5) that ${}_{n_1}^{n_2}\hat{\mu}_J = \frac{1}{J}\sum_{j=1}^J \hat{\mu}_j$. Let

$$S_{\hat{\mu}_j}^2 = \frac{1}{J-1}\sum_{j=1}^J (\hat{\mu}_j - {}_{n_1}^{n_2}\hat{\mu}_J)^2 \tag{9}$$

be the sample variance of the $\hat{\mu}_j$'s. According to Lemma 1,

$$E[S_{\hat{\mu}_j}^2] = \sigma_1(1-\rho) = \frac{1-\rho}{\rho + \frac{1-\rho}{J}}\sigma_1\left(\rho + \frac{1-\rho}{J}\right) = \frac{\sigma_1\left(\rho + \frac{1-\rho}{J}\right)}{\frac{1}{J} + \frac{\rho}{1-\rho}} = \frac{Var[\,{}_{n_1}^{n_2}\hat{\mu}_J]}{\frac{1}{J} + \frac{\rho}{1-\rho}}, \tag{10}$$

so that $\left(\frac{1}{J}+\frac{\rho}{1-\rho}\right)S^2_{\hat{\mu}_j}$ is an unbiased estimator of $Var[\ {}^{n_2}_{n_1}\hat{\mu}_J]$. The only problem is that $\rho=\rho(n_1,n_2)=\frac{\sigma_2(n_1,n_2)}{\sigma_1(n_1,n_2)}$, the correlation between the $\hat{\mu}_j$'s, is unknown and difficult to estimate. Indeed, neither σ_1 nor σ_2 can be written as a linear combination of ${}_{n_1}\mu$, $(\sigma_2+(\ {}_{n_1}\mu)^2)$, $(\sigma_0-\sigma_3)$ and $(\sigma_3-\sigma_2)$, the only quantities we know how to estimate unbiasedly (besides linear combinations of these). We use a very naive surrogate for ρ as follows. Let us recall that $\hat{\mu}_j=\frac{1}{n_2}\sum_{i\in S_j^c}\mathcal{L}(Z_{S_j};Z_i)$. For the purpose of building our estimator, let us proceed as if $\mathcal{L}(Z_{S_j};Z_i)$ depended only on Z_i and n_1, i.e. the loss does not depend on the actual n_1 examples (Z_{S_j}) used for training but only on the number of training examples (n_1) and on the testing example (Z_i). Then it is not hard to show that the correlation between the $\hat{\mu}_j$'s becomes $\frac{n_2}{n_1+n_2}$. Indeed, when $\mathcal{L}(Z_{S_j};Z_i)=f(Z_i)$, we have

$$\hat{\mu}_1=\frac{1}{n_2}\sum_{i=1}^{n}I_1(i)f(Z_i) \qquad \text{and} \qquad \hat{\mu}_2=\frac{1}{n_2}\sum_{k=1}^{n}I_2(k)f(Z_k),$$

where $I_1(i)$ is equal to 1 if Z_i is a test example for $\hat{\mu}_1$ and is equal to 0 otherwise. Naturally, $I_2(k)$ is defined similarly. We obviously have $Var[\hat{\mu}_1]=Var[\hat{\mu}_2]$ with

$$Var[\hat{\mu}_1]=E[Var[\hat{\mu}_1|I_1(.)]]+Var[E[\hat{\mu}_1|I_1(.)]]=E\left[\frac{Var[f(Z_1)]}{n_2}\right]+Var[E[f(Z_1)]]=\frac{Var[f(Z_1)]}{n_2},$$

where $I_1(.)$ denotes the $n\times 1$ vector made of the $I_1(i)$'s. Moreover,

$$\begin{aligned}Cov[\hat{\mu}_1,\hat{\mu}_2] &= E[Cov[\hat{\mu}_1,\hat{\mu}_2|I_1(.),I_2(.)]]+Cov[E[\hat{\mu}_1|I_1(.),I_2(.)],E[\hat{\mu}_2|I_1(.),I_2(.)]]\\ &= E\left[\frac{1}{n_2^2}\sum_{i=1}^{n}I_1(i)I_2(i)\,Var[f(Z_i)]\right]+Cov[E[f(Z_1)],E[f(Z_1)]]\\ &= \frac{Var[f(Z_1)]}{n_2^2}\sum_{i=1}^{n}\frac{n_2^2}{n^2}+0=\frac{Var[f(Z_1)]}{n},\end{aligned}$$

so that the correlation between $\hat{\mu}_1$ and $\hat{\mu}_2$ ($\hat{\mu}_j$ and $\hat{\mu}_{j'}$ with $j\neq j'$ in general) is $\frac{n_2}{n}$.

Therefore our first estimator of $Var[\ {}^{n_2}_{n_1}\hat{\mu}_J]$ is $\left(\frac{1}{J}+\frac{\rho_o}{1-\rho_o}\right)S^2_{\hat{\mu}_j}$ where $\rho_o=\rho_o(n_1,n_2)=\frac{n_2}{n_1+n_2}$, that is $\left(\frac{1}{J}+\frac{n_2}{n_1}\right)S^2_{\hat{\mu}_j}$. This will tend to overestimate or underestimate $Var[\ {}^{n_2}_{n_1}\hat{\mu}_J]$ according to whether $\rho_o>\rho$ or $\rho_o<\rho$.

By construction, ρ_o will be a good substitute for ρ when $\mathcal{L}(Z_{S_j};Z)$ does not depend much on the training set Z_{S_j}, that is when the decision function of the underlying algorithm does not change too much when different training sets are chosen. Here are instances where we might suspect this to be true.

- The capacity (VC dimension) of the algorithm is not too large relative to the size of the training set (for instance a parametric model that is not too complex).

- The algorithm is robust relative to perturbations in the training set. For instance, one could argue that the support vector machine (Burges, 1998) would tend to fall in this category. Classification and regression trees (Breiman, Friedman, Olshen, & Stone, 1984) however will typically not have this property as a slight modification in data may lead to substantially different tree growths so that for two different training sets, the corresponding decision functions (trees) obtained may differ substantially on some regions. K-nearest neighbors techniques will also lead to substantially different decision functions when different training sets are used, especially if K is small.

3.2 Second Method: Overestimating $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$

Our second method aims at overestimating $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$. As explained in Appendix A.0, this leads to conservative inference, that is tests of hypothesis with actual size less than the nominal size. This is important because techniques currently in use have the opposite defect, that is they tend to be liberal (tests with actual size exceeding the nominal size), which is normally regarded as less desirable than conservative tests.

We have shown at the end of Section 2 that ${}^{n_2}_{n_1}\sigma_J^2 = Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ could not be estimated unbiasedly without some prior knowledge about $\sigma_0, \sigma_1, \sigma_2, \sigma_3$ (we showed after (10) how this can be done when $\rho = \frac{\sigma_2}{\sigma_1}$ is known). However, as we show below, we may estimate unbiasedly ${}^{n_2}_{n_1'}\sigma_J^2 = Var[\,{}^{n_2}_{n_1'}\hat{\mu}_J]$ where $n_1' = \lfloor \frac{n}{2} \rfloor - n_2 < n_1$ (we assume $n_2 < \lfloor \frac{n}{2} \rfloor$). Let ${}^{n_2}_{n_1'}\hat{\sigma}_J^2$ be the unbiased estimator, developed below, of the above variance. Since $n_1' < n_1$, we have (according to Conjecture 1) $Var[\,{}^{n_2}_{n_1'}\hat{\mu}_J] \geq Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$, so that ${}^{n_2}_{n_1'}\hat{\sigma}_J^2$ will tend to overestimate ${}^{n_2}_{n_1}\sigma_J^2$, that is $E[\,{}^{n_2}_{n_1'}\hat{\sigma}_J^2] = {}^{n_2}_{n_1'}\sigma_J^2 \geq {}^{n_2}_{n_1}\sigma_J^2$.

Here's how we may estimate ${}^{n_2}_{n_1'}\sigma_J^2$ without bias. The main idea is that we can get two independent instances of ${}^{n_2}_{n_1'}\hat{\mu}_J$ which allows us to estimate ${}^{n_2}_{n_1'}\sigma_J^2$ without bias. Of course variance estimation from only two observations is noisy. Fortunately, the process by which this variance estimate is obtained can be repeated at will, so that we may have many unbiased estimates of ${}^{n_2}_{n_1'}\sigma_J^2$. Averaging these yields a more accurate estimate of ${}^{n_2}_{n_1'}\sigma_J^2$.

Obtaining a pair of independent ${}^{n_2}_{n_1'}\hat{\mu}_J$'s is simple. Suppose, as before, that our data set Z_1^n consists of $n = n_1 + n_2$ examples. For simplicity, assume that n is even[2]. We have to randomly split our data Z_1^n into two distinct data sets, D_1 and D_1^c, of size $\lfloor \frac{n}{2} \rfloor$ each. Let $\hat{\mu}_{(1)}$ be the statistic of interest (${}^{n_2}_{n_1'}\hat{\mu}_J$) computed on D_1. This involves, among other things, drawing J train/test subsets from D_1, respectively of size n_1' and n_2. Let $\hat{\mu}_{(1)}^c$ be the statistic computed on D_1^c. Then $\hat{\mu}_{(1)}$ and $\hat{\mu}_{(1)}^c$ are independent since D_1 and D_1^c are independent data sets [3], so that $(\hat{\mu}_{(1)} - \frac{\hat{\mu}_{(1)} + \hat{\mu}_{(1)}^c}{2})^2 + (\hat{\mu}_{(1)}^c - \frac{\hat{\mu}_{(1)} + \hat{\mu}_{(1)}^c}{2})^2 = \frac{1}{2}(\hat{\mu}_{(1)} - \hat{\mu}_{(1)}^c)^2$ is unbiased for ${}^{n_2}_{n_1'}\sigma_J^2$. This

[2]When n is odd, everything is the same except that splitting the data in two will result in a leftover observation that is ignored. Thus D_m and D_m^c are still disjoint subsets of size $\lfloor \frac{n}{2} \rfloor$ from Z_1^n, but $Z_1^n \setminus (D_m \cup D_m^c)$ is a singleton instead of being the empty set.

[3]Independence holds if the train/test subsets selection process in D_1 is independent of the process in D_1^c. Otherwise, $\hat{\mu}_1$ and $\hat{\mu}_1^c$ may not be independent, but they are uncorrelated, which is all we actually need.

splitting process may be repeated M times. This yields D_m and D_m^c, with $D_m \cup D_m^c = Z_1^n$, $D_m \cap D_m^c = \emptyset$ and $|D_m| = |D_m^c| = \lfloor \frac{n}{2} \rfloor$ for $m = 1, \ldots, M$. Each split yields a pair $(\hat{\mu}_{(m)}, \hat{\mu}^c_{(m)})$ that is such that

$$E\left[\frac{(\hat{\mu}_{(m)} - \hat{\mu}^c_{(m)})^2}{2}\right] = \frac{1}{2} Var[\hat{\mu}_{(m)} - \hat{\mu}^c_{(m)}] = \frac{Var[\hat{\mu}_{(m)}] + Var[\hat{\mu}^c_{(m)}]}{2} = {}^{n_2}_{n_1'}\sigma^2_J.$$

This allows us to use the following unbiased estimator of ${}^{n_2}_{n_1'}\sigma^2_J$:

$$ {}^{n_2}_{n_1'}\hat{\sigma}^2_J = \frac{1}{2M} \sum_{m=1}^{M} (\hat{\mu}_{(m)} - \hat{\mu}^c_{(m)})^2. \tag{11}$$

Note that, according to Lemma 1, the variance of the proposed estimator is $Var[\,{}^{n_2}_{n_1'}\hat{\sigma}^2_J] = \frac{1}{4} Var[(\hat{\mu}_{(m)} - \hat{\mu}^c_{(m)})^2] \left(r + \frac{1-r}{M}\right)$ with $r = Corr[(\hat{\mu}_{(m)} - \hat{\mu}^c_{(m)})^2, (\hat{\mu}_{(m')} - \hat{\mu}^c_{(m')})^2]$ for $m \neq m'$. We may deduce from Lemma 1 that $r > 0$, but simulations yielded r close to 0, so that $Var[\,{}^{n_2}_{n_1'}\hat{\sigma}^2_J]$ decreased roughly like $\frac{1}{M}$. We provide some guidance on the choice of M in Section 5.

Note that the computational effort to obtain this variance estimator is proportional to JM. We could reduce this by a factor J if we use ${}^{n_2}_{n_1'}\hat{\sigma}^2_1$ to overestimate $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$, but we suspect that overestimation might be too gross as $E[\,{}^{n_2}_{n_1'}\hat{\sigma}^2_1] = {}^{n_2}_{n_1'}\sigma^2_1 \geq {}^{n_2}_{n_1'}\sigma^2_J \geq {}^{n_2}_{n_1}\sigma^2_J$. We considered this ultra-conservative estimator of ${}^{n_2}_{n_1}\sigma^2_J$ when we performed the simulations presented in Section 5 but, as we suspected, the resulting inference was too conservative. We do not show the results to avoid overcrowding the paper. We could also have gone half way by using ${}^{n_2}_{n_1'}\hat{\sigma}^2_{J'}$ with $1 < J' < J$, but we did not pursue this for the same reason as above.

4 Inference about ${}_{n_1}\mu$

We present seven different techniques to perform inference (confidence interval or test) about ${}_{n_1}\mu$. The first three are methods already in use in the machine-learning community, the others are methods we put forward. Among these new methods, two were shown in the previous section; the other two are the "pseudo-bootstrap" and corrected "pseudo-bootstrap" (described later). Tests [4] of the hypothesis $H_0 :\ {}_{n_1}\mu = \mu_0$ (at significance level α) have the following form

$$\text{reject } H_0 \text{ if } |\hat{\mu} - \mu_0| > c\sqrt{\hat{\sigma}^2}, \tag{12}$$

while confidence intervals for ${}_{n_1}\mu$ (at confidence level $1 - \alpha$) will look like

$${}_{n_1}\mu \in [\hat{\mu} - c\sqrt{\hat{\sigma}^2}, \hat{\mu} + c\sqrt{\hat{\sigma}^2}]. \tag{13}$$

[4] At this point, we encourage readers to consult Appendix A.0.

Note that in (12) or (13), $\hat{\mu}$ will be an average, $\hat{\sigma}^2$ is meant to be a variance estimate of $\hat{\mu}$ and (using the central limit theorem to argue that the distribution of $\hat{\mu}$ is approximately Gaussian) c will be a percentile from the $N(0,1)$ distribution or from Student's t distribution. The only difference between the seven techniques is in the choice of $\hat{\mu}$, $\hat{\sigma}^2$ and c. In this section we lay out what $\hat{\mu}$, $\hat{\sigma}^2$ and c are for the seven techniques considered and comment on whether each technique should be liberal or conservative based on its political ratio. All this is summarized in Table 1. The properties (size and power of the tests) of those seven techniques shall be investigated in Section 5.

We are now ready to introduce the statistics we will consider in this paper.

1. **t test statistic**
   Let the available data Z_1^n be split into a training set Z_{S_1} of size n_1 and a test set $Z_{S_1^c}$ of size $n_2 = n - n_1$, with n_2 relatively large (a third or a quarter of n for instance). One may consider $\hat{\mu} = {}^{n_2}_{n_1}\hat{\mu}_1$ to estimate ${}_{n_1}\mu$ and $\hat{\sigma}^2 = \frac{S_L^2}{n_2}$ where S_L^2 is the sample variance of the $L(1,i)$'s involved in ${}^{n_2}_{n_1}\hat{\mu}_1 = n_2^{-1}\sum_{i \in S_1^c} L(1,i)$ [5]. Inference would be based on the (incorrect) belief that

$$\frac{{}^{n_2}_{n_1}\hat{\mu}_1 - {}_{n_1}\mu}{\sqrt{\frac{S_L^2}{n_2}}} \sim N(0,1). \tag{14}$$

   We use $N(0,1)$ here (instead of t_{n_2-1} for instance) as n_2 is meant to be fairly large (greater than 50, say).

   Lemma 1 tells us that the political ratio here is

$$\frac{Var[\,{}^{n_2}_{n_1}\hat{\mu}_1]}{E\left[\frac{S_L^2}{n_2}\right]} = \frac{n_2\sigma_3 + (\sigma_0 - \sigma_3)}{\sigma_0 - \sigma_3} > 1,$$

   so this approach leads to liberal inference. This phenomenon grows worse as n_2 increases. Note that S_L^2 is a biased estimator of σ_0 (the unconditional variance of $L(1,i) =$

[5] We note that this statistic is closely related to the McNemar statistic (Everitt, 1977) when the problem at hand is the comparison of two classification algorithms, i.e. L is of the form (4) with Q of the form (2). Indeed, let $L_{A-B}(1,i) = L_A(1,i) - L_B(1,i)$ where $L_A(1,i)$ indicates whether Z_i is misclassified ($L_A(1,i) = 1$) by algorithm A or not ($L_A(1,i) = 0$); $L_B(1,i)$ is defined likewise. Of course, algorithms A and B share the same training set (S_1) and testing set (S_1^c). We have ${}^{n_2}_{n_1}\hat{\mu}_1 = \frac{n_{10}-n_{01}}{n_2}$, with n_{jk} being the number of times $L_A(1,i) = j$ and $L_B(1,i) = k$, $j = 0,1$, $k = 0,1$. McNemar's statistic is devised for testing $H_0 : {}_{n_1}\mu = 0$ (i.e. the $L_{A-B}(1,i)$'s have expectation 0) so that one may estimate the variance of the $L_{A-B}(1,i)$'s with the mean of the $(L_{A-B}(1,i) - 0)^2$'s (which is $\frac{n_{01}+n_{10}}{n_2}$) rather than with S_L^2. Then (12) becomes

$$\text{reject } H_0 \text{ if } \left|\frac{n_{10} - n_{01}}{\sqrt{n_{10} + n_{01}}}\right| > z_{1-\alpha/2},$$

with z_p the quantile p of $N(0,1)$, which squared leads to the McNemar's test (not corrected for continuity).

$L(Z_{S_1}; Z_i)$, $i \notin S_1$), but is unbiased for the variance of $L(1,i)$ conditional on the training set Z_{S_1} [6]. That is so because, given Z_{S_1}, the $L(1,i)$'s are independent variates. Therefore, although (14) is wrong, we do have

$$\frac{{}_{n_1}^{n_2}\hat{\mu}_1 - E[{}_{n_1}^{n_2}\hat{\mu}_1 | Z_{S_1}]}{\sqrt{\frac{S_L^2}{n_2}}} \approx N(0,1)$$

in so far as n_2 is large enough for the central limit theorem to apply. Therefore this method really allows us to make inference about $E[{}_{n_1}^{n_2}\hat{\mu}_1 | Z_{S_1}] = E[L(1,i)|Z_{S_1}] = E[\mathcal{L}(Z_{S_1}; Z_i)|Z_{S_1}]$, $i \notin S_1$, that is the generalization error of the specific rule obtained by training the algorithm on Z_{S_1}, not the generalization error of the algorithm *per se*. That is, according to Dietterich's taxonomy (Dietterich, 1998) briefly explained in Section 1, it deals with questions of type 1 through 4 rather than questions of type 5 through 8.

2. **Resampled t test statistic**
   Let us refresh some notation from Section 1. Particularly, let us recall that ${}_{n_1}^{n_2}\hat{\mu}_J = \frac{1}{J}\sum_{j=1}^{J}\hat{\mu}_j$. The resampled t test technique[7] considers $\hat{\mu} = {}_{n_1}^{n_2}\hat{\mu}_J$ and $\hat{\sigma}^2 = \frac{S_{\hat{\mu}_j}^2}{J}$ where $S_{\hat{\mu}_j}^2$ is the sample variance of the $\hat{\mu}_j$'s (see (9)). Inference would be based on the (incorrect) belief that

$$\frac{{}_{n_1}^{n_2}\hat{\mu}_J - {}_{n_1}\mu}{\sqrt{\frac{S_{\hat{\mu}_j}^2}{J}}} \sim t_{J-1}. \tag{15}$$

   Combining (8) and Lemma 1 gives us the following political ratio

$$\frac{Var[{}_{n_1}^{n_2}\hat{\mu}_J]}{E\left[\frac{S_{\hat{\mu}_j}^2}{J}\right]} = \frac{J\,Var[{}_{n_1}^{n_2}\hat{\mu}_J]}{E[S_{\hat{\mu}_j}^2]} = \frac{J\sigma_2 + (\sigma_1 - \sigma_2)}{\sigma_1 - \sigma_2} > 1,$$

   so this approach leads to liberal inference, a phenomenon that grows worse as J increases. Dietterich (Dietterich, 1998) observed this empirically through simulations.

   As argued in Section 2, $S_{\hat{\mu}_j}^2/J$ actually estimates (without bias) the variance of ${}_{n_1}^{n_2}\hat{\mu}_J$

---

[6]From this, we can rederive that S_L^2 is biased for the unconditional variance as follows:

$$\begin{aligned} E[S_L^2] &= E[E[S_L^2|Z_{S_1}]] = E[Var[L(1,i)|Z_{S_1}]] \\ &\leq E[Var[L(1,i)|Z_{S_1}]] + Var[E[L(1,i)|Z_{S_1}]] = Var[L(1,i)]. \end{aligned}$$

[7]When the problem at hand is the comparison of two classification algorithms, i.e. L is of the form (4) with Q of the form (2), this approach is what Dietterich (Dietterich, 1998) calls the "resampled paired t test" statistic.

conditional on Z_1^n. Thus while (15) is wrong, we do have

$$\frac{{}_{n_1}^{n_2}\hat{\mu}_J - E[\,{}_{n_1}^{n_2}\hat{\mu}_J|Z_1^n]}{\sqrt{\frac{S^2_{\hat{\mu}_j}}{J}}} \approx t_{J-1}.$$

Recall from the discussion following (8) that $E[\,{}_{n_1}^{n_2}\hat{\mu}_J|Z_1^n] = {}_{n_1}^{n_2}\hat{\mu}_\infty$. Therefore this method really allows us to make inference about ${}_{n_1}^{n_2}\hat{\mu}_\infty$, which is not too useful because we want to make inference about ${}_{n_1}\mu$.

3. **5x2 cv t test**
Dietterich (Dietterich, 1998)[8] split Z_1^n in half $M = 5$ times to yield $D_1, D_1^c, \ldots, D_5, D_5^c$ as in Section 3 and let

$$\tilde{\mu}_{(m)} = \lfloor n/2 \rfloor^{-1} \sum_{i \in D_m^c} \mathcal{L}(D_m; Z_i), \qquad \tilde{\mu}^c_{(m)} = \lfloor n/2 \rfloor^{-1} \sum_{i \in D_m} \mathcal{L}(D_m^c; Z_i).$$

He then used $\hat{\mu} = \tilde{\mu}_{(1)}$, $\hat{\sigma}^2 = \hat{\sigma}^2_{Diet} = \frac{1}{10}\sum_{m=1}^5 (\tilde{\mu}_{(m)} - \tilde{\mu}^c_{(m)})^2$ and $c = t_{5,1-\alpha/2}$. Note that the political ratio is

$$\frac{Var[\tilde{\mu}_{(1)}]}{E[\hat{\sigma}^2]} = \frac{\sigma_1(\lfloor n/2 \rfloor, \lfloor n/2 \rfloor)}{\sigma_1(\lfloor n/2 \rfloor, \lfloor n/2 \rfloor) - \sigma_4}$$

where $\sigma_4 = Cov[\tilde{\mu}_{(m)}, \tilde{\mu}^c_{(m)}]$.
Remarks

- As Dietterich noted, this allows inference for ${}_{\lfloor n/2 \rfloor}\mu$ which may be substantially distant from ${}_n\mu$.
- The choice of $M = 5$ seems arbitrary.
- The statistic was developed under the assumption that the $\tilde{\mu}_{(m)}$'s and $\tilde{\mu}^c_{(m)}$'s are 10 independent and identically distributed Gaussian variates. Even in this ideal case,

$$t_D = \frac{\hat{\mu} - {}_{\lfloor n/2 \rfloor}\mu}{\sqrt{\hat{\sigma}^2}} = \frac{\tilde{\mu}_{(1)} - {}_{\lfloor n/2 \rfloor}\mu}{\sqrt{\frac{1}{10}\sum_{m=1}^5 (\tilde{\mu}_{(m)} - \tilde{\mu}^c_{(m)})^2}} \tag{16}$$

  is not distributed as t_5 as assumed in (Dietterich, 1998) because $\tilde{\mu}_{(1)}$ and $(\tilde{\mu}_{(1)} - \tilde{\mu}^c_{(1)})$ are not independent. That is easily fixed in two different ways:
  - Take the sum from $m = 2$ to $m = 5$ and replace 10 by 8 in the denominator of (16) which would result in $t_D \sim t_4$,
  - Replace the numerator by $\sqrt{2}(\frac{\tilde{\mu}_{(1)} + \tilde{\mu}^c_{(1)}}{2} - {}_{\lfloor n/2 \rfloor}\mu)$ which would lead to $t_D \sim t_5$ as $\tilde{\mu}_{(1)} + \tilde{\mu}^c_{(1)}$ and $\tilde{\mu}_{(1)} - \tilde{\mu}^c_{(1)}$ are independent.

[8]Dietterich only considered the comparison of two classification algorithms, that is L of the form (4) with Q of the form (2).

In all cases, more degrees of freedom could be exploited; statistics distributed as t_8 can be devised by appropriate use of the 10 (assumed) independent variates.

4. **Conservative** Z
We estimate ${}_{n_1}\mu$ by $\hat{\mu} = {}^{n_2}_{n_1}\hat{\mu}_J$ and use $\hat{\sigma}^2 = {}^{n_2}_{n'_1}\hat{\sigma}^2_J$ (equation 11) as its conservative variance estimate. Since ${}^{n_2}_{n_1}\hat{\mu}_J$ is the mean of many (Jn_2 to be exact) $L(j,i)$'s, we may expect that its distribution is approximatively normal. We then proceed as if
$$Z = \frac{{}^{n_2}_{n_1}\hat{\mu}_J - {}_{n_1}\mu}{\sqrt{{}^{n_2}_{n'_1}\hat{\sigma}^2_J}} \tag{17}$$
was a $N(0,1)$ variate (even though ${}^{n_2}_{n'_1}\hat{\sigma}^2_J$ is designed to overestimate $Var[{}^{n_2}_{n_1}\hat{\mu}_J]$) to perform inference, leading us to use $c = z_{1-\alpha/2}$ in (12) or (13), where $z_{1-\alpha/2}$ is the percentile $1-\alpha$ of the $N(0,1)$ distribution. Some would perhaps prefer to use percentile from the t distribution, but it is unclear what the degrees of freedom ought to be. People like to use the t distribution in approximate inference frameworks, such as the one we are dealing with, to yield conservative inference. This is unnecessary here as inference is already conservative via the variance overestimation. Indeed, the political ratio
$$\frac{Var[{}^{n_2}_{n_1}\hat{\mu}_J]}{E[{}^{n_2}_{n'_1}\hat{\sigma}^2_J]} = \frac{{}^{n_2}_{n_1}\sigma^2_J}{{}^{n_2}_{n'_1}\sigma^2_J}$$
is smaller than 1 if we believe in Conjecture 1.

Regarding the choice of n_2 (and thus n_1), we may take it to be small relatively to n (the total number of examples available). One may use $n_2 = \frac{n}{10}$ for instance provided J is not smallish.

5. **Pseudo-Bootstrap**
To estimate the variance of $\hat{\mu} = {}^{n_2}_{n_1}\hat{\mu}_J$ by a procedure similar to the bootstrap (Efron & Tibshirani, 1993), we obtain R other instances of that random variable, by redoing the computation with different splits *on the same data* Z_1^n; call these $\check{\mu}_1, \ldots, \check{\mu}_R$. Thus, in total, $(R+1)J$ training and testing sets are needed here. Then one could consider $\hat{\sigma}^2 = \check{\sigma}^2$, where $\check{\sigma}^2$ is the sample variance of $\check{\mu}_1, \ldots, \check{\mu}_R$, and take $c = t_{R-1,1-\alpha/2}$, as $\check{\sigma}^2$ has $R-1$ degrees of freedom. Of course ${}^{n_2}_{n_1}\hat{\mu}_J$, $\check{\mu}_1, \ldots, \check{\mu}_R$ are $R+1$ identically distributed random variables. But they are not independent as we find, from (7), that the covariance between them is σ_2. Using Lemma 1, we have
$$\frac{Var[{}^{n_2}_{n_1}\hat{\mu}_J]}{E[\check{\sigma}^2]} = \frac{{}^{n_2}_{n_1}\sigma^2_J}{{}^{n_2}_{n_1}\sigma^2_J - \sigma_2} = \frac{J\sigma_2 + (\sigma_1 - \sigma_2)}{\sigma_1 - \sigma_2} > 1.$$
Note that this political ratio is the same as its counterpart for the resampled t-test because $E[\check{\sigma}^2] = E[\frac{S^2_{\hat{\mu}_j}}{J}]$. So the pseudo-bootstrap leads to liberal inference that should

worsen with increasing J just like the resampled t test statistic. In other words, the pseudo-bootstrap only provides a second estimator of $\frac{\sigma_1-\sigma_2}{J}$ which is more complicated and harder to compute than $\frac{S^2_{\hat{\mu}_j}}{J}$ which is also unbiased for $\frac{\sigma_1-\sigma_2}{J}$.

6. **Corrected resampled t-test statistic**
From our discussion in Section 3, we know that an unbiased estimator of $\frac{n_2}{n_1}\sigma^2_J$ is $\left(\frac{1}{J}+\frac{\rho}{1-\rho}\right)S^2_{\hat{\mu}_j}$, where $S^2_{\hat{\mu}_j}$ is the sample variance of the $\hat{\mu}_j$'s. Unfortunately ρ, the correlation between the $\hat{\mu}_j$'s, is unknown. The resampled t-test boldly puts $\rho=0$. We propose here to proceed as if $\rho=\rho_0=\frac{n_2}{n_1+n_2}$ as our argument in Section 3 suggests. So we use $\hat{\sigma}^2=\left(\frac{1}{J}+\frac{n_2}{n_1}\right)S^2_{\hat{\mu}_j}$. We must say again that this approximation is gross, but we feel it is better than putting $\rho=0$. Furthermore, in the ideal case where the vector of the $\hat{\mu}_j$'s follows the multivariate Gaussian distribution and ρ is actually equal to ρ_0, Lemma 2 states that $\frac{\frac{n_2}{n_1}\hat{\mu}_J-\ _{n_1}\mu}{\sqrt{\hat{\sigma}^2}}\sim t_{J-1}$. This is why we use $c=t_{J-1,1-\alpha/2}$.
Finally, let us note that the political ratio
$$\frac{Var[\ \frac{n_2}{n_1}\hat{\mu}_J]}{E[\hat{\sigma}^2]}=\frac{\frac{1}{J}+\frac{\rho}{1-\rho}}{\frac{1}{J}+\frac{n_2}{n_1}}$$
will be greater than 1 (liberal inference) if $\rho>\rho_0$. If $\rho<\rho_0$, the above ratio is smaller than 1, so that we must expect the inference to be conservative. Having mentioned earlier that conservative inference is preferable to liberal inference, we therefore hope that the ad hoc $\rho_0=\frac{n_2}{n_1+n_2}$ will tend to be larger than the actual correlation ρ.

7. **Corrected pseudo-bootstrap statistic**
Naturally, the correction we made in the resampled t test can be applied to the pseudo-bootstrap procedure as well. Namely, we note that $\left(1+J\frac{\rho}{1-\rho}\right)\check{\sigma}^2$, where $\check{\sigma}^2$ is the sample variance of the $\check{\mu}_r$'s, is unbiased for $\frac{n_2}{n_1}\sigma^2_J$. Naively replacing ρ by ρ_0 leads us to use $\hat{\sigma}^2=\left(1+\frac{Jn_2}{n_1}\right)\check{\sigma}^2$. Furthermore, in the ideal case where ρ is actually equal to ρ_0, and the vector made of $\frac{n_2}{n_1}\hat{\mu}_J$, $\check{\mu}_1,\ldots\check{\mu}_R$ follows the multivariate Gaussian distribution, Lemma 2 states that $\frac{\frac{n_2}{n_1}\hat{\mu}_J-\ _{n_1}\mu}{\sqrt{\hat{\sigma}^2}}\sim t_{R-1}$. This is why we use $c=t_{R-1,1-\alpha/2}$. Finally note that, just like in the corrected resampled t-test, the political ratio is
$$\frac{Var[\ \frac{n_2}{n_1}\hat{\mu}_J]}{E[\hat{\sigma}^2]}=\frac{\frac{1}{J}+\frac{\rho}{1-\rho}}{\frac{1}{J}+\frac{n_2}{n_1}}.$$

We conclude this section by providing in Table 1 a summary of the seven inference methods considered in the present section.

Name	$\hat{\mu}$	$\hat{\sigma}^2$	c	$\frac{Var[\hat{\mu}]}{E[\hat{\sigma}^2]}$
1. t-test (McNemar)	${}_{n_1}^{n_2}\hat{\mu}_1$	$\frac{1}{n_2}S_L^2$	$z_{1-\alpha/2}$	$\frac{n_2\sigma_3+(\sigma_0-\sigma_3)}{\sigma_0-\sigma_3} > 1$
2. resampled t	${}_{n_1}^{n_2}\hat{\mu}_J$	$\frac{1}{J}S_{\hat{\mu}_j}^2$	$t_{J-1,1-\alpha/2}$	$1+J\frac{\rho}{1-\rho} > 1$
3. Dietterich's 5×2 cv	${}_{n/2}^{n/2}\hat{\mu}_1$	$\hat{\sigma}_{Diet}^2$	$t_{5,1-\alpha/2}$	$\frac{\sigma_1}{\sigma_1-\sigma_4}$
4. conservative Z	${}_{n_1}^{n_2}\hat{\mu}_J$	${}_{n_1'}^{n_2}\hat{\sigma}_J^2$	$z_{1-\alpha/2}$	$\frac{{}_{n_1}^{n_2}\sigma_J^2}{{}_{n_1'}^{n_2}\sigma_J^2} < 1$
5. pseudo-bootstrap	${}_{n_1}^{n_2}\hat{\mu}_J$	$\check{\sigma}^2$	$t_{R-1,1-\alpha/2}$	$1+J\frac{\rho}{1-\rho} > 1$
6. corrected resampled t	${}_{n_1}^{n_2}\hat{\mu}_J$	$\left(\frac{1}{J}+\frac{n_2}{n_1}\right)S_{\hat{\mu}_j}^2$	$t_{J-1,1-\alpha/2}$	$\frac{1+J\frac{\rho}{1-\rho}}{1+J\frac{n_2}{n_1}}$
7. corrected pseudo-bootstrap	${}_{n_1}^{n_2}\hat{\mu}_J$	$\left(1+\frac{Jn_2}{n_1}\right)\check{\sigma}^2$	$t_{R-1,1-\alpha/2}$	$\frac{1+J\frac{\rho}{1-\rho}}{1+J\frac{n_2}{n_1}}$

Table 1: Summary description of the seven inference methods considered in relation to the rejection criteria shown in (12) or the confidence interval shown in (13). z_p and $t_{k,p}$ refer to the quantile p of the $N(0,1)$ and Student t_k distributions respectively. The political ratio, that is $\frac{Var[\hat{\mu}]}{E[\hat{\sigma}^2]}$, indicates if inference according to the corresponding method will tend to be conservative (ratio less than 1) or liberal (ratio greater than 1). See Section 4 for further details.

5 Simulation study

We performed a simulation study to investigate the power and the size of the seven statistics considered in the previous section. We also want to make recommendations on the value of J to use for those methods that involve ${}_{n_1}^{n_2}\hat{\mu}_J$. Simulation results will also lead to a recommendation on the choice of M when the conservative Z is used.

We will soon introduce the three kinds of problems we considered to cover a good range of possible applications. For a given problem, we shall generate 1000 independent sets of data of the form $\{Z_1, \ldots, Z_n\}$. Once a data set $Z_1^n = \{Z_1, \ldots Z_n\}$ has been generated, we may compute confidence intervals and/or a tests of hypothesis based on the statistics laid out in Section 4 and summarized in Table 1. A difficulty arises however. For a given n, those seven methods don't aim at inference for the same generalization error. For instance, Dietterich's method aims at ${}_{n/2}\mu$ (we take n even for simplicity), while the others aim at ${}_{n_1}\mu$ where n_1 would usually be different for different methods (e.g. $n_1 = \frac{2n}{3}$ for the t-test and $n_1 = \frac{9n}{10}$ for methods using ${}_{n_1}^{n_2}\hat{\mu}_J$). In order to compare the different techniques, for a given n, we shall always aim at ${}_{n/2}\mu$. The use of the statistics other than Dietterich's 5×2 cv shall be modified as follows.

- **t test statistic**
  We take $n_1 = n_2 = \frac{n}{2}$. This deviates slightly from the normal usage of the t test where n_2 is one third, say, of n, not one half.

- **Methods other that the t-test and Dietterich's 5×2 cv**

For methods involving ${}^{n_2}_{n_1}\hat{\mu}_J$ where J is a free parameter, that is all methods except the t-test and Dietterich's 5×2 cv, we take $n_1 = n_2 = \frac{n}{2}$. This deviates substantially from the normal usage where n_1 would be 5 to 10 times larger than n_2, say. For that reason, we also take $n_1 = \frac{n}{2}$ and $n_2 = \frac{n}{10}$ (assume n is a multiple of 10 for simplicity). This is achieved by throwing away 40% of the data. Note that when we will address the question of the choice of J (and M for the conservative Z), we shall use $n_1 = \frac{9n}{10}$ and $n_2 = \frac{n}{10}$, more in line with the normal usage.

- **Conservative Z**
  For the conservative Z, we need to explain how we compute the variance estimate. Indeed, formula (11) suggests that we have to compute ${}^{n_2}_{0}\hat{\sigma}^2_J$ whenever $n_1 = n_2 = \frac{n}{2}$! What we do is that we choose n_2 as we would normally do (10% of n here) and do the variance calculation as usual (${}^{n_2}_{n/2-n_2}\hat{\sigma}^2_J = {}^{n/10}_{2n/5}\hat{\sigma}^2_J$). However, as explained above, we use ${}^{n/2}_{n/2}\hat{\mu}_J$ and ${}^{n_2}_{n/2}\hat{\mu}_J = {}^{n/10}_{n/2}\hat{\mu}_J$ instead of ${}^{n_2}_{n-n_2}\hat{\mu}_J$ as the cross-validation estimators. Recall that we have argued in Section 2 that ${}^{n_2}_{n_1}\sigma^2_J$ was decreasing in n_1 and n_2. Consequently the variances of ${}^{n/2}_{n/2}\hat{\mu}_J$ and ${}^{n_2}_{n/2}\hat{\mu}_J$ are smaller than ${}^{n_2}_{n/2-n_2}\sigma^2_J$, so that ${}^{n_2}_{n/2-n_2}\hat{\sigma}^2_J$ still acts as a conservative variance estimate, that is

$$E[{}^{n_2}_{n/2-n_2}\hat{\sigma}^2_J] = {}^{n_2}_{n/2-n_2}\sigma^2_J = Var[{}^{n_2}_{n/2-n_2}\hat{\mu}_J] \geq Var[{}^{n_2}_{n/2}\hat{\mu}_J] \geq Var[{}^{n/2}_{n/2}\hat{\mu}_J].$$

  Thus the variance overestimation will be more severe in the case of ${}^{n/2}_{n/2}\hat{\mu}_J$.

We consider three kinds of problems to cover a good range of possible applications:

1. **Prediction in simple normal linear regression**

   We consider the problem of estimating the generalization error in a simple Gaussian regression problem. We thus have $Z = (X, Y)$ with $X \sim N(\mu_X, \sigma^2_X)$ and $Y|X \sim N(a + bX, \sigma^2_{Y|X})$ where $\sigma^2_{Y|X}$ is constant (does not depend on X). The learning algorithms are

   (A) **Sample mean**
   The decision function is $F_A(Z_S)(X) = \frac{1}{n_1}\sum_{i \in S} Y_i = \bar{Y}_S$, that is the mean of the Y's in the training set Z_S. Note that this decision function does not depend on X. We use a quadratic loss, so that $L_A(j, i) = (F_A(Z_{S_j})(X_i) - Y_i)^2 = (\bar{Y}_{S_j} - Y_i)^2$.

   (B) **Linear regression**
   The decision function is $F_B(Z_S)(X) = \hat{a}_S + \hat{b}_S X$ where $\hat{a}_S$ and $\hat{b}_S$ are the intercept and the slope of the ordinary least squares regression of Y on X performed on the training set Z_S. Since we use a quadratic loss, we therefore have $L_B(j, i) = (F_B(Z_{S_j})(X_i) - Y_i)^2 = (\hat{a}_{S_j} + \hat{b}_{S_j} X_i - Y_i)^2$.

   On top of inference about the generalization errors of algorithm A (${}_{n_1}\mu_A$) and algorithm B (${}_{n_1}\mu_B$), we also consider inference about ${}_{n_1}\mu_{A-B} = {}_{n_1}\mu_A - {}_{n_1}\mu_B$, the difference of those generalization errors. This inference is achieved by considering $L_{A-B}(j,i) = L_A(j,i) - L_B(j,i)$.

Table 2 describes the four simulations we performed for the regression problem. For instance, in Simulation 1, we generated 1000 samples of size 200, with $\mu_x = 10$, $\sigma_X^2 = 1$, $a = 100$, $b = 1$ and $\sigma_{Y|X}^2 = 97$. It is shown in (Nadeau & Bengio, 1999) that ${}_{n_1}\mu_A = \frac{n_1+1}{n_1}(\sigma_{Y|X}^2 + b^2\sigma_X^2)$ and ${}_{n_1}\mu_B = \frac{n_1+1}{n_1}\frac{n_1-2}{n_1-3}\sigma_{Y|X}^2$. Thus the first and third simulation correspond to cases where the two algorithms generalize equally well (for $n_1 = \frac{n}{2}$); in the second and fourth case, the linear regression generalizes better than the sample mean [9]. The table also provides some summary confidence intervals for quantities of interest, namely ${}_{n_1}\mu$, $\rho(n_1,n_2) = \frac{\sigma_2(n_1,n_2)}{\sigma_1(n_1,n_2)}$ and r.

2. **Classification of two Gaussian populations**
We consider the problem of estimating the generalization error in a classification problem with two classes. We thus have $Z = (X,Y)$ with $Prob(Y=1) = Prob(Y=0) = \frac{1}{2}$, $X|Y=0 \sim N(\mu_0, \Sigma_0)$ and $X|Y=1 \sim N(\mu_1, \Sigma_1)$. The learning algorithms are

(A) **Regression tree**
We train a least square regression tree [10] (Breiman et al., 1984) of Y against X and the decision function is $F_A(Z_S)(X) = I[N_{Z_S}(X) > 0.5]$ where $N_{Z_S}(X)$ is the leaf value corresponding to X of the tree obtained when training on Z_S. Thus $L_A(j,i) = I[F_A(Z_{S_j})(X_i) \neq Y_i]$ is equal to 1 whenever this algorithm misclassifies example i when the training set is Z_{S_j}; otherwise it is 0.

(B) **Ordinary least squares linear regression**
We perform the regression of Y against X and the decision function is $F_B(Z_S)(X) = I[\hat{\beta}'_{Z_S}X > \frac{1}{2}]$ where $\hat{\beta}_S$ is the ordinary least squares regression coefficient estimates[11] obtained by training on the set Z_S. Thus $L_B(j,i) = I[F_B(Z_{S_j})(X_i) \neq Y_i]$ is equal to 1 whenever this algorithm misclassifies example i when the training set is Z_{S_j}; otherwise it is 0.

On top of inference about the generalization errors ${}_{n_1}\mu_A$ and ${}_{n_1}\mu_B$ associated with those two algorithms, we also consider inference about ${}_{n_1}\mu_{A-B} = {}_{n_1}\mu_A - {}_{n_1}\mu_B = E[L_{A-B}(j,i)]$ where $L_{A-B}(j,i) = L_A(j,i) - L_B(j,i)$.

Table 3 describes the four simulations we performed for the Gaussian populations classification problem. Again, we considered two simulations with $n = 200$ and two simulations with $n = 2000$. We also chose the parameters μ_0, μ_1, Σ_0 and Σ_1 in such a

[9] The parameters of the simulations displayed in Tables 2, 3 and 4 were more or less chosen arbitrarily. However, efforts were made so that one or two simulations for each problem would correspond to ${}_{n_1}\mu_{A-B} = 0$ (i.e. ${}_{n_1}\mu_A = {}_{n_1}\mu_B$).

[10] The function **tree** in Splus 4.5 for Windows with default options and no pruning was used to train the regression tree.

[11] $\hat{\beta}_{Z_S}$ includes an intercept and correspondingly 1 was included in the input vector X.

	Simulation 1	Simulation 2	Simulation 3	Simulation 4
n	200	200	2000	2000
μ_X	10	10	10	10
a	100	100	100	100
b	1	2	0.1	0.1
σ_X^2	1	2	1	5
$\sigma_{Y\|X}^2$	97	64	9.97	9
${}_{n/2}\mu_A$ *	[98.77,100.03]	[72.30,73.25]	[9.961,10.002]	[9.040,9.075]
${}_{n/2}\mu_B$	[98.69,99.96]	[64.89,65.73]	[9.961,10.002]	[8.999,9.034]
${}_{n/2}\mu_{A-B}$ *	[-0.03,0.19]	[7.25,7.68]	[-0.001,0.001]	[0.039,0.043]
${}_{9n/10}\mu_A$ *	[98.19, 99.64]	[71.92, 72.99]	[9.952,9.998]	[9.026,9.067]
${}_{9n/10}\mu_B$	[97.71, 99.16]	[64.30,65.24]	[9.948,9.993]	[8.982,9.023]
${}_{9n/10}\mu_{A-B}$ *	[0.36,0.60]	[7.45,7.93]	[0.003,0.006]	[0.042,0.047]
$\rho_A(\frac{n}{2},\frac{n}{2})$	[0.466,0.512]	[0.487,0.531]	[0.484,0.531]	[0.471,0.515]
$\rho_B(\frac{n}{2},\frac{n}{2})$	[0.467,0.514]	[0.473,0.517]	[0.483,0.530]	[0.472,0.517]
$\rho_{A-B}(\frac{n}{2},\frac{n}{2})$	[0.225,0.298]	[0.426,0.482]	[0.226,0.282]	[0.399,0.455]
$\rho_A(\frac{n}{2},\frac{n}{10})$	[0.148,0.179]	[0.165,0.193]	[0.162,0.194]	[0.147,0.176]
$\rho_B(\frac{n}{2},\frac{n}{10})$	[0.152,0.183]	[0.156,0.183]	[0.162,0.194]	[0.147,0.175]
$\rho_{A-B}(\frac{n}{2},\frac{n}{10})$	[0.103,0.143]	[0.146,0.184]	[0.089,0.128]	[0.131,0.165]
$\rho_A(\frac{9n}{10},\frac{n}{10})$	[0.090,0.115]	[0.094,0.117]	[0.090,0.111]	[0.088,0.108]
$\rho_B(\frac{9n}{10},\frac{n}{10})$	[0.092,0.117]	[0.089,0.111]	[0.090,0.111]	[0.088,0.108]
$\rho_{A-B}(\frac{9n}{10},\frac{n}{10})$	[0.062,0.091]	[0.084,0.109]	[0.059,0.085]	[0.086,0.109]
r_A	[0.021,0.034]	[0.027,0.040]	[-0.003,0.008]	[-0.001,0.008]
r_B	[0.022,0.034]	[0.028,0.043]	[-0.003,0.008]	[-0.001,0.009]
r_{A-B}	[0.154,0.203]	[0.071,0.095]	[0.163,0.202]	[0.087,0.114]

Table 2: Description of four simulations for the simple linear regression problem. In each of the four simulations, 1000 independent samples of size n where generated with μ_X, a, b, σ_X^2 and $\sigma_{Y|X}^2$ as shown in the table. 95% confidence intervals for ${}_{n_1}\mu$, $\rho(n_1,n_2)=\frac{\sigma_2(n_1,n_2)}{\sigma_1(n_1)}$ and $r = Corr[(\hat{\mu}_{(m)}-\hat{\mu}^c_{(m)})^2,(\hat{\mu}_{(m')}-\hat{\mu}^c_{(m')})^2]$ defined after (11) are provided. The subscripts ${}_A$, ${}_B$ and ${}_{A-B}$ indicates whether we are working with L_A, L_B or L_{A-B}. An asterisk besides μ indicates that powers of tests for that μ are displayed in a figure.

	Simulation 1	Simulation 2	Simulation 3	Simulation 4
n	200	200	2000	2000
μ_0	(0,0)	(0,0)	(0,0)	(0,0)
μ_1	(1,1)	(1,1)	(1,1)	(1,1)
Σ_0	I_2	I_2	I_2	I_2
Σ_1	$\frac{1}{2}I_2$	$\frac{1}{6}I_2$	$\frac{1}{2}I_2$	$0.173I_2$
${}_{n/2}\mu_A$ *	[0.249,0.253]	[0.146,0.149]	[0.247,0.248]	[0.142,0.143]
${}_{n/2}\mu_B$	[0.204,0.208]	[0.146,0.148]	[0.200,0.201]	[0.142,0.143]
${}_{n/2}\mu_{A-B}$ *	[0.044,0.046]	[-0.001,0.002]	[0.0467,0.0475]	$[-1\times 10^{-4}, 8\times 10^{-4}]$
${}_{9n/10}\mu_A$ *	[0.247,0.252]	[0.142,0.147]	[0.235,0.237]	[0.132,0.133]
${}_{9n/10}\mu_B$	[0.201,0.205]	[0.142,0.145]	[0.199,0.200]	[0.142,0.143]
${}_{9n/10}\mu_{A-B}$ *	[0.044,0.049]	[-0.001,0.003]	[0.036,0.037]	[-0.011,-0.009]
$\rho_A(\frac{n}{2},\frac{n}{2})$	[0.345,0.392]	[0.392,0.438]	[0.354,0.400]	[0.380,0.423]
$\rho_B(\frac{n}{2},\frac{n}{2})$	[0.418,0.469]	[0.369,0.417]	[0.462,0.508]	[0.388,0.432]
$\rho_{A-B}(\frac{n}{2},\frac{n}{2})$	[0.128,0.154]	[0.174,0.205]	[0.120,0.146]	[0.179,0.211]
$\rho_A(\frac{n}{2},\frac{n}{10})$	[0.189,0.223]	[0.224,0.260]	[0.190,0.225]	[0.207,0.242]
$\rho_B(\frac{n}{2},\frac{n}{10})$	[0.150,0.182]	[0.135,0.163]	[0.141,0.170]	[0.129,0.156]
$\rho_{A-B}(\frac{n}{2},\frac{n}{10})$	[0.100,0.124]	[0.130,0.157]	[0.087,0.106]	[0.112,0.138]
$\rho_A(\frac{9n}{10},\frac{n}{10})$	[0.137,0.166]	[0.156,0.187]	[0.113,0.137]	[0.126,0.153]
$\rho_B(\frac{9n}{10},\frac{n}{10})$	[0.089,0.112]	[0.077,0.097]	[0.080,0.102]	[0.081,0.100]
$\rho_{A-B}(\frac{9n}{10},\frac{n}{10})$	[0.077,0.096]	[0.090,0.111]	[0.049,0.065]	[0.078,0.100]
r_A	[0.007,0.018]	[0.025,0.039]	[-0.005,0.003]	[-0.003,0.006]
r_B	[0.006,0.017]	[0.023,0.037]	[-0.003,0.007]	[-0.003,0.006]
r_{A-B}	[0.010,0.021]	[0.007,0.017]	[-0.003,0.006]	[-0.001,0.009]

Table 3: Description of four simulations for the classification of two Gaussian populations. In each of the four simulations, 1000 independent samples of size n where generated with μ_0, μ_1, Σ_0, Σ_1 as shown in the table. 95% confidence intervals for ${}_{n_1}\mu$, $\rho(n_1,n_2)=\frac{\sigma_2(n_1,n_2)}{\sigma_1(n_1)}$ and $r = Corr[(\hat{\mu}_{(m)} - \hat{\mu}^c_{(m)})^2, (\hat{\mu}_{(m')} - \hat{\mu}^c_{(m')})^2]$ defined after (11) are provided. The subscripts ${}_A$, ${}_B$ and ${}_{A-B}$ indicates whether we are working with L_A, L_B or L_{A-B}. An asterisk besides μ indicates that powers of tests for that μ are displayed in a figure.

way that in Simulations 2 and 4, the two algorithms generalize equally well; in Simulations 1 and 3, the linear regression generalizes better than the regression tree. The table also provides some summary confidence intervals for quantities of interest, namely ${}_{n_1}\mu$, $\rho(n_1, n_2) = \frac{\sigma_2(n_1,n_2)}{\sigma_1(n_1,n_2)}$ and r.

3. **Classification of letters**
   We consider the problem of estimating generalization errors in the Letter Recognition classification problem (Blake, Keogh, & Merz, 1998). The learning algorithms are

   (A) **Classification tree**
   We train a classification tree (Breiman et al., 1984) [12] to obtain its decision function $F_A(Z_S)(X)$. Here the classification loss function $L_A(j,i) = I[F_A(Z_{S_j})(X_i) \neq Y_i]$ is equal to 1 whenever this algorithm misclassifies example i when the training set is Z_{S_j}; otherwise it is 0.

   (B) **First nearest neighbor**
   We apply the first nearest neighbor rule with a distorted distance metric to pull down the performance of this algorithm to the level of the classification tree (as in (Dietterich, 1998)). Specifically, the distance between two vectors of inputs $X^{(1)}$ and $X^{(2)}$ is

$$d(X^{(1)}, X^{(2)}) = \sum_{k=1}^{3} w^{2-k} \sum_{i \in C_k} (X_i^{(1)} - X_i^{(2)})^2$$

   where $C_1 = \{1, 3, 9, 16\}$, $C_2 = \{2, 4, 6, 7, 8, 10, 12, 14, 15\}$ and $C_3 = \{5, 11, 13\}$ denote the sets of components that are weighted by w, 1 and w^{-1} respectively. Table 4 shows the values of w considered. We have $L_B(j,i)$ equal to 1 whenever this algorithm misclassifies example i when the training set is Z_{S_j}; otherwise it is 0.

   In addition to inference about the generalization errors ${}_{n_1}\mu_A$ and ${}_{n_1}\mu_B$ associated with those two algorithms, we also consider inference about ${}_{n_1}\mu_{A-B} = {}_{n_1}\mu_A - {}_{n_1}\mu_B = E[L_{A-B}(j,i)]$ where $L_{A-B}(j,i) = L_A(j,i) - L_B(j,i)$. We sample, without replacement, 300 examples from the 20000 examples available in the Letter Recognition data base. Repeating this 1000 times, we obtain 1000 sets of data of the form $\{Z_1, \ldots, Z_{300}\}$. The table also provides some summary confidence intervals for quantities of interest, namely ${}_{n_1}\mu$, $\rho(n_1, n_2) = \frac{\sigma_2(n_1,n_2)}{\sigma_1(n_1,n_2)}$ and r.

Before we comment on Tables 2, 3 and 4, let us describe how confidence intervals shown in those tables were obtained. First, let us point out that confidence intervals for generalization errors in those tables have nothing to do with the confidence intervals that we may compute from the statistics shown in Section 4. Indeed, the latter can be computed on a single data

[12] We used the function **tree** in Splus version 4.5 for Windows. The default arguments were used and no pruning was performed. The function **predict** with option **type="class"** was used to retrieve the decision function of the tree.

	Simulation 1	Simulation 2	Simulation 3
n	300	300	300
w	1	5	10
${}_{n/2}\mu_B$ *	[0.539,0.542]	[0.593,0.596]	[0.632,0.635]
${}_{n/2}\mu_{A-B}$ *	[0.150,0.152]	[0.096,0.099]	[0.057,0.060]
${}_{9n/10}\mu_B$	[0.434,0.439]	[0.496,0.501]	[0.544,0.548]
${}_{9n/10}\mu_{A-B}$ *	[0.148,0.153]	[0.086,0.091]	[0.039,0.044]
$\rho_B(\frac{n}{2},\frac{n}{2})$	[0.310,0.355]	[0.334,0.376]	[0.349,0.393]
$\rho_{A-B}(\frac{n}{2},\frac{n}{2})$	[0.134,0.160]	[0.152,0.180]	[0.160,0.191]
$\rho_B(\frac{n}{2},\frac{n}{10})$	[0.167,0.198]	[0.182,0.214]	[0.197,0.232]
$\rho_{A-B}(\frac{n}{2},\frac{n}{10})$	[0.122,0.148]	[0.129,0.155]	[0.130,0.156]
$\rho_B(\frac{9n}{10},\frac{n}{10})$	[0.105,0.129]	[0.106,0.131]	[0.115,0.140]
$\rho_{A-B}(\frac{9n}{10},\frac{n}{10})$	[0.085,0.105]	[0.085,0.105]	[0.084,0.104]
r_B	[-0.006,0.001]	[-0.004,0.004]	[-0.004,0.005]
r_{A-B}	[-0.004,0.004]	[-0.004,0.004]	[-0.003,0.005]
	Simulation 4	Simulation 5	Simulation 6
n	300	300	300
w	17.25	25	2048
${}_{n/2}\mu_B$ *	[0.666,0.669]	[0.690,0.693]	[0.779,0.782]
${}_{n/2}\mu_{A-B}$ *	[0.023,0.026]	[-0.001,0.002]	[-0.089,-0.087]
${}_{9n/10}\mu_B$	[0.586,0.591]	[0.616,0.620]	[0.730,0.734]
${}_{9n/10}\mu_{A-B}$ *	[-0.003,0.001]	[-0.033,-0.028]	[-0.147,-0.142]
$\rho_B(\frac{n}{2},\frac{n}{2})$	[0.360,0.404]	[0.368,0.413]	[0.347,0.392]
$\rho_{A-B}(\frac{n}{2},\frac{n}{2})$	[0.167,0.198]	[0.170,0.202]	[0.178,0.211]
$\rho_B(\frac{n}{2},\frac{n}{10})$	[0.200,0.238]	[0.201,0.238]	[0.201,0.237]
$\rho_{A-B}(\frac{n}{2},\frac{n}{10})$	[0.130,0.156]	[0.129,0.155]	[0.133,0.162]
$\rho_B(\frac{9n}{10},\frac{n}{10})$	[0.118,0.143]	[0.125,0.151]	[0.119,0.145]
$\rho_{A-B}(\frac{9n}{10},\frac{n}{10})$	[0.085,0.106]	[0.087,0.108]	[0.094,0.116]
r_B	[-0.004,0.004]	[-0.005,0.004]	[0.002,0.012]
r_{A-B}	[-0.002,0.007]	[-0.001,0.009]	[-0.001,0.009]

Table 4: Description of six simulations for the letter recognition problem. In each of the six simulations, 1000 independent samples of size $n = 300$ where generated and algorithms A and B were used with B using the distorted metric factor w shown in the table. 95% confidence intervals for ${}_{n_1}\mu$, $\rho(n_1, n_2) = \frac{\sigma_2(n_1,n_2)}{\sigma_1(n_1)}$ and $r = Corr[(\hat{\mu}_{(m)} - \hat{\mu}^c_{(m)})^2, (\hat{\mu}_{(m')} - \hat{\mu}^c_{(m')})^2]$ defined after (11) are provided. The subscripts ${}_A$ ${}_B$ and ${}_{A-B}$ indicates whether we are working with L_A, L_B or L_{A-B}. An asterisk besides μ indicates that powers of tests for that μ are displayed in a figure. See Table 5 for the results obtained with algorithm A (the same for all 6 simulations).

${}_{n/2}\mu_A$	${}_{9n/10}\mu_A$	$\rho_A(\frac{n}{2},\frac{n}{2})$	$\rho_A(\frac{n}{2},\frac{n}{10})$	$\rho_A(\frac{9n}{10},\frac{n}{10})$	r_A
[0.691,0.694]	[0.585,0.589]	[0.223,0.259]	[0.137,0.164]	[0.099,0.123]	[0.002,0.013]

Table 5: Confidence intervals for the statistics measured with algorithm A for all 6 simulations with the letter recognition problem (see Table 4).

set Z_1^n, while the confidence intervals in the tables use 1000 data sets Z_1^n as we now explain. For a given data set, we may compute ${}^{n_2}_{n_1}\hat{\mu}_{25}$, which has expectation ${}_{n_1}\mu$. Recall, from (5) in Section 1, that ${}^{n_2}_{n_1}\hat{\mu}_{25} = \frac{1}{25}\sum_{j=1}^{25}\hat{\mu}_j$ is the average of 25 crude estimates of the generalization error. Also recall from Section 2 that those crude estimates have the moment structure displayed in Lemma 1 with $\beta = {}_{n_1}\mu$ and $\pi = \rho(n_1, n_2) = \frac{\sigma_2(n_1,n_2)}{\sigma_1(n_1,n_2)}$. Call $\vec{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_{25})'$ the vector of those crude estimates. Since we generate 1000 independent data sets, we have 1000 independent instances of such vectors. As may be seen in the Appendix A.3, appropriate use of the theory of estimating functions (White, 1982) then yields approximate confidence intervals for ${}_{n_1}\mu$ and $\rho(n_1, n_2)$. Confidence intervals for $r = Corr[(\hat{\mu}_{(m)} - \hat{\mu}^c_{(m)})^2, (\hat{\mu}_{(m')} - \hat{\mu}^c_{(m')})^2]$, defined in Section 3, are obtained in the same manner we get confidence intervals for $\rho(n_1, n_2)$. Namely, we have 1000 independent instances of the vector $((\hat{\mu}_{(1)} - \hat{\mu}^c_{(1)})^2, \ldots, (\hat{\mu}_{(20)} - \hat{\mu}^c_{(20)})^2)'$ where the $\hat{\mu}_{(m)}$'s and $\hat{\mu}^c_{(m)}$ are ${}^{n/10}_{2n/5}\hat{\mu}_{15}$'s as we advocate the use of $J = 15$ later in this section.

We see that ${}_{n_1}\mu$ may substantially differ for different n_1. This is most evident in Table 4 where confidence intervals for ${}_{150}\mu$ differ from confidence intervals for ${}_{270}\mu$ in a noticeable manner. We see that our very naive approximation $\rho_0(n_1, n_2) = \frac{n_2}{n_1+n_2}$ is not as bad as one could expect. Often the confidence intervals for the actual $\rho(n_1, n_2)$ contains $\rho_0(n_1, n_2)$ [13]. When this is not the case, the approximation $\rho_0(n_1, n_2)$ usually appears to be reasonably close to the actual value of the correlation $\rho(n_1, n_2)$. Furthermore, when we compare two algorithms, the approximation $\rho_0(n_1, n_2)$ is not smaller than the actual value of the correlation $\rho_{A-B}(n_1, n_2)$, which is good since that indicates that the inference based on the corrected pseudo-bootstrap and on the corrected resampled t-test will not be liberal, as argued in Section 4. We finally note that the correlation r appears to be fairly small, except when we compare algorithms A and B in the simple linear regression problem. Thus, as we stated at the end of Section 3, we should expect $Var[{}^{n_2}_{n'_1}\hat{\sigma}^2_J]$ to decrease like $\frac{1}{M}$.

5.1 Sizes and powers of tests

One of the most important thing to investigate is the size (probability of rejecting the null hypothesis when it is true) of the tests based on the statistics shown in Section 4 and compare their powers (probability of rejecting the null hypothesis when it is false). The four panels of Figure 1 show the estimated powers of the statistics for the hypothesis $H_0 : {}_{n/2}\mu_A = \mu_0$ for various values of μ_0 in the regression problem. We estimate powers (probabilities of rejection) by proportions of rejection observed in the simulation. We must underline that, despite appearances, these are not "power curves" in the usual sense of the term (see Appendix A.0). In a "power curve", the hypothesized value of ${}_{n/2}\mu_A$ is fixed and the actual value of ${}_{n/2}\mu_A$ varies. Here, it is the reverse that we see in a given panel: the actual value of ${}_{n/2}\mu_A$ is fixed while the hypothesized value of ${}_{n/2}\mu_A$ (i.e. μ_0) is varied. We may call this a pseudo-power curve. We do this because constructing "power curves" would be too computationally

[13] As mentioned before, the corrected pseudo-bootstrap and the corrected resampled t-test are typically used in cases where training sets are 5 or 10 times larger than test sets. So we must only be concerned with $\rho(\frac{n}{2}, \frac{n}{10})$ and $\rho(\frac{9n}{10}, \frac{n}{10})$.

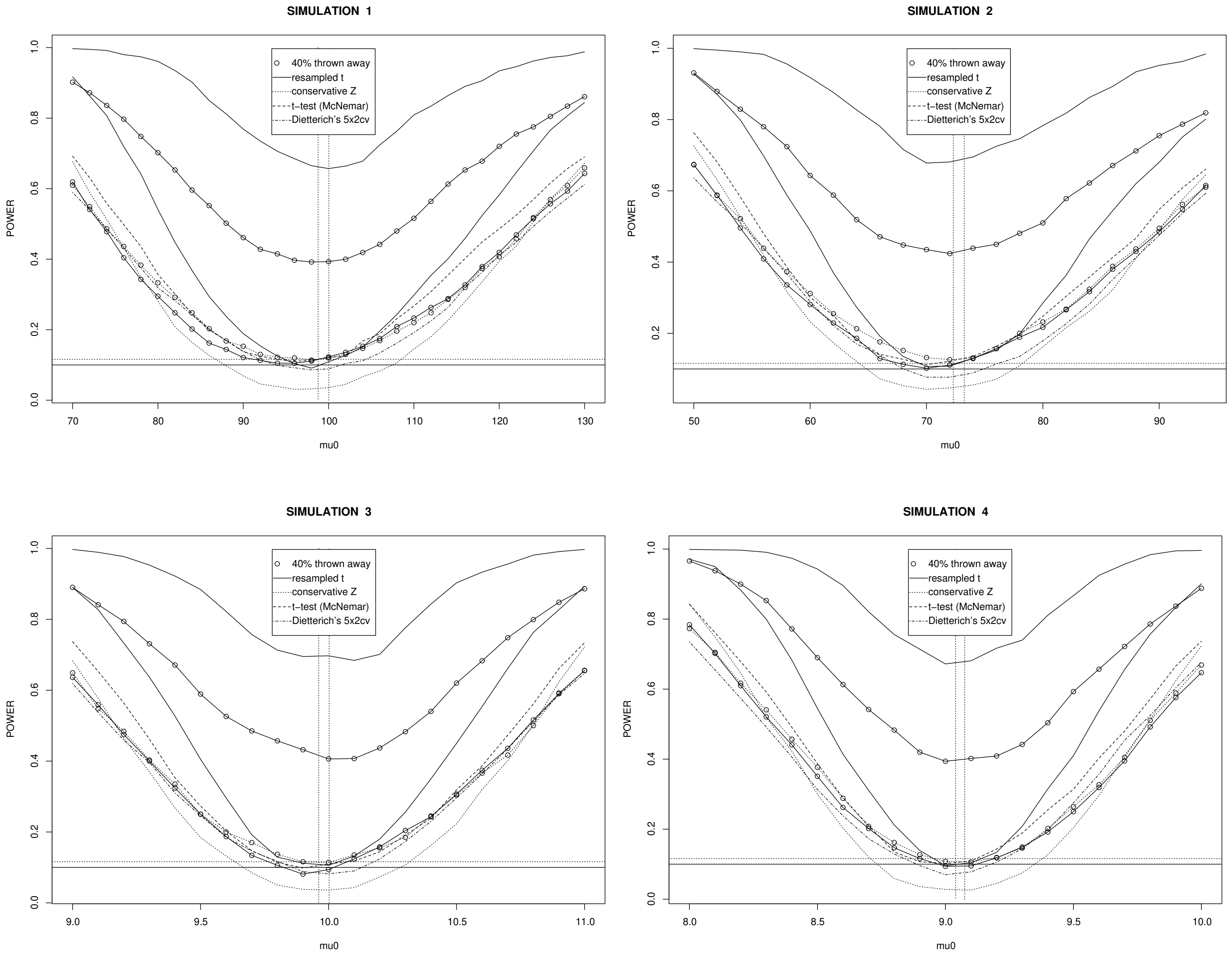


Figure 1: Powers of the tests about H_0 : $\frac{n}{2}\mu_A = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{n}{2}\mu_A$ shown in Table 2, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%). Where it matters $J = 15$, $M = 10$ and $R = 15$ were used.

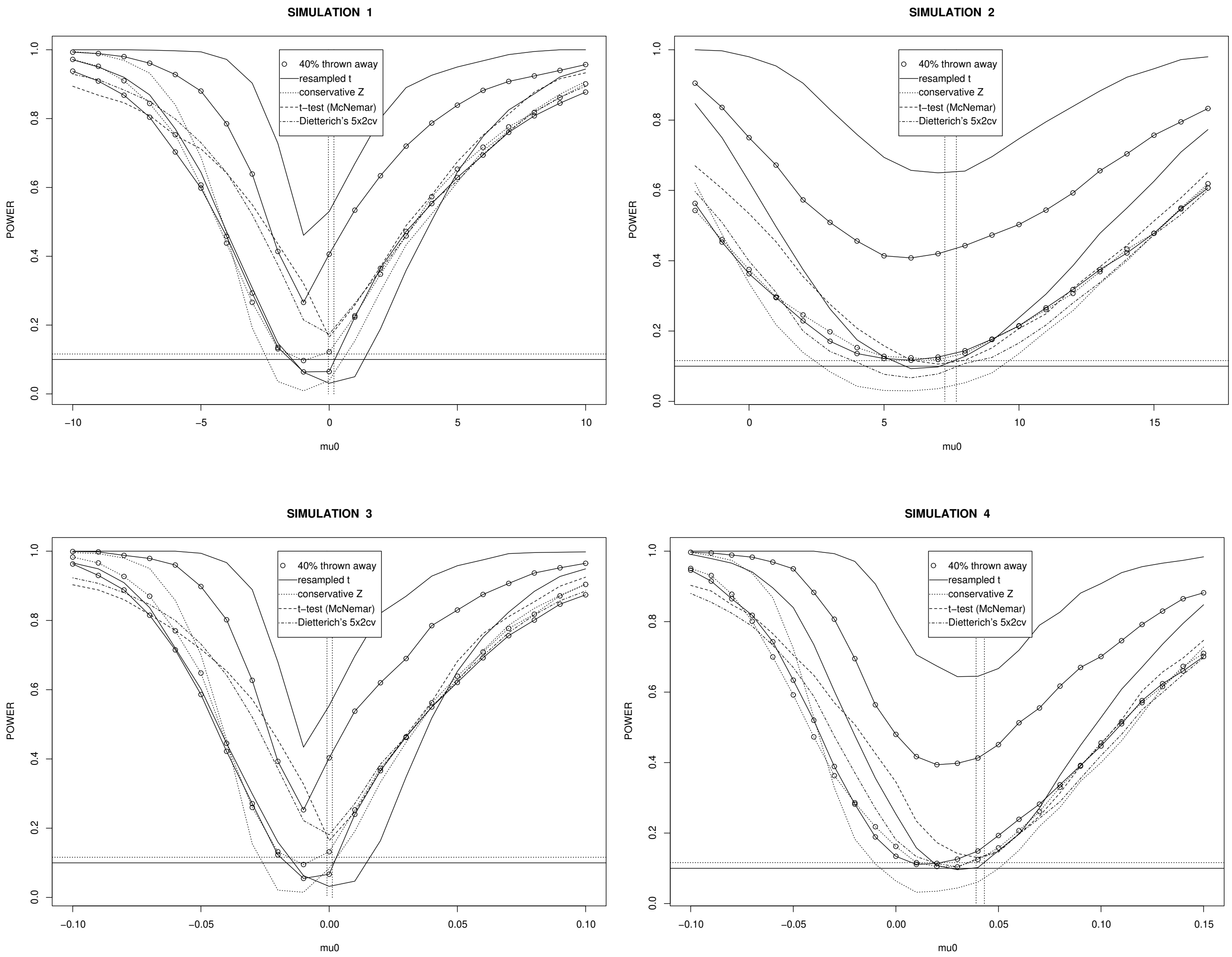


Figure 2: Powers of the tests about $H_0: \frac{n}{2}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{n}{2}\mu_{A-B}$ shown in Table 2, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%). Where it matters $J = 15$, $M = 10$ and $R = 15$ were used.

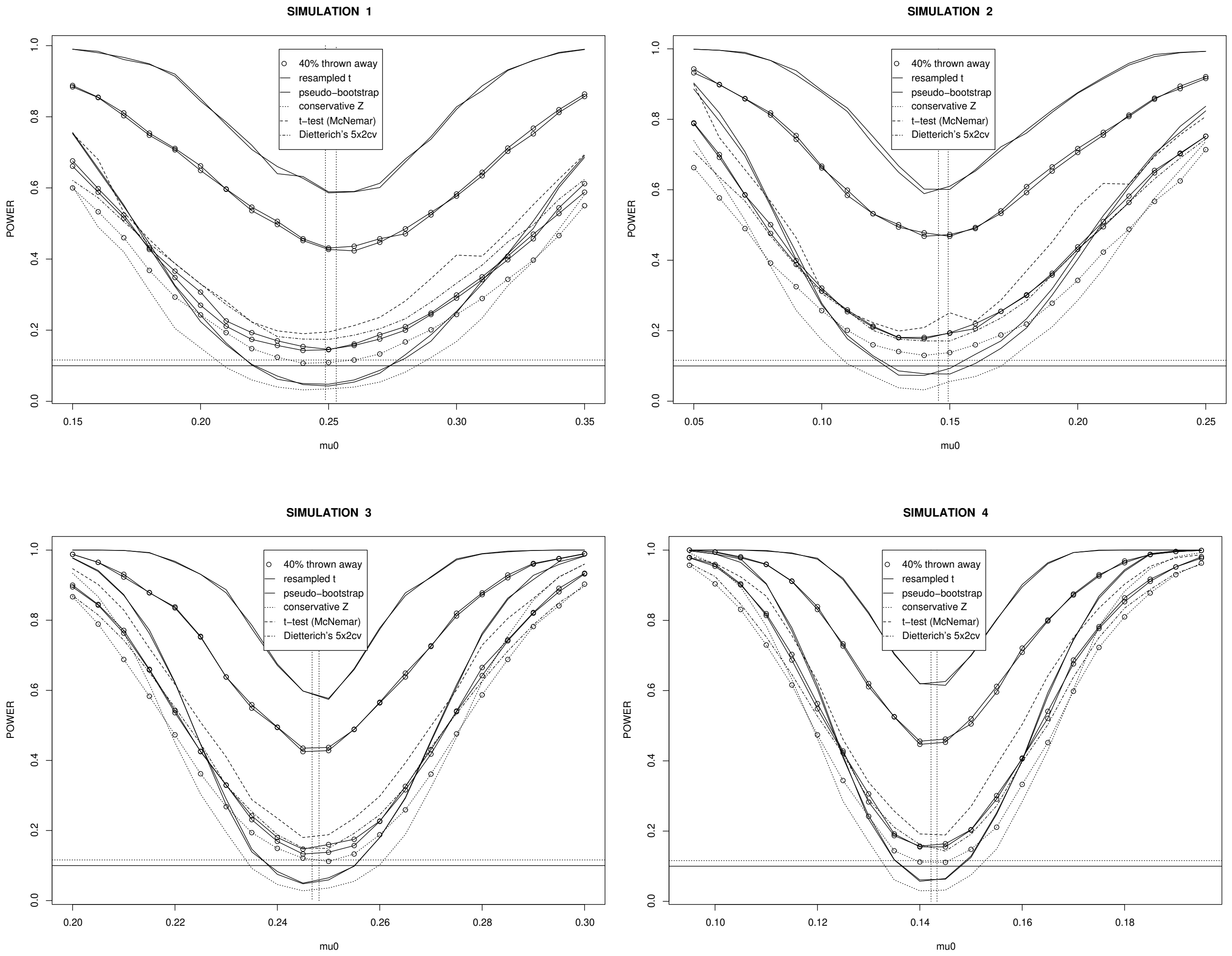


Figure 3: Powers of the tests about $H_0: \frac{n}{2}\mu_A = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{n}{2}\mu_A$ shown in Table 3, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%). Where it matters $J = 15$, $M = 10$ and $R = 15$ were used.

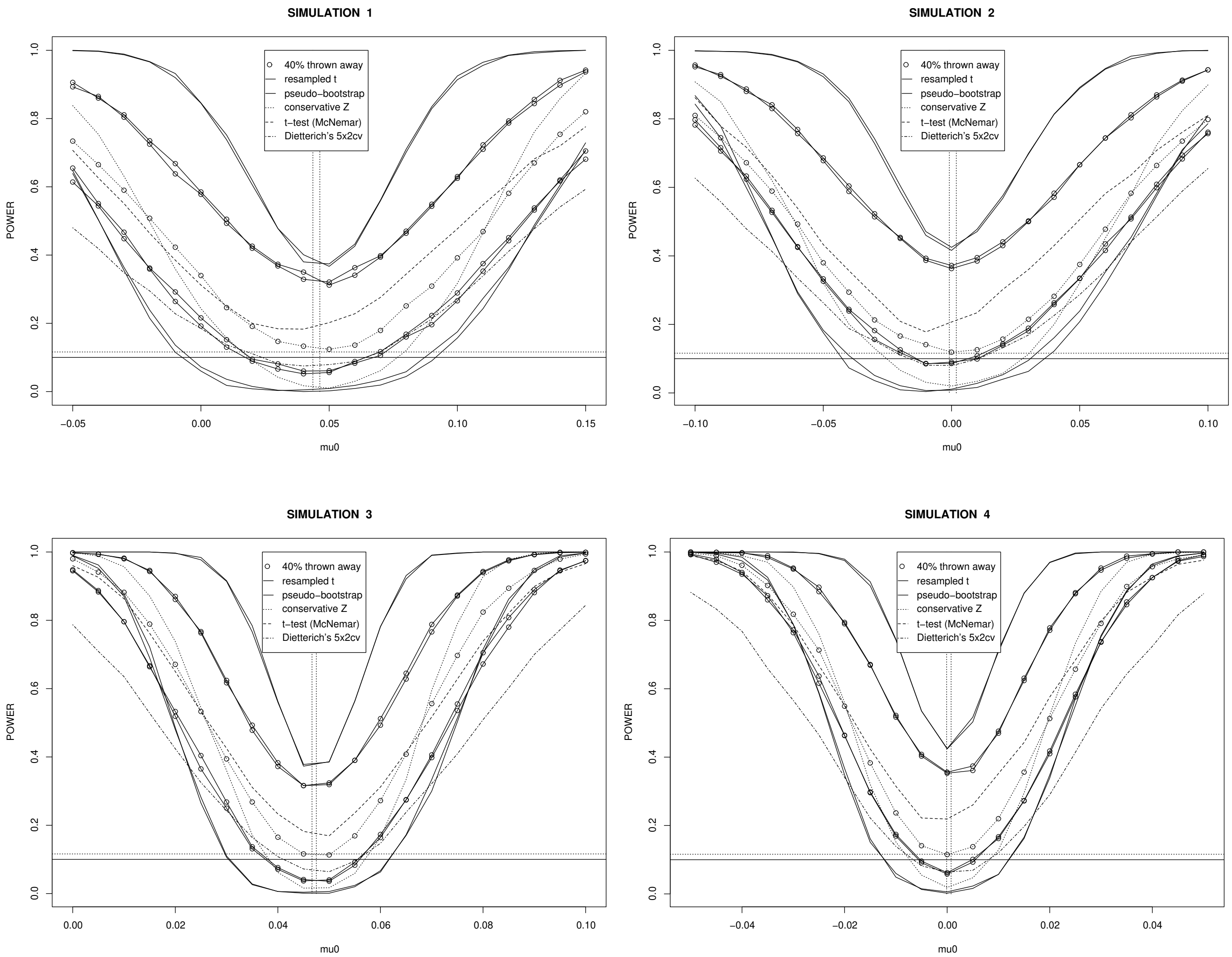


Figure 4: Powers of the tests about $H_0: \frac{n}{2}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{n}{2}\mu_{A-B}$ shown in Table 3, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%). Where it matters $J = 15$, $M = 10$ and $R = 15$ were used.

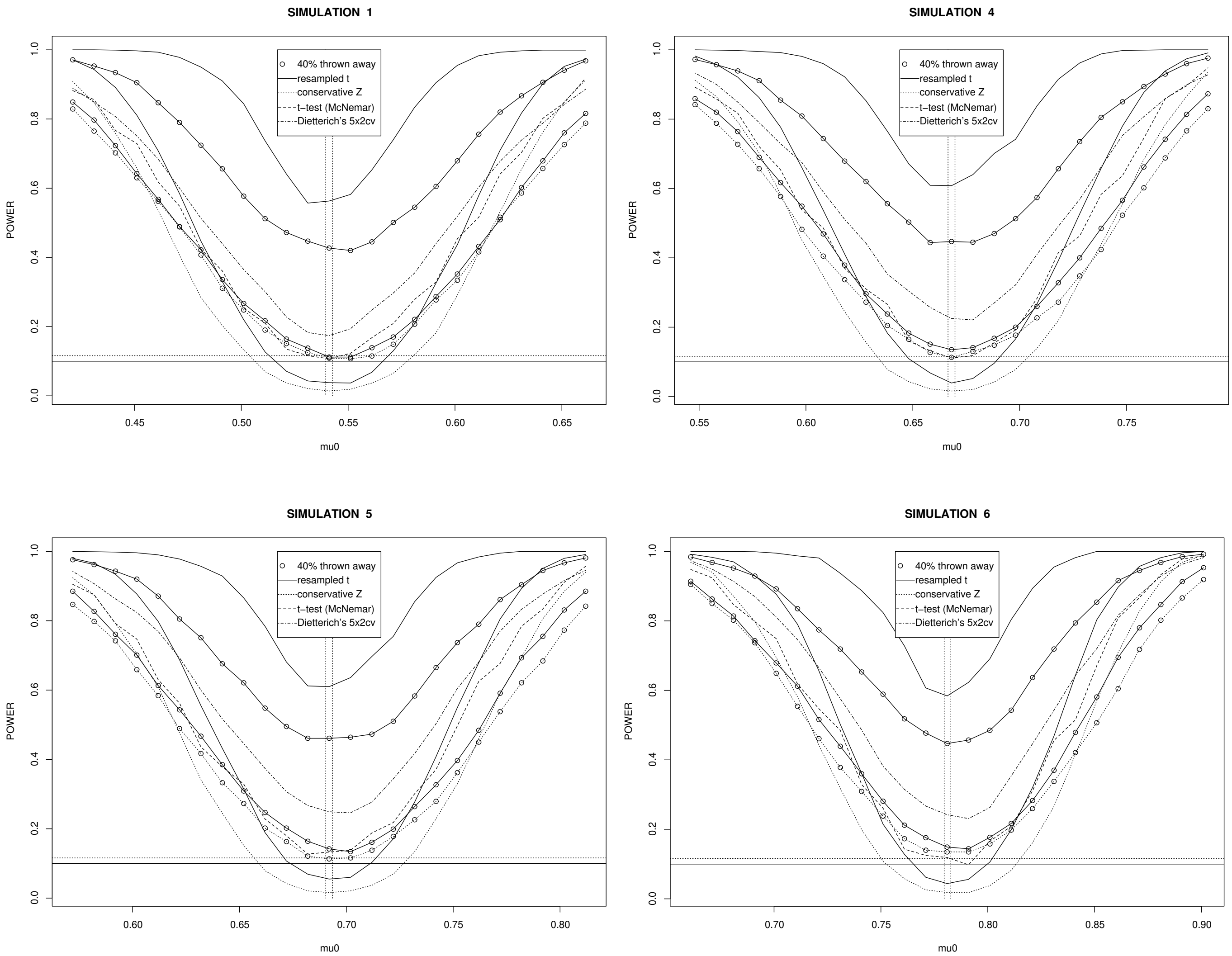


Figure 5: Powers of the tests about $H_0: \frac{n}{2}\mu_B = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the letter recognition problem. Each panel corresponds to one of the simulations design described in Table 4. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{n}{2}\mu_B$ shown in Table 4, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%). Where it matters $J = 15$, $M = 10$ and $R = 15$ were used.

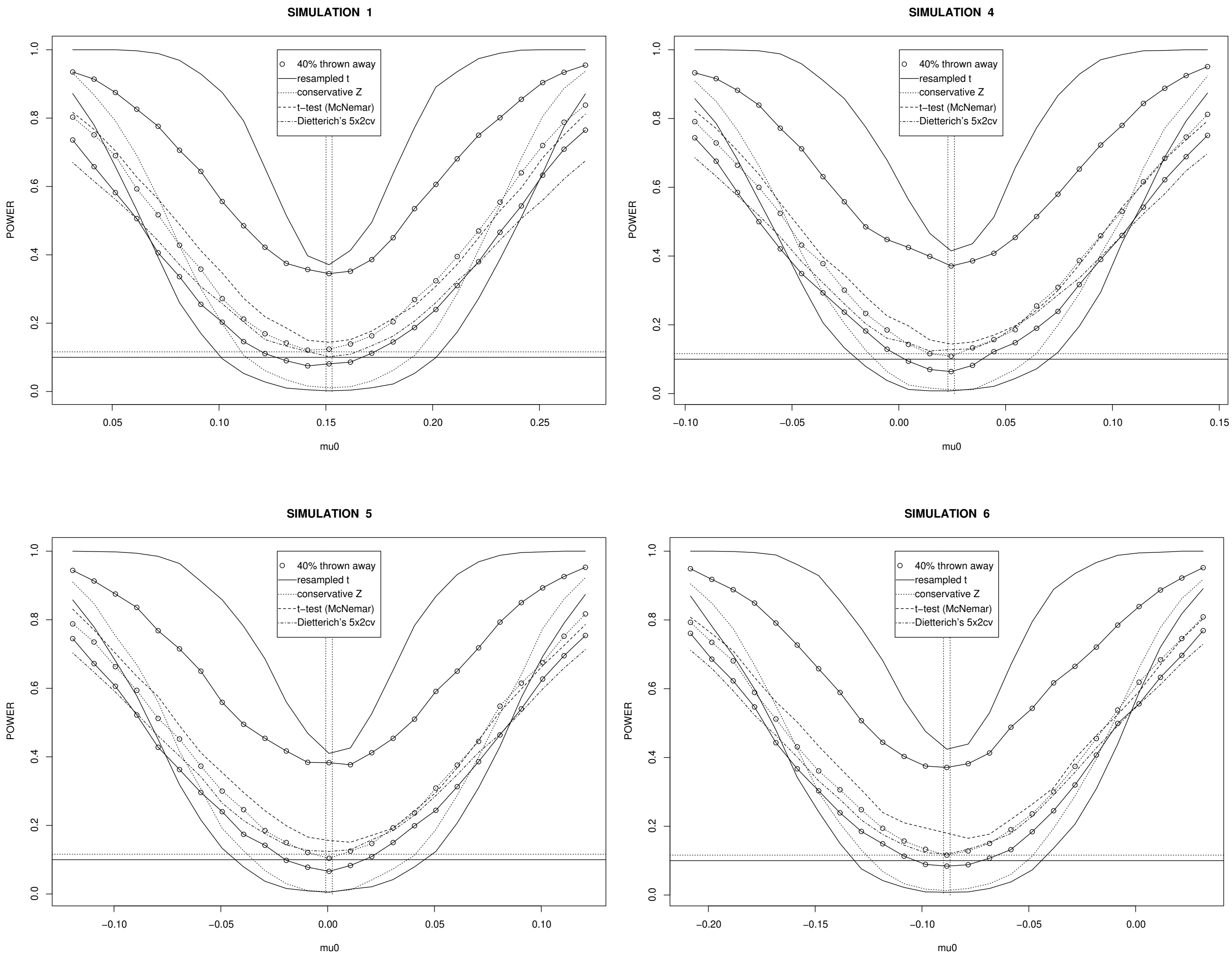


Figure 6: Powers of the tests about $H_0 : \frac{n}{2}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the letter recognition problem. Each panel corresponds to one of the simulations design described in Table 4. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{n}{2}\mu_{A-B}$ shown in Table 4, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%). Where it matters $J = 15$, $M = 10$ and $R = 15$ were used.

expensive. Nevertheless, pseudo-power curves shown in Figure 1 convey information similar to conventional "power curves". Indeed, we can find the size of a test by reading its pseudo-power curve at the actual value of ${}_{n/2}\mu_A$ (laying between the two vertical dotted lines). We can also appreciate the progression of the power as the hypothesized value of ${}_{n/2}\mu_A$ and the actual value of ${}_{n/2}\mu_A$ grow apart. We shall see in Figure 7 that those pseudo-power curves are good surrogate to "power curves".

Figures 2 through 6 are counterparts of Figure 1 for other problems and/or algorithms. Power plots corresponding to tests about ${}_{n/2}\mu_B$ in the regression problem and about ${}_{n/2}\mu_B$ in the classification of Gaussian populations problem are not shown since they convey the same information as Figure 1. However, missing figures are available in (Nadeau & Bengio, 1999).

Note that in order to reduce the number of displayed line types in Figure 1 and its counterparts appearing later, some curves share the same line type. So one must take note of the following.

- In a given panel, you will see four solid curves. They correspond to either the resampled t-test or the corrected resampled t-test with $n_2 = \frac{n}{10}$ or $n_2 = \frac{n}{2}$. Curves with circled points correspond to $n_2 = \frac{n}{10}$ (40% thrown away); curves without circled points correspond to $n_2 = \frac{n}{2}$. Telling apart the resampled t-test and the corrected resampled t-test is easy; the two curves that are well above all others correspond to the resampled t-test.

- The dotted curves depict the conservative Z test with either $n_2 = \frac{n}{10}$ (when it is circled) or $n_2 = \frac{n}{2}$ (when it is not circled).

- You might have noticed that the pseudo-bootstrap and the corrected pseudo-bootstrap do not appear in Figure 1 and all its counterparts (except Figure 3 and Figure 4). We ignored them because, as we anticipated from political ratios shown in Table 1, the pseudo-bootstrap test behaves like the resampled t-test and the corrected pseudo-bootstrap test behaves like the corrected resampled t-test. If we don't ignore the pseudo-bootstrap, some figures become too crowded. We made an exception and plotted curves corresponding to the pseudo-bootstrap in Figures 3 and 4. In those two figures, the pseudo-bootstrap and corrected pseudo-bootstrap curves are depicted with solid curves (just like the resampled t-test and corrected resampled t-test) and obey the same logic that applies to resampled t-test and corrected resampled t-test curves. What you must notice is that these figures look like the others except that where you would have seen a single solid curve, you now see two solid curves that nearly overlap. That shows how similar the resampled t-test and the pseudo-bootstrap are. This similitude is present for all problems, no just for the inference about ${}_{\frac{n}{2}}\mu_A$ or ${}_{\frac{n}{2}}\mu_{A-B}$ in the classification of Gaussian populations (Figures 3 and 4). We chose to show the pseudo-bootstrap curves in Figures 3 and 4 because this is where the plots looked the least messy when the pseudo-bootstrap curves were added.

Here's what we can draw from those figures.

- The most striking feature of those figures is that the actual size of the resampled t-test and the pseudo-bootstrap procedure are far away from the nominal size 10%. This is what we expected in Section 4. The fact that those two statistics are more liberal when $n_2 = \frac{n}{2}$ than they are when $n_2 = \frac{n}{10}$ (40% of the data thrown away) suggests that $\rho(n_1, n_2)$ is increasing in n_2. This is in line with what one can see in Tables 2, 3 and 4, and the simple approximation $\rho_0(n_1, n_2) = \frac{n_2}{n_1+n_2}$.

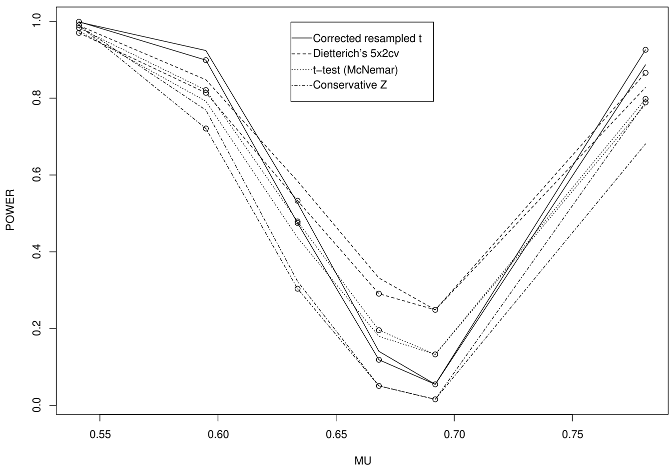
- We see that the sizes of the corrected resampled t-test (and corrected pseudo-bootstrap) are in line with what we could have forecasted from Tables 2, 3 and 4. Namely the test is liberal when $\rho(n_1, n_2) > \rho_0(n_1, n_2)$, conservative when $\rho(n_1, n_2) < \rho_0(n_1, n_2)$, and pretty much on target when $\rho(n_1, n_2)$ does not differ significantly from $\rho_0(n_1, n_2)$. For instance, on Figure 1 the sizes of the corrected resampled t-test are close to the nominal 10%. We see in Table 2 that $\rho_A(n_1, n_2)$ does not differ significantly from $\rho_0(n_1, n_2)$. Similarly, in Figures 3 and 5, the corrected resampled t-test appears to be significantly liberal when $n_2 = \frac{n}{10}$ (40% of the data thrown away) [14]. We see that $\rho_A(\frac{n}{2}, \frac{n}{10})$ is significantly greater than $\rho_0(\frac{n}{2}, \frac{n}{10}) = \frac{1}{6}$ in Table 3, and $\rho_B(\frac{n}{2}, \frac{n}{10})$ is significantly greater than $\rho_0(\frac{n}{2}, \frac{n}{10}) = \frac{1}{6}$ in Table 4. However, in those same figures, we see that the corrected resampled t-test that do not throw data away is conservative and, indeed, we can see that $\rho_A(\frac{n}{2}, \frac{n}{2})$ is significantly smaller than $\rho_0(\frac{n}{2}, \frac{n}{2}) = \frac{1}{2}$ in Table 3, and $\rho_B(\frac{n}{2}, \frac{n}{2})$ is significantly smaller than $\rho_0(\frac{n}{2}, \frac{n}{2}) = \frac{1}{2}$ in Table 4.

- The conservative Z with $n_2 = \frac{n}{2}$ is too conservative. However, when $n_2 = \frac{n}{10}$ (so that $\frac{n_1}{n_2} = 5$, more in line with normal usage), the conservative Z has more interesting properties. It does not quite live up to its name since it is at times liberal, but barely so. Its size is never very far from 10% (like 20% for instance), making it the best inference procedure among those considered in terms of size.

- The t-test and Dietterich's 5×2 cv are usually well behaved in term of size, but they are sometimes fairly liberal as can be seen in some panels of Figures 2, 3 and 4.

- When their sizes are comparable, the powers of the t-test, Dietterich's 5×2 cv, conservative Z throwing out 40% of the data and corrected resampled t-test throwing out 40% of the data are fairly similar. If we have to break the tie, it appears that the t-test is the most powerful, Dietterich's 5×2 cv is the least powerful procedure and the corrected resampled t-test and the corrected conservative Z lay in between. The fact that the conservative Z and the corrected resampled t-test perform well despite throwing 40% of

[14]Actually in Figure 2 we do see that the corrected resampled t-test with $n_2 = \frac{n}{10}$ is liberal in Simulations 2 and 4 despite the fact that $\rho_{A-B}(\frac{n}{2}, \frac{n}{10})$ do not differ significantly from $\frac{1}{6}$ in Simulation 2 and $\rho_{A-B}(\frac{n}{2}, \frac{n}{10})$ is barely significantly smaller than $\frac{1}{6}$ in Simulation 4. But, as mentioned in Appendix A.0, the political ratio $\frac{Var[\hat{\mu}]}{E[\hat{\sigma}^2]}$ is not the only thing determining whether inference is liberal or conservative. What happens in this particular case is that the distribution of ${}^{n_2}_{n_1}\hat{\mu}_{15}$ is asymmetric; ${}^{n_2}_{n_1}\hat{\mu}_1$ did not appear to suffer from this problem. The comparison of algorithm A and B for the regression problem is the only place where this phenomenon was substantial in our simulation. That is why curves (other than t-test and Dietterich's 5×2 cv that are based on ${}^{n_2}_{n_1}\hat{\mu}_1$) are asymmetric and bottom out before the actual value of ${}_{n/2}\mu_{A-B}$ (laying between the vertical dotted lines). We don't observe this in other figures.



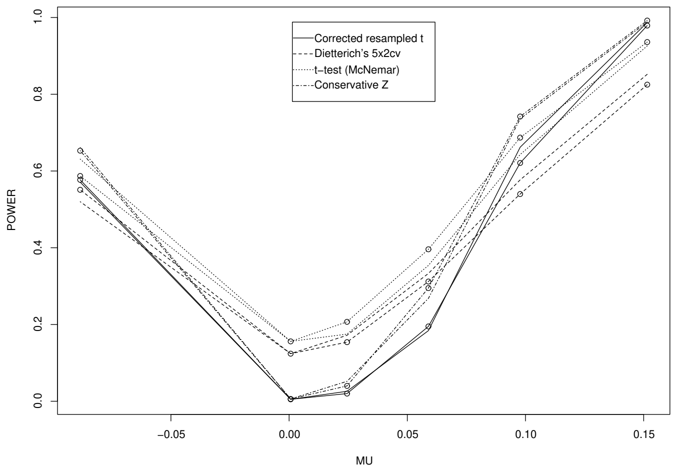


Figure 7: Real power curves (circle lines) and pseudo-power curves (not circled) in the letter recognition problem. In the left panel, we see "real" and "pseudo" power curves for the the null hypothesis $H_0 :\ {}_{150}\mu_B = 0.692$. In the right panel, we see "real" and "pseudo" power curves for the the null hypothesis $H_0 :\ {}_{150}\mu_{A-B} = 0.001$. See the end of Section 5.1 for more details on their constructions. Here, the "corrected resampled t" and the "conservative Z" statistics are those which do not throw away data.

the data indicates that these methods are very powerful compared to Dietterich's 5×2 cv and the t-test. This may be seen in Figure 1 where the size of the corrected resampled t-test with the full data is comparable to the size of other tests. The power of the corrected resampled t-test is then markedly superior to the powers of other tests with comparable size. In other figures, we see the power of the corrected resampled t-test with full data and/or conservative Z with full data catch on (as we move away from the null hypothesis) the powers of other methods that have larger size.

As promised earlier, we now illustrate that pseudo-power curves are good surrogates to actual real power curves. For the letter recognition problem, we have the opportunity to draw real power curves since we have simulated data under six different schemes. Recall from Table 4 that we have simulated data with ${}_{150}\mu_B$ approximatively equal to $0.541, 0.595, 0.634, 0.668, 0.692, 0.781$ and ${}_{150}\mu_{A-B}$ approximatively equal to $0.151, 0.098, 0.059, 0.025, 0.001, -0.088$ in Simulations 1 through 6 respectively. The circled lines in Figure 7 depict real power curves. For instance, in the left panel, the power of tests for $H_0 :\ {}_{150}\mu_B = 0.692$ has been obtained in all six simulations, enabling us to draw the circled curves. The non-circled curves correspond to what we have been plotting so far. Namely, in Simulation 5, we computed the powers of tests for $H_0 :\ {}_{150}\mu_B = \mu_0$ with $\mu_0 = 0.541, 0.595, 0.634, 0.668, 0.692, 0.781$, enabling us to draw the non-circled curves. We see that circled and non-circled curves agree relatively well, leading us to believe that our previous plots are good surrogates to real power curves.

5.2 The choice of J

In Section 5.1, the statistics involving ${}^{n_2}_{n_1}\hat{\mu}_J$ used $J = 15$. We look at how those statistics behave with varying J's, in order to formulate a recommendation on the choice of J. We are going to do so with $n_1 = \frac{9n}{10}$ and $n_2 = \frac{n}{10}$, which correspond to a more natural usage for these statistics. Of the seven statistics displayed in Section 4 (see also Table 1), five involved ${}^{n_2}_{n_1}\hat{\mu}_J$. We ignore the pseudo-bootstrap and the corrected pseudo-bootstrap as political ratios provided in Section 4 and empirical evidence in Section 5.1 suggest that these statistics are virtually identical to the resampled t-test and the corrected resampled t-test (but require a lot more computation). We therefore only consider the resampled t-test, the corrected resampled t-test and the conservative Z here.

The investigation of the properties of those statistics will again revolve around their sizes and powers. You will therefore see that figures in this section (Figures 8 to 12) are similar to those of the Section 5.1. Note that figures corresponding to ${}_{9n/10}\mu_B$ are not shown as they convey no additional information. However, missing figures are available in (Nadeau & Bengio, 1999). In a given plot, we see the powers of the three statistics when $J = 5$, $J = 10$, $J = 15$ and $J = 25$. Therefore a total of twelve curves are present in each plot.

Here's what we can draw from those figures.

- Again, the first thing that we see is that the resampled t-test is very liberal. However, things were even worse in Section 5.1. That is due to the fact that $\rho(\frac{9n}{10}, \frac{n}{10})$ is smaller than $\rho(\frac{n}{2}, \frac{n}{10})$ and $\rho(\frac{n}{2}, \frac{n}{2})$. We also see that the statistic is more liberal when J is large, as it should be according to the theoretical discussion of that statistic in Section 4.

- The conservative Z lives up to its name.

- Regarding the corrected resampled t-test, the plots again only confirm what we might have guessed from Tables 2, 3 and 4. Namely the resampled t-test is conservative when $\rho(\frac{9n}{10}, \frac{n}{10})$ is significantly greater than $\rho_0(\frac{9n}{10}, \frac{n}{10}) = 0.1$, liberal when $\rho(\frac{9n}{10}, \frac{n}{10})$ is significantly smaller then 0.1, and has size very close to 0.1 otherwise. When it is liberal or conservative, things tend to grow worse when J increases; see Figure 10 for the liberal case. That makes sense since the political ratio $\frac{Var[\hat{\mu}]}{E[\hat{\sigma}^2]} = \frac{1+J\frac{\rho}{1-\rho}}{1+J\frac{n_2}{n_1}}$ (see Table 1) is monotonic in J (increasing when $\rho > \frac{n_2}{n_1+n_2}$; decreasing when $\rho < \frac{n_2}{n_1+n_2}$).

- Obviously (from equation (8) or Proposition 2), the greater J is, the greater the power will be. Note that increasing J from 5 to 10 brings about half the improvement in the power obtained by increasing J from 5 to 25. Similarly, increasing J from 10 to 15 brings about half the improvement in the power obtained by increasing J from 10 to 25. With that in mind, we feel that one must take J to be at least equal to 10 as $J = 5$ leads to unsatisfactory power. Going beyond $J = 15$ gives little additional power and is probably not worth the computational effort. We could tackle this question from a

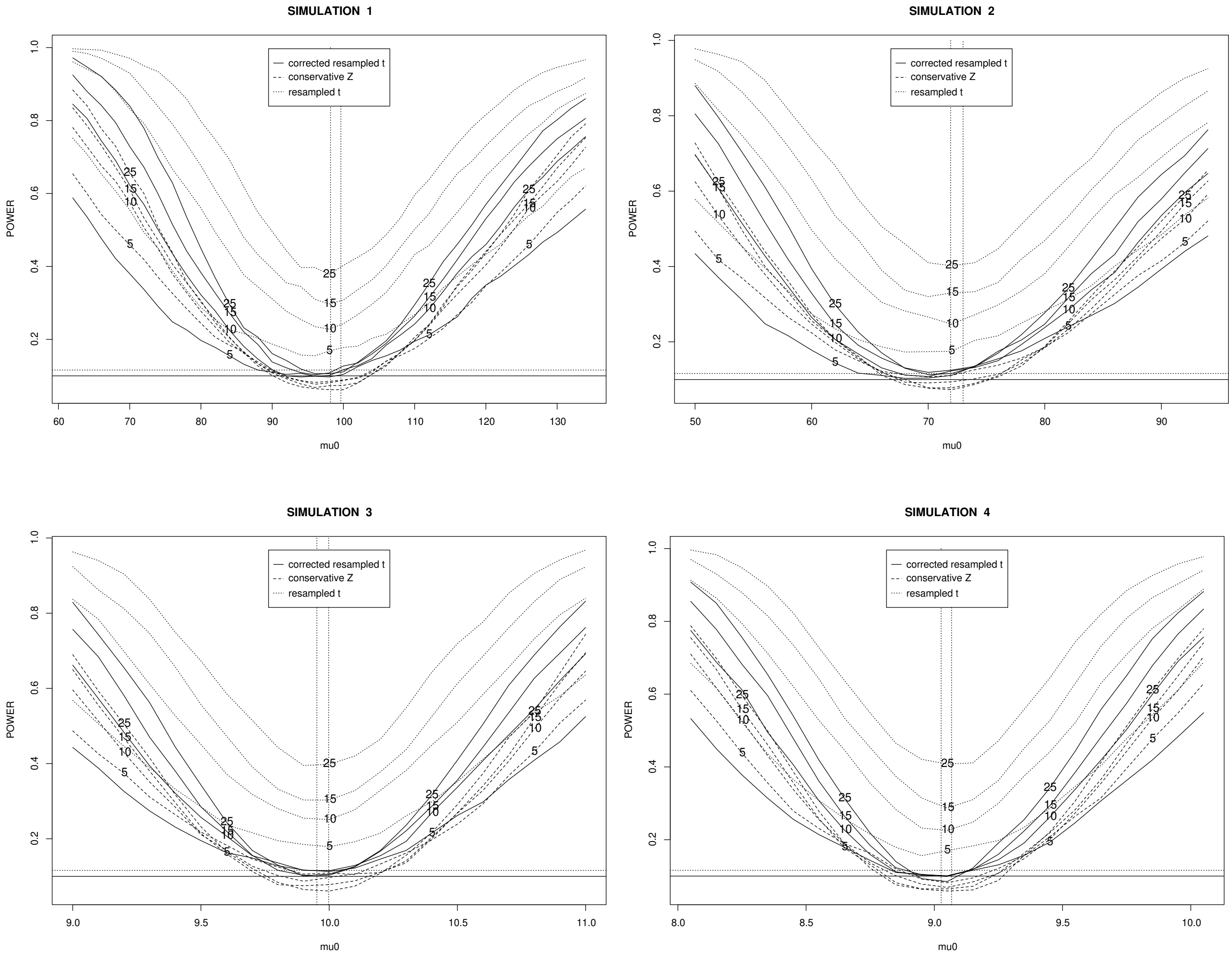


Figure 8: Powers of the tests about H_0 : ${}_{9n/10}\mu_A = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual ${}_{9n/10}\mu_A$ shown in Table 2, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%). For the conservative Z, $M = 10$ was used.

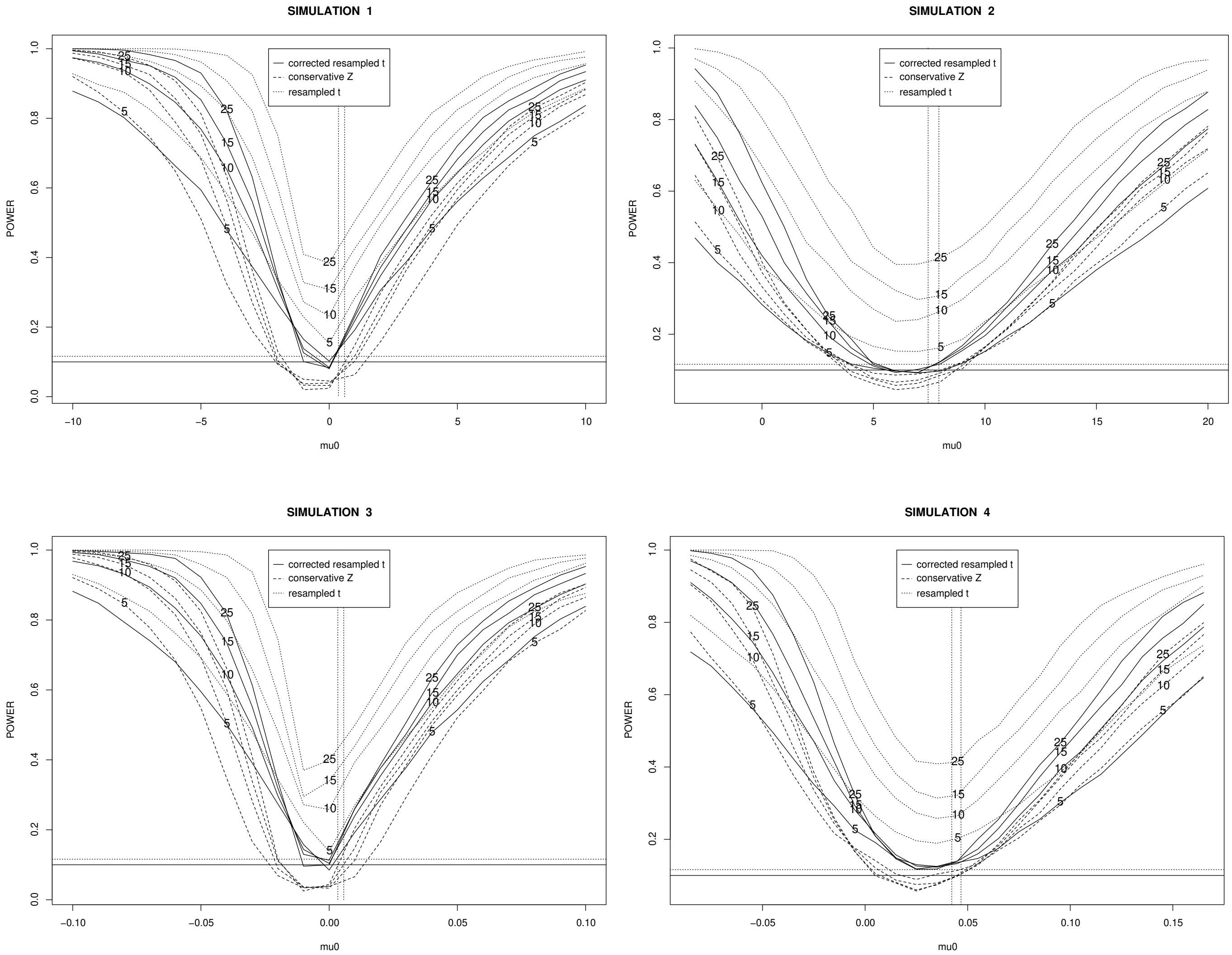


Figure 9: Powers of the tests about H_0 : ${}_{9n/10}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual ${}_{9n/10}\mu_{A-B}$ shown in Table 2, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%). For the conservative Z, $M = 10$ was used.

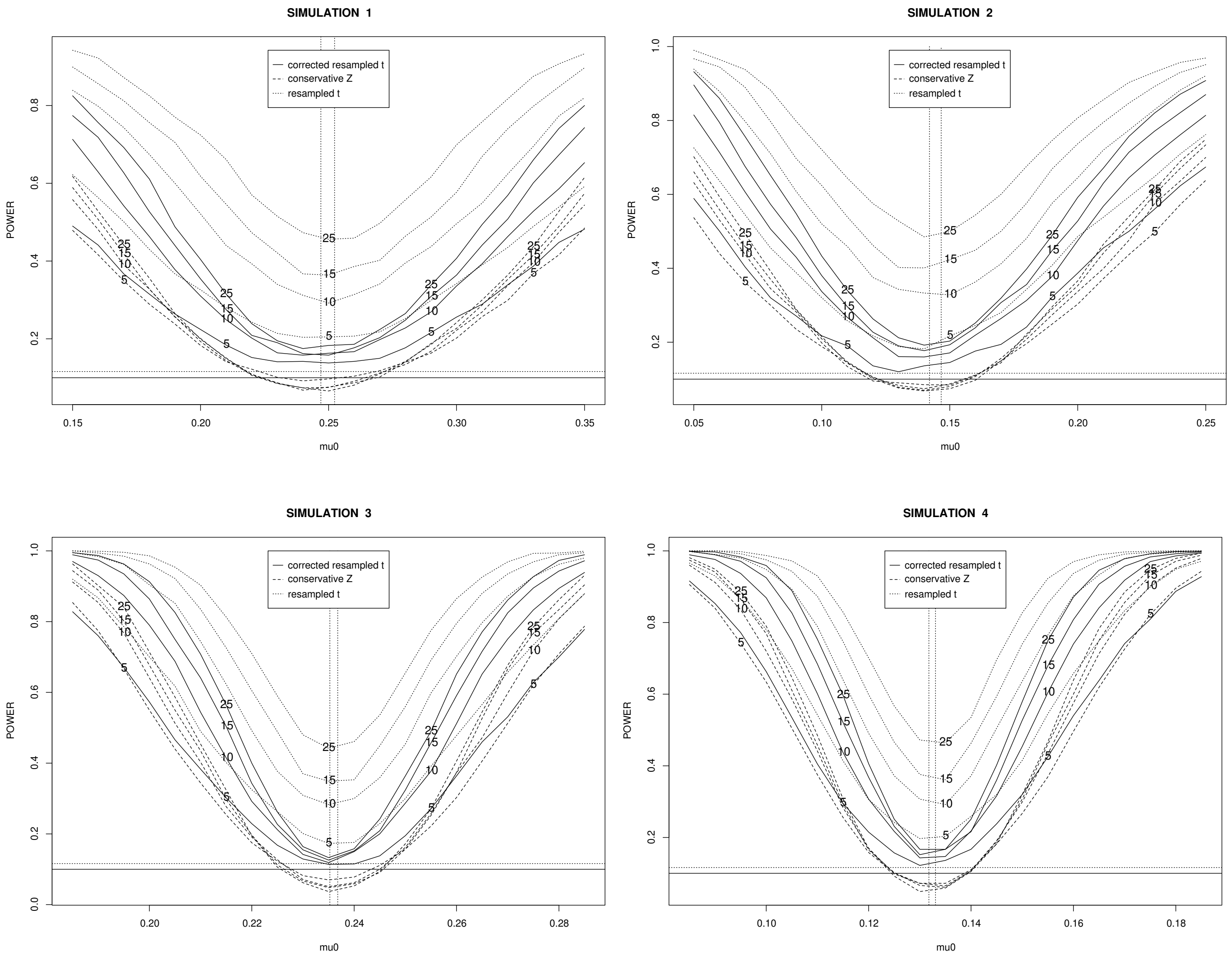


Figure 10: Powers of the tests about $H_0 : \frac{9n}{10}\mu_A = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{9n}{10}\mu_A$ shown in Table 3, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%). For the conservative Z, $M = 10$ was used.

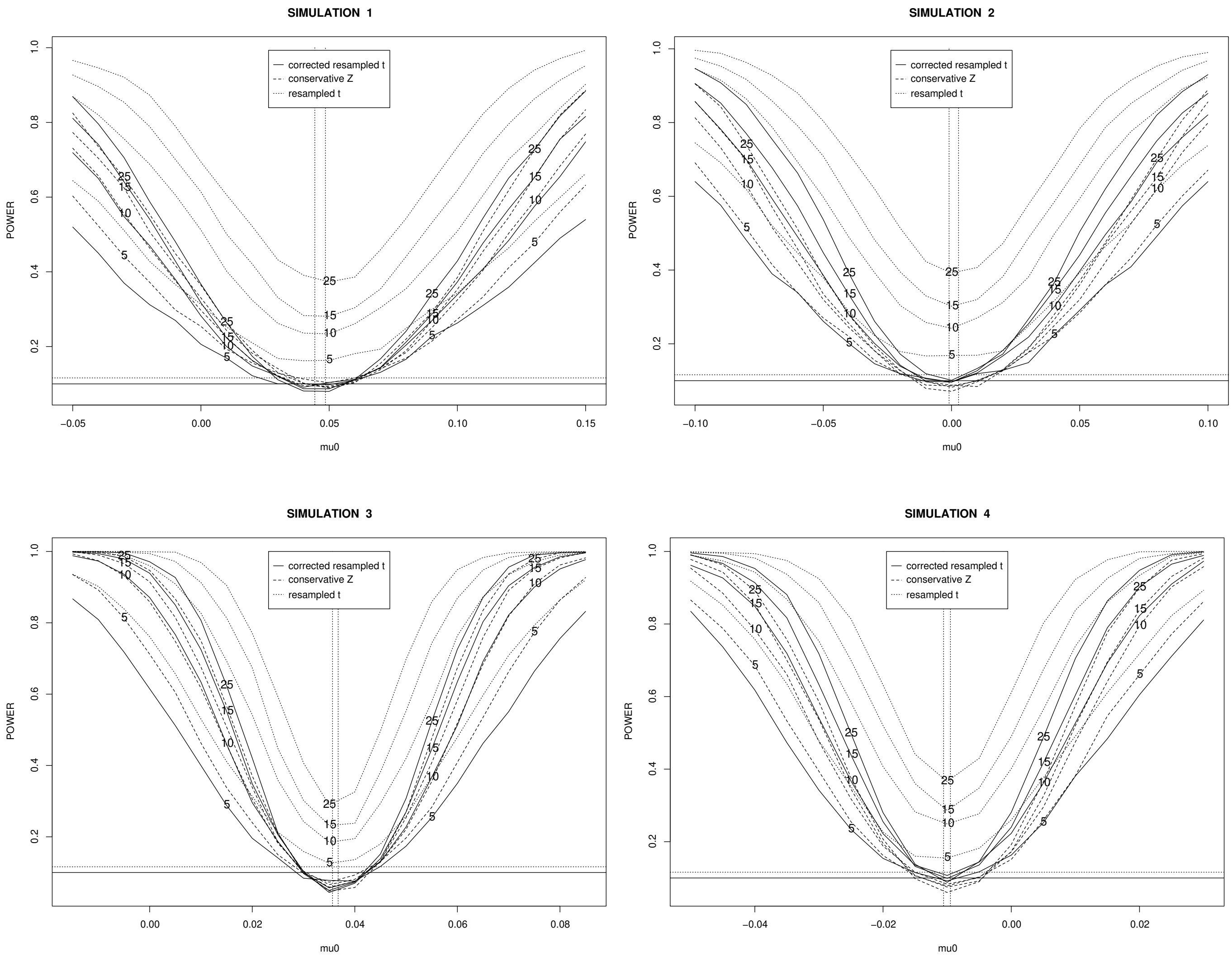


Figure 11: Powers of the tests about $H_0 : \frac{9n}{10}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{9n}{10}\mu_{A-B}$ shown in Table 3, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%). For the conservative Z, $M = 10$ was used.

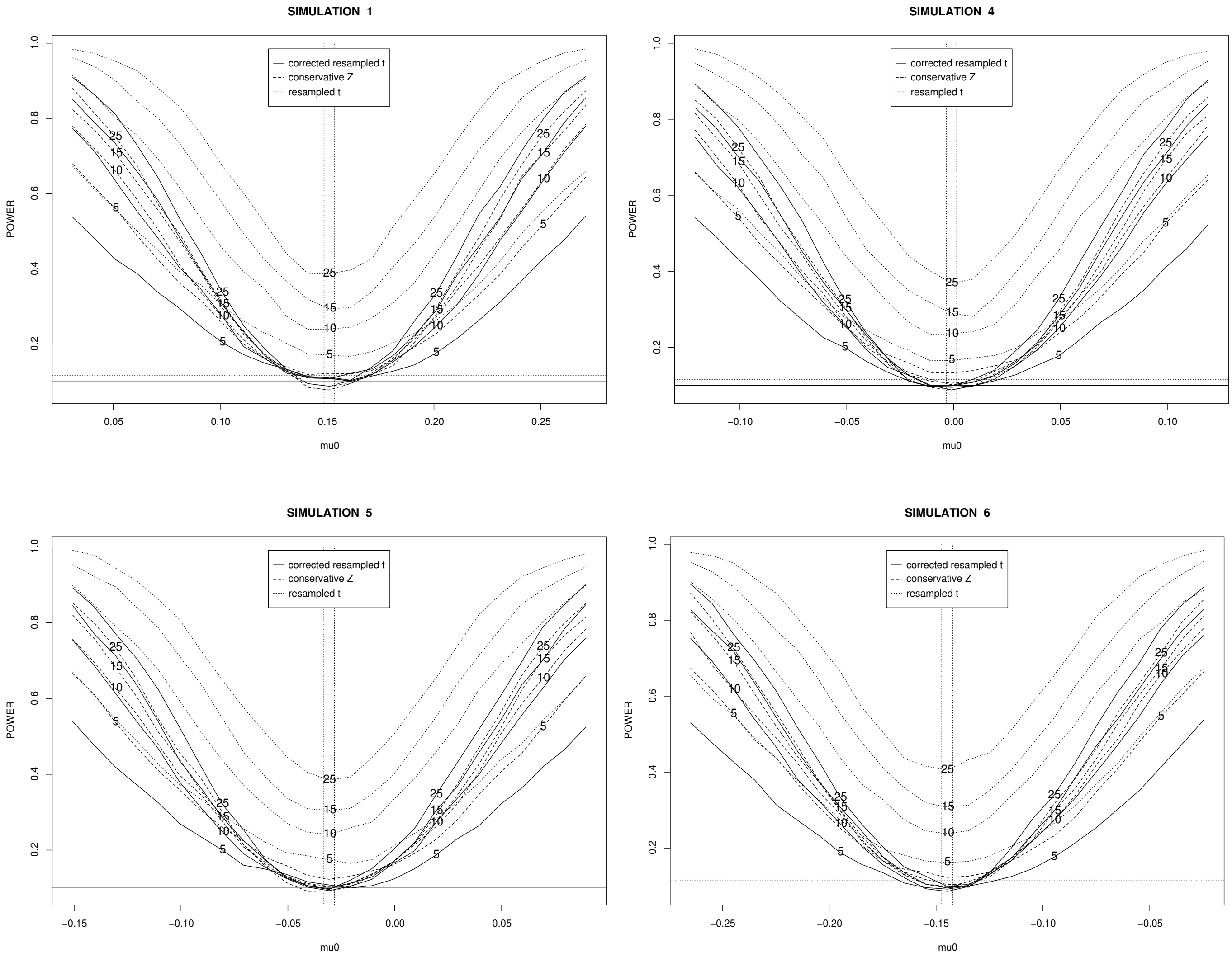


Figure 12: Powers of the tests about $H_0: \frac{9n}{10}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the letter recognition problem. Each panel corresponds to one of the simulations design described in Table 4. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{9n}{10}\mu_{A-B}$ shown in Table 4, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%). For the conservative Z, $M = 10$ was used.

theoretical point of view. We know from (8) that $Var[\ {}^{n_2}_{n_1}\hat{\mu}_J] = \sigma_1 \left(\rho + \frac{1-\rho}{J}\right)$. Take $\rho = 0.1$ for instance (that is $\rho_0(\frac{9n}{10}, \frac{n}{10})$). Increasing J from 1 to 3 reduces the variance by 60%. Increasing J from 3 to 9 further halves the variance. Increasing J from 9 to ∞ only halves the variance. We thus see that the benefit of increasing J quickly becomes faint.

- Since the conservative Z is fairly conservative, it rarely has the same size as the corrected resampled t-test, making power comparison somewhat difficult. But it appears that the two methods have equivalent powers which makes sense since they are both based on ${}^{n_2}_{n_1}\hat{\mu}_J$. We can see this in Figures 11 and 12 where the two tests have about the same size and similar power.

Based on the above observations, we believe that $J = 15$ is a good choice: it provides good power with reasonable computational effort. If computational effort is not an issue, one may take $J > 15$, but must not expect a great gain in power. Another reason in favor of not taking J too large is that the size of the resampled t-test gets worse with increasing J when that method is liberal or conservative.

Of course the choice of J is not totally independent of n_1 and n_2. Indeed, if one uses a larger test set (and thus a smaller train set), then we might expect ρ to be larger and therefore $J = 10$ might then be sufficiently large.

Although it is not related to the choice of J, we may comment on the choice of the inference procedure as figures in this section are the most informative in that regard. If one wants an inference procedure that is not liberal, the obvious choice is the conservative Z. However, if one prefers an inference procedure with size close to the nominal size α and is ready to accept departures in the liberal side as well as in the conservative side, then the corrected resampled t appears to be a good choice. However, as we shall see shortly, we can make the conservative Z more or less conservative by playing with M. The advantage of the corrected resampled t is that it requires little computing in comparison to the conservative Z.

Finally, as we did earlier, we may assess to what extent the pseudo-power curves shown in Figures 8 through 12 are good surrogates to actual real power curves. Remember that for the letter recognition problem, we have the opportunity to draw real power curves since we have simulated data under six different schemes. Recall from Table 4 that we have simulated data with ${}_{270}\mu_B$ approximatively equal to $0.437, 0.499, 0.546, 0.589, 0.618, 0.732$ and ${}_{270}\mu_{A-B}$ approximatively equal to $0.151, 0.089, 0.042, -0.001, -0.031, -0.145$ in Simulations 1 through 6 respectively. The circled lines in Figure 13 depict real power curves. For instance, in the left panel, the power of tests for $H_0:\ {}_{270}\mu_B = 0.589$ has been obtained in all six simulations, enabling us to draw the circled curves. The non-circled curves are the pseudo-power curves we usually plot in this paper. Namely, in Simulation 4, we computed the powers of tests for $H_0:\ {}_{270}\mu_B = \mu_0$ with $\mu_0 = 0.437, 0.499, 0.546, 0.589, 0.618, 0.732$, enabling us to draw the non-circled curves. We see that circled and non-circled curves agree very well, leading us to believe that our previous plots are good surrogates to real power curves.

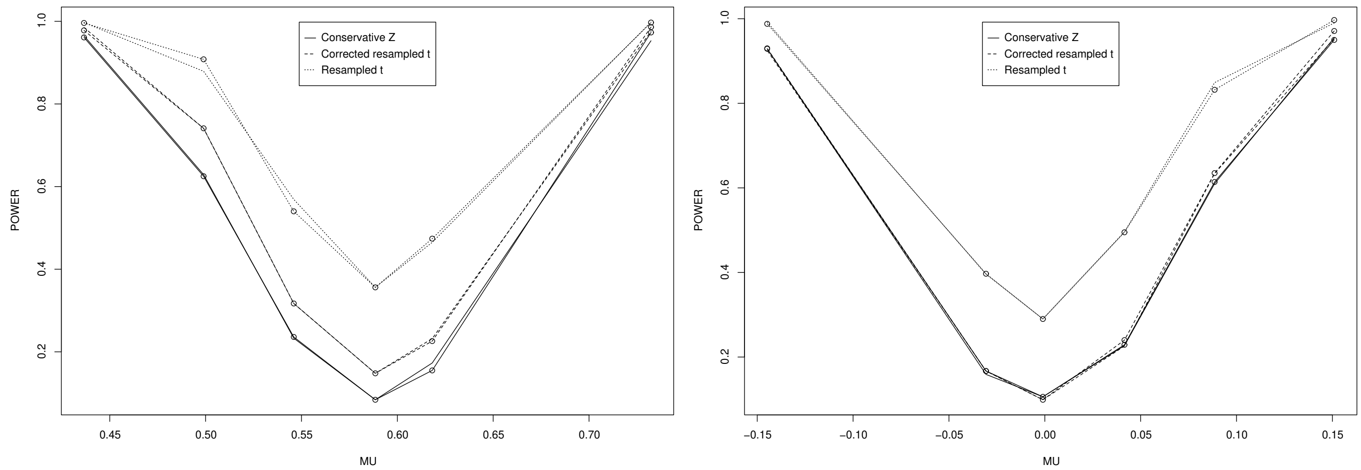


Figure 13: Real power curves (circle lines) and pseudo-power curves (not circled) in the letter recognition problem. In the left panel, we see "real" and "pseudo" power curves for the the null hypothesis H_0 : ${}_{270}\mu_B = 0.589$ In the right panel, we see "real" and "pseudo" power curves for the the null hypothesis H_0 : ${}_{270}\mu_{A-B} = -0.001$. See the end of Section 5.2 for more details on their constructions.

5.3 The choice of M

When using the conservative Z, we have so far always used $M = 10$. We study the behavior of this statistic for various values of M in order to formulate a recommendation on the choice of M. Again we consider the case where $n_1 = \frac{9n}{10}$ and $n_2 = \frac{n}{10}$. The investigation will again revolve around the size and power of the statistic. We see in Figure 14 (figures for other problems and/or algorithms convey the same information and are therefore not shown but are available in (Nadeau & Bengio, 1999)) that the conservative Z is more conservative when M is large. We see that there is not a great difference in the behavior of the conservative Z when $M = 10$ and when $M = 20$. For that reason, we recommend using $M \leq 10$. The difference between $M = 10$ and $M = 5$ is more noticeable, $M = 5$ leads to inference that is less conservative, which is not a bad thing considering that with $M = 10$ it tends to be a little bit too conservative. With $M = 5$, the conservative Z is sometimes liberal, but barely so. Using $M < 5$ would probably go against the primary goal of the statistic, that is provide inference that is not liberal. Thus $5 \leq M \leq 10$ appears to be a reasonable choice. Within this range, pick M large if non-liberal inference is important; otherwise take M small if you want the size of the test to be closer to the nominal size α (you then accept the risk of performing inference that could be slightly liberal). Of course, computational effort is linear in M so that taking M small has an additional appeal.

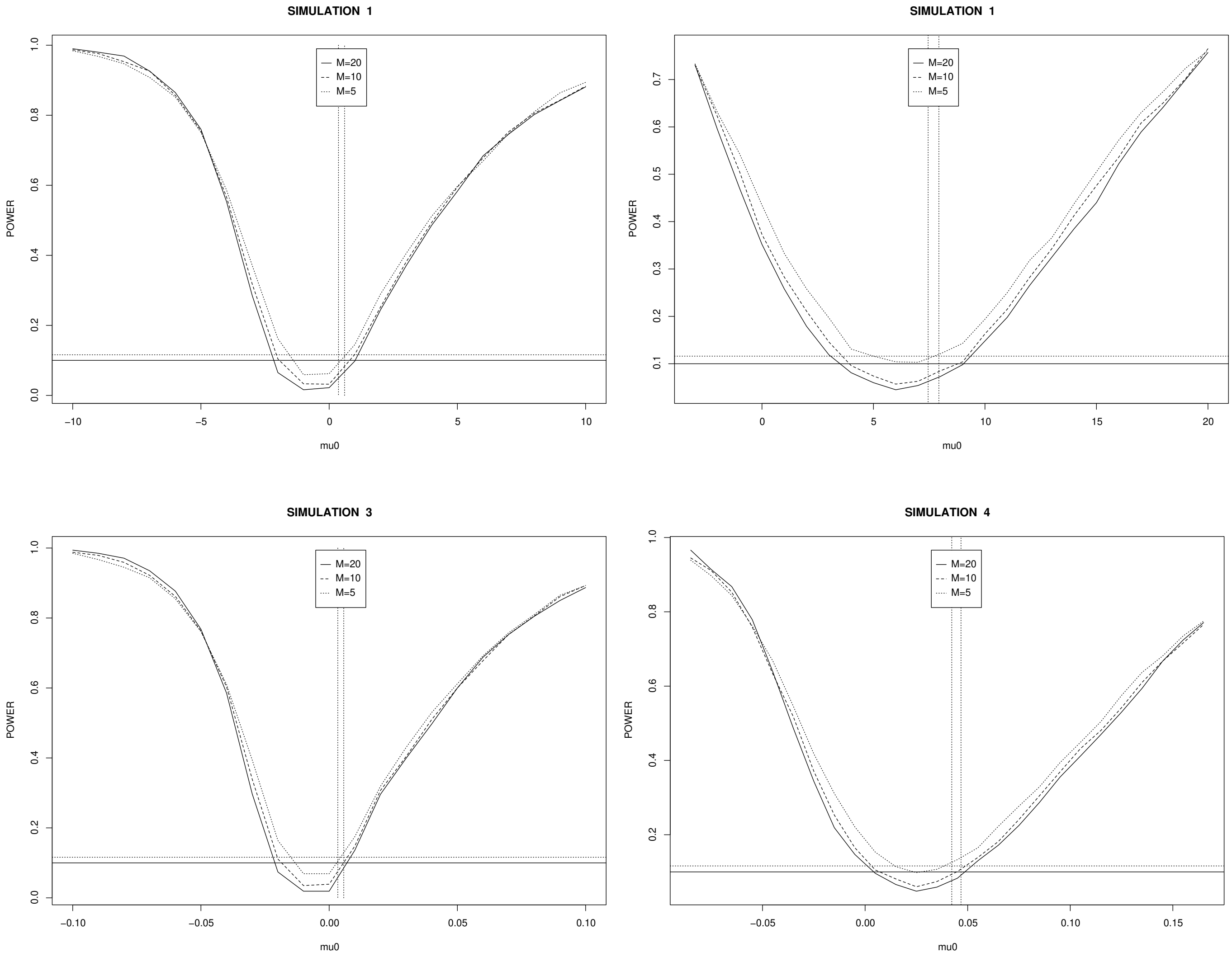


Figure 14: Powers of the conservative Z (with $J = 15$) about $H_0 : {}_{9n/10}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and M for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual ${}_{9n/10}\mu_{A-B}$ shown in Table 2, therefore that is where the actual size of the tests may be read. The solid horizontal line displays the nominal size of the tests, i.e. 10%. Estimated probabilities of rejection laying above the dotted horizontal line are significantly greater than 10% (at significance level 5%).

6 Conclusion

We have tackled the problem of estimating the variance of the cross-validation estimator of the generalization error. In this paper, we paid special attention to the variability introduced by the selection of a particular training set, whereas most empirical applications of machine learning methods concentrate on estimating the variability of the estimate of generalization error due to the finite test set.

A theoretical investigation of the variance to be estimated shed some valuable insight on reasons why some estimators currently in use underestimate the variance. We showed that no general unbiased estimator of the variance of the cross-validation estimator could be found. This analysis allowed us to construct two variance estimates that take into account both the variability due to the choice of the training sets and the choice of the test examples. One of the proposed estimators looks like the 5×2 cv method (Dietterich, 1998) and is specifically designed to overestimate the variance to yield conservative inference. The other may overestimate or underestimate the real variance, but is typically not too far off the target.

We performed a simulation where the new techniques put forward were compared to test statistics currently used in the machine learning community. We tackle both the inference for a generalization error of an algorithm and the comparison of the generalization errors of two algorithms. We considered two kinds of problems: classification and regression. Various algorithms were considered: linear regression, regression trees, classification trees and the nearest neighbor algorithm. Over this wide range of problems and algorithms, we found that the new tests behave better in terms of size and have powers that are unmatched by any known techniques (with comparable size).

The simulation also allowed us to recommend values for the parameters involved in the proposed techniques, namely J around 15 and (for the conservative Z) M between 5 and 10. If one wants an inference procedure that is not liberal, the natural choice is the conservative Z. However, if one prefers an inference procedure with size close to the nominal size α and is ready to accept small departures in the liberal side as well as in the conservative side, then the corrected resampled t test appears to be a good choice. The advantage of the latter is that it requires little computing in comparison to the conservative Z.

The paper revolved around a specific cross-validation estimator; one in which we split the data sets of n examples into a training set (of size n_1) and a testing set (of size $n_2 = n - n_1$), and repeat this process J times in an independent manner. So, for instance, the testing sets of two different splits may partially overlap. This contrasts with the most standard cross-validation estimator for which the testing sets are mutually exclusive. Analyzing the variance of this standard estimator and providing valid estimates of that variance would be valuable future work.

References

Blake, C., Keogh, E., & Merz, C. (1998). UCI repository of machine learning databases..

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, *24 (6)*, 2350–2383.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth International Group.

Burges, C. (1998). A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 1–47.

Devroye, L., Gyrfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag.

Dietterich, T. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10 (7)*, 1895–1924.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Monographs on Statistics and Applied Probability 57. Chapman & Hall, New York, NY.

Everitt, B. (1977). *The analysis of contengency tables*. Chapman & Hall, London.

Goutte, C. (1997). Note on free lunches and cross-validation. *Neural Computation*, *9*(6), 1053–1059.

Hinton, G., Neal, R., Tibshirani, R., & DELVE team members (1995). Assessing learning procedures using DELVE. Tech. rep., University of Toronto, Department of Computer Science.

Kearns, M., & Ron, D. (1997). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Tenth Annual Conference on Computational Learning Theory*, pp. 152–162. Morgan Kaufmann.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceeding of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1137–1143. Morgan Kaufmann.

Kolen, J., & Pollack, J. (1991). Back propagation is sensitive to initial conditions. In *Advances in Neural Information Processing Systems*, pp. 860–867. Morgan Kauffmann, San Francisco, CA.

Nadeau, C., & Bengio, Y. (1999). Inference for the generalisation error. Tech. rep. 99s-25, CIRANO.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Wolpert, D., & Macready, W. (1995). No free lunch theorems for search. Tech. rep. SFI-TR-95-02-010, The Santa Fe Institute.

Zhu, H., & Rohwer, R. (1996). No free lunch for cross validation. *Neural Computation*, *8*(7), 1421–1426.

Appendix

A.0 **Some statistical prerequisites**

Suppose that we observe data D that is generated by some probability mechanism $P(D)$. Suppose that we are interested in some quantity $\mu = \mu(P)$ and that we have derived an estimator $\hat{\mu}$ (based on D) of the quantity of interest μ. Although the estimator $\hat{\mu}$ may have some nice properties such as being unbiased (i.e. $E[\hat{\mu}] = \mu$), the value $\hat{\mu}$ alone is not sufficient to say something interesting about μ. There are two questions that we might want to answer: (i) what are the plausible values of μ?, (ii) is a given value μ_0 plausible for μ? Question (i) is answered with a confidence interval and question (ii) is resolved with a hypothesis test.

Quite often $\hat{\mu}$ will be approximatively distributed as $N(\mu, Var[\hat{\mu}])$. If that is the case, for question (i) we consider the random interval $I = [\hat{\mu} - z_{1-\alpha/2}\sqrt{Var[\hat{\mu}]}, \hat{\mu} + z_{1-\alpha/2}\sqrt{Var[\hat{\mu}]}]$, with $z_{1-\alpha/2}$ being the *percentile* $1 - \alpha/2$ of the normal $N(0, 1)$ distribution, and called *confidence interval for* μ *at confidence level* $1 - \alpha$. This random interval has the following property: $P(\mu \in I) \approx 1 - \alpha$. For question (ii), if we want to assess if a hypothesis $H_0 : \mu = \mu_0$ is plausible, we use the following criterion: reject H_0 if $|\hat{\mu} - \mu_0| > z_{1-\alpha/2}\sqrt{Var[\hat{\mu}]}$. This is a test at the *significance level* α as $\text{Prob}(\text{reject}H_0|\text{H}_0 \text{ is true}) \approx \alpha$.

However $Var[\hat{\mu}]$ is seldom known, it has to be estimated. If $\widehat{Var}[\hat{\mu}]$ is a good estimator of $Var[\hat{\mu}]$, the above confidence interval and hypothesis test can be carried out with $Var[\hat{\mu}]$ replaced by $\widehat{Var}[\hat{\mu}]$. Namely, the test of the hypothesis $H_0 : \mu = \mu_0$ (of size [15] α) has the following form

$$\text{reject} \quad H_0 \quad \text{if} \quad |\hat{\mu} - \mu_0| > z_{1-\alpha/2}\sqrt{\widehat{Var}[\hat{\mu}]}, \tag{18}$$

while the confidence interval for μ (at confidence level $1 - \alpha$) will look like

$$\mu \in \left[\hat{\mu} - z_{1-\alpha/2}\sqrt{\widehat{Var}[\hat{\mu}]}, \hat{\mu} + z_{1-\alpha/2}\sqrt{\widehat{Var}[\hat{\mu}]}\right]. \tag{19}$$

When $\widehat{Var}[\hat{\mu}]$ is used in lieu of $Var[\hat{\mu}]$ (as done above), some people might prefer to replace $z_{1-\alpha/2}$ by the percentile $1 - \alpha/2$ of Student's t distribution [16]; that is mostly motivated by the second result of Lemma 2.

But what if $\widehat{Var}[\hat{\mu}]$ is not a good estimator of $Var[\hat{\mu}]$? We say that a confidence interval is *liberal* if it covers the quantity of interest with probability smaller than the required $1 - \alpha$ ($\text{Prob}(\mu \in [\hat{\mu} - z_{1-\alpha/2}\sqrt{\widehat{Var}[\hat{\mu}]}, \hat{\mu} + z_{1-\alpha/2}\sqrt{\widehat{Var}[\hat{\mu}]}]) < 1 - \alpha$); if the above probability is greater

[15] Size is a synonym of significance level.

[16] The Student's distribution with k degrees of freedom refers to the law of $\frac{Z}{\sqrt{Y/k}}$ where $Z \sim N(0, 1)$ is independent of $Y \sim \chi^2_k$. Often, the Y random variable will take the form $Y = \sum_{i=1}^{N} R_i^2$. This will have $(N - p)$ degrees of freedom where p is the number of linear constraints the R_i's are subject to. For instance, if $R_i = X_i - \bar{X}$, then $p = 1$ as $\sum_{i=1}^{N} R_i = 0$.

than $1-\alpha$, it is said to be *conservative.* A test is liberal if it rejects the null hypothesis with probability greater than the required size α whenever the null hypothesis is actually true ($P(|\hat{\mu}-\mu_0| > c\sqrt{\widehat{Var}[\hat{\mu}]} | \mu = \mu_0) > \alpha$); if the above probability is smaller than α, the test is said to be conservative. To determine if an inference procedure is liberal or conservative, we will ask ourself if $\widehat{Var}[\hat{\mu}]$ tends to underestimate or overestimate $Var[\hat{\mu}]$. Let us consider these two cases carefully.

- If we have $\frac{Var[\hat{\mu}]}{E[\widehat{Var}[\hat{\mu}]]} > 1$, this means that $\widehat{Var}[\hat{\mu}]$ tends to underestimate the actual variance of $\hat{\mu}$ so that the confidence interval of the form (19) will tend to be shorter than it needs to be to cover μ with probability $(1-\alpha)$. So the confidence interval would cover the value μ with probability smaller than the required $(1-\alpha)$, i.e. the interval would be liberal. In terms of hypothesis testing, the criterion shown in (18) will be met too often since $\widehat{Var}[\hat{\mu}]$ tends to be smaller than it should. In other words, the probability of rejecting H_0 when H_0 is actually true will exceed the prescribed α. We then say that the (actual) size of the test is greater than the nominal (or desired) size α.
- Naturally, the reverse happens if $\frac{Var[\hat{\mu}]}{E[\widehat{Var}[\hat{\mu}]]} < 1$. So in this case, the confidence interval will tend to be larger than needed and thus will cover μ with probability greater than the required $(1-\alpha)$, and the test of hypothesis based on the criterion (18) will tend to reject the null hypothesis with probability smaller than α (the nominal level of the test) whenever the null hypothesis is true.

We shall call $\frac{Var[\hat{\mu}]}{E[\widehat{Var}[\hat{\mu}]]}$ the *political ratio* since it indicates that inference should be *liberal* when it is greater than 1, *conservative* when it is less than 1. Of course, the political ratio is not the only thing determining whether an inference procedure is liberal or conservative. For instance, if $\frac{Var[\hat{\mu}]}{E[\widehat{Var}[\hat{\mu}]]} = 1$, the inference may still be liberal or conservative if the wrong number of degrees of freedom is used, or if the distribution of $\hat{\mu}$ is not approximately Gaussian.

In hypothesis testing, whether or not the size of the test is close to the nominal size α is not the only concern. We also want to know how the test behaves when the null hypothesis is false. Ideally, we would want the criterion (18) to often lead to rejection when the hypothesis is false and seldom lead to rejection of H_0 when H_0 is actually true. We define the *power* function (or curve) as $\pi(\mu) = \text{Prob}(\text{reject } H_0 : \mu = \mu_0 \mid \mu)$. This is the probability of rejection when we hold the hypothesised value μ_0 constant and we let the true value of the quantity of interest μ vary. We thus want the size of the test $\text{Prob}(\text{reject } H_0 : \mu = \mu_0 \mid H_0 \text{ is true}) = \pi(\mu_0)$ to be close to the nominal size α and we want $\pi(\mu)$ to be large when $\mu \neq \mu_0$.

A.1 **Proof of Lemma 2**

Let $W = (W_1, \ldots, W_K, W_{K+1}) \overset{iid}{\sim} N(0, \delta - \gamma)$ and $\epsilon \sim N(\beta, \gamma)$ be $(K+2)$ independent Gaussian variates and let $U_k = W_k + \epsilon$, $\forall k$. One may easily show that $(U_1, \ldots, U_{K+1})$ follow the multivariate Gaussian distribution with the mean and covariance structure considered in Lemma 1. Let $\bar{W} = K^{-1}\sum_{k=1}^{K} W_k$. Observe that

- $S_U^2 = \frac{1}{K-1}\sum_{k=1}^K (U_k - \bar{U})^2 = \frac{1}{K-1}\sum_{k=1}^K (W_k - \bar{W})^2 = S_W^2$,
- $\bar{W}$ and S_W^2 are independent with $\frac{(K-1)S_W^2}{\delta-\gamma} \sim \chi^2_{K-1}$ (well known result),
- S_W^2, $\bar{W}$, ϵ and W_{k+1} are four independent random variables.

It then follows that

$$\sqrt{1-\pi}\ \ \frac{U_{K+1}-\beta}{\sqrt{S_U^2}} = \sqrt{\frac{\delta-\gamma}{\delta}}\ \ \frac{U_{K+1}-\beta}{\sqrt{S_U^2}} = \frac{\frac{W_{K+1}+\epsilon-\beta}{\sqrt{\delta-\gamma+\gamma}}}{\sqrt{\frac{1}{K-1}\ \frac{(K-1)S_W^2}{\delta-\gamma}}} = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{K-1}}{K-1}}} \sim t_{K-1},$$

$$\sqrt{\frac{1-\pi}{1+(K-1)\pi}}\ \ \frac{\sqrt{K}(\bar{U}-\beta)}{\sqrt{S_U^2}} = \frac{\sqrt{\delta-\gamma}}{\sqrt{\frac{\delta-\gamma}{K}+\gamma}}\ \ \frac{\bar{U}-\beta}{\sqrt{S_U^2}} = \frac{\frac{\bar{W}+\epsilon-\beta}{\sqrt{\frac{\delta-\gamma}{K}+\gamma}}}{\sqrt{\frac{1}{K-1}\ \frac{(K-1)S_W^2}{\delta-\gamma}}} = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{K-1}}{K-1}}} \sim t_{K-1}.$$

Recall that the t_r law is the distribution of $\frac{X}{\sqrt{Y/r}}$ where $X \sim N(0,1)$ is independent of $Y \sim \chi^2_r$.

A.2 **Proof of Proposition 2**

In order to show that σ_2 is non-increasing in n_2, it is sufficient to show that, for fixed n_1 and arbitrary $n_2 > 1$, we have $\sigma_2(n_1, n_2) \le \sigma_2(n_1, n_2 - 1)$.

We show later (we keep the fun part for the end) that

$$ {}^{n_2}_{n_1}\hat{\mu}_\infty = \frac{1}{n}\sum_{k=1}^{n}\ {}^{n_2-1}_{n_1}\hat{\mu}_\infty^{(-k)}, \tag{20}$$

where ${}^{n_2}_{n_1}\hat{\mu}_\infty$ is as introduced after (8) and reproduced below, and ${}^{n_2-1}_{n_1}\hat{\mu}_\infty^{(-k)}$ is the result of calculating ${}^{n_2-1}_{n_1}\hat{\mu}_\infty$ on Z_1^n with Z_k removed (leaving a data set of $n-1 = n_1+n_2-1$ examples). Obviously, the ${}^{n_2-1}_{n_1}\hat{\mu}_\infty^{(-k)}$'s are identically distributed so that [17]

$$\sigma_2(n_1,n_2) = \ Var[\ {}^{n_2}_{n_1}\hat{\mu}_\infty] \le \ Var[{}^{n_2-1}_{n_1}\hat{\mu}_\infty^{(-1)}] = \sigma_2(n_1, n_2-1).$$

To complete the proof, we only need to show that identity (20) is true.

Let $C(S, n_1)$ denote the set of all possible subsets of n_1 distinct elements from S, where S is itself a set of distinct positive integers (of course n_1 must not be greater than $|S|$, the cardinality of S). For instance, the cardinality of $C(S,n_1)$ is $|C(S,n_1)| = \binom{|S|}{n_1}$, i.e. the number of ways to choose n_1 distinct elements from S.

We have

$$ {}^{n_2}_{n_1}\hat{\mu}_\infty \quad = \quad \frac{1}{\binom{n}{n_1}} \sum_{s\in C(\{1,\ldots,n\},n_1)} \frac{1}{n_2} \sum_{i\in\{1,\ldots,n\}\setminus s} \mathcal{L}(Z_s; Z_i)$$

[17] Let $U_1, \ldots, U_n$ be variates with equal variance and let $\bar{U} = \frac{1}{n}\sum_{i=1}^n U_i$, then we have $Var[\bar{U}] \le Var[U_1]$.

$$
\begin{aligned}
&= \frac{1}{\binom{n}{n_1}} \sum_{s\in C(\{1,\ldots,n\},n_1)} \frac{1}{n_2(n_2-1)} \sum_{k\in\{1,\ldots,n\}\setminus s} \sum_{i\in(\{1,\ldots,n\}\setminus s)\setminus\{k\}} \mathcal{L}(Z_s;Z_i) \\
&= \frac{n_1!n_2!}{n!n_2} \sum_{k=1}^{n} \sum_{s\in C(\{1,\ldots,n\}\setminus\{k\},n_1)} \frac{1}{n_2-1} \sum_{i\in(\{1,\ldots,n\}\setminus\{k\})\setminus s} \mathcal{L}(Z_s;Z_i) \\
&= \frac{1}{n} \sum_{k=1}^{n} \frac{1}{\binom{n-1}{n_1}} \sum_{s\in C(\{1,\ldots,n\}\setminus\{k\},n_1)} \frac{1}{n_2-1} \sum_{i\in(\{1,\ldots,n\}\setminus\{k\})\setminus s} \mathcal{L}(Z_s;Z_i) \\
&= \frac{1}{n} \sum_{k=1}^{n} {}_{n_1}^{n_2-1}\hat{\mu}_{\infty}^{(-k)}.
\end{aligned}
$$

Note that to get from the first to second line in the above development, we used the following identity for the arithmetic mean (of $x_1,\ldots x_I$ say) : $\frac{1}{I}\sum_{i=1}^{I} x_i = \frac{1}{I(I-1)}\sum_{k=1}^{I}\sum_{i\neq k} x_i$.

A.3 **Inference when vectors have moments as in Lemma 1**

Suppose that we have n independent and identically distributed random vectors $T_1,\ldots,T_i,\ldots,T_n$ where $T_i = (T_{i,1},\ldots,T_{i,K})'$. Suppose that $T_{i,1},\ldots,T_{i,K}$ has the moment structure displayed in Lemma 1. Let $\bar{T}_i = \frac{1}{K}\sum_{k=1}^{K} T_{i,k}$. Let $\theta = (\beta,\delta,\pi)$ be the vector of parameters involved in Lemma 1. Consider the following unbiased estimating function

$$
g(\theta) = \sum_{i=1}^{n} g_i(\theta) = \sum_{i=1}^{n} \begin{pmatrix} \bar{T}_i - \beta \\ \sum_{k=1}^{K}[(T_{i,k}-\beta)^2 - \delta] \\ (\bar{T}_i-\beta)^2 - \delta\left(\pi + \frac{1-\pi}{K}\right) \end{pmatrix}.
$$

Let $B(\theta) = \sum_{i=1}^{n} g_i(\theta)g_i(\theta)'$ and

$$
A(\theta) = -E\left[\frac{\partial g(\theta)}{\partial\theta'}\right] = n \begin{bmatrix} 1 & 0 & 0 \\ 0 & \mathrm{K} & 0 \\ 0 & (\pi + \frac{1-\pi}{K}) & \frac{K-1}{K}\delta \end{bmatrix}.
$$

Let $\hat{\theta}$ be such that $g(\hat{\theta}) = \mathbf{0}_3$, then, according to (White, 1982),

$$
\left[\hat{\theta}_j \pm Z_{1-\alpha/2}\sqrt{\hat{V}[\hat{\theta}_j]}\right],
$$

with $\hat{V}[\hat{\theta}_j] = \left[A(\hat{\theta})^{-1}B(\hat{\theta})(A(\hat{\theta})^{-1})'\right]_{j,j}$, is a confidence interval at approximate confidence level $(1-\alpha)$. For instance, in the case of β, this yields

$$
\beta = \theta_1 \in \left[\bar{T}_{\cdot} \pm Z_{1-\alpha/2}\sqrt{\frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}(\bar{T}_i - \bar{T}_{\cdot})^2}\right],
$$

where $\bar{T}_{\cdot} = \frac{1}{n}\sum_{i=1}^{n}\bar{T}_i$ is the mean of all the $\bar{T}_i$'s.