99s-25

Inference for the Generalization Error

Claude Nadeau, Yoshua Bengio

Série Scientifique
Scientific Series

CIRANO
Centre interuniversitaire de recherche en analyse des organisations

Montréal
Juillet 1999

ISSN 1198-8177

Inference for the Generalization Error*

Claude Nadeau†, Yoshua Bengio‡

Résumé / Abstract

Nous considérons l'estimation par validation croisée de l'erreur de généralisation. Nous effectuons une étude théorique de la variance de ect estimateur en tenant compte de la variabilité due au choix des ensembles d'entraînement et des exemples de test. Cela nous permet de proposer deux nouveaux estimateurs de cette variance. Nous montrons, via des simulations, que ces nouvelles statistiques performent bien par rapport aux statistiques considérées dans Dietterich (1998). En particulier, ces nouvelles statistiques se démarquent des autres présentement utilisées par le fait qu'elles mènent à des tests d'hypothèses qui sont puissants sans avoir tendance à être trop libéraux.

We perform a theoretical investigation of the variance of the cross-validation estimate of the generalization error that takes into account the variability due to the choice of training sets and test examples. This allows us to propose two new estimators of this variance. We show, via simulations, that these new statistics perform well relative to the statistics considered in Dietterich (1998). In particular, tests of hypothesis based on these don't tend to be too liberal like other tests currently available, and have good power.

Mots Clés : Erreur de généralisation, validation croisée, estimation de la variance, test d'hypothèses, niveau, puissance

Keywords: Generalization error, cross-validation, variance estimation, hypothesis tests, size, power

---

* Corresponding Author: Yoshua Bengio, CIRANO, 2020 University Street 25th floor, Montréal, Qc, Canada H3A 2A5 Tel: (514) 985-4025 Fax: (514) 985-4039 email: bengioy@cirano.umontreal.ca
Both authors benefited from financial support from NSERC. The second author would also like to thank the Canadian Networks of Centers of Excellence (MITACS and IRIS).

† CIRANO

‡ Université de Montréal and CIRANO

1 Generalization Error and its Estimation

When applying a learning algorithm (or comparing several algorithms), one is typically interested in estimating its generalization error. Its point estimation is rather trivial through cross-validation. Providing a variance estimate of that estimation, so that hypothesis testing and/or confidence intervals are possible, is more difficult, especially, as pointed out in (Hinton et al., 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 and Pollack, 1991). A notable effort in that direction is Dietterich's work (Dietterich, 1998). Building upon this work, 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 estimates.

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 $\mathcal{Z}^{n_1} \times \mathcal{Z} \to \mathbb{R}$. 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. According to Dietterich's taxonomy (Dietterich, 1998), we deal with problems of type 5 through 8, rather then type 1 through 4. 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

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

  where F represents a learning algorithm that yields $F(D)$ ($F(D) : \mathbb{R}^p \to \mathbb{R}^q$), when training the algorithm on D, and Q is a loss function measuring the inaccuracy of a decision. 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 ${}_n\mu$ is what most people call the generalization error.

- **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 represents the size of the training set and we shall let $n_2 = n - n_1$ be the size of the corresponding test set.
- Let $S_1, \ldots S_J$ be such random index sets, 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 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 ${}^{n_2}_{n_1}\hat{\mu}_J$. We first study theoretically this variance 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.

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, common variance δ and $Cov[U_k, U_{k'}] = \gamma$, $\forall k \neq k'$. Let $\pi = \frac{\gamma}{\delta}$ be the correlation between U_k and $U_{k'}$ ($k \neq k'$). Let $\bar{U} = k^{-1} \sum_{k=1}^{K} U_i$ 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 any K (with γ and δ not depending on K), then $\gamma \geq 0$.*
3. *$E[S_U^2] = \delta - \gamma$.*

Proof

1. *This results 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 $\gamma \geq 0$. Note 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 $\bar{U} = K^{-1} \sum_{i=k}^{K} U_k$ 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. *$\sqrt{1-\pi} \quad \frac{U_{K+1}-\beta}{\sqrt{S_U^2}} \sim t_{K-1}$,*
2. *$\sqrt{\frac{1-\pi}{1+(K-1)\pi}} \quad \frac{\sqrt{K}(\bar{U}-\beta)}{\sqrt{S_U^2}} \sim t_{K-1}$,*

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.

- Let $\sigma_0 = \sigma_0(n_1) = Var[L(j,i)]$ when i is randomly drawn from S_j^c.
- Let [1] $\sigma_2 = \sigma_2(n_1, n_2) = Cov[L(j,i), L(j',i')]$, with $j \neq j'$, i and i' randomly and independently drawn from S_j^c and $S_{j'}^c$ respectively.

[1] Yes, we know how to count! We just save σ_1 for another quantity to be introduced in (6).

- 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.

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

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

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 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 $\lim_{J\to\infty} Var[{}^{n_2}_{n_1}\hat{\mu}_J] = \sigma_2$ 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 ${}^{n_2}_{n_1}\hat{\mu}_J$ is how n_1, n_2 and J affect its variance.

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

- *$Var[{}^{n_2}_{n_1}\hat{\mu}_J]$ 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 $Var[{}^{n_2}_{n_1}\hat{\mu}_J]$ 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).*

Regarding σ_2, we show in Appendix A.2 that $\sigma_2(n_1, n_2) \leq \sigma_2(n_1, n_2')$ if $n_2' < n_2$. 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 that in most situations it should decrease in n_1**. Our arguments go as follows [2].

- 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 say that the decision function yielded by a learning algorithm is less variable when the training set is larger. We conclude that 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.
- Note that when L is a test error or a difference of test errors, and when the learning algorithms have a finite capacity, it can be shown that ${}_{n_1}\mu$ is bounded with a given high probability by a decreasing function of n_1 (Vapnik, 1982), converging to the asymptotic training error (which is both the training error and the expected generalization error when $n_1 \to \infty$). This argument is based on bounds on the cumulative distribution of the difference between the training error and the expected generalization error. When n_1 increases, the mass of the distribution of ${}_{n_1}\mu$ gets concentrated closer to the training error (and asymptotically it becomes a Dirac at the training error). We conjecture that the same argument can be used to show that the variance of ${}_{n_1}\mu$ is a decreasing function of n_1.

Regarding the estimation of $Var[{}_{n_1}^{n_2}\hat{\mu}_J]$, we note that we can easily estimate the following quantities.

- From Lemma 1, we obtain readily that the sample variance of the $\hat{\mu}_j$'s (call it $S^2_{\hat{\mu}_j}$) 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
$$\begin{aligned} E[S^2_{\hat{\mu}_j}] &= E[E[S^2_{\hat{\mu}_j} | Z_1^n]] = E[Var[\hat{\mu}_j | Z_1^n]] \\ &= Var[\hat{\mu}_j] - Var[E[\hat{\mu}_j | Z_1^n]] = \sigma_1 - Var[{}_{n_1}^{n_2}\hat{\mu}_\infty] = \sigma_1 - \sigma_2. \end{aligned}$$
Note that $E[\hat{\mu}_j | Z_1^n] = {}_{n_1}^{n_2}\hat{\mu}_\infty$ and $Var[{}_{n_1}^{n_2}\hat{\mu}_\infty] = \sigma_2$ both come from Appendix A.2.
- 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 $\sigma_0 - \sigma_3$.

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

We are thus able to estimate unbiasedly any linear combination of $\sigma_0 - \sigma_3$ and $\sigma_3 - \sigma_2$. This turns out to be all we can hope to estimate unbiasedly as we show in Proposition 3. This is not sufficient to estimate (8) unbiasedly as we know no identity involving σ_0, σ_2 and σ_3.

Proposition 3 *There is no general non-negative unbiased estimator of $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ based on the $L(j,i)$'s involved in ${}^{n_2}_{n_1}\hat{\mu}_J$.*

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 stacking the $\vec{L}_j$'s; $\vec{L}$ is thus a vector of length n_2J. We know that $\vec{L}$ has expectation ${}_{n_1}\mu \mathbf{1}_{n_2J}$ and variance*

$$Var[\vec{L}] = \sigma_2 \mathbf{1}_{n_2J}\mathbf{1}'_{n_2J} + (\sigma_3 - \sigma_2) I_J \otimes (\mathbf{1}_{n_2}\mathbf{1}'_{n_2}) + (\sigma_0 - \sigma_3) I_{n_2J},$$

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 generally don't know anything about the higher moments of $\vec{L}$ or expectations of other non-linear functions of $\vec{L}$; these will involve ${}_{n_1}\mu$ and the σ's (and possibly other things) in an unknown manner. This forces us to only consider estimators of $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ of the following form

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

We have

$$E[\hat{V}[\,{}^{n_2}_{n_1}\hat{\mu}_J]] = trace(A\,Var[\vec{L}]) + {}_{n_1}\mu^2 \mathbf{1}'_{n_2J} A \mathbf{1}_{n_2J} + {}_{n_1}\mu b' \mathbf{1}_{n_2J}.$$

Since we wish $\hat{V}[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ to be unbiased for $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$, 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. We take $b = \mathbf{0}_{n_2J}$ as any other choice of b such that $0 = b'\mathbf{1}_{n_2J}$ simply adds noise of expectation 0 to the estimator. Then, in order to have a non-negative estimator of $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$, we have to take A to be non-negative definite. Then $\mathbf{1}'_{n_2J}A\mathbf{1}_{n_2J} = 0 \Rightarrow A\mathbf{1}_{n_2J} = \mathbf{0}_{n_2J}$. So

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

This means that only linear combinations of $(\sigma_3 - \sigma_2)$ and $(\sigma_0 - \sigma_3)$ can be estimated.

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

We are interested in estimating ${}^{n_2}_{n_1}\sigma^2_J = Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ where ${}^{n_2}_{n_1}\hat{\mu}_J$ is as defined in (5). We provide two new estimators of $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$ 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 $Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]$. The second is meant to be conservative, that is, if our conjecture following Proposition 2 is correct, its expected value exceeds the actual variance.

3.1 First Method: Approximating ρ

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

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

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

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

so that $\left(\frac{1}{J}+\frac{\rho}{1-\rho}\right)S^2_{\hat{\mu}_j}$ is an unbiased estimator of $Var[\,\frac{n_2}{n_1}\hat{\mu}_J]$. 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, ρ is a function of σ_0, σ_2 et σ_3 that can not be written as a function of $(\sigma_0-\sigma_3)$ 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 do as if $\mathcal{L}(Z_{S_j};Z_i)$ depended only on Z_i and n_1. 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 $Var[\,\frac{n_2}{n_1}\hat{\mu}_J]$ 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 $Var[\,\frac{n_2}{n_1}\hat{\mu}_J]$ 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 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 et al., 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 the next section, 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 typically regarded as more undesirable than conservative tests.

We have shown in the previous section that ${}^{n_2}_{n_1}\sigma^2_J$ could not be estimated unbiasedly. However we may estimate unbiasedly ${}^{n_2}_{n_1'}\sigma^2_J = Var[\, {}^{n_2}_{n_1'}\hat{\mu}_J]$ where $n_1' = \lfloor \frac{n}{2} \rfloor - n_2 < n_1$. Let ${}^{n_2}_{n_1'}\hat{\sigma}^2_J$ be the unbiased estimator, developed below, of the above variance. We argued in the previous section that, because $n_1' < n_1$, $Var[\, {}^{n_2}_{n_1'}\hat{\mu}_J] \geq Var[\, {}^{n_2}_{n_1}\hat{\mu}_J]$, so that ${}^{n_2}_{n_1'}\hat{\sigma}^2_J$ will tend to overestimate ${}^{n_2}_{n_1}\sigma^2_J$, that is $E[\, {}^{n_2}_{n_1'}\hat{\sigma}^2_J] = {}^{n_2}_{n_1'}\sigma^2_J \geq {}^{n_2}_{n_1}\sigma^2_J$.

Here's how we may estimate ${}^{n_2}_{n_1'}\sigma^2_J$ 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 ${}^{n_2}_{n_1'}\sigma^2_J$ 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 ${}^{n_2}_{n_1'}\sigma^2_J$. Averaging these yields a more accurate estimate of ${}^{n_2}_{n_1'}\sigma^2_J$.

Obtaining a pair of independent ${}^{n_2}_{n_1'}\hat{\mu}_J$ 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[3]. 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. Let $\hat{\mu}^c_{(1)}$ be the statistic computed on D_1^c. Then $\hat{\mu}_{(1)}$ and $\hat{\mu}^c_{(1)}$ are independent since D_1 and D_1^c are independent data sets [4], so that $(\hat{\mu}_{(1)} - \frac{\hat{\mu}_{(1)} + \hat{\mu}^c_{(1)}}{2})^2 + (\hat{\mu}^c_{(1)} - \frac{\hat{\mu}_{(1)} + \hat{\mu}^c_{(1)}}{2})^2 = \frac{1}{2}(\hat{\mu}_{(1)} - \hat{\mu}^c_{(1)})^2$ is unbiased for ${}^{n_2}_{n_1'}\sigma^2_J$. This splitting process may be repeated M times. This yields D_m and D^c_m, with $D_m \cup D^c_m = Z_1^n$, $D_m \cap D^c_m = \emptyset$ and $|D_m| = |D^c_m| = \lfloor \frac{n}{2} \rfloor$ for $m = 1, \ldots, M$. Each split yields a pair $(\hat{\mu}_{(m)}, \hat{\mu}^c_{(m)})$ that is such that

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

[3] 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^c_m are still disjoint subsets of size $\lfloor \frac{n}{2} \rfloor$ from Z_1^n, but $Z_1^n \setminus (D_m \cup D^c_m)$ is a singleton instead of being the empty set.

[4] 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.

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 $\frac{1}{M}$.

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 bootstrap and corrected bootstrap. Tests of the hypothesis $H_0 :\ {}_{n_1}\mu = \mu_0$ (at significance level α) have the following form

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

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

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

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. 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.

Before we go through all these statistics, we need to introduce the concept of liberal and conservative inference. We say that a confidence interval is liberal if it covers the quantity of interest with probability smaller than the required $1-\alpha$; if the above probability is greater 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 α whenever the null hypothesis is actually true; 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 $\hat{\sigma}^2$ tends to underestimate or overestimate $Var[\hat{\mu}]$. Let us consider these two cases carefully.

- If we have $\frac{Var[\hat{\mu}]}{E[\hat{\sigma}^2]} > 1$, this means that $\hat{\sigma}^2$ tends to underestimate the actual variance of $\hat{\mu}$ so that a confidence interval of the form (13) will tend to be shorter then it needs to be to cover ${}_{n_1}\mu$ with probability $(1-\alpha)$. So the confidence interval would cover the value ${}_{n_1}\mu$ with probability smaller than the required $(1-\alpha)$. Such an interval is called liberal in Statistics. In terms of hypothesis testing, the criterion shown in (12) will be met too often since $\hat{\sigma}^2$ 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.

- Naturally, the reverse happens if $\frac{Var[\hat{\mu}]}{E[\hat{\sigma}^2]} < 1$. So in this case, the confidence interval will tend to be larger then needed and thus will cover ${}_{n_1}\mu$ with probability greater than the required $(1-\alpha)$, and tests of hypothesis based on the criterion (12) 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 $\frac{Var[\hat{\mu}]}{E[\hat{\sigma}^2]}$ 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 on conservative. For instance, if $\frac{Var[\hat{\mu}]}{E[\hat{\sigma}^2]} = 1$, 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.

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 fact 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 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.

[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

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

which squared leads to the McNemar's test (not corrected for continuity).

Note that S_L^2 is a biased estimator of σ_0 (the unconditional variance of $L(1,i) = L(Z_{S_1}; Z_i),\ i \notin S_1$), 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

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

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 $E[\ {}_{n_1}^{n_2}\hat{\mu}_1 | Z_{S_1}] = E[L(1,i)|Z_{S_1}] = E[L(Z_{S_1}; Z_i)|Z_{S_1}],\ i \notin S_1$, 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), it deals with questions 1 through 4, rather than questions 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. Inference would be based on the fact that

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

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, $S_{\hat{\mu}_j}^2$ actually estimates (without bias) the variance of ${}_{n_1}^{n_2}\hat{\mu}_J$ conditional on Z_1^n. Thus while (15) is wrong, we do have

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

Recall from the proof of Proposition 2 in Appendix A.2 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 ${}_{n_1}\mu$.

[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.

3. **5x2 cv t test**
Dietterich (Dietterich, 1998) used[8] $\hat{\mu} = \tilde{\mu}_{(1)}$, $\hat{\sigma}^2 = \hat{\sigma}^2_{Diet} = \frac{1}{10}\sum_{m=1}^{5}(\tilde{\mu}_{(m)} - \tilde{\mu}^c_{(m)})^2$ and $c = t_{5,1-\alpha/2}$, where the $\tilde{\mu}_{(m)}$'s and $\tilde{\mu}^c_{(m)}$'s are $\frac{\lfloor n/2 \rfloor}{\lfloor n/2 \rfloor}\hat{\mu}_1$'s somewhat similar to the $\hat{\mu}_{(m)}$'s and $\hat{\mu}^c_{(m)}$'s used in (11). Specifically, Z_1^n is split in half $M = 5$ times to yield $D_1, D_1^c, \ldots, D_5, D_5^c$ as in Section 3. Then let
$$\hat{\mu}_{(m)} = \lfloor n/2 \rfloor^{-1} \sum_{i \in D_m^c} \mathcal{L}(D_m; Z_i), \qquad \hat{\mu}^c_{(m)} = \lfloor n/2 \rfloor^{-1} \sum_{i \in D_m} \mathcal{L}(D_m^c; Z_i).$$
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 allow 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 $\tilde{\mu}^2_{(m)^c}$'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.

   In all cases, more degrees of freedom could be exploited; statistics distributed as t_8 can be devised by appropriate use of the 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 may then use
$$Z = \frac{{}^{n_2}_{n_1}\hat{\mu}_J - {}_{n_1}\mu}{\sqrt{{}^{n_2}_{n_1'}\hat{\sigma}^2_J}} \tag{17}$$
as a $N(0,1)$ variate 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

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

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 is

$$\frac{Var[\ {}^{n_2}_{n_1}\hat{\mu}_J]}{E[\ {}^{n_2}_{n_1'}\hat{\sigma}^2_J]} = \frac{{}^{n_2}_{n_1}\sigma^2_J}{{}^{n_2}_{n_1'}\sigma^2_J} < 1,$$

according to the argument following Proposition 2.

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. **Bootstrap**

To estimate the variance of $\hat{\mu} = {}^{n_2}_{n_1}\hat{\mu}_J$ by the bootstrap (Efron and Tibshirani, 1993), we must obtain R other instances of that random variable, by redoing the computation with different splits; call these $\breve{\mu}_1, \ldots, \breve{\mu}_R$. Thus, in total, $(R+1)J$ training and testing sets are needed here. Then one could consider $\hat{\sigma}^2 = \breve{\sigma}^2$, where $\breve{\sigma}^2$ is the sample variance of $\breve{\mu}_1, \ldots, \breve{\mu}_R$, and take $c = t_{R-1,1-\alpha/2}$, as $\breve{\sigma}^2$ has $R-1$ degrees of freedom. Of course ${}^{n_2}_{n_1}\hat{\mu}_J$, $\breve{\mu}_1, \ldots, \breve{\mu}_R$ 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

$$\frac{Var[\ {}^{n_2}_{n_1}\hat{\mu}_J]}{E[\breve{\sigma}^2]} = \frac{{}^{n_2}_{n_1}\sigma^2_J}{{}^{n_2}_{n_1}\sigma^2_J - \sigma_2} = \frac{J\sigma_2 + (\sigma_1 - \sigma_2)}{\sigma_1 - \sigma_2} > 1.$$

Note that this political ratio is the same as its counterpart for the resampled t-test because $E[\breve{\sigma}^2] = E[\frac{S^2_{\hat{\mu}_j}}{J}]$. So the bootstrap leads to liberal inference that should worsen with increasing J just like the resampled t test statistic. In other words, the 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 ${}^{n_2}_{n_1}\sigma^2_J$ 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 do 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 $\frac{{}^{n_2}_{n_1}\hat{\mu}_J - {}_{n_1}\mu}{\sqrt{\hat{\sigma}^2}} \sim t_{J-1}$.

Finally, let us note that the political ratio

$$\frac{Var[\ {}^{n_2}_{n_1}\hat{\mu}_J]}{E[\hat{\sigma}^2]} = \frac{\frac{1}{J} + \frac{\rho}{1-\rho}}{\frac{1}{J} + \frac{n_2}{n_1}}$$

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 bootstrap statistic**
Naturally, the correction we made in the resampled t test can be applied to the 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 ${}^{n_2}_{n_1}\sigma^2_J$. 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 ${}^{n_2}_{n_1}\hat{\mu}_J$, $\check{\mu}_1, \ldots \check{\mu}_R$ follows the multivariate Gaussian distribution, Lemma 2 states that $\frac{{}^{n_2}_{n_1}\hat{\mu}_J - {}_{n_1}\mu}{\sqrt{\hat{\sigma}^2}} \sim t_{R-1}$. Finally note that, just like in the corrected resampled t-test, the political ratio is

$$\frac{Var[\,{}^{n_2}_{n_1}\hat{\mu}_J]}{E[\hat{\sigma}^2]} = \frac{\frac{1}{J} + \frac{\rho}{1-\rho}}{\frac{1}{J} + \frac{n_2}{n_1}}.$$

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)	${}^{n_2}_{n_1}\hat{\mu}_1$	$\frac{1}{n_2}S^2_L$	$Z_{1-\alpha/2}$	$\frac{n_2\sigma_3+(\sigma_0-\sigma_3)}{\sigma_0-\sigma_3} > 1$
2. resampled t	${}^{n_2}_{n_1}\hat{\mu}_J$	$\frac{1}{J}S^2_{\hat{\mu}_j}$	$t_{J-1,1-\alpha/2}$	$1 + J\frac{\rho}{1-\rho} > 1$
3. Dietterich's 5×2 cv	${}^{n/2}_{n/2}\hat{\mu}_1$	$\hat{\sigma}^2_{Diet}$	$t_{5,1-\alpha/2}$	$\frac{\sigma_1}{\sigma_1-\sigma_4}$
4. conservative Z	${}^{n_2}_{n_1}\hat{\mu}_J$	${}^{n_2}_{n'_1}\hat{\sigma}^2_J$	$Z_{1-\alpha/2}$	$\frac{{}^{n_2}_{n_1}\sigma^2_J}{{}^{n_2}_{n'_1}\sigma^2_J} < 1$
5. bootstrap	${}^{n_2}_{n_1}\hat{\mu}_J$	$\check{\sigma}^2$	$t_{R-1,1-\alpha/2}$	$1 + J\frac{\rho}{1-\rho} > 1$
6. corrected resampled t	${}^{n_2}_{n_1}\hat{\mu}_J$	$\left(\frac{1}{J} + \frac{n_2}{n_1}\right)S^2_{\hat{\mu}_j}$	$t_{J-1,1-\alpha/2}$	$\frac{1+J\frac{\rho}{1-\rho}}{1+J\frac{n_2}{n_1}}$
7. corrected bootstrap	${}^{n_2}_{n_1}\hat{\mu}_J$	$\left(1 + \frac{Jn_2}{n_1}\right)\check{\sigma}^2$	$t_{R-1,1-\alpha/2}$	$\frac{1+J\frac{\rho}{1-\rho}}{1+J\frac{n_2}{n_1}}$

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 ${}^{n_2}_{n_1}\hat{\mu}_J$. 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 $^{n_2}_{n_1}\hat{\mu}_J$). 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, in the numerator of (17), we compute both $^{n/2}_{n/2}\hat{\mu}_J$ 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 explained above. 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 $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2_{Y|X})$ 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 $F_B(Z_S)(X) = \hat{\alpha}_S + \hat{\beta}_S X$ where $\hat{\alpha}_S$ and $\hat{\beta}_S$ are the intercept and the slope of the ordinary least squares regression of Y on X performed on the training set S. Since we use a quadratic loss, we therefore have $L_B(j,i) = (F_B(Z_{S_j})(X_i) - Y_i)^2 = (\hat{\alpha}_{S_j} + \hat{\beta}_{S_j} X_i - Y_i)^2$.

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 $L_{A-B}(j,i) = L_A(j,i) - L_B(j,i)$.

What's interesting in the present case is that we can derive analytically the actual generalization errors ${}_{n_1}\mu_A$ and ${}_{n_1}\mu_B$. Indeed we show in Appendix A.4 that

$$ {}_{n_1}\mu_A = \frac{n_1+1}{n_1}(\sigma^2_{Y|X} + \beta^2\sigma^2_X) \tag{18} $$

and

$$ {}_{n_1}\mu_B = \frac{n_1+1}{n_1}\frac{n_1-2}{n_1-3}\sigma^2_{Y|X}. \tag{19} $$

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 $\mu_x = 10$, $\sigma^2_X = 1$, $\alpha = 100$, $\beta = 1$ and $\sigma^2_{Y|X} = 97$ (so that ${}_{100}\mu_A = {}_{100}\mu_B = \frac{101}{100}98$, and therefore ${}_{100}\mu_{A-B} = 0$). 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. The table also provides some summary confidence intervals [9] 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 perform 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.

[9] Of course confidence intervals for the generalization errors are not interesting here because we know analytically what they are. For other kind of problems, this will not be the case.

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

	Simulation 1	Simulation 2	Simulation 3	Simulation 4
n	200	200	2000	2000
μ_X	10	10	10	10
α	100	100	100	100
β	1	2	0.1	0.1
σ^2_X	1	2	1	5
$\sigma^2_{Y\|X}$	97	64	9.97	9
$_{n/2}\mu_A$	98.98	72.72	9.99	9.06
$_{n/2}\mu_B$	98.98	65.31	9.99	9.02
$_{n/2}\mu_{A-B}$	0	7.41	0	0.04
$_{n/2}\mu_A$	[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]
$_{9n/10}\mu_A$	98.54	72.40	9.99	9.06
$_{9n/10}\mu_B$	98.09	64.72	9.98	9.01
$_{9n/10}\mu_{A-B}$	0.45	7.68	0.004	0.045
$_{9n/10}\mu_A$	[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, α, β, σ^2_X and $\sigma^2_{Y|X}$ as shown in the table. Actual values of the generalization errors $_{n_1}\mu$ are given according to formulas 18 and 19. 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}.

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 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 training 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 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 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 and Merz, 1998). The learning algorithms are

(a) **Classification tree**
We perform 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 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 S_j; otherwise it is 0.

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

[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	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 problem. 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}.

	Simulation 1	Simulation 2	Simulation 3	Simulation 4	Simulation 5	Simulation 6
n	300	300	300	300	300	300
w	1	5	10	17.25	25	2048
${}_{n/2}\mu_A$	[0.6911,0.6942]	[0.6911,0.6942]	[0.6911,0.6942]	[0.6911,0.6942]	[0.6911,0.6942]	[0.6911,0.6942]
${}_{n/2}\mu_B$	[0.5395,0.5427]	[0.5932,0.5965]	[0.6320,0.6353]	[0.6665,0.6697]	[0.6903,0.6936]	[0.7796,0.7824]
${}_{n/2}\mu_{A-B}$	[0.1501,0.1529]	[0.0963,0.0992]	[0.0574,0.0604]	[0.0230,0.0260]	[-0.0009,0.0022]	[-0.0899,-0.0868]
${}_{9n/10}\mu_A$	[0.5853,0.5897]	[0.5853,0.5897]	[0.5853,0.5897]	[0.5853,0.5897]	[0.5853,0.5897]	[0.5853,0.5897]
${}_{9n/10}\mu_B$	[0.4343,0.4388]	[0.4967,0.5012]	[0.5437,0.5483]	[0.5862,0.5908]	[0.6159,0.6205]	[0.7303,0.7344]
${}_{9n/10}\mu_{A-B}$	[0.1486,0.1533]	[0.0861,0.0910]	[0.0391,0.0440]	[-0.0034,0.0014]	[-0.0332,-0.0282]	[-0.1473,-0.1423]
$\rho_A(\frac{n}{2},\frac{n}{2})$	[0.2231,0.2592]	[0.2231,0.2592]	[0.2231,0.2592]	[0.2231,0.2592]	[0.2231,0.2592]	[0.2231,0.2592]
$\rho_B(\frac{n}{2},\frac{n}{2})$	[0.3103,0.3549]	[0.3336,0.3761]	[0.3495,0.3926]	[0.3597,0.4044]	[0.3679,0.4135]	[0.3474,0.3916]
$\rho_{A-B}(\frac{n}{2},\frac{n}{2})$	[0.1336,0.1605]	[0.1517,0.1801]	[0.1604,0.1906]	[0.1671,0.1982]	[0.1700,0.2020]	[0.1779,0.2112]
$\rho_A(\frac{n}{2},\frac{n}{10})$	[0.1369,0.1645]	[0.1369,0.1645]	[0.1369,0.1645]	[0.1369,0.1645]	[0.1369,0.1645]	[0.1369,0.1645]
$\rho_B(\frac{n}{2},\frac{n}{10})$	[0.1672,0.1984]	[0.1822,0.2145]	[0.1966,0.2318]	[0.2005,0.2382]	[0.2006,0.2385]	[0.2014,0.2371]
$\rho_{A-B}(\frac{n}{2},\frac{n}{10})$	[0.1222,0.1481]	[0.1291,0.1555]	[0.1300,0.1562]	[0.1305,0.1564]	[0.1295,0.1554]	[0.1334,0.1617]
$\rho_A(\frac{9n}{10},\frac{n}{10})$	[0.0988,0.1230]	[0.0988,0.1230]	[0.0988,0.1230]	[0.0988,0.1230]	[0.0988,0.1230]	[0.0988,0.1230]
$\rho_B(\frac{9n}{10},\frac{n}{10})$	[0.1046,0.1290]	[0.1064,0.1310]	[0.1153,0.1401]	[0.1184,0.1427]	[0.1250,0.1509]	[0.1196,0.1448]
$\rho_{A-B}(\frac{9n}{10},\frac{n}{10})$	[0.0849,0.1055]	[0.0849,0.1055]	[0.0836,0.1038]	[0.0848,0.1056]	[0.0875,0.1085]	[0.0936,0.1160]
r_A	[0.0022,0.0128]	[0.0022,0.0128]	[0.0022,0.0128]	[0.0022,0.0128]	[0.0022,0.0128]	[0.0022,0.0128]
r_B	[-0.0065,0.0012]	[-0.0044,0.0044]	[-0.0037,0.0055]	[-0.0044,0.0042]	[-0.0047,0.0041]	[0.0022,0.0124]
r_{A-B}	[-0.0047,0.0041]	[-0.0043,0.0043]	[-0.0035,0.0053]	[-0.0024,0.0066]	[-0.0007,0.0092]	[-0.0005,0.0095]

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}.

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 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 interval 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.

Table 2 confirms our calculations about ${}_{n_1}\mu$ for the simple linear regression problem as the confidence intervals for ${}_{n_1}\mu$ cover the actual values found according to formulas 18 and 19. 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 bootstrap and on the corrected resampled t-test will not be liberal. 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 $\frac{1}{M}$.

[13] As mentioned before, the corrected 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 $\rho(\frac{9n}{10}, \frac{n}{10})$.

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 value 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. 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 do this because constructing "power curves" would be too computationally expensive. Nevertheless, Figure 1 conveys information similar to conventional "power curves". Indeed, we can find the size of a test by reading its curves 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 9 that those curves are good surrogate to "power curves".

Figures 2 through 8 are counterparts of Figure 1 for other problems and/or algorithms.

Note that in order to keep the number of line types down 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 $n_2 = \frac{n}{10}$; curves without circled points correspond to $n_2 = \frac{n}{2}$ (40% thrown away). 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 bootstrap and the corrected bootstrap do not appear in Figure 1 and all its counterparts (except Figure 4 and Figure 6). We ignored them because, as we anticipated from political ratios shown in Table 1, the bootstrap test behaves like the resampled t-test and the corrected bootstrap test behaves like the corrected resampled t-test. If we don't ignore the bootstrap, some figures become too crowded. We made an exception and plotted curves corresponding to the bootstrap in Figures 4 and 6. In those two figures, the bootstrap and corrected 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 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

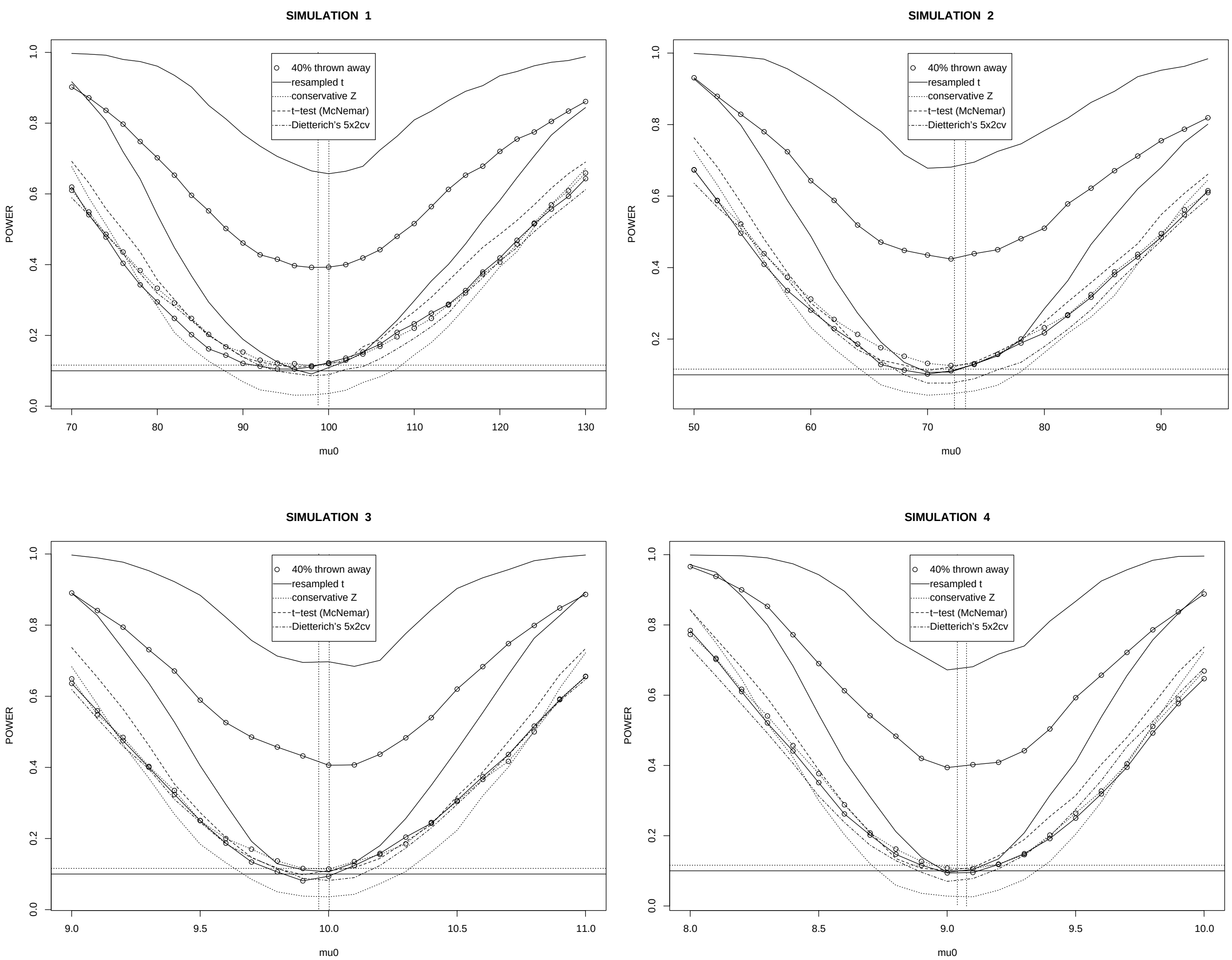

Figure 1: Powers of the tests about $H_0: \frac{n}{2}\mu_A = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{n}{2}\mu_A$ 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.

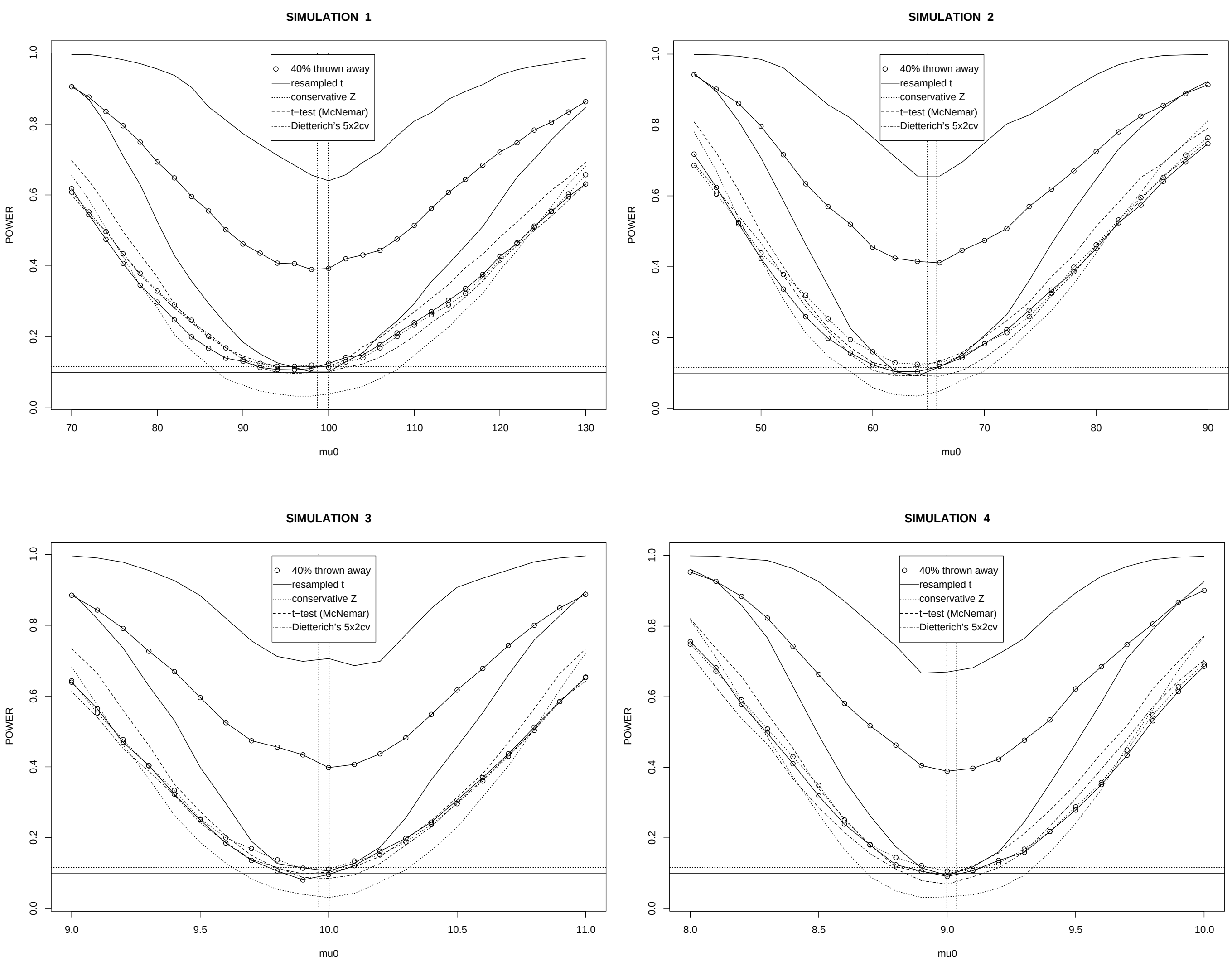

Figure 2: Powers of the tests about $H_0: \frac{n}{2}\mu_B = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{n}{2}\mu_B$ 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.

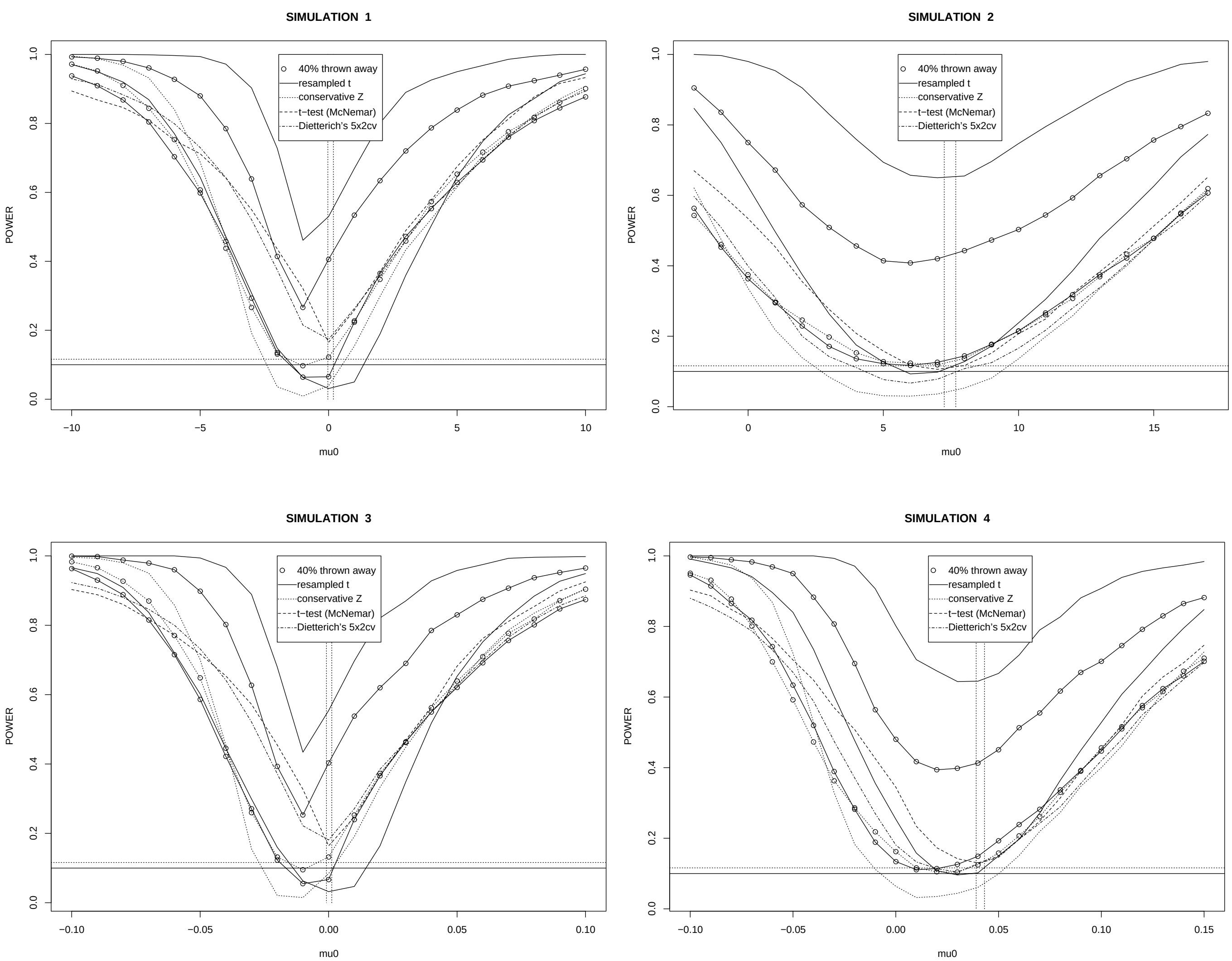

Figure 3: Powers of the tests about $H_0: \frac{n}{2}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{n}{2}\mu_{A-B}$ 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.

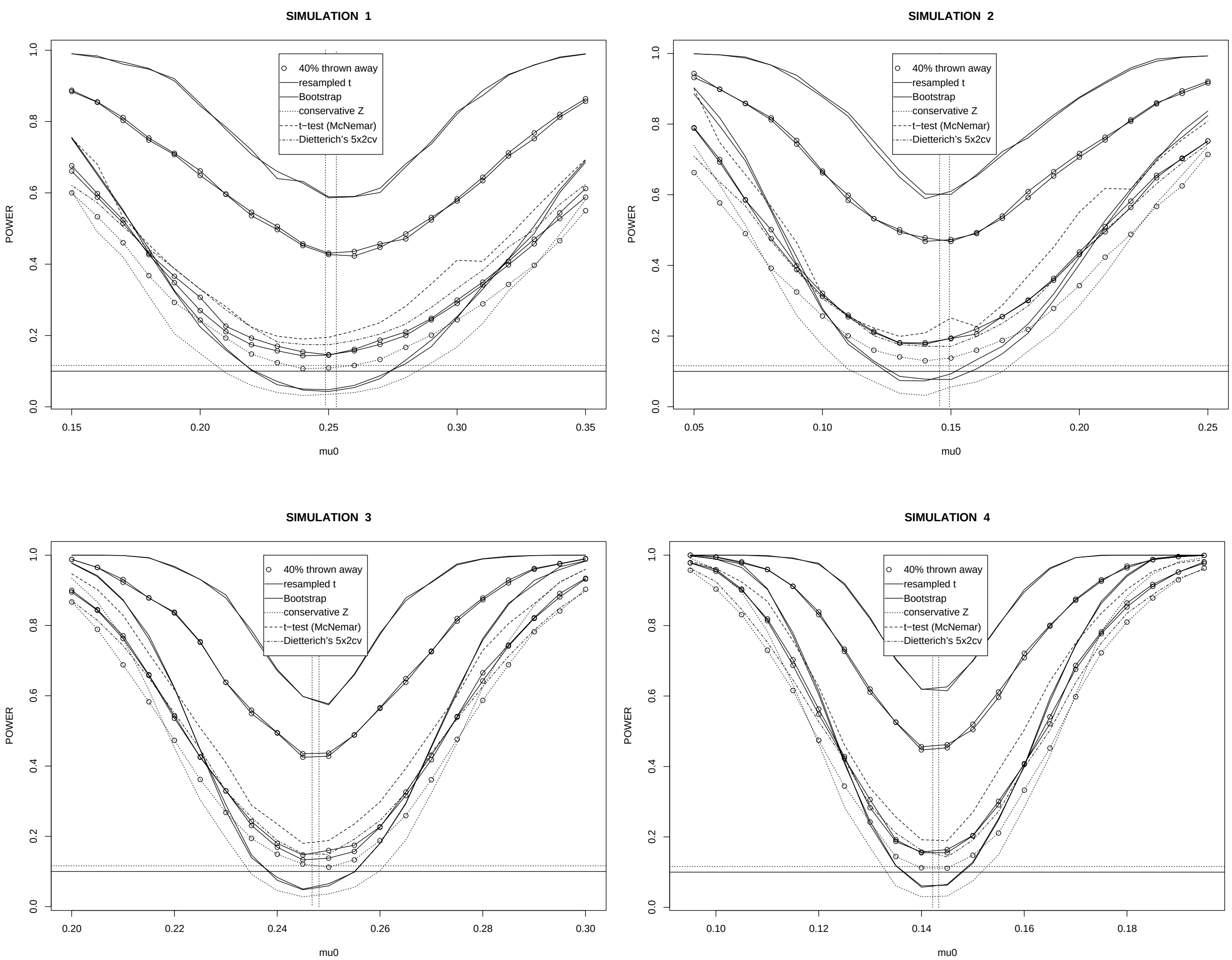

Figure 4: Powers of the tests about H_0 : $\frac{n}{2}\mu_A = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines correspond to the 95% confidence 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.

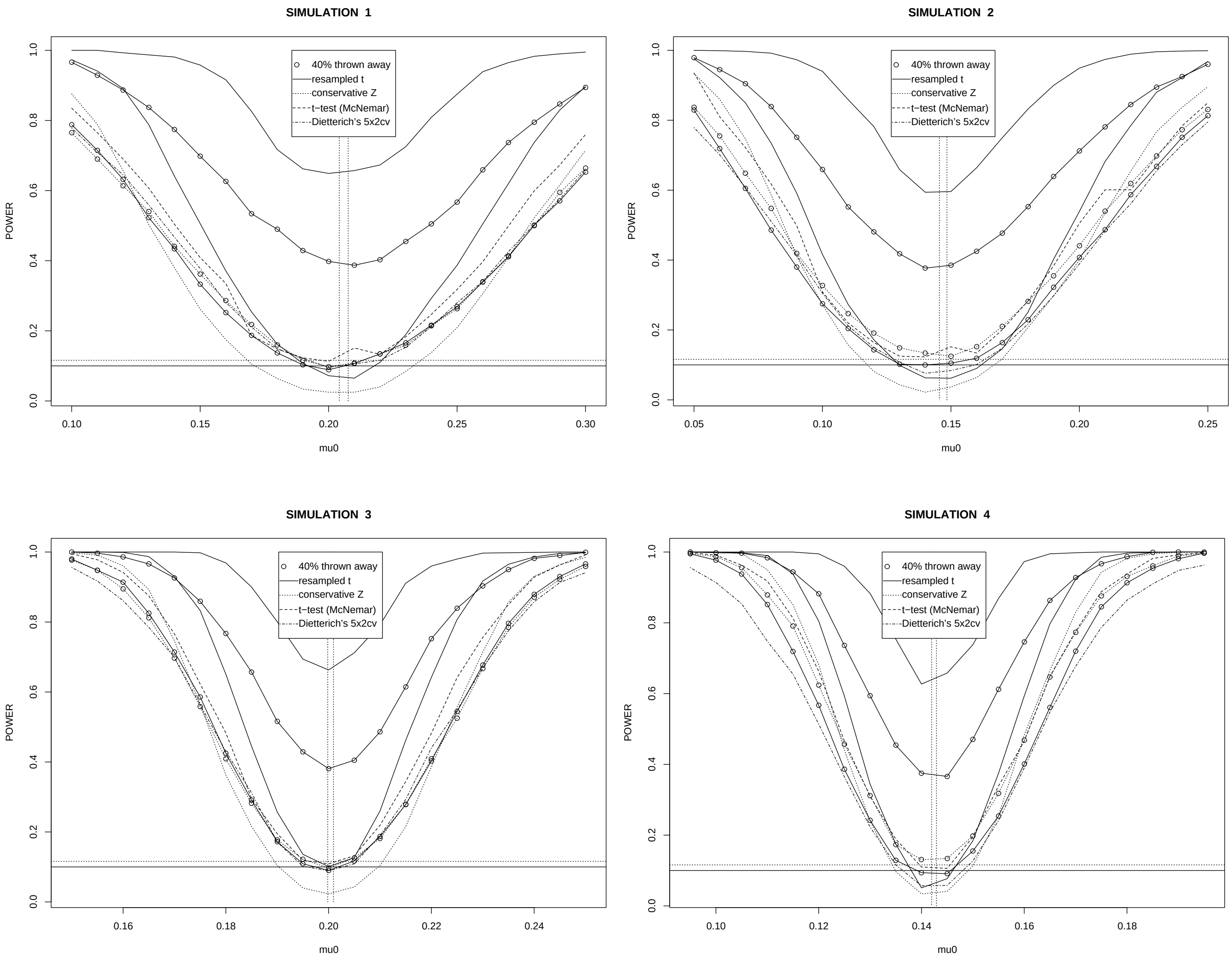

Figure 5: Powers of the tests about $H_0 : \frac{n}{2}\mu_B = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{n}{2}\mu_B$ 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.

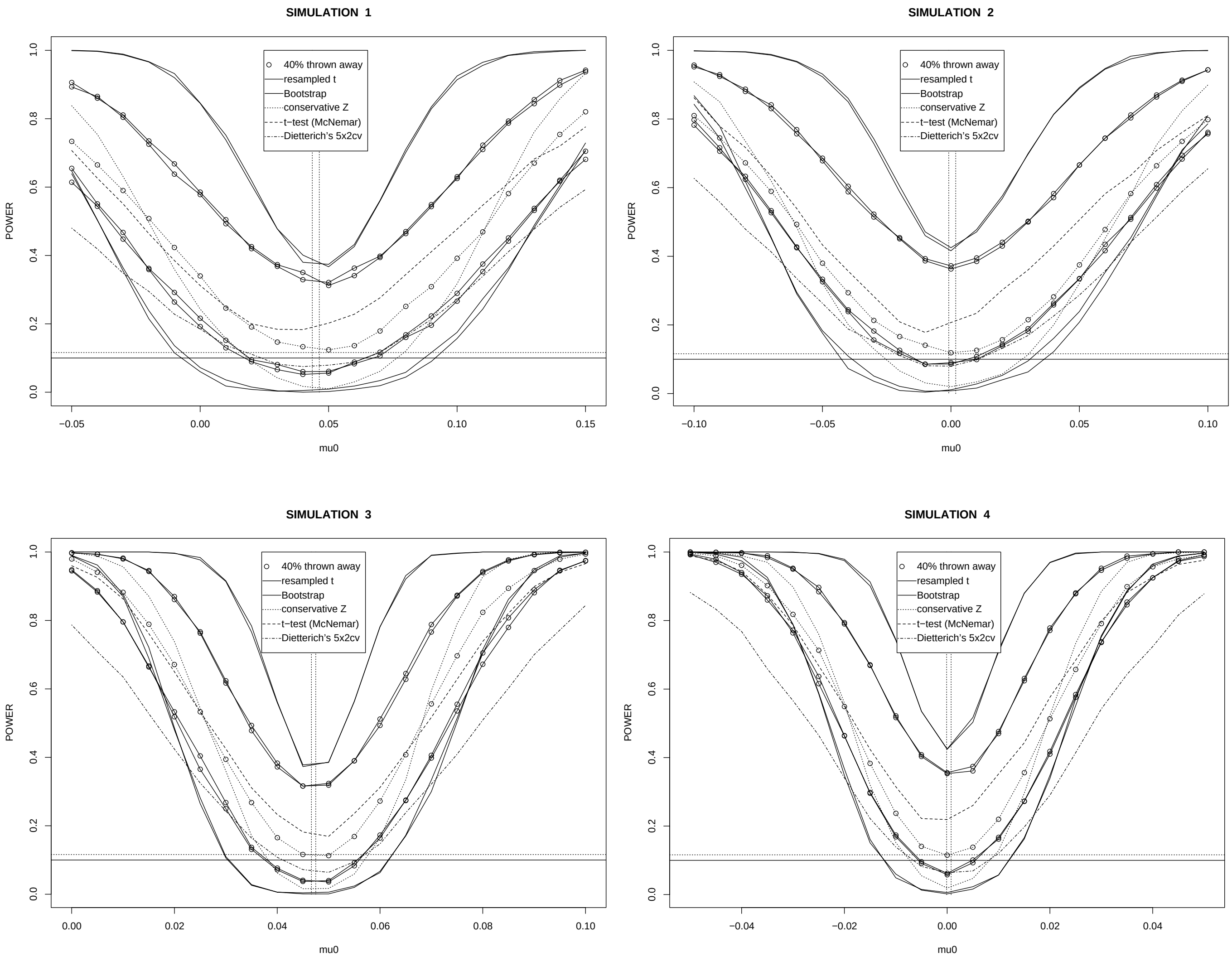

Figure 6: Powers of the tests about H_0 : $\frac{n}{2}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines correspond to the 95% confidence 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.

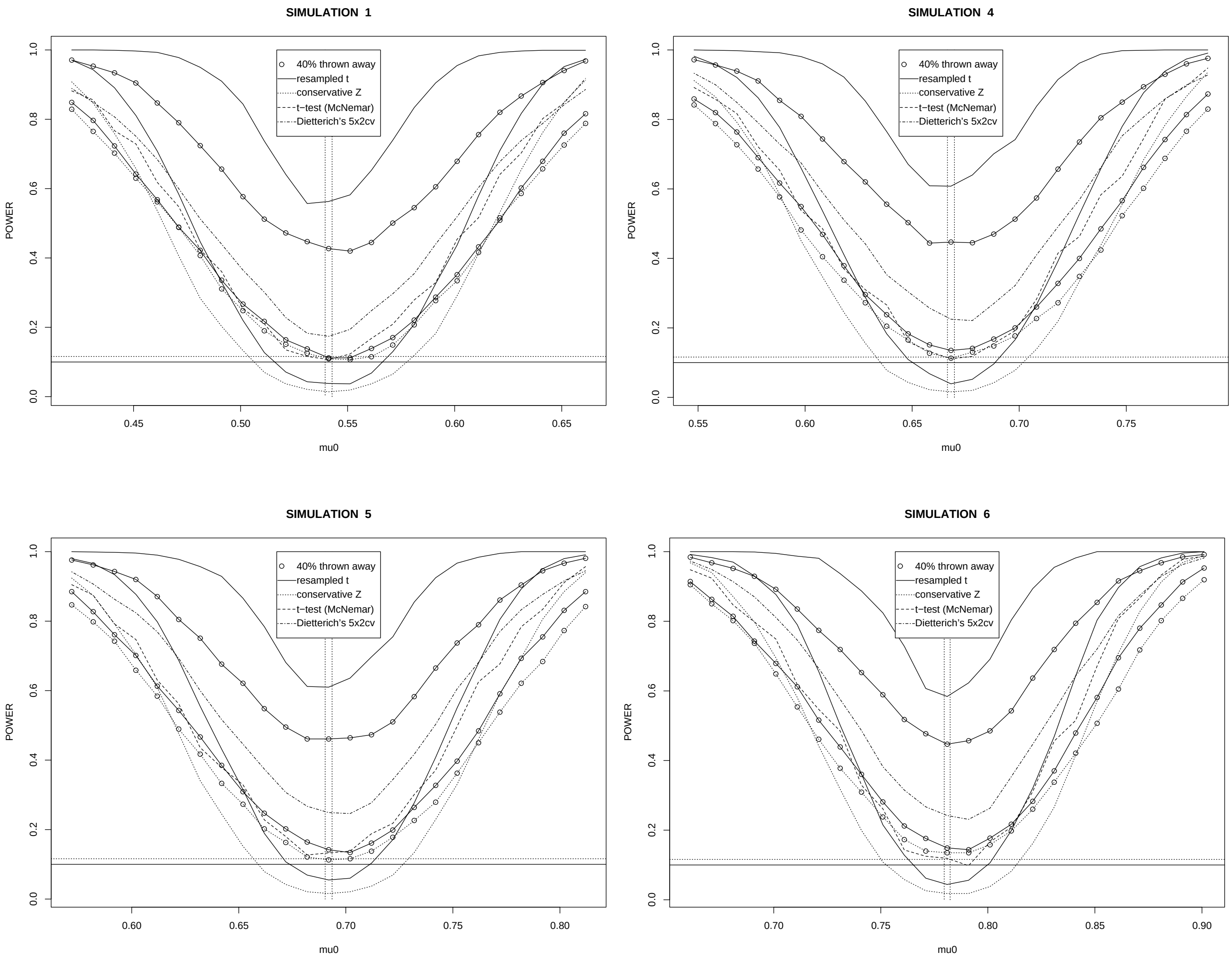

Figure 7: Powers of the tests about $H_0: \frac{n}{2}\mu_B = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the letter recognition problem. Each panel corresponds to one of the simulations design described in Table 4. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{n}{2}\mu_B$ 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.

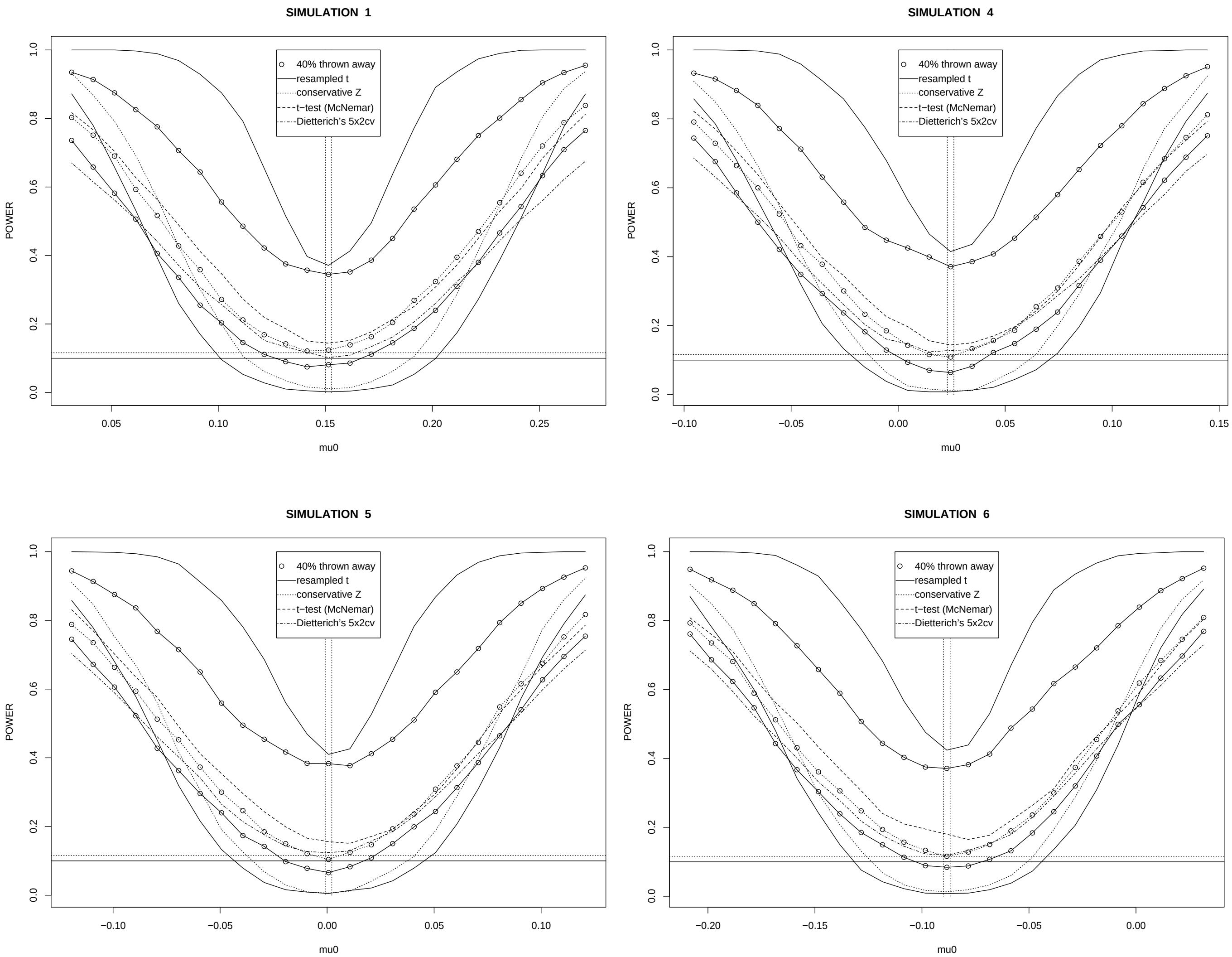

Figure 8: Powers of the tests about H_0 : $\frac{n}{2}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 for the letter recognition problem. Each panel corresponds to one of the simulations design described in Table 4. The dotted vertical lines correspond to the 95% confidence interval for the 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.

(Figures 4 and 6). We chose to show the bootstrap curves in Figures 4 and 6 because this is where the plots looked the least messy when the 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 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 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 4 and 7, 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 3, 4 and 6.
- 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

[14] Actually in Figure 3 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 we mentioned before, 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 vertical dotted lines. We don't observe this in other figures.

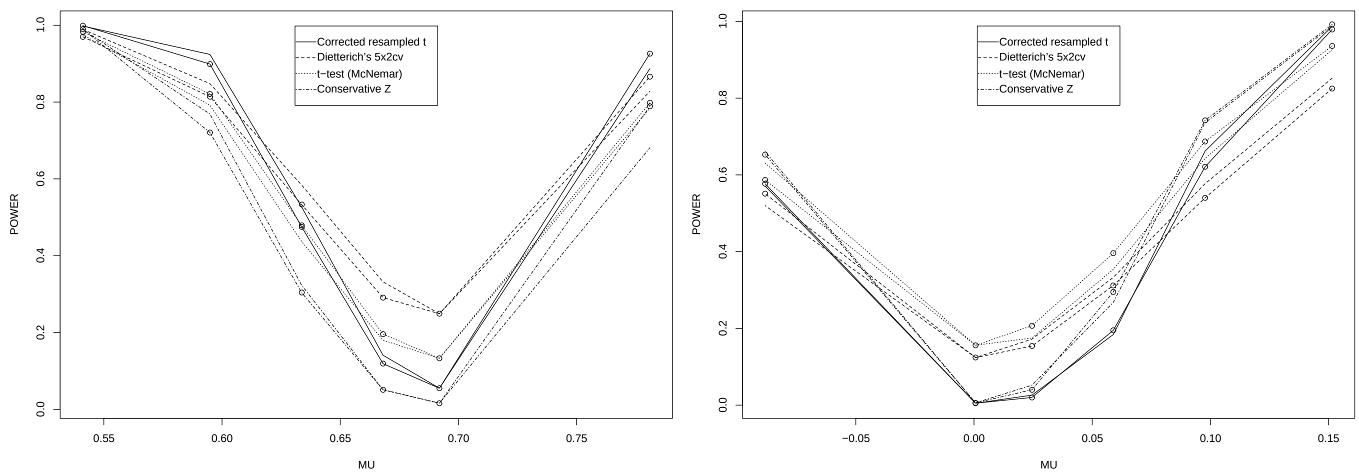

Figure 9: Real power curves (circle lines) and their surrogates (not circled) in the letter recognition problem. In the left panel, we see "real" and "surrogate" power curves for the the null hypothesis $H_0 : {}_{150}\mu_B = 0.692$. In the right panel, we see "real" and "surrogate" power curves for 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 "conservative Z" and the "corrected resampled t" statistics are those which do not throw away data.

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 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 the figures shown so far 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 9 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, with 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 bootstrap and the corrected 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 10 to 17) are similar to those of the Section 5.1. 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 worst 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 corrected resampled t-test, the plots again only confirms 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 worst when J increases; see Figure 13 for the liberal case. That makes sense since the 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, 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 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

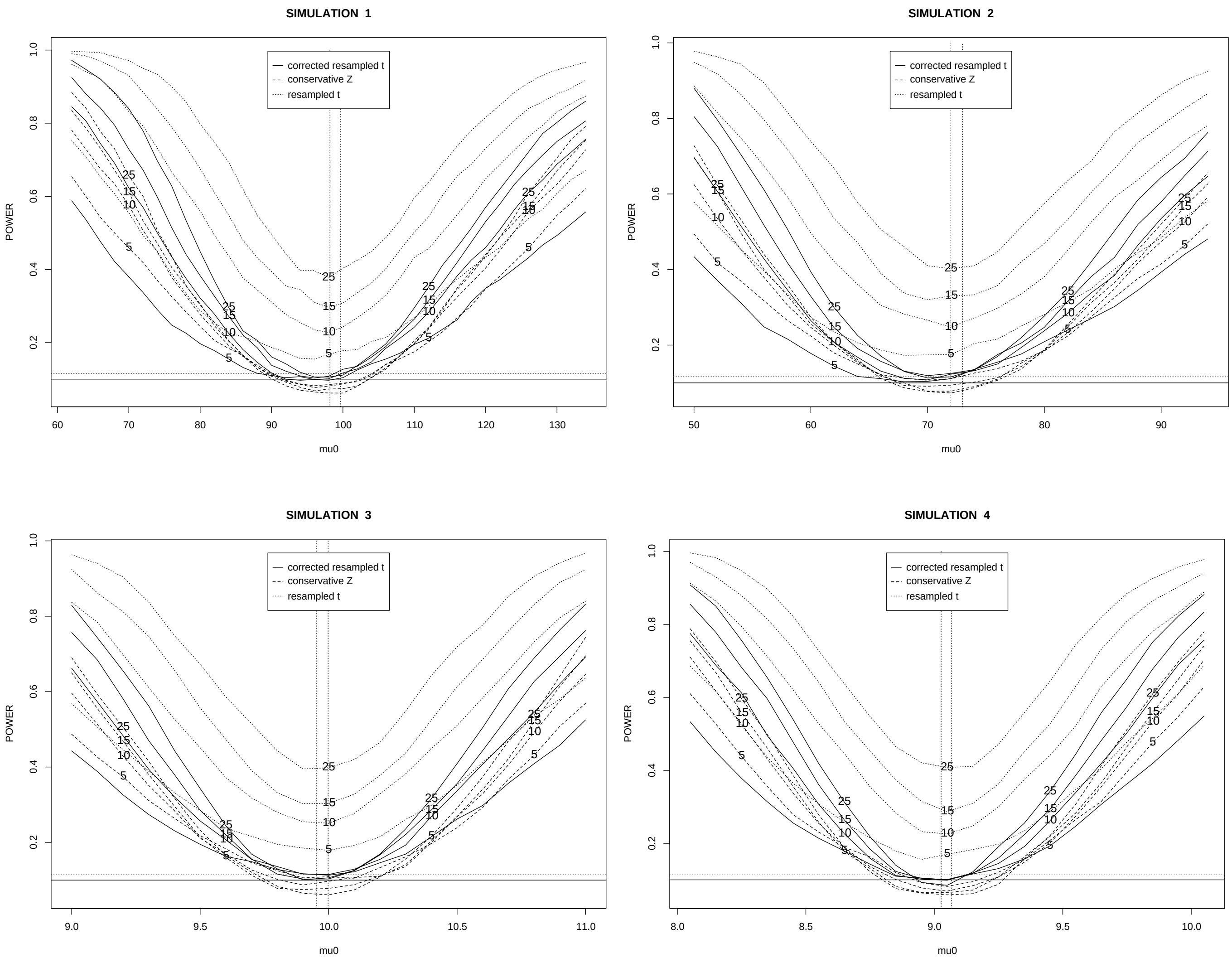

Figure 10: Powers of the tests about H_0 : $_{9n/10}\mu_A = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual $_{9n/10}\mu_A$ 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.

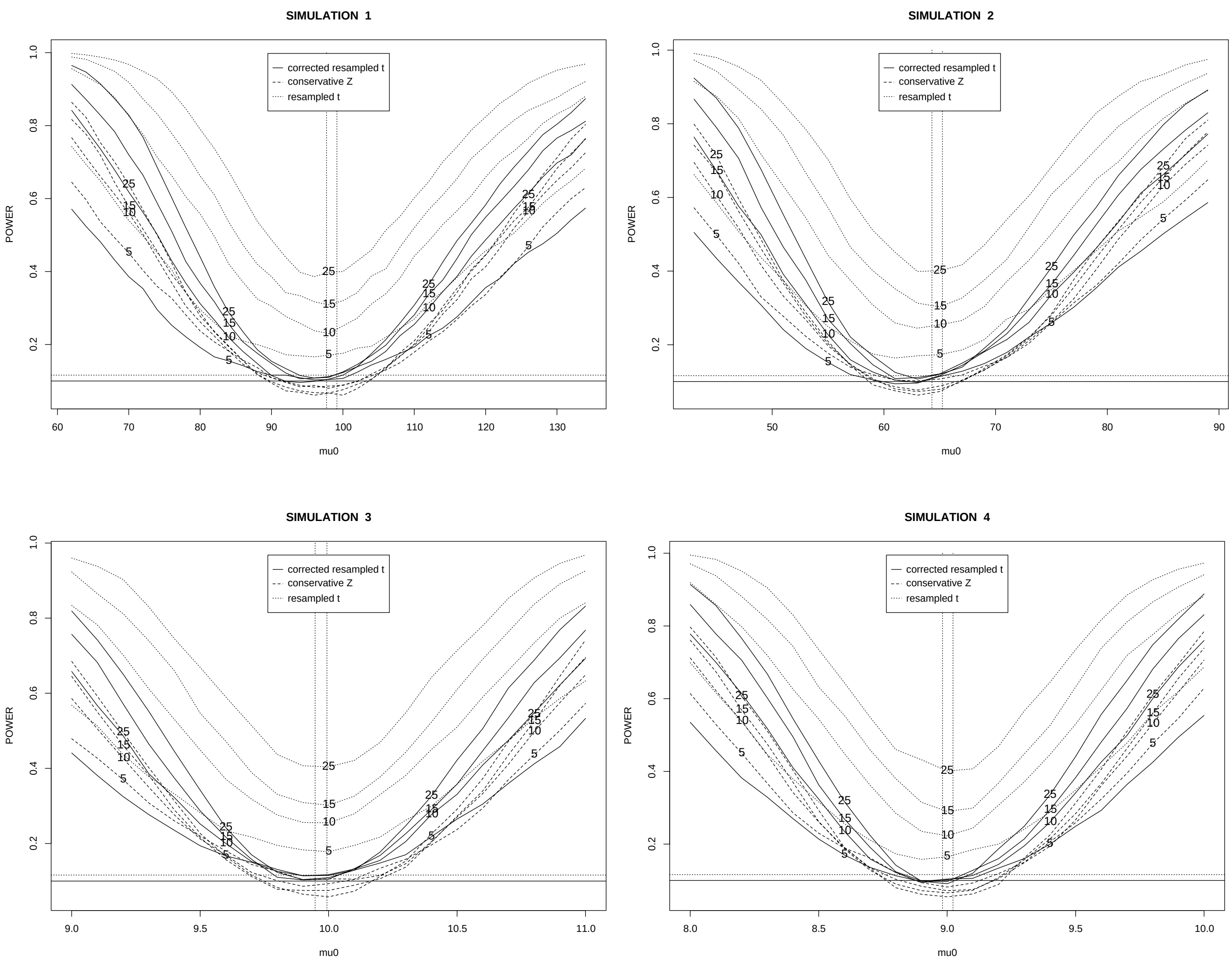

Figure 11: Powers of the tests about H_0 : ${}_{9n/10}\mu_B = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual ${}_{9n/10}\mu_B$ 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.

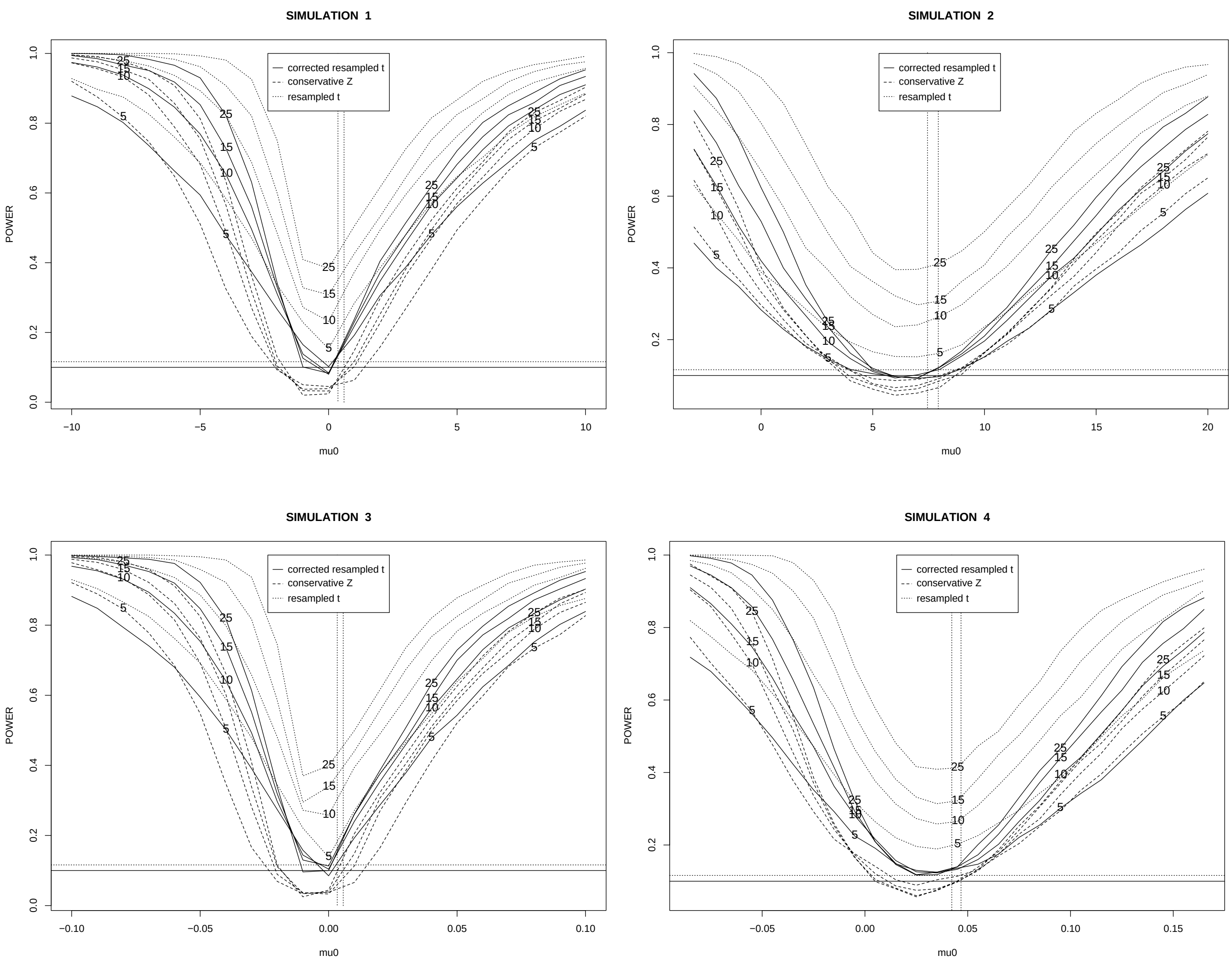

Figure 12: Powers of the tests about $H_0 : {}_{9n/10}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the 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.

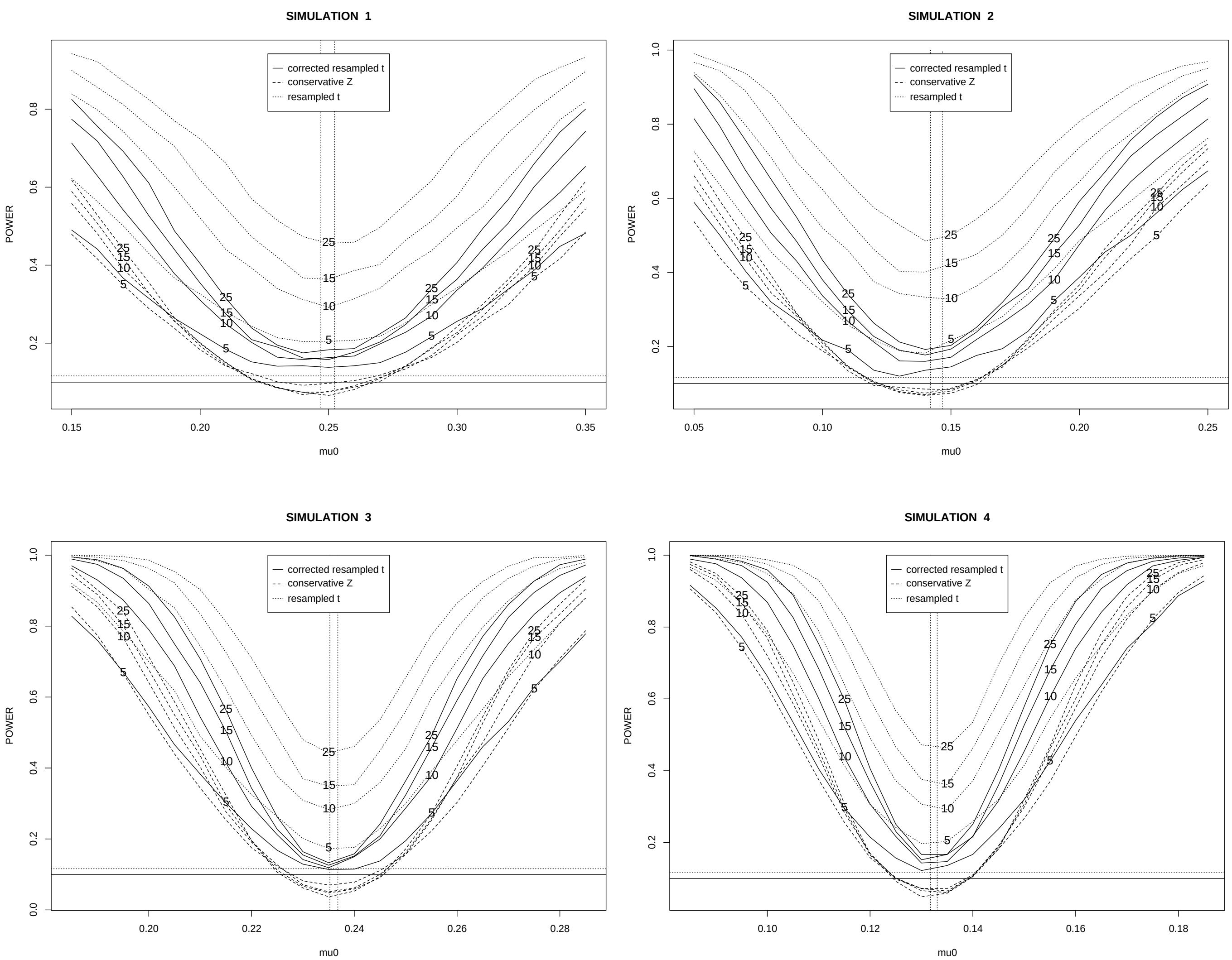

Figure 13: Powers of the tests about $H_0: \frac{9n}{10}\mu_A = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines correspond to the 95% confidence 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.

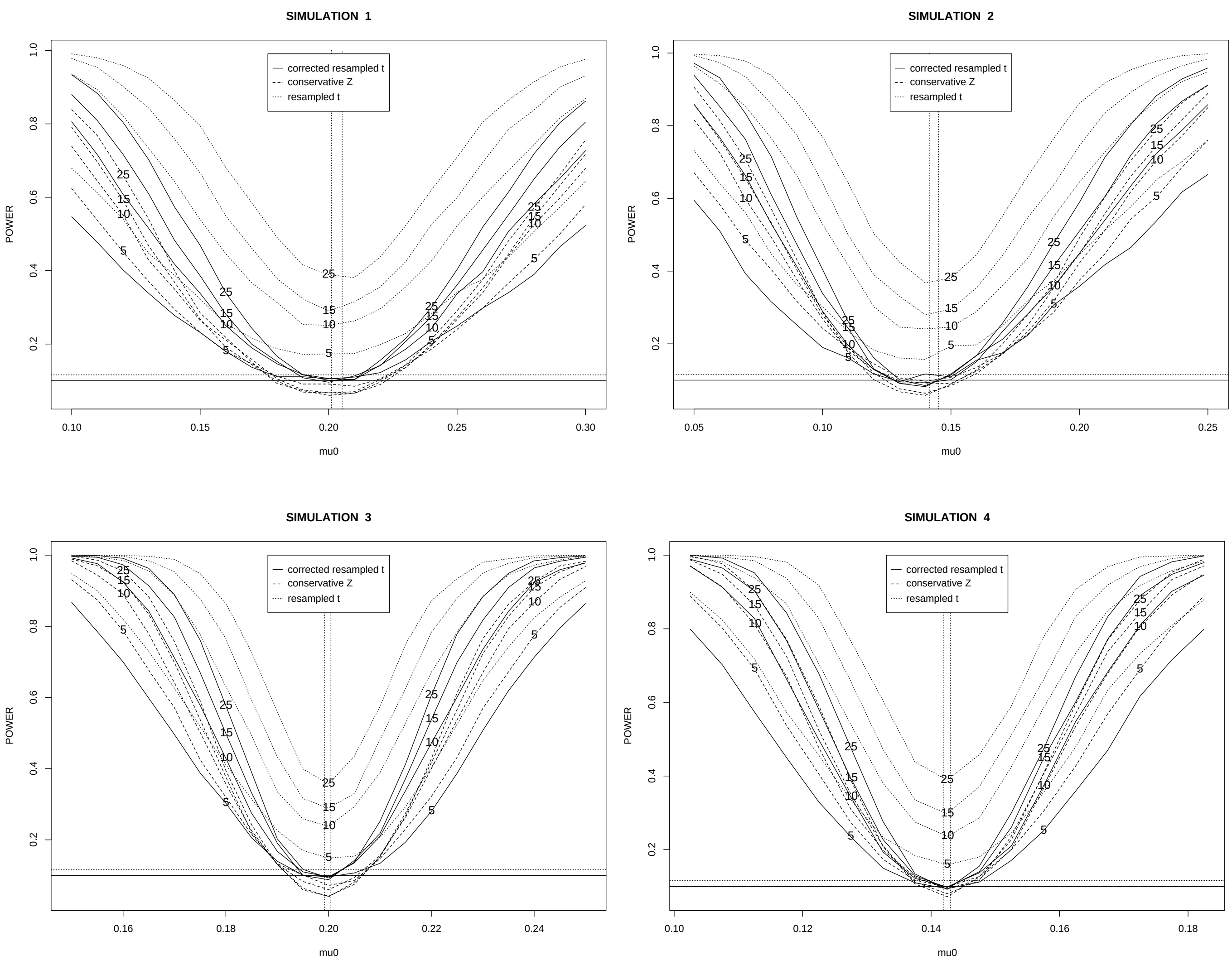

Figure 14: Powers of the tests about H_0 : $\frac{9n}{10}\mu_B = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{9n}{10}\mu_B$ 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.

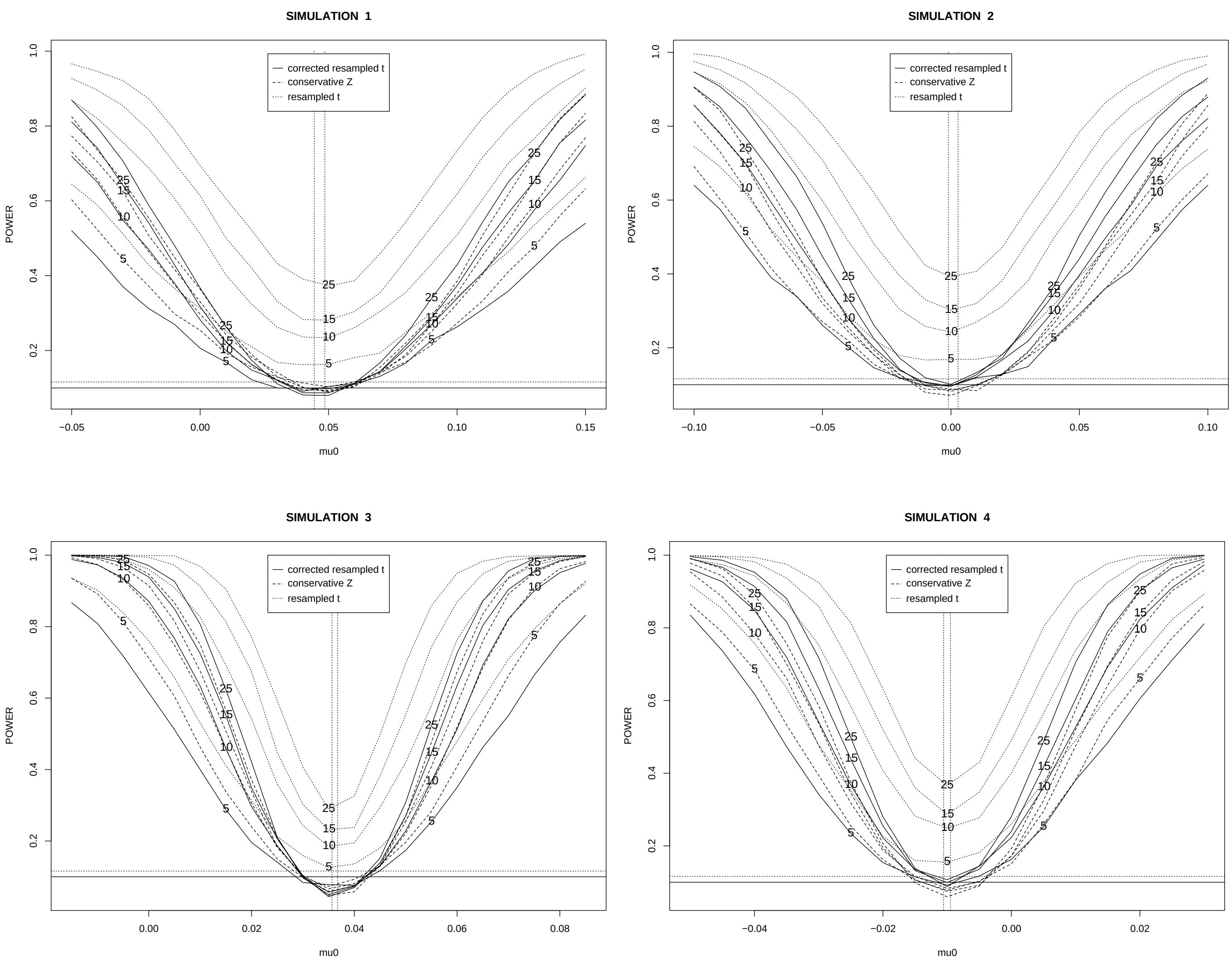

Figure 15: Powers of the tests about $H_0: \frac{9n}{10}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines correspond to the 95% 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.

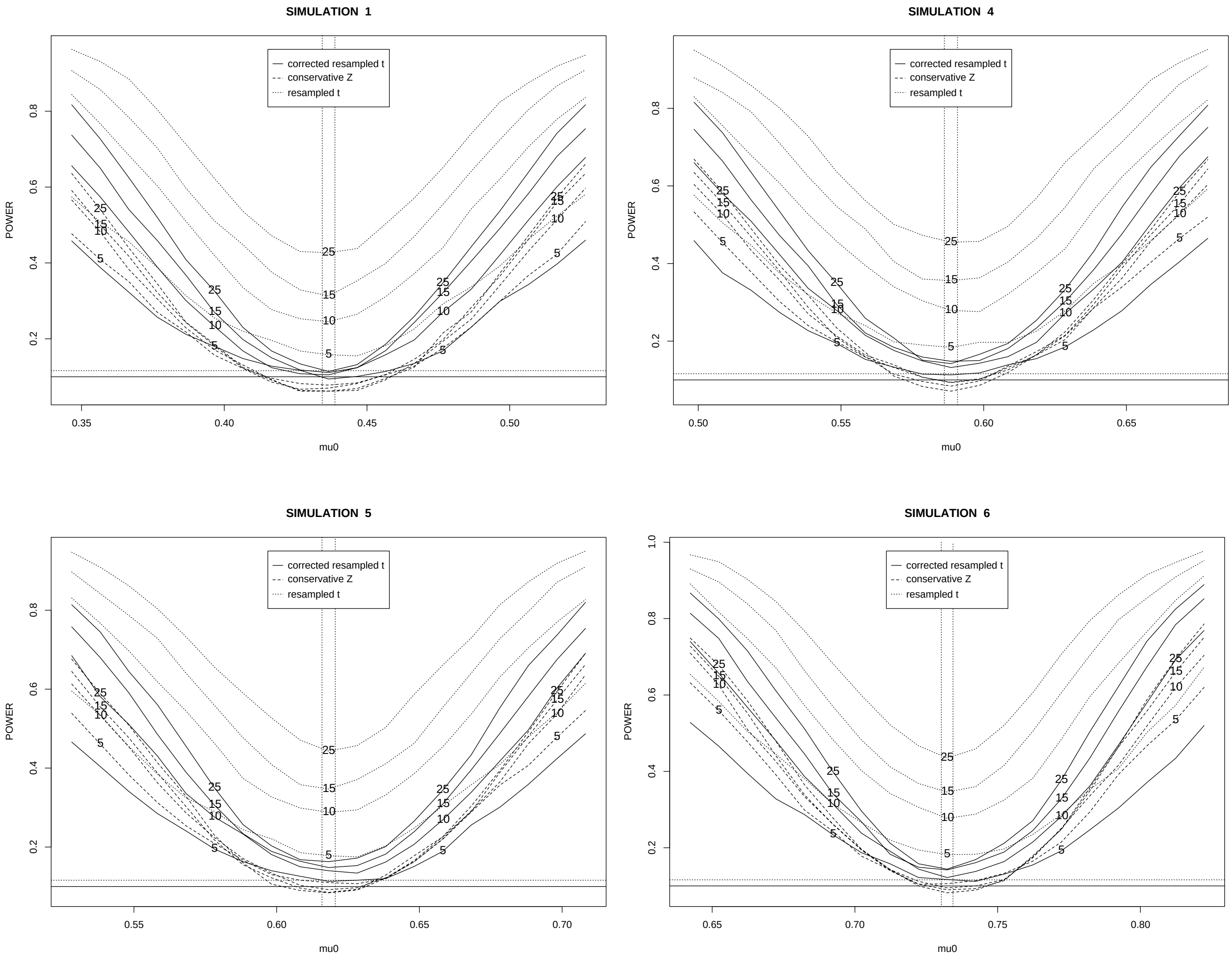

Figure 16: Powers of the tests about $H_0 : \frac{9n}{10}\mu_B = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the letter recognition problem. Each panel corresponds to one of the simulations design described in Table 4. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{9n}{10}\mu_B$ 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.

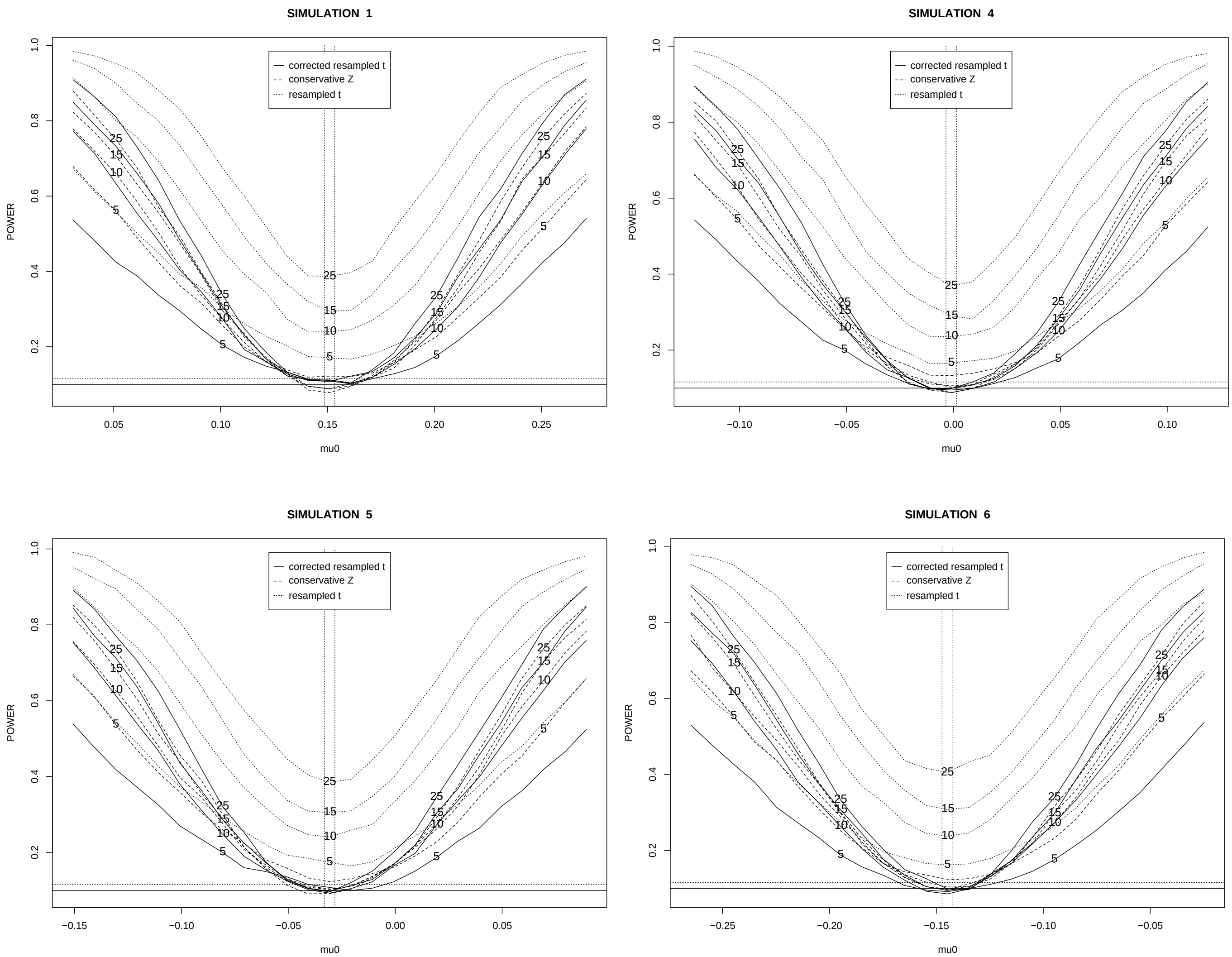

Figure 17: Powers of the tests about $H_0: \frac{9n}{10}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and J for the letter recognition problem. Each panel corresponds to one of the simulations design described in Table 4. The dotted vertical lines correspond to the 95% confidence interval 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.

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 15 and 17 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 worst 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 in 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 Figures 10 through 17 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 18 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 correspond to what 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.

5.3 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

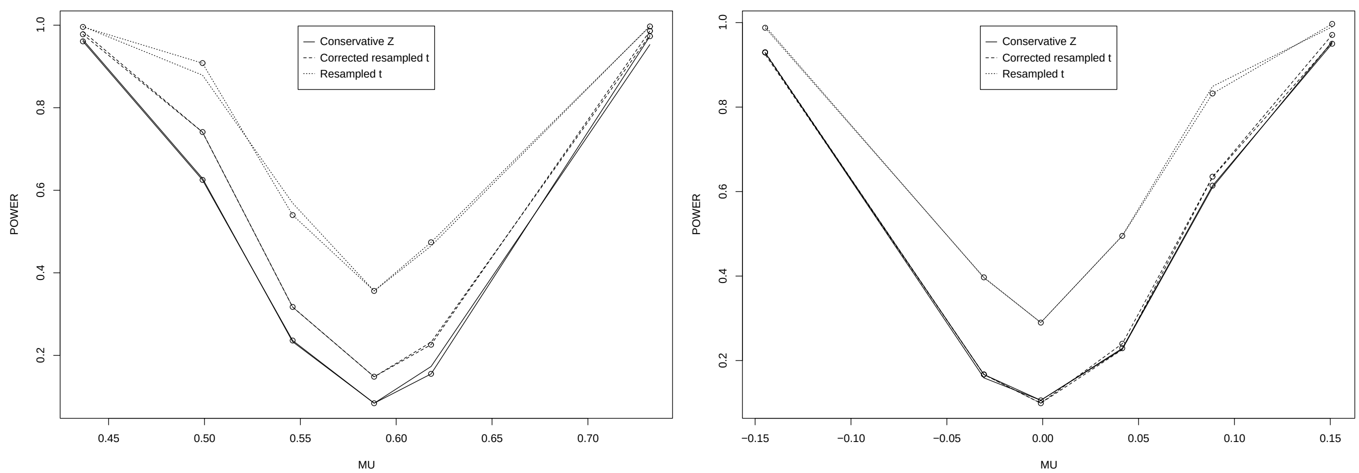

Figure 18: Real power curves (circle lines) and their surrogates (not circled) in the letter recognition problem. In the left panel, we see "real" and "surrogate" power curves for the the null hypothesis H_O : ${}_{270}\mu_B = 0.589$ In the right panel, we see "real" and "surrogate" power curves for the the null hypothesis H_O : ${}_{270}\mu_{A-B} = -0.001$. See the end of Section 5.2 for more details on their constructions.

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 Figures 19 through 26 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.

6 Conclusion

We have tackled the problem of estimating the variance of the cross-validation estimator of the generalization error. In this paper, we pay 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.

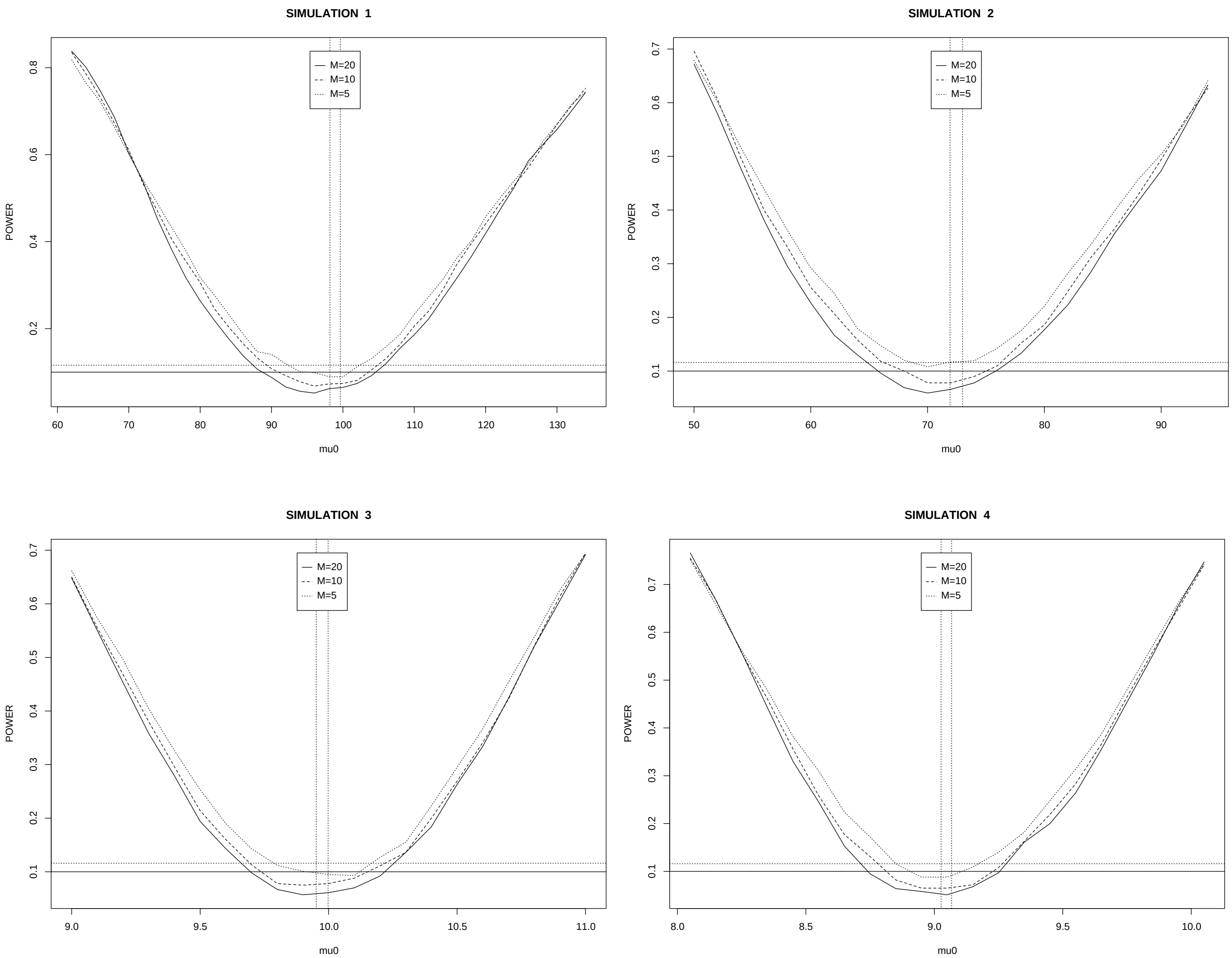

Figure 19: Powers of the conservative Z (with $J = 15$) about H_0 : ${}_{9n/10}\mu_A = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and M for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual ${}_{9n/10}\mu_A$ 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%).

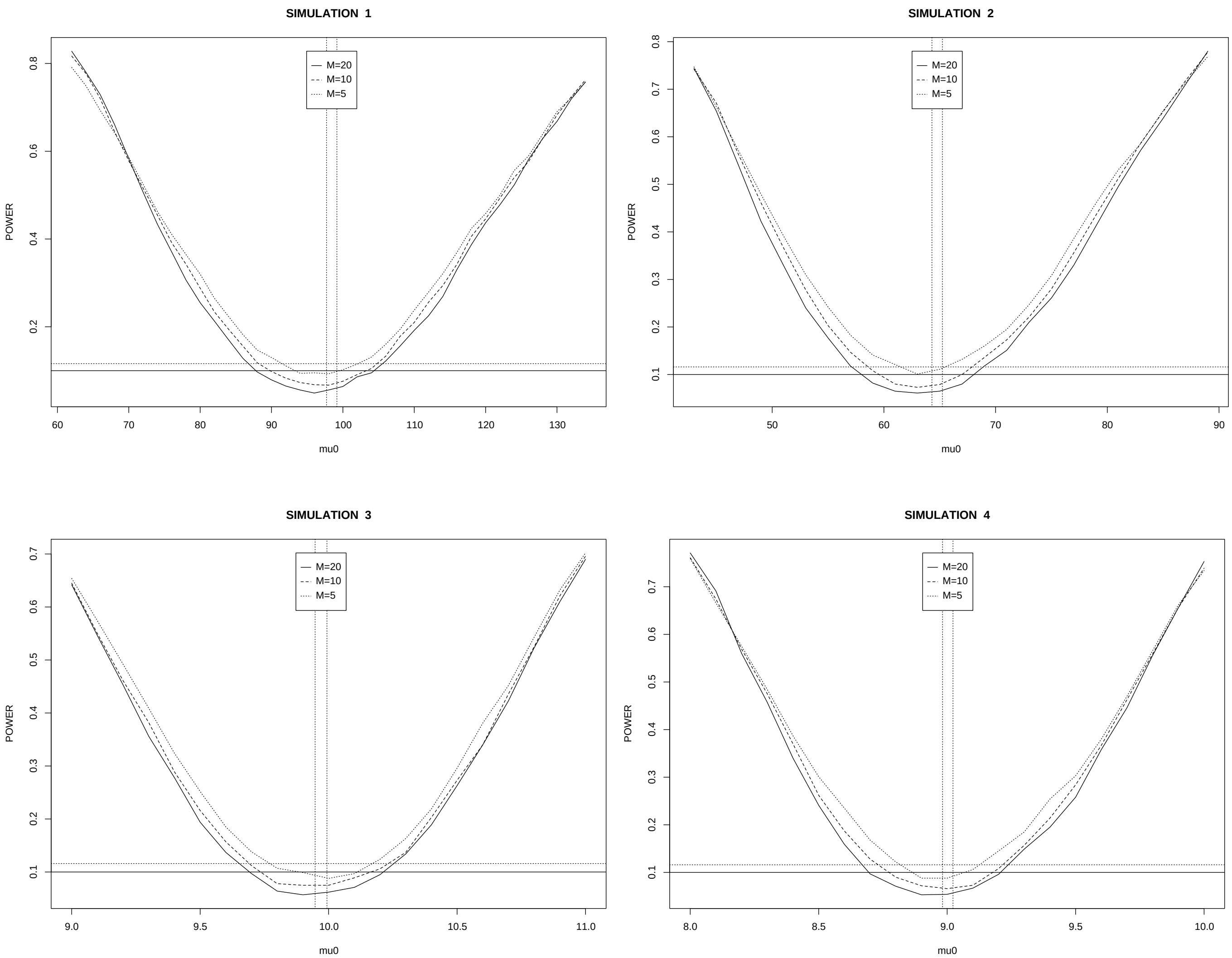

Figure 20: Powers of the conservative Z (with $J = 15$) about H_0 : ${}_{9n/10}\mu_B = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and M for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence interval for the actual ${}_{9n/10}\mu_B$ 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%).

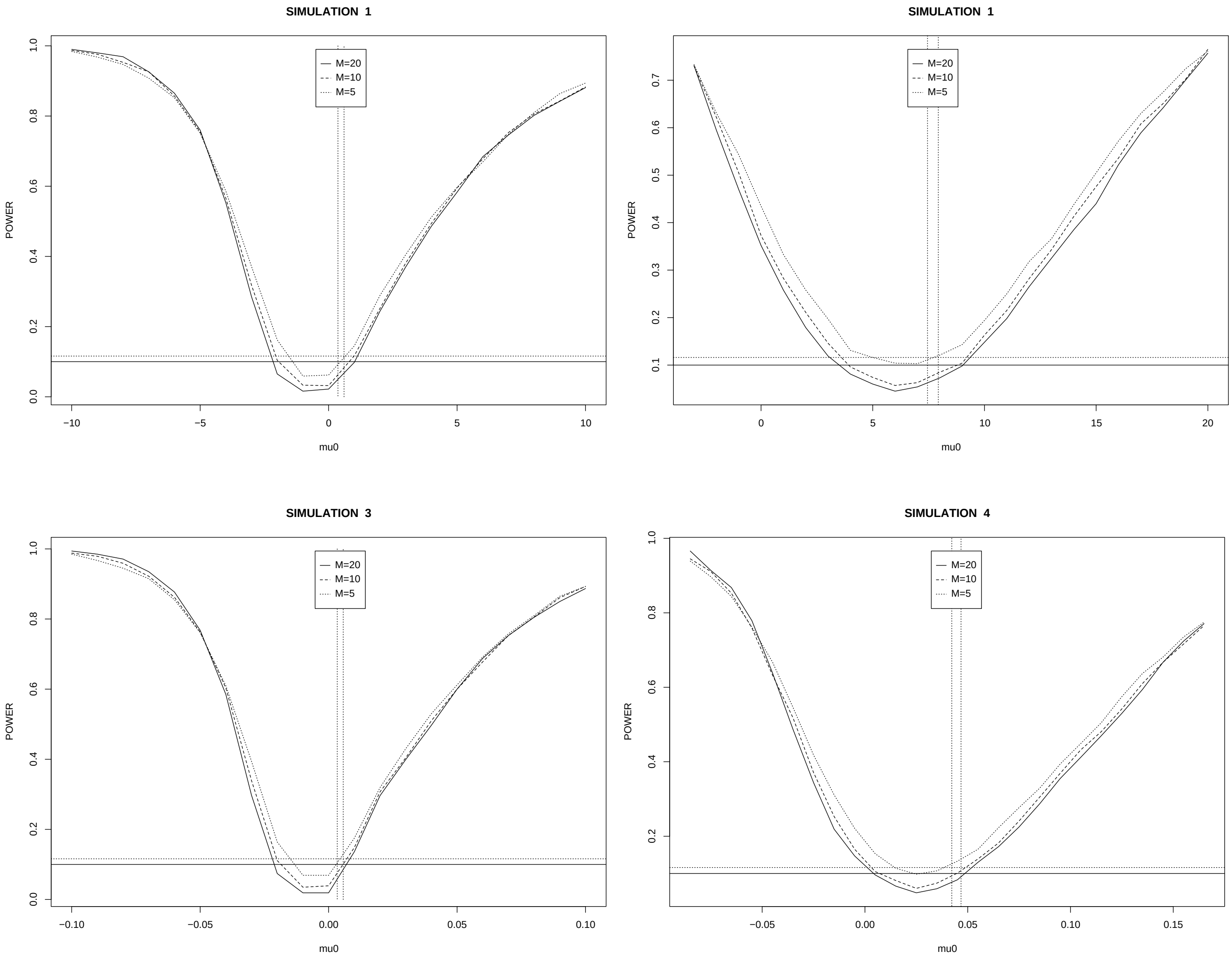

Figure 21: Powers of the conservative Z (with $J = 15$) about $H_0 : \ _{9n/10}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and M for the regression problem. Each panel corresponds to one of the simulations design described in Table 2. The dotted vertical lines correspond to the 95% confidence 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%).

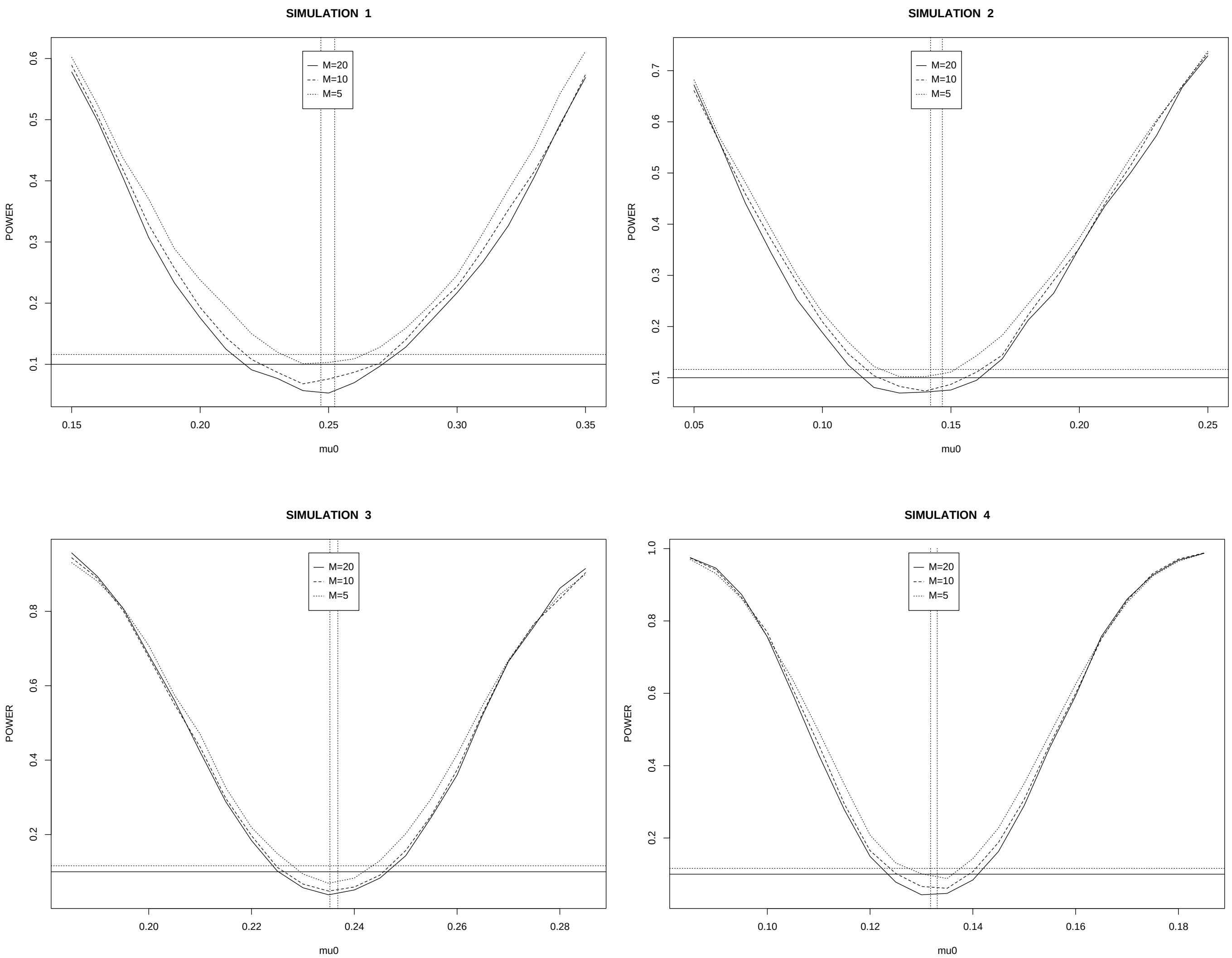

Figure 22: Powers of the conservative Z (with $J = 15$) about H_0 : $\frac{9n}{10}\mu_A = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and M for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines 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%).

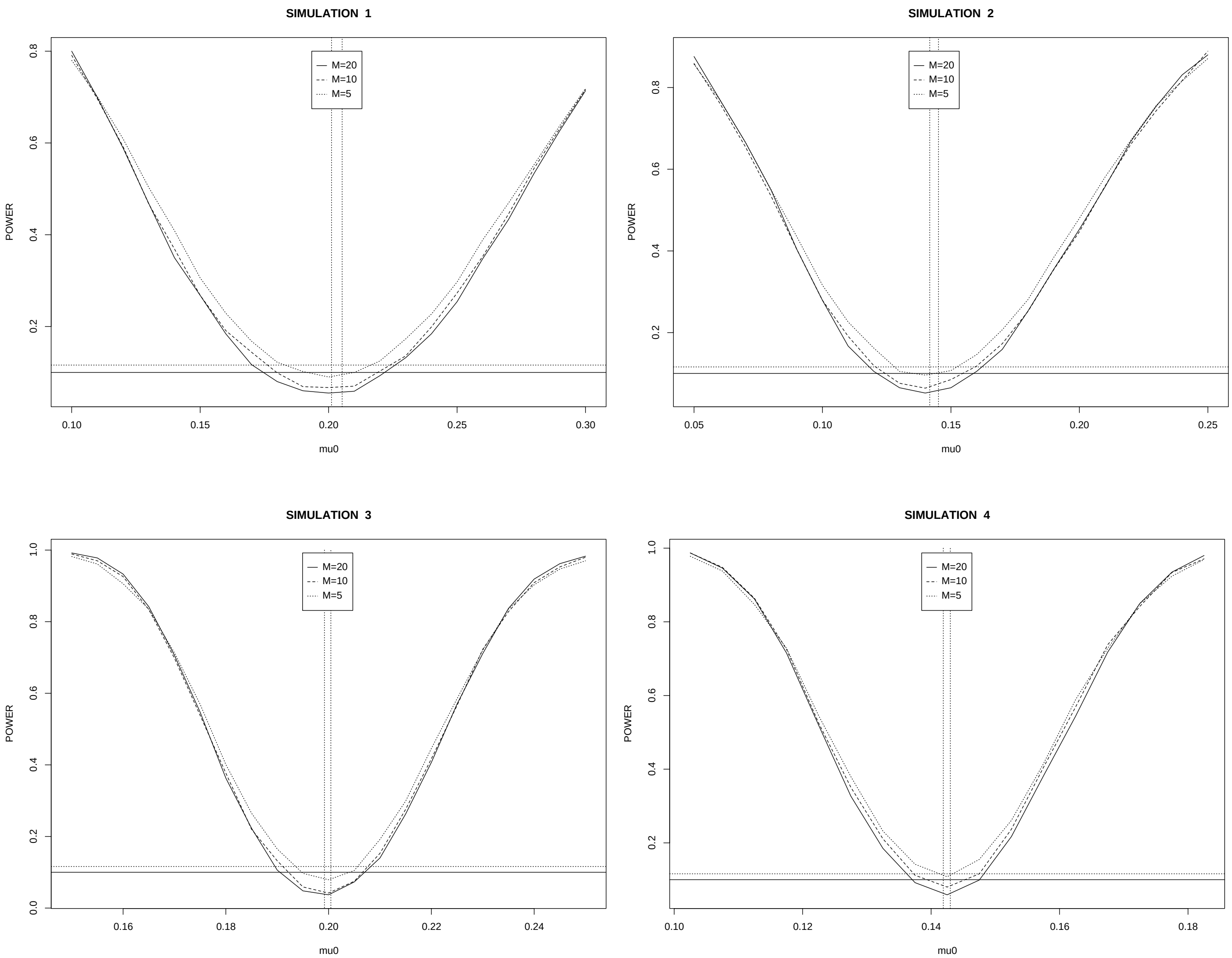

Figure 23: Powers of the conservative Z (with $J = 15$) about H_0 : $\frac{9n}{10}\mu_B = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and M for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{9n}{10}\mu_B$ 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%).

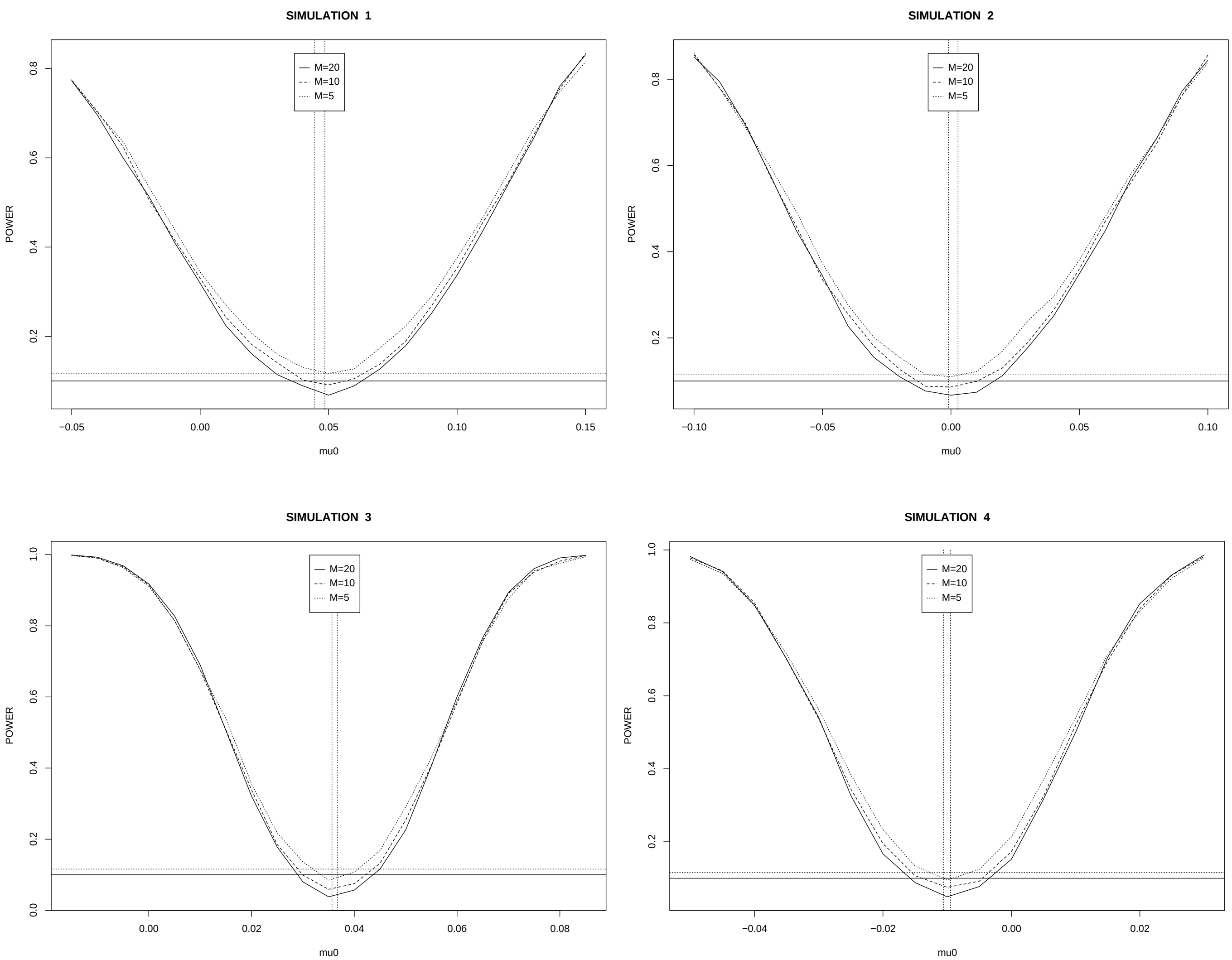

Figure 24: Powers of the conservative Z (with $J = 15$) about H_0 : $\frac{9n}{10}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and M for the classification of Gaussian populations problem. Each panel corresponds to one of the simulations design described in Table 3. The dotted vertical 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%).

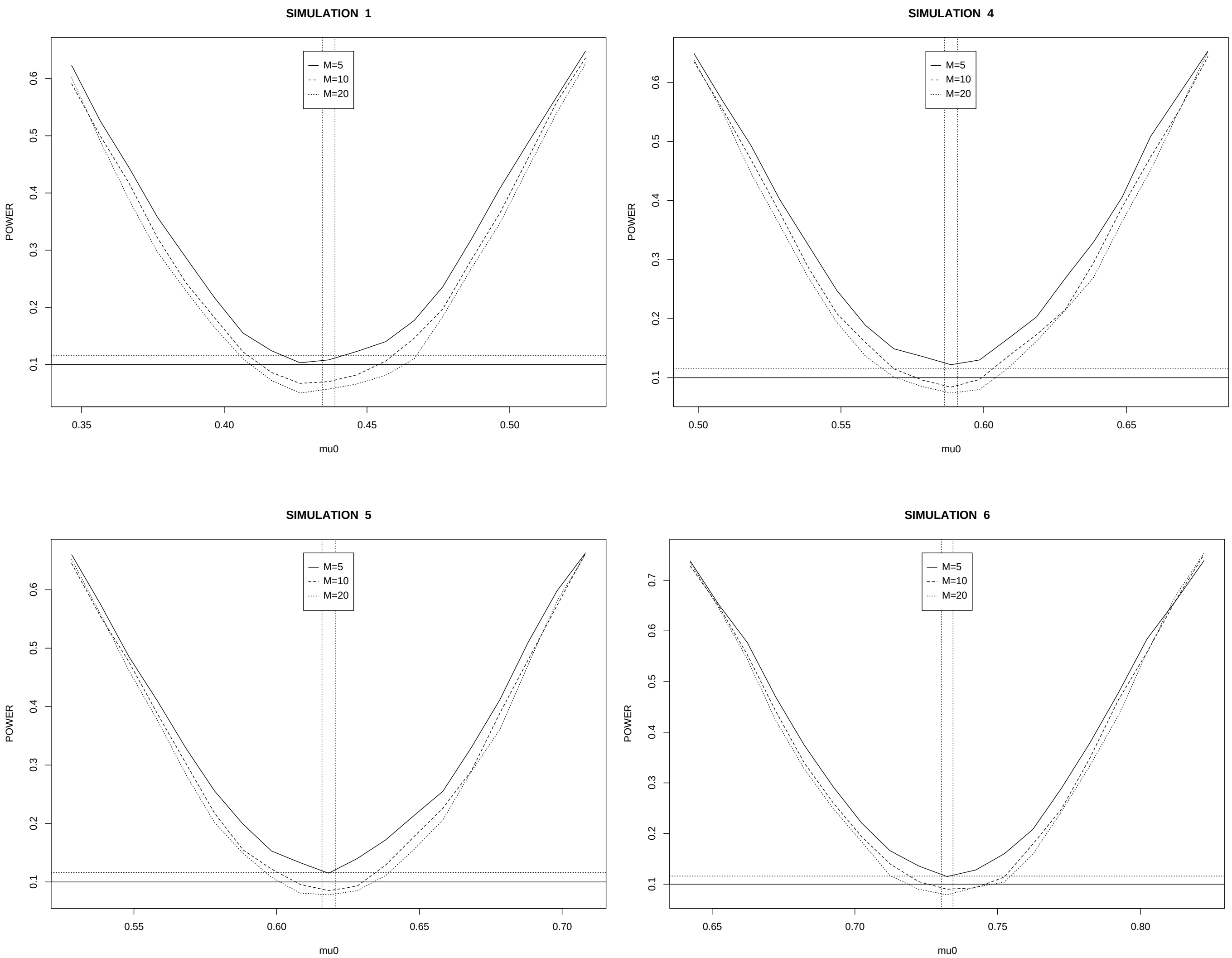

Figure 25: Powers of the conservative Z (with $J = 15$) about H_0 : $\frac{9n}{10}\mu_B = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and M for the letter recognition problem. Each panel corresponds to one of the simulations design described in Table 4. The dotted vertical lines correspond to the 95% confidence interval for the actual $\frac{9n}{10}\mu_B$ 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%).

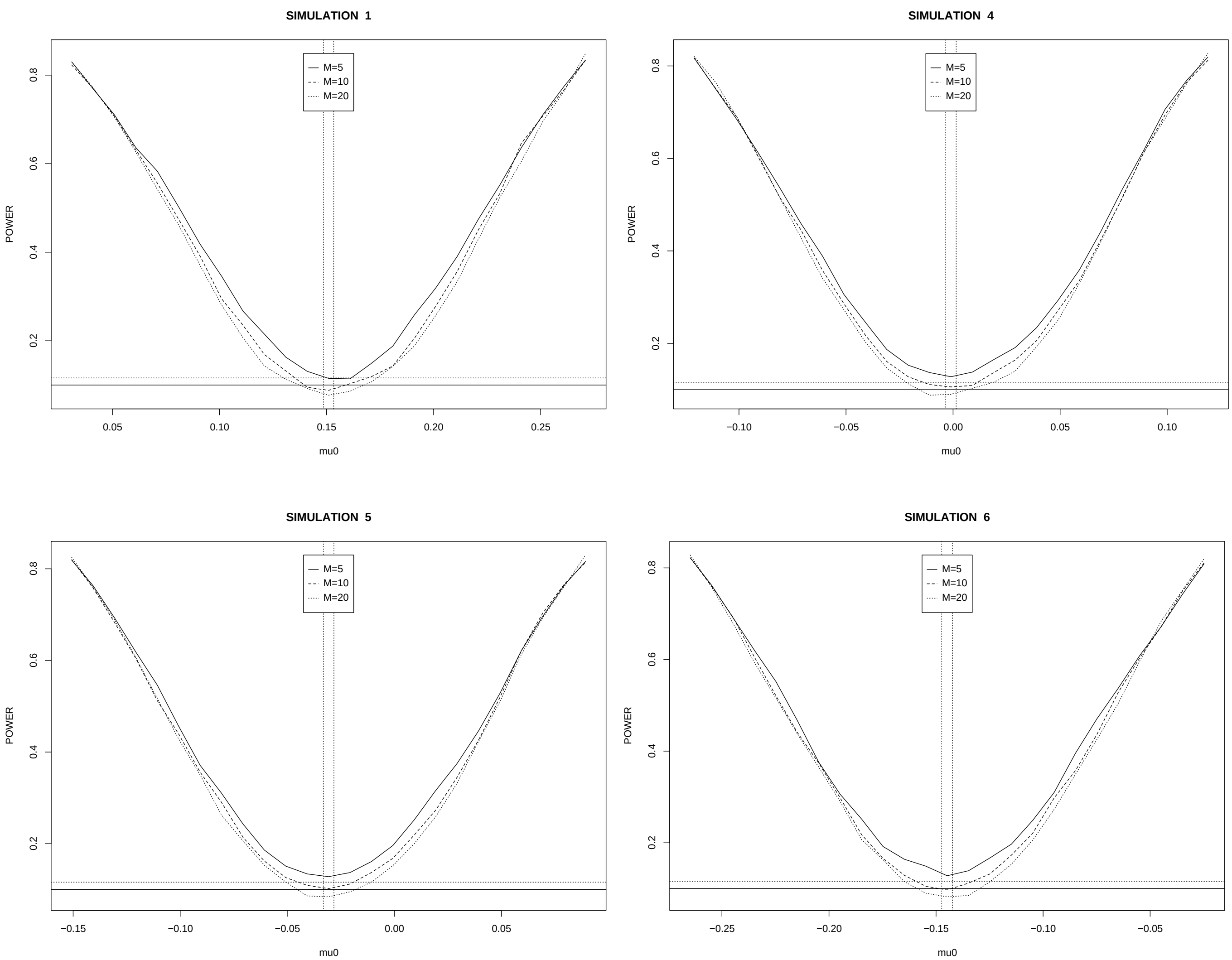

Figure 26: Powers of the conservative Z (with $J = 15$) about $H_0 : \frac{9n}{10}\mu_{A-B} = \mu_0$ at level $\alpha = 0.1$ for varying μ_0 and M for the letter recognition problem. Each panel corresponds to one of the simulations design described in Table 4. The dotted vertical lines correspond to 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%).

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 found that there is no general non-negative unbiased estimator of the variance of a large class of cross-validation estimates based only on the individual test errors involved in the computation of this estimate. 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 similar to 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 prevision of a continuous output. 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).

References

Blake, C., Keogh, E., and 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., and 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.

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

Efron, B. and 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.

Hinton, G., Neal, R., Tibshirani, R., and DELVE team members (1995). Assessing learning procedures using DELVE. Technical report, University of Toronto, Department of Computer Science.

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

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.

Appendix

A.1 **Proof of Lemma 2**

Let $U = (U_1, \ldots, U_K, U_{K+1})$ and $U_- = (U_1, \ldots, U_K)$. Let H be a $(K-1) \times K$ matrix such that $HH' = I_{K-1}$ and $H\mathbf{1}_K = \mathbf{0}_{K-1}$ (the lines of H form an orthogonal basis of the space orthogonal to $\mathbf{1}_K$). Let A be the following $((K-1)+1+1) \times (K+1)$ matrix:

$$A = \left[\begin{array}{c|c} H & \mathbf{0}_{K-1} \\ \hline \frac{1}{\sqrt{K}}\mathbf{1}'_K & 0 \\ \hline \mathbf{0}'_K & 1 \end{array}\right].$$

Since $U \sim N_{K+1}(E[U], \mathit{Var}[U])$, we have

$$AU = \begin{pmatrix} HU_- \\ \sqrt{K}\bar{U} \\ U_{K+1} \end{pmatrix} \sim N_{K+1}(E[AU], \mathit{Var}[AU]),$$

with

$$E[AU] = AE[U] = A(\beta \mathbf{1}_{K+1}) = \beta \begin{pmatrix} \mathbf{0}_{K-1} \\ \sqrt{K} \\ 1 \end{pmatrix} = \begin{pmatrix} E[HU_-] \\ E[\sqrt{K}\bar{U}] \\ E[U_{K+1}] \end{pmatrix}.$$

Regarding the variance, note that A is an orthonormal matrix since its lines are orthogonal to each other and have unit lengths. Therefore

$$\begin{aligned} \mathit{Var}[AU] &= A\,\mathit{Var}[U]A' = A[\delta(1-\pi)I_{K+1} + \gamma \mathbf{1}_{K+1}^{\otimes 2}]A' \\ &= \delta(1-\pi)I_{K+1} + \gamma(A\mathbf{1}_{K+1})^{\otimes 2} = \delta(1-\pi)I_{K+1} + \delta\pi \begin{pmatrix} \mathbf{0}_{K-1} \\ \sqrt{K} \\ 1 \end{pmatrix}^{\otimes 2} \\ &= \left[\begin{array}{c|cc} \delta(1-\pi)I_{K-1} & \multicolumn{2}{c}{\mathbf{0}_{(K-1)\times 2}} \\ \hline \mathbf{0}_{2\times(K-1)} & \delta(K\pi + 1 - \pi) & \sqrt{K}\delta\pi \\ & \sqrt{K}\delta\pi & \delta(1-\pi+\pi) \end{array}\right] \\ &= \left[\begin{array}{c|cc} \mathit{Var}[HU_-] & \multicolumn{2}{c}{\mathbf{0}_{(K-1)\times 2}} \\ \hline \mathbf{0}_{2\times(K-1)} & \mathit{Var}[\sqrt{K}\bar{U}] & \mathit{Cov}[\sqrt{K}\bar{U}, U_{K+1}] \\ & \mathit{Cov}[U_{K+1}, \sqrt{K}\bar{U}] & \mathit{Var}[U_{K+1}] \end{array}\right] \end{aligned}$$

where $v^{\otimes 2} = vv'$.

Define $T = (HU_-)'HU_-$. Since $HU_- \sim N_{K-1}(\mathbf{0}_{K-1}, \delta(1-\pi)I_{K-1})$, we have $\frac{T}{\delta(1-\pi)} \sim \chi^2_{K-1}$. From the structure of $\mathit{Var}[AU]$ we have that

- HU_- is independent of $\sqrt{K}\bar{U}$ and thus T and $\bar{U}$ are independent,
- HU_- is independent of U_{K+1} and thus T and U_{K+1} are independent.

Therefore we have

$$\frac{\frac{\sqrt{K}\bar{U} - \sqrt{K}\beta}{\sqrt{\delta(K\pi+1-\pi)}}}{\sqrt{\frac{T}{(K-1)\delta(1-\pi)}}} = \sqrt{\frac{1-\pi}{1-\pi+K\pi}} \frac{\sqrt{K}(\bar{U}-\beta)}{\sqrt{\frac{T}{K-1}}} \sim t_{K-1}$$

and

$$\frac{\frac{U_{K+1}-\beta}{\sqrt{\delta}}}{\sqrt{\frac{T}{(K-1)\delta(1-\pi)}}} = \sqrt{1-\pi}\ \ \frac{U_{K+1}-\beta}{\sqrt{\frac{T}{K-1}}} \sim t_{K-1}.$$

To complete the proof, we have to show that $T = \sum_{k=1}^{K}(U_k - \bar{U})^2 = \sum_{k=1}^{K} U_k^2 - K\bar{U}^2$. Let B be the upper left $K \times K$ sub-matrix of A. Note that B is an orthonormal matrix since its lines are orthogonal to each other and have unit Euclidean norm. Thus

$$\begin{aligned}\sum_{k=1}^{K} U_k^2 &= U_-'(B'B)U_- = U_-'[H'H + K^{-1}\mathbf{1}_{K\times 1}^{\otimes 2}]U_- \\ &= (HU_-)'HU_- + K^{-1}U_-'\mathbf{1}_{K\times 1}\mathbf{1}_{K\times 1}'U_- = T + K\bar{U}^2.\end{aligned}$$

A.2 **Proof of Proposition 2**

We first need to introduce two objects.

- 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.
- Let
$$\breve{\mu}(S, n_1) = \frac{1}{|C(S, n_1)|(|S| - n_1)} \sum_{s\in C(S,n_1)} \sum_{i\in S\setminus s} \mathcal{L}(Z_s; Z_i).$$

Let

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

where $n = n_1 + n_2$ and $C(\{1,\ldots,n\}, n_1)$ represents, here, all the possible ways to choose n_1 integers from $\{1,\ldots,n\}$ for the purpose of constructing training sets. We note that ${}_{n_1}^{n_2}\hat{\mu}_\infty$ represents two different things.

- First ${}_{n_1}^{n_2}\hat{\mu}_\infty = \lim_{J\to\infty} {}_{n_1}^{n_2}\hat{\mu}_J = \lim_{J\to\infty} \frac{1}{J}\sum_{j=1}^{J} \hat{\mu}_j$. Indeed, what happens when J goes to infinity is that all possible errors (there are $\binom{n_1+n_2}{n_1}n_2$ different ways to choose a training set and a training example) appear with relative frequency $\frac{1}{\binom{n_1+n_2}{n_1}n_2}$. In other words, ${}_{n_1}^{n_2}\hat{\mu}_\infty$ is like ${}_{n_1}^{n_2}\hat{\mu}_J$ except that all $\binom{n_1+n_2}{n_1}$ possible training sets are chosen exactly once. Briefly, sampling infinitely often with replacement is equivalent to sampling exhaustively without replacement (i.e. a census). From (8) we have
$$\sigma_2 = \sigma_2(n_1, n_2) = \lim_{J\to\infty} Var[\,{}_{n_1}^{n_2}\hat{\mu}_J] = Var[\,{}_{n_1}^{n_2}\hat{\mu}_\infty].$$
- We also have ${}_{n_1}^{n_2}\hat{\mu}_\infty = E[\mathcal{L}(Z_{S_j}; Z_i)|Z_1^n]$, where the expectation is taken over the random index set S_j and $i \notin S_j$.

We show later (we keep the fun part for the end) that, for $0 < n_2' < n_2 = n - n_1$,

$$\check{\mu}(\{1,\dots,n\},n_1) = \frac{1}{|C(\{1,\dots,n\},n_1+n_2')|} \sum_{s\in C(\{1,\dots,n\},n_1+n_2')} \check{\mu}(s,n_1) \tag{20}$$

Obviously, the $\check{\mu}(s,n_1)$'s are identically distributed so that [15]

$$\sigma_2(n_1,n_2) = Var[\check{\mu}(\{1,\dots,n\},n_1)] \le Var[\check{\mu}(\{1,\dots,n_1+n_2'\},n_1)] = \sigma_2(n_1,n_2').$$

To complete the proof, we only need to show that identity (20) is true. We must first observe that choosing a random (training) set of size n_1 and a test example outside the training set can be performed in the following way.

- Choose $S \in C(\{1,\dots,n\},n_1+n_2')$ at random.
- Choose a training set $T \in C(S,n_1)$ at random and a test example $K \in S \setminus T$ at random.

Indeed, $\forall t \in C(\{1,\dots,n\},n_1)$ and $k \in \{1,\dots,n\}\setminus t$,

$$\begin{aligned}
P(T=t,K=k) &= \sum_{s\in C(\{1,\dots,n\},n_1+n_2')} P(T=t,K=k|S=s)P(S=s) \\
&= \sum_{s\in C(\{1,\dots,n\},n_1+n_2')} \frac{I[t\subset s, k\in s]}{\binom{n_1+n_2'}{n_1}n_2'} \frac{1}{|C(\{1,\dots,n\},n_1+n_2')|} \\
&= \frac{\binom{n_2-1}{n_2'-1}}{\binom{n_1+n_2'}{n_1}n_2'\binom{n}{n_1+n_2'}} = \frac{1}{\binom{n}{n_1}n_2}.
\end{aligned}$$

This being established, we finally have

$$\begin{aligned}
\check{\mu}(\{1,\dots,n\},n_1) &= E[\mathcal{L}(Z_T;Z_K)|Z_1^n] \\
&= \sum_{s\in C(\{1,\dots,n\},n_1+n_2')} E[\mathcal{L}(Z_T;Z_K)|Z_1^n, S=s]P(S=s) \\
&= \frac{1}{|C(\{1,\dots,n\},n_1+n_2')|} \sum_{s\in C(\{1,\dots,n\},n_1+n_2')} \check{\mu}(s,n_1).
\end{aligned}$$

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

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

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

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

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

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

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.

A.4 ${}_{n_1}\mu_A$ and ${}_{n_1}\mu_B$ for the Gaussian regression problem

We have (n_1+1) couples $Z_i = (X_i, Y_i)$ as described above (18). To obtain ${}_{n_1}\mu_A$, we first note that $Y_i \sim N(\beta_0 + \beta_1 \mu_X, \sigma^2_{Y|X} + \beta_1^2 \sigma_X^2)$, so that

$${}_{n_1}\mu_A = E[(\bar{Y} - Y_{n_1+1})^2] = Var[\bar{Y} - Y_{n_1+1}] = Var[\bar{Y}] + Var[Y_{n_1+1}] = \left(\frac{1}{n_1} + 1\right)(\sigma^2_{Y|X} + \beta_1^2\sigma_X^2),$$

where $\bar{Y} = n_1^{-1}\sum_{i=1}^{n_1} Y_i$.

Things are a little more complicated for ${}_{n_1}\mu_B$. They go as follows.

$$\begin{aligned} {}_{n_1}\mu_B &= E[(Y_{n_1+1} - \hat{\beta}_0 - \hat{\beta}_1 X_{n_1+1})^2] = Var[Y_{n_1+1} - \hat{\beta}_0 - \hat{\beta}_1 X_{n_1+1}] \\ &= E[Var[Y_{n_1+1} - \hat{\beta}_0 - \hat{\beta}_1 X_{n_1+1} | X_{n_1+1}]] + Var[E[Y_{n_1+1} - \hat{\beta}_0 - \hat{\beta}_1 X_{n_1+1} | X_{n_1+1}]] \\ &= E[Var[Y_{n_1+1} - \hat{\beta}_0 - \hat{\beta}_1 X_{n_1+1} | X_{n_1+1}]] + Var[0] \\ &= E\left[\sigma^2_{Y|X} + (1, X_{n_1+1})\, Var[\hat{\beta}](1, X_{n_1+1})'\right] \end{aligned}$$

with $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)'$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the intercept and the slope of the ordinary least squares regression of Y on X performed on the training set $Z_1^{n_1} = \{Z_1, \ldots, Z_{n_1}\}$. Now

$$\begin{aligned} Var[\hat{\beta}] &= E[Var[\hat{\beta}|Z_1^{n_1}]] + Var[E[\hat{\beta}|Z_1^{n_1}]] = E[\sigma^2_{Y|X}(X'X)^{-1}] + Var[(\beta_0, \beta_1)'] \\ &= \sigma^2_{Y|X} E\left[\begin{array}{cc} \frac{1}{n_1} + \frac{\bar{X}^2}{T} & \frac{-\bar{X}}{T} \\ \frac{-\bar{X}}{T} & \frac{1}{T} \end{array}\right], \end{aligned}$$

where X denotes the usual design matrix, $\bar{X} = n_1^{-1}\sum_{i=1}^{n_1} X_i$ and $T = \sum_{i=1}^{n_1}(X_i - \bar{X})^2$. Since the X_i's are independent and identically distributed normal variates, then we know (from Appendix A.1 or textbooks) that $\bar{X}$ and T are independent with $\frac{T}{\sigma_X^2} \sim \chi_{n_1-1}$ so that

$E[\frac{1}{T}] = \frac{1}{(n_1-3)\sigma_X^2}$. Putting all this together leads us to

$$\begin{aligned} Var[\hat{\beta}] &= \sigma^2_{Y|X} \left[\begin{array}{cc} \frac{1}{n_1} + \frac{\mu_X^2 + \frac{\sigma_X^2}{n_1}}{(n_1-3)\sigma_X^2} & \frac{-\mu_X}{(n_1-3)\sigma_X^2} \\ \frac{-\mu_X}{(n_1-3)\sigma_X^2} & \frac{1}{(n_1-3)\sigma_X^2} \end{array} \right] \\ &= \frac{\sigma^2_{Y|X}}{(n_1-3)\sigma_X^2} \left(\begin{array}{c} -\mu_X \\ 1 \end{array} \right)^{\otimes 2} + \frac{\sigma^2_{Y|X}}{n_1} \left[\begin{array}{cc} \frac{n_1-2}{n_1-3} & 0 \\ 0 & 0 \end{array} \right] \end{aligned}$$

So we have

$$\begin{aligned} {}_{n_1}\mu_B &= \sigma^2_{Y|X} + E[(1, X_{n_1+1})\, Var[\hat{\beta}](1, X_{n_1+1})'] \\ &= \sigma^2_{Y|X} + \frac{\sigma^2_{Y|X}}{(n_1-3)\sigma_X^2} E[(X_{n_1+1} - \mu_X)^2] + \frac{\sigma^2_{Y|X}}{n_1} \frac{n_1-2}{n_1-3} \\ &= \frac{\sigma^2_{Y|X}}{n_1(n_1-3)} [n_1(n_1-3) + n_1 + (n_1-2)]. \end{aligned}$$

Liste des publications au CIRANO *

Cahiers CIRANO / *CIRANO Papers* (ISSN 1198-8169)

99c-1 Les Expos, l'OSM, les universités, les hôpitaux : Le coût d'un déficit de 400 000 emplois au Québec — Expos, Montréal Symphony Orchestra, Universities, Hospitals: The Cost of a 400,000-Job Shortfall in Québec / Marcel Boyer

96c-1 Peut-on créer des emplois en réglementant le temps de travail ? / Robert Lacroix

95c-2 Anomalies de marché et sélection des titres au Canada / Richard Guay, Jean-François L'Her et Jean-Marc Suret

95c-1 La réglementation incitative / Marcel Boyer

94c-3 L'importance relative des gouvernements : causes, conséquences et organisations alternative / Claude Montmarquette

94c-2 Commercial Bankruptcy and Financial Reorganization in Canada / Jocelyn Martel

94c-1 Faire ou faire faire : La perspective de l'économie des organisations / Michel Patry

Série Scientifique / *Scientific Series* (ISSN 1198-8177)

99s-24 Mobility and Cooperation: On the Run / Karl-Martin Ehrhart et Claudia Keser

99s-23 Input Price Discrimination, Access Pricing, and Bypass / Ngo Van Long et Antoine Soubeyran

99s-22 Existence and Uniqueness of Cournot Equilibrium: A Contraction Mapping Approach / Ngo Van Long et Antoine Soubeyran

99s-21 Sources of Productivity Growth: Technology, Terms of Trade, and Preference Shifts / Thijs ten Raa et Pierre Mohnen

99s-20 Remarks on Environmental Regulation, Firm Behavior and Innovation / Bernard Sinclair-Desgagné

99s-19 Subvention gouvernementale et partage du travail : Une analyse économique – II / Paul Lanoie et François Raymond

99s-18 Subvention gouvernementale et partage du travail : Une analyse économique – I / Paul Lanoie et Ali Béjaoui

99s-17 Content Horizons for Forecasts of Economic Time Series / John W. Galbraith

99s-16 Modelling the Role of Organizational Justice: Effects on Satisfaction and Unionization Propensity of Canadian Managers / Michel Tremblay et Patrice Roussel

99s-15 Pricing Discretely Monitored Barrier Options by a Markov Chain / Jin-Chuan Duan, Evan Dudley, Geneviève Gauthier et Jean-Guy Simonato

99s-14 Shame and Guilt in Lancashire: Enforcing Piece-Rate Contracts / Michael Huberman

99s-13 Cost Manipulation Games in Oligopoly, with Costs of Manipulations / Ngo Van Long et Antoine Soubeyran

* Vous pouvez consulter la liste complète des publications du CIRANO et les publications elles-mêmes sur notre site World Wide Web à l'adresse suivante : **http://www.cirano.umontreal.ca/publication/documents.html**