Biostatistics (2019), **0**, 0, *pp.* 1–16
doi:10.1093/biostatistics/supplementary

Better-Than-Chance Classification for Signal Detection– Supplementary Material

Jonathan D. Rosenblatt*

Department of IE&M and Zlotowsky Center for Neuroscience, Ben Gurion University of the Negev, Israel.

Yuval Benjamini

Department of Statistics, Hebrew University, Israel

Roee Gilron

Movement Disorders and Neuromodulation Center, University of California, San Francisco.

Roy Mukamel

School of Psychological Science Tel Aviv University, Israel.

Jelle Goeman

Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, The Netherlands.

1. Large Sample

We have focused on the him-dim–small-sample setup because it is appropriate for many problems in neuroimaging and genetics. To show that our conclusions are not due to the *small-sample*, but rather, to the *high-dim*, we scale our basic setup ten-fold. Fixing p/n, we simulate with $p = 230$ and $n = 400$. The results, reported in Figure 1, are qualitatively similar to the high-dim–small-

*johnros@bgu.ac.il

sample in the main text. In particular with respect to the dominance of two-group tests.

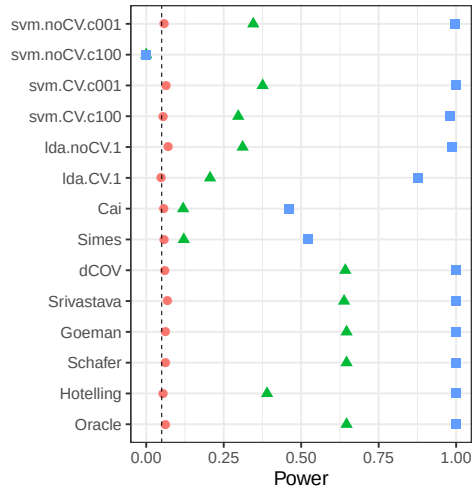


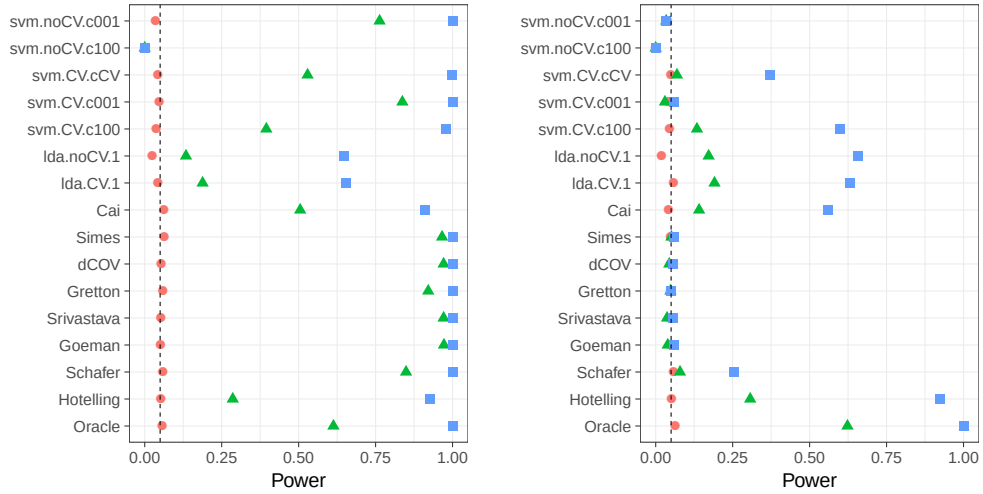
Fig. 1: **Large sample**: The basic simulation setup scaled ten-fold: $n = 400; p = 230$. $\frac{n}{2}\|\mu\|^2_\Sigma$ was set to 0 (red circle), 100 (green triangle), 400 (blue square) for comparable power to other simulations.

2. Departure From Sphericity

In the main text we have departed from the sphericity assumption by allowing Σ to be an $AR(1)$ covariance. We now try other covariance structures: a long-memory Brownian motion correlation, and an arbitrary (random) covariance structure. As seen in Figures 2 and 3, the findings in the main test hold also for the "long-memory", and "arbitrary" correlation structures. In particular: two-group tests dominate accuracy tests, and signal in the low PCs of the noise is masked.

3. Departure From Shift Alternatives

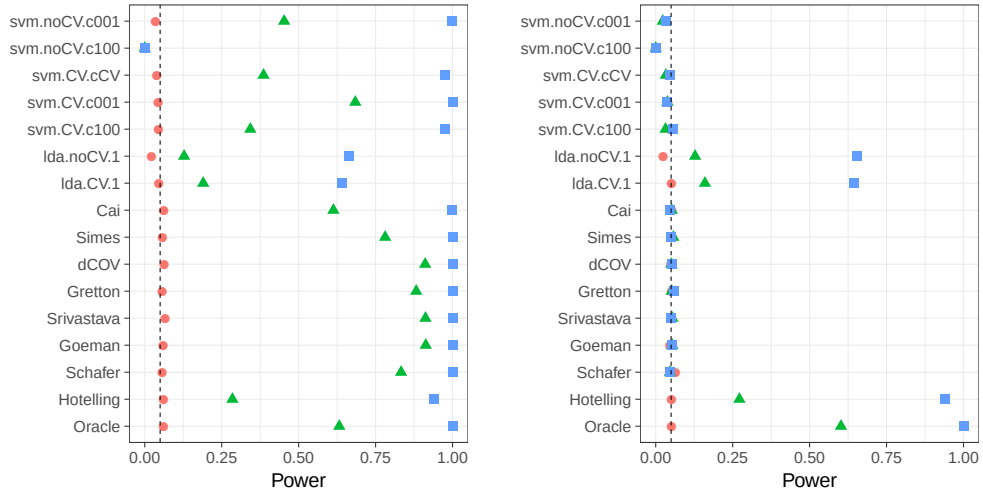
Shift alternatives are the most common signal model in the univariate statistical literature. They are also very common in the multivariate literature, as they are implied by Fisher's LDA problem setup, known as *Gaussian Bayes* in the machine learning literature. On the other hand, effects may manifest themselves in many ways, not necessarily in location. We now verify our claims in



(a) Signal in direction of highest variance PC of Σ.

(b) Signal in direction of lowest lowest variance PC of Σ.

Fig. 2: Long-memory Brownian motion correlation: $\Sigma = D^{-1}RD^{-1}$ where D is diagonal with $D_{jj} = \sqrt{R_{jj}}$, and $R_{k,l} = \min\{k, l\}$.



(a) Signal in direction of highest variance PC of Σ.

(b) Signal in direction of lowest variance PC of Σ.

Fig. 3: Arbitrary Correlation. $\Sigma = D^{-1}RD^{-1}$ where D is diagonal with $D_{jj} = \sqrt{R_{jj}}$, and $R = A'A$ where A is a Gaussian $p \times p$ random matrix with independent $\mathcal{N}(0, 1)$ entries.

models which are not "pure shifts". These include logistic regression, and a mixture class.

3.1 *Logistic Regression*

In Figure 4 we report the usual power simulation, when generating from a logistic regression setup with both main effects, and second order interactions. This setup is also reported in the main text.

Formally, the logistic assumption implies that $P(y=1|x) = \exp(\eta)/[1+\exp(\eta)]$. Main-effects and second order interactions imply that $\eta = \beta_0 + x'\beta + x'Bx$, for some p-vector β, and symmetric $p \times p$ matrix B. We also assume $x \sim \mathcal{N}(0, I_{p\times p})$. We perform the various tests in the original space, x, but also in the 276 dimensional space of main effects and second order interactions:

$$\tilde{x} := \Phi(x) = (x_1, \ldots, x_j, \ldots, x_p, \ldots, x_1x_1, \ldots x_jx_{j'}, \ldots, x_px_p).$$

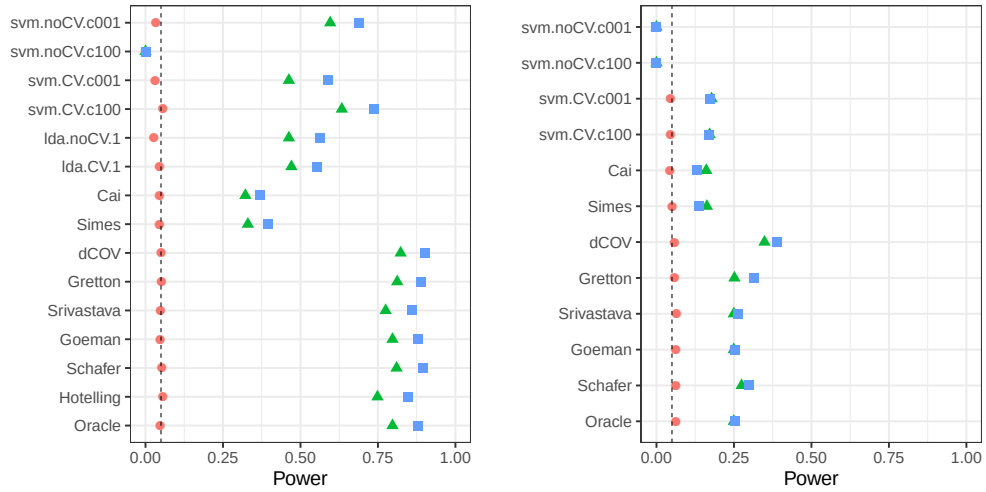
From Figure 4 we learn that two-group tests dominate accuracy tests also in the logistic setup. Was this to be expected, given our conclusions on multivariate shifts? In the logistic setup x_1 cannot be a shifted version of x_0, but their means certainly differ. The larger the main effects, the larger the difference in means. This suggests that a main-effect-only model ($B = 0$) will be roughly similar to a shift class, and a second-order-interactions-only model may be very different than a shift class. To isolate these two cases, we simulate a main-effects-only setup (Fig.5), and an interaction-only setup (Fig.6).

The main-effect-only case in Figure 5 is not very surprising: the groups differ in their first moment, so that our conclusions are almost identical to the "pure shift" examples in the main text. Analyzing the data in the augmented 276 dimensional space decreases power, since the problem increases in dimension, and many more parameters need to be estimated.

It may also be surprising that in the original space, x, signal detection is easier in the absence of second order interactions. Put differently, all tests have more power when $B = 0$ than when $B \neq 0$ (Fig.4a vs. Fig.5a). Why does more signal/effects reduce power? The signal added is not in the span of the x space, so for detection algorithms operating in x (not $\tilde{x}$), this is actually part

of the noise. This intuition is confirmed when analyzing in $\tilde{x}$: it is indeed easier to detect signal with interactions, than main effects only (Fig.4b vs. Fig.5b).

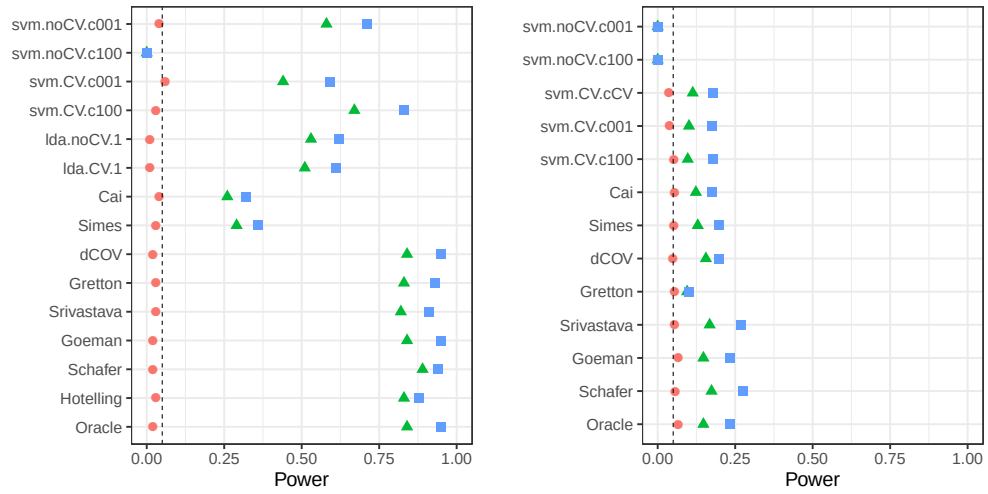
The second-order-interactions-only case in Figure 6 is of particular interest. From Figure 6a we see that in this setup x_1 is roughly a scaled version of x_0. In the context of machine-learning, and in particular learning in kernel spaces, this is sometimes known as a *ring problem*. When all signal is in the scatter of the distribution, we expect that tests for shifts would have poor performance. This intuition is confirmed: in the original space x_0 and x_1 have the same location and shift detectors have no power at all (Fig.6b). In the augmented space, x_0 and x_1 differ in location so that shift detectors are re-powered (Fig.6c). Most importantly for practitioners, the high-dim GOF tests detect the change in scatter in all these cases. Why the GOF tests perform so much better in the augmented space may be the topic of future research.



(a) Data analyzed in the original space (x).

(b) Data analyzed in augmented interactions space ($\tilde{x}$).

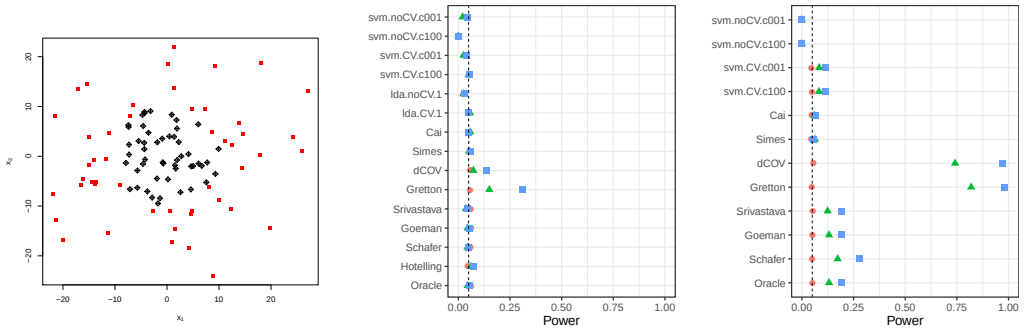
Fig. 4: **Logistic regression. Main effects and interactions.** Data generated via $y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1+\exp(\eta)]; \eta = \beta_0 + x'\beta + x'Bx$ where β is a scaled vector of ones, B a scaled identity matrix, and β_0 set so that the median $\eta \approx 0$. Finally, $x \sim \mathcal{N}(0, I_{p\times p})$.



(a) Data analyzed in the original space (x).

(b) Data analyzed in augmented interactions space ($\tilde{x}$).

Fig. 5: **Logistic Regression. Main effects only.** Data generated via $y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1+\exp(\eta)]; \eta = \beta_0 + x'\beta$ where β is a scaled vector of ones, and β_0 set so that the median $\eta \approx 0$. Finally, $x \sim \mathcal{N}(0, I_{p\times p})$.



(a) Illustration with $p = 2$. $y = 1$ in red squares. $y = 0$ in black diamonds.

(b) Detection power in **original** space.

(c) Detection power in **augmented** space.

Fig. 6: **Logistic regression. Second order interactions only.** Data generated via $y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1+\exp(\eta)]; \eta = \beta_0 + x'Bx$ where B is a scaled identity matrix. Finally, $x \sim \mathcal{N}(0, I_{p\times p})$.

3.2 *Mixture Class*

Another example where x_1 is not a shifted version of x_0 is a mixture class. Golland and Fischl [2003] and Golland et al. [2005] study accuracy-tests using simulation, neuroimaging data, genetic data, and analytically. The finite Vapnik–Chervonenkis dimension requirement [Golland et al.,

2005, Sec 4.3] implies a the problem is low dimensional and prevents the permutation p-value from (asymptotically) concentrating near 1. They find that the power increases with the size of the test set. This is seen in Fig.4 of Golland et al. [2005], where the size of the test-set, K, governs the discretization. We attribute this to the reduced discretization of the accuracy statistic.

When discussing the power of the resubstitution accuracy, Golland et al. [2005] simulate power by sampling from a Gaussian mixture family of models. Under their model (with some abuse of notation)

$$x_1 \sim \pi\mathcal{N}(\mu_1, I) + (1-\pi)\mathcal{N}(\mu_2, I),$$
$$x_0 \sim (1-\pi)\mathcal{N}(\mu_1, I) + \pi\mathcal{N}(\mu_2, I).$$

Varying π interpolates between the null distribution ($\pi = 0.5$) and a shift model ($\pi = 0$). We now perform the same simulation as Golland et al. [2005], but in the same dimensionality of our previous simulations. We re-parameterize so that $\pi = 0$ corresponds to the null model:

$$\begin{aligned} x_1 &\sim (1/2-\pi)\mathcal{N}(\mu_1, I) + (1/2+\pi)\mathcal{N}(\mu_2, I), \\ x_0 &\sim (1/2+\pi)\mathcal{N}(\mu_1, I) + (1/2-\pi)\mathcal{N}(\mu_2, I). \end{aligned} \tag{3.1}$$

From Figure 7, we see that for the mixture class of Golland et al. [2005] location tests are still preferred over accuracy-tests.

4. Sparse Alternatives

In our set of simulations we discussed "dense" alternatives. Dense alternatives are motivated by neuroimaging where most brain locations in a region carry signal. In a genetic application, a sparse alternative may be more plausible. Figure 8 reports power when μ is sparse. As usual, two-group tests dominate accuracy-tests, only this time, the winners are not the T^2 type statistics, but rather, the tests for sparse shifts (*Cai*, *Simes*).

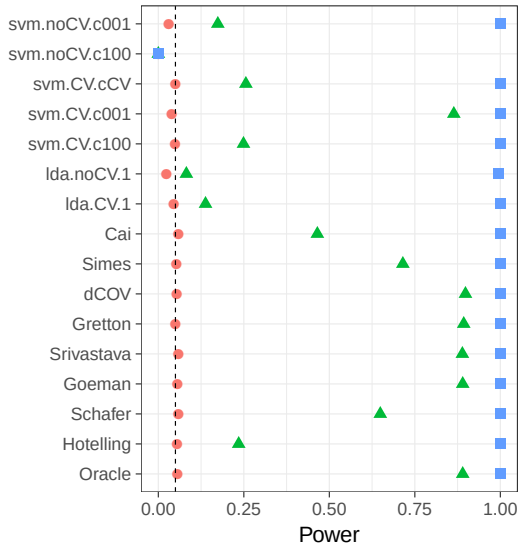


Fig. 7: **Mixture Alternatives.** $\mathbf{x}_i$ is distributed as in Eq.(3.1). μ is a p-vector with $3/\sqrt{p}$ in all coordinates. The effect, π, is color and shape coded and varies over 0 (red circle), 1/4 (green triangle) and 1/2 (blue square).

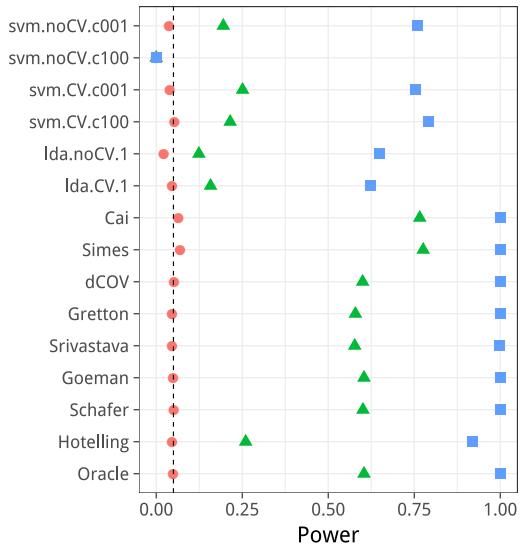


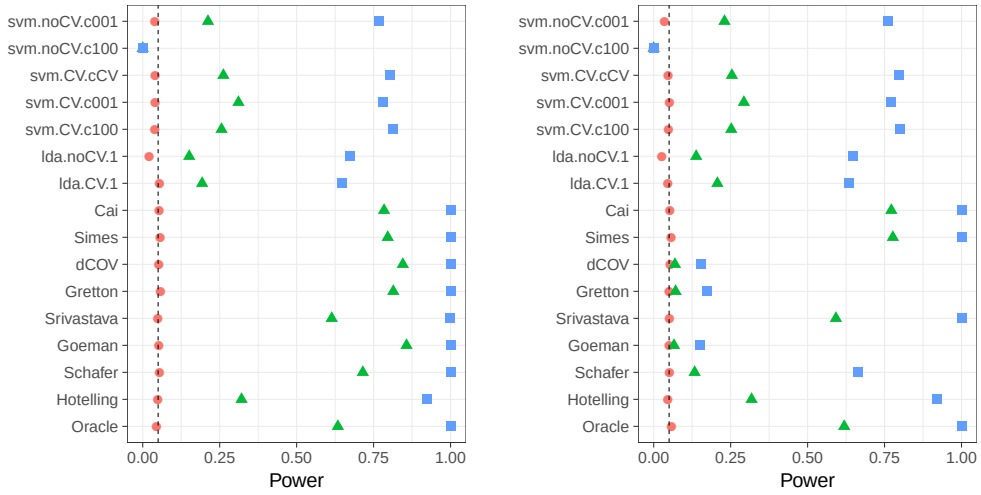
Fig. 8: Sparse μ.

5. Departure from Homoskedasticity and Scalar Invariance

In our simulations variables have unit variance. Practitioners are already accustomed to z-score features before learning a regularized predictor (e.g. ridge regression) so this is not an unrealistic

setup. Implicit z-scoring is sometime an integral part of a test statistic. This is known as *scalar invariance*. The *Srivastava* statistic, for instance, is scalar invariant. It can be (roughly) thought of as the l_2 norm of the p-vector of coordinate-wise t-statistics. The *Goeman* statistic, for instance, is not scalar invariant. It can be (roughly) thought of as the l_2 norm of the p-vector of variable-wise mean differences. Under heteroskedasticity, the *Goeman* statistic will give less importance to signal in the high-variance directions than signal in the low-variance directions. *Srivastava* will give all coordinates the same importance.

In Figure 9a we can see the difference between the *Goeman* statistic, and the scalar-invariant *Srivastava* statistic. We also see that two-group tests dominate accuracy-tests also in the heteroskedastic case.



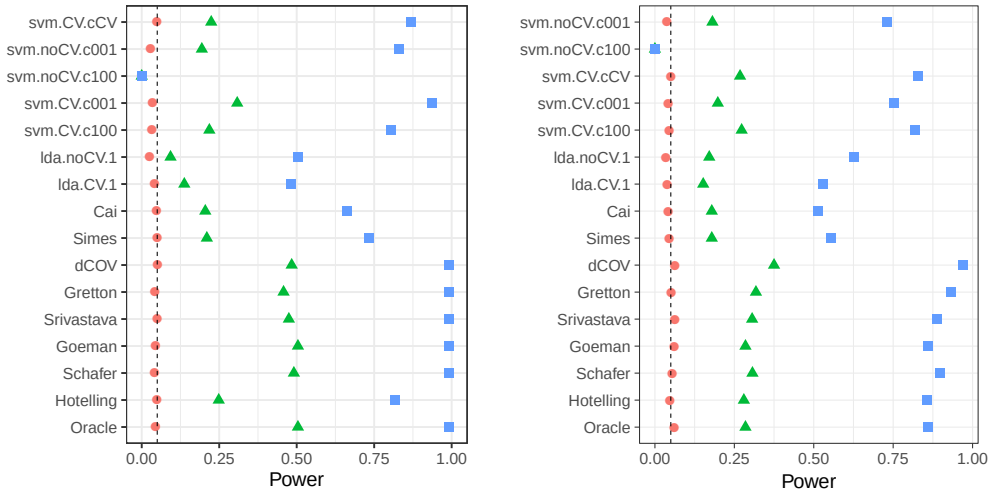
(a) μ in the high variance PC of Σ. **(b)** μ in the low variance PC of Σ.

Fig. 9: Heteroskedasticity: Σ is diagonal with $\Sigma_{jj} = j$.

6. Departure from Gaussianity

Hotelling's T^2 is a generalized likelihood ratio test in the Gaussian shift class. This Neyman-Pearson Lemma (NPL) type reasoning that favors two-group location-tests over accuracy-tests

in our simulations may fail when the data is not Gaussian. To verify our conclusions in the non-Gaussian case, we replaced the multivariate Gaussian distribution of η with a heavy-tailed multivariate-t distribution with 3 degrees of freedom. In this heavytailed setup, the dominance of the two-group tests was preserved, even if less evident than in the light-tailed Gaussian case.



(a) Heavytailed. η_i is p-variate t distributed, with $df = 3$.

(b) Heavytailed. η_i is p-variate t distributed, with $df = 3$.

Fig. 10: Short memory, AR(1) correlation. $\Sigma_{k,l} = \rho^{|k-l|}; \rho = 0.6$.

7. Designing The Cross Validation

We will now address the design of the cross validation scheme, while bearing in mind that unlike the typical supervised learning setup, we are not interested in an unbiased estimate of $\mathcal{E}_{\mathcal{A}}$, but rather in the detection of its departure from chance level.

Cross-validate or not? For the purpose of statistical testing, bias in $\hat{\mathcal{E}}_{\mathcal{A}}$ is not a problem, since it does not inflate the error rates of the accuracy-tests. The underlying intuition is that if the same bias is introduced in all permutations, it will not affect the properties of the permutation test. We will thus be considering both unbiased cross-validated accuracies, and biased resubstitution

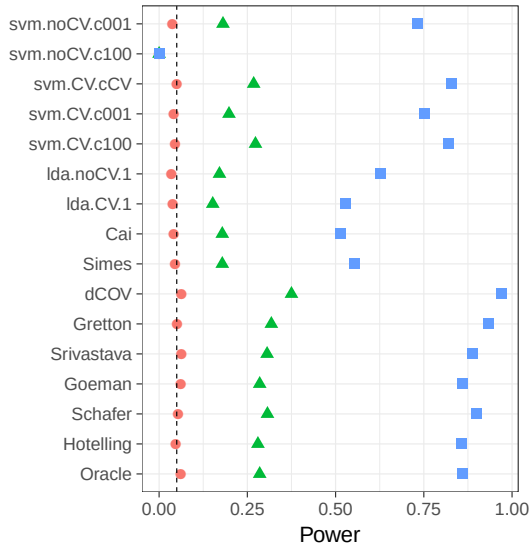


Fig. 11: **Heavytailed.** η_i is p-variate t distributed, with $df = 3$.

accuracies.

Balanced folding The standard practice in V-fold CV is to constrain the data folds to be balanced, a.k.a. stratified [Ojala and Garriga, 2010, for example]. This means that each fold has the same number of samples from each class. We will report results only with balanced folding, mostly because we will conclude that V-fold CV should not be used for our detection problem.

Refolding In V-fold CV, *folding* the data means assigning each observation to one of the V data folds. The standard practice in neuroimaging is to permute labels and refold the data after each permutation. This is done because permuting labels will unbalance the original balanced folding. We will adhere to this practice due to its popularity, even though it is computationally more efficient to permute features instead of labels [e.g. Golland et al., 2005] .

How many folds Different authors suggest different rules for the number of folds. We fix the number of folds to $V = 4$. A different number of folds does not change our conclusions. We do not discuss the effect of V because we will ultimately show that V-fold CV is dominated by other

cross-validation procedures, and thus, never recommended.

8. Tie Breaking

Discrete test statistics lose power by not exhausting the permissible false positive rate. A common remedy is a *randomized test* with tie-breaking, in which the rejection of the null is decided at random in a manner that exhausts the false positive rate. Formally, denoting by $\mathcal{T}$ the observed test statistic, by $\mathcal{T}_\pi$, its value after under permutation π, and by $\mathbb{P}\{A\}$ the proportion of permutations satisfying A then the randomized version of our tests imply that if the permutation p-value, $\mathbb{P}\{\mathcal{T}_\pi \geqslant \mathcal{T}\}$, is greater than α then we reject the null with probability

$$max\left\{\frac{\alpha - \mathbb{P}\{\mathcal{T}_\pi > \mathcal{T}\}}{\mathbb{P}\{\mathcal{T}_\pi = \mathcal{T}\}}, 0\right\}.$$

Figure 12 reports the basic simulation setup while allowing for random tie breaking. It demonstrates that the power disadvantage of accuracy-tests cannot be remedied by random tie breaking.

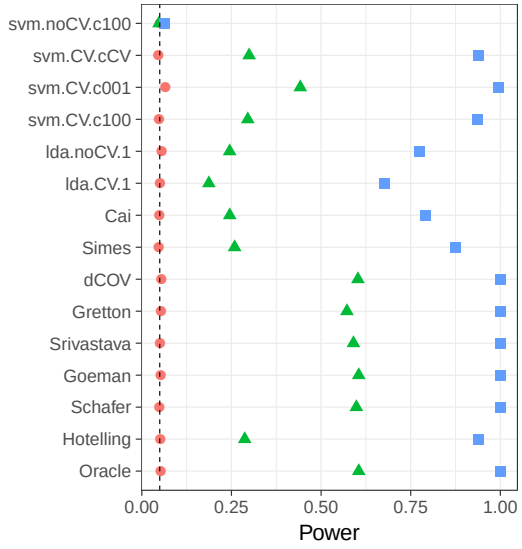


Fig. 12: **Tie breaking**: The basic simulation setup with random tie breaking.

9. Fixed SNR

For a fair comparison between simulations, in particular between those with different Σ, we needed to fix the difficulty of the problem. We fix the Kullback–Leibler Divergence between distributions of sample means. Formally, the Kullback–Leibler Divergence between two Gaussian populations is given by

$$KL[x_1, x_0] = \frac{1}{2}\left(\log\frac{\det\Sigma_0}{\det\Sigma_1} - p + Tr(\Sigma_0^{-1}\Sigma_1) + (\mu_0-\mu_1)'\Sigma_0^{-1}(\mu_0-\mu_1)\right), \tag{9.2}$$

where $x_y \sim \mathcal{N}(\mu_y, \Sigma_y)$. In the case of the sample means of two shifted groups of size n, then

$$KL[\bar{x}_1, \bar{x}_0] = \frac{n}{2}\mu'\Sigma^{-1}\mu = \frac{n}{2}\|\mu\|_\Sigma^2, \tag{9.3}$$

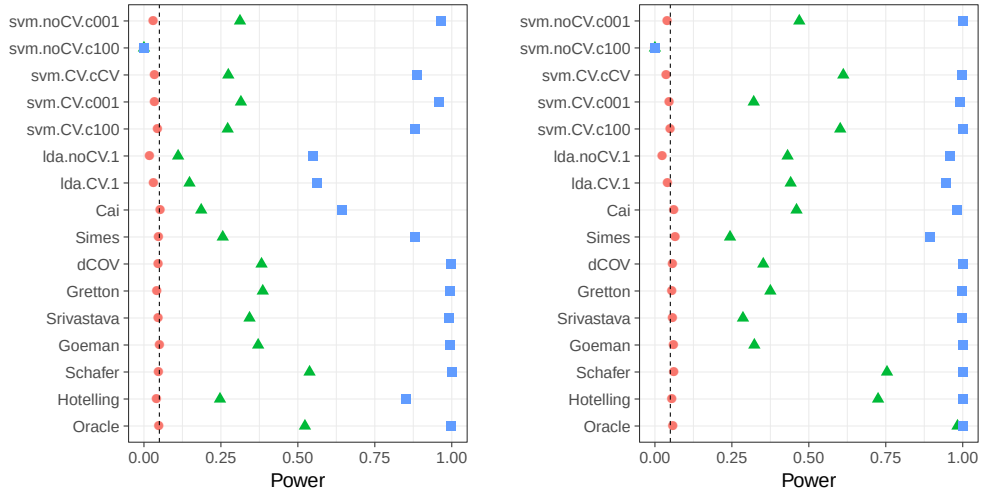
where $\mu := \mu_1 - \mu_0$.

In most of our simulations we fixed $n\|\mu\|_\Sigma^2$. The logistic regression setup is an exception because the signal is not a shift. We did set effect sizes so that power in the logistic regression is comparable to power in the other examples.

Fixing $n\|\mu\|_\Sigma^2$ implies that the Euclidean norm of μ varies with Σ, with the sample size, and with the direction of the signal. An initial intuition may suggest that detecting signal in the low variance PCs is easier than in the high variance PCs. This is true when fixing $\|\mu\|_2$, but not when fixing $\|\mu\|_\Sigma$.

For completeness, Figure 13 reports the power analysis under $AR(1)$ correlations, but with $\|\mu\|_2$ fixed. We compare the power of a shift in the direction of some high variance PC (Figure 13a), versus a shift in the direction of a low variance PC (Figure 13b). The intuition that it is easier to detect signal in the low variance directions is confirmed.

Other authors have also observed the need for fixing the SNR for a fair comparison between tests. In Ramdas et al. [2015], authors prefer to use sparse alternatives. With sparse alternatives, the difficulty of the problem is governed by the sparsity of the signal and not only the dimension of the data. In Chen et al. [2010], authors fix $\|\mu\|_2^2/\|\Sigma\|_{Frob}^2$ where $\|\Sigma\|_{Frob}^2 = \mathrm{Tr}(\Sigma'\Sigma)$ is the



(a) μ in PC7 of Σ. **(b)** μ in PC15 of Σ.

Fig. 13: Short memory, AR(1) correlation. $\|\mu\|_2$ fixed.

Frobenius matrix norm. Clearly, $\|\mu\|_2^2/\|\Sigma\|_{Frob}^2$ is invariant to the direction of the signal with respect to the noise. For this reason, we prefer fixing $\|\mu\|_\Sigma$.

10. R Packages

Many R packages were used in this project. The following deserve many thanks: Genz et al. [2018], Weston and Microsoft [2017], Curran [2018], Eddelbuettel et al. [2017], Ramey [2017], Friedman et al. [2010], Goeman and Oosting [2018], Karatzoglou et al. [2004], Rizzo and Szekely [2018], Cao et al. [2018]

References

M. Cao, T. He, and W. Zhou. *HDtest: High Dimensional Hypothesis Testing for Mean Vectors, Covariance Matrices, and White Noise of Vector Time Series*, 2018. URL `https://CRAN.R-project.org/package=HDtest`. R package version 2.1.

S. X. Chen, Y.-L. Qin, et al. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835, 2010.

J. Curran. *Hotelling: Hotelling's T2 Test and Variants*, 2018. URL `https://CRAN.R-project.org/package=Hotelling`. R package version 1.0-5.

D. Eddelbuettel, B. Evans, M. Birdgeneau, H. Bengtsson, and S. Wenchel. *RPushbullet: R Interface to the Pushbullet Messaging Service*, 2017. URL `https://CRAN.R-project.org/package=RPushbullet`. R package version 0.3.1.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2018. URL `https://CRAN.R-project.org/package=mvtnorm`. R package version 1.0-8.

J. J. Goeman and J. Oosting. *Globaltest: testing association of a group of genes with a clinical variable*, 2018. R package version 5.36.0.

P. Golland and B. Fischl. Permutation tests for classification: towards statistical significance in image-based studies. In *IPMI*, volume 3, pages 330–341. Springer, 2003.

P. Golland, F. Liang, S. Mukherjee, and D. Panchenko. Permutation Tests for Classification. In P. Auer and R. Meir, editors, *Learning Theory*, number 3559 in Lecture Notes in Computer

Science, pages 501–515. Springer Berlin Heidelberg, June 2005. ISBN 978-3-540-26556-6 978-3-540-31892-7. doi: 10.1007/11503415_34.

A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004. URL `http://www.jstatsoft.org/v11/i09/`.

M. Ojala and G. C. Garriga. Permutation Tests for Studying Classifier Performance. *Journal of Machine Learning Research*, 11(Jun):1833–1863, 2010. ISSN ISSN 1533-7928.

A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. A. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, pages 3571–3577, 2015.

J. Ramey. *Sparse and Regularized Discriminant Analysis*, 2017. URL `https://github.com/ramhiser/sparsediscrim`. R package version 0.2.5.

M. Rizzo and G. Szekely. *energy: E-Statistics: Multivariate Inference via the Energy of Data*, 2018. URL `https://CRAN.R-project.org/package=energy`. R package version 1.7-5.

S. Weston and Microsoft. *foreach: Provides Foreach Looping Construct for R*, 2017. URL `https://CRAN.R-project.org/package=foreach`. R package version 1.4.4.

[*xxx*]