Biostatistics (2019), **0**, 0, *pp.* 1–9
doi:10.1093/biostatistics/supplementary

Better-Than-Chance Classification for Signal Detection– Supplementary

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

The results, reported in Figure 1, are qualitatively similar to the high-dim–small-sample of Figure **??**.

2. Departure From Sphericity

3. Departure From Shift Alternatives

3.1 *Logistic Regression and Interactions*

In Figure 4 we report the usual power simulation, when generating from the logistic setup, with both main effects and second order interactions. We emphasize that all tests are performed in the original x space, i.e., ignore the presence of interactions.

Formally, the logistic assumption implies that $P(Y = 1|x) = \exp(\eta)/[1+\exp(\eta)]$. Main-effects and second order interaction imply that $\eta = x'\beta + x'Bx$, for some p-vector β, and symmetric matrix B. We also assume $X \sim \mathcal{N}(0, I_{p\times p})$.

From Figure 4 we see that in the logistic setup, with interactions, two-group tests still dominate.

The logistic assumption differs from our original setup in that the logistic assumption states $Y|x$, instead of $X|y$, as in Fisher's LDA setup. The logistic assumption implies that X_0 is no longer a shifted versions of X_1, even in the presence of main effects alone. I.e., when B is a $p \times p$ matrix of zeroes. While not a "pure shift", we expect that lacking interactions, the logistic model is very close to a shift. This is verified in Figure 5.

*johnros@bgu.ac.il

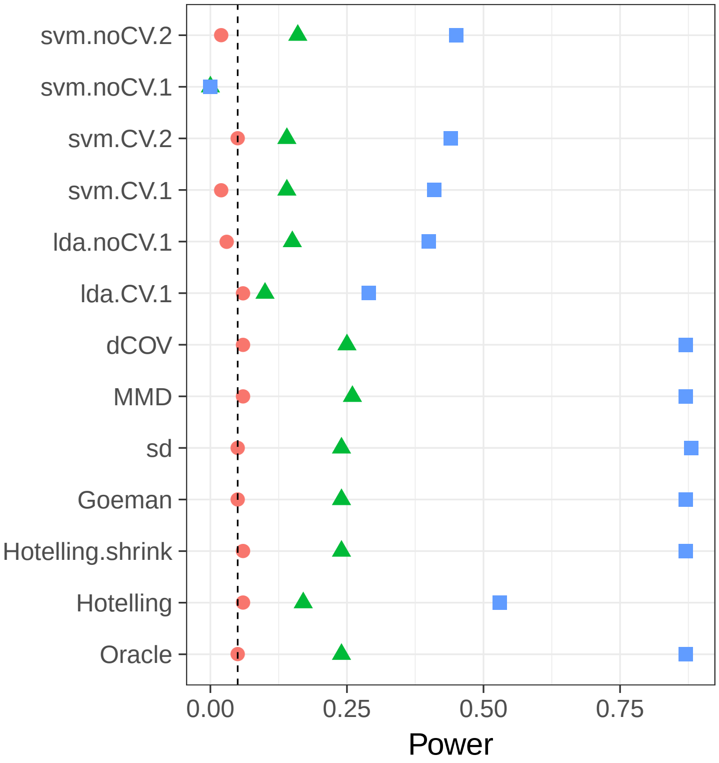


Fig. 1: The same as Figure **??** with $n = 4,000; p = 2,300$.

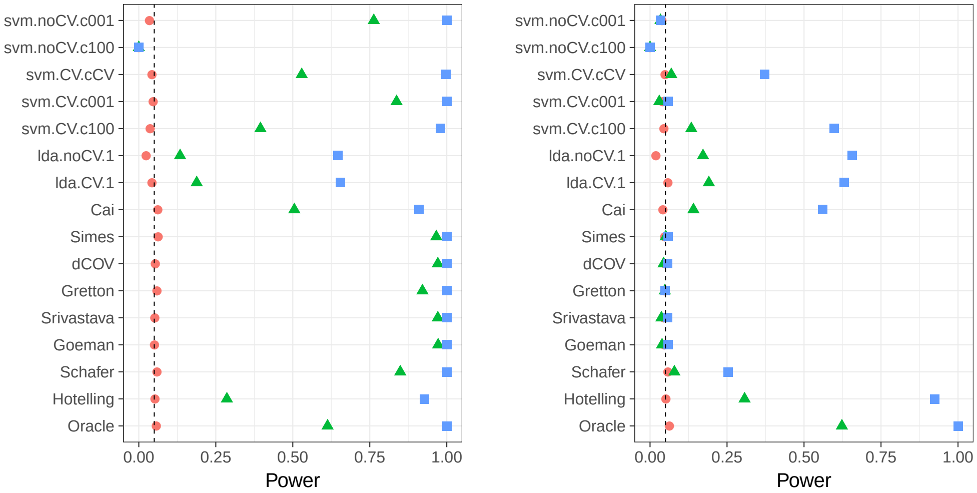
It is possible to use the logistic setup to generate data with no shift at all. For instance, if $\eta = \beta_0 + x'x$. In this case, X_1 is a rescaled version X_0, depicted for $p = 2$ in Figure 6a. This example is typically encountered in the machine learning literature, to motivate learning with kernels.

Interactions are understood as statements on $Y|x$. In particular, that the gradient of $P(Y = 1|x)$ is a function of multiple coordinates. In this sense, the shift model includes interactions.

3.2 *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.

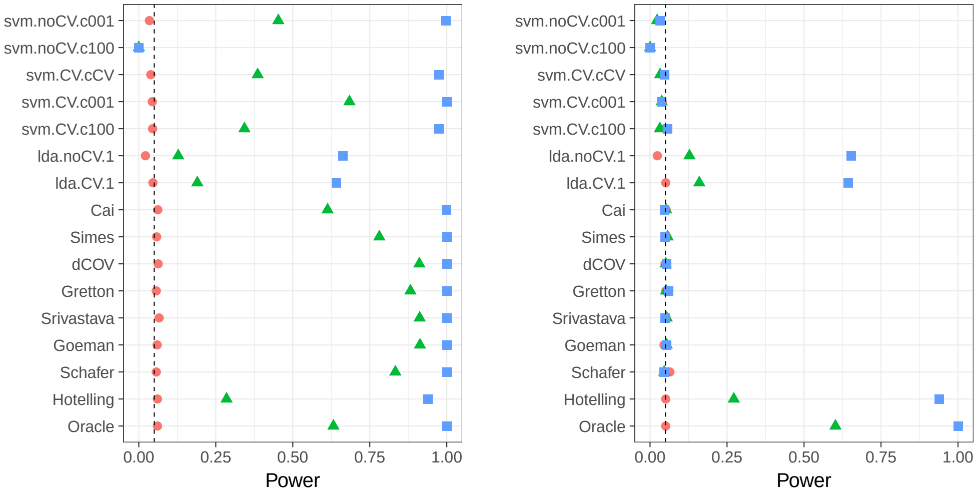
When discussing the power of the resubstitution accuracy, Golland et al. [2005] simulate power by sampling from a Gaussian mixture family of models, and not from a location family as our



(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.

own simulations. Under their model (with some abuse of notation)

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

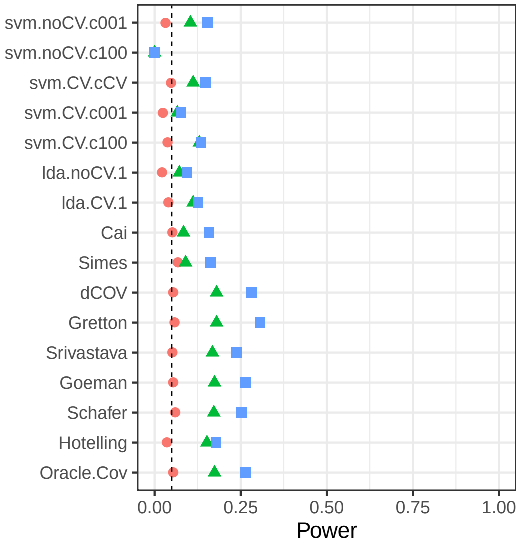


Fig. 4: **Logistic Regression. Main effects and interactions.** Data generated via $Y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1 + \exp(\eta)]; \eta = x'\beta + x'Bx; X \sim \mathcal{N}(0, I_{p\times p})$.

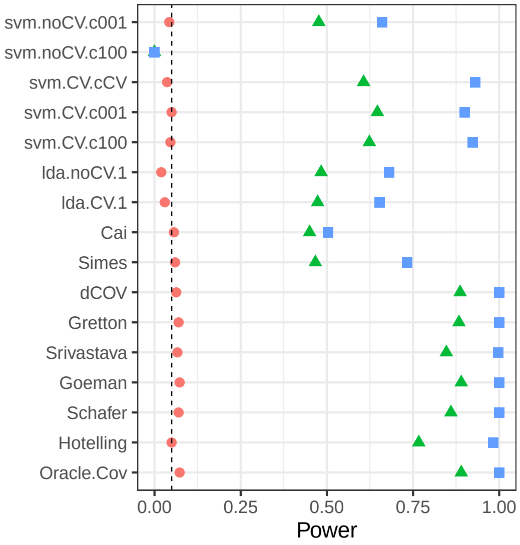
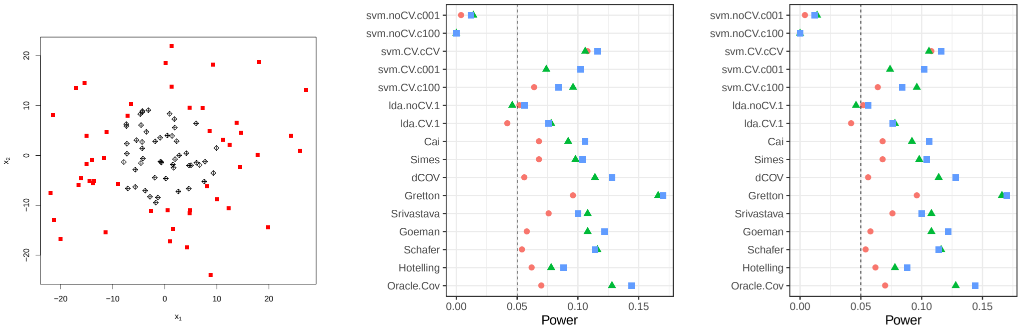


Fig. 5: **Logistic Regression. Main effects only.** Data generated via $Y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1+\exp(\eta)]; \eta = x'\beta; X \sim \mathcal{N}(0, I_{p\times p})$.

Varying π interpolates between the null distribution ($\pi = 0.5$) and a location shift model ($\pi = 0$). We now perform the same simulation as Golland et al. [2005], but in the same dimensionality of



(a) Illustration with $p = 2$.

(b) Detection power in original space.

(c) Detection power in augmented space.

Fig. 6: **Logistic Regression. No main effects.** An example of a "pure rescaling", with no shift. First class in red squares. Second class in black rhombus. Data generated via $Y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1+\exp(\eta)]; \eta = \beta_0 + x'x; X \sim \mathcal{N}(0, \sigma^2 I_{p \times p})$.

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] locations tests are still preferred over accuracy-tests.

4. Departure from Homoskedasticity and Scalar Invariance

Our previous simulations assume 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 8a we can see the difference between the scalar-invariant *Srivastava* and *Goeman* statistics. We also see that two-group tests dominate accuracy-tests also in the heteroskedastic case.

4.1 *Tie Breaking*

As already stated in the introduction, the accuracy statistic is highly discrete. Especially the resubstitution accuracy-tests. Discrete test statistics lose power by not exhausting the permissible false positive rate. A common remedy is a *randomized test*, 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

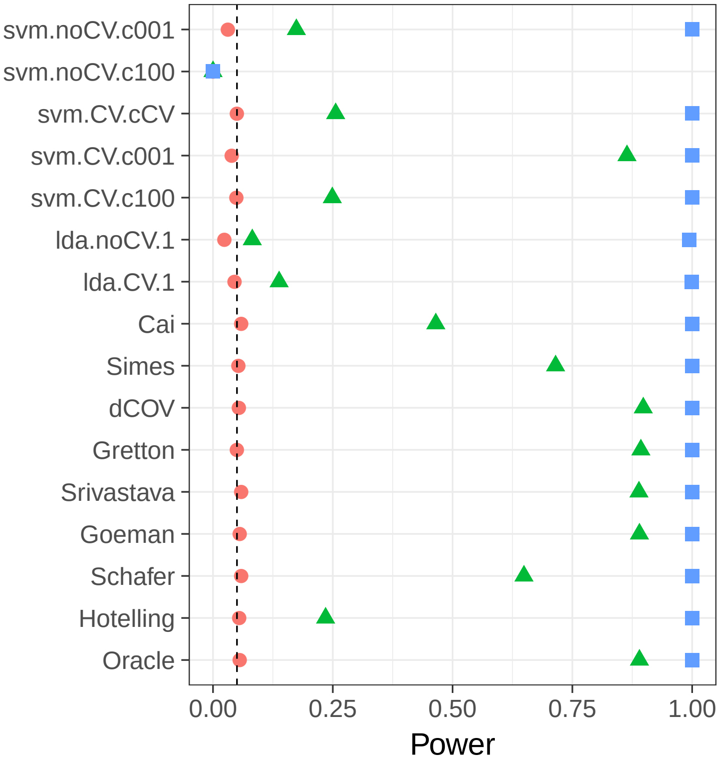


Fig. 7: **Mixture Alternatives.** $\mathbf{x}_i$ is distributed as in Eq.(3.1). μ is a p-vector with $3/\sqrt{p}$ in all coordinates. The effect, π, is color and shape coded and varies over 0 (red circle), 1/4 (green triangle) and 1/2 (blue square).

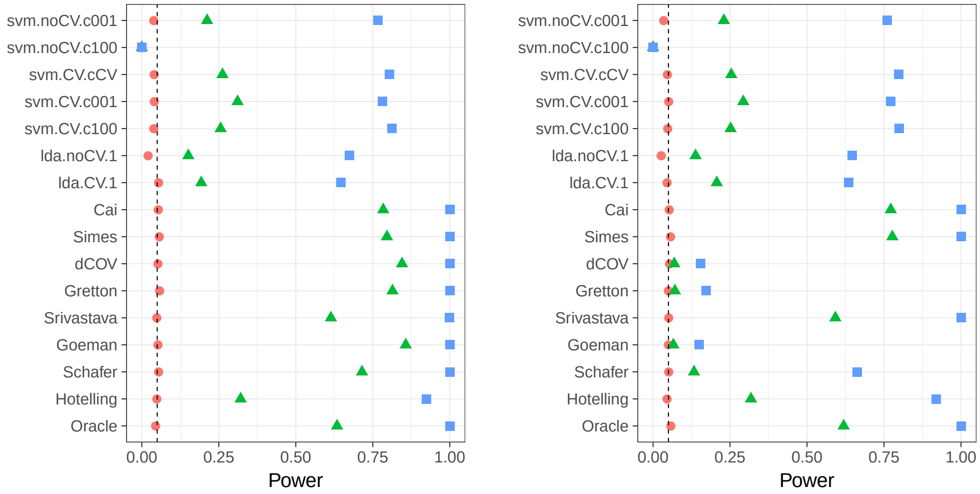
permutations satisfying A then the randomized version of our tests imply that if the permutation p-value, $\mathbb{P}\{\mathcal{T}_\pi \geqslant \mathcal{T}\}$, is greater than α then we reject the null with probability

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

Figure 9 reports the same analysis as in Figure **??**, after allowing for random tie breaking. It demonstrates that the power disadvantage of accuracy-tests, cannot be remedied by random tie breaking.

5. Sparse Alternatives

In our set of simulations we discussed "dense" alternatives, in the sense that all coordinates carry signal. Dense alternatives are motivated by neuroimaging where most brain locations in a regions carry signal. In a genetic application, a "sparse" alternative may be more plausible. Figure 10 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*).



(a) μ in the high variance PC of Σ.

(b) μ in the low variance PC of Σ.

Fig. 8: Heteroskedasticity: Σ is diagonal with $\Sigma_{jj} = j$.

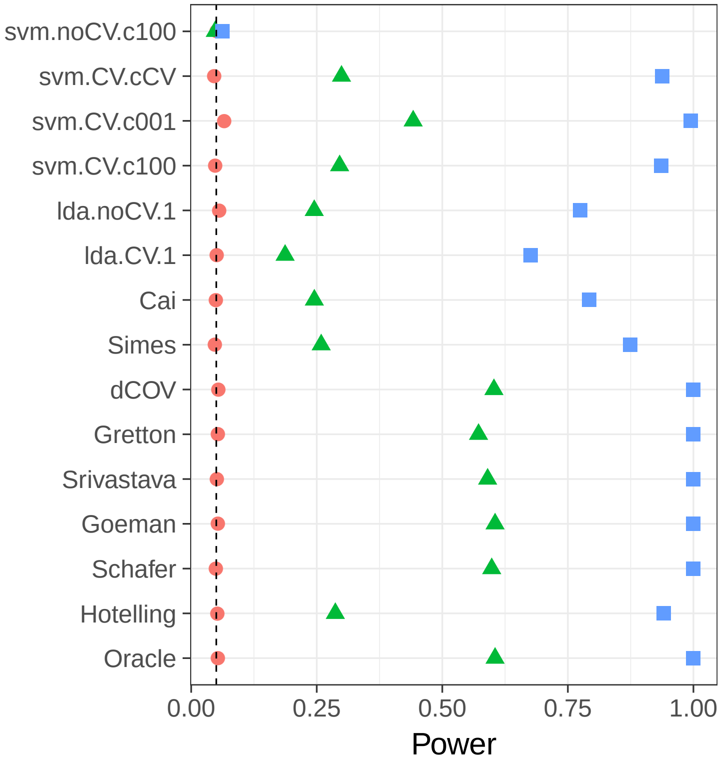


Fig. 9: The same as Figure **??**, with random tie breaking.

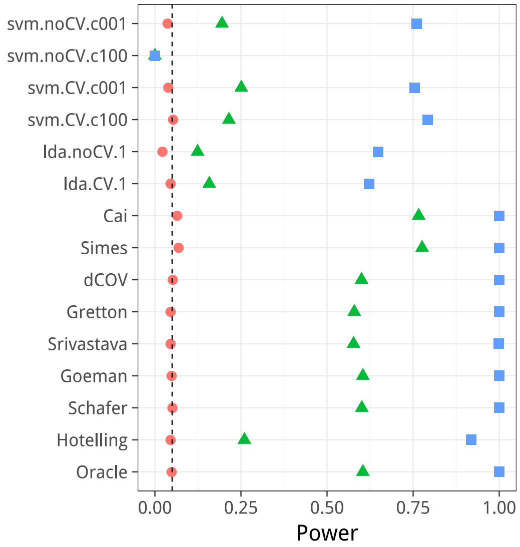


Fig. 10: Sparse μ.

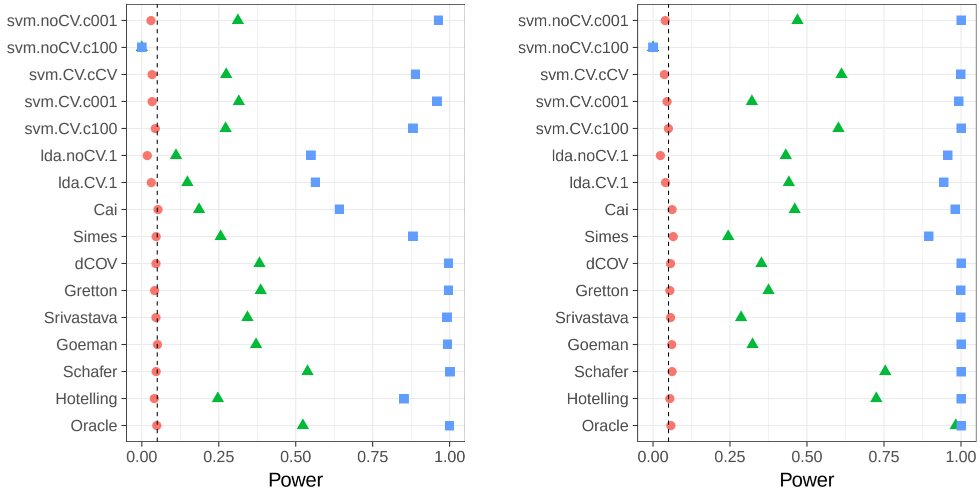
6. Fixed SNR

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. Abusing notation, we fix $KL[\bar{x}_1, \bar{x}_0] = c^2 p$, which is the same as fixing $\|\mu\|^2_\Theta$, with the exception of the large sample (**??**) and the heavytailed analysis (**??**).

Our choice implies that the Euclidean norm of $\mu := \mathbb{E}(x_1) - \mathbb{E}(x_0)$ 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 $\|\mu\|_\Theta$.

For completeness, Figure 11 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 11a), versus a shift in the direction of a low variance PC (Figure 11b). 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 $\|\mu\|^2_2 / \|\Sigma\|^2_{Frob}$ where $\|\Sigma\|^2_{Frob} = \text{Tr}(\Sigma'\Sigma)$ is the Frobenius matrix norm. Clearly, $\|\mu\|^2_2 / \|\Sigma\|^2_{Frob}$ is invariant to the direction of the signal with respect to the noise. For this reason, we prefer fixing $\|\mu\|_\Theta$.



(a) μ in PC7 of Σ.

(b) μ in PC15 of Σ.

Fig. 11: Short memory, AR(1) correlation. $\|\mu\|_2$ fixed.

References

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.

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. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

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.