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

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

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, our findings hold for these correlation structures. In particular: two-group tests dominate accuracy tests, and signal is masked in the low PCs of the noise.

3. Departure From Shift Alternatives

In the main text we have argued that shift alternatives are the most common in the univariate statistical literature. They are also very common in the multivariate literature, as they are implied by Fisher's LDA model, and in multivariate analysis of variance (MANOVA). On the other hand, effects may manifest themselves in many ways. We thus verify our statements in models which

*johnros@bgu.ac.il

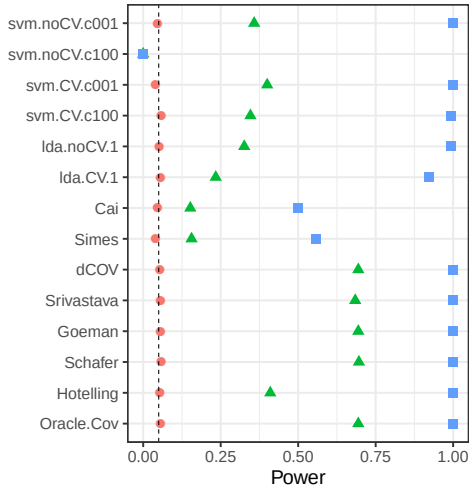
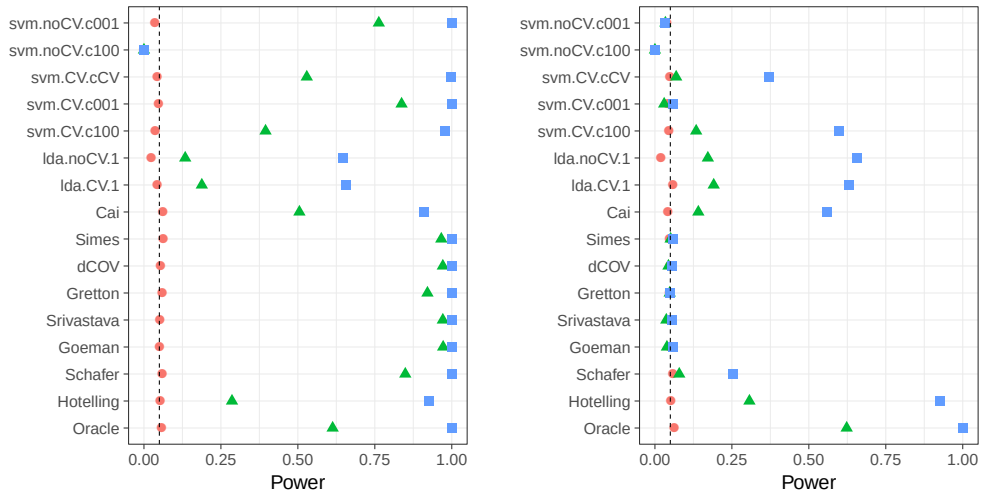


Fig. 1: **Large sample**: The basic simulation setup scaled ten-fold: $n = 400; p = 230$.



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

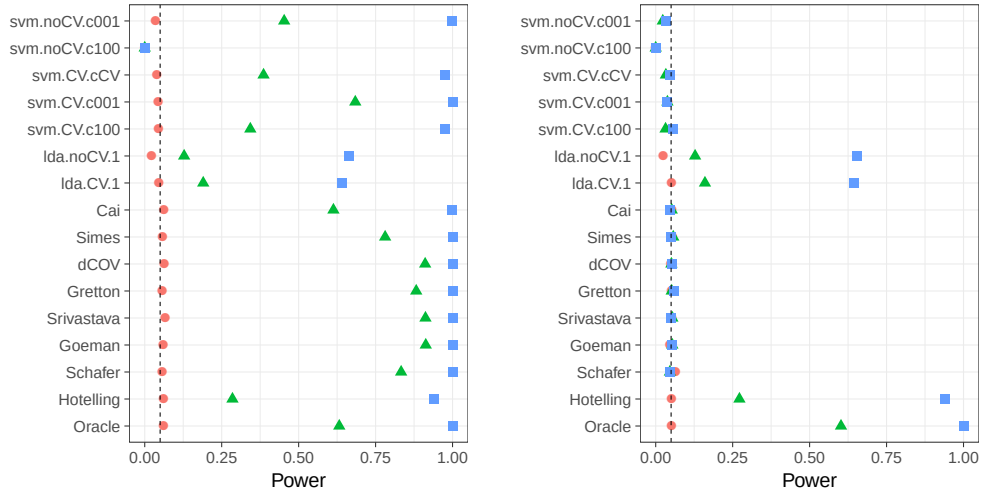
(b) Signal in direction of lowest lowest variance PC of Σ.

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

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 exact setup is also reported in the main text.



(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.

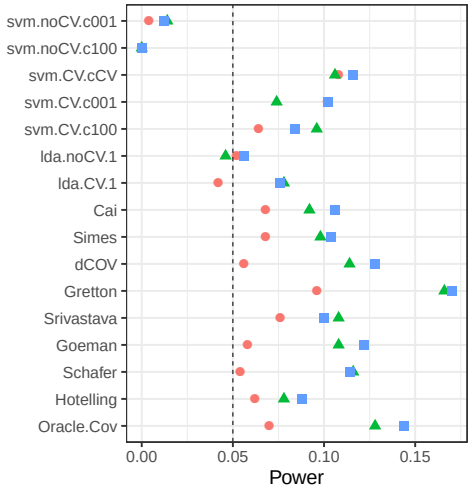
Formally, the logistic assumption implies that $P(y = 1|x) = \exp(\eta)/[1 + \exp(\eta)]$. Main-effects and second order interaction imply that $\eta = x'\beta + x'Bx$, for some p-vector β, and symmetric $p \times p$ matrix B. We also assume $x \sim \mathcal{N}(0, I_{p\times p})$. We perform the various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

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

The logistic assumption differs from our original setup in that it states $y|x$, instead of $x|y$. In the terms of Ng and Jordan [2002], the logistic is a *discriminative model* whereas Fisher's LDA is a *generative model*. The logistic assumption implies that x_0 is no longer a shifted versions of x_1, even in the presence of main effects alone ($B = 0$). While not a "pure shift", the logistic model with main effects has a strong shift component. We thus expect it to behave like the basic setup. This is verified in Figure 5, where two-group tests dominate accuracy-tests in the original space, and in the augmented space.

It is possible to use the logistic setup to generate data with no shift at all. For instance, if effects have a quadratic form in link scale: $\eta = \beta_0 + x'x$. This is depicted for $p = 2$ in Figure 6a. This example is typically encountered in the machine learning literature, to motivate learning with kernels.

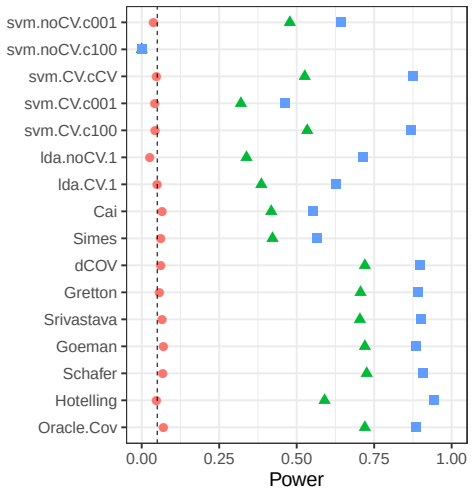
We need to distinguish between analysis in the original x space, and in the augmented interaction space. In the original space, x_1 is more akin to a *rescaling* of x_0 than a shift. In the augmented space $\tilde{x}$, x_1 is more akin to a *shifting* of x_0. We thus expect that all tests will perform poorly in the original space, and two-group tests to dominate in the augmented space. This is confirmed in Figure 6.



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

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

Fig. 4: **Logistic regression with second order interactions.** Data generated via $y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1+\exp(\eta)]; \eta = x'\beta + x'Bx; x \sim \mathcal{N}(0, I_{p\times p})$.



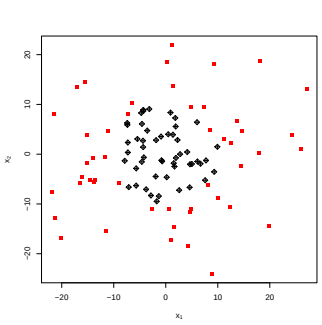
(a) Data analyzed in the original space (x).

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

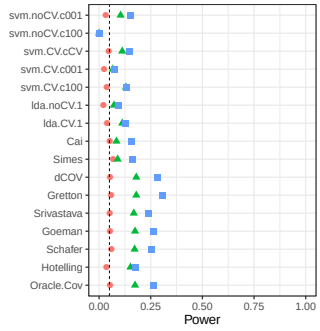
Fig. 5: **Logistic Regression. Main effects only.** Data generated via $y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1+\exp(\eta)]; \eta = x'\beta; x \sim \mathcal{N}(0, I_{p\times p})$.

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.,



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



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

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

Fig. 6: **Logistic regression. Second order interactions only.** Data generated via $y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1 + \exp(\eta)]; \eta = \beta_0 + x'x; x \sim \mathcal{N}(0, \sigma^2 I_{p\times p})$.

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 location 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] locations tests are still preferred over accuracy-tests.

4. 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

$$KL[x_1, x_0] = \frac{1}{2}\left(\log\frac{\det\Sigma_0}{\det\Sigma_1} - p + Tr(\Sigma_0^{-1}\Sigma_1) + (\mu_0-\mu_1)'\Sigma_0^{-1}(\mu_0-\mu_1)\right), \tag{4.2}$$

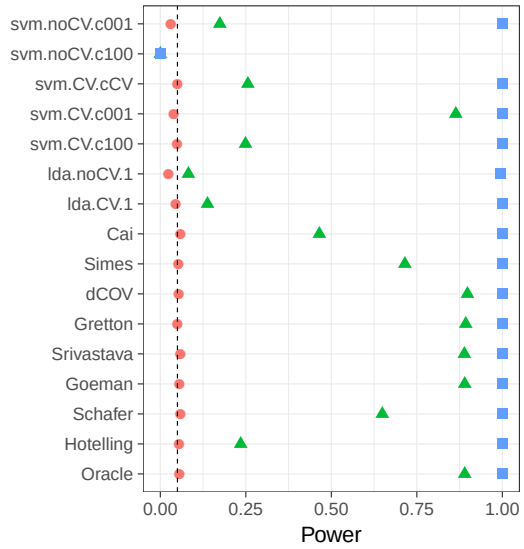


Fig. 7: **Mixture Alternatives.** $\mathbf{x}_i$ is distributed as in Eq.(3.1). μ is a p-vector with $3/\sqrt{p}$ in all coordinates. The effect, π, is color and shape coded and varies over 0 (red circle), 1/4 (green triangle) and 1/2 (blue square).

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

$$KL[\bar{x}_1, \bar{x}_0] = \frac{n}{2}\mu'\Sigma^{-1}\mu = \frac{n}{2}\|\mu\|^2_\Theta, \tag{4.3}$$

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

In most of our simulations we fixed $n\|\mu\|^2_\Theta$. The logistic regression setup is an exception because... [TODO: relate to logistic regression]

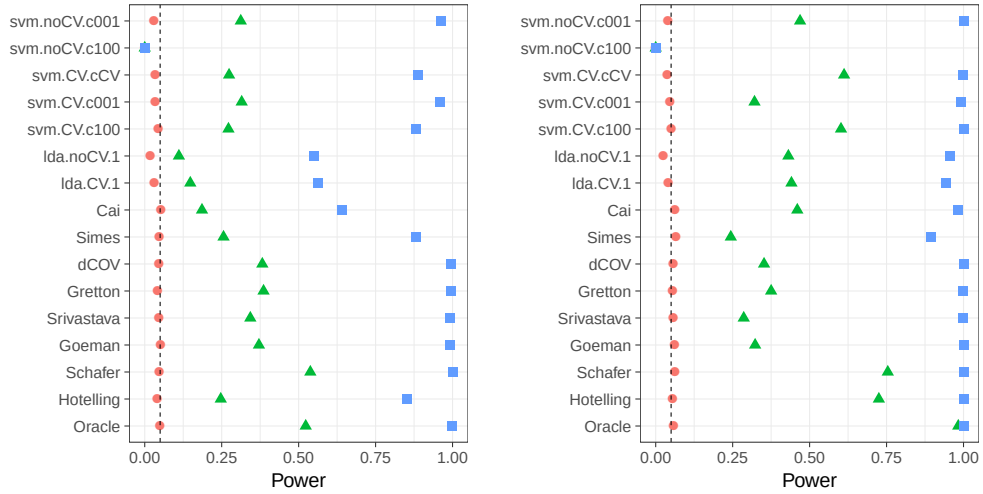
Fixing $n\|\mu\|^2_\Theta$ implies that the Euclidean norm of μ varies with Σ, with the sample size, and with the direction of the signal. An initial intuition may suggest that detecting signal in the low variance PCs is easier than in the high variance PCs. This is true when fixing $\|\mu\|_2$, but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

For completeness, Figure 8 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 8a), versus a shift in the direction of a low variance PC (Figure 8b). 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 $\|\mu\|_\Theta$.

5. 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



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

Fig. 8: Short memory, AR(1) correlation. $\|\mu\|_2$ fixed.

sparse alternative may be more plausible. Figure 9 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*).

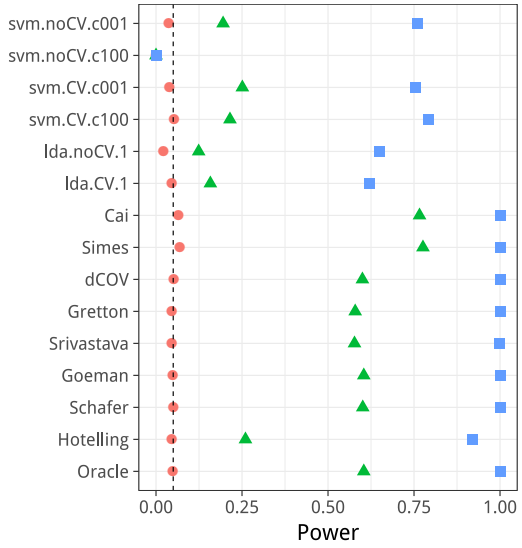
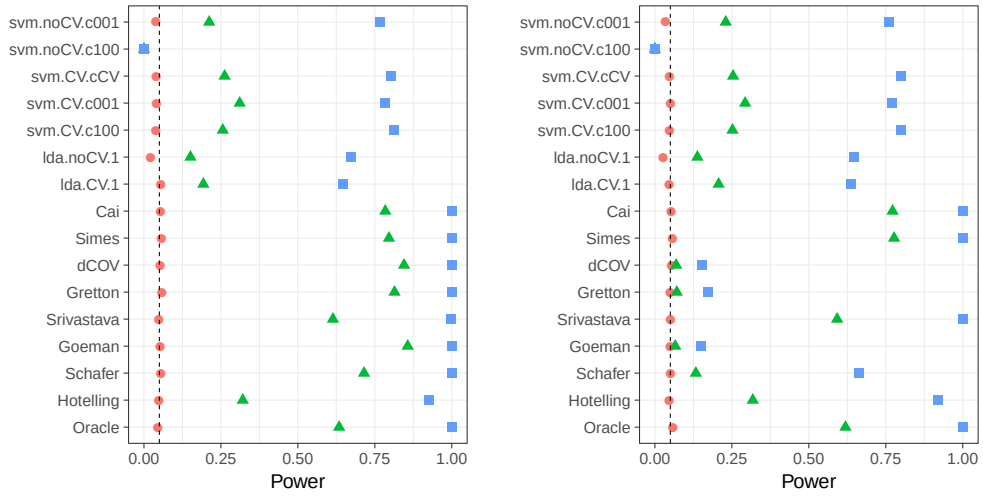


Fig. 9: Sparse μ.

6. Departure from Homoskedasticity and Scalar Invariance

Our previous simulations assume variables have unit variance (diagonal Σ). 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 10a 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.



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

Fig. 10: Heteroskedasticity: Σ is diagonal with $\Sigma_{jj} = j$.

7. Tie Breaking

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 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 11 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.

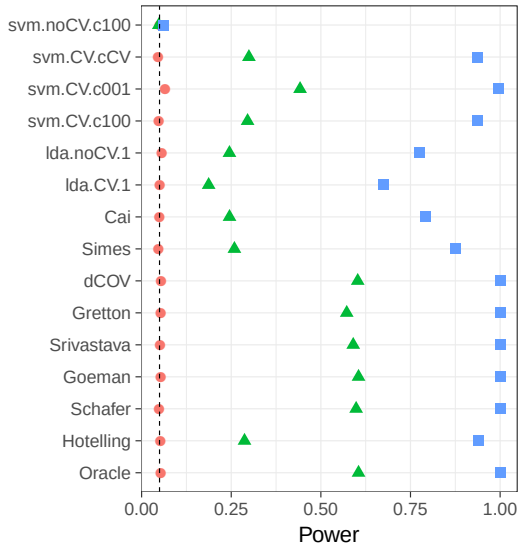


Fig. 11: **Tie breaking**: The basic simulation setup with random tie breaking.

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.

[*xxx*]