Biostatistics (2019), **0**, 0, *pp.* 1–8
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 about here.]

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.

[Fig. 2 about here.]

[Fig. 3 about here.]

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 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.

[Fig. 4 about here.]

[Fig. 5 about here.]

[Fig. 6 about here.]

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.

[Fig. 7 about here.]

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. 8 about here.]

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.

[Fig. 9 about here.]

6. 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 10 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. 10 about here.]

7. 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

$$KL[x_1, x_0] = \frac{1}{2}\left(\log\frac{\det\Sigma_0}{\det\Sigma_1} - p + Tr(\Sigma_0^{-1}\Sigma_1) + (\mu_0 - \mu_1)'\Sigma_0^{-1}(\mu_0 - \mu_1)\right), \tag{7.2}$$

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

$$KL[\bar{x}_1, \bar{x}_0] = \frac{n}{2}\mu'\Sigma^{-1}\mu = \frac{n}{2}\|\mu\|^2_\Sigma, \tag{7.3}$$

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

In most of our simulations we fixed $n\|\mu\|^2_\Sigma$. 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 $n\|\mu\|^2_\Sigma$ 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 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.

[Fig. 11 about here.]

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 $\|\mu\|_\Sigma$.

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. 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*]

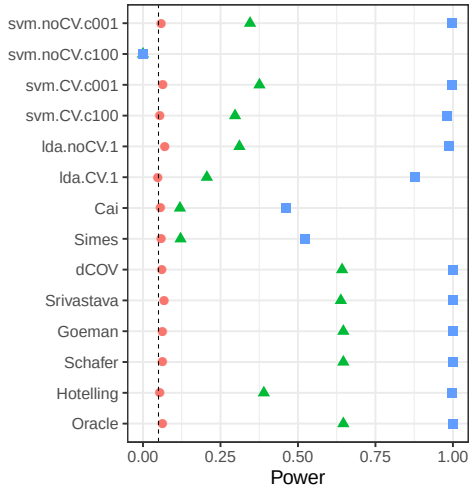
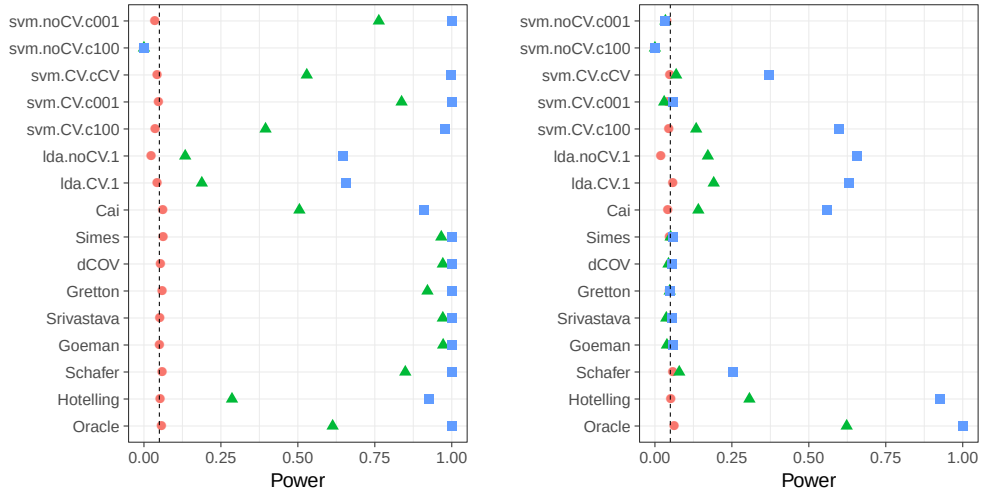


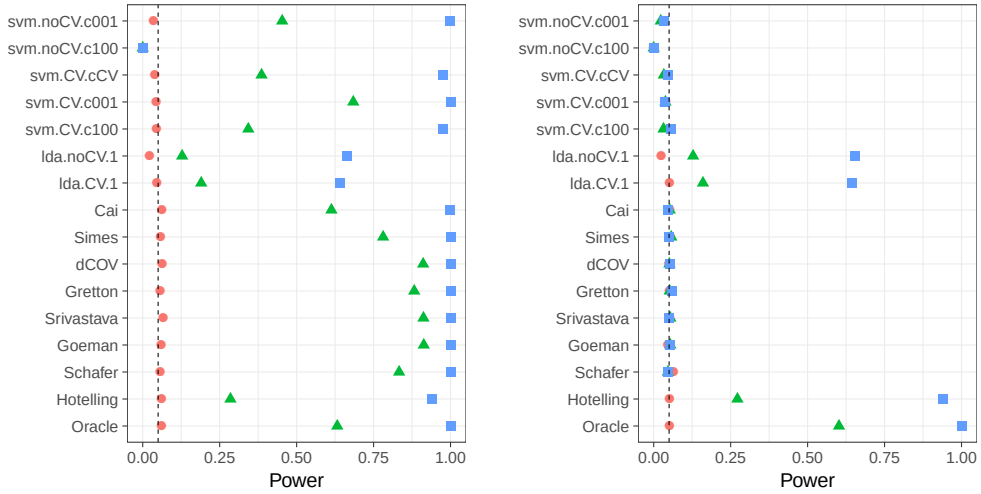
Fig. 1: **Large sample**: The basic simulation setup scaled ten-fold: $n = 400; p = 230$. $\frac{n}{2}\|\mu\|_{\Sigma}^2$ was set to 0 (red circle), 100 (green triangle), 400 (blue square) for comparable power to other simulations.



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

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

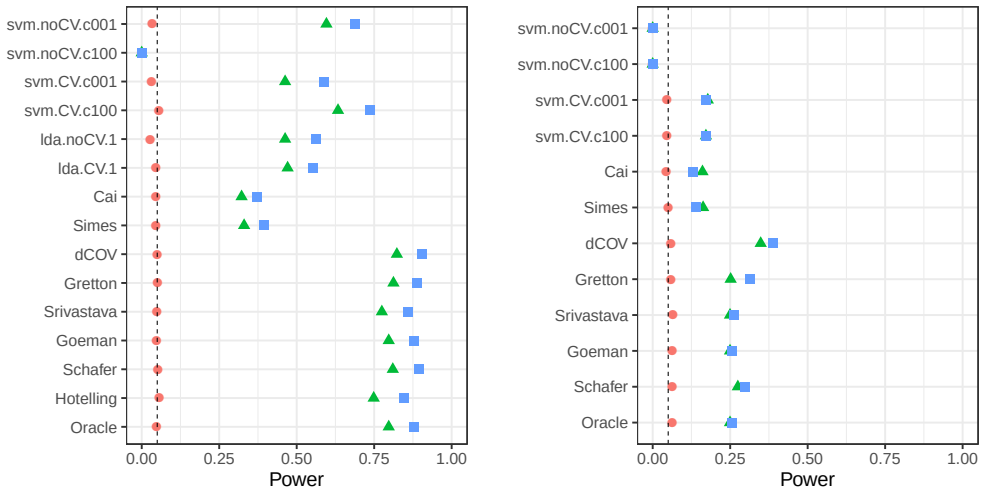
Fig. 2: Long-memory Brownian motion correlation: $\Sigma = D^{-1}RD^{-1}$ where D is diagonal with $D_{jj} = \sqrt{R_{jj}}$, and $R_{k,l} = \min\{k, l\}$.



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

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

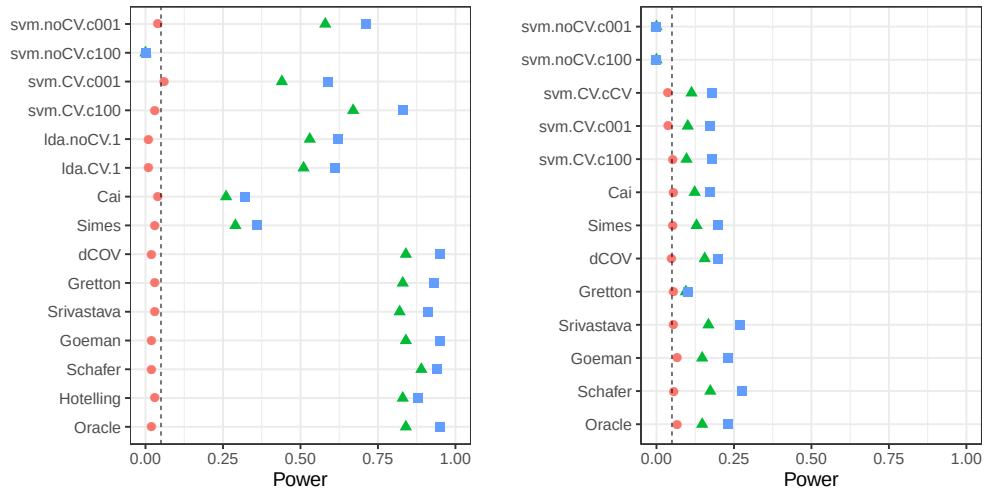
Fig. 3: Arbitrary Correlation. $\Sigma = D^{-1}RD^{-1}$ where D is diagonal with $D_{jj} = \sqrt{R_{jj}}$, and $R = A'A$ where A is a Gaussian $p \times p$ random matrix with independent $\mathcal{N}(0, 1)$ entries.



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

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

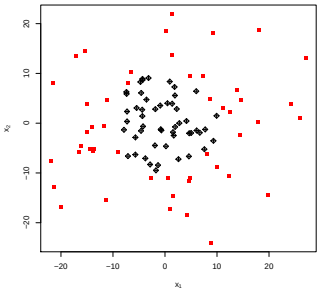
Fig. 4: **Logistic regression. Main effects and interactions.** Data generated via $y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1+\exp(\eta)]; \eta = \beta_0 + x'\beta + x'Bx$ where β is a scaled vector of ones, B a scaled identity matrix, and β_0 set so that the median $\eta \approx 0$. Finally, $x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



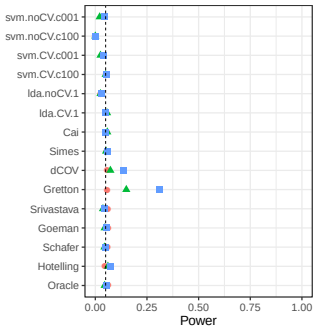
(a) Data analyzed in the original space (x).

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

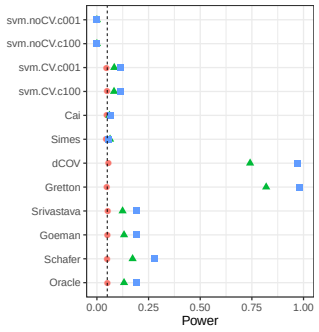
Fig. 5: **Logistic Regression. Main effects only.** Data generated via $y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1+\exp(\eta)]; \eta = \beta_0 + x'\beta$ where β is a scaled vector of ones, and β_0 set so that the median $\eta \approx 0$. Finally, $x \sim \mathcal{N}(0, I_{p\times p})$.



(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, $x \sim \mathcal{N}(0, I_{p\times p})$.

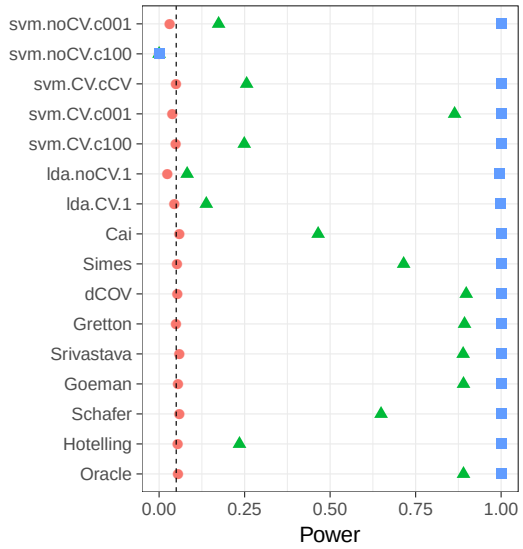


Fig. 7: **Mixture Alternatives.** $\mathbf{x}_i$ is distributed as in Eq.(3.1). μ is a p-vector with $3/\sqrt{p}$ in all coordinates. The effect, π, is color and shape coded and varies over 0 (red circle), 1/4 (green triangle) and 1/2 (blue square).

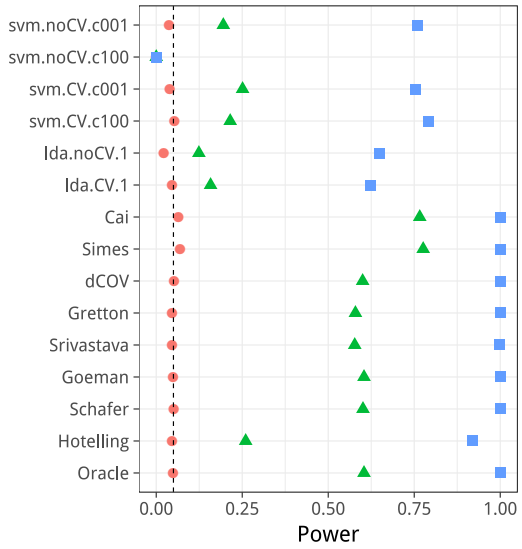
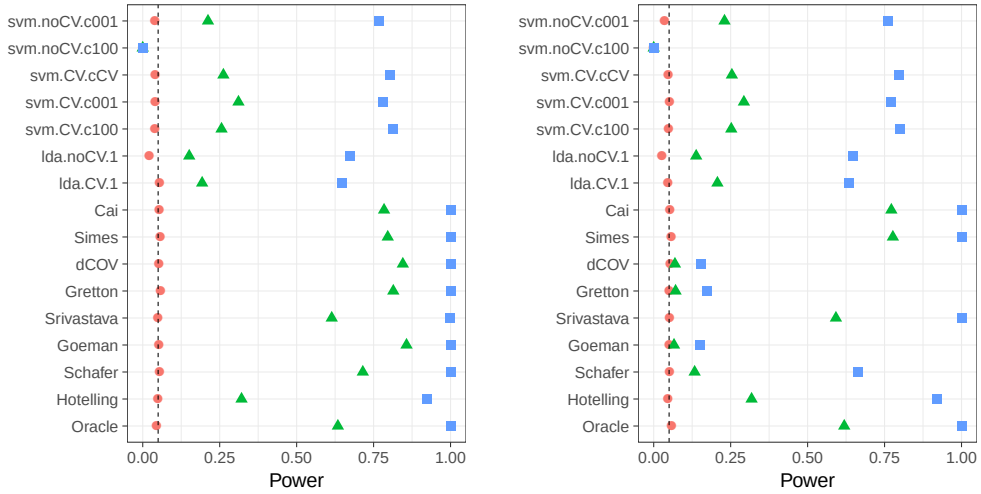


Fig. 8: Sparse μ.



(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 $\Sigma_{jj} = j$.

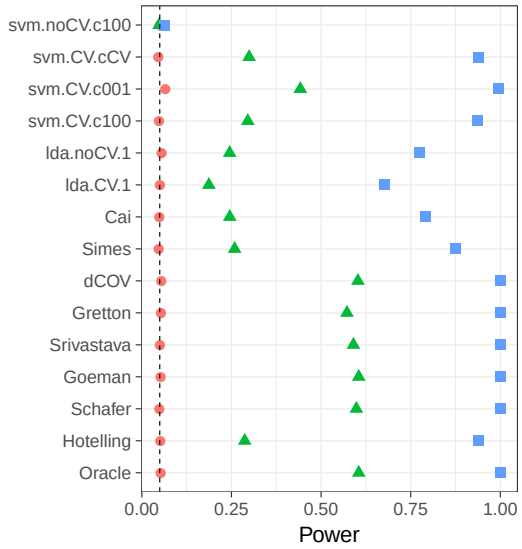
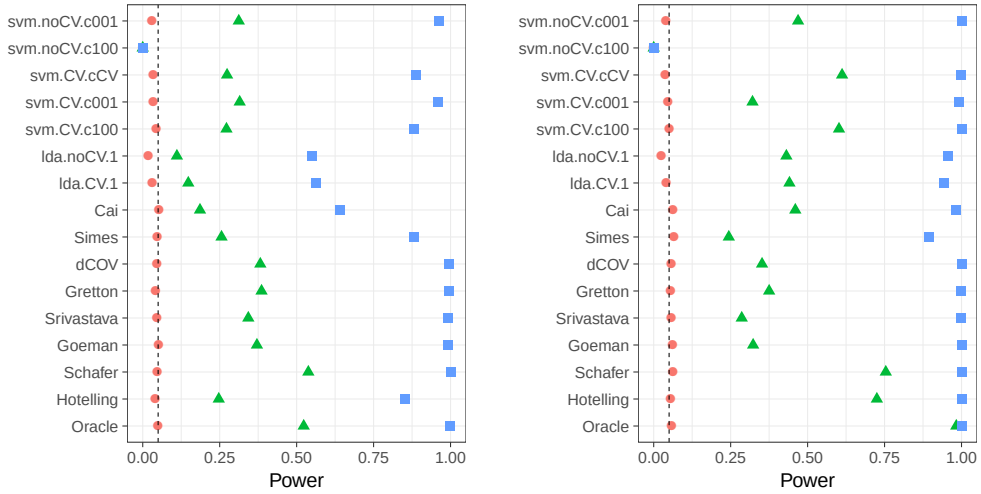


Fig. 10: **Tie breaking**: The basic simulation setup with random tie breaking.



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

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