A framework for paired-sample hypothesis testing for high-dimensional data

Ioannis Bargiotas, Argyris Kalogeratos, Nicolas Vayatis

Centre Borelli, ENS-Paris Saclay

6 Nov 2023

Ioannis.Bargiotas@ens-paris-saclay.fr

ICTAI (2023) 6-8 November 2023

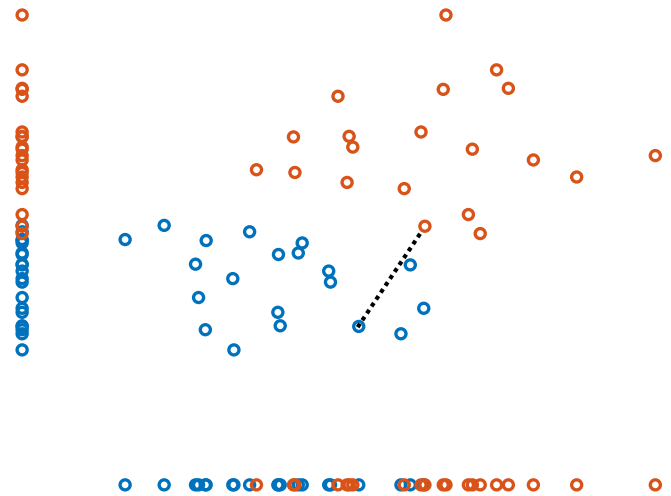
CENTRE BORELLI

école normale supérieure paris–saclay

université PARIS-SACLAY

1.**Background:** Common Paired Sample Setting

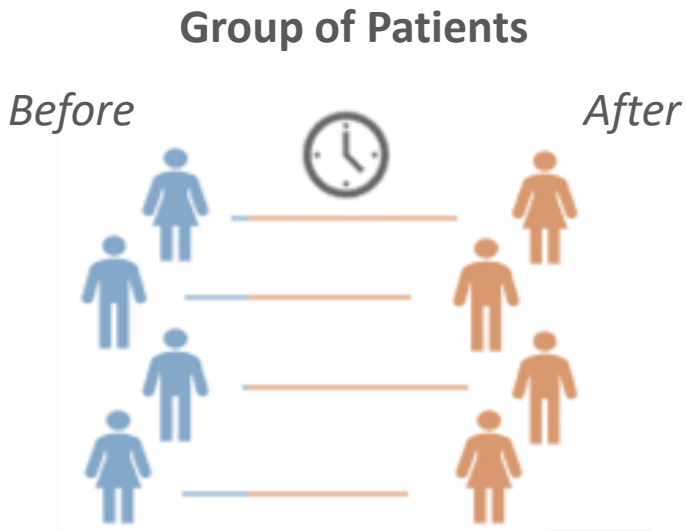
Definition: Paired data is where **natural matching or coupling** is possible. Every data point in one sample would be paired—uniquely—to a data point in another sample.

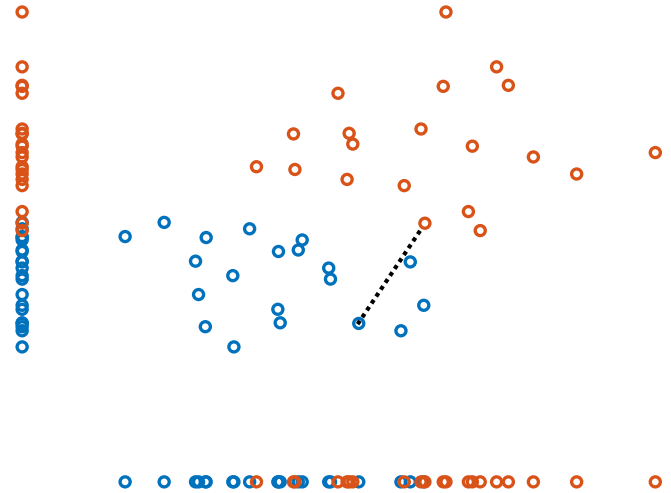


1.**Background:** Common Paired Sample Setting

Definition: Paired data is where **natural matching or coupling** is possible. Every data point in one sample would be paired—uniquely—to a data point in another sample.

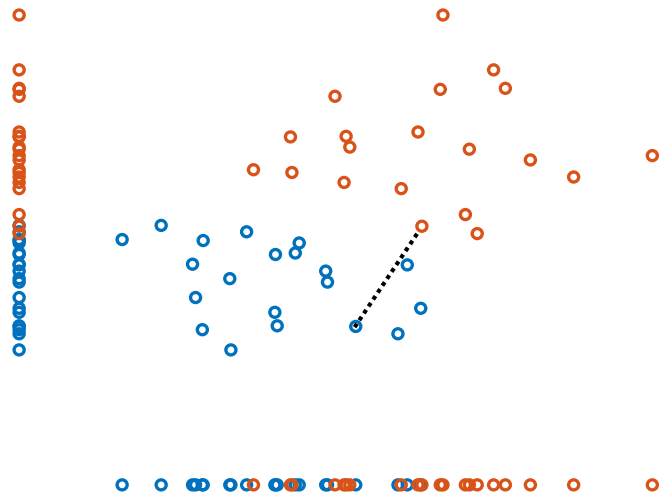
e.g. Sequential measurements (pre-treatment/post-treatment).



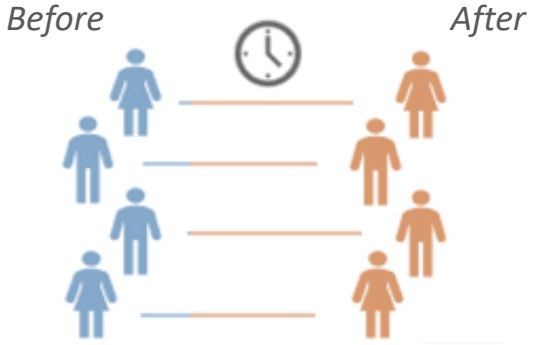


1.Background: Common Practices in various fields

- For $D \in \mathbb{R}^1$
  - Use of Univariate hypothesis tests such as T-tests (parametric) or rank statistics (non-parametric).
- For $D \in \mathbb{R}^M, where\ M > 1$
  - Use of multivariate hypothesis tests such as Hotelling T2-test (HT2) or,
  - (**Most of the times**) independent consecutive univariate hypothesis tests, or else Multiple testing (MT)

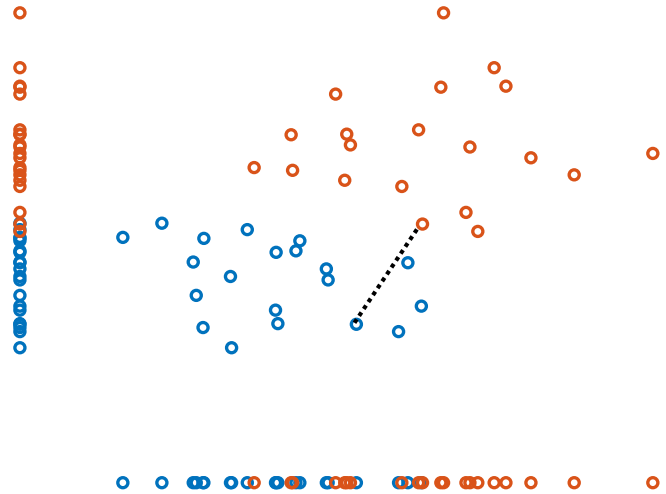


Paired sample setting



1.**Background:** Multiple Testing (MT) workflow

- Multiple features
- Multiple univariate testing (Parametric/Non parametric)
- p-value calculation per feature
- p-values adjustment (or not ?)*
- Report
  - Are the groups statistically different?
  - In which dimension (feature) ?



*Multiple hypothesis testing increases the likelihood of observing a significant result purely by chance (Type I error). To counteract this inflation of false positives, p-values are adjusted using **false discovery rate** (FDR) or **family wise error rate** (FWER) control methods (e.g. Bonferroni) to maintain a desired overall significance level.

1.Motivation: p-value debate…..To adjust or not to adjust?

Debate | Open Access | Published: 17 June 2002

Do multiple outcome measures require p-value adjustment?

Ronald J Feise

BMC Medical Research Methodology **2**, Article number: 8 (2002) | Cite this article

67k Accesses | **759** Citations | **9** Altmetric | Metrics

Summary

Readers should balance a study's statistical significance with the magnitude of effect, the quality of the study and with findings from other studies. Researchers facing multiple outcome measures might want to either select a primary outcome measure or use a global assessment measure, rather than adjusting the p-value.

BMJ
https://www.bmj.com › content

What's wrong with Bonferroni adjustments

by TV Perneger · 1998 · Cited by 6631 — This paper advances the view, widely held by epidemiologists, that **Bonferroni adjustments are, at best, unnecessary and, at worst,**…

When to use the B onferroni correction

RA Armstrong - Ophthalmic and Physiological Optics, 2014 - Wiley Online Library

Purpose The B onferroni correction adjusts probability (p) values because of the increased risk of a type I error when making multiple statistical tests. The routine use of this test has …

☆ Save Cite Cited by 2666 Related articles All 6 versions

nature › nature human behaviour › comment › article

Comment | Published: 01 September 2017

Redefine statistical significance

Daniel J. Benjamin, James O. Berger, … Valen E. Johnson + Show authors

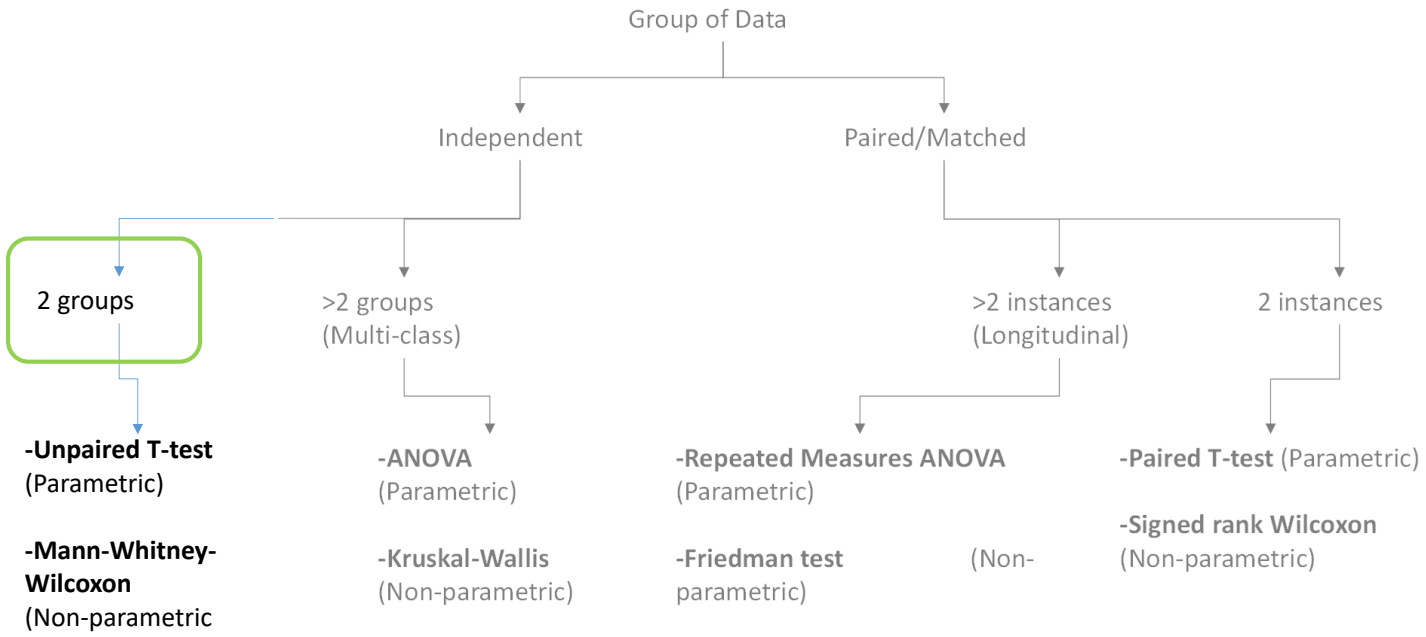
Nature Human Behaviour **2**, 6–10 (2018) | Cite this article

149k Accesses | **1158** Citations | **885** Altmetric | Metrics

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

The lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on 'statistically significant' findings. There has been much progress toward documenting and addressing several causes of this lack of reproducibility (for example, multiple testing, *P*-hacking, publication bias and under-powered studies). However, we believe that a leading cause of non-reproducibility has not yet been adequately addressed: statistical standards of evidence for claiming new discoveries in many fields of science are simply too low. Associating statistically significant findings with *P* <

1.Motivation: Machine learning Framework for the two-sample problem



ARTICLE FREE ACCESS

A kernel two-sample test

Authors: Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, Alexander Smola

Authors Info & Claims

The Journal of Machine Learning Research, Volume 13 • 3/1/2012 • pp 723–773

Corpus ID: 2049706

AUC optimization and the two-sample problem

S. Clémençon, N. Vayatis, M. Depecker • Published in NIPS 7 December 2009 • Computer Science

Classification accuracy as a proxy for two-sample testing

Ilmun Kim, Aaditya Ramdas, Aarti Singh, Larry Wasserman

Ann. Statist. 49(1): 411-434 (February 2021). DOI: 10.1214/20-AOS1962

Two-sample Testing Using Deep Learning

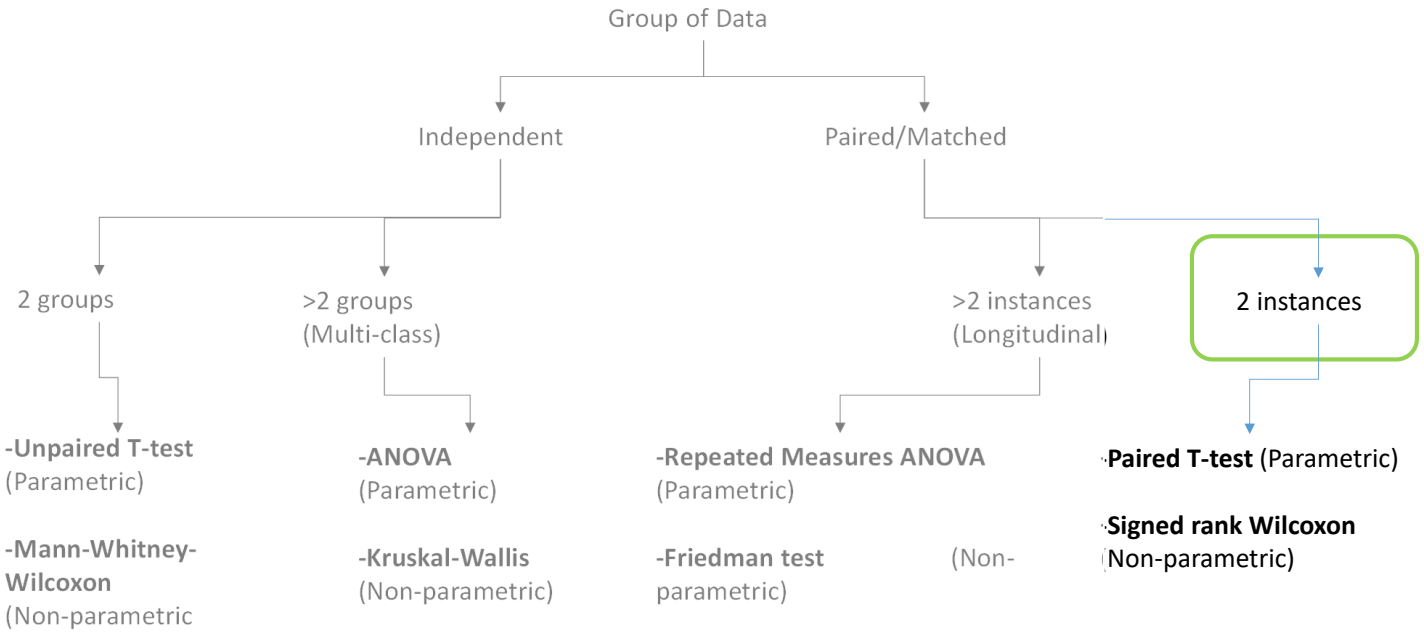
Matthias Kirchler, Shahryar Khorasani, Marius Kloft, Christoph Lippert *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, PMLR 108:1387-1398, 2020.

Revealing posturographic profile of patients with Parkinsonian syndromes through a novel hypothesis testing framework based on machine learning

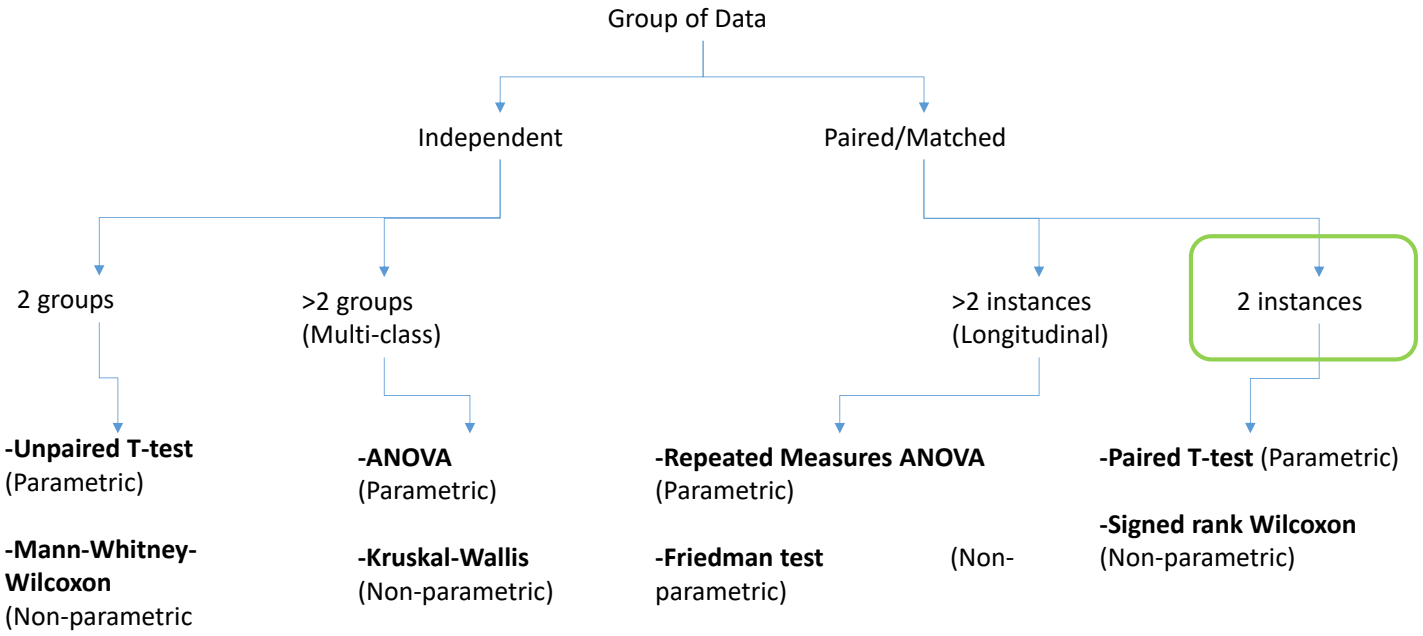
Ioannis Bargiotas, Argyris Kalogeratos, Myrto Limnios, Pierre-Paul Vidal, Damien Ricard, Nicolas Vayatis

Published: February 25, 2021 • https://doi.org/10.1371/journal.pone.0246790

1.**Motivation:** Machine learning Framework for paired samples?



1.**Motivation:** Machine learning Framework for paired samples?



Objectives: Develop a test that:

- extends the use of a well known test (Wilcoxon signed rank) to higher dimensions
- Provide well-known outputs (p-values, effect sizes, significant features)
- is easy-implemented and understandable from non-experts
- “bypasses” the p-value adjustment discussion.

2.Methodology: The 2-step framework for the paired-sample problem

1 **Step 1 - Scoring:** A decision rule is **"learned"**, using specific "properties", and scores the instances.

(Here the aggregation manner of Hodges-Lehmann estimator calculation)

2 **Step 2 - Testing:** A univariate test is applied to produced scores.

(Here the Wilcoxon sign rank (WSR) test)

Key point: The creation of the decision rule in **Step 1** should be linked to the statistic of the applied test in **Step 2**.

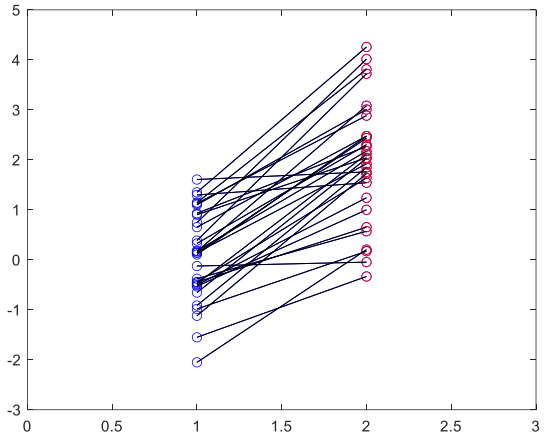
2.Methodology: Univariate Case - (Hodges-Lehmann and WSR)

Wilcoxon Signed Rank (WSR) test

For paired *X* and *Y*, WSR test can be transformed to One sample test, and tests if the median (θ) of $Z(Z_i = Y_i - X_i)$ differences is significantly different from *0*.

$$\mathrm{H}_0 : \theta = 0, \quad \text{the } Z_i\text{'s are symmetric around } \theta = 0;$$
$$\mathrm{H}_1 : \theta \neq 0, \quad \text{the } Z_i\text{'s are symmetric around } \theta \neq 0.$$

WSR ranks the $|Z_i|$ in ascending order, ignoring the signs and sums the ranks of the positive differences (greater than zero) to get **T⁺**.



Hodges-Lehmann estimator

The effect size θ between the paired samples, is estimated by the pseudomedian of the differences, which is in turn estimated by the Hodges-Lehmann estimator $\hat{\theta}$.

$$\mathrm{HL}(Z) = \hat{\theta} = median\left\{\frac{1}{2}(Z_i + Z_j);\ \forall i \leq j = 1, \ldots, N\right\}$$

Therefore, as the median of the pairwise average differences, or else the **Walsh Averages (W)**.

Key point 1: Supposing there are no ties and no zeros among the Z_i's, the number of positive Walsh averages (W^+) is equal to the WSR statistic **T⁺**

(Hoyland, A. Robustness of the Hodges-Lehmann estimates for shift. Annals of Math. Stat., 36(1):174–197, 1965.)

2.Methodology: From 1-*D* to *d-D*

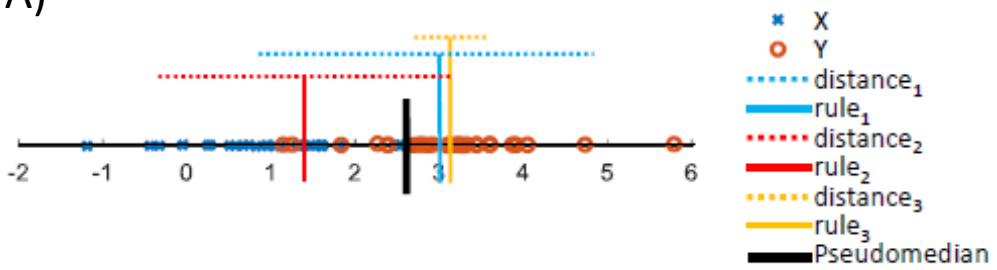
Multidimensional extension:

1. The multidimensional Euclidean distance between two paired instances, X_i and Y_i, can be seen as an analogy to the difference $Z_i = Y_i - X_i$ in a unidimensional setup.

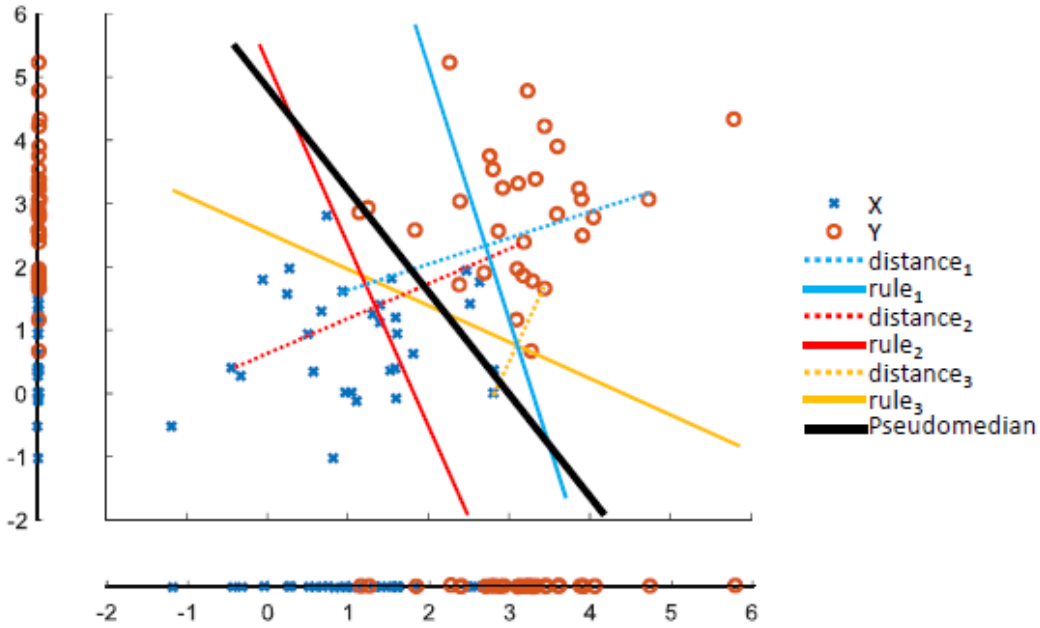
2. Each separating rule associated with a midpoint of the *1-D* case, now becomes a ***(d–1)*-dimensional perpendicular bisecting hyperplane**.

3. Each such hyperplane is computed by taking into account **only one specific pair of instances**, yet it splits the space in two parts, and therefore it can be seen as a **decision rule** that could hopefully classify the data in two parts, the X and the Y part.

A)

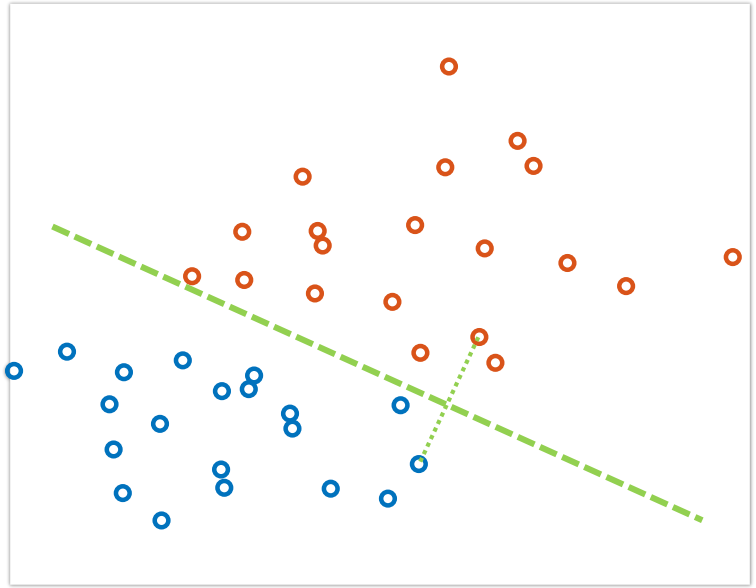


B)



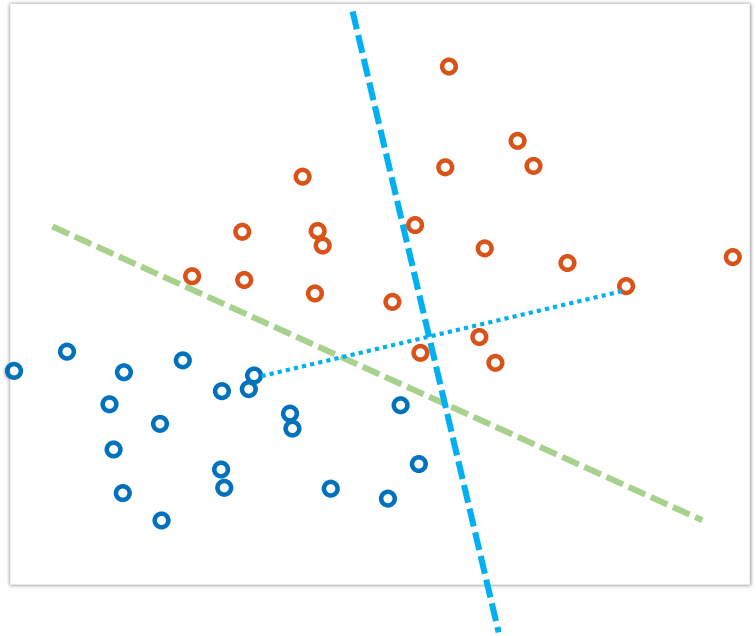
2.Methodology: MWSR test

1. Find the **(*d*–*1*)-dimensional perpendicular bisecting hyperplane** for every pair (X_i, Y_i)



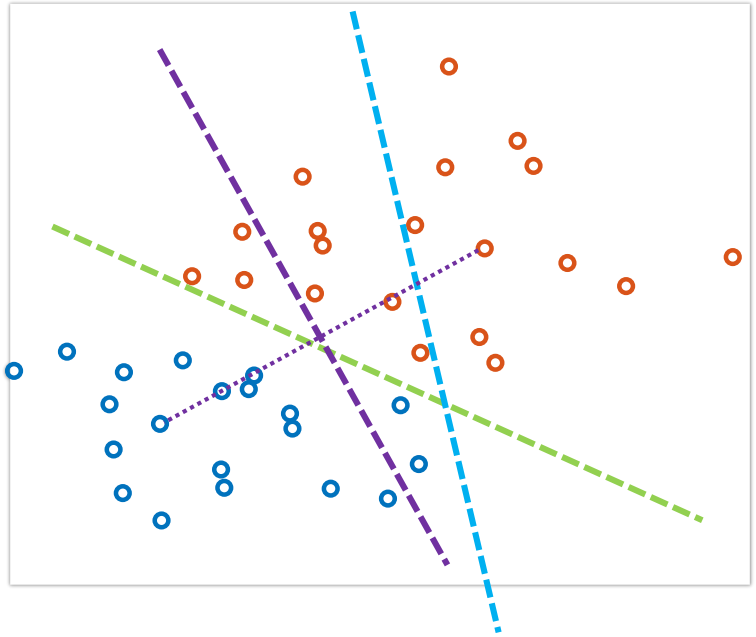
2.Methodology: MWSR test

1. Find the **(*d*–*1*)-dimensional perpendicular bisecting hyperplane** for every pair (X_i, Y_i)



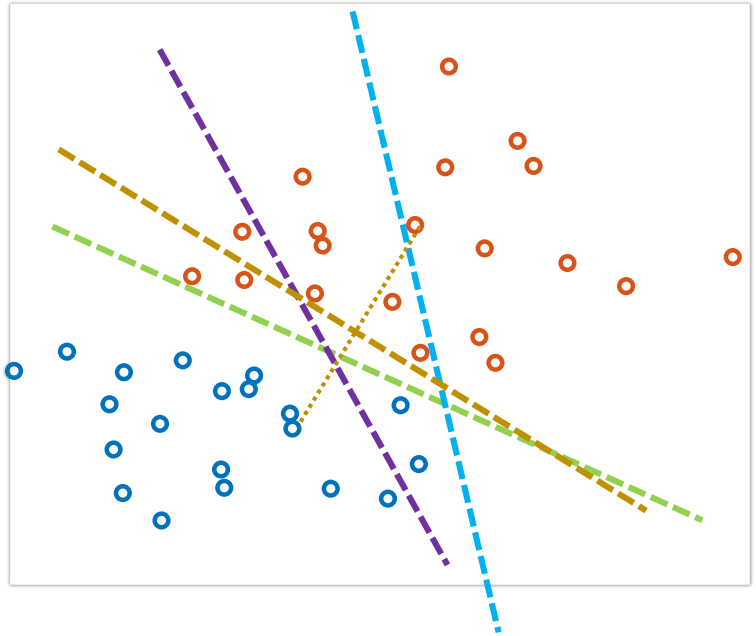
2.Methodology: MWSR test

1. Find the **(*d*–*1*)-dimensional perpendicular bisecting hyperplane** for every pair (X_i, Y_i)



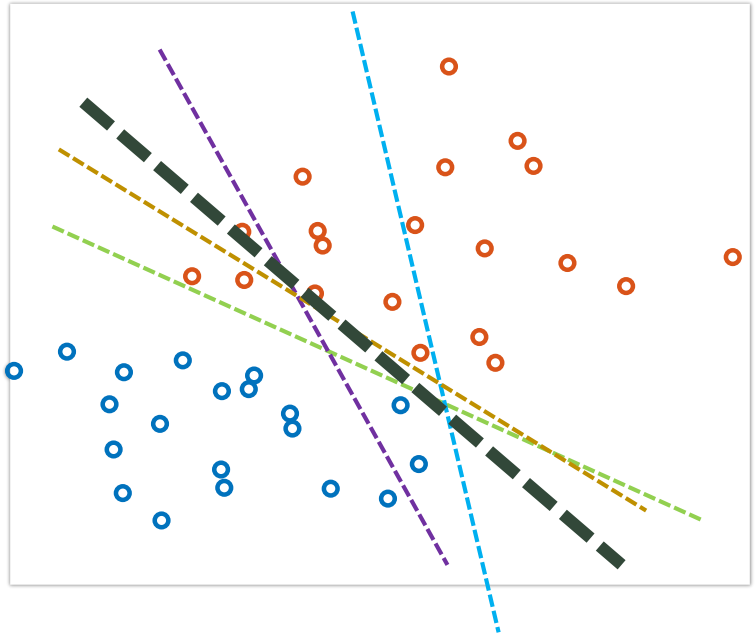
2.Methodology: MWSR test

1. Find the **(*d*–*1*)-dimensional perpendicular bisecting hyperplane** for every pair (X_i, Y_i)



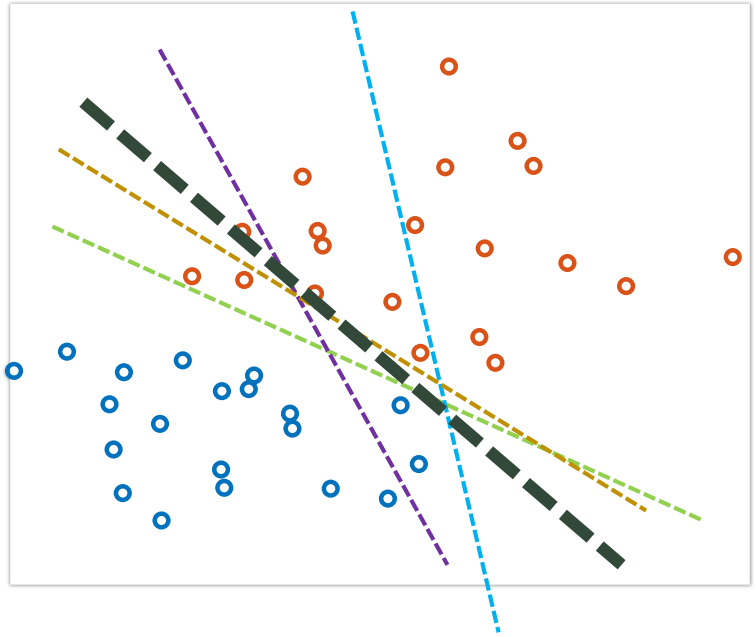
2.Methodology: MWSR test

1. Find the **(*d*–*1*)-dimensional perpendicular bisecting hyperplane** for every pair (X_i, Y_i)
2. Aggregate these decisions to a classifier $\hat{C}^*$ in a Hodges-Lehmann sense ($\hat{C}^*$ pseudomedian classifier)



2.Methodology: MWSR test

1. Find the **(*d*–*1*)-dimensional perpendicular bisecting hyperplane** for every pair (X_i, Y_i)
2. Aggregate these decisions to a classifier $\hat{C}^*$ in a Hodges-Lehmann sense ($\hat{C}^*$ pseudomedian classifier)
3. Score all instances S_X^*, S_Y^* (*1-D* representation)
4. Apply Wilcoxon Sign Rank WSR(S_X^*, S_Y^*)



Outputs:

1. p-value, size effect
2. Feature importance index from $\hat{C}^*$ coefficients.

2.Methodology: MWSR algorithm

Algorithm 1 The MWSR paired-sample testing framework

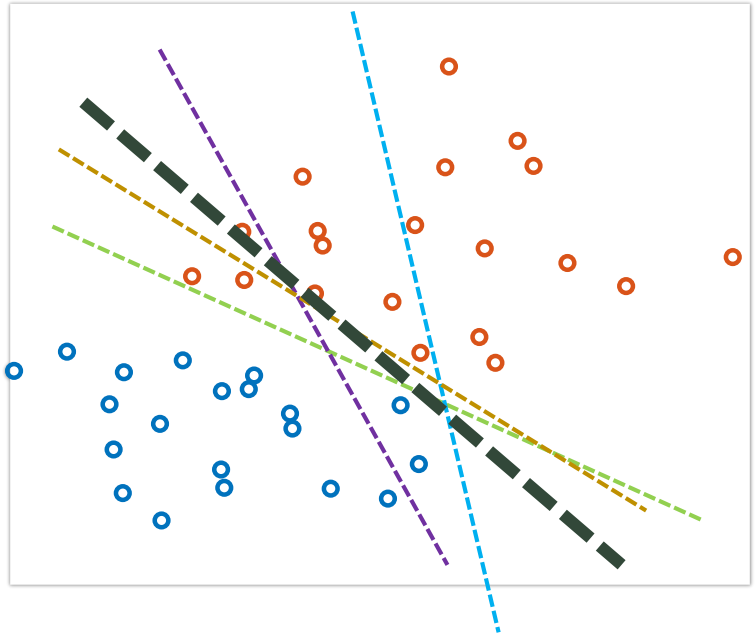
Input: $X, Y \in \mathbb{R}^{N \times d}$ are the $2 \cdot N$ paired samples;
Output: $\hat{C}^*$, $(\hat{S}_1^*, \hat{S}_2^*)$, p^*-*value*, θ^*, I^*

■ ***First step: Compute a scoring***

for $i = 1, .., N$ **do**
  $C_i \leftarrow$ perpendicular_bisector(X_i, Y_i)
end for
$k \leftarrow 1$; $M \leftarrow \mathbf{0}_{N \times N}$
for $i = 1, .., N$ **do**
  for $j = i, .., N$ **do**
    $W_{C,k} \leftarrow \frac{1}{2}(C_i + C_j)$ ▷ the Walsh average of hyperplanes
    $k \leftarrow k + 1$
  end for
end for
$\hat{C}^* \leftarrow$ median(W_C) ▷ the pseudomedian aggregate, see Eq. 5
$\hat{S}_1^*, \hat{S}_2^* \leftarrow$ get_scores$(\hat{C}^*(X, Y))$ ▷ classification-based scoring

■ ***Second step: Paired-sample test over the computed scores***

p^*-*value*, $\theta^* \leftarrow$ WSR$(\hat{S}_1^*, \hat{S}_2^*)$ ▷ p-value and effect size
$I^* \leftarrow w(\hat{C}^*)$ ▷ feature importance index
return $\hat{C}^*$, $(\hat{S}_1^*, \hat{S}_2^*)$, p^*-*value*, θ^*, I^*



2.Results: Synthetic dataset with progressive shift of mean.

Synthetic datasets

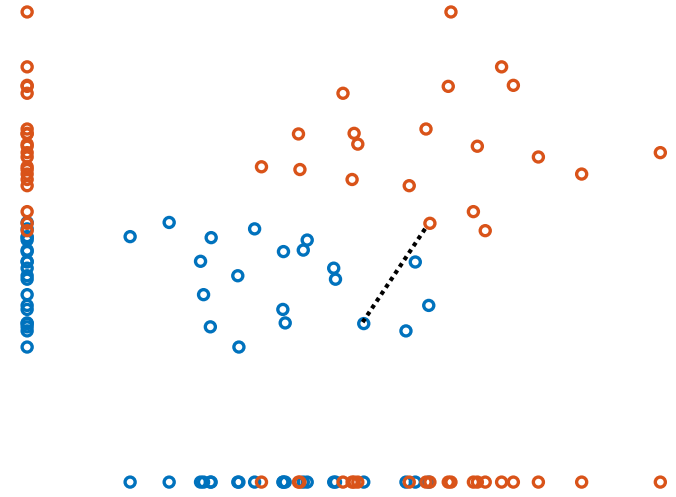
Synthetic data are simulated by pairing two samples coming from two Gaussian distributions, with feature-wise correlation:

$$\mathcal{R}_{X^{(k)},Y^{(k)}} = \frac{\mathrm{cov}(X^{(k)}, Y^{(k)})}{\sigma_{X^{(k)}}\sigma_{Y^{(k)}}} = 0.5, \qquad (6)$$

where $X^{(k)}$, $Y^{(k)}$ are vectors representing the paired samples with only the k-th feature, $\mathrm{cov}(\cdot,\cdot)$ is their covariance, and $\sigma_{X^{(k)}}$, $\sigma_{Y^{(k)}}$ are the respective standard deviations. To produce the dataset for each scenario, we set:

- a fixed population size $N = 30$ pairs of data instances;
- the dimensions $d = \{10, 20, 30, 60\}$ mimicking the number of features that a usual study may have;
- a standard deviation $std = \{1, 2\}$;
- the first 90% of the dimensions to have no statistical difference between the two samples, and hence to be randomly drawn (separately) from $\mathcal{N}(0, std)$;
- the last 10% of the dimensions to present the same difference in mean value, hence producing a linear shift. We allow this shift to also increase progressively from 0 to 1 to investigate the detection sensitivity of the methods..

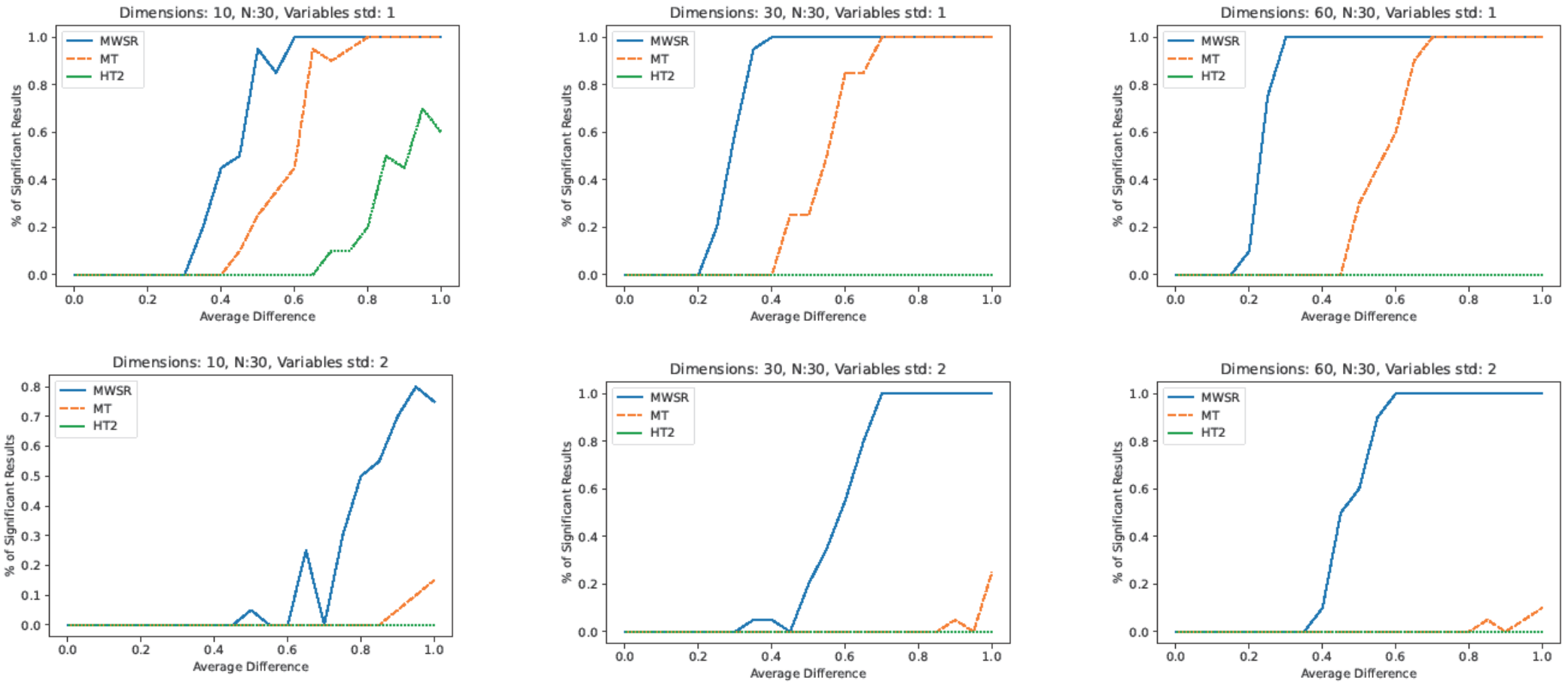
Given a scenario, we generate 20 cases and apply all statistical tests. We report the average performance, namely the percentage of cases with a significant shift obtained by each statistical test, as a function of the size of the shift (referred to as average difference in the figures). Moreover, we provide results regarding the inferred feature importance. Both elements should be examined jointly to validate the acquired results further.



Comparison between:

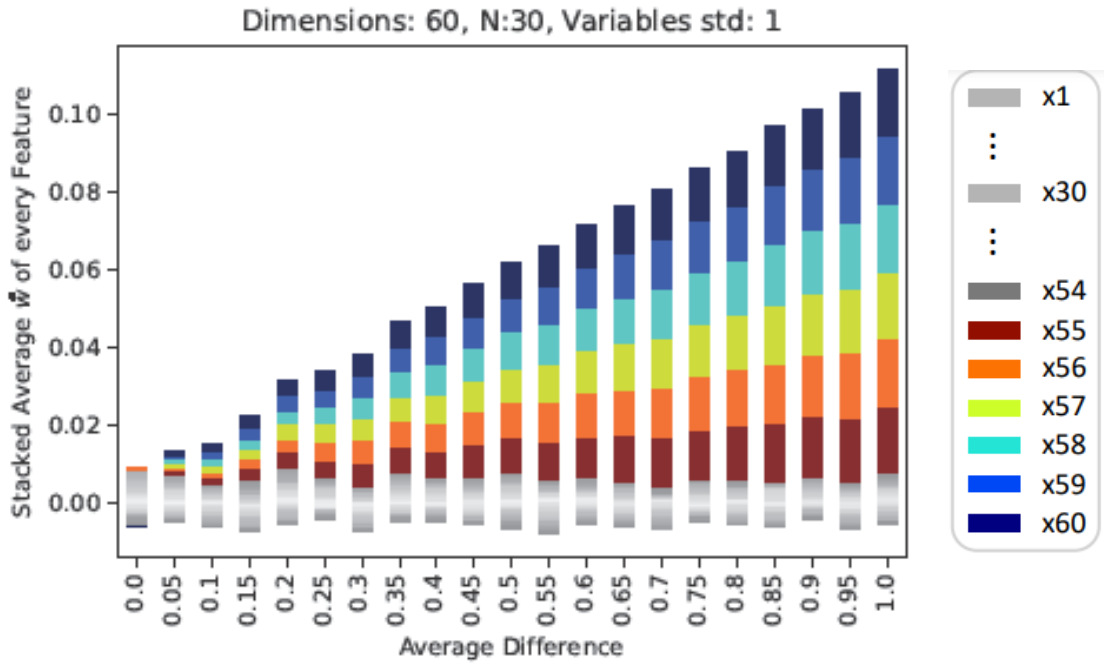
- MWSR
- Multiple Testing (MT) with Bonferroni adjustment
- Hotelling T2 test (HT2)

3. Results: Effect of dimensionality and variance



The performance is presented as a function of the separation distance between the two distributions. On the x-axis the progressive difference in the mean value for the 10% of the dimensions (D) while on the y-axis it appears the **%** of significant results detected by each method.

3. Results: Feature « importance » index



The relative feature importance for MWSR per feature. The feature importance performance is presented as a function of the separation distance between the two distributions (average difference on x-axis).

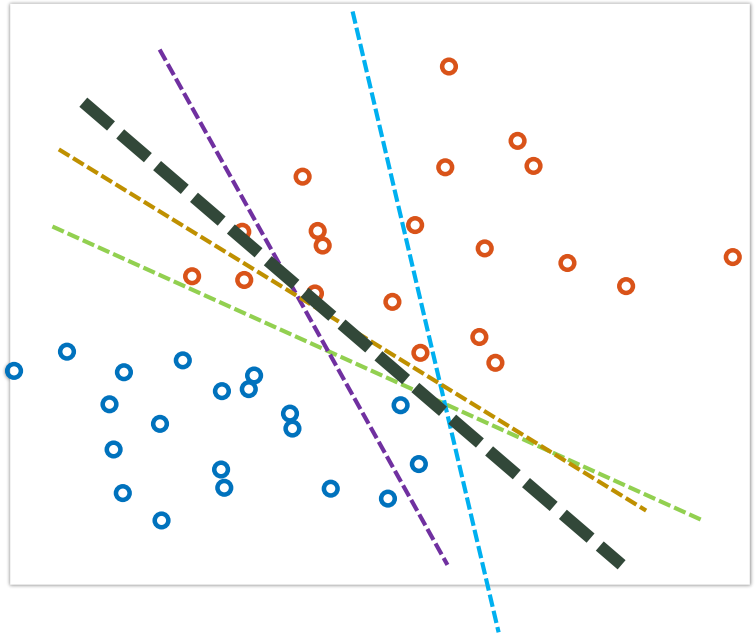
Conclusions/Perspectives

Conclusions

1. MWSR outperforms the classical multivariate approaches (Hotelling T2 test, the multiple testing with p-value adjustment) in our experiments.
2. MWSR is generally simple understandable
3. MWSR allows the user to interpret the actual contribution of every feature to the final result.
4. MWSR is customizable with more sophisticated approches

Perspectives

1. Extension of framework to non-linear deformations
2. Extension of framework to more than one pair (E.G. repeated measures)



3. Results: Real Dataset

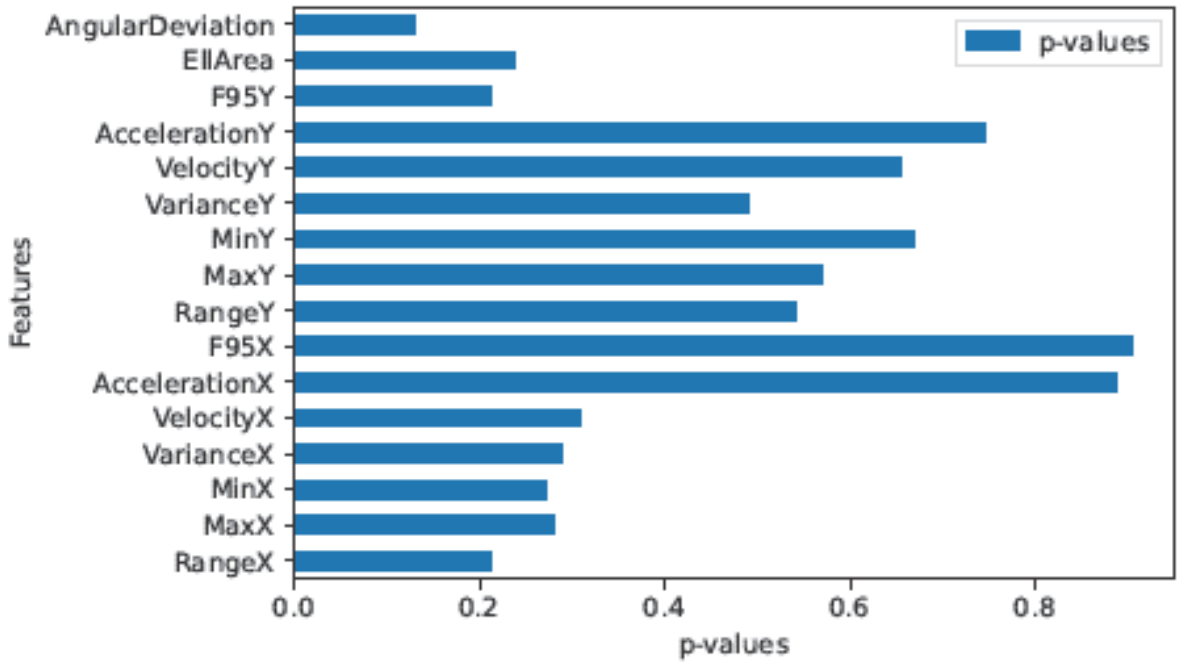
Real dataset

We extend our empirical validation by employing a typical real clinical dataset, with a relatively low population and multiple features. It concerns posturographic assessment for subjects with Parkinson's syndrome (PS). This dataset, initially used in [3], includes 30 subjects (mean age: 79.6 ± 4.4 years) from the Neurology department of the HIA, Percy hospital in Clamart, France, who were diagnosed with PS. The subjects underwent a posturography assessment using a force platform (here a Wii Balance Board (Nintendo, Kyoto, Japan)) that captures the trajectory of the center of pressure (CoP) exerted by the entire body over time when an individual stands on them. The assessment comprises two examinations with a 6-month difference in time, which are the paired samples we use in our experiment. Each time their postural stability was recorded for 25 seconds while maintaining an upright position on a force platform with eyes open.

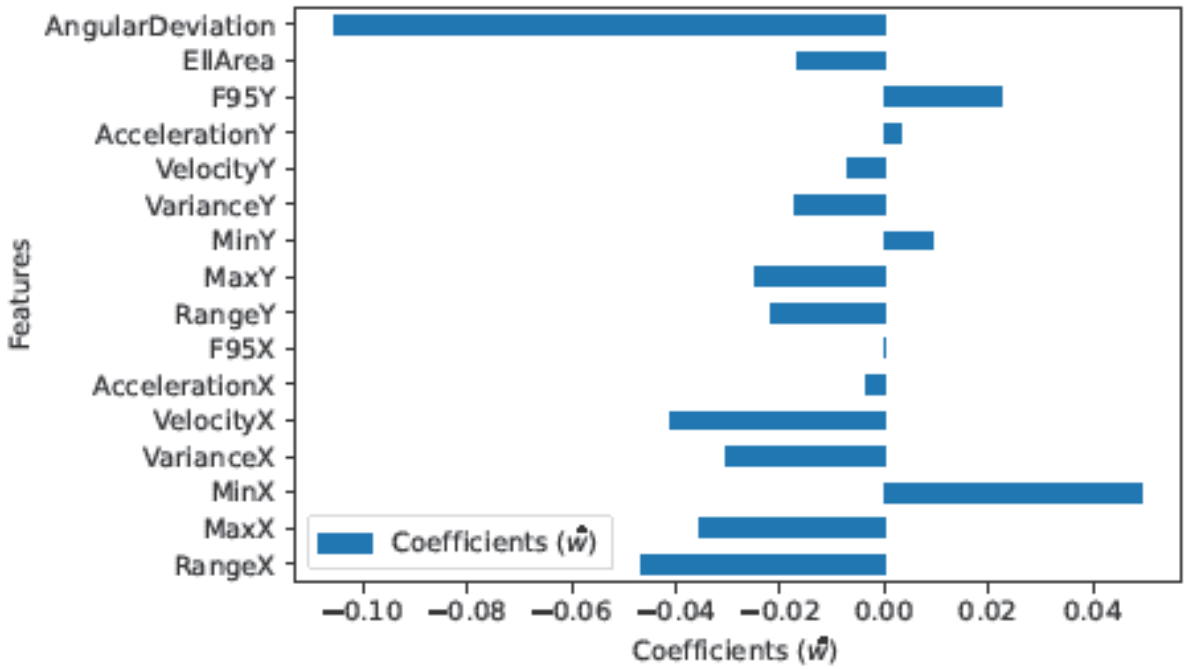
To characterize subjects' postural control, the dataset provides 16 features that had been previously proposed as indicators of postural stability [20]. In detail: Percentiles (95% and 5%) (cm), Range (cm), Variance (cm^2), Mean Instant Velocity (cm/s), Acceleration (cm/s^2) and Frequency (Hz) below which 95% of the signal energy is found, for both X-medio-lateral (ML) and Y-antero-posterior (AP) axes, confidence ellipse area (cm^2) that covers the 95% of the points of the trajectory and the angular deviation (in degrees°).

3. **Results:** Real Dataset

MT



MWSR



The relative importance of each feature as indicated by MT and MWSR on the posturographic dataset (D = 16 features)

Thank you for your attention

CENTRE BORELLI

école normale supérieure paris–saclay

université PARIS-SACLAY

Proof (Hodges-Lehmann estimator and WSR statistic association)

Proof: Suppose we have continuous data $\mathrm{Z} = (Z_1, \dots, Z_n)$, assumed to be symmetric around the mean μ. We consider the hypothesis $H_0: \mu_0 = \mu$. The signed rank statistic is:

$$T^+(Z) = \sum_{i=1}^{n} s(Z_i) R_i \qquad \textbf{(1)}$$

Where $s(Z_i) = \mathbb{I}(Z_1 > \mu_0)$

And R_i is the rank of $|Z_i - \mu_0|$ in $\{|Z_1 - \mu_0|, \dots, |Z_n - \mu_0|\}$. That is,

$$R_i = \sum_{j=1}^{n} \mathbb{I}\left(\left|Z_j - \mu_0\right| \le |Z_i - \mu_0|\right) \qquad \textbf{(2)}$$

Now , from **(1), (2)**

$$T^+(Z) = \sum_{i=1}^{n} (R_i, Z_i > \mu_0) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{I}\left(\left|Z_j - \mu_0\right| \le |Z_i - \mu_0|, Z_i > \mu_0\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{I}\left(\left|Z_j - \mu_0\right| \le Z_i - \mu_0, Z_i > \mu_0\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{I}\left(\left|Z_j - \mu_0\right| \le Z_i - \mu_0\right) \qquad \textbf{(3)}$$

We can remove dependence on the $Z_1 > \mu_0$ as the remaining indicator function will return zero if $Z_1 < \mu_0$. From **(3),**

$$T^+(Z) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{I}\left(\mu_0 - Z_i \le Z_j - \mu_0 \le Z_i - \mu_0\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{I}\left(2\mu_0 \le Z_j + Z_i \le 2Z_i\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{I}\left(\frac{Z_j + Z_i}{2} \ge \mu_0, Z_j \le Z_i\right) = \sum_{1 \le i \le j \le n} \mathbb{I}\left(\frac{Z_j + Z_i}{2} \ge \mu_0\right)$$