A Maximum Wilcoxon-Mann-Whitney Test

In High Dimensions

Lukas Mödl

Frank Konietschke

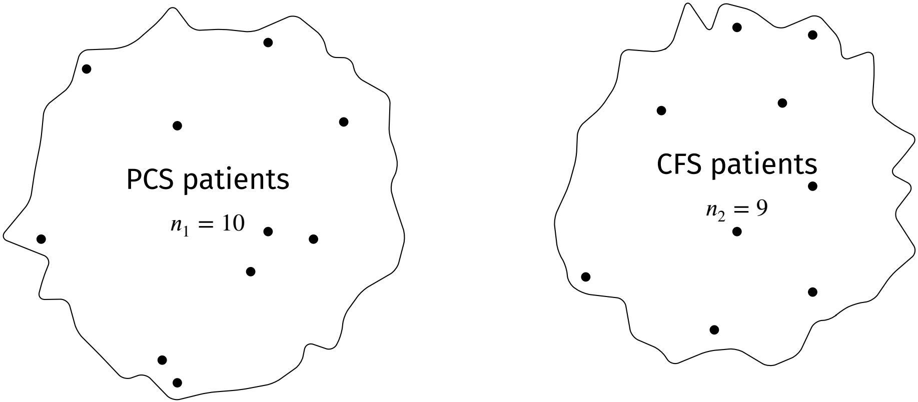
Symposium Statistical Planning of Translational Studies

Mar 21, 2024

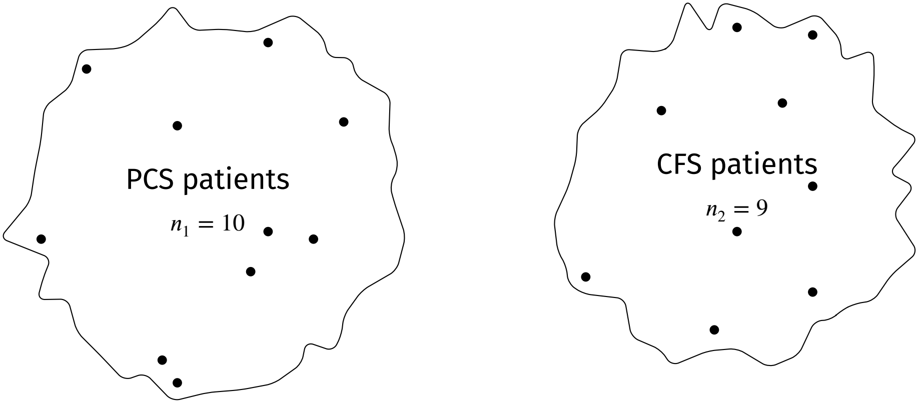
Example:

Post Covid Syndrome (PCS) and Chronique Fatigue Syndrome (CFS)

SETTING



SETTING

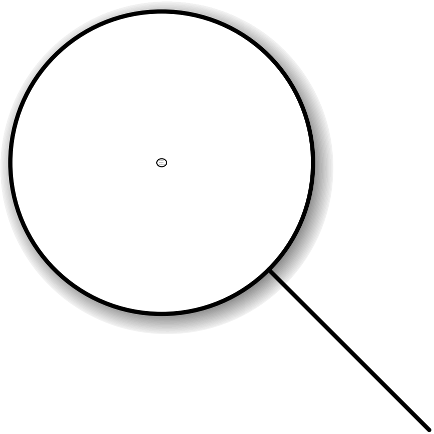


→ Goal: compare both groups

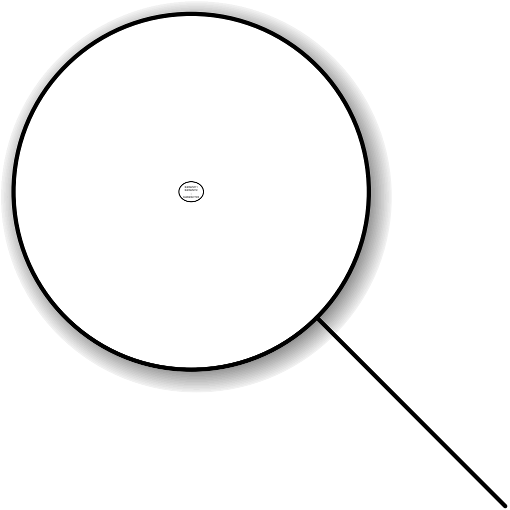
A CLOSER LOOK

•

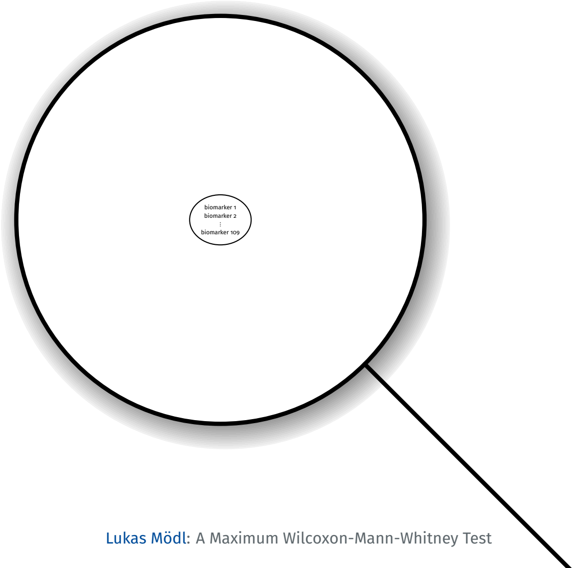
A CLOSER LOOK



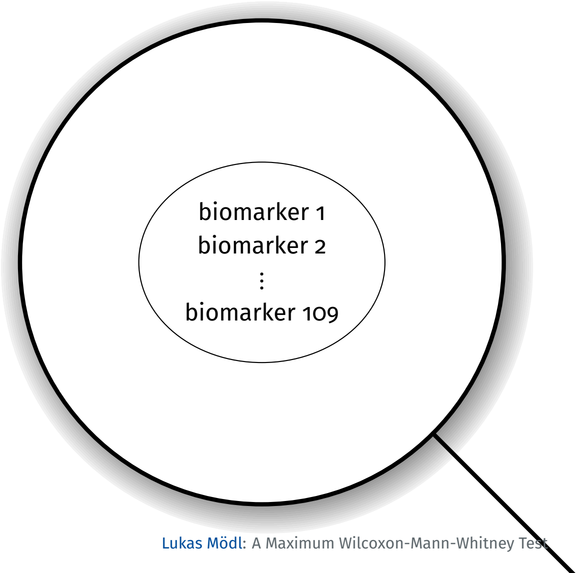
A CLOSER LOOK



A CLOSER LOOK



A CLOSER LOOK



COMPARING THE GROUPS

Test the global hypothesis

$$\text{distribution of biomarker 1 in PCS patients} = \text{distribution of biomarker 1 in CFS patients}$$
$$\text{distribution of biomarker 2 in PCS patients} = \text{distribution of biomarker 2 in CFS patients}$$
$$\vdots$$
$$\text{distribution of biomarker 109 in PCS patients} = \text{distribution of biomarker 109 in CFS patients}$$

at a type 1 error rate α

How to test this hypothesis?

COMMON APPROACHES

- Bonferroni-type test
  - → Run 109 separate Wilcoxon-Mann-Whitney tests with adjusted type 1 error rate α
  - → Inference is based on the univarite standard normal distribution

COMMON APPROACHES

- Bonferroni-type test
  - → Run 109 separate Wilcoxon-Mann-Whitney tests with adjusted type 1 error rate α
  - → Inference is based on the univarite standard normal distribution

- ANOVA-type test
  - → Run 109 separate Wilcoxon-Mann-Whitney tests and take the (weighted, squared) ℓ^2-norm
  - → Inference is based on the centered 109-variate normal distribution with the (scaled) empirical correlation matrix of ranks as variance-covariance matrix

COMMON APPROACHES

- Bonferroni-type test
  - → Run 109 separate Wilcoxon-Mann-Whitney tests with adjusted type 1 error rate α
  - → Inference is based on the univarite standard normal distribution
- ANOVA-type test
  - → Run 109 separate Wilcoxon-Mann-Whitney tests and take the (weighted, squared) ℓ^2-norm
  - → Inference is based on the centered 109-variate normal distribution with the (scaled) empirical correlation matrix of ranks as variance-covariance matrix
- MCTP-type test
  - → Run 109 separate Wilcoxon-Mann-Whitney tests and take the ℓ^∞-norm
  - → Inference is based on the centered 109-variate normal distribution with the empirical correlation matrix of ranks as variance-covariance matrix

TYPE 1 ERROR RATE CONTROL

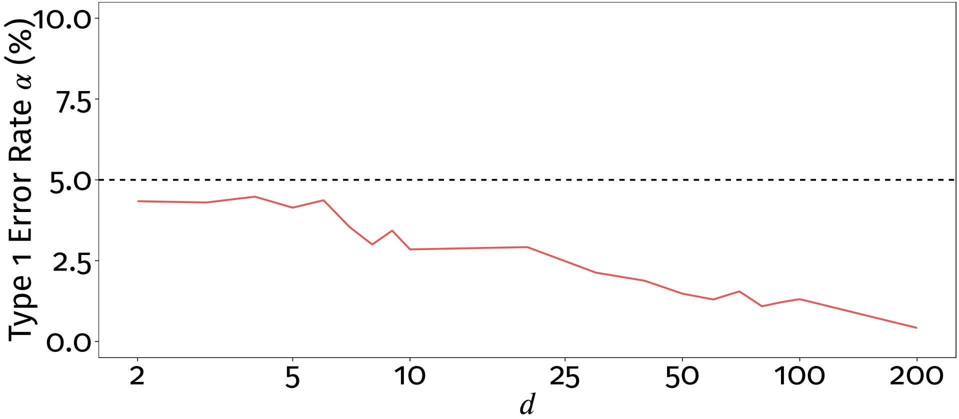


Fig. 1: Simulated data from multivariate normal distribution with $n_1 = 10$, $n_2 = 9$; $10{,}000$ runs for each d

Can we do something about it?

Can we do something about it?

$\rightsquigarrow$ Wild Bootstrap

STEP 1

Data matrix:

$$\begin{matrix} X_{111} & X_{121} & \cdots & X_{1d1} \\ X_{112} & X_{122} & \cdots & X_{1d2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{11n_1} & X_{12n_1} & \cdots & X_{1dn_1} \\ X_{211} & X_{221} & \cdots & X_{2d1} \\ X_{212} & X_{222} & \cdots & X_{2d2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{21n_2} & X_{22n_2} & \cdots & X_{2dn_2} \end{matrix}$$

STEP 1

Assign ranks:

$$\begin{matrix} R_{111} & R_{121} & \cdots & R_{1d1} \\ R_{112} & R_{122} & \cdots & R_{1d2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{11n_1} & R_{12n_1} & \cdots & R_{1dn_1} \\ R_{211} & R_{221} & \cdots & R_{2d1} \\ R_{212} & R_{222} & \cdots & R_{2d2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{21n_2} & R_{22n_2} & \cdots & R_{2dn_2} \end{matrix}$$

STEP 2

Center ranks by rank mean $c := \frac{n_1+n_2+1}{2}$:

$$
\begin{matrix}
(R_{111}-c) & (R_{121}-c) & \cdots & (R_{1d1}-c) \\
(R_{112}-c) & (R_{122}-c) & \cdots & (R_{1d2}-c) \\
\vdots & \vdots & \ddots & \vdots \\
(R_{11n_1}-c) & (R_{12n_1}-c) & \cdots & (R_{1dn_1}-c) \\
(R_{211}-c) & (R_{221}-c) & \cdots & (R_{2d1}-c) \\
(R_{212}-c) & (R_{222}-c) & \cdots & (R_{2d2}-c) \\
\vdots & \vdots & \ddots & \vdots \\
(R_{21n_2}-c) & (R_{22n_2}-c) & \cdots & (R_{2dn_2}-c)
\end{matrix}
$$

STEP 3

Pertubate centered ranks of each row by a random sign W_{ik} (+ or –):

$$
\begin{matrix}
W_{11}(R_{111}-c) & W_{11}(R_{121}-c) & \cdots & W_{11}(R_{1d1}-c) \\
W_{12}(R_{112}-c) & W_{12}(R_{122}-c) & \cdots & W_{12}(R_{1d2}-c) \\
\vdots & \vdots & \ddots & \vdots \\
W_{1n_1}(R_{11n_1}-c) & W_{1n_1}(R_{12n_1}-c) & \cdots & W_{1n_1}(R_{1dn_1}-c) \\
W_{21}(R_{211}-c) & W_{21}(R_{221}-c) & \cdots & W_{21}(R_{2d1}-c) \\
W_{22}(R_{212}-c) & W_{22}(R_{222}-c) & \cdots & W_{22}(R_{2d2}-c) \\
\vdots & \vdots & \ddots & \vdots \\
W_{2n_2}(R_{21n_2}-c) & W_{2n_2}(R_{22n_2}-c) & \cdots & W_{2n_2}(R_{2dn_2}-c)
\end{matrix}
$$

STEP 4

Compute column means for both groups:

$$\begin{matrix}
W_{11}(R_{111} - c) & W_{11}(R_{121} - c) & \cdots & W_{11}(R_{1d1} - c) \\
W_{12}(R_{112} - c) & W_{12}(R_{122} - c) & \cdots & W_{12}(R_{1d2} - c) \\
\vdots & \vdots & \ddots & \vdots \\
W_{1n_1}(R_{11n_1} - c) & W_{1n_1}(R_{12n_1} - c) & \cdots & W_{1n_1}(R_{1dn_1} - c) \\
W_{21}(R_{211} - c) & W_{21}(R_{221} - c) & \cdots & W_{21}(R_{2d1} - c) \\
W_{22}(R_{212} - c) & W_{22}(R_{222} - c) & \cdots & W_{22}(R_{2d2} - c) \\
\vdots & \vdots & \ddots & \vdots \\
W_{2n_2}(R_{21n_2} - c) & W_{2n_2}(R_{22n_2} - c) & \cdots & W_{2n_2}(R_{2dn_2} - c)
\end{matrix}$$

STEP 4

Compute column means for both groups:

$$
\begin{matrix}
W_{11}(R_{111}-c) & W_{11}(R_{121}-c) & \cdots & W_{11}(R_{1d1}-c) \\
W_{12}(R_{112}-c) & W_{12}(R_{122}-c) & \cdots & W_{12}(R_{1d2}-c) \\
\vdots & \vdots & \ddots & \vdots \\
W_{1n_1}(R_{11n_1}-c) & W_{1n_1}(R_{12n_1}-c) & \cdots & W_{1n_1}(R_{1dn_1}-c) \\
W_{21}(R_{211}-c) & W_{21}(R_{221}-c) & \cdots & W_{21}(R_{2d1}-c) \\
W_{22}(R_{212}-c) & W_{22}(R_{222}-c) & \cdots & W_{22}(R_{2d2}-c) \\
\vdots & \vdots & \ddots & \vdots \\
W_{2n_2}(R_{21n_2}-c) & W_{2n_2}(R_{22n_2}-c) & \cdots & W_{2n_2}(R_{2dn_2}-c)
\end{matrix}
$$

M_{11} M_{12} M_{1d}

M_{21} M_{22} M_{2d}

STEP 5

Compute standard deviations column-wise:

$$
\begin{matrix}
W_{11}(R_{111}-c) & W_{11}(R_{121}-c) & \cdots & W_{11}(R_{1d1}-c) \\
W_{12}(R_{112}-c) & W_{12}(R_{122}-c) & \cdots & W_{12}(R_{1d2}-c) \\
\vdots & \vdots & \ddots & \vdots \\
W_{1n_1}(R_{11n_1}-c) & W_{1n_1}(R_{12n_1}-c) & \cdots & W_{1n_1}(R_{1dn_1}-c) \\
W_{21}(R_{211}-c) & W_{21}(R_{221}-c) & \cdots & W_{21}(R_{2d1}-c) \\
W_{22}(R_{212}-c) & W_{22}(R_{222}-c) & \cdots & W_{22}(R_{2d2}-c) \\
\vdots & \vdots & \ddots & \vdots \\
W_{2n_2}(R_{21n_2}-c) & W_{2n_2}(R_{22n_2}-c) & \cdots & W_{2n_2}(R_{2dn_2}-c)
\end{matrix}
$$

STEP 5

Compute standard deviations column-wise:

$$
\begin{array}{cccc}
W_{11}(R_{111}-c) & W_{11}(R_{121}-c) & \cdots & W_{11}(R_{1d1}-c) \\
W_{12}(R_{112}-c) & W_{12}(R_{122}-c) & \cdots & W_{12}(R_{1d2}-c) \\
\vdots & \vdots & \ddots & \vdots \\
W_{1n_1}(R_{11n_1}-c) & W_{1n_1}(R_{12n_1}-c) & \cdots & W_{1n_1}(R_{1dn_1}-c) \\
W_{21}(R_{211}-c) & W_{21}(R_{221}-c) & \cdots & W_{21}(R_{2d1}-c) \\
W_{22}(R_{212}-c) & W_{22}(R_{222}-c) & \cdots & W_{22}(R_{2d2}-c) \\
\vdots & \vdots & \ddots & \vdots \\
W_{2n_2}(R_{21n_2}-c) & W_{2n_2}(R_{22n_2}-c) & \cdots & W_{2n_2}(R_{2dn_2}-c) \\
\downarrow S_1 & \downarrow S_2 & & \downarrow S_d
\end{array}
$$

STEP 6

Repeat steps 3 to 5 B times and obtain the empirical p-value by the proportion of

$$T_0 < T_0^*,$$

where

- $T_0 := \max\left\{|T_1|, |T_2|, \dots, |T_d|\right\}$ with T_j being the usual Wilcoxon-Mann-Whitney test statistic for testing the local hypothesis

$$\text{distribution of outcome } j \text{ in group 1} = \text{distribution of outcome } j \text{ in group 2}$$

- $T_0^* := \max\left\{|T_1^*|, |T_2^*|, \dots, |T_d^*|\right\}$ with

$$T_j^* := \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{M_{1j} - M_{2j}}{S_j}$$

TYPE 1 ERROR RATE CONTROL REVISITED

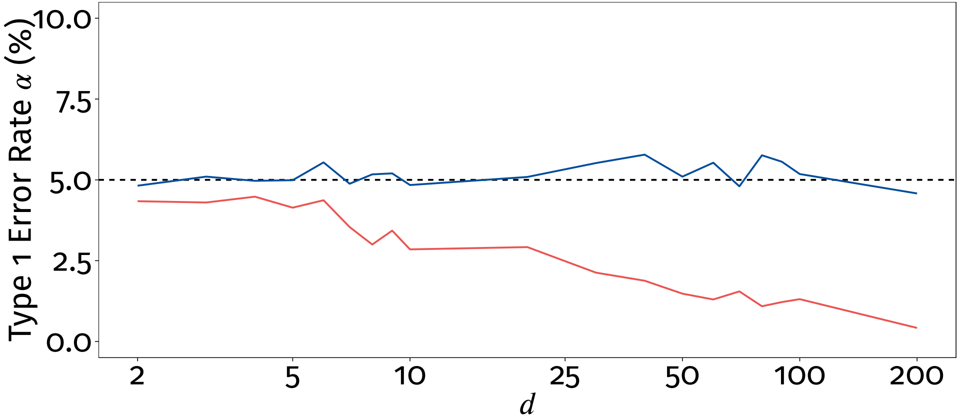


Fig. 2: Same setting as for Fig. 1; $b = 10{,}000$ were used for the wild bootstrap approach

Conclusion

CONCLUSION

Summary:

- The wild bootstrap approach controls type 1 error rate α quite well
  → even in high dimensions

CONCLUSION

Summary:

- The wild bootstrap approach controls type 1 error rate α quite well
  → even in high dimensions

Outlook:

- Extend to multiple groups
- Consider other tests (e.g., Brunner-Munzel test)
- Consider other norms
- ...

Discussion

References

REFERENCES I

Konietschke F, Schwab K, Pauly M. Small sample sizes: A big data problem in high-dimensional data analysis. *Statistical Methods in Medical Research*. 2021;30(3):687-701.

Brunner E, Bathke A, Konietschke F. *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs: Using R and SAS*. Springer International Publishing. 2019.

Vaart, A. W. van der. *Asymptotic Statistics*. Cambridge University Press. 2000.

Appendix

APPENDIX A

Theorem 1 – Consistency of Wild Bootstrap Approach

Let Q_n denote the joint distribution of $(T_1, T_2, \dots, T_d)$, Q_n^* the joint distribution of $(T_1^*, T_2^*, \dots, T_d^*)$ given the data, and suppose that $\frac{n_1+n_2}{n_i} \to \lambda_i$ with $0 < \lambda_i < 1$ for $i = 1, 2$ as $n_1 \wedge n_2 \to \infty$. Then

$$\varrho(Q_n, Q_n^*) \to 0$$

in probability as $n_1 \wedge n_2 \to \infty$, where ϱ denotes the Lévy-Prokhorov metric on the space of probability measures defined on the usual Borel σ-algebra over $\mathbb{R}^d$.