# A Primer on Permutation Tests (not only) for MVPA

Joram Soch

( with slides by Carsten Allefeld )

Bernstein Center for Computational Neuroscience, Berlin, Germany
German Center for Neurodegenerative Diseases, Göttingen, Germany
E-Mail: joram.soch@bccn-berlin.de, Joram.Soch@DZNE.de

# introduction

why use a permutation test?

- sometimes precise distributions are **not known**, especially in MVPA
- a permutation test makes **weaker assumptions** about distributions than parametric tests
- permutation tests provide **exact inference**
- permutation testing applies **in the same way** to univariate and multivariate analysis

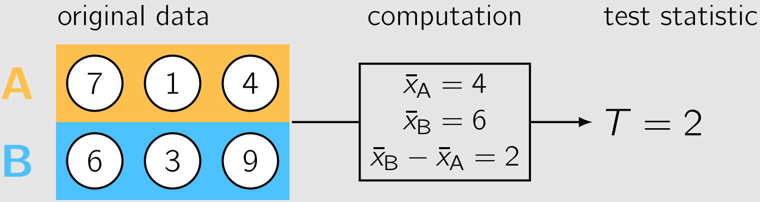
this talk

- no recipes, but the **basic principles** underlying permutation tests, especially exchangeability

# A simple example

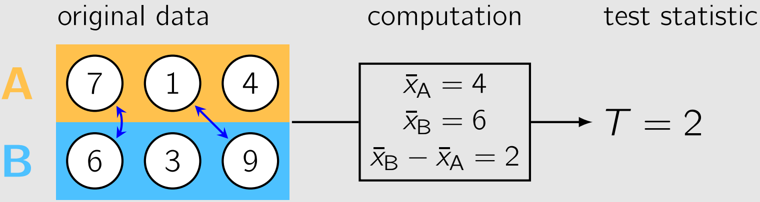
# a univariate example: two-sample test

is there a mean difference between groups A and B?



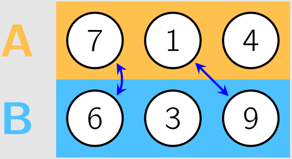
# two-sample test

if not, values can be **exchanged** between conditions

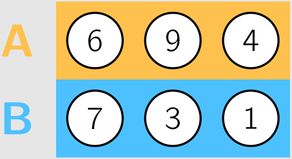


# two-sample test

the same test statistic is computed from **permuted data**

| original data | computation | test statistic |
|---|---|---|
|  | $\bar{x}_A = 4$<br>$\bar{x}_B = 6$<br>$\bar{x}_B - \bar{x}_A = 2$ | $T = 2$ |



| permuted data | computation | test statistic |
|---|---|---|
|  | $\bar{x}_A = 6.33$<br>$\bar{x}_B = 3.67$<br>$\bar{x}_B - \bar{x}_A = -2.67$ | $T = -2.67$ |

# two-sample test

compute the test statistic for **all permutations**

| perm | data (A) | data (B) | $T = \bar{x}_B - \bar{x}_A$ |
|---|---|---|---|
| 1 | 7 1 4 | 6 3 9 | 2.00 |
| 2 | 7 1 6 | 4 3 9 | 0.67 |
| 3 | 7 1 3 | 4 6 9 | 2.67 |
| 4 | 7 1 9 | 4 6 3 | -1.33 |
| 5 | 7 4 6 | 1 3 9 | -1.33 |
| 6 | 7 4 3 | 1 6 9 | 0.67 |
| 7 | 7 4 9 | 1 6 3 | -3.33 |
| 8 | 7 6 3 | 1 4 9 | -0.67 |
| 9 | 7 6 9 | 1 4 3 | -4.67 |
| 10 | 7 3 9 | 1 4 6 | -2.67 |
| 11 | 1 4 6 | 7 3 9 | 2.67 |
| 12 | 1 4 3 | 7 6 9 | 4.67 |
| 13 | 1 4 9 | 7 6 3 | 0.67 |
| 14 | 1 6 3 | 7 4 9 | 3.33 |
| 15 | 1 6 9 | 7 4 3 | -0.67 |
| 16 | 1 3 9 | 7 4 6 | 1.33 |
| 17 | 4 6 3 | 7 1 9 | 1.33 |
| 18 | 4 6 9 | 7 1 3 | -2.67 |
| 19 | 4 3 9 | 7 1 6 | -0.67 |
| 20 | 6 3 9 | 7 1 4 | -2.00 |

# two-sample test

determine **ranks** by sorting in descending order

| perm | data (A) | data (B) | $T = \bar{x}_B - \bar{x}_A$ | sorted | rank |
|---|---|---|---|---|---|
| 1 | 7 1 4 | 6 3 9 | 2.00 | 4.67 | 1 |
| 2 | 7 1 6 | 4 3 9 | 0.67 | 3.33 | 2 |
| 3 | 7 1 3 | 4 6 9 | 2.67 | 2.67 | 3 |
| 4 | 7 1 9 | 4 6 3 | -1.33 | 2.67 | 4 |
| 5 | 7 4 6 | 1 3 9 | -1.33 | 2.00 | 5 |
| 6 | 7 4 3 | 1 6 9 | 0.67 | 1.33 | 6 |
| 7 | 7 4 9 | 1 6 3 | -3.33 | 1.33 | 7 |
| 8 | 7 6 3 | 1 4 9 | -0.67 | 0.67 | 8 |
| 9 | 7 6 9 | 1 4 3 | -4.67 | 0.67 | 9 |
| 10 | 7 3 9 | 1 4 6 | -2.67 | 0.67 | 10 |
| 11 | 1 4 6 | 7 3 9 | 2.67 | -0.67 | 11 |
| 12 | 1 4 3 | 7 6 9 | 4.67 | -0.67 | 12 |
| 13 | 1 4 9 | 7 6 3 | 0.67 | -0.67 | 13 |
| 14 | 1 6 3 | 7 4 9 | 3.33 | -1.33 | 14 |
| 15 | 1 6 9 | 7 4 3 | -0.67 | -1.33 | 15 |
| 16 | 1 3 9 | 7 4 6 | 1.33 | -2.00 | 16 |
| 17 | 4 6 3 | 7 1 9 | 1.33 | -2.67 | 17 |
| 18 | 4 6 9 | 7 1 3 | -2.67 | -2.67 | 18 |
| 19 | 4 3 9 | 7 1 6 | -0.67 | -3.33 | 19 |
| 20 | 6 3 9 | 7 1 4 | -2.00 | -4.67 | 20 |

# two-sample test

the **neutral permutation** is part of the set

| perm | data | | $T = \bar{x}_B - \bar{x}_A$ | sorted | rank |
|---|---|---|---|---|---|
| 1 | 7 1 4 | 6 3 9 | 2.00 | 4.67 | 1 |
| 2 | 7 1 6 | 4 3 9 | 0.67 | 3.33 | 2 |
| 3 | 7 1 3 | 4 6 9 | 2.67 | 2.67 | 3 |
| 4 | 7 1 9 | 4 6 3 | -1.33 | 2.67 | 4 |
| 5 | 7 4 6 | 1 3 9 | -1.33 | 2.00 | 5 |
| 6 | 7 4 3 | 1 6 9 | 0.67 | 1.33 | 6 |
| 7 | 7 4 9 | 1 6 3 | -3.33 | 1.33 | 7 |
| 8 | 7 6 3 | 1 4 9 | -0.67 | 0.67 | 8 |
| 9 | 7 6 9 | 1 4 3 | -4.67 | 0.67 | 9 |
| 10 | 7 3 9 | 1 4 6 | -2.67 | 0.67 | 10 |
| 11 | 1 4 6 | 7 3 9 | 2.67 | -0.67 | 11 |
| 12 | 1 4 3 | 7 6 9 | 4.67 | -0.67 | 12 |
| 13 | 1 4 9 | 7 6 3 | 0.67 | -0.67 | 13 |
| 14 | 1 6 3 | 7 4 9 | 3.33 | -1.33 | 14 |
| 15 | 1 6 9 | 7 4 3 | -0.67 | -1.33 | 15 |
| 16 | 1 3 9 | 7 4 6 | 1.33 | -2.00 | 16 |
| 17 | 4 6 3 | 7 1 9 | 1.33 | -2.67 | 17 |
| 18 | 4 6 9 | 7 1 3 | -2.67 | -2.67 | 18 |
| 19 | 4 3 9 | 7 1 6 | -0.67 | -3.33 | 19 |
| 20 | 6 3 9 | 7 1 4 | -2.00 | -4.67 | 20 |

# two-sample test

the rank of the original value determines **significance**

| perm | data (A) | data (B) | $T = \bar{x}_B - \bar{x}_A$ | sorted | rank | |
|---|---|---|---|---|---|---|
| 1 | 7 1 4 | 6 3 9 | 2.00 | 4.67 | 1 | significant |
| 2 | 7 1 6 | 4 3 9 | 0.67 | 3.33 | 2 | |
| 3 | 7 1 3 | 4 6 9 | 2.67 | 2.67 | 3 | |
| 4 | 7 1 9 | 4 6 3 | -1.33 | 2.67 | 4 | |
| 5 | 7 4 6 | 1 3 9 | -1.33 | 2.00 | 5 | |
| 6 | 7 4 3 | 1 6 9 | 0.67 | 1.33 | 6 | |
| 7 | 7 4 9 | 1 6 3 | -3.33 | 1.33 | 7 | |
| 8 | 7 6 3 | 1 4 9 | -0.67 | 0.67 | 8 | |
| 9 | 7 6 9 | 1 4 3 | -4.67 | 0.67 | 9 | |
| 10 | 7 3 9 | 1 4 6 | -2.67 | 0.67 | 10 | |
| 11 | 1 4 6 | 7 3 9 | 2.67 | -0.67 | 11 | |
| 12 | 1 4 3 | 7 6 9 | 4.67 | -0.67 | 12 | |
| 13 | 1 4 9 | 7 6 3 | 0.67 | -0.67 | 13 | |
| 14 | 1 6 3 | 7 4 9 | 3.33 | -1.33 | 14 | |
| 15 | 1 6 9 | 7 4 3 | -0.67 | -1.33 | 15 | |
| 16 | 1 3 9 | 7 4 6 | 1.33 | -2.00 | 16 | |
| 17 | 4 6 3 | 7 1 9 | 1.33 | -2.67 | 17 | |
| 18 | 4 6 9 | 7 1 3 | -2.67 | -2.67 | 18 | |
| 19 | 4 3 9 | 7 1 6 | -0.67 | -3.33 | 19 | |
| 20 | 6 3 9 | 7 1 4 | -2.00 | -4.67 | 20 | |

# two-sample test

equivalently, a **p-value** can be computed

| perm | data (A) | data (B) | $T = \bar{x}_B - \bar{x}_A$ | sorted | rank | |
|---|---|---|---|---|---|---|
| 1 | 7 1 4 | 6 3 9 | 2.00 | 4.67 | 1 | significant |
| 2 | 7 1 6 | 4 3 9 | 0.67 | 3.33 | 2 | |
| 3 | 7 1 3 | 4 6 9 | 2.67 | 2.67 | 3 | |
| 4 | 7 1 9 | 4 6 3 | -1.33 | 2.67 | 4 | |
| 5 | 7 4 6 | 1 3 9 | -1.33 | 2.00 | 5 | $p = \frac{5}{20} = 0.25$ |
| 6 | 7 4 3 | 1 6 9 | 0.67 | 1.33 | 6 | |
| 7 | 7 4 9 | 1 6 3 | -3.33 | 1.33 | 7 | |
| 8 | 7 6 3 | 1 4 9 | -0.67 | 0.67 | 8 | |
| 9 | 7 6 9 | 1 4 3 | -4.67 | 0.67 | 9 | |
| 10 | 7 3 9 | 1 4 6 | -2.67 | 0.67 | 10 | |
| 11 | 1 4 6 | 7 3 9 | 2.67 | -0.67 | 11 | |
| 12 | 1 4 3 | 7 6 9 | 4.67 | -0.67 | 12 | |
| 13 | 1 4 9 | 7 6 3 | 0.67 | -0.67 | 13 | |
| 14 | 1 6 3 | 7 4 9 | 3.33 | -1.33 | 14 | |
| 15 | 1 6 9 | 7 4 3 | -0.67 | -1.33 | 15 | |
| 16 | 1 3 9 | 7 4 6 | 1.33 | -2.00 | 16 | |
| 17 | 4 6 3 | 7 1 9 | 1.33 | -2.67 | 17 | |
| 18 | 4 6 9 | 7 1 3 | -2.67 | -2.67 | 18 | |
| 19 | 4 3 9 | 7 1 6 | -0.67 | -3.33 | 19 | |
| 20 | 6 3 9 | 7 1 4 | -2.00 | -4.67 | 20 | |

# two-sample test

test for **decrease**: use the negative test statistic

| perm | data (A) | data (B) | $T = \bar{x}_A - \bar{x}_B$ | sorted | rank | |
|---|---|---|---|---|---|---|
| 1 | 7 1 4 | 6 3 9 | -2.00 | 4.67 | 1 | significant |
| 2 | 7 1 6 | 4 3 9 | -0.67 | 3.33 | 2 | |
| 3 | 7 1 3 | 4 6 9 | -2.67 | 2.67 | 3 | |
| 4 | 7 1 9 | 4 6 3 | 1.33 | 2.67 | 4 | |
| 5 | 7 4 6 | 1 3 9 | 1.33 | 2.00 | 5 | |
| 6 | 7 4 3 | 1 6 9 | -0.67 | 1.33 | 6 | |
| 7 | 7 4 9 | 1 6 3 | 3.33 | 1.33 | 7 | |
| 8 | 7 6 3 | 1 4 9 | 0.67 | 0.67 | 8 | |
| 9 | 7 6 9 | 1 4 3 | 4.67 | 0.67 | 9 | |
| 10 | 7 3 9 | 1 4 6 | 2.67 | 0.67 | 10 | |
| 11 | 1 4 6 | 7 3 9 | -2.67 | -0.67 | 11 | |
| 12 | 1 4 3 | 7 6 9 | -4.67 | -0.67 | 12 | |
| 13 | 1 4 9 | 7 6 3 | -0.67 | -0.67 | 13 | |
| 14 | 1 6 3 | 7 4 9 | -3.33 | -1.33 | 14 | |
| 15 | 1 6 9 | 7 4 3 | 0.67 | -1.33 | 15 | |
| 16 | 1 3 9 | 7 4 6 | -1.33 | -2.00 | 16 | $p = \frac{16}{20} = 0.80$ |
| 17 | 4 6 3 | 7 1 9 | -1.33 | -2.67 | 17 | |
| 18 | 4 6 9 | 7 1 3 | 2.67 | -2.67 | 18 | |
| 19 | 4 3 9 | 7 1 6 | 0.67 | -3.33 | 19 | |
| 20 | 6 3 9 | 7 1 4 | 2.00 | -4.67 | 20 | |

# two-sample test

**two-sided test**: use the absolute value of the test statistic

| perm | data | | $T = \lvert \bar{x}_B - \bar{x}_A \rvert$ |
|---|---|---|---|
| 1 | 7 1 4 | 6 3 9 | 2.00 |
| 2 | 7 1 6 | 4 3 9 | 0.67 |
| 3 | 7 1 3 | 4 6 9 | 2.67 |
| 4 | 7 1 9 | 4 6 3 | 1.33 |
| 5 | 7 4 6 | 1 3 9 | 1.33 |
| 6 | 7 4 3 | 1 6 9 | 0.67 |
| 7 | 7 4 9 | 1 6 3 | 3.33 |
| 8 | 7 6 3 | 1 4 9 | 0.67 |
| 9 | 7 6 9 | 1 4 3 | 4.67 |
| 10 | 7 3 9 | 1 4 6 | 2.67 |
| 11 | 1 4 6 | 7 3 9 | 2.67 |
| 12 | 1 4 3 | 7 6 9 | 4.67 |
| 13 | 1 4 9 | 7 6 3 | 0.67 |
| 14 | 1 6 3 | 7 4 9 | 3.33 |
| 15 | 1 6 9 | 7 4 3 | 0.67 |
| 16 | 1 3 9 | 7 4 6 | 1.33 |
| 17 | 4 6 3 | 7 1 9 | 1.33 |
| 18 | 4 6 9 | 7 1 3 | 2.67 |
| 19 | 4 3 9 | 7 1 6 | 0.67 |
| 20 | 6 3 9 | 7 1 4 | 2.00 |

| sorted | rank | |
|---|---|---|
| 4.67 | 1 | significant |
| 4.67 | 2 | |
| 3.33 | 3 | |
| 3.33 | 4 | |
| 2.67 | 5 | |
| 2.67 | 6 | |
| 2.67 | 7 | |
| 2.67 | 8 | |
| 2.00 | 9 | |
| 2.00 | 10 | $p = \frac{10}{20} = 0.50$ |
| 1.33 | 11 | |
| 1.33 | 12 | |
| 1.33 | 13 | |
| 1.33 | 14 | |
| 0.67 | 15 | |
| 0.67 | 16 | |
| 0.67 | 17 | |
| 0.67 | 18 | |
| 0.67 | 19 | |
| 0.67 | 20 | |

# formal procedure for a permutation test

given: data, test statistic, sig. level $\alpha$, **possible exchanges**

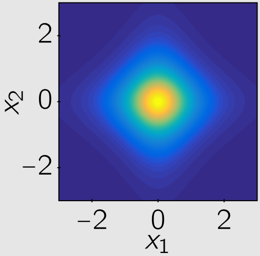
- compute test statistic $T_1$ for original data
- compute test statistic $T_i$ for all permutations $i = 2 \dots n_\mathrm{P}$
  (or a random selection $\longrightarrow$ 'Monte-Carlo')
  $i = 1$ is the 'neutral permutation'
- determine the rank of $T_1$: $$r = \sum_{i=1}^{n_\mathrm{P}} [T_i \geq T_1]$$
  where $[\text{true}] = 1$ and $[\text{false}] = 0$
- determine the $p$-value: $$p = \frac{r}{n_\mathrm{P}}$$
- if $p \leq \alpha$, reject $\mathrm{H}_0$

the test is exact if $\alpha$ is a multiple of $\frac{1}{n_\mathrm{P}}$ and there are no ties, otherwise it is conservative

# Exchangeability

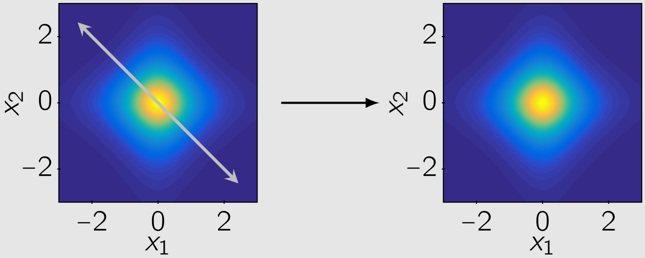
# exchangeability

**independent** samples from the **same distribution** (i.i.d.) ...



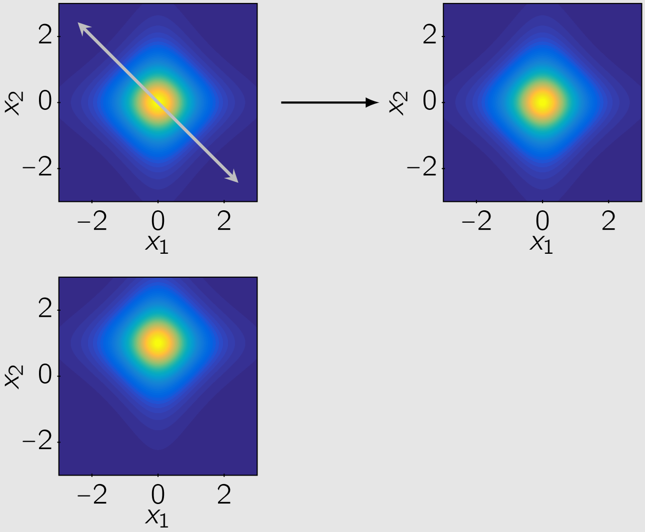
# exchangeability

... can be **exchanged**: $x_1 \leftrightarrow x_2$



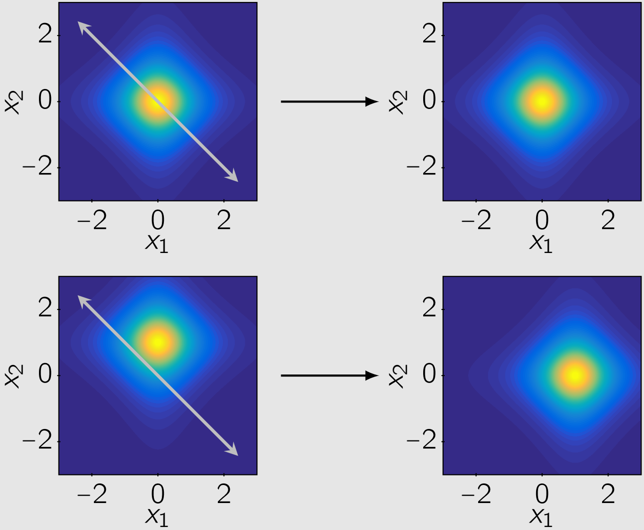
# exchangeability

**independent** samples from **different distributions** . . .



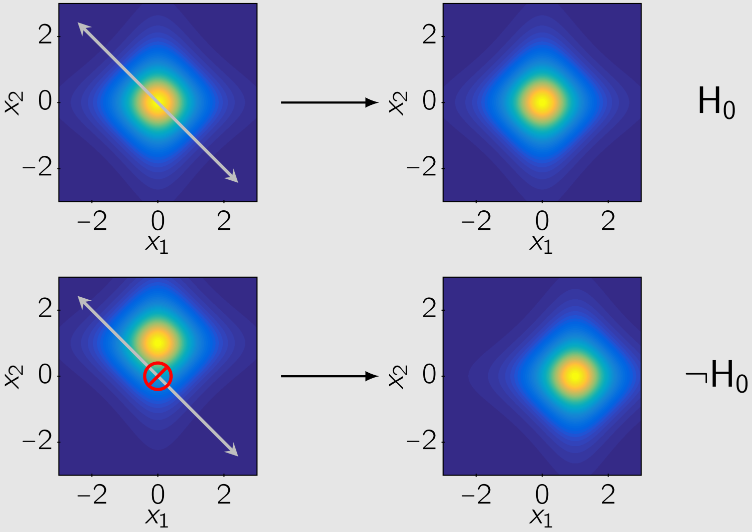
# exchangeability

. . . **cannot** be **exchanged**



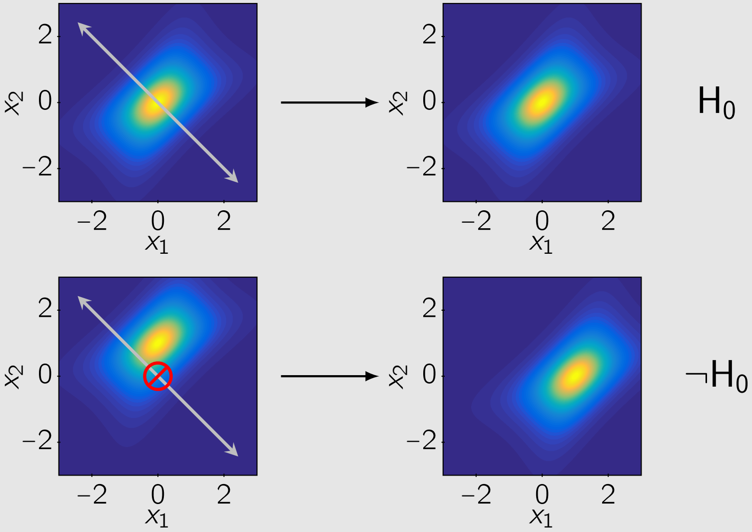
# exchangeability

exchangeability refers to the **null hypothesis**!

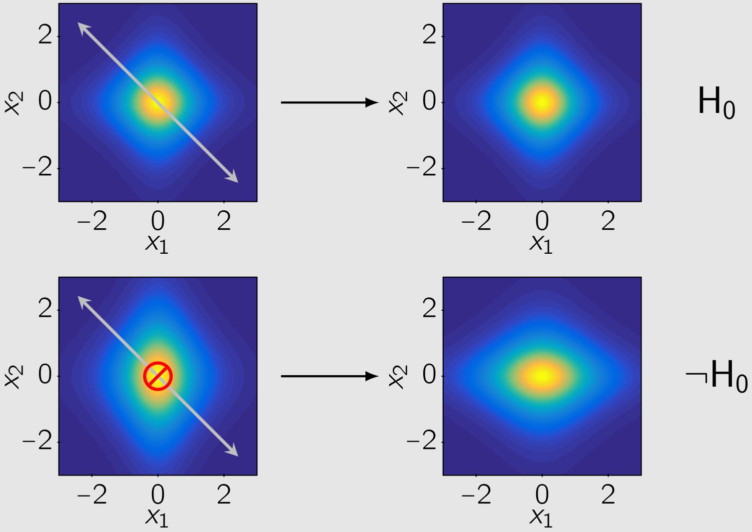


# exchangeability

exchangeable also if there is dependency but **symmetric**



**caution:** null hypothesis may be wrong **in unexpected ways**

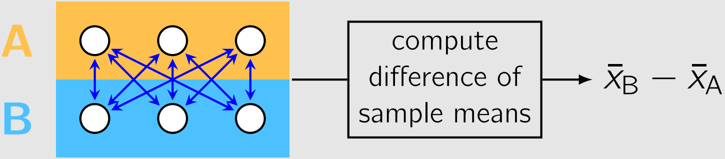


# Some test examples

# tests: univariate two-sample

independent values sampled from two univariate distributions

$H_0$: two samples come from the **same distribution**
$\longrightarrow$ i.i.d. case: values can be exchanged freely

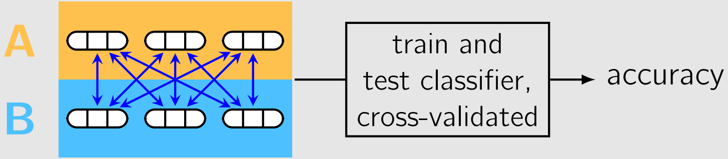



test statistic is sensitive to difference in means:
permutation analogue of the **two-sample t-test**

# tests: multivariate two-sample

independent vectors from two multivariate distributions

$H_0$: two samples from the **same multivariate distribution**
$\longrightarrow$ i.i.d. case: **vector values** can be exchanged freely



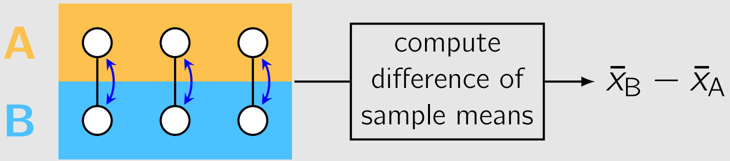

**Problem with accuracies:** few possible values $\longrightarrow$ ties!

**Example:** classification of subject-specific patterns,
e.g. of patients based on structural MRI

# tests: univariate paired

**paired** values sampled from two univariate distributions

$H_0$: two samples come from the same distribution
$\longrightarrow$ partial dependency: values can be exchanged **within pairs**

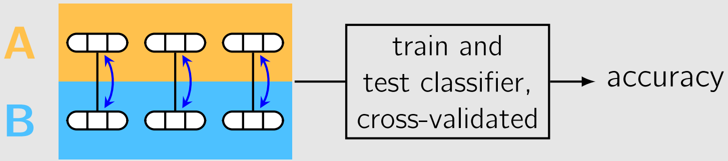



test statistic is sensitive to difference in means:
permutation analogue of the **paired t-test**

# tests: multivariate paired

paired vectors sampled from two multivariate distributions

$H_0$: two samples from the same multivariate distribution
$\longrightarrow$ partial dependency: vectors can be exchanged within pairs




**Example:**

– classification of condition-specific patterns across subjects
– classification of patterns across runs, within subject

# alternative to accuracy: multivariate test

multivariate samples can also be compared using significance tests ⟶ potentially **more powerful**

NeuroImage 146 (2017) 113–120

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

ELSEVIER


What's in a pattern? Examining the type of signal multivariate analysis uncovers at the group level☆

CrossMark

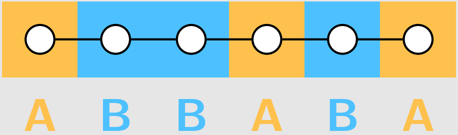
Roee Gilron[a,b,*], Jonathan Rosenblatt[c,d], Oluwasanmi Koyejo[e], Russell A. Poldrack[e], Roy Mukamel[a,b]

[a] *School of Psychological Sciences, Tel Aviv University, Tel Aviv, Israel*
[b] *Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel*
[c] *Zlotowsky Center for Neuroscience Ben Gurion University, Beer Sheva, Israel*
[d] *Department of Industrial Engineering and Management, Beer Sheva, Israel*
[e] *Department of Psychology, Stanford University, Stanford, CA, United States*

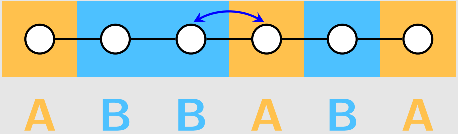
# Limited exchangeability

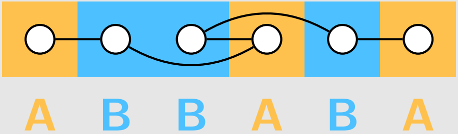
# time series

fMRI trials occur in a time series with **serial dependency**

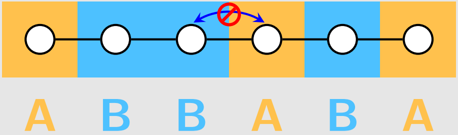


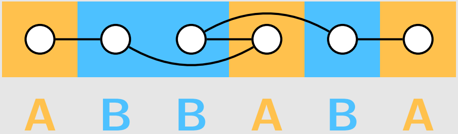
**problem:** exchanges change the **dependency structure**



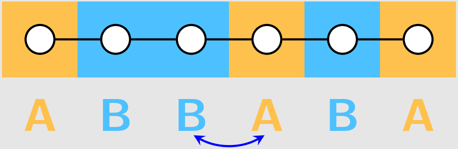
**no exchangeability** of values in fMRI time series!



# time series

for a **randomized** trial sequence, **labels** can be exchanged



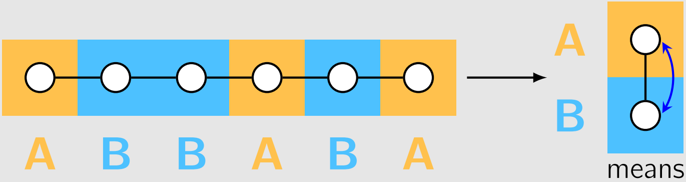

$\longrightarrow$ **randomization test**, $\neq$ permutation test

relies on the distribution of the labels, not the data

disadvantage: not exact for given randomized design

# time series

run-wise **means** or GLM estimates are often exchangeable




⟶ run-wise classification may be a better alternative

# alternative to classification: cvMANOVA

cross-validated multivariate ANOVA allows for arbitrary time series designs ⟶ **more flexible** than decoding

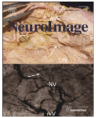
NeuroImage 89 (2014) 345–357

ELSEVIER

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg




## Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA

CrossMark

Carsten Allefeld [a,b,*], John-Dylan Haynes [a,b,c,d,e,f]

[a] Bernstein Center for Computational Neuroscience, Charité Universitätsmedizin Berlin, Germany
[b] Berlin Center of Advanced Neuroimaging, Charité Universitätsmedizin Berlin, Germany
[c] Department of Neurology, Charité Universitätsmedizin Berlin, Germany
[d] Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Germany
[e] Excellence Cluster NeuroCure, Charité Universitätsmedizin Berlin, Germany
[f] Department of Psychology, Humboldt-Universität zu Berlin, Germany

# Further topics

# multiple comparisons

several tests in parallel (voxels): control for **family-wise error**

precise correction depends on dependency between tests, accounted for by using the **same permutation** across tests

test statistic $T_{ij}$, permutations $i = 1 \dots n_{\mathrm{P}}$, tests $j = 1 \dots n_{\mathrm{T}}$

- test whether there is an effect somewhere (omnibus $\mathrm{H}_0$): use the **maximum statistic**

$$M_i = \max_{j=1}^{n_{\mathrm{T}}} T_{ij}, \qquad p_{\mathrm{omnibus}} = \frac{1}{n_{\mathrm{P}}} \sum_{i=1}^{n_{\mathrm{P}}} [M_i \geq M_1]$$

- test whether there is an effect at $j$, corrected

$$p_j = \frac{1}{n_{\mathrm{P}}} \sum_{i=1}^{n_{\mathrm{P}}} [M_i \geq T_{1j}]$$

# beyond permutations

the principle underlying permutation tests:

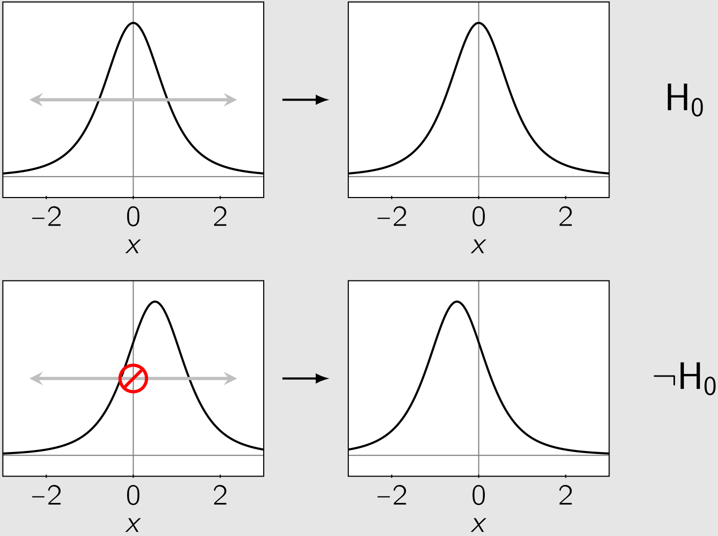
- find a group of **transformations** under which the $H_0$-distribution is **invariant** (includes the neutral transformation)
- compute the test statistic from data after **each transformation** has been applied
- ...

permutation is a special case of transformation,
exchangeability is a special case of invariance

**“group”**: the combination of two transformations is another transformation (is also in the group)

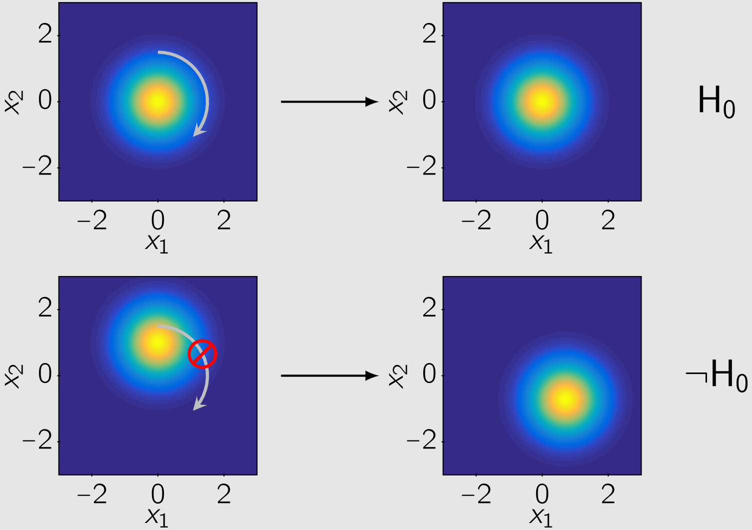
# beyond permutations: mirror symmetry

'sign-flip test': $x \leftrightarrow -x$



# beyond permutations: sphericity

'rotation test': `https://f1000research.com/posters/6-1085`



# further reading

permutation tests, general

– Ernst, Permutation methods: A basis for exact inference, *Statistical Science* 2004
– Good, *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, Springer 2005
– Lehmann & Romano, *Testing Statistical Hypotheses*, Springer 2005, Sec. 5.8

neuroimaging

– Nichols & Holmes, Nonparametric permutation tests for functional neuroimaging: A primer with examples, *Human Brain Mapping* 2001
– Winkler et al., Permutation inference for the general linear model, *NeuroImage* 2014
– Winkler et al., Non-parametric combination and related permutation tests for neuroimaging, *Human Brain Mapping* 2016

MVPA (caution!)

– Golland & Fischl, Permutation tests for classification: Towards statistical significance in image-based studies, *Information processing in medical imaging* 2003
– Etzel & Braver, MVPA permutation schemes: Permutation testing in the land of cross-validation, *PRNI* 2013
– Schreiber & Krekelberg, The statistical analysis of multi-voxel patterns in functional imaging, *PLOS ONE* 2013
– Stelzer et al., Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA), *NeuroImage* 2013
– Allefeld et al., Valid population inference for information-based imaging: From the second-level *t*-test to prevalence inference, *NeuroImage* 2016

randomization tests

– Lehmann & Romano, *Testing Statistical Hypotheses*, Springer 2005, Sec. 5.10
– Edgington & Onghena, *Randomization Tests*, Chapman & Hall 2007