Change detection in multi-dimensional datasets and time series

Andrea De Simone

andrea.desimone@sissa.it

Univ. Camerino, 2019-02-26

[DS, Jacques – arXiv:1807.06038]

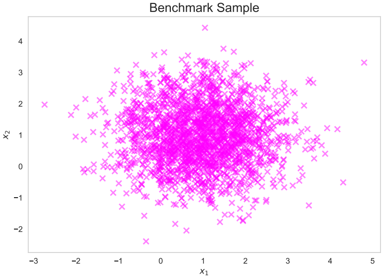
⟩ Outline

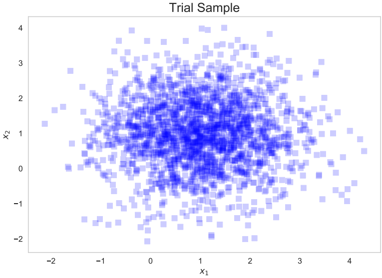
〉Two-Sample Test

Two sets:

$$\text{Trial:} \quad \mathcal{T} \equiv \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_T}\} \overset{\text{iid}}{\sim} p_T \,, \qquad \boldsymbol{x}_i, \boldsymbol{x}'_i \in \mathbb{R}^D$$

$$\text{Benchmark:} \quad \mathcal{B} \equiv \{\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{N_B}\} \overset{\text{iid}}{\sim} p_B \,. \qquad p_B, p_T \text{ unknown}$$





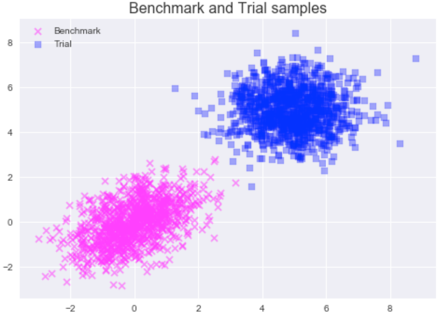
$\rangle$ Two-Sample Test

Two sets:

$$\text{Trial:} \quad \mathcal{T} \equiv \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_T}\} \overset{\text{iid}}{\sim} p_T\,, \qquad \boldsymbol{x}_i, \boldsymbol{x}'_i \in \mathbb{R}^D$$

$$\text{Benchmark:} \quad \mathcal{B} \equiv \{\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{N_B}\} \overset{\text{iid}}{\sim} p_B\,. \qquad p_B, p_T \text{ unknown}$$

« Are $\mathcal{B}, \mathcal{T}$ drawn from the same probability distribution? »



easy. . .

$\rangle$ Two-Sample Test

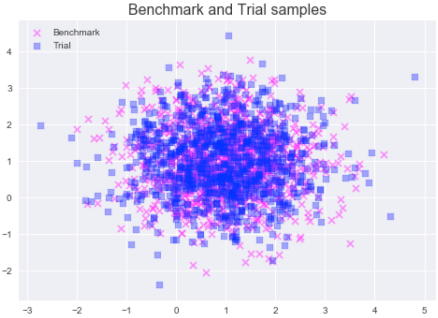
Two sets:

$$\begin{aligned} \text{Trial:} \quad & \mathcal{T} \equiv \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_T}\} \overset{\text{iid}}{\sim} p_T\,, \\ \text{Benchmark:} \quad & \mathcal{B} \equiv \{\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{N_B}\} \overset{\text{iid}}{\sim} p_B\,. \end{aligned}$$

$\boldsymbol{x}_i, \boldsymbol{x}'_i \in \mathbb{R}^D$

p_B, p_T unknown

« Are $\mathcal{B}, \mathcal{T}$ drawn from the same probability distribution? »



. . . hard

〉Two-Sample Test

Why is it important?

- detect departures from benchmark
- find anomalous points (outliers)
- check if observed data are compatible with expectations
- detect changes in underlying distributions
- real-time detect events/shifts in time series

$\rangle$ Two-Sample Test

Desiderata for a statistical test

(1) **model-independent**
no assumption about underlying physical model to interpret data
$\longrightarrow$ more general

(2) **non-parametric**
compare two samples as a whole (not just their means, etc.)
$\longrightarrow$ fewer assumptions, no max likelihood estim.

(3) **un-binned**
high-dim feature space partitioned without rectangular bins
$\longrightarrow$ retain full multi-dim info of data

⟩ Two-Sample Test

Recipe

(1) **Density Estimator**
$\longrightarrow$ reconstruct PDF from samples

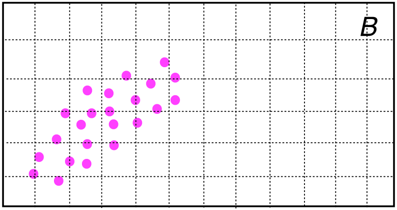
(2) **Test Statistic (TS)**
$\longrightarrow$ "measure distance" between PDFs

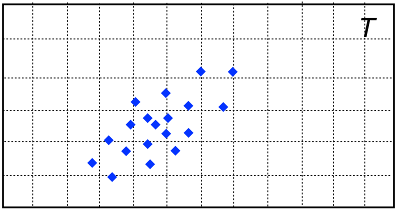
(3) **TS distribution**
$\longrightarrow$ associate probabilities to TS
under null hypothesis $H_0 : p_B = p_T$

(4) p**-value**
$\longrightarrow$ if $p < \alpha$ then reject H_0

Let's build the **Nearest Neighbors Two-Sample Test (NN2ST)**

〉1. Density Estimator





Divide space in square bins?

- ✓ easy
- ✓ can use simple statistics (e.g. χ^2)
- ✗ hard/slow/impossible in high-D

Need un-binned, multi-variate approach

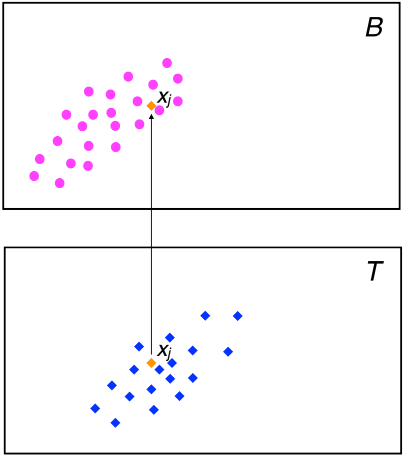
Find PDF *estimators* $\hat{p}_B, \hat{p}_T$, e.g. based on density of points

$$\hat{p}_{B,T}(\boldsymbol{x}) = \frac{\rho_{B,T}(\boldsymbol{x})}{N_{B,T}}$$

Nearest Neighbors!

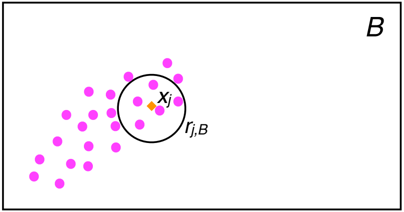
[Schilling 1986, Henze 1988]

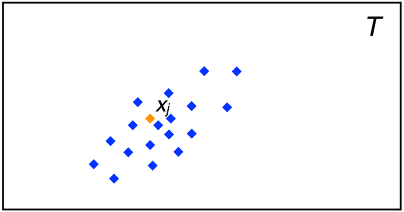
[Wang et al. 2005-2006, Perez-Cruz. 2008]



- Fix integer K.
- Choose query point $\boldsymbol{x}_j$ in $\mathcal{T}$ and draw it in $\mathcal{B}$.

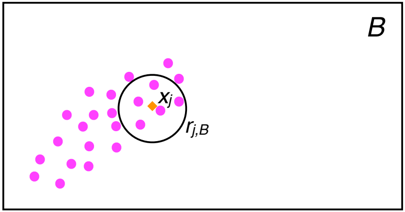
〉1. Density Estimator

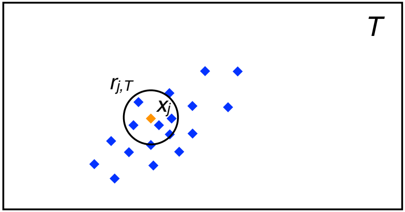




- Fix integer K.
- Choose query point $\boldsymbol{x}_j$ in $\mathcal{T}$ and draw it in $\mathcal{B}$.
- Find the distance $r_{j,B}$ of the K^{th}-NN of $\boldsymbol{x}_j$ in $\mathcal{B}$.

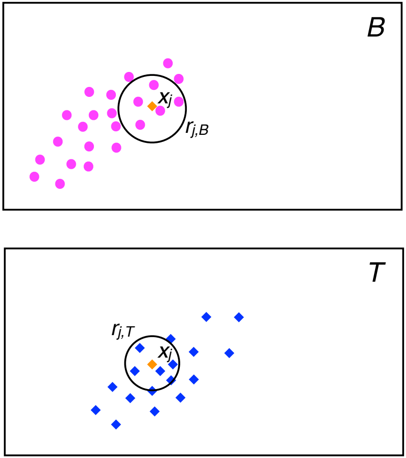
〉 1. Density Estimator





- Fix integer K.
- Choose query point $\boldsymbol{x}_j$ in $\mathcal{T}$ and draw it in $\mathcal{B}$.
- Find the distance $r_{j,B}$ of the K^{th}-NN of $\boldsymbol{x}_j$ in $\mathcal{B}$.
- Find the distance $r_{j,T}$ of the K^{th}-NN of $\boldsymbol{x}_j$ in $\mathcal{T}$.

〉1. Density Estimator



- Fix integer K.
- Choose query point $\boldsymbol{x}_j$ in $\mathcal{T}$ and draw it in $\mathcal{B}$.
- Find the distance $r_{j,B}$ of the K^{th}-NN of $\boldsymbol{x}_j$ in $\mathcal{B}$.
- Find the distance $r_{j,T}$ of the K^{th}-NN of $\boldsymbol{x}_j$ in $\mathcal{T}$.
- Estimate PDFs:

$$\begin{aligned}\hat{p}_B(\boldsymbol{x}_j) &= \frac{K}{N_B}\frac{1}{\omega_D r_{j,B}^D} \\ \hat{p}_T(\boldsymbol{x}_j) &= \frac{K}{N_T - 1}\frac{1}{\omega_D r_{j,T}^D}\end{aligned}$$

〉2. Test Statistic

- Measure the "distance" between 2 PDFs
- Define **Test Statistic** (to detect under-/over-densities)

$$\boxed{\mathrm{TS}(\mathcal{T}) \equiv \frac{1}{N_T}\sum_{j=1}^{N_T} \log \frac{\hat{p}_T(\boldsymbol{x}_j)}{\hat{p}_B(\boldsymbol{x}_j)}}$$

- Form NN-estimated PDFs:

$$\mathrm{TS}(\mathcal{T}) = \frac{D}{N_T}\sum_{j=1}^{N_T} \log \frac{r_{j,B}}{r_{j,T}} + \log \frac{N_B}{N_T - 1}$$

- Related to *Kullback-Leibler* divergence as: $\mathrm{TS}(\mathcal{T}) = \hat{D}_{\mathrm{KL}}(\hat{p}_T || \hat{p}_B)$

$\left[D_{\mathrm{KL}}(p||q) \equiv \int_{\mathbb{R}^D} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}\right]$

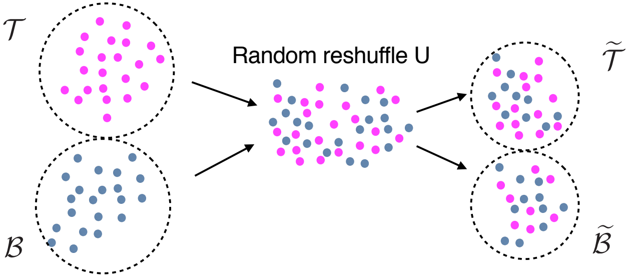
- **Theorem:**
  this estimator converges to $D_{\mathrm{KL}}(p_B || p_T)$, in the large sample limit

[Wang et al. – 2005, 2006]

〉3. Test Statistic Distribution

How is TS distributed? **Permutation test!**

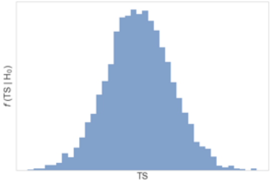
Assume $p_B = p_T$. Union set $\mathcal{U} = \mathcal{T} \cup \mathcal{B}$.



Compute the test statistic TS_n on $(\widetilde{\mathcal{B}}, \widetilde{\mathcal{T}})$.

Repeat many times.

Distribution of TS under H_0 : $f(\mathrm{TS}|H_0) \leftarrow \{\mathrm{TS}_n\}$
[asymptotically normal with zero mean]



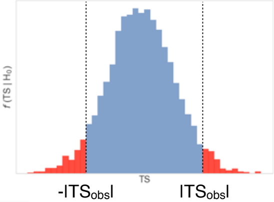
4. p-value

- Find $\hat{\mu}, \hat{\sigma}$: mean, variance of $f(\mathrm{TS}|H_0)$
- Standardize the TS:

$$\mathrm{TS} \to \mathrm{TS}' \equiv \frac{\mathrm{TS} - \hat{\mu}}{\hat{\sigma}}$$

- TS$'$ distributed according to $f'(\mathrm{TS}'|H_0) = \hat{\sigma} f(\hat{\mu} + \hat{\sigma}\mathrm{TS}'|H_0)$
- Two-sided p-value

$$\boxed{p = 2\int_{|\mathrm{TS}_{\mathrm{obs}}|}^{\infty} f'(\mathrm{TS}'|H_0) d\mathrm{TS}'}$$



$\rangle$ NN2ST: Summary

INPUT:

$$\begin{aligned} \text{Trial sample:} \quad & \mathcal{T} \equiv \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_T}\} \overset{\text{iid}}{\sim} p_T\,, \\ \text{Benchmark sample:} \quad & \mathcal{B} \equiv \{\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{N_B}\} \overset{\text{iid}}{\sim} p_B \\ K: \quad & \text{number of nearest neighbors} \\ N_{\text{perm}}: \quad & \text{number of permutations} \end{aligned}$$

$\boldsymbol{x}_i, \boldsymbol{x}'_i \in \mathbb{R}^D$
p_B, p_T
unknown

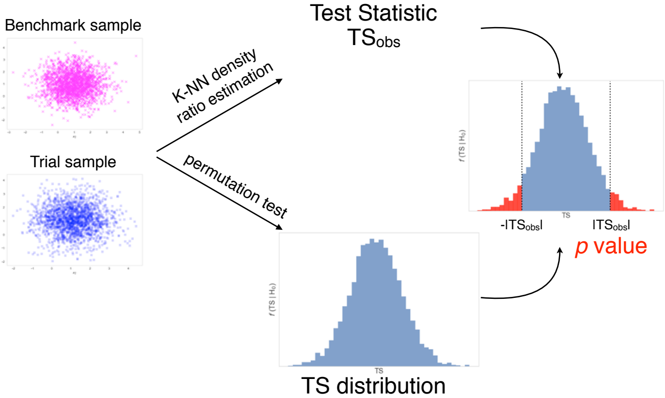
OUTPUT:

p-value of the null hypothesis $H_0 : p_B = p_T$

[check compatibility between 2 samples]

[detect changes in underlying distributions]

〉NN2ST: Summary



Python code:
github.com/de-simone/NN2ST

[DS, Jacques – arXiv:1807.06038]

〉NN2ST: Summary

✓ general, model-independent

✓ solid math foundations

✓ fast, no optimization

✓ sensitive to unspecified signals

✗ need to run for each sample pair

✗ permutation test is bottleneck

⟩ NN2ST on Gaussian Samples

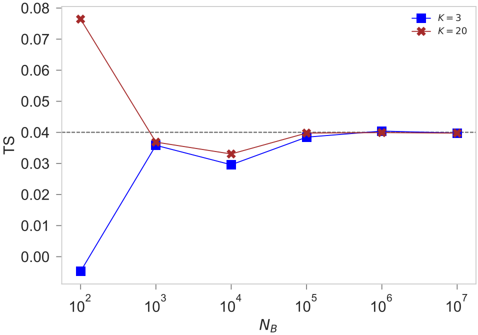
Random samples from
D-dimensional Gaussians

$$D = 2,$$

$$p_B = \mathcal{N}(\boldsymbol{\mu}_B, \Sigma_B)\,,$$
$$p_T = \mathcal{N}(\boldsymbol{\mu}_T, \Sigma_T)\,.$$

$$\boldsymbol{\mu}_B = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}, \qquad \boldsymbol{\mu}_T = \begin{pmatrix} 1.2 \\ 1.2 \end{pmatrix},$$

$$\Sigma_B = \Sigma_T = \boldsymbol{I}_2\,.$$

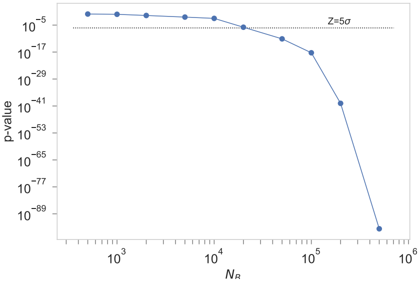


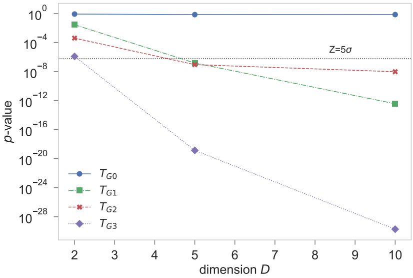
Convergence to exact KL divergence

〉NN2ST on Gaussian Samples

Dataset	$\boldsymbol{\mu}$	Σ
$\mathcal{B}$	1_D	$\boldsymbol{I}_D$
$\mathcal{T}_{G0}$	1_D	$\boldsymbol{I}_D$
$\mathcal{T}_{G1}$	1.12_D	$\boldsymbol{I}_D$
$\mathcal{T}_{G2}$	1_D	$\left(\begin{array}{cc\|c} 0.95 & 0.1 & \multirow{2}{*}{$\mathbf{0}$} \\ 0.1 & 0.8 & \\ \hline \multicolumn{2}{c\|}{\mathbf{0}} & \boldsymbol{I}_{D-2} \end{array}\right)$
$\mathcal{T}_{G3}$	1.15_D	$\boldsymbol{I}_D$

$$\begin{aligned} N_B = N_T &= 20\,000 \\ K &= 5 \\ N_{\rm perm} &= 1\,000 \end{aligned}$$





more data, more power

higher D, more power

⟩ Outlook: time series data

`[Caveat Emptor: very preliminary!]`

Real-time detection of changes in data streams:
variation in underlying mechanism generating data.

$\mathcal{T}, \mathcal{B}$ samples: windows of time series data, ending at discrete times t, t'

$$\begin{aligned} \mathcal{T}_t &= \{\boldsymbol{x}_{t-N+1}, \ldots, \boldsymbol{x}_t\}\,, \\ \mathcal{B}_{t'} &= \{\boldsymbol{x}_{t'-N+1}, \ldots, \boldsymbol{x}_{t'}\}\,, \qquad (N_B = N_T \equiv N)\,. \end{aligned}$$

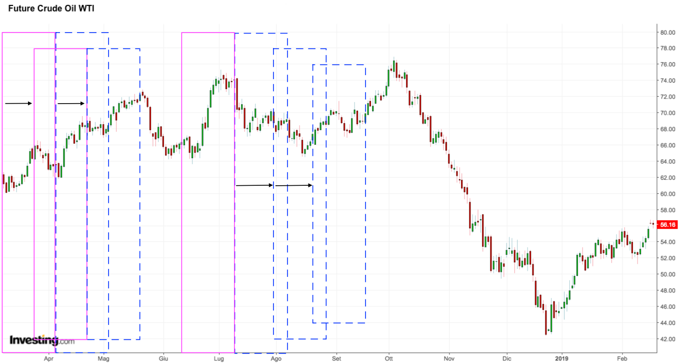
Trial window sliding forward with time.
Benchmark window anchored or rolling.

- anchored $\mathcal{B}$ window: $t' = N \longrightarrow \mathcal{B}_{t'} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$
  Captures cumulative changes over time.

- adjacent windows: $t' = t - N \longrightarrow \mathcal{B}_{t'} = \{\boldsymbol{x}_{t-2N+1}, \ldots, \boldsymbol{x}_{t-N}\}$
  Captures “rate of change” at current time.

〉Outlook: time series data

〉Outlook: time series data



adjacent vs. anchored windows

⟩ Outlook: time series data

- **Feature space** can be high-dimensional: prices (OHLC), prices of related markets, indicators, volumes, …

- Reduce false alarms with **persistence factor** γ (~ 1)%.

H_0 rejected $\gamma \cdot N$ times in a row

$\longrightarrow$ **detected change in market conditions**

⟩ Take-Home Messages

(1) **Proposed a new statistical test: NN2ST**

(2) **Model-independent and suitable for high-D data**

(3) **Excellent results on static datasets**

(4) **Promising applications for change detection in time series data**