Collaborative non-parametric two-sample testing

Alejandro de la Concha, **Nicolas Vayatis**, **Argyris Kalogeratos**

Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, France

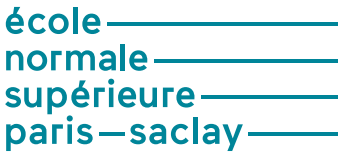
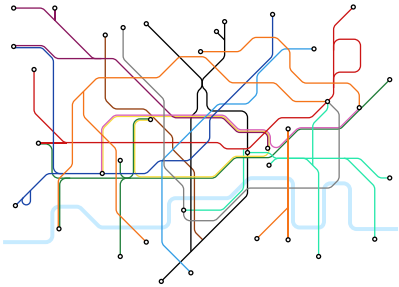


Table of contents

Table of contents

Complex systems in real life

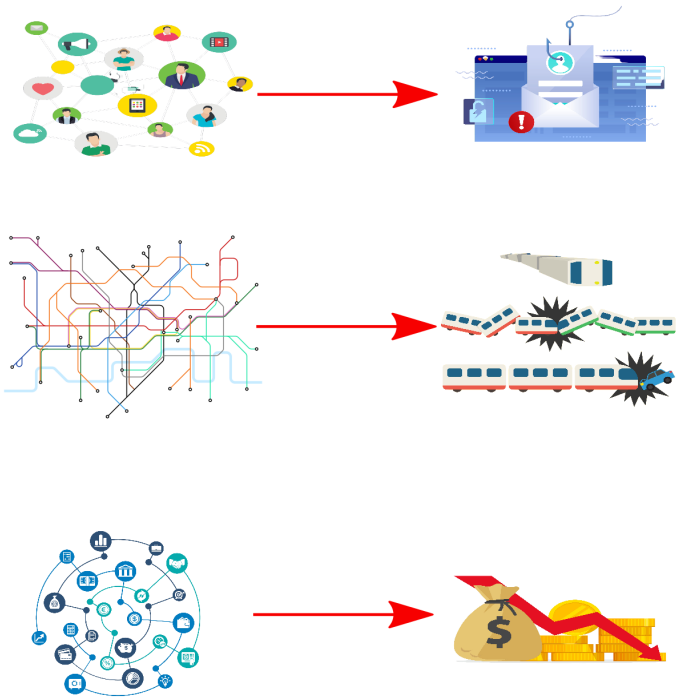
Complex systems in which nodes generate data are everywhere.



[1] Source: Shutterstock

Detecting changes in complex systems

Goal: Compare the behavior of a complex system at two time-stamps τ_1 and τ_2 or two different experimental conditions:



Question: Is network topology relevant for a comparison task?

Spatial statistics

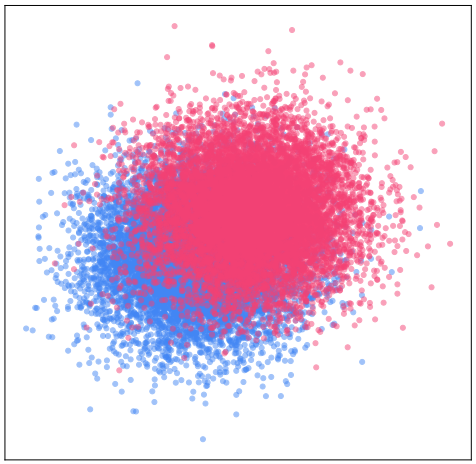
Waldo Tobler (1930-2018)

Neuroscience

Donald Hebb (1904-1985)

Table of contents

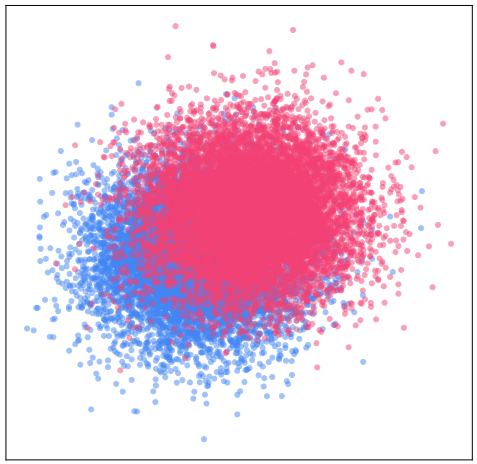
Problem



$$\mathrm{H}_{\mathrm{null}}: \quad p = q \quad \text{vs.} \quad \mathrm{H}_{\mathrm{alt}}: \quad p \neq q$$

Statistics 101

Problem



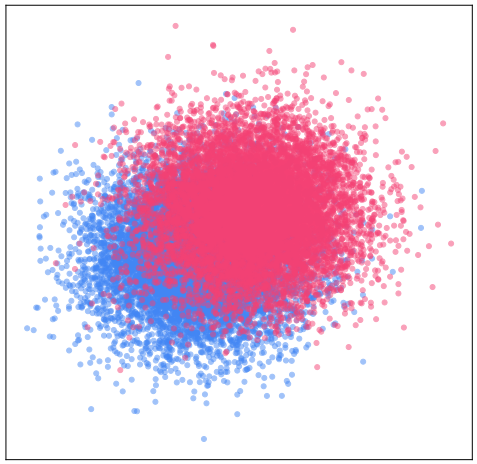
$$\mathrm{H}_{\text{null}}: \quad p = q \quad \text{vs.} \quad \mathrm{H}_{\text{alt}}: \quad p \neq q$$

Solution

1. A test statistic S to quantify the difference between p and q.
2. Compute π-value of S. Reject if $\hat{\pi} \leq \pi^*$.

Example: Hotelling's T^2

Problem



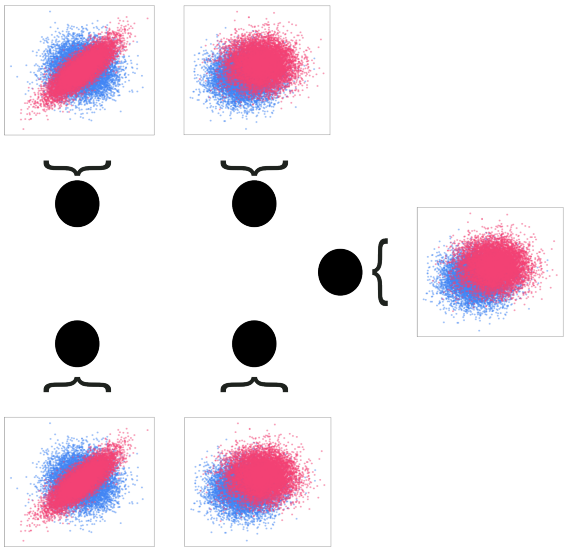
$$\mathrm{H}_{\text{null}} : \mu_1 = \mu_2 \text{ vs. } \mathrm{H}_{\text{alt}} : \mu_1 \neq \mu_2$$

Solution

1. $\hat{S} = (\hat{\mu}_1 - \hat{\mu}_2)^{\mathsf{T}}((\frac{1}{n} + \frac{1}{n'})\hat{\Sigma})^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$
2. The distribution of S is known when p and q are Gaussians. Then $\hat{\pi} = P(S > \hat{S} | \mathrm{H}_{\text{null}})$. If $\hat{\pi} \leq \pi^*$, reject H_{null}.

Multiple hypothesis testing

Problem

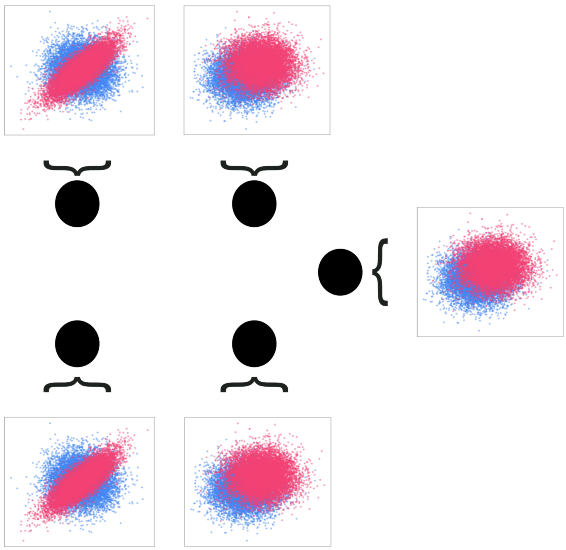


Given a set of two-sample testing problems $v \in \{1, ..., N\}$, identify the null hypotheses to be rejected:

$$\mathrm{H}_{\mathrm{null},v} : p_v = q_v \quad \text{vs.} \quad \mathrm{H}_{\mathrm{alt},v} : p_v \neq q_v$$

Multiple hypothesis testing

Problem



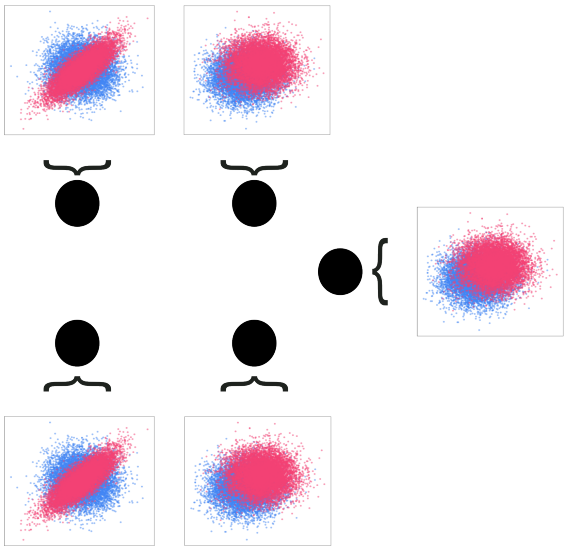
Given a set of two-sample testing problems $v \in \{1, ..., N\}$, identify the null hypotheses to be rejected:

$$\mathrm{H}_{\mathrm{null},v} : p_v = q_v \quad \text{vs.} \quad \mathrm{H}_{\mathrm{alt},v} : p_v \neq q_v$$

Multiple Comparison Problem: Given a level of confidence π^*, we expect to reject $\pi^* N$ null hypothesis just by chance.

Multiple hypothesis testing

Problem

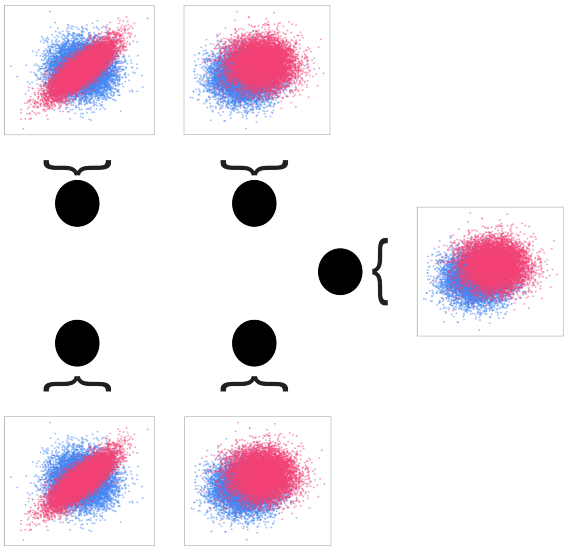


Given a set of two-sample testing problems $v \in \{1, ..., N\}$, identify the null hypotheses to be rejected:

$$\mathrm{H}_{\mathrm{null},v} : p_v = q_v \quad \text{vs.} \quad \mathrm{H}_{\mathrm{alt},v} : p_v \neq q_v$$

Multiple hypothesis testing

Problem



Given a set of two-sample testing problems $v \in \{1, ..., N\}$, identify the null hypotheses to be rejected:

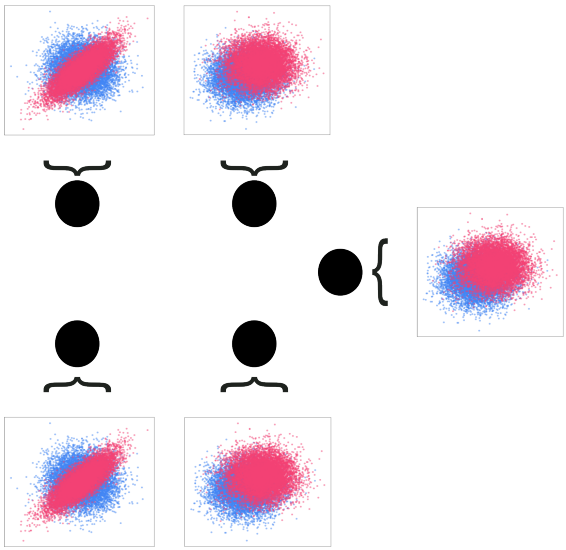
$$\mathrm{H}_{\mathrm{null},v} : p_v = q_v \quad \text{vs.} \quad \mathrm{H}_{\mathrm{alt},v} : p_v \neq q_v$$

Naive Solution

1. Estimate a test statistic independently for each pair of data-sets $\{\hat{S}_v\}_{v \in \{1,...,N\}}$.
2. Estimate a π-value independently for each test $\{\hat{\pi}_v\}_{v \in \{1,...,N\}}$.
3. A strategy to address MCP.

Example: Hotelling's T^2

Problem



$$\mathrm{H}_{\mathrm{null},v} : \mu_{1,v} = \mu_{2,v} \text{ vs. } \mathrm{H}_{\mathrm{alt},v} : \mu_{1,v} \neq \mu_{2,v}$$

Naive solution

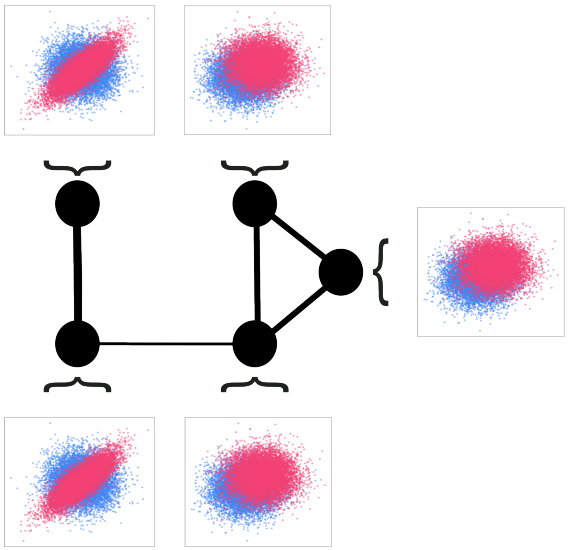
1. For each pair of datasets $\hat{S}_v = (\hat{\mu}_{1,v} - \hat{\mu}_{2,v})^{\mathsf{T}} \left(\left(\frac{1}{n} + \frac{1}{n'} \right) \hat{\Sigma} \right)^{-1} (\hat{\mu}_{1,v} - \hat{\mu}_{2,v})$
2. For each test: $\hat{\pi}_v = P(S_v > \hat{S}_v | \mathrm{H}_{\mathrm{null},v})$.
3. (Bonferroni correction) Fix the level of confidence: $\frac{\pi^*}{N}$.Reject hypotheses $R_{MT} = \{v | \hat{\pi}_v \leq \frac{\pi^*}{N}\}$.

Multiple hypothesis testing

1. Parametric tests such as Hotelling's T^2 require strong assumptions on the distributions p_v and q_v and the way they differ.
2. Estimating the π-value independently for each test ignores the potential correlation between datasets.
3. Bonferroni correction ignores any structure of the problem. Consequently, it may be overly conservative and often fails to detect phenomena of interest.

Collaborative non-parametric two-sample testing

Problem



Consider a weighted graph $G = (V, E, W)$, and to each node $v \in V$, associate a two-sample testing problem. Identify those hypotheses that should be rejected:

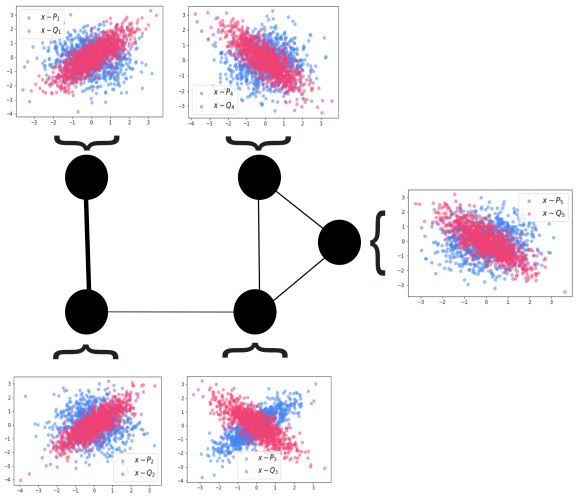
$$\mathrm{H}_{\mathrm{null},v} : p_v = q_v \quad \text{vs.} \quad \mathrm{H}_{\mathrm{alt},v} : p_v \neq q_v$$

Graph smoothness: Connected nodes will have similar two-sample test problems.

Table of contents

Collaborative non-parametric two-sample testing

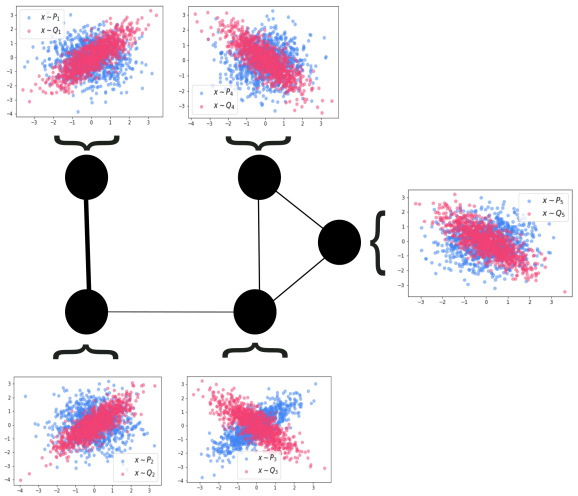
Solution



1. **Collaborative estimation**: Estimate jointly the dissimilarity between p_v and q_v by non-parametric methods. This minimizes the assumptions on p_v and q_v.

Collaborative non-parametric two-sample testing

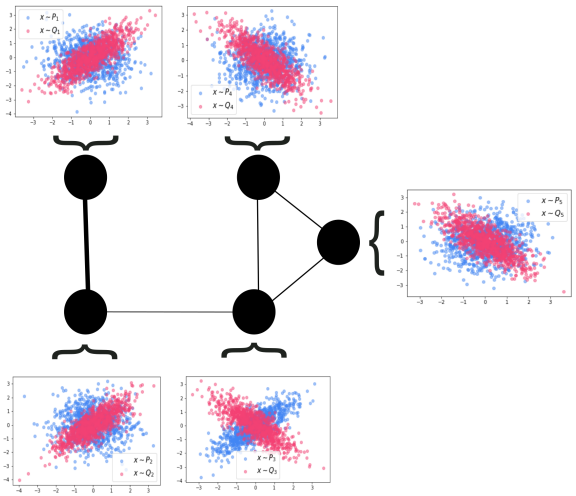
Solution



1. **Collaborative estimation**: Estimate jointly the dissimilarity between p_v and q_v by non-parametric methods. This minimizes the assumptions on p_v and q_v.
2. **Node-level π-value**: Run a permutation test that accounts for the correlation between test statistics $\{\hat{\pi}_v\}_{v \in G}$.

Collaborative non-parametric two-sample testing

Solution



1. **Collaborative estimation**: Estimate jointly the dissimilarity between p_v and q_v by non-parametric methods. This minimizes the assumptions on p_v and q_v.
2. **Node-level π-value**: Run a permutation test that accounts for the correlation between test statistics $\{\hat{\pi}_v\}_{v \in G}$.
3. **MCP control**: Control the probability of wrongly rejecting at least one null hypothesis (FWER).

Pearson divergence

Let us consider $\mathbf{X} \subset \mathrm{R}^n$. And two q, p probability density functions w.r.t the Lebesgue measure. Then the PE-divergence [**Pearson1900**] between p and q is defined as:

$$PE(p\|q) := \int \frac{(r(x)-1)^2}{2} p(x)dx$$

where $r(x) = \frac{q(x)}{p(x)}$ (likelihood-ratio).

Pearson divergence

Let us consider $\mathbf{X} \subset \mathrm{R}^n$. And two q, p probability density functions w.r.t the Lebesgue measure. Then the PE-divergence [**Pearson1900**] between p and q is defined as:

$$PE(p\|q) := \int \frac{(r(x)-1)^2}{2} p(x)dx$$

where $r(x) = \frac{q(x)}{p(x)}$ (likelihood-ratio).

Key property: $PE(p\|q) \geq 0$, $PE(p\|q) = 0$ if and only if $p = q$.
Main idea: Use $PE(p\|q)$ as a test statistic.

Variational Representation

Variational representation: Computing $PE(p\|q)$ is equivalent to solve an optimization problem:

$$PE(p\|q) \geq \sup_{f \in \mathcal{F}} \int f(x)q(x)dx - \int \frac{f^2(x)}{2} p(x)dx - \frac{1}{2},$$

where $\mathcal{F}$ is a functional space and f approximates the likelihood-ratio r.

Variational Representation

Variational representation: Computing $PE(p\|q)$ is equivalent to solve an optimization problem:

$$PE(p\|q) \geq \sup_{f \in \mathcal{F}} \int f(x)q(x)dx - \int \frac{f^2(x)}{2} p(x)dx - \frac{1}{2},$$

where $\mathcal{F}$ is a functional space and f approximates the likelihood-ratio r.

Computing $PE(p\|q)$ is equivalent to estimate the likelihood-ratio.

Variational Representation

Variational representation: Computing $PE(p\|q)$ is equivalent to solve an optimization problem:

$$PE(p\|q) \geq \sup_{f \in \mathcal{F}} \int f(x)q(x)dx - \int \frac{f^2(x)}{2} p(x)dx - \frac{1}{2},$$

where $\mathcal{F}$ is a functional space and f approximates the likelihood-ratio r.

Computing $PE(p\|q)$ is equivalent to estimate the likelihood-ratio.

Question: Which functional space should be employed?

Relative Likelihood-Ratio Estimation

Required conditions:

1. The Variational formulation of $PE(p\|q)$ requires $PE(p\|q) < \infty$.
2. The existence of the likelihood-ratio requires $Q \ll P$.

Relative Likelihood-Ratio Estimation

Required conditions:

1. The Variational formulation of $PE(p\|q)$ requires $PE(p\|q) < \infty$.
2. The existence of the likelihood-ratio requires $Q \ll P$.

Practice: If $\mathrm{Supp}(Q) \not\subset \mathrm{Supp}(P)$ then:

1. r does not exist.
2. $PE(p\|q) = \infty$

Relative Likelihood-Ratio Estimation

Required conditions:

1. The Variational formulation of $PE(p\|q)$ requires $PE(p\|q) < \infty$.
2. The existence of the likelihood-ratio requires $Q \ll P$.

Practice: If $\mathrm{Supp}(Q) \not\subset \mathrm{Supp}(P)$ then:

1. r does not exist.
2. $PE(p\|q) = \infty$

Solution: Use $P^\alpha = (1-\alpha)P + \alpha Q$, $0 < \alpha < 1$.

1. $Q \ll P^\alpha$, $r^\alpha(x) = \frac{q(x)}{(1-\alpha)p(x)+\alpha Q}$, $0 < \alpha < 1$.
2. $r^\alpha(x) \leq \frac{1}{\alpha}$ implying $PE(p^\alpha\|q) < \infty$.
3. It still holds $PE(p^\alpha\|q) = 0$ iff $P = Q$.

Likelihood-ratio estimation

Reproducing Kernel Hilbert Space.

1. $\mathbb{H}$ is equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}} : \mathbb{H} \times \mathbb{H} \to \mathrm{R}$, which will be reproduced by a kernel function $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathrm{R}$.
2. $\langle K(x, \cdot), f \rangle_{\mathbb{H}} = f(x)$, for any $f \in \mathbb{H}$
3. $\mathbb{H} = \overline{\mathrm{span}}(\{K(x, \cdot) : \forall x \in \mathcal{X}\})$

Likelihood-ratio estimation

Reproducing Kernel Hilbert Space.

1. $\mathbb{H}$ is equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}} : \mathbb{H} \times \mathbb{H} \to \mathrm{R}$, which will be reproduced by a kernel function $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathrm{R}$.
2. $\langle K(x, \cdot), f \rangle_{\mathbb{H}} = f(x)$, for any $f \in \mathbb{H}$
3. $\mathbb{H} = \overline{\mathrm{span}}(\{K(x, \cdot) : \forall x \in \mathcal{X}\})$

Why Kernel Methods? Nice geometry. It is easier to encode graph smoothness:

$$|f_u(x) - f_v(x)| = |\langle \mathrm{K}(x, \cdot), f_u - f_v \rangle_{\mathbb{H}}| \leq C \, \|f_u - f_v\|_{\mathbb{H}}, \tag{1}$$

Connected nodes should have similar likelihood-ratios

Collaborative estimation in practice

Optimization problem:

$$\min_{\Theta \in \mathbb{R}^{N\hat{M}}} \frac{1}{N} \overbrace{\sum_{v \in V} \left((1-\alpha) \frac{\theta_v^\top H_{\mathcal{Z},v} \theta_v}{2} + \alpha \frac{\theta_v^\top H'_{\mathcal{Z},v} \theta_v}{2} - h'_{\mathcal{Z},v} \theta_v \right)}^{\text{Pearson divergence estimate}}$$

$$+ \overbrace{\frac{\lambda}{4} \sum_{u,v \in V} W_{uv} \left\| \theta_v - \theta_u \right\|^2}^{\text{Graph-smoothness hypothesis}} + \overbrace{\frac{\lambda\gamma}{2} \sum_{v \in V} \left\| \theta_v \right\|^2}^{\text{Node-level regularization}},$$

The problem is quadratic in Θ ! $\Theta = (\theta_1, ..., \theta_v)$

Collaborative estimation in practice

Optimization problem:

$$\min_{\Theta \in \mathbb{R}^{N\hat{M}}} \frac{1}{N} \overbrace{\sum_{v \in V} \left((1-\alpha) \frac{\theta_v^\mathsf{T} H_{\mathcal{Z},v} \theta_v}{2} + \alpha \frac{\theta_v^\mathsf{T} H'_{\mathcal{Z},v} \theta_v}{2} - h'_{\mathcal{Z},v} \theta_v \right)}^{\text{Pearson divergence estimate}}$$

$$+ \overbrace{\frac{\lambda}{4} \sum_{u,v \in V} W_{uv} \left\| \theta_v - \theta_u \right\|^2}^{\text{Graph-smoothness hypothesis}} + \overbrace{\frac{\lambda\gamma}{2} \sum_{v \in V} \left\| \theta_v \right\|^2}^{\text{Node-level regularization}},$$

The problem is quadratic in Θ ! $\Theta = (\theta_1, ..., \theta_v)$

Likelihood-ratio estimate: $r_v^\alpha(x) \approx \hat{f}_v(x) = \sum_{m=1}^{M} \hat{\theta}_{v,m} \mathsf{K}(x, x_m)$

Collaborative estimation in practice

Optimization problem:

$$\min_{\Theta \in \mathbb{R}^{N\hat{M}}} \frac{1}{N} \overbrace{\sum_{v \in V} \left((1-\alpha) \frac{\theta_v^\mathsf{T} H_{\mathcal{Z},v} \theta_v}{2} + \alpha \frac{\theta_v^\mathsf{T} H'_{\mathcal{Z},v} \theta_v}{2} - h'_{\mathcal{Z},v} \theta_v \right)}^{\text{Pearson divergence estimate}}$$

$$+ \overbrace{\frac{\lambda}{4} \sum_{u,v \in V} W_{uv} \left\| \theta_v - \theta_u \right\|^2}^{\text{Graph-smoothness hypothesis}} + \overbrace{\frac{\lambda\gamma}{2} \sum_{v \in V} \left\| \theta_v \right\|^2}^{\text{Node-level regularization}},$$

The problem is quadratic in Θ ! $\Theta = (\theta_1, ..., \theta_v)$

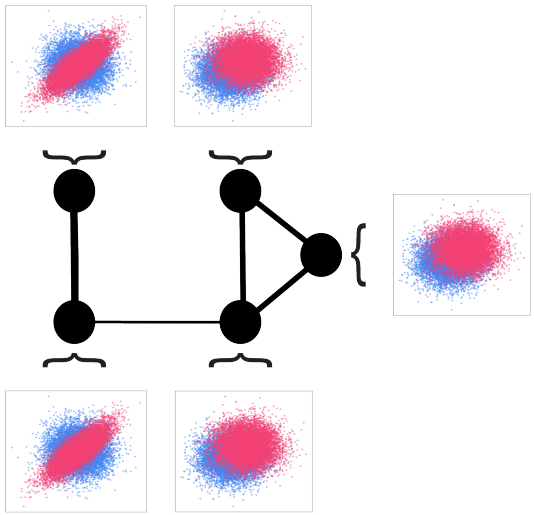
Likelihood-ratio estimate: $r_v^\alpha(x) \approx \hat{f}_v(x) = \sum_{m=1}^{M} \hat{\theta}_{v,m} \mathsf{K}(x, x_m)$

Pearson divergence estimate:

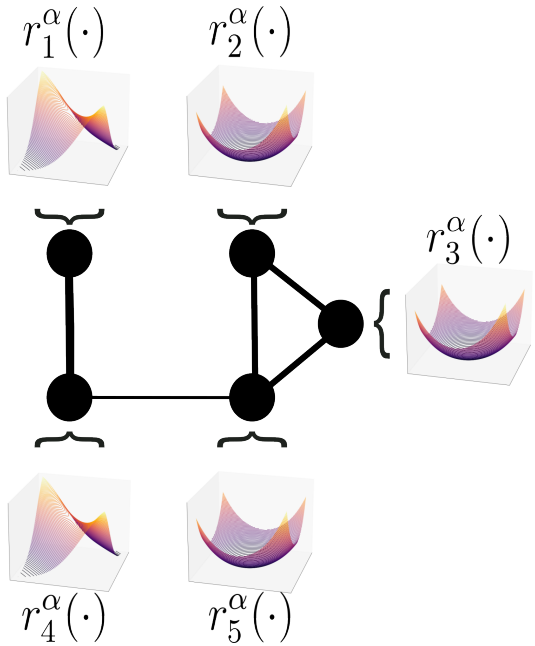
$$\hat{PE}_v^\alpha(\mathbf{X}_v \| \mathbf{X}'_v) = h'^\mathsf{T}_{\psi,v} \hat{\theta}_v - \frac{1-\alpha}{2} \hat{\theta}_v^\mathsf{T} H_{\psi,v} \hat{\theta}_v - \frac{\alpha}{2} \hat{\theta}_v^\mathsf{T} H'_{\psi,v} \hat{\theta}_v - \frac{1}{2}$$

Summary

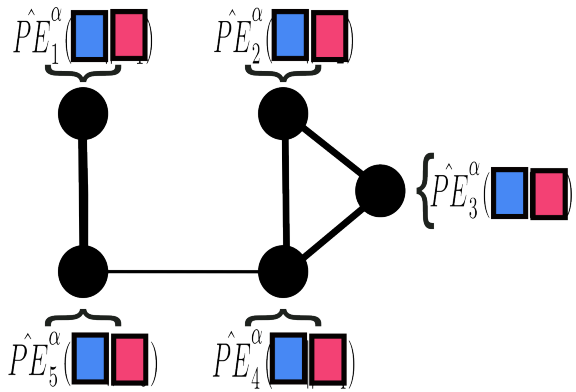
1) Datasets



2) Likelihood-ratio estimation



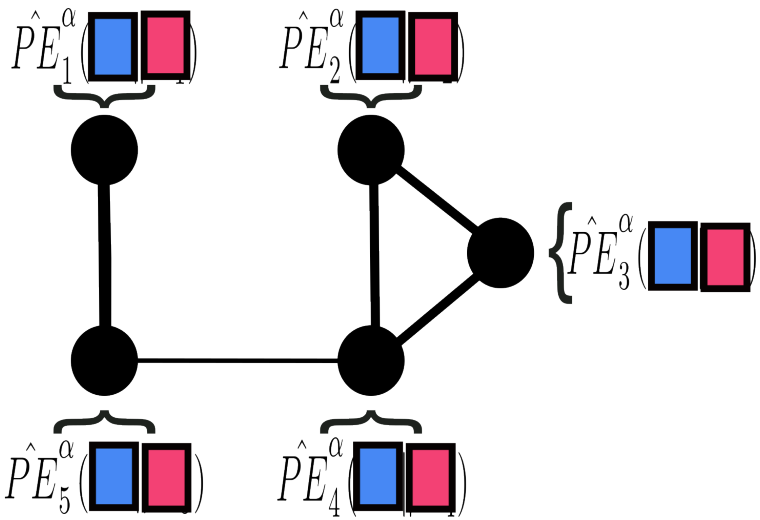
3) Dissimilarity between p_v and q_v

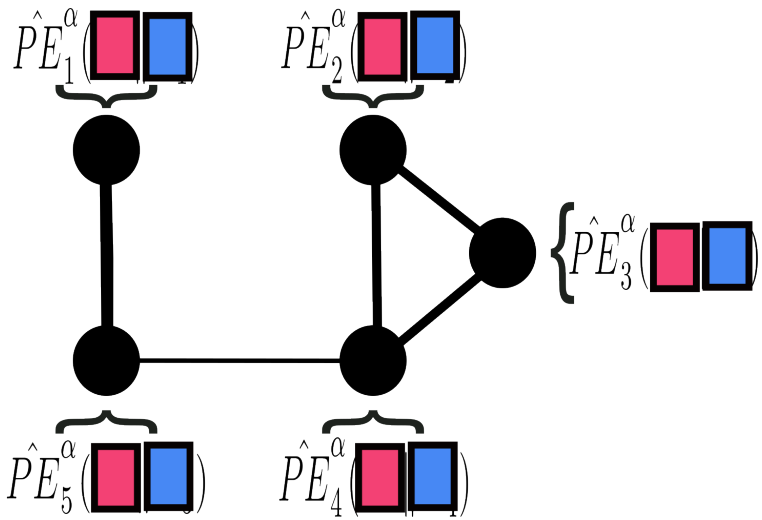


More details: **delaConcha2024**

Node level test statistic

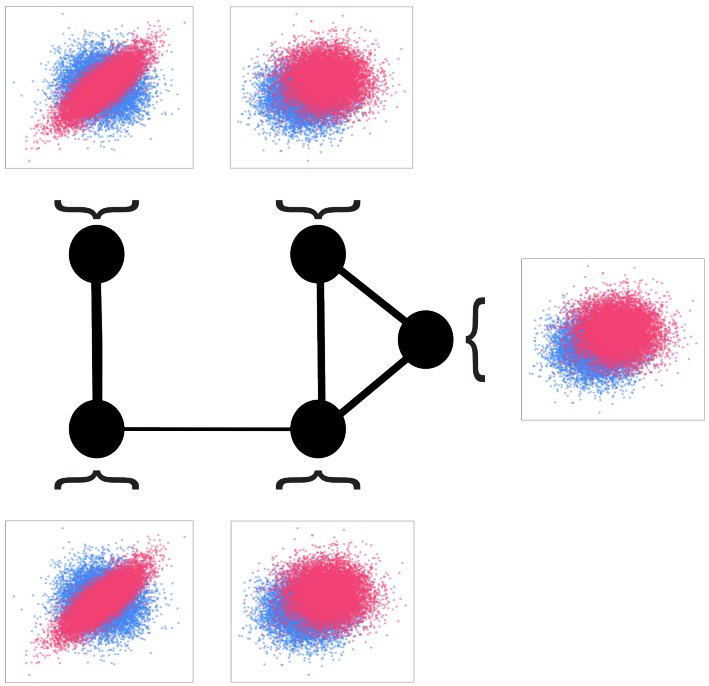
Problem: Pearson divergence is not symmetric. One side is more sensitive than the other.
Solution: Estimate $\hat{PE}_v^{\alpha}(\mathbf{X}_v \| \mathbf{X}'_v)$ and $\hat{PE}_v^{\alpha}(\mathbf{X}'_v \| \mathbf{X}_v)$.

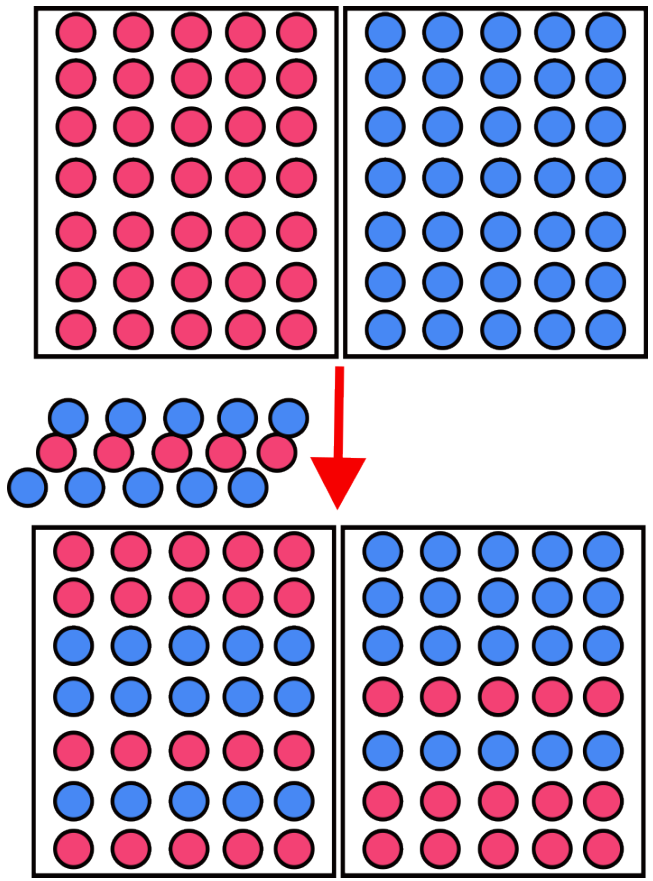




Permutation test

A permutation test that respects the structure of the problem. Test statistics will be correlated.

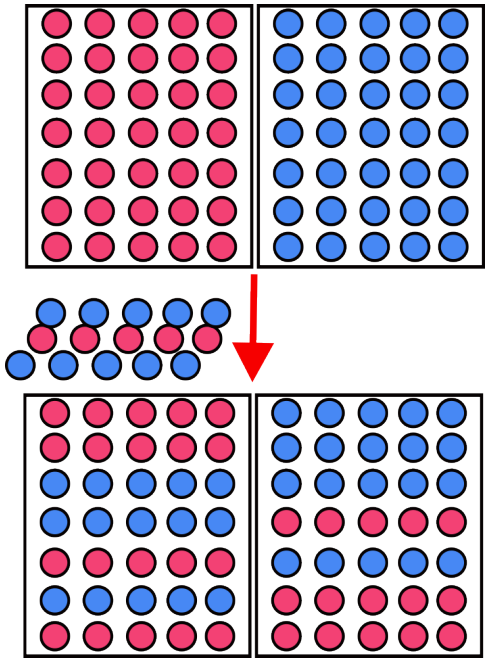




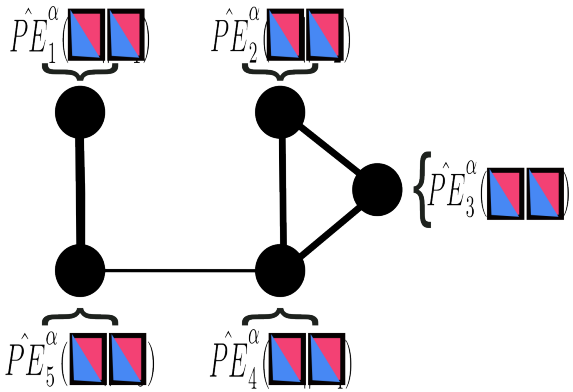
Permutation test

Repeat I times:

1) Permute datasets



2) Collaborative estimation



3) Take the maximum over the entire graph

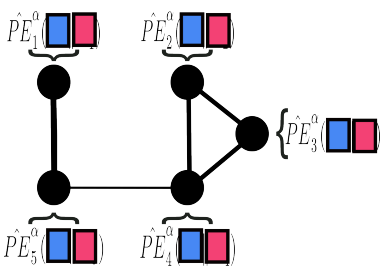
$$S_i = \max_{v \in V} \hat{PE}_v^\alpha(\cdot, \cdot)$$

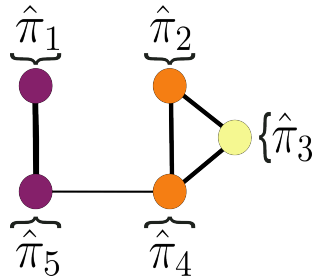
Output: $\{S_i\}_{i=1}^I$.

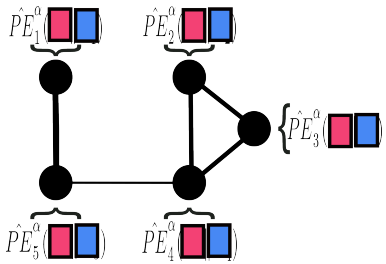
π-value estimation

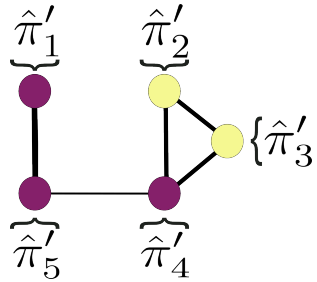
$$\hat{\pi}_v = \frac{1}{I}\#\{i|S_i > \hat{PE}_v^{\alpha}(\mathbf{X}_v\|\mathbf{X}'_v)\} \qquad \hat{\pi}'_v = \frac{1}{I}\#\{i|S_i > \hat{PE}_v^{\alpha}(\mathbf{X}'_v\|\mathbf{X}_v)\}$$

1) Collaborative estimation **2) π-value estimation**









Node identification

Question: Which nodes to pick? Which hypotheses to reject? Which nodes are likely to have seen $p_v \neq q_v$?

- Avoid MCP
- In complex systems we are interested in the global question: is there any change in the system?

Family Wise Error Rate Probability of finding at least one false positive.

Node identification

Question: Which nodes to pick? Which hypotheses to reject? Which nodes are likely to have seen $p_v \neq q_v$?

- Avoid MCP
- In complex systems we are interested in the global question: is there any change in the system?

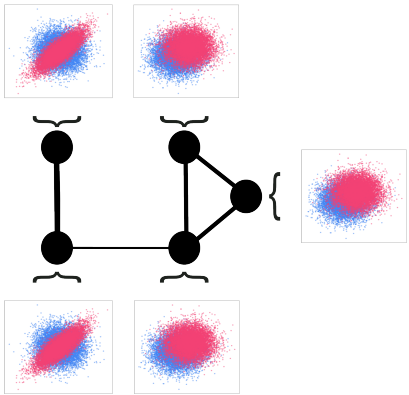
Family Wise Error Rate Probability of finding at least one false positive.
Result: Our permutation algorithm guarantees $\mathrm{FWER} \leq \pi^*$ if:

$$R_{\mathrm{CTST}} = \{v \in V \,|\, \hat{\pi}_v \leq \frac{\pi^*}{2} \quad \text{or} \quad \hat{\pi}'_v \leq \frac{\pi^*}{2}\}$$

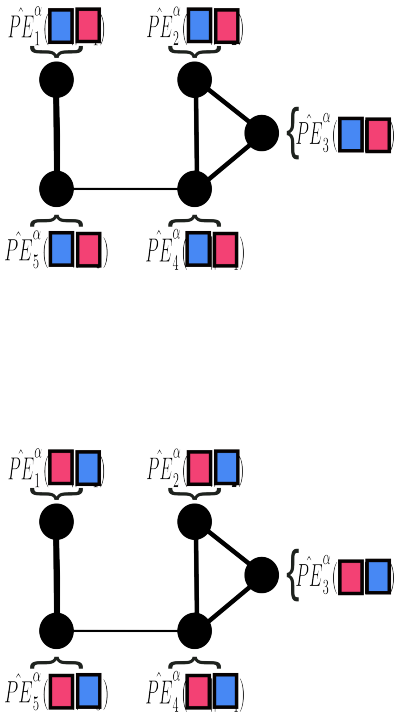
where π^* rate fixed by the user.

Collaborative non-parametric two-sample test

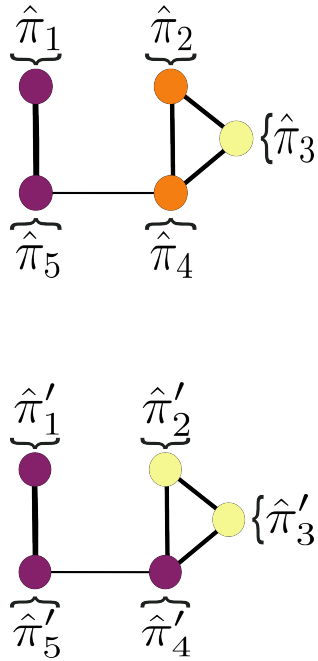
1) Datasets



2) Collaborative estimation



3) π-value Estimation



4) Nodes where $p_v \neq q_v$

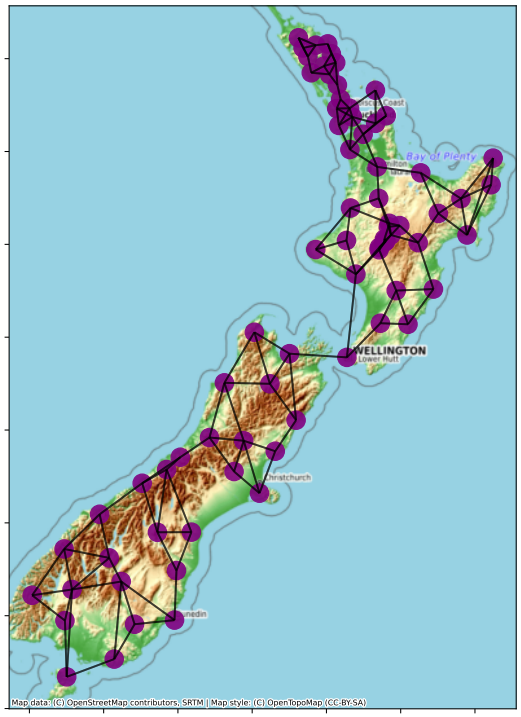
Fix a FWER π^* and chose the nodes $v \in V$ such that $\hat{\pi}_v \leq \frac{\pi^*}{2}$ or $\hat{\pi}'_v \leq \frac{\pi^*}{2}$.

Table of contents

Understanding the behaviour of a earthquake

Question ? Was there an earthquake? If so, which stations were more sensitive to the event, and when?

Nodes: Seismic stations across New-Zealand.
Signal: Each station contains three geophones in three mutually perpendicular directions. ($\mathbb{R}^3$)
Probabilistic models: The state of the system 50 seconds before and 50 seconds after the event.



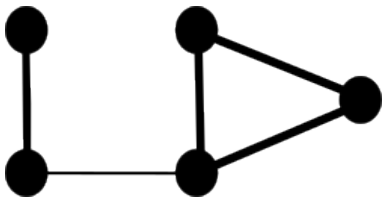
The similarity graph

Question: Which graph to use?

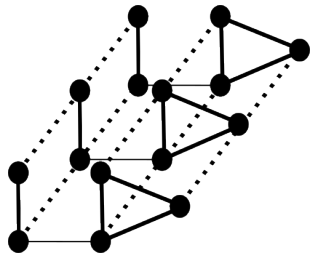
The graph encodes our prior knowledge of how similar tests will be.

1. **Spatial similarity:** Stations close to the epicenter will react similarly.
2. **Temporal similarity:** The seism will propagate through the territory and gradually fade away.

1) K-nearest neighbors



Temporal multiplex



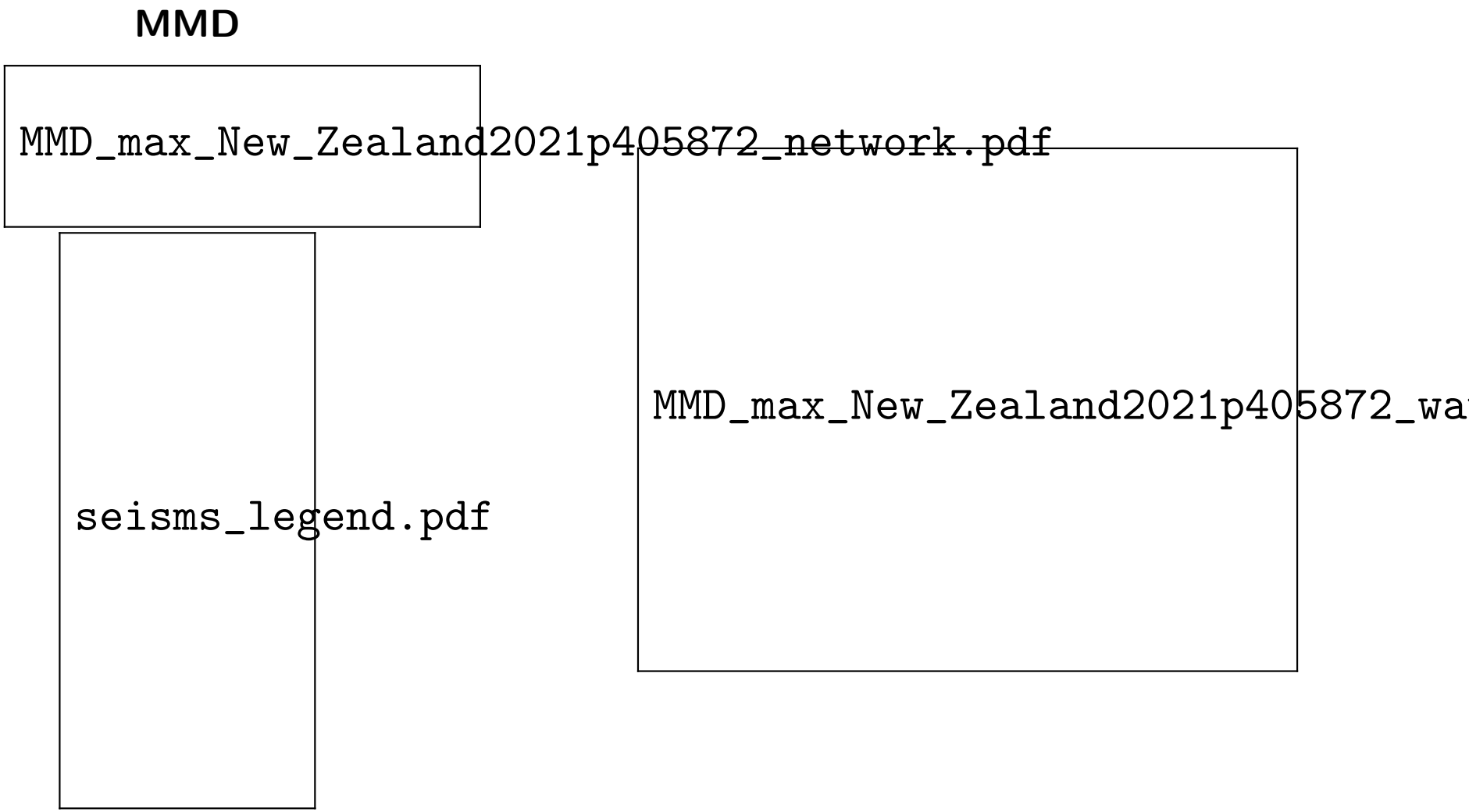
Seism of magnitude 5.5 occurred on 03/31/2021

CTST $\alpha = 0.1$

Seism of magnitude 5.5 occurred on 03/31/2021

RULSIF $\alpha = 0.1$

Seism of magnitude 5.5 occurred on 03/31/2021



Seism of magnitude 2.6, occurred on 10/02/2023

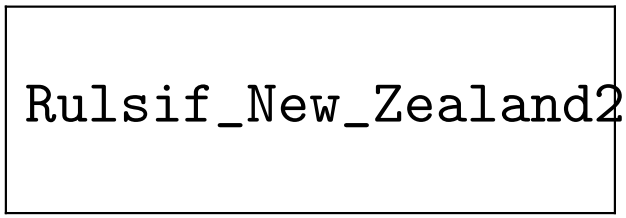
CTST $\alpha = 0.1$

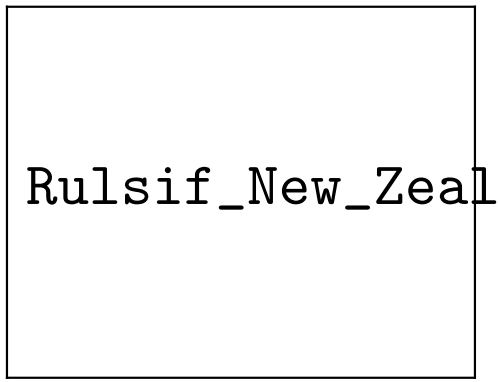
Grulsif_New_Zealand2023p741652_network.pdf

seisms_legend.pdf

Seism of magnitude 2.6 occurred on 10/02/2023

RULSIF $\alpha = 0.1$





Seism of magnitude 2.6 occurred on 10/02/2023

MMD

MMD_max_New_Zealand2023p741652_network.pdf

seisms_legend.pdf

MMD_max_New_Zealand2023p741652_wa

Table of contents

Conclusions

1. Novel graph-structured non-parametric test designed for multiple two-sample testing over the nodes of a graph.
2. A method built upon collaborative likelihood-ratio estimation to compute jointly node-level test statistics and identify null hypotheses to be rejected, under a graph smoothness hypothesis.
3. Data at every node can be multivariate, the nature of the difference between the compared p.d.f.'s is unknown and it is allowed a certain amount of heterogeneity among the tests of the nodes.

More details: **delaconcha2024collaborative**

References 1