Collaborative likelihood-ratio estimation over graphs

Alejandro de la Concha, Argyris Kalogeratos, Nicolas Vayatis

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

1. Motivation

Consider a feature space $\mathcal{X} \subset \mathrm{R}^n$. The **likelihood-ratio** between two density functions $p(x)$ and $q(x)$ is: $r(x) = \frac{q(x)}{p(x)} \quad x \in \mathcal{X}$

Applications of likelihood-ratio estimation (LRE): *Hypothesis Testing* (Neyman-Pearson Lemma, [2]), *Sequential Change-point detection* [4], *Transfer Learning* (*Importance sampling* [1]), ...

Question: LRE techniques are only used on single-source or aggregated data. How can we extend LRE to complex systems such as network of sensors, transport networks, public health surveillance, etc?

Contribution: A graph-based collaborative framework that capitalizes over the similarities between data sources to infer $(r_1(\cdot), ..., r_N(\cdot))$ for all the nodes of a graph

2. Problem statement and framework

Setting

- $G = (V, E, W)$ is a given weighted undirected graph, and W is its weighted adjacency matrix encoding similarity between nodes
- Each node $v \in V$ has access to observations $x_1, x_2, ..., x_n \overset{\text{iid}}{\sim} P_v$ and $x'_1, x'_2, ..., x'_{n'} \overset{\text{iid}}{\sim} Q_v$

Framework [?]

- **Non-parametric LRE:** Infer the node-level relative likelihood-ratios $r_v^\alpha(\cdot) = \frac{q_v(\cdot)}{(1-\alpha)p_v(\cdot)+\alpha q_v(\cdot)}$ via the variational formulation of the Pearson's PE-divergence minimization [3]:

$$PE(p^\alpha, q) = \int \frac{(r^\alpha(x) - 1)^2}{2} p^\alpha(x)dx$$
$$\geq \sup_{f \in \mathbb{H}} \int f(x)q(x)dx - \int \frac{f^2(x)}{2} p^\alpha(x)dx - \frac{1}{2}$$

- **Reproducing Kernel Hilbert Space (RKHS):** The space $\mathbb{H}$ is equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}} : \mathbb{H} \times \mathbb{H} \to \mathrm{R}$, which is induced by a symmetric and positive semi-definite kernel function $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathrm{R}$. $\mathbb{H}$ satisfies the reproducing property, that is $\forall x \in \mathcal{X}$ and $f \in \mathbb{H}$: $f(x) = \langle f(\cdot), K(x, \cdot) \rangle_{\mathbb{H}}$. $\phi(X)$ denotes the associated feature map.
- **Integrate graph component via multitasking:** $\|r_u - r_v\|_{\mathbb{H}} < \epsilon$ if $u \sim v$

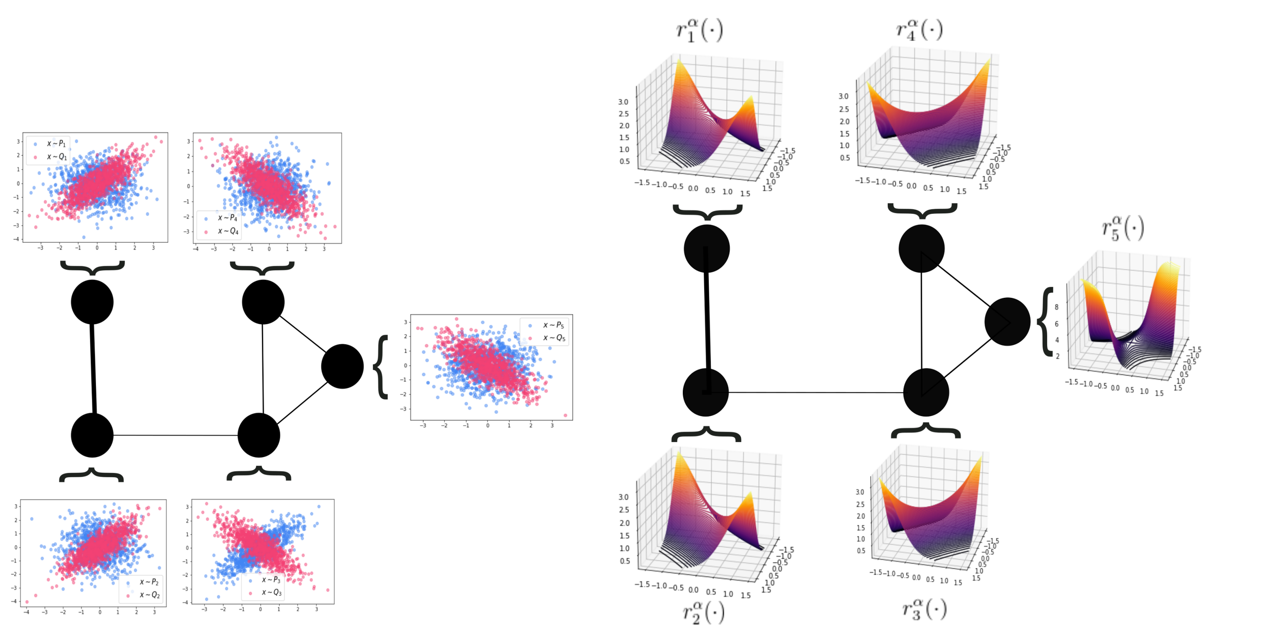
Application: Collaborative two sample test

$$\mathrm{H}_{\text{null}}: \quad p_v = q_v, \quad \forall v \in V \quad \text{vs}$$
$$\mathrm{H}_{\text{alt}}: \quad p_v \neq q_v, \quad \forall v \in C,$$

where C is a subset of nodes

3. GRULSIF: Graph-based Relative Unconstrained Least-Squares Importance Fitting

From node-level data sets to node-level likelihood-ratio functions



By the reproducing property of $\mathbb{H}$, $\forall v \in V$, f_v takes the form $f_v(x) = \sum_{l=1}^{L} \theta_{v,i} K(x, x_i)$. Define the terms:

$$H_v = \frac{1}{n_v} \sum_{x \in \mathbf{X}_v} \phi(x)\phi(x)^\top, H'_v = \frac{1}{n'_v} \sum_{x \in \mathbf{X}'_v} \phi(x)\phi(x)^\top$$
$$h'_v = \frac{1}{n'_v} \sum_{x \in \mathbf{X}'_v} \phi(x).$$

Multitasking formulation of the problem via PE-divergence minimization

$$\min_{\Theta \in \mathbb{R}^{NL}} \frac{1}{N} \sum_{v \in V} \overbrace{\left((1-\alpha)\frac{\theta_v^\top H_v \theta_v}{2} + \alpha \frac{\theta_v^\top H'_v \theta_v}{2} - h'_v \theta_v \right)}^{\text{PE-divergence node-level}}$$
$$+ \overbrace{\frac{\lambda\gamma}{2} \sum_{v \in V} \|\theta_v\|^2}^{\text{node-level regularization term}} + \overbrace{\frac{\lambda}{4} \sum_{u,v \in V} W_{uv} \|\theta_u - \theta_v\|^2}^{\text{graph-level regularization term}}$$

Implementation

The problem is **quadratic** and we solve it via block gradient descent, whose number of iterations scales in $\mathcal{O}(\log^2(NL))$. The i-th cycle of updates for node v can be written as:

$$\hat{\theta}_v^{(i)} = \frac{1}{\eta_v + \lambda\gamma} \left[\eta_v \hat{\theta}_v^{(i-1)} - \overbrace{\left(\frac{(1-\alpha)H_v + \alpha H'_v}{N} \hat{\theta}_v^{(i-1)} - \frac{h'_v}{N} \right)}^{\text{component depending on node } v} \right.$$
$$\left. - \overbrace{\lambda \left(d_v \hat{\theta}_v^{(i-1)} - \sum_{u \in (v)} W_{uv} \left(\hat{\theta}_u^{(i)} \mathbb{1}_{u<v} + \hat{\theta}_u^{(i-1)} \mathbb{1}_{u \geq v} \right) \right)}^{\text{component depending on the graph}} \right]$$

4. Application: Collaborative two-sample test

From likelihood-ratios to p-values

Main hypothesis: C is such that the vector $(r_1^\alpha(x), ..., r_N^\alpha(x))$ is smooth over the graph

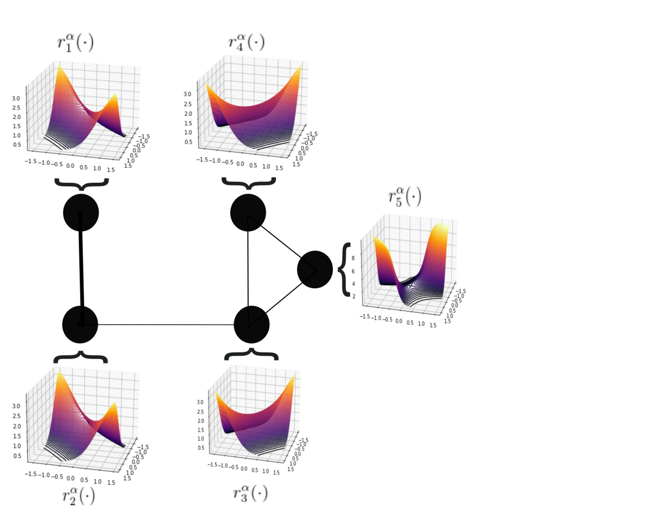
Statistical scores S_v based on PE-divergence approximation

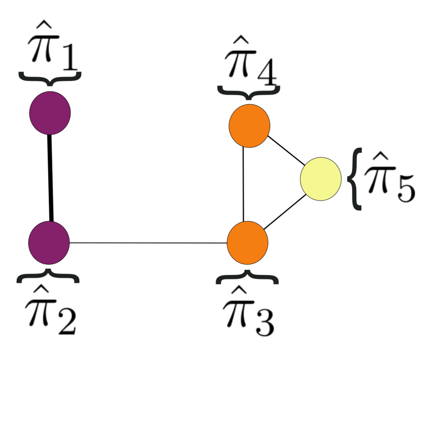
$$\hat{PE}_v^\alpha(\mathbf{X}, \mathbf{X}') = \sum_{x' \in \mathbf{X}'_v} \frac{\hat{f}_v(x)}{n'_v} - \frac{(1-\alpha)}{2} \sum_{x \in \mathbf{X}_v} \frac{\hat{f}_v(x)^2}{n_v}$$
$$- \frac{\alpha}{2} \sum_{x' \in \mathbf{X}'_v} \frac{\hat{f}_v(x)^2}{n'_v} - \frac{1}{2}$$

To address the lack of symmetry, we compute the node-level score $S_v = \hat{PE}_v^\alpha(\mathbf{X}, \mathbf{X}') + \hat{PE}_v^\alpha(\mathbf{X}', \mathbf{X})$

Identify the nodes in C

- Run a permutation test to estimate the p-value $\hat{\pi}_v$ associated with the statistic S_v
- Identify the nodes with the p-values $\hat{\pi}_v$ lower than a prefixed value π^*





6. Experiments on semi-synthetic examples

Synthetic graph structure: Stochastic Block Model to generate graphs with 4 clusters (C_1, C_2, C_3, C_4), made of 20 nodes each. The probability of intra-cluster link is fixed at 0.5, and that of inter-cluster link at 0.01

Node-level dataset: MNIST digits dataset

$$\begin{cases} \mathrm{H}_{\text{null}} & \rightarrow & \mathrm{H}_{\text{alt}} & \text{Selected clusters} \\ x_v \in \text{digits}\{0,1\} & \rightarrow & x'_v \in \text{digits}\{8,9\}, & \text{if } v \in C_1; \\ x_v \in \text{digits}\{2,3\} & \rightarrow & x'_v \in \text{digits}\{8,9\}, & \text{if } v \in C_2; \\ x_v \in \text{digits}\{4,5\} & \rightarrow & x'_v \in \text{digits}\{8,9\}, & \text{if } v \in C_3; \\ x_v \in \text{digits}\{6,7\} & \rightarrow & x'_v \in \text{digits}\{8,9\}, & \text{if } v \in C_1. \end{cases}$$

		$\pi^* = 0.01$		
Method	n=n′	**Recall** (↑)	**Precision** (↑)	**F1** (↑)
GRULSIF α=0.1	25	1.00 (0.01)	0.98 (0.03)	**0.99 (0.01)**
Pool α=0.1	25	0.98 (0.13)	0.60 (0.21)	0.71 (0.18)
GRULSIF α=0.5	25	0.98 (0.05)	0.98 (0.03)	0.98 (0.03)
Pool α=0.5	25	1.00 (0.00)	0.45 (0.12)	0.61 (0.11)
RULSIF α=0.1	25	0.99 (0.03)	0.86 (0.06)	0.92 (0.04)
ULSIF	25	0.97 (0.05)	0.90 (0.05)	0.93 (0.04)
KLIEP	25	0.99 (0.02)	0.43 (0.05)	0.60 (0.05)
MMD median	25	0.33 (0.31)	0.91 (0.22)	0.42 (0.31)
MMD aggreg	25	0.33 (0.29)	0.92 (0.19)	0.43 (0.29)
GRULSIF α=0.1	50	1.00 (0.00)	0.97 (0.04)	**0.98 (0.02)**
Pool α=0.1	50	1.00 (0.00)	0.33 (0.09)	0.50 (0.08)
GRULSIF α=0.5	50	1.00 (0.00)	0.96 (0.04)	0.98 (0.02)
Pool α=0.5	50	1.00 (0.00)	0.33 (0.05)	0.49 (0.05)
RULSIF α=0.1	50	1.00 (0.00)	0.85 (0.06)	0.91 (0.04)
ULSIF	50	1.00 (0.00)	0.88 (0.06)	0.94 (0.03)
KLIEP	50	0.99 (0.02)	0.62 (0.07)	0.76 (0.06)
MMD median	50	0.50 (0.32)	0.94 (0.10)	0.59 (0.27)
MMD aggreg	50	0.57 (0.28)	0.95 (0.07)	0.68 (0.21)
		$\pi^* = 0.05$		
Method	n=n′	**Recall** (↑)	**Precision** (↑)	**F1** (↑)
GRULSIF α=0.1	25	1.00 (0.00)	0.94 (0.05)	**0.97 (0.03)**
Pool α=0.1	25	0.98 (0.13)	0.46 (0.15)	0.60 (0.15)
GRULSIF α=0.5	25	1.00 (0.02)	0.93 (0.06)	0.96 (0.03)
Pool α=0.5	25	1.00 (0.00)	0.34 (0.07)	0.51 (0.08)
RULSIF α=0.1	25	1.00 (0.00)	0.55 (0.05)	0.71 (0.44)
ULSIF	25	1.00 (0.00)	0.62 (0.06)	0.76 (0.04)
KLIEP	25	1.00 (0.00)	0.32 (0.04)	0.49 (0.04)
MMD median	25	0.49 (0.30)	0.82 (0..13)	0.57 (0.24)
MMD aggreg	25	0.55 (0.28)	0.82 (0.10)	0.57 (0.24)
GRULSIF α=0.1	50	1.00 (0.00)	0.92 (0.06)	**0.96 (0.03)**
Pool α=0.1	50	1.00 (0.00)	0.28 (0.07)	0.44 (0.07)
GRULSIF α=0.5	50	1.00 (0.00)	0.89 (0.06)	0.94 (0.03)
Pool α=0.5	50	1.00 (0.00)	0.27 (0.02)	0.43 (0.03)
RULSIF α=0.1	50	1.00 (0.00)	0.55 (0.71)	0.71 (0.04)
ULSIF	50	1.00 (0.00)	0.60 (0.05)	0.75 (0.04)
KLIEP	50	1.00 (0.00)	0.58 (0.05)	0.40 (0.05)
MMD median	50	0.66 (0.24)	0.83 (0.11)	0.72 (0.17)
MMD aggreg	50	0.79 (0.15)	0.86 (0.08)	0.82 (0.09)

7. Conclusions

GRULSIF

- A novel non-parametric framework for multiple likelihood-ratios estimation
- A detailed and efficient implementation that is conveniently scalable for big graphs

Collaborative two-sample test:

- A detailed procedure that identifies the best hyperparameters and estimates node-level p-values
- A collaborative two-sample test, which outperforms non-parametric approaches that does not take the graph into account

8. Acknowledgments

This work was supported by the Industrial Data Analytics and Machine Learning (IdAML) Chair hosted at ENS Paris-Saclay, University Paris-Saclay, and grants from Région Ile-de-France.

References

[1] A. de la Concha, A. Kalogeratos, and N. Vayatis. Collaborative likelihood-ratio estimation over graphs, 2022.

[2] G. S. Fishman. *Monte Carlo*. Springer New York, 1996.

[3] J. Neyman, E. S. Pearson, and K. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[4] X. Nguyen, M. J. Wainwright, and M. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, 2008.

[5] A. Tartakovsky, I. Nikiforov, and M. Basseville. *Sequential Analysis: Hypothesis Testing and Change-point Detection*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, CRC Press, 2014.