Collaborative likelihood-ratio estimation over graphs

Alejandro de la Concha, Argyris Kalogeratos, Nicolas Vayatis

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

Likelihood-ratio and statistics

Definition

Consider a feature space $\mathcal{X} \subset \mathbb{R}^n$ and two probability distributions P and Q $(Q \ll P)$ such that they admit density functions $p(x)$ and $q(x)$ with respect to dx, then the **likelihood-ratio** is defined as:

$$r(x) = \frac{q(x)}{p(x)} \quad x \in \mathcal{X}$$

Applications

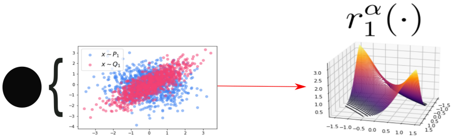
- *Hypothesis Testing* (Neyman-Pearson Lemma [Neyman et al., 1933])
- *Sequential Change-point detection* [Page, 1954, Shiryaev, 1963]
- *Transfer Learning (Importance sampling [Fishman, 1996])*

Existing methods

Usual non-parametric approach

$r(\cdot)$ is the solution of an optimization problem defined in terms of a functional space $\mathbb{H}$ (RKHS) [Nguyen et al., 2008a, Sugiyama et al., 2012]

Single source or Data aggregation



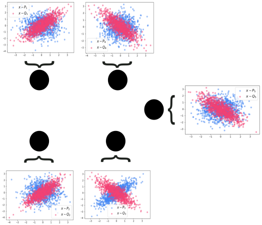
What about multiple interrelated sources of data?

Indicative use-cases

- Spatial statistics and learning:
  e.g. Monitoring geophysical phenomena using a sensor network, public health surveillance, transport network analysis, ...

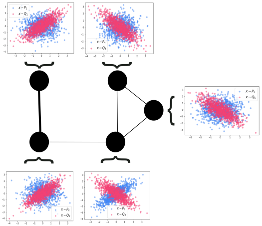
Motivation and main contribution

1. Local information can be important and **data aggregation** may hide phenomena of interest
2. **Ignoring inter-dependencies** may lead to a weaker performance than a multitasking approach



Main contribution

Graph-based Relative Unconstrained Least-Squares Importance Fitting (GRULSIF): a framework addressing both limitations by integrating a graph component into the estimation of multiple likelihood-ratios



Problem statement

Question How to infer a vector of likelihood-ratio functions $\mathbf{r}(X) = (r_1(x_1), ..., r_N(x_N))$ from interrelated data sources?

Setting

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

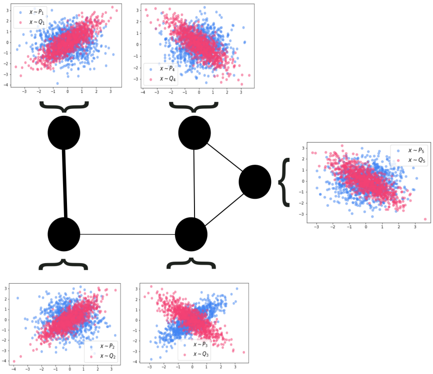


Figure: Example of different interrelated sources of information. In this case $\mathcal{X} = \mathbb{R}^2$

Problem statement

Framework

- Infer the node-level relative likelihood-ratios $r_v^{\alpha}(\cdot) = \frac{q_v(\cdot)}{(1-\alpha)p_v(\cdot)+\alpha q_v(\cdot)}$
- Non-parametric estimation based on a kernel function $K(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{+}$ to avoid making hypotheses on the nature of $q_v(\cdot)$ and $p_v(\cdot)$
- Capitalize over the information provided apriori by the graph: $\|r_u - r_v\|_{\mathbb{H}} < \epsilon$ if $W_{uv} \neq 0$

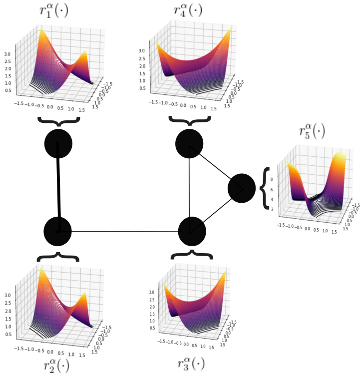


Figure: We aim to estimate the set of functions $\{r_v^{\alpha}(\cdot)\}_{v \in V}$ by exploiting $G = (V, E)$

Application: Collaborative two-sample testing

Problem statement

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

where C is a subset of nodes, such that the vector $r^{\alpha}(X) = (r_1^{\alpha}(x), \ldots, r_N^{\alpha}(x))$ is a smooth signal over the graph G

Potential applications:

1. Detection of an earthquake
2. Characterization of the nature of an epidemic
3. Spot regions of pollution peak

Non-parametric estimation with a single data source

Definition

A ϕ-divergence functional quantifies the similarity between two probability measures that are described by their $p(x)$, $q(x)$ in X:

$$\mathcal{D}_\phi(P,Q) = \int p(x)\,\phi(r(x))\,dx \geq \sup_{f \in \mathcal{F}} \int f dQ - \int \phi^*(f) dP \quad (1)$$

where $\mathcal{F}$ is a functional space [Nguyen et al., 2008b].

If we choose the Pearson-Divergence $\phi(x) = \frac{(x-1)^2}{2}$ and $F = \mathbb{H}$ (RKHS), then we aim to infer the likelihood-ratio $r(\cdot)$ by solving:

$$\min_{f \in \mathbb{H}} \int \frac{f^2(x)}{2} dP(x) - \int f(x) dQ(x)$$

Multitasking formulation of the problem

By the reproduction property of $\mathbb{H}$, for each $v \in V$, f_v takes the form:

$$f_v(x) = \sum_{l=1}^{L} \theta_{v,i} K(x, x_i)$$

Then we estimate the vector-valued function $f = (f_1, ..., f_N) \in \mathbb{H}^N$ under the hypothesis that θ_u and θ_v are expected to be similar if nodes u and v are connected in G:

$$\min_{\Theta \in \mathbb{R}^{NL}} \frac{1}{N} \sum_{v \in V} \frac{\mathbb{E}_{p_v^\alpha(x)}[(r_v^\alpha(x) - f_v(x))^2]}{2} + \frac{\lambda\gamma}{2} \sum_{v \in V} \|\theta_v\|^2 + \frac{\lambda}{2} \sum_{u,v \in V} W_{uv} \|\theta_u - \theta_v\|^2$$

Main results and contribution [de la Concha et al., 2022]

❶ **GRULSIF: Graph-based Relative Unconstrained Least-Squares Importance Fitting**

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

❷ **Collaborative two-sample test**

- A detailed procedure that includes best hyperparameters indentification and p-values estimation
- A collaborative two-sample test that outperforms non-parametric approaches that does not take the graph into account: KLIEP [Sugiyama et al., 2007], ULSIF [Sugiyama et al., 2011] RULSIF [Yamada et al., 2011], MMD [Gretton et al., 2012] MMD aggregated [Schrab et al., 2021]

Thank you

References I

Alejandro de la Concha, Argyris Kalogeratos, and Nicolas Vayatis. Collaborative likelihood-ratio estimation over graphs, 2022. URL `https://arxiv.org/abs/2205.14461`.

George S. Fishman. *Monte Carlo*. Springer New York, 1996.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25): 723–773, 2012.

Jerzy Neyman, Egon Sharpe Pearson, and Karl 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.

XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, 2008a.

XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, 2008b.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954.

References II

Alejandro de la Concha, Argyris Kalogeratos, and Nicolas Vayatis. Collaborative likelihood-ratio estimation over graphs, 2022. URL https://arxiv.org/abs/2205.14461.

George S. Fishman. *Monte Carlo*. Springer New York, 1996.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25): 723–773, 2012.

Jerzy Neyman, Egon Sharpe Pearson, and Karl 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.

XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, 2008a.

XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, 2008b.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954.