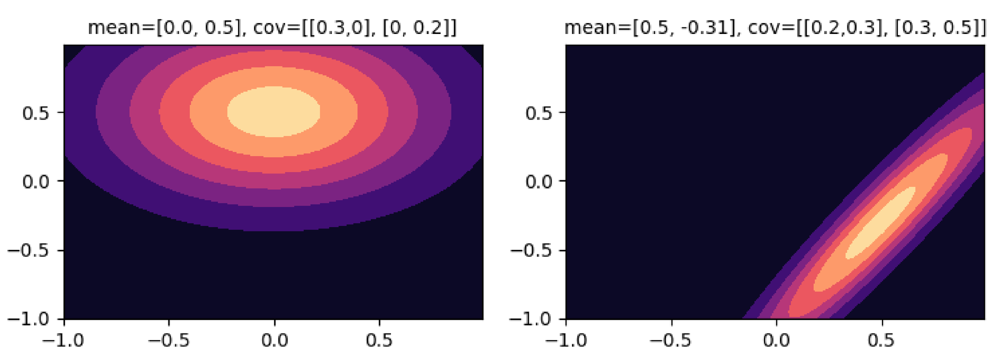
Bayesian optimization of the PC algorithm for learning Gaussian Bayesian networks

Irene Córdoba, Eduardo C. Garrido Merchán, Daniel Hernández-Lobato, Concha Bielza, Pedro Larrañaga
Universidad Politécnica de Madrid, Universidad Autónoma de Madrid

1 - Introduction

Bayesian Networks (BN) serves as a **compact representation** between variables in a domain:

- **Conditional Independences** are encoded by missing edges in a directed acyclic graph (DAG).
- They yield a **Modular Factorization of the Joint Probability Distribution** over the data.
- **Gaussian Bayesian Networks** are **Gaussian Multivariate Distributions**.



Gaussian Bayesian Network fitting to the data includes:

- **Structure Learning**: Recovering the graph structure. *(Combinatorial Space Search)*
- The **PC algorithm** determines absent edges in the DAG, using *statistical tests* and a *significance level*.

We can solve the *search for the optimum* in this space using *an optimization scenario*!

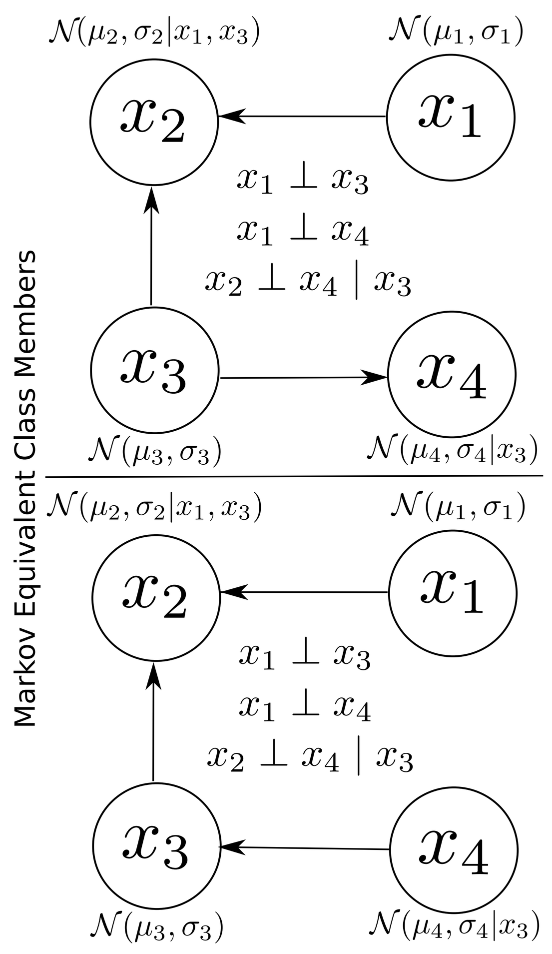
Very expensive evaluations	Noisy evaluations	Gradients are not available

Bayesian Optimization (BO) arises as an *ideal solution*.

2 - Gaussian Bayesian Networks and the PC algorithm

We want to *reconstruct* the skeleton of a **GBN** from data optimizing the *PC algorithm*.

- The *PC algorithm* first estimates the skeleton and then orientates it.
- Starts with the *complete graph* and in a *backward stepwise elimination* fashion it **removes edges**.
- For every *node* X_i from the graph, it looks every neighbor of it X_j and test $X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_C$.
- If the test succeds, the edge is removed.
- If the DAG is big, the PC algorithm is computationally very expensive and in order to select its hyperparameters, BO is a good solution.



Algorithm 1 The PC algorithm in its population version

Input: Conditional independence information about $\boldsymbol{X} = (X_1, \ldots, X_p)$
Output: Skeleton of the Gaussian Bayesian network

1: $G \leftarrow$ complete undirected graph on $\{1, \ldots, p\}$
2: $l \leftarrow -1$
3: **repeat**
4: $\quad l \leftarrow l + 1$
5: $\quad$ **repeat**
6: $\quad\quad$ Select i such that $(i, j) \in E$ and $|\mathrm{ne}(i) \setminus \{j\}| \geq l$
7: $\quad\quad$ **repeat**
8: $\quad\quad\quad$ Choose new $C \subseteq \mathrm{ne}(i) \setminus \{j\}$ with $|C| = l$
9: $\quad\quad\quad$ **if** $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_C$ **then**
10: $\quad\quad\quad\quad E \leftarrow E \setminus \{(i, j), (j, i)\}$
11: $\quad\quad\quad$ **end if**
12: $\quad\quad$ **until** (i, j) has been deleted or all neighbor subsets of size l have been tested
13: $\quad$ **until** All $(i, j) \in E$ such that $|\mathrm{ne}(i) \setminus \{j\}| \geq l$ have been tested
14: **until** $|\mathrm{ne}(i) \setminus \{j\}| < l$ for all $(i, j) \in E$

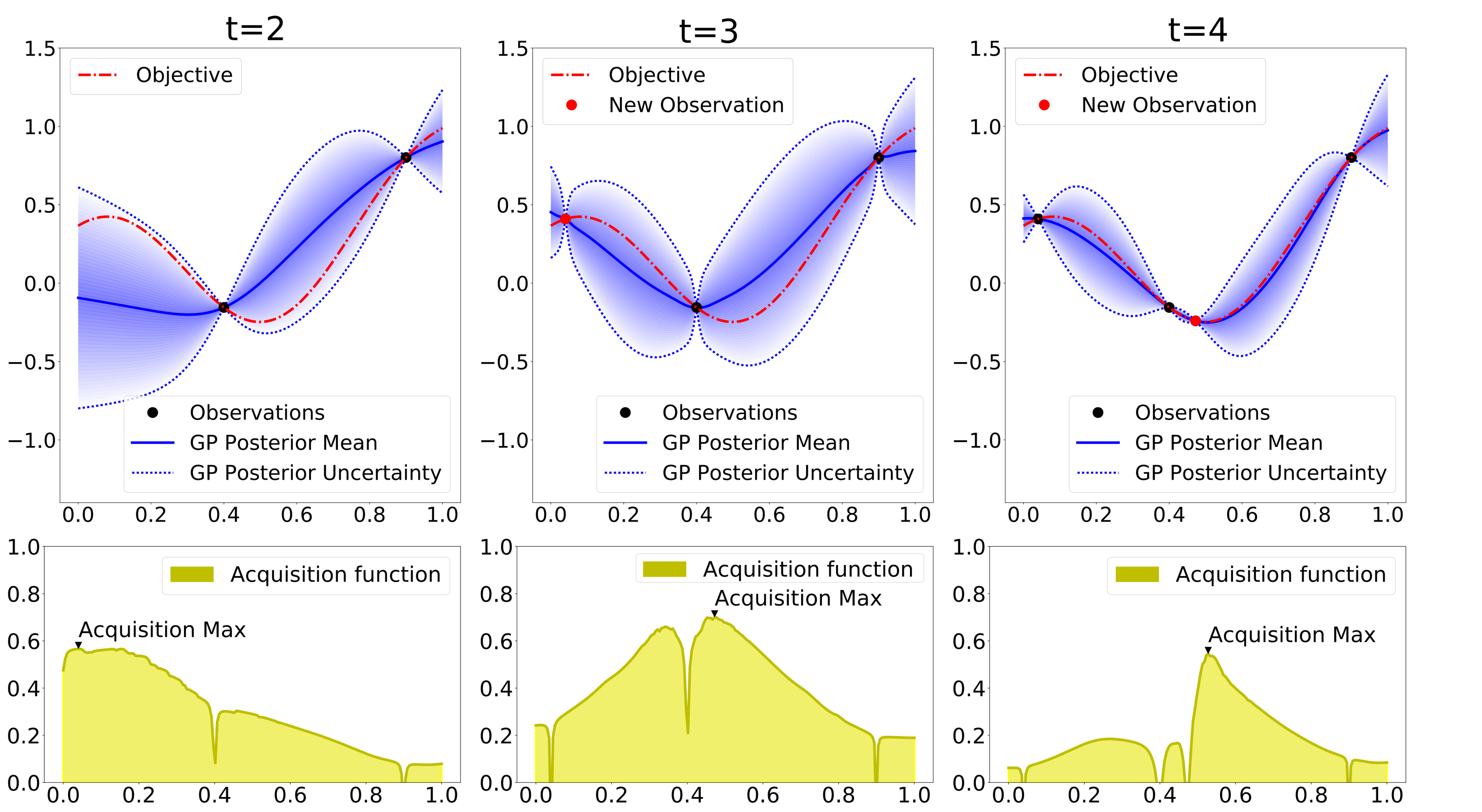
- We will consider different size networks and number of neighbours to retrieve the optimum significance level and statistical test for the PC algorithm.
- From a number of *samples* from a **BN**, we want to obtain through the *PC algorithm* a Markov equivalent **BN**.

3 - Bayesian Optimization and the normalized SHD metric

We optimize the **normalized SHD metric** in the search space of significance level and statistical tests using **BO**.

for $t = 1, 2, 3, \ldots,$ *max_steps* **do**
1: Find the next point to evaluate by optimizing the acquisition function:
$$\mathbf{x}_t = \arg\max_{\mathbf{x}} \alpha(\mathbf{x}|\mathcal{D}_{1:t-1}).$$
2: Evaluate the black-box objective $f(\cdot)$ at $\mathbf{x}_t$: $y_t = f(\mathbf{x_t}) + \epsilon_t$.
3: Augment the observed data $\mathcal{D}_{1:t} = \mathcal{D}_{1:t-1} \bigcup \{\mathbf{x}_t, y_t\}$.
4: Update the Gaussian process model using $\mathcal{D}_{1:t}$.
end
Result: Optimize the mean of the Gaussian process to find the solution.

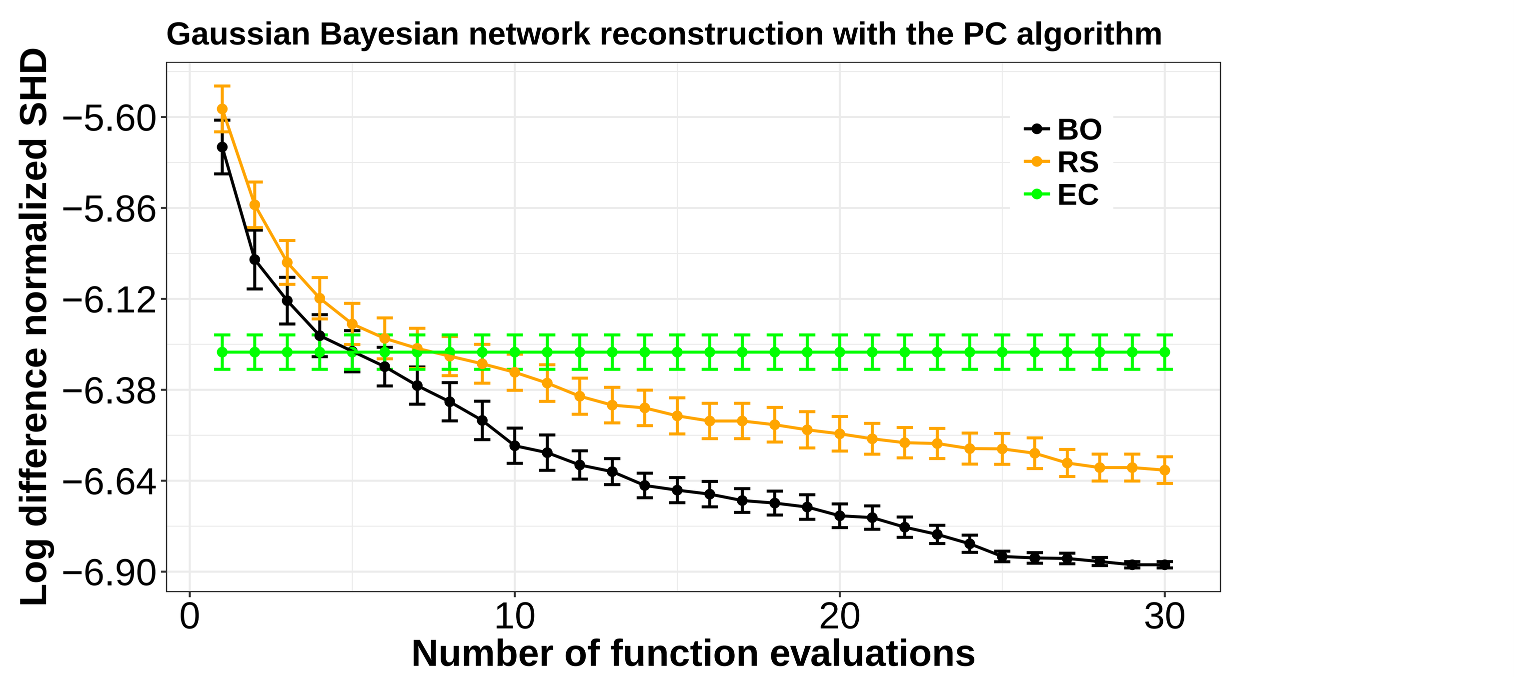
- The **normalized SHD metric** takes into account the **Markov equivalence** property of **DAGs**.
- It counts *the operations* to transform the **Markov equivalence class** of a **DAG** into another.
- Minimizing the **SHD** through **BO** will obtain a similar **BN**!



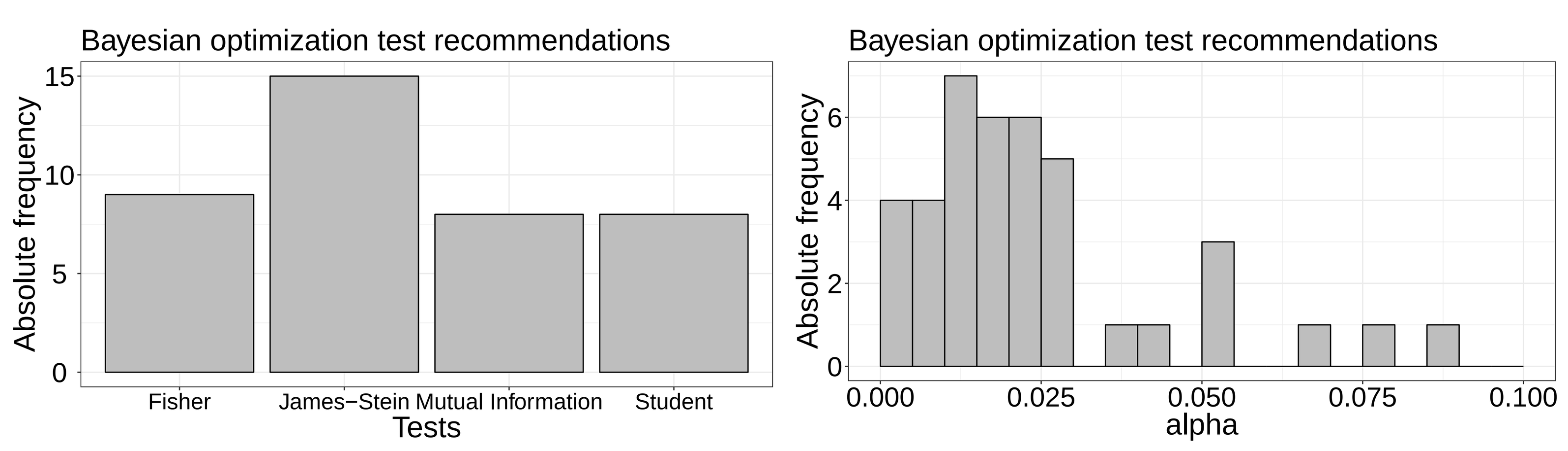
- The significance level is a continuous variable in the logarithmic space $[-5, -1]$.
- The statistical tests is a categorical variable with four options appearing in the literature.

4 - Experiments

We use **Spearmint** and the **PES** acquisition function to launch 100 different experiments obtaining the following results.



The **Normalized SHD metric** given by **BO** outperforms the expert and random search metrics.
We observe that the **James-Stein** test and a significance level of 0.01 are the preferred values for the hyperparameters.



5 - Conclusions and further work

- We used **BO** for selecting the optimal parameters of the **PC algorithm** for structure recovery in **BNs**.
- Expert suggestion is *outperformed*, surprising result of the *statistical test*.
- We plan to explore other objective measures, not relying on the true graph structure.
- Extend to *constrained multi-objective scenarios* for creating *specific Bayesian Networks* and *Graphical Models*.