An Oracle Approach for Interaction Neighborhood Estimation

Matthieu Lerasle

NUMEC, IME, USP

with D. Y. Takahashi (Princeton)

A motivation from neuroscience

We study a population of neurons, hereafter represented by a discrete set S. We want to infer, from the observation of the neural activity, the structure of the connected neurons.
The neural activity is represented by a configuration x in A^S, where A is a finite alphabet of answers.
The relations between neurons are encoded by a probability measure P on A^S. More precisely, we say that a site j is connected to i if there exists a configuration x such that

$$P(x(i)|x(k),\ k \in S/\{i\}) \neq P(x(i)|x(k),\ k \in S/\{i,j\}).$$

The system is partially observed, *i.e.* we can only observe the values of the configurations on a finite subset V_M of S. The sample is defined by $X_{1:n}(V_M) = (X_1(j), ..., X_n(j))_{j \in V_M}$, where $X_1, ..., X_n$ are i.i.d P.

Estimation of interacting neighborhood, some previous works

For all i in S, let us denote by S_i the set of j interacting with i. In order to estimate S_i, the methods usually required strong assumptions on P or on S_i.

S_i should be finite, with a known upper bound in *Bresler et. al. (2008), Csizar and Talata (2006), Bento and Montanari (2009)*.

S_i can be infinite in *Galves et. al. (2010), Ravikumar et. al (2009)* but P has to be a ferromagnetic Gibbs measure. Moreover, in *Ravikumar et. al (2009)*, S_i should satisfy an incoherence assumption and in *Galves et. al. (2010)*, the strongest interacting points have to belong to a small (of order $O(\ln n)$) fixed neighbor of i.

The oracle Approach

We propose an alternative approach to the problem. Instead of estimating directly the set of interacting points, we focus on the estimation of the function $P_{i|S}$ defined, for all x in A^S by

$$P_{i|S}(x) = P(x(i)|x(j),\ j \in S/\{i\}).$$

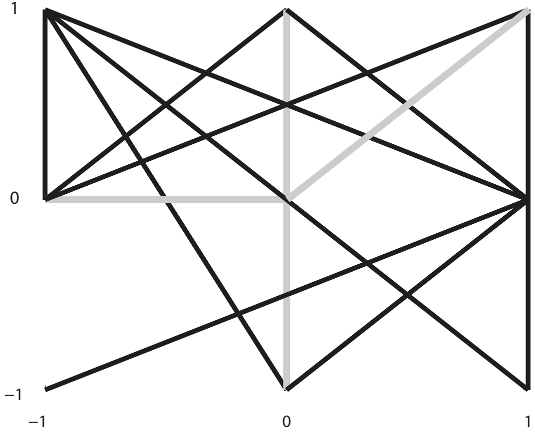
Let $\widehat{P}$ be the empirical measure. For all subsets V of V_M, $\widehat{P}_{i|V}$, defined, for all configurations x by

$$\widehat{P}_{i|V}(x) = \widehat{P}(x(i)|x(j),\ j \in V/\{i\})$$

is an estimator of $P_{i|S}$. The oracle approach consists in the research of the set $\widehat{V}$ such that the risk of $\widehat{P}_{i|\widehat{V}}$ is the smallest among all the estimator $(\widehat{P}_{i|V})_{V \subset V_M}$.

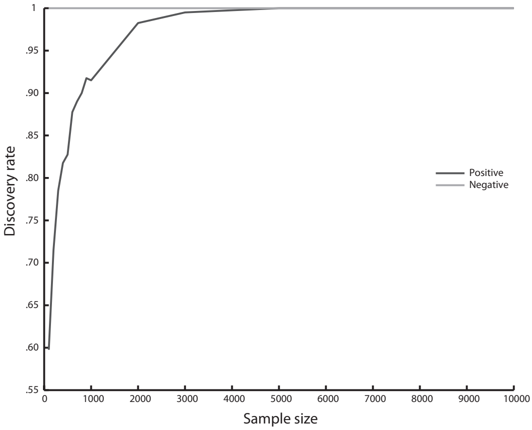
Example

Consider the Ising model with the following interaction graph



Example

The following graph shows the positive and negative discovery rates of an oracle



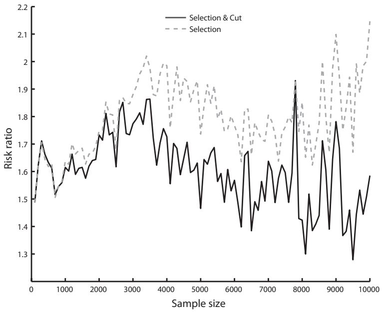
Overview of the results

The following results have been obtained.

- *In the L_∞ case, the following results are proved in a generalization of the Ising models. see arXiv:1010.4783*
    1. We build an estimator $\widehat{V}$ and we prove that it satisfies an oracle inequality, *i.e.*, with large probability
    $$\left\|P_{i|S} - \widehat{P}_{i|\widehat{V}}\right\|_\infty \leq C_\star \inf_{V \subset V_M} \left\|P_{i|S} - \widehat{P}_{i|V}\right\|_\infty .$$
    2. We deduce from an oracle $\widehat{V}$ a consistent estimator of S_i (two-steps method).
    3. We obtain an efficient algorithm to compute our estimator and prove its validity in the Ising model.
- *In the L_2 case and in the Küllback case, the following results have been obtained without any restriction on the random field.*
    1. We build an estimator $\widehat{V}$ and we prove that it satisfies an oracle inequality.
    2. We justify the slope heuristic, that allows to optimize our penalization methods for model selection.
    3. The links between oracle and consistent estimator, and the efficient algorithm can still be used but we do not justify it at this level of generality.

Example

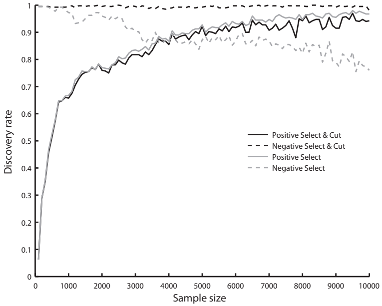
The following graph shows the performances, in the L_∞-case, of $\widehat{V}$ in terms of oracle properties.



Plot of the number of samples n against the average of ratio $\frac{\left\|\widehat{P}_{i|\widehat{V}} - P_{i|S}\right\|_\infty}{\inf_{V\subset S}\left\|\hat{P}_{i|V} - P_{i|S}\right\|_\infty}$.

The IN Problem

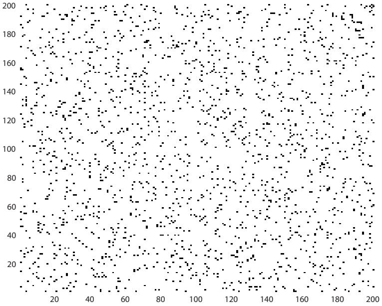
The following graphs show the performances, in the L_∞-case, of our selected model in terms of discovering properties.



Discovery rates of $\widehat{V}$ and the two-steps estimator.

The efficient algorithm

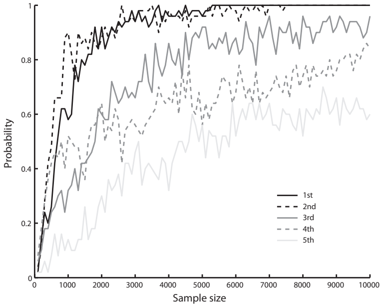
In order to emphasize the advantages of the efficient algorithm, we run it on an Ising model with the following graph of interaction.



A more realistic example

Discovery rates

The following graph shows the discovery rates of the strongest interaction, the 2nd strongest,... of our efficient algorithm when the number of data increases.



Discovery rates

The end

Thank you very much!!!