INRIA Saclay, Équipe Parietal http://team.inria.fr/parietal bat 145, CEA Saclay

Statistical control of sparse linear models

Research theme: machine learning, functional brain imaging
Keywords: regression, fMRI, optimal transport.
Duration & salary: 4 to 6 months, between 500 €and 800 € monthly
Research team: Parietal (INRIA Saclay and CEA)
Adviser: Bertrand Thirion
Contact: bertrand.thirion@inria.fr
Application: Interested candidate should send CV and motivation letter

Context: Inferential statistics give a probabilistic control on the selection of variables associated with a target of interest. However, they are mostly limited to univariate models, that ignore the structure of the input variables (smoothness, correlation, complementary information). By contrast, multivariate inference has become popular in the framework of regularized regression (lasso, ridge Elastic Net and more complex avatars [2]), however, they do not bring guarantees on the variables selected. While some solutions have recently been proposed in the case where the number of features is not too large, they are inefficient in the *large p, small n* regime encountered in many domains, e.g. brain imaging or genomics.
The most promising solutions so far consist in a combination of geometric arguments with convex optimization [4], yet its usefulness has not yet been assessed in large p settings. Moreover, it likely suffers from standard false negative issues in the presence of correlations.
Proposed work: We will blend together the procedures described in [3] with those of [4], to achieve some guarantees on the detections without resorting to non-parametric inference. We will consider both the ideas in [4] and [1] to build the most effective and powerful test that gives accurate specificity control. We will consider carefully the hypotheses (independence, pivotality) that underlie the test.
We will then proceed by first establishing that the statistical procedures have an accurate control of false detections (i.e. in the absence of any true association). We will consider both theoretical and empirical arguments. Whenever possible, we will target the expected false discovery rate rather than the false alarms rate, as it is more meaningful to practitioners.
We will then compare the sensitivity of the tests to detect true effects on simulation and real data. When working with real data, we will rely on expensive, yet reliable non-parametric approaches to provide a solid comparison.
If the attempt is successful, we will release an implementation of the test in the Nilearn neuroimaging library nilearn.github.io.

Required skills: The successful candidate will know statistics and machine learning, and be able to implement analysis methods. In addition, he or she should be interested in understanding functional brain images and validating approaches on real data. Knowledge of scientific computing in Python (Numpy, Scipy) is encouraged. All the work will be done in Python.

[1] R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical Science.*, 2015.

[2] Michael Eickenberg, Elvis Dohmatob, Bertrand Thirion, and Gaël Varoquaux. Total Variation meets Sparsity: statistical learning with segmenting penalties. In *Medical Image Computing and Computer Assisted Intervention*, July 2015.

[3] Benoit Da Mota, Virgile Fritsch, Gaël Varoquaux, Tobias Banaschewski, Gareth J. Barker, Arun L.W. Bokde, Uli Bromberg, Patricia Conrod, Jürgen Gallinat, Hugh Garavan, Jean-Luc Martinot, Frauke Nees, Tomas Paus, Zdenka Pausova, Marcella Rietschel, Michael N. Smolka, Andreas Ströhle, Vincent Frouin, Jean-Baptiste Poline, and Bertrand Thirion. Randomized parcellation based inference. *NeuroImage*, 89:203 – 215, 2014.

[4] Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.