Approximate Message Passing for HIgh-dimensional data analysis (AMPHI)

Post-doc project

Teams implied, coordinator and partners

Partners

Research unit	Inria Parietal team	LTCI, Telecom Paris	Institut de Physique Théorique, CEA
Partner	Bertrand Thirion	Jospeh Salmon	Lenka Zdeborová.
Tel	+33 1 69 08 81 14	+33 1 45 81 75 47	+33 1 69 08 79 92
mail	bertrand.thirion@inria.fr	joseph.salmon@telecom-paristech.fr	lenka.zdeborova@cea.fr

Duration : 12 months (with possible extensions), starting end 2017.
Salary : 2100€per month net.

Summary of the project

In many scientific fields, the data acquisition devices have benefited of hardware improvement to increase the resolution of the observed phenomena, leading to ever larger datasets. While the dimensionality has increased, the number of samples available is often limited, due to physical or financial limits. This is a problem when these data are processed with estimators that have a large sample complexity, such as multivariate statistical models. In that case it is very useful to rely on structured priors, so that the results reflect the state of knowledge on the phenomena of interest. The study of the human brain activity through neuroimaging belongs among these problems, with up to 10^6 features, yet a set of observations limited by cost and participant comfort.

We are missing fast estimators for *multivariate models with structured priors*, that furthermore provide statistical control on the solution.

We want to join forces to design a new generation of inverse problem solvers that can take into account the complex structure of brain images and provide guarantees in the low-sample-complexity regime. To this end, we will first adapt alternating direction method of multipliers (ADMM) or Approximate Message Passing (AMP) methods to the brain mapping setting, using first simple convex priors regularizations. We will then consider more complex structured priors that control the variation of the learned image patterns in space [16] and non-convex priors. Crucial gains are expected from the use of the EM algorithm for parameter setting. We will also examine the estimation of parametric and non-parametric confidence intervals about the estimates.

AMPHI will design a reference inference toolbox released as a generic open source library. We expect a 3- to 10-fold improvement in CPU time with respect to current solutions, that will benefit to large-scale brain mapping investigations.

Detailed information

Context In many fields of physics or life sciences, improving the resolution of observations (signal, images, spectra) is a major endeavor, as it is a pre-requisite toward more accurate information. Data acquisition devices benefit thus of hardware improvement to increase the resolution of the observed phenomena, leading to ever larger datasets. From a statistical perspective, these datasets have a high dimensionality and the signals show some prominent structures that require adequate modeling. While the **dimensionality has increased**, the **number of samples available is sometimes limited**, due to physical or financial limits. This becomes a problem when these data are processed with estimators that have a large sample complexity, such as many multivariate estimators (classifiers, regression models, covariance estimators, structure learning). In that case it is very useful to rely on **structured priors**, so that the resulting models reflect the state of knowledge on the phenomena of interest. Well-chosen priors improve the accuracy of the models and decrease the sample complexity of the estimators.

The study of the human brain activity through functional Magnetic Resonance Imaging (MRI) belongs among these problems. The number of features per image reaches 10^5 to 10^6 —thanks to the rise of high-field MRI acquisitions that cross the mm scale— yet the number of observations is limited by the duration of scanning sessions and the number of subjects that can be included in studies. The key challenge addressed here is to **set up a novel generation of efficient techniques to enforce structured priors on high-resolution MRI datasets** to improve statistical analysis.

State-of-the art and positioning The members of the consortium have set up analytic tools for predictive modeling on these images, using the framework of convex M-estimators [11, 5, 2, 1, 3], obtaining state-of-the-art and nearly computationally optimal solutions. An example is given in Fig. 1.

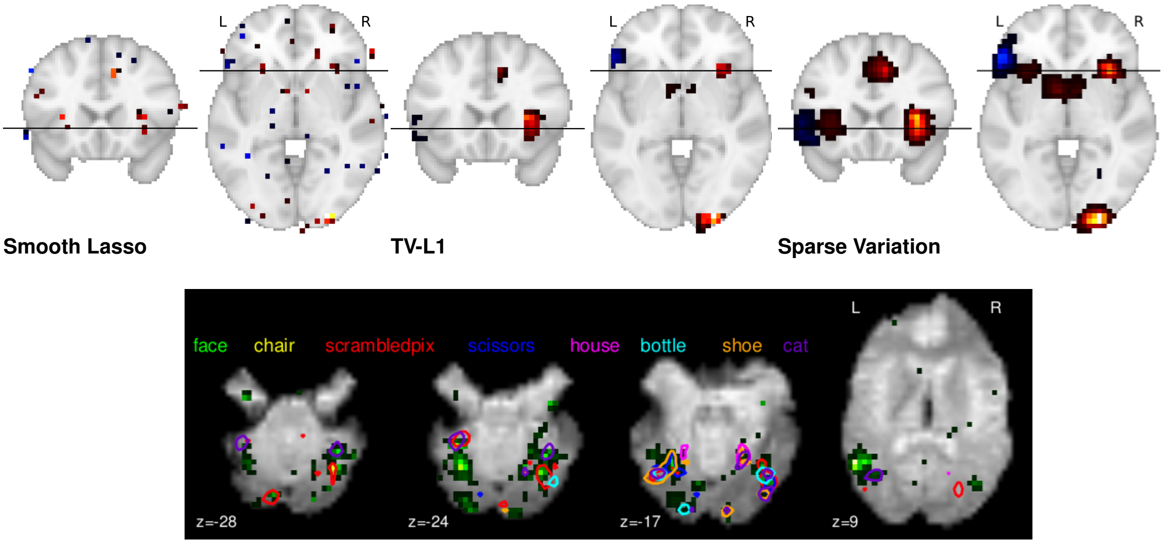

FIGURE 1 – (Top) Weight vectors from estimating gain from brain functional activity, on a mixed gambles task [15]. Prediction target is the gain proposed in a series of gambles proposed to the subjects. This inter-subject analysis shows broad regions of activation. Sophisticated estimators, such as Sparse variation and TV-l1, yield a clean map and obtain high accuracy. Parameter setting for these estimators is also easier than the cheaper Smooth Lasso estimator. See [3] for details. (Bottom) These effective estimators can be used to single out specialized regions in the ventral occipital cortex from [6], providing an elegant solution to fundamental brain mapping problems.

In brief, researchers of the consortium have applied structured penalties to multivariate estimators of brain activity with great success. Yet an essential problem remains computation time [1], because of the non-smooth optimization steps involved when imposing these penalties. Another major open question is to gain statistical control (e.g. confidence intervals, p-values) on the solutions of this problem : for instance, if a certain pattern of activity predicts an autism diagnosis, one would like to test whether a given brain region is significantly loaded by this

[1] Remember that neuroscientists have no easy access to intensive computation facilities, that require technical skills typically rare in neuroscience labs. The research is performed by PhD students that cannot develop expertise on all technical aspects, such as computer science.

pattern. High dimensional regression models rely on convex optimization tools. In particular, recent developments have allowed to achieve noticeable speed-ups for Lasso and multi-task Lasso solvers [13, 10]. Expertise on noise level estimation is also brought by the consortium, as this could be a key element for predictive tasks [12].

An important challenge is to set up efficient estimators for the tractability of the computation. Approximate Message Passing (AMP) techniques [14, 7, 8, 9] are known to be efficient, but fast implementations cannot deal with the complex correlation structure of the data, and extensions have to be developed to enforce their convergence. The challenge is thus to maintain computational efficiency while ensuring convergence.

Alternating direction method of multipliers (ADMM) approaches, that lead to a setting close to AMP, are known to be generic and convergent, but they suffer from slow convergence and hard parameter tuning [18]. However, recent works have addressed these limitations, making ADMM attractive again [18].

Jointly considering these approaches may bring new opportunities, as they have distinct advantages. For instance, AMP offers natural hyper-parameters tuning strategies, bypassing cumbersome and expensive cross-validation runs.

Objectives, planning and deliverable We thus propose to design **a new generation of inverse problem solvers** that can take into account the complex structure of brain images and provide guarantees in the low-sample regime where they are used. To this end, we will first adapt ADMM/AMP tools to the neuroimaging setting, using standard sparsity priors (a.k.a. ℓ_0 norm, Gauss-Bernoulli etc.) on the estimated model. In this we will follow the approach described in [19]. We will then consider more complex structured priors, that control the variation of the learned image patterns in space [17, 16]. This is related to the well known *analysis sparsity* framework [4] : the signals of interest should have a sparse representation in a well-chosen signal basis.

Crucial gains are expected from the use of the variational strategy for parameter setting (EM type of algorithm) that comes naturally with AMP approaches [8]. An important task will be to benchmark the resulting estimator against alternatives, in terms of prediction accuracy, support recovery and computation time. We expect different regimes, depending on the number of features or samples in the dataset and the noise level in the data. We will also examine the statistical guarantees provided by the AMP approach : are the confidence intervals returned by the estimators reliable, compared to e.g. bootstrapped estimators, so that non-expert neuroscientists can rely on them ?

The project will deliver a **technical publication** and a **software implementation** in Python, which, after a thorough assessment (code quality control, efficiency, simplicity of the API, documentation), will be included in the Nilearn open source library. In parallel, these contributions will serve as a basis for large-scale analyses carried out on human cognition based on open data repositories (http://neurovault.org, http://openfmri.org) ; in the longer term they will also be considered for other problems (image reconstruction, compressed sensing). The efficiency of the estimator will be crucial for the feasibility of the approach on large-scale data, as existing datasets size now scale in tens of Gigabytes and keep increasing.

Références

[1] Elvis Dohmatob, Michael Eickenberg, Bertrand Thirion, and Gaël Varoquaux. Speeding-up model-selection in GraphNet via early-stopping and univariate feature-screening. In *PRNI*, Stanford, United States, June 2015.

[2] Elvis Dohmatob, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Benchmarking solvers for TV-l1 least-squares and logistic regression in brain imaging. In *Pattern Recoginition in Neuroimaging (PRNI)*, Tübingen, Allemagne, April 2014. IEEE.

[3] Michael Eickenberg, Elvis Dohmatob, Bertrand Thirion, and Gaël Varoquaux. Total Variation meets Sparsity : statistical learning with segmenting penalties. In *Medical Image Computing and Computer Aided Intervention (MICCAI)*, Proceedings of MICCAI 2015, München, Germany, October 2015.

[4] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse problems*, 23(3) :947–968, 2007.

[5] Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *Pattern Recognition in Neuroimaging (PRNI)*, Philadelphia, États-Unis, June 2013. IEEE. ANR grant BrainPedia, ANR-10-JCJC 1408-01, FMJH Program Gaspard Monge in optimization and operation research with support from EDF.

[6] James V. Haxby, Ida M. Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293 :2425, 2001.

[7] Florent Krzakala, Marc Mézard, François Sausset, YF Sun, and Lenka Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2) :021005, 2012.

[8] Florent Krzakala, Marc Mézard, Francois Sausset, Yifan Sun, and Lenka Zdeborová. Probabilistic reconstruction in compressed sensing : algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics : Theory and Experiment*, 2012(08) :P08009, 2012.

[9] Andre Manoel, Florent Krzakala, Eric Tramel, and Lenka Zdeborovà. Swept approximate message passing for sparse estimation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1123–1132, 2015.

[10] M. Massias, A. Gramfort, and J. Salmon. From safe screening rules to working sets for faster lasso-type solvers. *CoRR*, abs/1703.07285, 2017.

[11] Vincent Michel, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, and Bertrand Thirion. Total variation regularization for fMRI-based prediction of behaviour. *IEEE Transactions on Medical Imaging*, 30(7) :1328 – 1340, February 2011.

[12] E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon. Efficient smoothed concomitant lasso estimation for high dimensional regression. In *NCMIP*, 2017.

[13] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparse multi-task and multi-class models. In *NIPS*, pages 811–819, 2015.

[14] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2168–2172. IEEE, 2011.

[15] Sabrina M Tom, Craig R Fox, Christopher Trepel, and Russell A Poldrack. The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811) :515–8, Jan 2007.

[16] Eric W Tramel, Angélique Drémeau, and Florent Krzakala. Approximate message passing with restricted boltzmann machine priors. *arXiv preprint arXiv :1502.06470*, 2015.

[17] Xing Wang and Jie Liang. Approximate message passing-based compressed sensing reconstruction with generalized elastic net prior. *Image Commun.*, 37(C) :19–33, September 2015.

[18] Zheng Xu, Mário A. T. Figueiredo, and Thomas Goldstein. Adaptive ADMM with spectral penalty parameter selection. *CoRR*, abs/1605.07246, 2016.

[19] Justin Ziniel, Philip Schniter, and Per Sederberg. Binary linear classification and feature selection via generalized approximate message passing. In *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*, pages 1–6. IEEE, 2014.