FORUM STIC Paris-Saclay 2014

Labex DigiCosme · digiteo · SystemX

ELVIS DOHMATOB, ALEXANDRE GRAMFORT, BERTRAND THIRION, GAEL VAROQUAUX

N° 4.16

BENCHMARKING SOLVERS FOR TV-ℓ_1 PENALIZED LOGISTIC AND LEAST SQUARES REGRESSION: APPLICATION TO BRAIN DATA

Learning predictive models from brain imaging data, as in decoding cognitive states from fMRI (functional Magnetic Resonance Imaging), is typically an ill-posed problem as it entails estimating many more parameters than available sample points. This estimation problem thus requires regularization. Total variation regularization, combined with sparse models, has been shown to yield good predictive performance, as well as stable and interpretable maps. However, the corresponding optimization problem is very challenging. Here we explore a wide variety of solvers and exhibit their convergence properties on fMRI data. Our findings show that care must be taken in solving TV$-\ell_1$ estimation in brain imaging. We highlight the successful strategies.

PROBLEM STATEMENT

$$\left.\begin{array}{l}\textbf{Minimize } E(w) := \mathcal{L}(X, y, w) + \alpha\left(\rho\|w\|_1 + (1-\rho)\, TV(w)\right) \\ \quad\textbf{For } w \in \mathbb{R}^p\end{array}\right\} \tag{1}$$

where :

- $\mathcal{L}(X, y, w)$ is the loss term = the loss incurred by using the brain map $w \in \mathbb{R}^p$ to predict a sample $y \in \mathbb{R}^n$ of n responses from a sample $X \in \mathbb{R}^{n\times p}$ of n corresponding brain images.
  - X is commonly called the design matrix whilst y is the response variate.
  - $\mathcal{L}(X, y, w) \equiv \frac{1}{2}\|y - Xw\|_2^2$ for linear regression, etc.
  - $n \ll p$ for brain data (high-dimensional problem) $\implies$ need for regularization
    - Typically, $n \sim 10 - 10^2$ brain images and $p \sim 10^4 - 10^6$ voxels
- $\alpha\left(\rho\|w\|_1 + (1-\rho)\, TV(w)\right)$ is the regularization (aka penalty) term (Michel et al. (2011); Baldassarre et al. (2012); Gramfort et al. (2013)) :
  - Encodes "sparsity + spatial structure" prior on the optimal brain map $\hat{w}$.
  - $\alpha \geq 0$: overall amount of regularization (tradeoff between data and regularization).
  - $\|w\|_1$: ℓ_1-norm of w defined by $\|w\|_1 := \sum_{j\in voxels}|w_j|$.
  - TV(w) : the isotropic Total-variation of w defined by
    $TV(w) := \|\nabla(w)\|_{21} := \sum_{j\in voxels}\sqrt{(\nabla^x w)_j^2 + (\nabla^y w)_j^2 + (\nabla^z w)_j^2}$, and
    $\nabla := [\nabla^x, \nabla^y, \nabla^z]^T \in \mathbb{R}^{3p\times p}$ is the 3D discrete spatial gradient operator.
  - $\rho \in [0, 1]$: also called the ℓ_1-ratio controls the tradeoff between sparsity (enforced by the minimizing $\ell_1(w)$) and spatial structure (enforced by minimizing $TV(w)$).

NEED FOR FAST SOLVERS

Problem (1) is a high-dimensional non-smooth convex optimization problem and calls for novel optimization techniques.

- We focus on convergence time for small tolerance
- Lack of fast solver and explicit control of tolerance for problem (1) can lead to brain maps and conclusions that reflect properties of the solver more than of the TV-ℓ_1 solution !

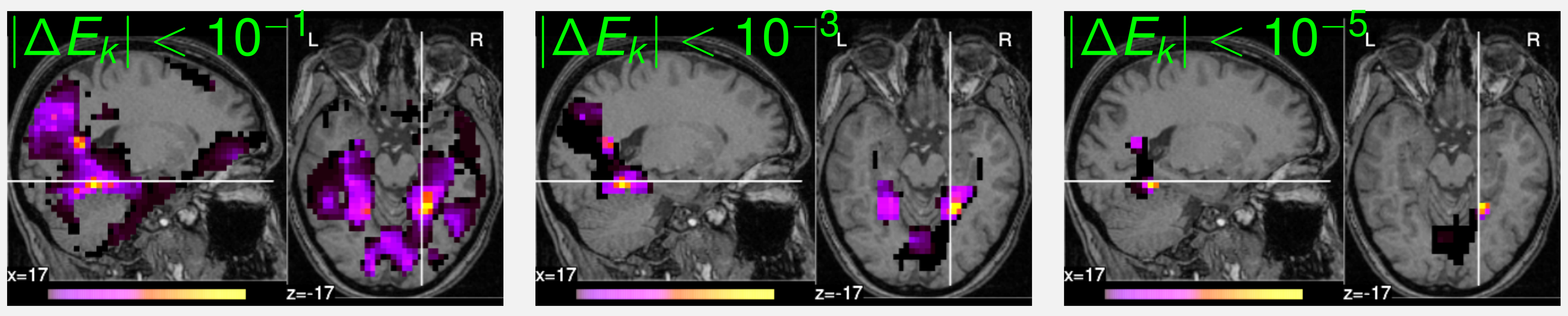

FIGURE : Optimal TV$-\ell_1$ brain maps for the face-house discrimination task on the Haxby et al. (2001) dataset, for various levels of tolerance. Dohmatob et al. (2014)

RESULTS

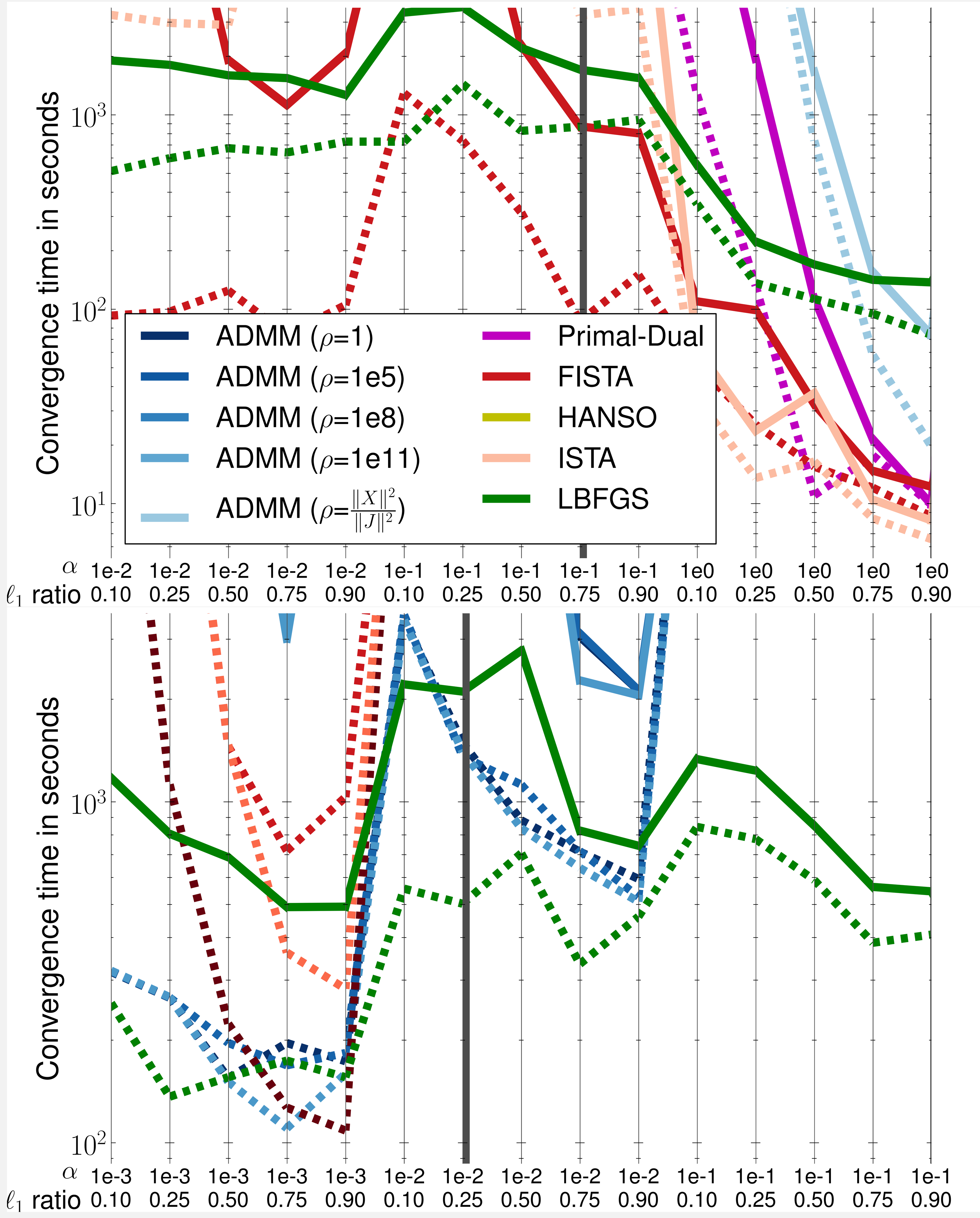

FIGURE : Benchmarks on Haxby et al. (2001) dataset. **Top** : TV-ℓ_1 penalized least squares regression. **Bottom** : Logistic regression. See Dohmatob et al. (2014)

.

CONCLUSION

- TV$-\ell_1$ penalized regression for brain imaging leads to very high-dimensional, non-smooth and very ill-conditioned optimization problems.
- We have presented a comprehensive comparison of state-of-the-art solvers (ADMM, ISTA, FISTA, HANSO, LBFGS, etc.) in these settings.
- Solvers were implemented with all known algorithmic improvements and implementation were carefully profiled and optimized.
- The implemented solvers are part of the open-source Python library *nilearn* : `http://www.github.com/nilearn/nilearn`.
- Our results outline best solvers : monotonous FISTA with a adaptive control of the tolerance of the TV proximal operator, in the case of squared loss ; smoothed quasi-newton (for example LBFGS here) based on surrogate upper-bounds of the non-smooth TV-ℓ_1 penalty, for logistic loss.

REFERENCES

Baldassarre et al. "Structured sparsity models for brain decoding from fMRI data". In *PRNI*, page 5, 2012.
Dohmatob et al. "Benchmarking solvers for TV-l1 least-squares and logistic regression in brain imaging". *PRNI*, 2014.
Gramfort et al. "Identifying predictive regions from fMRI with TV-L1 prior". In *PRNI*, 2013.
Haxby et al. "Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex". *Science*, 293 :2425, 2001.
Michel et al. "Total variation regularization for fMRI-based prediction of behaviour". *IEEE Transactions on Medical Imaging*, 30 :1328, 2011.

PARIETAL · Inria informatics mathematics · NeuroSpin