Spatial regularization and sparsity for brain mapping

Bertrand Thirion,
INRIA Saclay-Île-de-France, Parietal team
http://parietal.saclay.inria.fr
bertrand.thirion@inria.fr
FMRI data analysis pipeline
Statistical inference & MVPA

Question 1 : Is there any effect ? → omnibus test
  MVPA: Can I discriminate btw the two conditions ?
Question 2 : What regions actually display a difference btw the two conditions ?
  MVPA: Support of the discriminative pattern ?
Outline

• Machine learning techniques for MVPA in neuroimaging
• Improving the decoder: smoothness and sparsity
• Recovery and randomness.
Reverse inference: combining the information from different regions

Aims at decoding brain activities → predicting a cognitive variable

[Dehaene et al. 1998], [Haxby et al. 2001], [Cox et al. 2003]
Predictive linear model

\[ y = f(X, w, b) + \text{noise} \]

\( y \) is the behavioral variable.

\( X \in \mathbb{R}^{n \times p} \) is the data matrix, i.e. the activations maps

\((w, b)\) are the parameters to be estimated.

\( n \) activation maps (samples), \( p \) voxels (features).

\[ y \in \mathbb{R}^n \rightarrow \text{regression setting:} \quad f(X, w, b) = Xw + b, \]

\[ y \in \{-1, 1\}^n \rightarrow \text{classification setting:} \quad f(X, w, b) = \text{sign}(Xw + b), \]

where “sign” denotes the sign function.
Curse of dimensionality in MVPA

- **Problem:** $p \gg n$
  - Overfit the noise on the training data

- **Solutions**
  - **Prior region selection**
    - Prior selection of brain regions → prior-bound result

- **Data-driven feature selection** (e.g. Anova, RFE):
  - Univariate methods (Anova) → no optimality?
  - Multivariate methods → combinatorial pb, computational cost

- **Regularization** (e.g. Lasso, Elastic net):
  - Shrink $w$ according to your prior
Training a predictive model

- **Learning $w$ from a given training set $(y, X)$**
  \[ \hat{w} = \arg\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i, X_i w) + \lambda J(w) \]

- **Choice of the loss**
  - Regression: Least-squares, Hinge, Huber
  - Classification: Hinge, logistic

- **Choice of the regularizer**
  - Convex setting: a norm on $w$
  - Bayesian setting: prior distribution on $w$
Evaluation of the decoding

Prediction accuracy

Coefficient of determination $R^2$:

$$R^2(y^t, \hat{y}^t) = \frac{1}{N} \sum_{i=1}^{N} \frac{(y_i^t - \hat{y}_i^t)^2}{\text{var}(y^t)}$$

Classification accuracy $\kappa$:

$$\kappa(y^t, \hat{y}^t) = \frac{1}{N} \sum_{i=1}^{N} \delta(y_i^t - \hat{y}_i^t)$$

→ Quantify the amount of information shared by the pattern and $y$.

Layout of the resulting maps of weights: Do we have any guarantee to recover the true discriminative pattern?

Common hypothesis = segregation into functionally specific territories

→ sparse: few relevant regions implied

→ compact structure: grouping into connected clusters.
You said: recovery?

× MVPA cannot recover the true sources as it aims at finding a good discriminative model ("filters"), not at estimating the signal.
× A correction taking covariance structure is necessary

✔ However, this can be improved by choosing relevant priors
✔ You might want to have a discriminative model that makes sense to you

[Haufe et al. NIMG 2013]
Outline

- Machine learning techniques for MVPA in neuroimaging
- Improving the decoder: smoothness and sparsity
- Recovery and randomness.
Regularization framework

\( w = \text{the discriminative pattern} \)

Constrain \( w \) to select few parameters that explain well the data.

→ Penalized regression

\[
\hat{w} = \arg\min_{w \in \mathbb{R}^p} \ell(y, Xw) + \lambda J(w), \quad \lambda \geq 0
\]

\( \ell(y, Xw) \) is the loss function, usually \( \|y - Xw\|^2 \) for regression

\( \lambda J(w) \) is the penalization term.

\[
\begin{align*}
\lambda J(w) &= \lambda \|w\|_2^2 & \text{Ridge (no sparsity)} \\
\lambda J(w) &= \lambda \|w\|_1 & \text{Lasso (very sparse)} \\
\lambda J(w) &= \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2 & \text{Elastic net (sparsity + grouping)} \\
\lambda J(w) &= \lambda_1 \|w\|_1 + \lambda_2 \|\nabla w\|_2^2 & \text{Smooth lasso (sparsity + smoothness)} \\
\lambda J(w) &= \lambda_1 \|w\|_1 + \lambda_2 \|\nabla w\|_1 & \text{Total variation (piecewise sparsity)}
\end{align*}
\]
Priors and penalization: Brain decoding = engineering problem?

Prior on the relevant activation maps

Penalization in regularized regression

Design of a norm $\|w\|$ to be minimized

Example: Total Variation penalization

[Michel et al. 2011]
Do we need to bother about sparsity?

Is brain activation (connectivity,..) “sparse”? No!
But...

In neuroscience, people estimate discriminative patterns that look like:

But in a neuroimaging article, it will look more like

If you want to show the truly discriminative pattern, you need it to be sparse!
Solution: (F)ISTA

Gradient descent on the smooth terms

projection on the non-smooth constrains

\[ w^{t+1} = \text{prox}_\Omega (w^t - \nabla \ell(w^t)) \]

Lasso: the proximal operator is simply soft-thresholding

FISTA = accelerated ISTA (much faster convergence)
The smooth lasso: the proximal operator

Stronger penalty
Sparse total variation: the proximal operator

Stronger penalty

Small TV

sparsity
What do the results look like?

Can nevertheless be improved with adapted techniques

$$(\hat{w}) = \text{argmin}_w \ell(X, Yw) + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$

$$(\hat{w}) = \text{argmin}_w \ell(X, Yw) + \lambda_1 \|w\|_1 + \lambda_2 \|\nabla w\|_1$$

[Gramfort et al PRNI 2013]
Performance on recovery (simulation)

Example of recovery (simulated data): The TV-L1 prior outperforms alternatives
Caveat: resulting map depends on convergence tolerance

- TV-l1 estimator: stricter convergence → a different sparser map!

[Dohmatob et al. PRNI 2014]
Discussion

- **Bayesian alternatives** (ARD, smooth ARD) [Sabuncu et al.]
  - You lose the convexity
  - Empirical Bayes: adapts well to new data
- **Cost** of these methods
  - Convergence monitoring is hard
  - Smoothing + ANOVA selection + SVM is a good competitor...
- Other approaches: use of clustering for **structured sparsity** [Jenatton et al. SIAM 2012], even more costly!
Outline

- Machine learning techniques for MVPA in neuroimaging
- Improving the decoder: smoothness and sparsity
- Recovery and randomness
Recovery...

- **Prediction vs. Identification**
  - **Prediction**: estimate $w$ that maximizes the prediction accuracy
  - **Identification** or Recovery: estimate $\hat{w}$ such that $\text{supp}(\hat{w}) = \text{supp}(w)$

- **Compressive sensing**:
  - detection of $k$ signals out of $p$ (voxels)
  - with only $n$ observations $<< k$

- **Problem**: data are correlated

How to measure the recovery of the set of regions? How to improve recovery
Small sample recovery

[Haxby Science 2001] dataset:

Trying to discriminate faces vs houses: level of performance achieved with limited number of samples.
Randomization

\[ \hat{w}^{lasso} = \arg\min_{w \in \mathbb{R}^p} \| y - Xw \|^2 + \lambda \| w \|_1 \]

- Stability selection = randomization of the features + bootstrap on the samples
- Improved feature recovery... for few, weakly correlated features

[Meinshausen and Bühlman, 2009]
Hierarchical clustering and randomized selection

Algorithm Randomized-Ward-Logistic

1. **Loop**: randomly perturb the data

2. Ward agglomeration to form q features

3. sparse linear model on reduced features

4. accumulate non-zero features

5. threshold map of selection counts

[Gramfort et al. MLINI 2011]
Simulation study

Slice 9

Slice 1

Ground truth

F test

Randomized Ward logistic
The best approach for feature recovery depends on the problem.

- The response depends on the characteristics of the problem: **smoothness** (coupling between signal and noise) and **clustering** (redundancy of features).

128 samples

256 samples

[Varoquaux et al. ICML 2012]
Simulation study

Identification accuracy

Prediction accuracy

Improves both prediction and identification!
Examples on real data

Regression task
Jimura et al. 2011

Classification task
Haxby et al. 2001
Conclusion

✔ SVM and sparse models less powerful than univariate methods for recovery.
✔ Sparsity + clustering + randomization: excellent recovery
  ⇒ Multivariate brain mapping
✔ Simultaneous prediction and recovery

✗ High computational cost (parameter setting)
Acknowledgements

• Many thanks to my co-workers: V. Michel, G. Varoquaux, A. Gramfort, F. Pedregosa, P. Fillard, J.B. Poline, V.Fritsch, V. Siless, S.Medina, R. Bricquet

• To People who provide data: E.Eger, R. Poldrack, K. Jimura, J. Haxby
All this will land into...

- Scikit-learn-like API
- BSD, Python, OSS
  - Classification of neuroimaging data (decoding)
  - Functional connectivity analysis
Thank you for your attention

http://parietal.saclay.inria.fr