Statistical learning and optimization
for functional MRI data mining

Alexandre Gramfort
alexandre.gramfort@telecom-paristech.fr

Assistant Professor
LTCI, Télécom ParisTech, Université Paris-Saclay
Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders

Yoichi Miyawaki,1,2,6 Hajime Uchida,2,3,6 Okito Yamashita,2 Masa-aki Sato,2 Yusuke Morito,4,5 Hiroki C. Tanabe,4,5 Norihiro Sadato,4,5 and Yukiyasu Kamitani2,3,*

Subject S1

http://www.youtube.com/watch?v=h1GuLYSoDaY
Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies

Shinji Nishimoto,^1^ An T. Vu,^2^ Thomas Naselaris,^1^ Yuval Benjamini,^3^ Bin Yu,^3^ and Jack L. Gallant^1,2,4,*

---

Presented clip

http://www.youtube.com/watch?v=nsjDnYxJ0bo
Outline

• Background

• Estimating the hemodynamic response function [Pedregosa et al. Neuroimage 2015]

• Mapping the visual pathways with computational models and fMRI [Eickenberg et al. Neuroimage 2016]

• Optimal transport barycenter for group studies [Gramfort et al. IPMI 2015]
Functional MRI

Neurons

Oxy. Hb
Deoxy. Hb

Scanner

Magnetic resonance imaging

Time

\[ t \rightarrow t + k \]
**fMRI supervised learning (decoding)**

**Challenge:** Predict a behavioral variable from the fMRI data

**Objective:** Predict $y$ given $X$ or learn a function $f : X \rightarrow y$
The objective is to be able to predict Patient or Controls given an fMRI activation map.

Faces vs. Houses
...

1 vs. -1

ie. \( y = \{ -1, 1 \} \)

objective: Predict \( y = \{ -1, 1 \} \) given \( x \in \mathbb{R}^p \)
fMRI supervised learning (Encoding)

**Challenge:** Predict the BOLD response from the stimuli descriptors

**Objective:** Predict $y$ given $X$ or learn a function $f : X \rightarrow y$

[Thirion et al. 06, Kay et al. 08, Naselaris et al. 11, Nishimoto et al. 2011, Schoenmakers et al. 13 ...]
Learning the hemodynamic response function (HRF) for encoding and decoding models

thanks to Fabian Pedregosa  Michael Eickenberg

Data-driven HRF estimation for encoding and decoding models, Fabian Pedregosa, Michael Eickenberg, Philippe Ciuciu, Bertrand Thirion and Alexandre Gramfort, Neuroimage 2015

PDF: https://hal.inria.fr/hal-00952554/en
Code: https://pypi.python.org/pypi/hrf_estimation
fMRI paradigm and HRF

Condition 1

Condition 2

HRF: Hemodynamic response function

Time: 25.00

Amplitude

Time (seconds)

Time since stimulus (seconds)
fMRI paradigm and HRF

condition 1

condition 2

condition 1

condition 2

aggregate
General Linear Model (GLM)

\[ y = \text{Observed BOLD} \]

\[ X = \text{Design Matrix} \]

\[ \beta = \text{Activation coefficients} \]

\[ \varepsilon = \text{Noise} \]

\[ \begin{bmatrix}
\end{bmatrix} \begin{bmatrix}
X_1 \\
X_2 \\
X_3 \\
\vdots \\
X_k
\end{bmatrix} +
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\vdots \\
\beta_k
\end{bmatrix} +
\begin{bmatrix}
\end{bmatrix} \]
Basis constrained HRF

Hemodynamic response function (HRF) is known to vary substantially across subjects, brain regions and age.


Two basis-constrained models of the HRF: FIR and 3HRF
Rank1-GLM
### Rank I - GLM

**From 1 HRF per condition**

\[
\begin{pmatrix}
  h_1 & h_4 \\
  h_2 & h_5 \\
  h_3 & h_6 \\
  \vdots & \vdots \\
  h_{3k-2} & h_{3k-1} \\
  h_{3k} & h_{3k}
\end{pmatrix}
\]

**From 1 HRF shared between all conditions**

\[
\begin{pmatrix}
  h_1 & h_1 & h_1 \\
  h_2 & h_2 & h_2 \\
  h_3 & h_3 & h_3
\end{pmatrix}
\]
Assuming 1 HRF shared between all conditions and a different amplitude/scale per condition this leads to:

\[
\begin{pmatrix}
\beta_1 h_1 & \beta_2 h_1 \\
\beta_1 h_2 & \beta_2 h_2 \\
\vdots & \vdots \\
\beta_1 h_3 & \beta_2 h_3 \\
\end{pmatrix}
\begin{pmatrix}
\beta_k h_1 \\
\beta_k h_2 \\
\beta_k h_3 \\
\end{pmatrix}
= h\beta^T
\]
Rank1-GLM

\[
\begin{pmatrix}
\beta_1 h_1 & \beta_2 h_1 \\
\beta_1 h_2 & \beta_2 h_2 \\
\beta_1 h_3 & \beta_2 h_3 \\
\end{pmatrix}
\begin{pmatrix}
\vdots \\
\end{pmatrix}
= h\beta^T
\]

\[
\begin{align*}
\arg\min_{\mathbf{h}, \mathbf{\beta}} & \quad \| \mathbf{y} - \mathbf{X} \text{vec}(\mathbf{h}\beta^T) \|^2 \\
\text{subject to} & \quad \| \mathbf{h} \|_\infty = 1 \text{ and } \langle \mathbf{h}, \mathbf{h}_{\text{ref}} \rangle > 0 \\
\implies & \quad \text{solved locally using quasi-Newton methods}
\end{align*}
\]

**Challenge:** This optimization problem is not big yet it needs to be done tens of thousands of time (typically 30,000 to 50,000 times for each voxel)

**Remark:** Worked better than alternated optimization or 1st order methods
Results

Cross-validation score in two different datasets


Encoding (mean correlation) score

- R1-GLM (FIR basis): 0.219**
- Standard GLM: 0.067

Average Decoding score

- R1-GLM (3HRF basis): 0.276**
- Standard GLM: 0.247

p-value = *< 0.05, **< 10^{-3}
Results

**Measure:** voxel-wise encoding score. Correlation with the BOLD at each voxel on left-out data.

R1-GLM (FIR basis) improves voxel-wise encoding score on more than 98% of the voxels.
Results

GLM

R1-GLM, 3HRF basis

R1-GLMS, FIR basis
Results

Time to peak on top voxels
Convolutional Networks Map the Architecture of the Human Visual System

work of Michael Eickenberg

joint work with Bertrand Thirion and Gaël Varoquaux

“Seeing it all: Convolutional network layers map the function of the human visual system”
Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, Bertrand Thirion (submitted)
Convolutional Nets for Computer Vision

[Krizhevski et al, 2012]
V1 functionality comprises edge detection
- Convolutional nets learn edge detectors, color boundary detectors and blob detectors

[Hubel & Wiesel, 1959]

[Sermanet 2013]
Can we use computer vision models and a large fMRI data to better understand human vision?
Approach

Nonlinear Feature Extraction Via Convolutional Net Layers

Voxel-Wise Prediction Using Linear Model (Ridge Regression)

Forward Model Setup:
- Encoding model [Naselaris et al., 2011]
- Make sure complexity resides in feature extraction
Convolutional Net Forward Models

Convolutional Net

Linear predictive models

Score maps ($R^2$)
Best Predicting Layers per Voxel
Score Fingerprints per Region of Interest
Score Fingerprints per Region of Interest

Subject 1

Subject 2

Percent of maximum score per region

First convolutional layers

First convolutional layers

V1
V2
V3
V3A
V3B
V4
LatOcc

V1
V2
V3
V3A
V3B
V4
LatOcc
Score Fingerprints per Region of Interest

Subject 1

Percent of maximum score per region

First convolutional layers

Subject 2

Percent of maximum score per region

First convolutional layers

V1
V2
V3
V3A
V3B
V4
LatOcc

Percent of maximum score per region

First convolutional layers

V1
V2
V3
V3A
V3B
V4
LatOcc
Score Fingerprints per Region of Interest

Subject 1

Percent of maximum score per region

First convolutional layers

V1, V2, V3, V3A, V3B, V4, LatOcc

Subject 2

Percent of maximum score per region

First convolutional layers

V1, V2, V3, V3A, V3B, V4, LatOcc
Fingerprints summary statistic

A fingerprint summaries for Kay2008
Fingerprints summary statistic

A Fingerprint summaries for Kay2008

B Fingerprint summaries for Huth2012

Photos

Videos
If our model is strong enough, we can use it to reproduce known experiments.
Generate BOLD response, do GLM analysis.

New stimuli → Convolutional net forward model → Activation Maps
High-level Validation: Faces / Places

A stimuli from Kay2008

B stimuli from Haxby2001

Convolutional Net Forward Model

Activation Maps

GLM Contrast Maps
Faces vs Places: Ground Truth

Stimuli from [Kay 2008]
Close-up faces and scenes

Contrast of stimuli from [Kay 2008]
Close-up faces and scenes
Faces vs Places

Simulation on [Kay 2008] Left out stimuli

BOLD ground truth
Fast Optimal Transport Averaging of Neuroimaging Data

Joint work with: Gabriel Peyré Marco Cuturi

[Fast Optimal Transport Averaging of Neuroimaging Data
Alexandre Gramfort, Gabriel Peyré, Marco Cuturi, Proc. IPMI 2015]
The overall goal

What is an “average activation”?

Functional neuroimaging experiment
20 subjects
with Magnetoencephalography (MEG)

From sensors to sources at every ms for each subject
Motivation

Imagine a 2D flat brain with 4 activations…

4 points in $\mathbb{R}^2$

$x_1, x_2, x_3, x_4$
Motivation

Their mean is \( \frac{x_1 + x_2 + x_3 + x_4}{4} \).
Motivation

Consider for each point the function $\|\cdot - x_i\|_2^2$
The mean is the $\text{argmin} \frac{1}{4} \sum_{i=1}^{4} \| \cdot - x_i \|_2^2$. 
Now if the domain is not flat: you have a ground metric.
Assume that each datum is now an **empirical measure**. What could be the mean of these 4 measures?
From points to probability measures

\[ \text{naive mean of } \textit{all} \text{ observations.} \]

\[ \text{Mean of 4 measures } \text{=} \text{ a point?} \]

Should preserve the uncertainty & take into account the metric
Given a discrepancy function $\Delta$ between probabilities, compute their mean: $\arg\min \sum_{i} \Delta(\cdot, \nu_i)$

**Remark:** If discrepancy is a squared Riemannian distance it’s a Fréchet mean.
Optimal Transport distances rely on 2 key concepts:

- **A metric** $D : \Omega \times \Omega \rightarrow \mathbb{R}_+$;

- **$\Pi(\mu, \nu)$**: joint probabilities with marginals $\mu, \nu$. 

$$(\Omega, D)$$

$\mu$ and $\nu$ represent probability measures, and $D(x, y)$ is the distance between $x$ and $y$. The diagram illustrates the concept of optimal transport, where the optimal transport plan is shown as the shortest path between the two probability distributions $\mu$ and $\nu$.
Example of joint probabilities

...on the real line

$$\Pi(\mu, \nu) = \text{probability measures on } \Omega^2$$

with marginals $\mu$ and $\nu$. 
$\Pi(\mu, \nu) = \text{probability measures on } \Omega^2$

with marginals $\mu$ and $\nu$. 

...on the real line
Optimal Transport

$p$-Wasserstein distance for $p \geq 1$ is:

$$W_p(\mu, \nu) = \left( \inf_{P \in \Pi(\mu, \nu)} \int \int_{\Omega \times \Omega} D(x, y)^p dP(x, y) \right)^{1/p}.$$ 

[Monge-Kantorovich, Kantorovich-Rubinstein, Wasserstein, Earth Mover’s Distance, Mallows ...]
Optimal Transport in dimension $d$

$W^p_p(\mu, \nu)$ can be cast as a linear program

1. $M_{XY}^p \overset{\text{def}}{=} [D(x_i, y_j)^p]_{i,j} \in \mathbb{R}_+^{d \times d}$ (metric information)

2. Transportation Polytope (joint probabilities)

$U(a, b) \overset{\text{def}}{=} \{ T \in \mathbb{R}_+^{d \times d} \mid T1_d = a, T^T1_d = b \}$.

Example:

$$T = \begin{bmatrix}
.1 & 0 & .1 \\
.1 & 0 & .1 \\
.2 & .1 & .3
\end{bmatrix} \in U\left(\left[\begin{array}{c}
.2 \\
.2 \\
.6
\end{array}\right], \left[\begin{array}{c}
.4 \\
.1 \\
.5
\end{array}\right]\right)$$
Optimal Transport in dimension $d$

$W^p_p(\mu, \nu)$ can be cast as a linear program

Optimal transport problem reads:

$$W^p_p(a, b) = \text{OT}(a, b, M^p) \overset{\text{def}}{=} \min_{T \in U(a,b)} \langle T, M^p \rangle$$

$$\langle T, M^p \rangle = \sum_{i=1}^{d} \sum_{j=1}^{d} T_{i,j} M_{i,j}^p$$

$T$ is the transport plan

Problem: No solution if

$$|a|_1 = \sum_{i=1}^{d} |a_i| \neq |b|_1$$

Need to add and remove mass
non-negative and non-normalized data

Add a virtual point \( \omega \) whose distance to element \( i \) in \( \Omega \)

\[
D(i, \omega) = D(\omega, i) = \Delta_i
\]

Scale each observation \( b^j, 1 \leq j \leq n \) so that \( |b^j|_1 \leq 1 \)

Map each \( a \) to \([a, \|a\|_1 - 1]\) (Kind of feature map)

Use as metric \( \hat{M} = \begin{bmatrix} M & \Delta \\ \Delta^T & 0 \end{bmatrix} \in \mathbb{R}^{d+1 \times d+1} \)

\[
\arg\min_{u \in S_d} \frac{1}{N} \sum_{j=1}^{N} \text{OT}(\begin{bmatrix} 1 - |u|_1 \\ \beta_j \end{bmatrix}, b^j, \hat{M}^p).
\]

\( S_d = \{ u \in \mathbb{R}^d_+, \|u\|_1 \leq 1 \} \) … but a huge linear program
Idea: Regularize cost with entropy

\[ \text{OT}_\lambda(a, b, M^p) \overset{\text{def}}{=} \min_{T \in U(a,b)} \langle T, M^p \rangle - \frac{1}{\lambda} H(T). \]

Strongly convex with unique minimum

Problem reads:

\[ \arg\min_{a \in S_d} \frac{1}{N} \sum_{j} \text{OT}_\lambda(a, b^j, \hat{M}^p) \]

In practice: solved with an exponentiated gradient with projection in the dual (matrix-matrix computations and element wise multiplications which are GPGPU friendly)
$n = 100 \quad d = 10242$
Results fMRI

- 20 subjects
- Left hand button press
- Averaging of standardized effect size

[Pinel et al. 2007]

Sharp activation foci & less amplitude reduction
Results MEG

- 16 subjects
- Visual presentation of faces and scrambled faces
- Averaging of dSPM source estimates

[Henson et al. 2011]
Results MEG

• Contrast between faces and scrambled faces

With Tesla K40 GPU card (< a minute of computation)
• The world of neuroimaging is full of challenging maths and computer science problems ...

• … look at the data to find the relevant ones

• … but don’t be scared if they are not well posed

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem. ~ John Tukey"
Some refs:

Fabian Pedregosa, Michael Eickenberg, Philippe Ciuciu, Bertrand Thirion and Alexandre Gramfort, *Data-driven HRF estimation for encoding and decoding models*, Neuroimage 2015


Contact

http://alexandre.gramfort.net

GitHub : @agramfort  
Twitter : @agramfort