Modelling inter-subject functional variability

Elvis Dohmatob
(PhD supervised by B. Thirion and G. Varoquaux)

Parietal Team, INRIA

September 26, 2017

Context

- A major goal of human neuroscience is to understand
  - the structure,
  - function, and
  - inter-subject variability of the human brain
- We will focus on **inter-subject functional variability**

Context

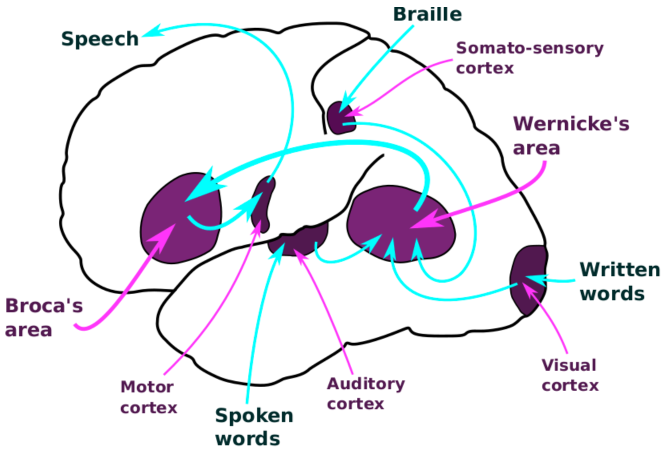
- A major goal of human neuroscience is to understand
  - the structure,
  - function, and
  - inter-subject variability of the human brain
- We will focus on **inter-subject functional variability**

Table of contents

Introduction

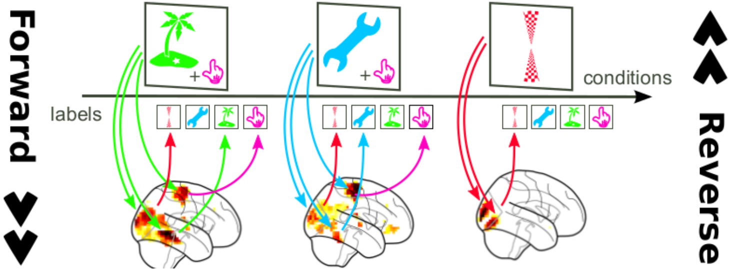
Brain function regions and networks

Part of the language network



(Picture is courtesy of Gael Varoquaux)

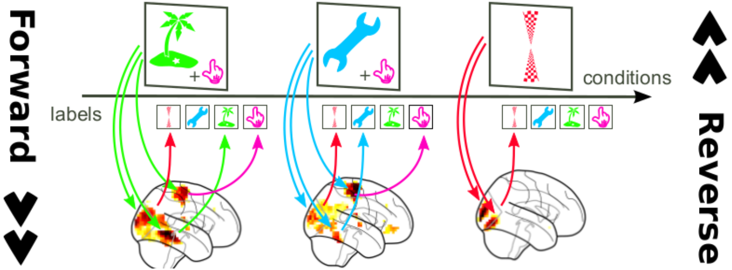
Mapping cognitive circuits in the brain



(Picture is courtesy of Yannick Schwarz)

- **Forward inference** [Friston '95] detects voxels responding to a given experimental condition
- **Reverse inference / brain-decoding** [Dehaene 98; Cox 03] predicts the experimental condition from brain signals

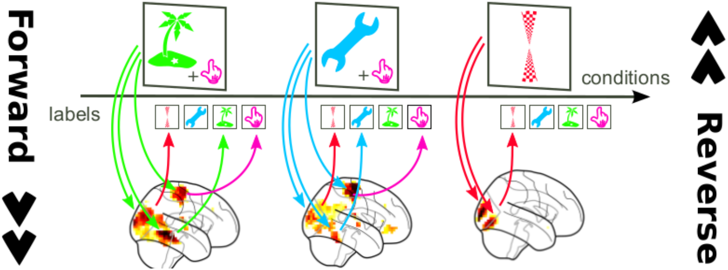
Mapping cognitive circuits in the brain



(Picture is courtesy of Yannick Schwarz)

- **Forward inference** [Friston '95] detects voxels responding to a given experimental condition
- **Reverse inference / brain-decoding** [Dehaene 98; Cox 03] predicts the experimental condition from brain signals
- We will focus on **reverse-inference / brain-decoding**

Mapping cognitive circuits in the brain

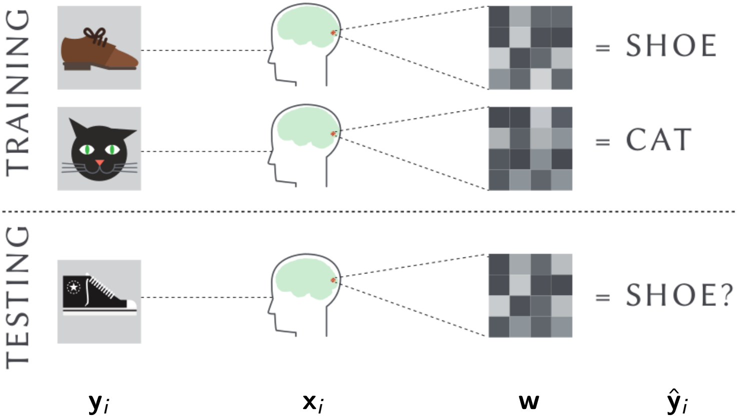


(Picture is courtesy of Yannick Schwarz)

- **Forward inference** [Friston '95'] detects voxels responding to a given experimental condition
- **Reverse inference / brain-decoding** [Dehaene 98; Cox 03] predicts the experimental condition from brain signals
- We will focus on **reverse-inference / brain-decoding**

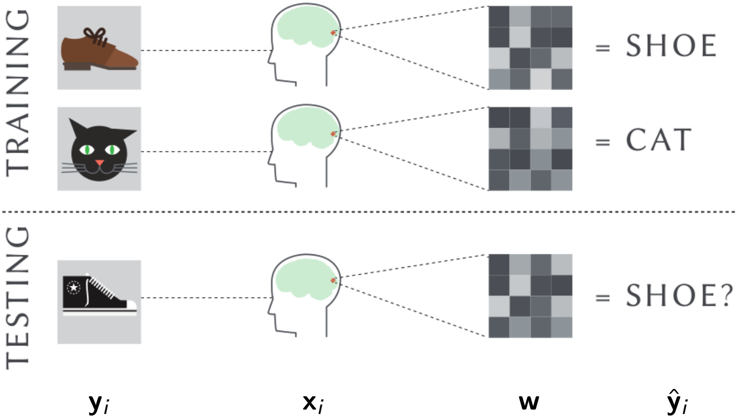
A zoom on brain-decoding

(Picture is courtesy of F. Pedregosa)



A zoom on brain-decoding

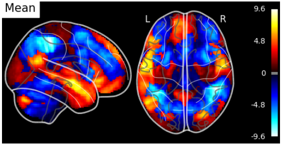
(Picture is courtesy of F. Pedregosa)



- This is **supervised machine-learning**
- We don't just want good predictions, we want **regions**

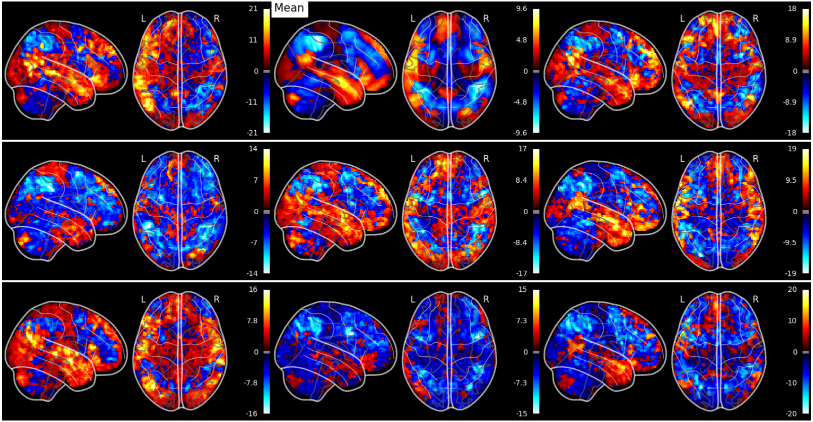
Variability in both location and magnitude of activations

- Story vs Math language contrast of HCP dataset [van Essen '12]



Variability in both location and magnitude of activations

- Story vs Math language contrast of HCP dataset [van Essen '12]



Variability in both location and magnitude of activations

- Inter-subject functional variability $\neq$ noise!
  - Usually (incorrectly) discarded in standard analysis
  - Is predictive of behavioral differences between individuals
- Cannot be corrected via **spatial normalization**, etc.
  - E.g spatial normalization cannot correct for differences in activation magnitude
- Driven by genetic and behavioral inter-individual differences
- Functional diseases can be seen as extremes of this variation

Variability in both location and magnitude of activations

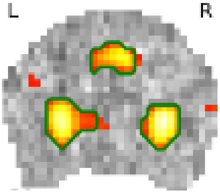
- Inter-subject functional variability $\neq$ noise!
  - Usually (incorrectly) discarded in standard analysis
  - Is predictive of behavioral differences between individuals
- Cannot be corrected via **spatial normalization**, etc.
  - E.g spatial normalization cannot correct for differences in activation magnitude
- Driven by genetic and behavioral inter-individual differences
- Functional diseases can be seen as extremes of this variation

Mapping the brain with structured multi-variate models

What we mean by “structured”

Definition:

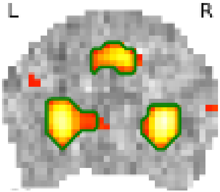
- Localized activation patterns – **sparsity**
- Clusters of active voxels – **smoothness**



What we mean by “structured”

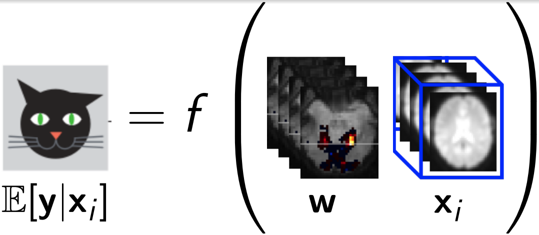
Definition:

- Localized activation patterns – **sparsity**
- Clusters of active voxels – **smoothness**



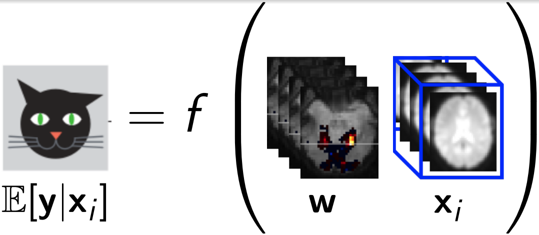
- Such a model is much more **interpretable** (i.e small number of **regions**) than classical methods like SVM, Ridge regression, Lasso
- Performs **model-estimation** and **feature-selection** jointly
- Fights the **curse-of-dimensionality**, via dimensionality reduction.

Generalized linear models with structured penalties



- **Samples** $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$
  - # samples $n \sim 10^3$
  - # **features** $p \sim 10^6$ voxels

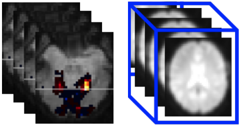
Generalized linear models with structured penalties



- **Samples** $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$
  - $\#$ samples $n \sim 10^3$
  - $\#$ **features** $p \sim 10^6$ voxels
- Model **weights** $\mathbf{w} \in \mathbb{R}^p$

Generalized linear models with structured penalties

$\mathbb{E}[\mathbf{y}|\mathbf{x}_i] = f\left(\right.$

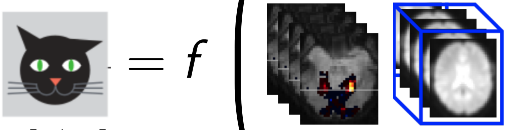


$\mathbf{w}$ $\mathbf{x}_i$ $\left.\right)$

- **Samples** $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$
  - # samples $n \sim 10^3$
  - # **features** $p \sim 10^6$ voxels
- Model **weights** $\mathbf{w} \in \mathbb{R}^p$
- f = "logit" in **classification**
- f = "id" in **regression**
- Optimization problem:

$$\min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(\langle \mathbf{w}, \mathbf{x}_i \rangle))}_{\text{data / loss term}} + \underbrace{\alpha \mathcal{P}(\mathbf{w})}_{\text{penalty}}$$

Generalized linear models with structured penalties



$\mathbb{E}[\mathbf{y}|\mathbf{x}_i] = f\left(\langle \mathbf{w}, \mathbf{x}_i \rangle\right)$

- **Samples** $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$
  - # samples $n \sim 10^3$
  - # **features** $p \sim 10^6$ voxels
- Model **weights** $\mathbf{w} \in \mathbb{R}^p$
- f = "logit" in **classification**
- f = "id" in **regression**

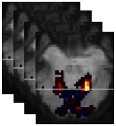
- Optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\langle \mathbf{w}, \mathbf{x}_i \rangle))}_{\text{data / loss term}} + \underbrace{\alpha\mathcal{P}(\mathbf{w})}_{\text{penalty}}$$

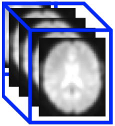
- $\ell(y_i, f(\mathbf{x}_i^T \mathbf{w})) = \begin{cases} \frac{1}{2}(y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2, & \text{in regression,} \\ \log(1 + \exp(-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)), & \text{in classif. (OvR)} \end{cases}$

Generalized linear models with structured penalties

$\mathbb{E}[\mathbf{y}|\mathbf{x}_i] = f\Big($



$\mathbf{w}$



$\mathbf{x}_i\Big)$

- **Samples** $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$
  - \# samples $n \sim 10^3$
  - \# **features** $p \sim 10^6$ voxels
- Model **weights** $\mathbf{w} \in \mathbb{R}^p$
- f = "logit" in **classification**
- f = "id" in **regression**
- Optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(\langle \mathbf{w}, \mathbf{x}_i \rangle))}_{\textbf{data / loss term}} + \underbrace{\alpha \mathcal{P}(\mathbf{w})}_{\text{penalty}}$$

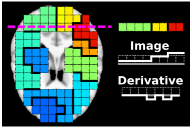
- $\ell(y_i, f(\mathbf{x}_i^T \mathbf{w})) = \begin{cases} \frac{1}{2}(y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2, & \text{in regression,} \\ \log(1 + \exp(-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle)), & \text{in classif. (OvR)} \end{cases}$

Spatial penalties impose structure in the model

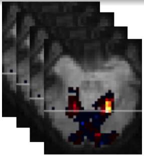
$$\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{y}_i, f(\langle\mathbf{w},\mathbf{x}_i\rangle))}_{\text{data / loss term}} + \underbrace{\alpha\mathcal{P}(\mathbf{w})}_{\textbf{penalty}}$$

Spatial penalties impose structure in the model

$$\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{y}_i, f(\langle \mathbf{w}, \mathbf{x}_i\rangle))}_{\text{data / loss term}} + \underbrace{\alpha\mathcal{P}(\mathbf{w})}_{\text{penalty}}$$



structured penalty

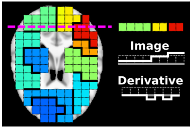


w

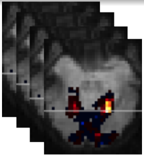
$$\mathcal{P}(\mathbf{w}) = \begin{cases} \sum_{j\in[\![p]\!]} \rho|\mathbf{w}_j| + \frac{1}{2}(1-\rho)\|(\nabla\mathbf{w})_j\|_2^2, & \text{GraphNet,} \\ \sum_{j\in[\![p]\!]} \rho|\mathbf{w}_j| + (1-\rho)\|(\nabla\mathbf{w})_j\|_2, & \text{isotropic TV-L1,} \\ \sum_{j\in[\![p]\!]} (\rho^2|\mathbf{w}_j|^2 + (1-\rho)^2\|(\nabla\mathbf{w})_j\|_2^2)^{1/2}, & \text{Sparse Variation,} \\ \vdots \end{cases}$$

Spatial penalties impose structure in the model

$$\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{y}_i, f(\langle \mathbf{w}, \mathbf{x}_i\rangle))}_{\text{data / loss term}} + \underbrace{\alpha\mathcal{P}(\mathbf{w})}_{\text{penalty}}$$



structured penalty



w

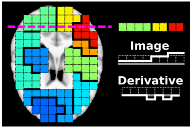
$$\mathcal{P}(\mathbf{w}) = \begin{cases} \sum_{j\in[\![p]\!]} \rho|\mathbf{w}_j| + \frac{1}{2}(1-\rho)\|(\nabla\mathbf{w})_j\|_2^2, & \textbf{GraphNet}, \\ \sum_{j\in[\![p]\!]} \rho|\mathbf{w}_j| + (1-\rho)\|(\nabla\mathbf{w})_j\|_2, & \textbf{isotropic TV-L1}, \\ \sum_{j\in[\![p]\!]} (\rho^2|\mathbf{w}_j|^2 + (1-\rho)^2\|(\nabla\mathbf{w})_j\|_2^2)^{1/2}, & \textbf{Sparse Variation}, \\ \vdots \end{cases}$$

- Bayesian interpretation

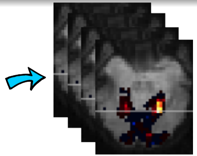
$$\underbrace{P(\mathbf{w}|\mathbf{x}_i, y_i)}_{\text{posterior}} \propto \underbrace{P(y_i|\mathbf{x}_i, \mathbf{w})}_{\text{likelihood}}\underbrace{P(\mathbf{w})}_{\text{prior}} \propto \exp(-\ell(y_i, f(\langle\mathbf{w}, \mathbf{x}_i\rangle)))\exp(-\alpha\mathcal{P}(\mathbf{w}))$$

Spatial penalties impose structure in the model

$$\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{y}_i, f(\langle \mathbf{w}, \mathbf{x}_i\rangle))}_{\text{data / loss term}} + \underbrace{\alpha\mathcal{P}(\mathbf{w})}_{\text{penalty}}$$



structured penalty



$$\mathcal{P}(\mathbf{w}) = \begin{cases} \sum_{j\in[\![p]\!]} \rho|\mathbf{w}_j| + \frac{1}{2}(1-\rho)\|(\nabla\mathbf{w})_j\|_2^2, & \textbf{GraphNet}, \\ \sum_{j\in[\![p]\!]} \rho|\mathbf{w}_j| + (1-\rho)\|(\nabla\mathbf{w})_j\|_2, & \textbf{isotropic TV-L1}, \\ \sum_{j\in[\![p]\!]} (\rho^2|\mathbf{w}_j|^2 + (1-\rho)^2\|(\nabla\mathbf{w})_j\|_2^2)^{1/2}, & \textbf{Sparse Variation}, \\ \vdots \end{cases}$$

- Bayesian interpretation

$$\underbrace{P(\mathbf{w}|\mathbf{x}_i, y_i)}_{\text{posterior}} \propto \underbrace{P(y_i|\mathbf{x}_i, \mathbf{w})}_{\text{likelihood}}\underbrace{P(\mathbf{w})}_{\text{prior}} \propto \exp(-\ell(y_i, f(\langle \mathbf{w}, \mathbf{x}_i\rangle)))\exp(-\alpha\mathcal{P}(\mathbf{w}))$$

References for the penalties

- Total-Variation (TV) [Michel '11]
- TV-L1 [Baldassare '12, Gramfort '13]
- GraphNet / S-Lasso [Hebiri '11, Grosenick '13]
- Sparse-Variation [Eickenberg '15]

Some notes

- TV is a very tight convex relaxation of Markovian prior
- GraphNet ("Dirichlet energy") is weaker, but easier to optimize (smooth convex optimization problem)

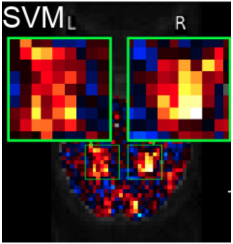
References for the penalties

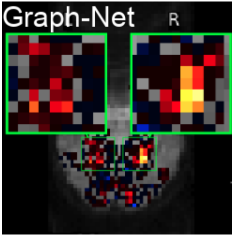
- Total-Variation (TV) [Michel '11]
- TV-L1 [Baldassare '12, Gramfort '13]
- GraphNet / S-Lasso [Hebiri '11, Grosenick '13]
- Sparse-Variation [Eickenberg '15]

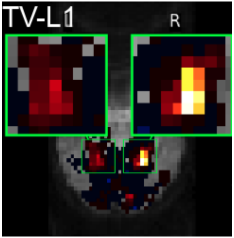
Some notes

- TV is a very tight convex relaxation of Markovian prior
- GraphNet ("Dirichlet energy") is weaker, but easier to optimize (smooth convex optimization problem)

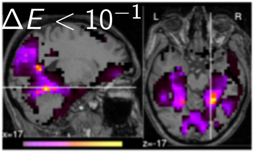
Spatial penalties $\Longrightarrow$ more interpretable brain maps



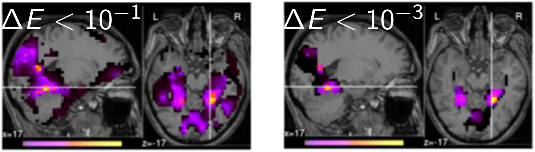




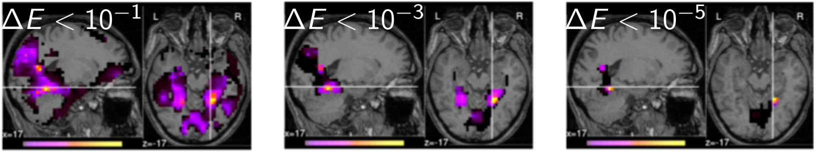
Penalties $\implies$ interpretable maps only if well-optimized



Penalties $\implies$ interpretable maps only if well-optimized

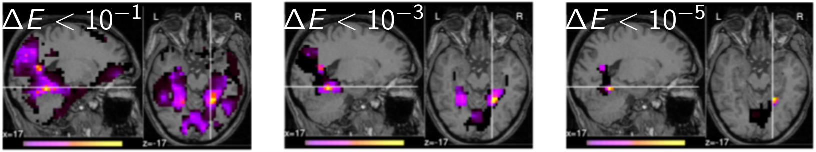


Penalties $\implies$ interpretable maps only if well-optimized



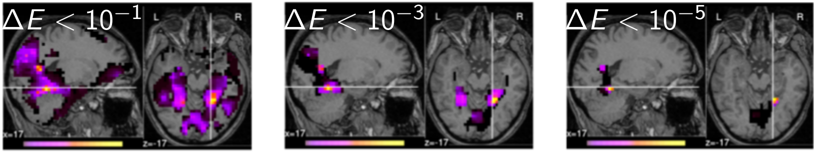
- Structured penalties $\implies$ **more interpretable models**

Penalties $\implies$ interpretable maps only if well-optimized



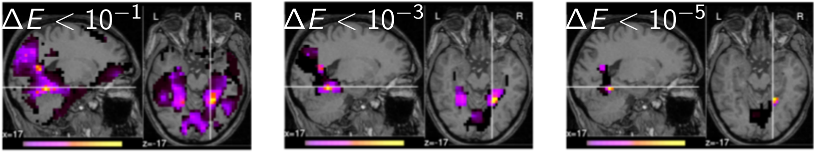
- Structured penalties $\implies$ **more interpretable models**
- Corresponding optim. problem is much **harder** (than SVM, etc.)
  - **high-dimensional non-smooth ill-conditioned** problem

Penalties $\implies$ interpretable maps only if well-optimized



- Structured penalties $\implies$ **more interpretable models**
- Corresponding optim. problem is much **harder** (than SVM, etc.)
  - **high-dimensional non-smooth ill-conditioned** problem
- Lack of fast solver can lead to **wrong conclusions about model**
- We need **fast solvers!**

Penalties $\implies$ interpretable maps only if well-optimized



- Structured penalties $\implies$ **more interpretable models**
- Corresponding optim. problem is much **harder** (than SVM, etc.)
  - **high-dimensional non-smooth ill-conditioned** problem
- Lack of fast solver can lead to **wrong conclusions about model**
- We need **fast solvers!**

Our contributions

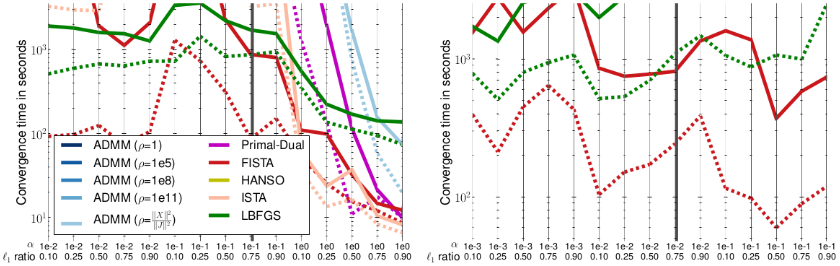
Faster, better, stronger!

We propose a combination of **algorithmic** and **implementation** improvements that make these models usable out-of-the-box

Looking for the ideal solver

[Dohmatob '14, '15 (PRNI); Varoquaux '15 (Gretsi)]

- Solver speed sensitive to hyper-parameter
- Retained strategy is nested FISTA [Beck '09] algorithm



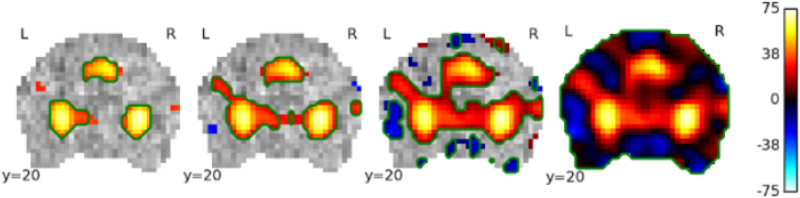
Benchmarks on "mixed-gambles" task [Jimura '12]

More speed via univariate feature-screening

[Dohmatob '15 (PRNI)]

- t_k := kth percentile of the vector $|\mathbf{X}^T\mathbf{y}| := (|\mathbf{x}_1^T\mathbf{y}|, \ldots, |\mathbf{x}_p^T\mathbf{y}|)$.
- Discard jth voxel if $|\mathbf{x}_j^T\mathbf{y}| < t_k$

$k = 10\%$ $k = 20\%$ $k = 50\%$ $k = 100\%$



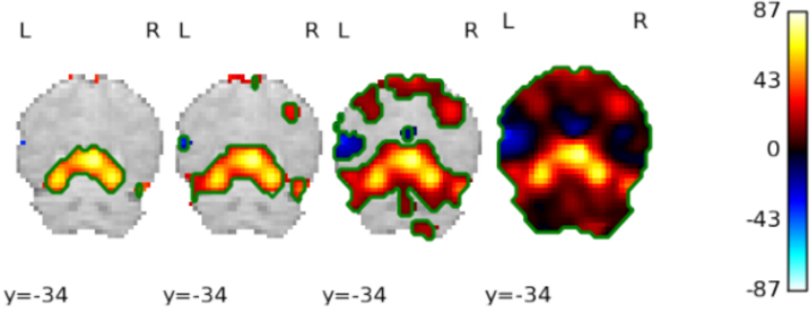
Mixed gambling

More speed via univariate feature-screening

[Dohmatob '15 (PRNI)]

- t_k := kth percentile of the vector $|\mathbf{X}^T\mathbf{y}| := (|\mathbf{x}_1^T\mathbf{y}|, \ldots, |\mathbf{x}_p^T\mathbf{y}|)$.
- Discard jth voxel if $|\mathbf{x}_j^T\mathbf{y}| < t_k$

$k = 10\%$ $k = 20\%$ $k = 50\%$ $k = 100\%$



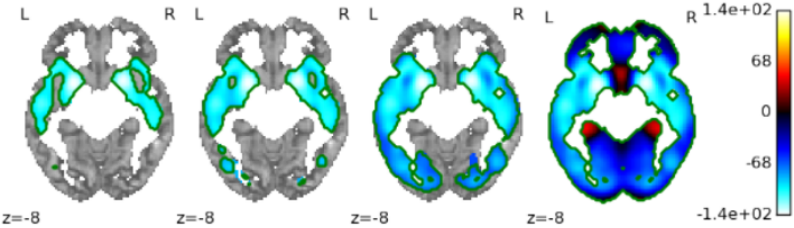
Visual recognition

More speed via univariate feature-screening

[Dohmatob '15 (PRNI)]

- $t_k :=$ kth percentile of the vector $|\mathbf{X}^T\mathbf{y}| := (|\mathbf{x}_1^T\mathbf{y}|, \ldots, |\mathbf{x}_p^T\mathbf{y}|)$.
- Discard jth voxel if $|\mathbf{x}_j^T\mathbf{y}| < t_k$

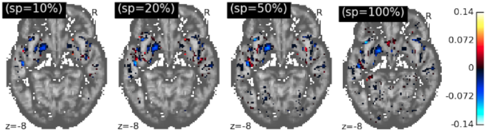
$k = 10\%$ $k = 20\%$ $k = 50\%$ $k = 100\%$



Age prediction from gray-matter maps

More speed via univariate feature-screening: results

- Age prediction



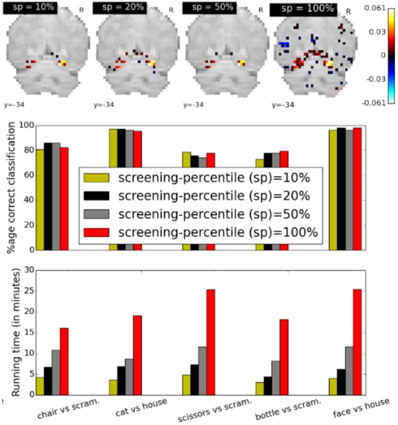
p	100%	50%	20%	10%
MSE	8.37	9.10	9.23	9.19

[Dohmatob '15 (PRNI)]

- Solve on **subset of features**
- Reduced **training time**

More speed via univariate feature-screening: results

- Visual object recognition

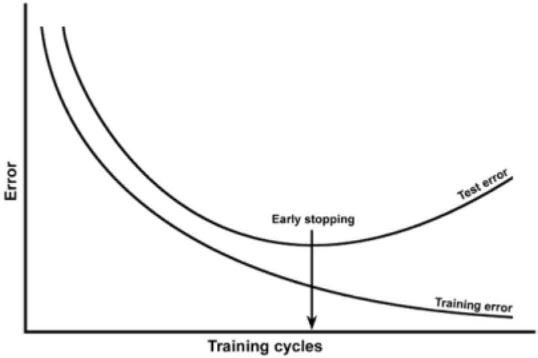


[Dohmatob '15 (PRNI)]

- Solve on **subset of features**
- Reduced **training time**

Early-stopping

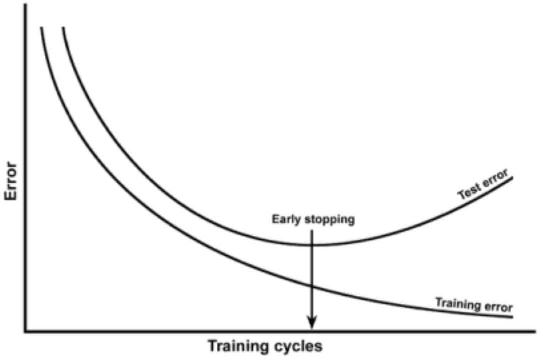
- Stop optimization if accuracy on validation data stops improving [Dohmatob '15 (PRNI)]



- old idea (e.g [Bottou '07])
- saves training time
- implicit regularization
- helps against overfitting
- it's a compromise
  - it doesn't destroy accuracy
  - but may lead to sub-optimal brain maps

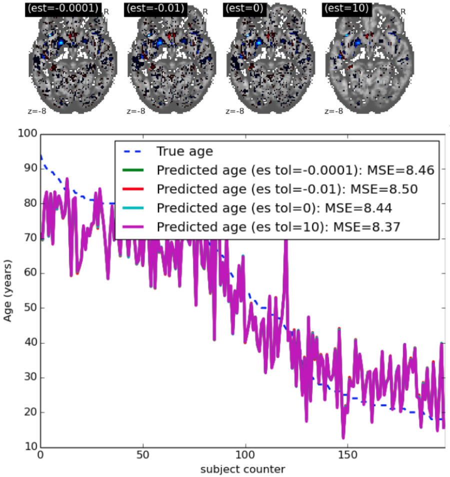
Early-stopping

- Stop optimization if accuracy on validation data stops improving [Dohmatob '15 (PRNI)]



- old idea (e.g [Bottou '07])
- saves training time
- implicit regularization
- helps against overfitting
- it's a compromise
  - it doesn't destroy accuracy
  - but may lead to sub-optimal brain maps

Early-stopping: results

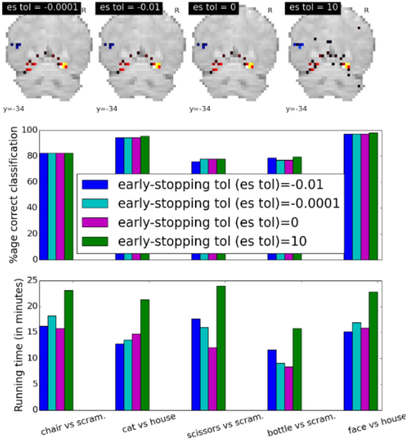


[Dohmatob '15 (PRNI)]

- Solve on subset of features
- Yields up to **x10 speedup!**
- No significant loss in accuracy

Early-stopping: results

- Visual object recognition



[Dohmatob '15 (PRNI)]

- Solve on subset of features
- Yields up to **x10 speedup!**
- No significant loss in accuracy

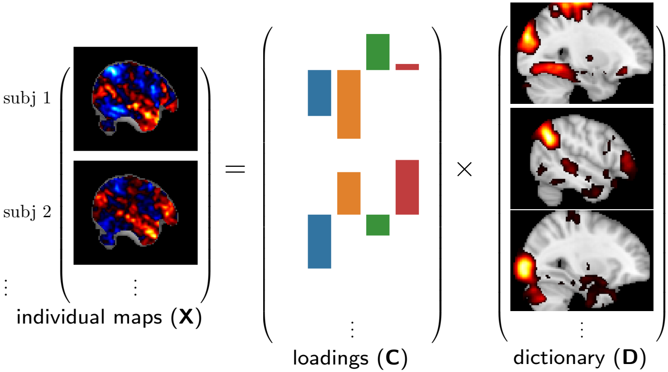
Section wrap-up

- Building on prior work, we have developed enhanced structured penalties for multi-variate brain-decoding
- Such penalties lead to more interpretable brain maps (a small number of smooth spatially localized regions)
- Focus on practical usability (fast model training)
- Our contributions are available as part of **Nilearn** toolkit.

Modelling inter-subject variability via dictionary-learning

Learn latent model for inter-subject variability

- **Goal:** Learn a latent model of inter-subject functional variability



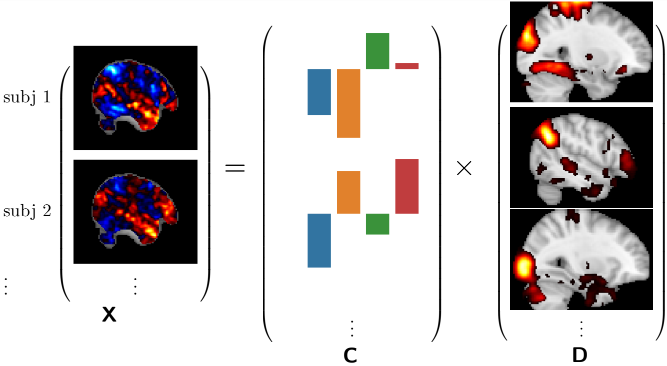
- Each **cognitive map** $\mathbf{x}_i$ with p voxels gets **encoded** over a **dictionary** $\mathbf{D}$ as k **loading coefficients** $\mathbf{c}_i$, with $k \ll p$

The challenge

[Dohmatob '16 (NIPS)]

- Sparsity: spatially localized atoms
- Smooth regions: each atom = interpretable blobs
- Scalable / online: model should trainable online

Introducing the proposed model [Dohmatob '16 (NIPS)]

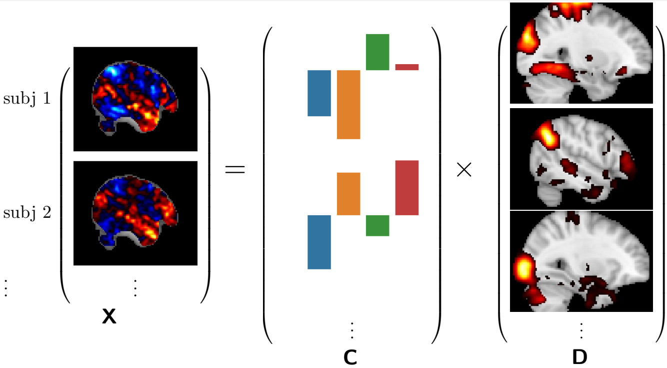


$$\min_{\mathbf{D}\in\mathbb{R}^{p\times k}} \left(\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\min_{\mathbf{c}_t\in\mathbb{R}^k}\frac{1}{2}\|\mathbf{x}_t-\mathbf{D}\mathbf{c}_t\|_2^2+\frac{1}{2}\alpha\|\mathbf{c}_t\|_2^2\right)$$

subject to $\mathbf{d}^1,\ldots,\mathbf{d}^k \in \mathcal{K}$ [Mairal '09']

- $\mathcal{K} \subseteq \mathbb{R}^p$ is an ℓ_1 ball

Introducing the proposed model [Dohmatob '16 (NIPS)]

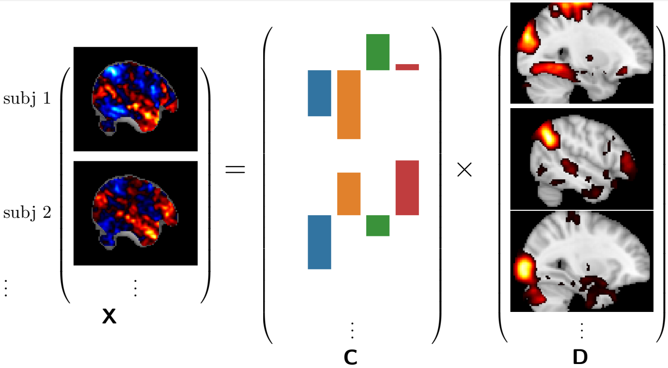


$$\min_{\mathbf{D}\in\mathbb{R}^{p\times k}} \left(\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\min_{\mathbf{c}_t\in\mathbb{R}^k}\frac{1}{2}\|\mathbf{x}_t-\mathbf{D}\mathbf{c}_t\|_2^2+\frac{1}{2}\alpha\|\mathbf{c}_t\|_2^2\right)+\gamma\sum_{j=1}^{k}\Omega_{\mathrm{Lap}}(\mathbf{d}^j)$$

subject to $\mathbf{d}^1, \ldots, \mathbf{d}^k \in \mathcal{K}$ [Mairal '09] [Dohmatob '16']

- $\mathcal{K} \subseteq \mathbb{R}^p$ is an ℓ_1 ball

Introducing the proposed model [Dohmatob '16 (NIPS)]



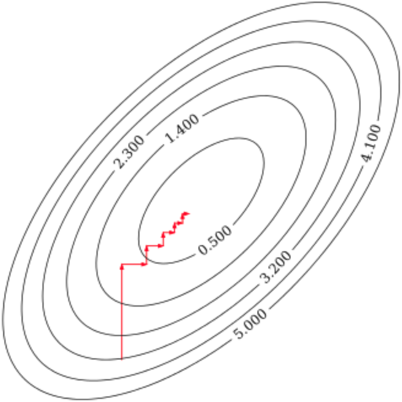
$$\min_{\mathbf{D} \in \mathbb{R}^{p \times k}} \left(\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \min_{\mathbf{c}_t \in \mathbb{R}^k} \frac{1}{2} \|\mathbf{x}_t - \mathbf{D}\mathbf{c}_t\|_2^2 + \frac{1}{2} \alpha \|\mathbf{c}_t\|_2^2 \right) + \gamma \sum_{j=1}^{k} \Omega_{\text{Lap}}(\mathbf{d}^j)$$

subject to $\mathbf{d}^1, \ldots, \mathbf{d}^k \in \mathcal{K}$ [Mairal '09] [Dohmatob '16']

- $\mathcal{K} \subseteq \mathbb{R}^p$ is an ℓ_1 ball

Reminder on coordinate-descent (CD)

- Optimize w.r.t a variable, and then w.r.t to another, and so on ...



The proposed algorithm

- **Draw a sample** 3D brain image (or mini-batch) $\mathbf{x}_t \in \mathbb{R}^p$

The proposed algorithm

- **Draw a sample** 3D brain image (or mini-batch) $\mathbf{x}_t \in \mathbb{R}^p$
- **Compute loadings** (i.e representation w.r.t current dict. $\mathbf{D}$)

$$\mathbf{c}_t \leftarrow \text{argmin}_{\mathbf{u} \in \mathbb{R}^k} \frac{1}{2}\|\mathbf{x}_t - \mathbf{D}\mathbf{u}\|_2^2 + \frac{1}{2}\alpha\|\mathbf{u}\|_2^2.$$

The proposed algorithm

- **Draw a sample** 3D brain image (or mini-batch) $\mathbf{x}_t \in \mathbb{R}^p$
- **Compute loadings** (i.e representation w.r.t current dict. $\mathbf{D}$)

$$\mathbf{c}_t \leftarrow \text{argmin}_{\mathbf{u}\in\mathbb{R}^k} \frac{1}{2}\|\mathbf{x}_t - \mathbf{D}\mathbf{u}\|_2^2 + \frac{1}{2}\alpha\|\mathbf{u}\|_2^2.$$

- **Rank-1** updates: $\mathbf{A}_t = \mathbf{A}_{t-1} + \mathbf{c}_t\mathbf{c}_t^T$, $\mathbf{B}_t := \mathbf{B}_{t-1} + \mathbf{x}_t\mathbf{c}_t^T$

The proposed algorithm

- **Draw a sample** 3D brain image (or mini-batch) $\mathbf{x}_t \in \mathbb{R}^p$
- **Compute loadings** (i.e representation w.r.t current dict. $\mathbf{D}$)

$$\mathbf{c}_t \leftarrow \text{argmin}_{\mathbf{u}\in\mathbb{R}^k} \frac{1}{2}\|\mathbf{x}_t - \mathbf{D}\mathbf{u}\|_2^2 + \frac{1}{2}\alpha\|\mathbf{u}\|_2^2.$$

- **Rank-1** updates: $\mathbf{A}_t = \mathbf{A}_{t-1} + \mathbf{c}_t\mathbf{c}_t^T$, $\mathbf{B}_t := \mathbf{B}_{t-1} + \mathbf{x}_t\mathbf{c}_t^T$
- **BCD dictionary update** of dictionary atoms
  - **Precompute** $\mathbf{R} \leftarrow \mathbf{B} - \mathbf{D}\mathbf{A}$
  - **for** $j = 1, 2, \ldots, k$

The proposed algorithm

- **Draw a sample** 3D brain image (or mini-batch) $\mathbf{x}_t \in \mathbb{R}^p$
- **Compute loadings** (i.e representation w.r.t current dict. $\mathbf{D}$)

$$\mathbf{c}_t \leftarrow \mathrm{argmin}_{\mathbf{u} \in \mathbb{R}^k} \frac{1}{2}\|\mathbf{x}_t - \mathbf{D}\mathbf{u}\|_2^2 + \frac{1}{2}\alpha\|\mathbf{u}\|_2^2.$$

- **Rank-1** updates: $\mathbf{A}_t = \mathbf{A}_{t-1} + \mathbf{c}_t\mathbf{c}_t^T$, $\mathbf{B}_t := \mathbf{B}_{t-1} + \mathbf{x}_t\mathbf{c}_t^T$
- **BCD dictionary update** of dictionary atoms
  - **Precompute** $\mathbf{R} \leftarrow \mathbf{B} - \mathbf{D}\mathbf{A}$
  - **for** $j = 1, 2, \ldots, k$
    - **Rank-1** update: $\mathbf{R} \leftarrow \mathbf{R} + \mathbf{d}^j \circ \mathbf{a}^j$
    - **FISTA loop:** $\mathbf{d}^j \leftarrow \mathrm{argmin}_{\mathbf{d} \in \mathcal{K}} F_{\gamma_t}(\mathbf{d}, a_{j,j}^{-1}\mathbf{r}^j)$
    - **Rank-1** update: $\mathbf{R} \leftarrow \mathbf{R} - \mathbf{d}^j \circ \mathbf{a}^j$

The proposed algorithm

- **Draw a sample** 3D brain image (or mini-batch) $\mathbf{x}_t \in \mathbb{R}^p$
- **Compute loadings** (i.e representation w.r.t current dict. $\mathbf{D}$)

$$\mathbf{c}_t \leftarrow \text{argmin}_{\mathbf{u} \in \mathbb{R}^k} \frac{1}{2}\|\mathbf{x}_t - \mathbf{D}\mathbf{u}\|_2^2 + \frac{1}{2}\alpha\|\mathbf{u}\|_2^2.$$

- **Rank-1** updates: $\mathbf{A}_t = \mathbf{A}_{t-1} + \mathbf{c}_t\mathbf{c}_t^T$, $\mathbf{B}_t := \mathbf{B}_{t-1} + \mathbf{x}_t\mathbf{c}_t^T$
- **BCD dictionary update** of dictionary atoms
  - **Precompute** $\mathbf{R} \leftarrow \mathbf{B} - \mathbf{D}\mathbf{A}$
  - **for** $j = 1, 2, \ldots, k$
    - **Rank-1** update: $\mathbf{R} \leftarrow \mathbf{R} + \mathbf{d}^j \circ \mathbf{a}^j$
    - **FISTA loop:** $\mathbf{d}^j \leftarrow \text{argmin}_{\mathbf{d} \in \mathcal{K}} F_{\gamma_t}(\mathbf{d}, a_{j,j}^{-1}\mathbf{r}^j)$
    - **Rank-1** update: $\mathbf{R} \leftarrow \mathbf{R} - \mathbf{d}^j \circ \mathbf{a}^j$

N.B.: $F_{\gamma_t}(\mathbf{d}, \mathbf{z}) := \frac{1}{2}\|\mathbf{d} - \mathbf{z}\|_2^2 + \frac{1}{2}\gamma_t\|\nabla\mathbf{d}\|_F^2$, $\gamma_t := \gamma(a_{j,j}/t)^{-1}$

The proposed algorithm

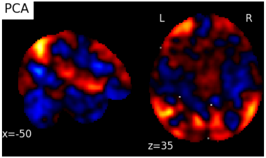
- **Draw a sample** 3D brain image (or mini-batch) $\mathbf{x}_t \in \mathbb{R}^p$
- **Compute loadings** (i.e representation w.r.t current dict. $\mathbf{D}$)

$$\mathbf{c}_t \leftarrow \mathrm{argmin}_{\mathbf{u} \in \mathbb{R}^k} \frac{1}{2}\|\mathbf{x}_t - \mathbf{D}\mathbf{u}\|_2^2 + \frac{1}{2}\alpha\|\mathbf{u}\|_2^2.$$

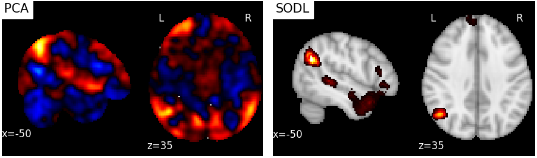
- **Rank-1** updates: $\mathbf{A}_t = \mathbf{A}_{t-1} + \mathbf{c}_t\mathbf{c}_t^T$, $\mathbf{B}_t := \mathbf{B}_{t-1} + \mathbf{x}_t\mathbf{c}_t^T$
- **BCD dictionary update** of dictionary atoms
  - **Precompute** $\mathbf{R} \leftarrow \mathbf{B} - \mathbf{D}\mathbf{A}$
  - **for** $j = 1, 2, \ldots, k$
    - **Rank-1** update: $\mathbf{R} \leftarrow \mathbf{R} + \mathbf{d}^j \circ \mathbf{a}^j$
    - **FISTA loop:** $\mathbf{d}^j \leftarrow \mathrm{argmin}_{\mathbf{d} \in \mathcal{K}} F_{\gamma_t}(\mathbf{d}, a_{j,j}^{-1}\mathbf{r}^j)$
    - **Rank-1** update: $\mathbf{R} \leftarrow \mathbf{R} - \mathbf{d}^j \circ \mathbf{a}^j$

- **N.B.:** $F_{\gamma_t}(\mathbf{d}, \mathbf{z}) := \frac{1}{2}\|\mathbf{d} - \mathbf{z}\|_2^2 + \frac{1}{2}\gamma_t\|\nabla\mathbf{d}\|_F^2$, $\gamma_t := \gamma(a_{j,j}/t)^{-1}$

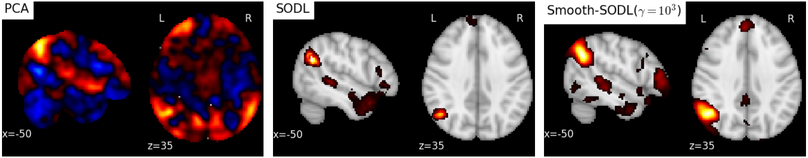
Experimental results on HCP fMRI data: qualitative



Experimental results on HCP fMRI data: qualitative

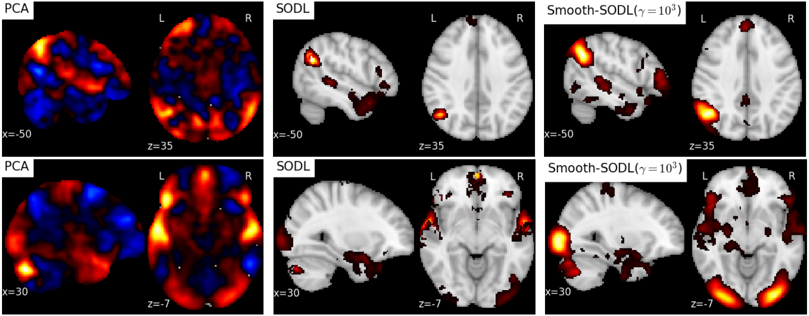


Experimental results on HCP fMRI data: qualitative



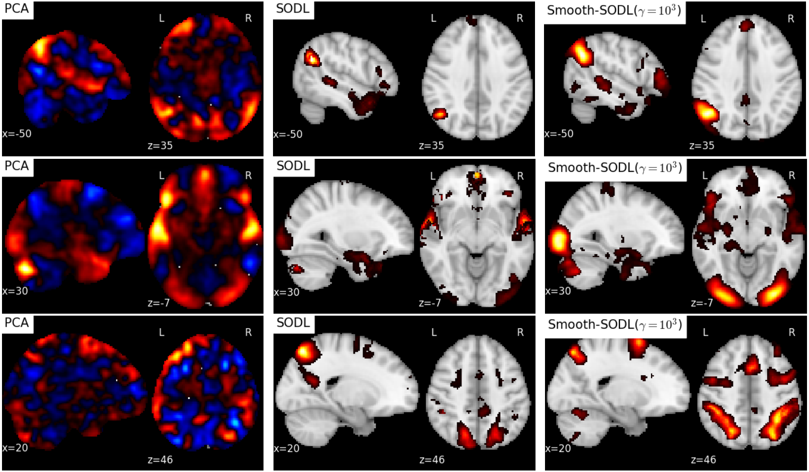
- Our method produces localized and smooth decompositions

Experimental results on HCP fMRI data: qualitative



- Our method produces localized and smooth decompositions

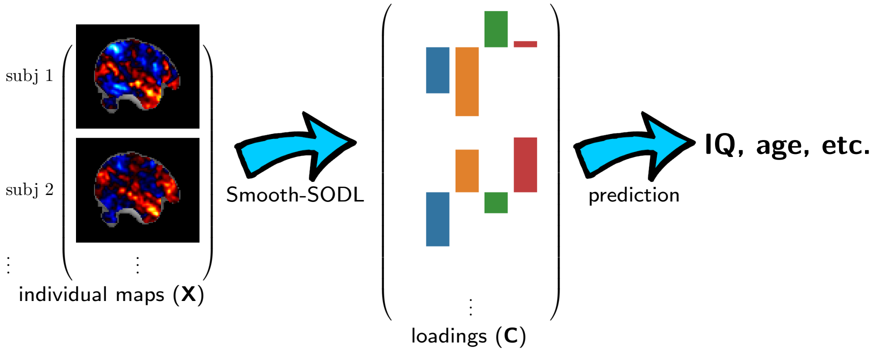
Experimental results on HCP fMRI data: qualitative



- Our method produces localized and smooth decompositions

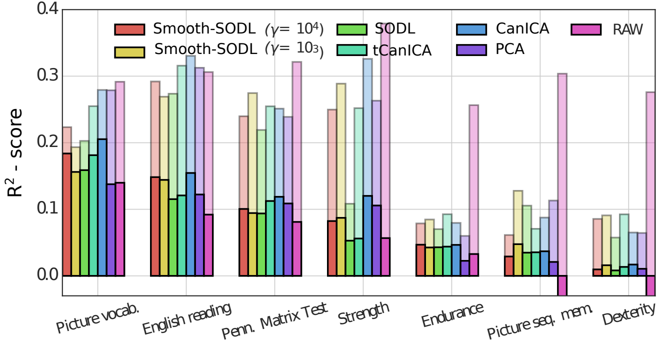
Learned latent dimensions capture inter-subject variability

- Predicting behavior from **compressed representation** of Story vs Math contrast of language task maps [van Essen '12]



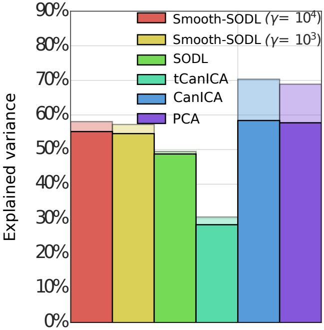
Learned latent dimensions capture inter-subject variability

- Predicting behavior from **compressed representation** of Story vs Math contrast of language task maps [van Essen '12]



- Thick bars $\implies$ scores on **test** set; faint bars $\implies$ on **train**
- Proposed **Smooth-SODL** overfits the least (i.e generalizes best)

What's happening



- Unregularized models **overfit**
- Models thresholded post-training **underfit**

Spatial prior reduces sample-complexity

Nb. subjects	vanilla [Mairal '10]	Proposed model	gain factor
17	2%	**31%**	**13.8**
92	37%	**50%**	**1.35**
167	47%	**54%**	**1.15**
241	49%	**55%**	**1.11**

Learning-curve for "boost" in explained variance of our proposed Smooth-SODL model over the reference SODL model.

Concluding remarks

Concluding remarks

- The goal of this thesis was to develop models for **inter-subject variability**
- "Regions" emerged as the right scale at which to work
  - A more stable representation of activity patterns across subjects, etc.

Concluding remarks

- The goal of this thesis was to develop models for **inter-subject variability**
- "Regions" emerged as the right scale at which to work
  - A more stable representation of activity patterns across subjects, etc.
- We proposed enhanced models and algorithms for **structured penalized multi-variate models** for brain decoding

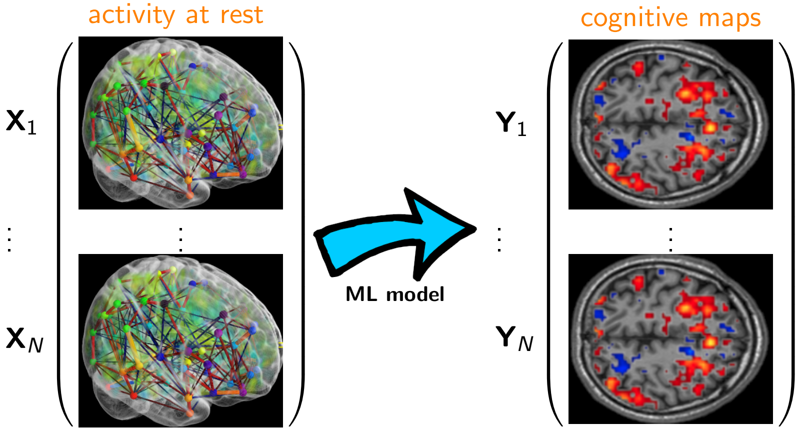
Concluding remarks

- The goal of this thesis was to develop models for **inter-subject variability**
- "Regions" emerged as the right scale at which to work
  - A more stable representation of activity patterns across subjects, etc.
- We proposed enhanced models and algorithms for **structured penalized multi-variate models** for brain decoding
- The notion of regions (via structured priors) was used to develop as the basis for a latent model of inter-subject variability [Dohmatob NIPS '16]

Concluding remarks

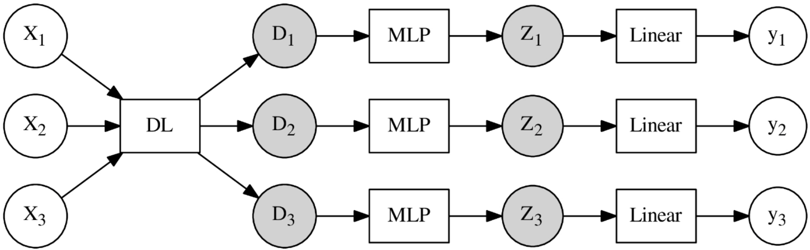
- The goal of this thesis was to develop models for **inter-subject variability**
- "Regions" emerged as the right scale at which to work
  - A more stable representation of activity patterns across subjects, etc.
- We proposed enhanced models and algorithms for **structured penalized multi-variate models** for brain decoding
- The notion of regions (via structured priors) was used to develop as the basis for a latent model of inter-subject variability [Dohmatob NIPS '16]

Can we predict task maps from resting-state data ?



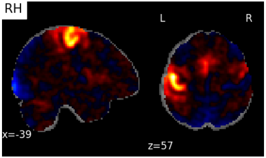
- $\mathbf{X}_s$: resting-state functional connectivity graph for subject s
- $\mathbf{Y}_s$: task-specific activation maps for subject s

Proposal: Deep semi-supervised voxel encoding

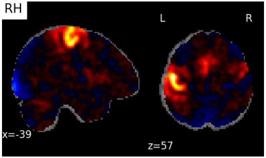


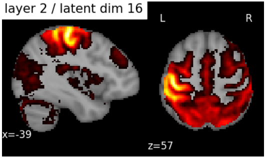
- $\mathbf{Y} \in \mathbb{R}^{p \times C}$: subject-specific GLM maps of brain activity
- $\mathbf{X} \in \mathbb{R}^{p \times T}$: resting-state fMRI data

Preliminary results: learned features

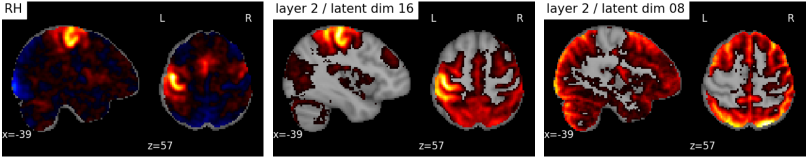


Preliminary results: learned features

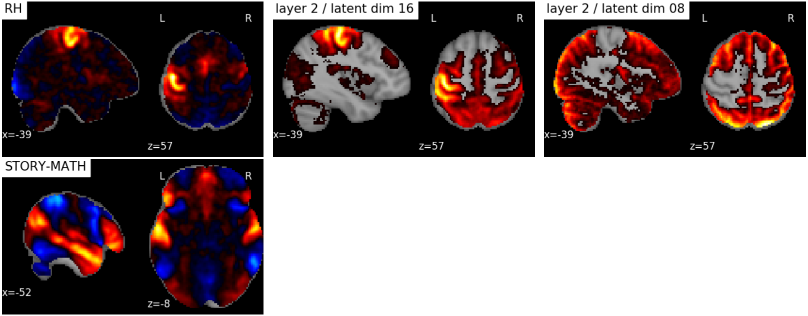




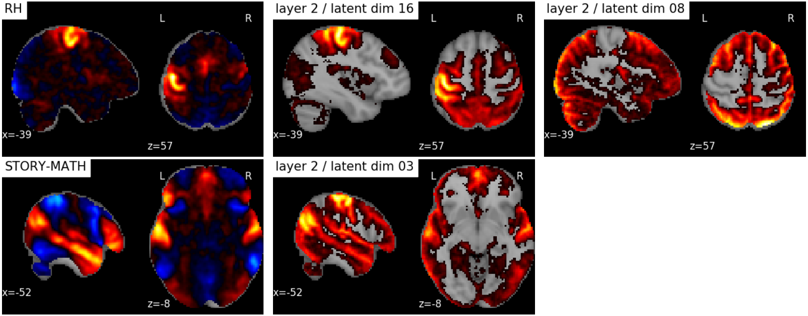
Preliminary results: learned features



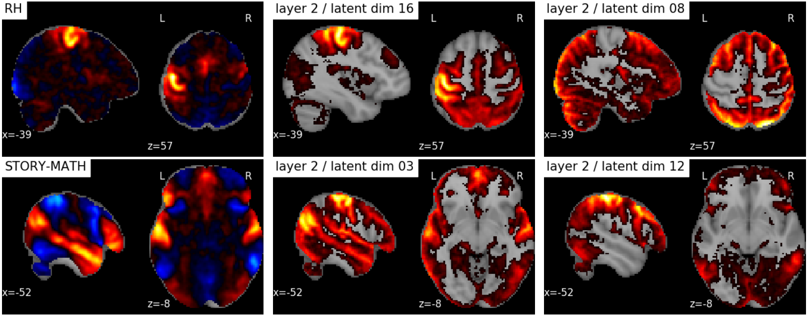
Preliminary results: learned features



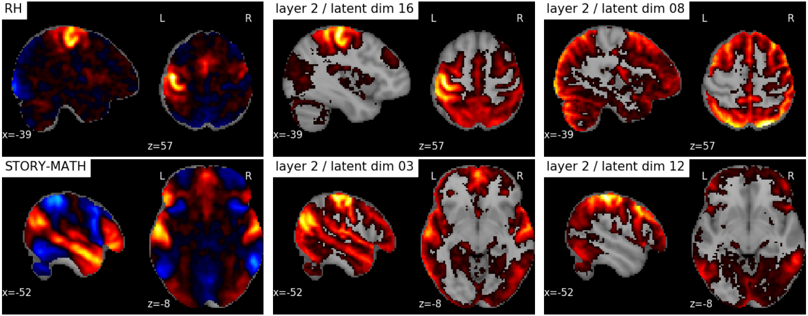
Preliminary results: learned features



Preliminary results: learned features

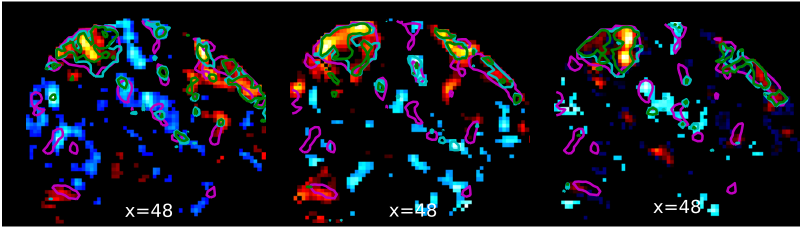


Preliminary results: learned features



- Learned the a presentation of task activity in resting-state space!
- This is ongoing application of models developed in previous sections!

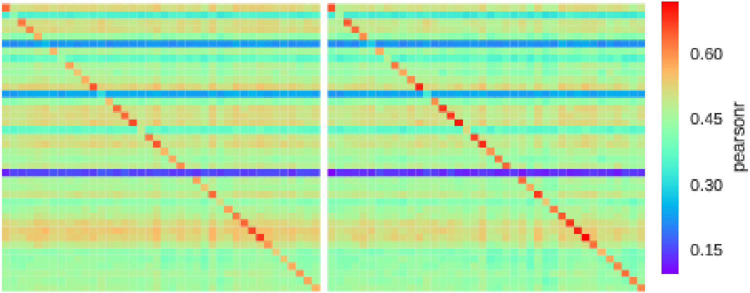
Preliminary results: predicted individual maps



2BK vs 0BK contrast of the Working Memory task
[van Essen '12]

- magenta = population mean
- reference method [Tavor '16]
- proposed method
  - Prediction agrees with subject's topography more faithfully

Preliminary results: quantitative



Confusion matrix for predicted versus true activation maps

Relevant contributions I