Tales from fMRI
Learning from limited labeled data

Gaël Varoquaux
fMRI data

- $p \sim 100,000$ voxels per map
- Heavily correlated + structured noise
- Low SNR: $\sim 5\% \sim -13$ dB

Brain response maps (activation)
- $n \sim$ Hundreds, maybe thousands

Resting-state (no cognitive labels)
- $n \sim 100$–10,000 per subject
- Thousands of subjects
- No salient structure
TL;DR

- Estimators with small sample complexity
- Increasing the amount of data
Outline of this talk

1. Regularizing linear models
2. Covariance estimation
3. Merging data sources
1 Regularizing linear models

\[ \text{sign}(Xw + e) = y \]

Design matrix × Coefficients = Target

\( p \sim 50,000 \)
\( n \sim 100 \) per category
Regularizing linear models

From sparsity to structure, to ensembling

Design matrix $\times$ Coefficients = Target

$p \sim 50,000$  
$n \sim 100$ per category
Sample complexity, $\ell_1$ versus $\ell_2$ regularization

Def Sample complexity: $n$ required for small error w.h.p.

**Rotationally invariant estimators are data hungry**

Thm For rotational invariant estimators,

\[ \text{sample complexity} > O(p) \]

[Ng 2004]
Sample complexity, $\ell_1$ versus $\ell_2$ regularization

**Def** Sample complexity: $n$ required for small error w.h.p.

**Rotationally invariant estimators are data hungry**

Thm For rotational invariant estimators, sample complexity $> O(p)$  

[Ng 2004]

**Sparsity, compressive sensing**

To recover $k$ non-zero coefficients, $n \sim k \log p$  

[Wainwright 2009]
Sample complexity, $\ell_1$ versus $\ell_2$ regularization

Def Sample complexity: $n$ required for small error $w.h.p.$

Rotationally invariant estimators are data hungry

Thm For rotational invariant estimators, sample complexity $> \mathcal{O}(p)$ [Ng 2004]

Sparsity, compressive sensing

To recover $k$ non-zero coefficients, $n \sim k \log p$ [Wainwright 2009]

Fragile to correlations in the design
Correlated design on support breaks $\ell_1$ beyond repair
Total-variation penalization

Impose sparsity on the gradient of the image:

\[ p(w) = \ell_1(\nabla w) \]

In fMRI: [Michel... 2011]
Structured sparsity: variations on Total variation

**Total-variation penalization**

Impose sparsity on the gradient of the image:

\[ p(w) = \ell_1(\nabla w) \]

In fMRI: [Michel... 2011]

\[ \hat{w} = \arg\min_w l(y - Xw) + \lambda \left( \rho \ell_1(w) + (1 - \rho) TV(w) \right) \]

\( l \): data-fit term  
\[ [\text{Baldassarre... 2012, Gramfort... 2013}] \]

TV-\( \ell_1 \): \text{Sparsity} + \text{regions}
Structured sparsity: variations on Total variation

**Total-variation penalization**

Impose sparsity on the gradient of the image:

\[ p(w) = \ell_1(\nabla w) \]

In fMRI: [Michel... 2011]

**Analysis sparsity:** \[ \|Kw\|_{21} \] [Eickenberg... 2015]

**TV-\ell_1:** Sparsity + regions

\[ \hat{w} = \arg\min_w l(y - Xw) + \lambda \left( \rho \ell_1(w) + (1 - \rho) TV(w) \right) \]

\( l: \) data-fit term [Baldassarre... 2012, Gramfort... 2013]
1. TV-\( \ell_1 \) works

- Good prediction performance

- Segment the relevant regions

SVM  ridge  sparse  TV-\( \ell_1 \)
TV-$\ell_1$ works

- Good prediction performance
- Segment the relevant regions

Computational costly
- Hyper-parameter selection brittle
- Tedious convergence
Hyper-parameter selection

Hyper-parameter setting important for $\ell_1$ models

- Cross-validation, rather than hold-out, for small $n$
Hyper-parameter selection

Cross-validation

Subsampling

Fitting with each hyperparameter

Average models with the same hyperparameter

Select the best hyperparameter

Refit the model (final model)
Hyper-parameter selection

CV-bagging

Subsampling

Fitting with each hyperparameter

Select the best model per CV fold

Averaging (final model)

Bagging reduces variance

[Maillard... 2017, McInerney 2017]


Fixing sparsity with clustering

**Idea:** cluster together correlated features

1. clustering to form reduced features
2. sparse linear model on reduced features

[Varoquaux... 2012]
Fixing sparsity with clustering and bagging

**Idea:** cluster together correlated features

1. Loop: perturb randomly data
2. Clustering to form reduced features
3. Sparse linear model on reduced features
4. Accumulate non-zero features

[Varoquaux... 2012]

G Varoquaux
FREM: ensembling everything

Bagging:
- Clustering for dimension reduction
- Feature selection
- Linear model
- Hyper-parameter selection (CV-bagging)

Empirical results

<table>
<thead>
<tr>
<th></th>
<th>prediction score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-10%</td>
</tr>
<tr>
<td>Graph-net</td>
<td></td>
</tr>
<tr>
<td>TV-$\ell_1$</td>
<td></td>
</tr>
<tr>
<td>Log-enet</td>
<td></td>
</tr>
<tr>
<td>SVM-$\ell_2$</td>
<td></td>
</tr>
<tr>
<td>SVM-$\ell_1$</td>
<td></td>
</tr>
<tr>
<td>FReM</td>
<td></td>
</tr>
</tbody>
</table>
**FREM**: ensembling everything [Hoyos-Idrobo... 2017]

**Bagging:**
- Clustering for dimension reduction
- Feature selection
- Linear model
- Hyper-parameter selection (CV-bagging)

**Empirical results**

<table>
<thead>
<tr>
<th></th>
<th>Prediction Score</th>
<th>Weight Stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph-net</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TV-$\ell_1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log-enet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM-$\ell_2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM-$\ell_1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FReM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
FREM: ensembling everything [Hoyos-Idrobo... 2017]

Bagging:
- Clustering for dimension reduction
- Feature selection
- Linear model
- Hyper-parameter selection (CV-bagging)

Empirical results
Lessons learned trying to regularize

- Sparsity is not enough: structure is needed
- Optimal sparse-structured is finicky and expensive
- Ensemble greedy approaches

Empirical results

<table>
<thead>
<tr>
<th>Model</th>
<th>Prediction Score</th>
<th>Weight Stability</th>
<th>Computation Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph-net</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TV-$\ell_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log-enet</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM-$\ell_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM-$\ell_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FReM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Graphs of brain function
Covariances capture interactions between regions
Gaussian graphical models

- Multivariate normal:

\[ \mathcal{P}(\mathbf{X}) \propto \sqrt{|\Sigma^{-1}|} e^{-\frac{1}{2} \mathbf{X}^T \Sigma^{-1} \mathbf{X}} \]

- Model parametrized by inverse covariance matrix, \( \mathbf{K} = \Sigma^{-1} \): *conditional* covariances

\[ \mathbf{X}_i \perp \perp \mathbf{X}_j \iff K_{i,j} = 0 \]

- Graphical lasso: \( \ell_1 \)-penalized MLE

Maximum-likelihood of \( \mathbf{K} \) needs \( O(p^2) \) samples.

\( \ell_1 \) enables support recovery [Ravikumar... 2011]
2 Gaussian graphical models

- Multivariate normal:

$$\mathcal{P}(X) \propto \sqrt{|\Sigma^{-1}|} e^{-\frac{1}{2} X^T \Sigma^{-1} X}$$

- Model parametrized by inverse covariance matrix, $K = \Sigma^{-1}$: conditional covariances

  $$X_i \perp \perp X_j \iff K_{i,j} = 0$$

  Sample complexity of recovering $s$ edges

  $$n = \mathcal{O}((s + p) \log(p)) \iff s = o(\sqrt{p})$$

- Graphical lasso: $\ell_1$-penalized MLE

  Maximum-likelihood of $K$ needs $\mathcal{O}(p^2)$ samples.

  $\ell_1$ enables support recovery $[Ravikumar... 2011]$
Common independence structure but different connection values

\[
\{ K^s \} = \arg\min_{\{ K^s \succ 0 \}} \sum_s \mathcal{L}(\hat{\Sigma}^s | K^s) + \lambda \ell_{21}(\{ K^s \})
\]

Multi-subject data fit, Likelihood

Group-lasso penalization

[Varoquaux... 2010]
Common independence structure but different connection values

\[
\{K^s\} = \arg\min_{\{K^s \succ 0\}} \sum_s \mathcal{L}(\hat{\Sigma}^s | K^s) + \lambda \ell_{21}(\{K^s\})
\]

Multi-subject data fit, Likelihood

\(\ell_1\) on the connections of the \(\ell_2\) on the subjects
Is sparse recovery the right question?

Our goal may be to compare patients

- $\ell_1$ recovery is unstable
- Brain graphs are not that sparse
  Between-subject differences may be sparse

[Belilovsky... 2016]

Which risk should we minimize on the covariance?
James-Stein shrinkage

To estimate a mean $\theta$:

$$\hat{\theta}_{JS} = (1 - \alpha) \hat{\theta}_{MLE} + \alpha \hat{\theta}_{\text{guess}}$$

whith $\alpha \sim \frac{\sigma^2}{n\|\theta - \theta_{\text{guess}}\|}$

$\hat{\theta}_{JS}$ dominates $\hat{\theta}_{MLE}$ for the MSE

Ledoit-Wolf covariance shrinkage estimator

$$\hat{\Sigma}_{LW} = (1 - \alpha) \Sigma_{MLE} + \alpha \text{trace}(\Sigma_{MLE}) I$$

with $\alpha$ oracle for $n \to \infty, \frac{n}{p} \to \text{cst}$

$\hat{\Sigma}_{LW}$ dominates $\hat{\Sigma}_{MLE}$ for the MSE

[Ledoit and Wolf 2004]
James-Stein shrinkage
To estimate a mean $\theta$:

$$\hat{\theta}_{JS} = (1 - \alpha) \theta_{MLE} + \alpha \theta_{\text{guess}}$$

with $\alpha \sim \frac{\sigma^2}{n\|\theta - \theta_{\text{guess}}\|}$

For inter-subject comparison, Ledoit-Wolf performs as well as $\ell_1$ estimators, but faster & less brittle.

Ledoit-Wolf covariance shrinkage estimator

$$\hat{\Sigma}_{LW} = (1 - \alpha) \Sigma_{MLE} + \alpha \text{trace}(\Sigma_{MLE}) I$$

with $\alpha$ oracle for $n \to \infty, \frac{n}{p} \to \text{cst}$

$\hat{\Sigma}_{LW}$ dominates $\hat{\Sigma}_{MLE}$ for the MSE

[Ledoit and Wolf 2004]
Shrinkage with order-2 moment

Shrinkage = MMSE = Bayesian posterior mean for Gaussian distribution 4.1.2 [Lehmann and Casella 2006]

⇒ Use prior $\mathcal{N}(\Sigma_0, \Lambda_0)$ learned on population

* $\Lambda_0$ is a covariance on covariances
Shrinkage with order-2 moment

- Shrinkage $= \text{MMSE} = \text{Bayesian posterior mean}$ for Gaussian distribution 4.1.2 [Lehmann and Casella 2006]
  $\Rightarrow$ Use prior $\mathcal{N}(\Sigma_0, \Lambda_0)$ learned on population
  * $\Lambda_0$ is a covariance on covariances

Information geometry / covariance manifold

- Covariances are not a vector space
- Computations on the manifold
- Turns MLE into an MSE

PoSCE: Population shrinkage of covariance

G Varoquaux [Rahim... 2017, 2018]
James-Stein shrinkage for population models

Inter-session reproducibility within subjects

- Correlation matrix
- GraphLasso CV
- Ledoit-Wolf
- Identity shrinkage CV
- Prior shrinkage CV
- PoSCE

Anisotropic shrinkage for the win

PoSCE: Population shrinkage of covariance

G Varoquaux [Rahim... 2017, 2018]
3 Merging data sources

More data trumps fancy regularizations

[Mensch... 2017]
There is plenty of fMRI data
Dozens of thousands of fMRI sessions, but terribly heterogeneous.

Measurements of brain activity

Descriptions of behavior and cognition

Heterogeneity in the behavior
Formal modeling of behavior is an open knowledge representation problem.

G Varoquaux
There is plenty of fMRI data. Dozens of thousands of fMRI sessions, but terribly heterogeneous.

- Unsupervised learning on fMRI data
- Multi-task learning across studies

Heterogeneity in the behavior
Formal modeling of behavior is an open knowledge representation problem.
Mapping cognition across studies labels

Cognitive label across many studies?
Very difficult to assign

- Visual
- Auditory
- Foot
- Hand
- Calculation
- Reading
- Checkboard
- Face
- Place
- Object
- Digit
- Saccade

...
Mapping cognition across studies labels

Cognitive label across many studies?
Very difficult to assign

"Multi-task learning"
- Solve many related but different problems
- Learn commonalities

Visual
Auditory
Foot
Hand
Calculation
Reading
Checkboard
Face
Place
Object
Digit
Saccade

G Varoquaux
Sharing representations across tasks

Great for multiple output (tasks)

[Bengio 2009]
Great for multiple output (tasks)

Millions of parameters, thousands of data points

[Bengio 2009]
3 Sharing representations across tasks

- Great for multiple output (tasks)
- Millions of parameters, thousands of data points

Simplify
Sharing representations across tasks

- Great for multiple output (tasks)
- Millions of parameters, thousands of data points

Simplify simplify more

G Varoquaux
Sharing representations across tasks

- Great for multiple output (tasks)
- Millions of parameters, thousands of data points

Simplify simplify more

[Bzdok... 2015]
Unsupervised learning for spatial atoms

Decomposing time series into spatial maps with sparsity to localize atoms

\[ Y = E \cdot S + N \]
Unsupervised learning for spatial atoms

Decomposing time series into spatial maps with sparsity to localize atoms

Adapted representations that capture local correlations
Unsupervised learning for spatial atoms

\[ Y = E \cdot S + N \]

Decomposing time series into spatial maps with sparsity to localize atoms

- Adapted representations that capture local correlations
- More data is always better

1Tb data

50Gb data

computational cost [Mensch... 2016]
3 Multi-task across studies

- Decode in each study
- But learn representations across

Loss-engineering & regularization

[Mensch... 2017]
Multi-task across studies

- Decode in each study
- But learn representations across studies
- Loss-engineering & regularization

[Mensch... 2017]

G Varoquaux
Learning with limited labeled data: fMRI lessons

- Sparse models are unstable and need ensembling
- Parameter selection is unstable and needs ensembling
- $\ell_2$ shrinkage is powerful, in particular with good mean & covariance
- Unsupervised learning of representations
- Multi-task to pool data

@GaelVaroquaux
Learning with limited labeled data: fMRI lessons

- Sparse models are unstable and need ensembling
- Parameter selection is unstable and needs ensembling
- $\ell_2$ shrinkage is powerful, in particular with good mean & covariance
- Unsupervised learning of representations
- Multi-task to pool data
- Software for machine learning in neuroimaging: http://nilearn.github.io

@GaelVaroquaux


