Machine learning for brain imaging

Gaël Varoquaux

INRIA/Parietal

INRIA

PARIETAL

NeuroSpin
1. An introduction to brain imaging
2. Learning to diagnose
3. Understanding brain function
4. Spatial penalties for learning from images
5. Learning from spontaneous activity
6. Beyond equations: software
1 An introduction to brain imaging

- The brain: its anatomy and function
- Imaging the brain
The brain: its anatomy and function

- Two hemispheres
- “sucli”
Inside a brain: neuroanatomy in one minute
Inside a brain: neuroanatomy in one minute

Grey matter

White matter
Inside a brain: neuroanatomy in one minute

Grey matter

White matter

Cortex

G Varoquaux
How the brain works: brain function

The language circuit:

- Speech
- Braille
- Somato-sensory cortex
- Wernicke's area
- Written words
- Visual cortex
- Auditory cortex
- Motor cortex
- Spoken words
- Broca's area

G Varoquaux
The language circuit:

Historically, regions ↔ function link discovered via impact of brain lesions
General neuroscience view of brain function:

- Specialized “brain modules” for atomic function
- Recruited together in “functional networks” for high-level function
The visual system: a computational model

- V1 cortex
- V2 cortex
- Inferior temporal cortex
- Fusiform face area
- Image

Jack?
The visual system: a computational model

Close ties to convolutional neural networks in computer vision

1. How the brain works: brain function

G Varoquaux
Imaging the brain

3D images
White matter, grey matter

Cortex
functional MRI (fMRI)

Time-resolved recordings of brain activity
Blood Oxygen Level Dependent effect

- MRI probes local magnetic properties
- oxyHemoglobin and deoxyHemoglobin have different magnetic susceptibility

- Neural activity consumes oxygen
  ⇒ Initial dip in oxyHemoglobin

- Metabolism compensates
  ⇒ Increase in oxyHemoglobin
**Blood Oxygen Level Dependent effect**

- MRI probes local magnetic properties
- oxyHemoglobin and deoxyHemoglobin have different magnetic susceptibility

- Neural activity consumes oxygen ⇒ Initial dip in oxyHemoglobin
- Metabolism compensates ⇒ Increase in oxyHemoglobin

- Relative effect, not absolute ⇒ Need to contrast values
- Very indirect effect ⇒ Non-linearities, inhomogeneities, lags...

All models are wrong
Mapping cognition with fMRI

Careful crafting of contrasts to isolate high-level cognition.
Mapping cognition with fMRI

Stimulus

Activation maps

Contrast
Cognitive processes

Careful crafting of contrasts to isolate high-level cognition

G Varoquaux
Mapping cognition with fMRI

Stimulus

Activation maps

Contrast
Cognitive processes

Careful crafting of contrasts to isolate high-level cognition
1 Mapping cognition with fMRI

Stimulus

Activation maps

Careful crafting of contrasts to isolate high-level cognition
Magneto and electro encephalography

- Measure electromagnetic field created by neural fixing

- “Interesting” inverse problem to reconstruct sources on cortex
  ⇒ Poor spatial resolution

- Great temporal resolution
Learning to diagnose
Diagnostic applications: promises and challenges
Diagnosis
Finding the nature or cause of a disease condition

Pronosis
Predicting the future evolution of the condition
⇒ Therapeutic indications

Early biomarkers
Measures enabling the detection of disease before standard symptoms
⇒ Population screening if cheap

Quantitative biomarkers
Metric to follow disease progression
⇒ Drug development
2 Specificity – sensitivity trade-offs

Depending on application, different types of error may need to be weighted differently

*E.g.*: Screening before human check
  Low false negative rate, high false positive rate

Detection leading to surgery
  Low false positive rate
2 More than prediction accuracy

Cannot replace the physician:
- Patient history
- Therapeutic strategies subject to logistics ...

⇒ No black-box

Segmentation, denoising task

as much as prediction
More than prediction accuracy

Cannot replace the physician:
- Patient history
- Therapeutic strategies subject to logistics

⇒ No black-box

Segmentation, denoising task

as much as prediction
The training set is not representative

- Early or weakly-symptomatic patients not represented; these are the most interesting
- Label noise: wrong diagnostic on difficult patients ⇒ validation difficult
- Confounding factors for patients
  ⇒ Epidemiological studies (biobank, UK)
  select subjects randomly from normal population, and follow them
  but imaging cost prohibitive
The training set is not representative

- Early or weakly-symptomatic patients not represented; these are the most interesting.
- Label noise: wrong diagnostic on difficult patients → validation difficult.
- Confounding factors for patients.

Real problems are not patient/control classifications, but multi-pathology classifications.

Datasets are small (not many subjects)
- Researchers don’t share data
- Large inter-site variability
Features from brain images

Brains are not translation-invariant
We don't understand much about the brain
but we still know something
Features from brain images

Not a standard computer vision pipeline!

- Brains are not translation-invariant
- We don’t understand much about the brain but we still know something
What are we trying to capture?

- Changes in brain shape?
- Local grey-matter changes?
- Micro-lesions?
Features in anatomical images

Changes in brain shape?
Local grey-matter changes?
Micro-lesions?

Across subjects
Feature extraction in anatomical images

Segment structures / tissues

Realign / learn correspondences

Features in structures or in correspondence vectors
- Voxel-based morphometry: grey matter density
- Cortical thickness
- Shape descriptors of cortex
- Realignment transformation field

G Varoquaux
Functional images are low SNR
Why use them?
Functional images are low SNR
Why use them?

- Behavioral deficit vary for a given lesion
- Neuro-psychiatric diseases (autism, schizophrenia) = deficit of function with no known anatomy

Des troubles de la parole
Features in functional images

Activation maps to dissect cognitive effects

Images
Features in functional images

Activation maps to dissect cognitive effects

Images
Reading
Counting

Often: one map per stimuli presentation

[Poldrack 2011]
3 Understanding brain function

Cognitive neuroimaging
Learn a bilateral link between brain activity and cognitive function
Predicting neural response: encoding models
Predicting neural response: encoding models
“Brain reading”: decoding
3 Brain mapping

fMRI data > 50,000 voxels

stimuli

Standard analysis
Detect voxels that correlate to the stimuli
Brain mapping ⇔ brain reading

- Predicting the object category viewed
  
  [Haxby 2001, Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex]

**Supervised learning task**

**Predictive modeling**
Find combinations of voxels to best predict
Brain mapping ⇔ brain reading

Predicting the object category viewed

[Haxby 2001, Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex ]

Take home message:

brain regions, not prediction

Face area
Place area

Predictive modeling
Find combinations of voxels to best predict
Recovery rather than prediction

Regions matter as much as prediction score
Recovery rather than prediction

Regions matter as much as prediction score

Danger of solving the wrong problem

Lost in formalization
4 Spatial penalties for learning from images

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

Design matrix $\times$ Coefficients $= \text{Target}$

$p \sim 50000$

$n \sim 100 \text{ per category}$
Spatial penalties for learning from images

- Small sample linear model estimation
- Random correlated design

Design matrix $\times$ Coefficients $=$ Target
Minimize an error term:

$$\hat{w} = \arg\min_w l(y - X w)$$

Ill-posed: $X$ is not full rank

Inject prior: regularize

$$\hat{w} = \arg\min_w l(y - X w) + p(w)$$
Estimation: statistical learning

**Inverse problem**

- Minimize an error term:

\[
\hat{w} = \text{argmin}_w l(y - Xw)
\]

Ill-posed: \(X\) is not full rank

Inject prior: regularize

\[
\hat{w} = \text{argmin}_w l(y - Xw) + p(w)
\]

**Example: Lasso = sparse regression**

\[
\hat{w} = \text{argmin}_w \|y - Xw\|^2_2 + \ell_1(w)
\]

\[
\ell_1(w) = \sum_i |w_i|
\]

\[
\text{Error term} + \text{Penalty}
\]
Good prediction ≠ good recovery

Simulations
Ground truth

Lasso
Prediction: 0.78
Recovery: 0.429

SVM
Prediction: 0.71
Recovery: 0.486

Need a method suited for recovery
Brain mapping & $\ell_1$ sparse recovery

Recovering brain regions
Recovering \( k \) non-zero coefficients

- \( n_{\text{min}} \sim 2k \log p \)

- Restricted-isometry-like property:
  The design matrix is well-conditioned on sub-matrices of size \( > k \)

- Mutual incoherence:
  Relevant features \( S \) and irrelevant ones \( \overline{S} \) are not too correlated

**Violated by spatial correlations in our design** 😞

Candes 2006

Tropp 2004

Wainwright 2009
Randomized sparsity

[Meinshausen and Buhlmann 2010, Bach 2008]

- Perturb the design matrix:
  - Subsample the data
  - Randomly rescale features

- Run sparse estimator

- Keep features that are often selected

⇒ **Good recovery without mutual incoherence**
  But RIP-like condition

**Cannot recover large correlated groups**

For $m$ correlated features, selection frequency divided by $m$
Spatially-connected hierarchical clustering

⇒ reduces voxel numbers [Michel Pat Rec 2011]

Replace features by corresponding cluster average

+ Use a supervised learner on reduced problem

Cluster choice sub-optimal for regression
1st approach: randomized clustering

Combining Clustering

Sparsity

[Varoquaux ICML 2012]
**Hypothesis**: clustering compatible with support($w$)

**Benefits of clustering**
- Reduced $k$ and $p$
  - $n > n_{\text{min}}$: good side of the “sharp threshold”
- Cluster together correlated features
  - Improves RIP-like conditions

Recovery possible on reduced features
Randomized parcellations + sparsity

Randomization + Stability scores

- Marginalize the cluster choice
- Relaxes mutual incoherence requirement
Algorithm

1. set $n_{clusters}$ and sparsity by cross-validation

2. loop: perturb randomly data

3. clustering to form reduced features

4. sparse linear model on reduced features

5. accumulate non-zero features

6. threshold map of apparition counts
Simulations

- \( p = 2048, k = 64, n = 256 \) \( (n_{\text{min}} > 1000) \)
- Weights \( \mathbf{w} \): patches of varying size
- Design matrix: 2D Gaussian random images of varying smoothness

Estimators
- Randomized lasso
- Elastic Net

- Our approach
- Univariate F test
4 When can we recover patches?

- Smoothness helps (reduces noise degrees of freedom)
- Small patches are hard to recover
What is the best method for patch recovery?

- For small patches: elastic net
- For large patches: randomized-clustered sparsity
- Large patches and very smooth images: F-test
Randomizing clusters matters!

- Non-random (Ward) clustering inefficient
- Fully-random performs quite well
- Randomized Ward gives an extra gain
Randomizing clusters matters!

Non-random (Ward) clustering inefficient

Fully-random performs quite well

Randomized Ward gives an extra gain

Degenerate family of cluster assignments
Univariate F-scores

[Haxby 2001]
fMRI: face vs house discrimination [Haxby 2001]

$l_1$ Logistic

G Varoquaux
Randomized $\ell_1$ Logistic
Randomized Clustered $\ell_1$ Logistic
fMRI: face vs house discrimination [Haxby 2001]

F-scores

G Varoquaux
1st approach: randomized clustering

Sparse recovery of patches on spatially-correlated designs

Ingredients: Clustering + Randomization
⇒ Reduced feature set compatible with recovery:
matches sparsity pattern + recovery conditions
[Varoquaux ICML 2012]

How to fit in compressive sensing theory?
2nd approach: analysis sparsity

\[ \ell_1 \text{ penalty with an } \textit{analysis} \text{ operator: } p(w) = \ell_1(Kw) \]

- Spatial regularization necessary
- Neuroscientists think in terms of brain regions

Total-variation penalization

Impose sparsity on the gradient of the image:

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

In fMRI: [Michel TMI 2011]
2nd approach: analysis sparsity

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

Spatial regularization necessary

Neuroscientists think in terms of brain regions

Total-variation penalization

Impose sparsity on the gradient of the image:

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

In fMRI: [Michel TMI 2011]
\[ \hat{w} = \arg\min_w l(y - Xw) + TV(w) \]

- \( l \): least-square or logistic-regression
- \( p \): TV: isotropic total variation: \( \ell_{21}(\nabla w) \)

**Prediction performance:**

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature screening + SVC</td>
<td>0.77</td>
</tr>
<tr>
<td>Sparse regression</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>Total Variation</strong></td>
<td><strong>0.84</strong></td>
</tr>
<tr>
<td>(explained variance)</td>
<td></td>
</tr>
</tbody>
</table>

[Michel TMI 2011]
\[
\hat{w} = \arg\min_w l(y - Xw) + TV(w)
\]

- \(l\): least-square or logistic-regression
- \(p\): TV: isotropic total variation: \(\ell_{21}(\nabla w)\)

Prediction performance:
- Feature screening + SVC: 0.77
- Sparse regression: 0.78
- Total Variation: 0.84 (explained variance)

[Michel TMI 2011]
\[
\hat{w} = \arg\min_w \|y - Xw\| + TV(w) + \ell_1(w)
\]

- Adding \(\ell_1\) = extending analysis operator
- Retreives sparsity in the original coefficients

On simulations: sparse recovery precision-recall
Convergence matters

Stopping:
\[ \Delta E < 10^{-1} \]

Stopping:
\[ \Delta E < 10^{-3} \]
Convergence matters

Stopping:
\[ \Delta E < 10^{-1} \]

Stopping:
\[ \Delta E < 10^{-5} \]
Optimization algorithms: FISTA

- FISTA loop for regression:
  1. Gradient descent on the datafit term
  2. Proximal operator for TV, computed with an inner FISTA loop

**Bottleneck**: Gradient descent step costly

\[
\hat{w} = \arg\min_w \|y - Xw\|_2^2 + TV(w) + \ell_1(w)
\]

Gradient: \(X^t X w\) with \(X\) big and dense

[Domathob PRNI 2014]
Take home messages

- Recovery $\neq$ prediction

- Univariate tests work very well to recover
  Also in genomics [Haury PLOS One 2011]

- Some form of spatial regularization usefull

- Speed matters! $\mathbf{X} : (100,000, 600)$
Learning from spontaneous activity from connectomes
5 Resting-state: Spontaneous activity

In the absence of explicit tasks, cognitive circuits are recruited spontaneously.

⇒ Meaningful cofluctuation patterns

Diagnostic interest for disabled patients
In the absence of explicit tasks cognitive circuits are recruited spontaneously
⇒ Meaningful co-fluctuation patterns

**Diagnostic interest for disabled patients**

1. Learn spatial maps/regions – *Unsupervised learning*
Resting-state: Spontaneous activity

- In the absence of explicit tasks, cognitive circuits are recruited spontaneously

$\Rightarrow$ Meaningful co-fluctuation patterns

Diagnostic interest for disabled patients

1. Learn spatial maps/regions – *Unsupervised learning*
2. Learn interaction graph – *Unsupervised learning*
In the absence of explicit tasks, cognitive circuits are recruited spontaneously, leading to meaningful co-fluctuation patterns. This has diagnostic interest for disabled patients.

1. Learn spatial maps/regions – *Unsupervised learning*
2. Learn interaction graph – *Unsupervised learning*
3. Inter-subject/inter-condition prediction – *Supervised learning*

[Varoquaux MICCAI 2010, NIPS 2010, 2011 IPMI, ...]
Working hypothesis:
Observing linear mixtures of networks at rest

Time courses
Working hypothesis:
Observing linear mixtures of networks at rest
Working hypothesis:

Observing linear mixtures of networks at rest
**Working hypothesis:**

Observing linear mixtures of networks at rest
Working hypothesis:
Observing linear mixtures of networks at rest
Working hypothesis:
Observing linear mixtures of networks at rest
Working hypothesis:
Observing linear mixtures of networks at rest
**Working hypothesis:**

Observing linear mixtures of networks at rest

Time courses

Ventral Att.
**Working hypothesis:**
Observing linear mixtures of networks at rest
**Working hypothesis:**
Observing linear mixtures of networks at rest

How to unmix networks?
Segmenting regions from spontaneous activity

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

Linear decomposition model
- spatial maps, \( S \)
- residuals, \( N \)
- time series, \( E \)

Independant Component Analysis

[Beckman TMI 2004, Varoquaux NeuroImage 2010]
Segmenting regions from spontaneous activity

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

**Linear decomposition model**
- spatial maps, \( S \)
- residuals, \( N \)
- time series, \( E \)

**TV-\( l_1 \) penalty on maps \( S \)**
The connectome classification pipeline

1. RS-fMRI
2. Functional connectivity
3. Diagnosis
4. Leave-one-site out cross-validation

Preliminary results on Autism
Leave-one-site out cross-validation
⇒ 72% accuracy
Beyond equations: software

- How to we reach our target audience (neuroscientists)?
- How do we disseminate our ideas?
- How do we facilitate new ideas?
Python as a scientific environment

- General purpose
- Easy, readable syntax
- Interactive (ipython)
- Great scientific libraries (numpy, scipy, matplotlib...)

G Varoquaux
Growing a software stack

- Code lines are costly
  - Open source + community driven
  - Need quality and impact
  - Focus on the general purpose libraries first

**Scikit-learn: machine learning in Python**
http://scikit-learn.org

[Pedregosa 2011]
### Computational performance

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>scikit-learn</th>
<th>mlpy</th>
<th>pybrain</th>
<th>pymvpa</th>
<th>mdp</th>
<th>shogun</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>5.2</td>
<td>9.47</td>
<td>17.5</td>
<td>11.52</td>
<td>40.48</td>
<td>5.63</td>
</tr>
<tr>
<td>LARS</td>
<td>1.17</td>
<td>105.3</td>
<td>-</td>
<td>37.35</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>0.52</td>
<td>73.7</td>
<td>-</td>
<td>1.44</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>kNN</td>
<td>0.57</td>
<td>1.41</td>
<td>-</td>
<td>0.56</td>
<td>0.58</td>
<td>1.36</td>
</tr>
<tr>
<td>PCA</td>
<td>0.18</td>
<td>-</td>
<td>-</td>
<td>8.93</td>
<td>0.47</td>
<td>0.33</td>
</tr>
<tr>
<td>k-Means</td>
<td>1.34</td>
<td>0.79</td>
<td>∞</td>
<td>-</td>
<td>35.75</td>
<td>0.68</td>
</tr>
</tbody>
</table>

- Algorithms rather than low-level optimization
- Convex optimization + machine learning
- Avoid memory copies
Community
- 200 contributors since 2008, 1500 github forks
- 25 contributors in latest release (3 months span)

Why this success?
- Trendy topic?
- Low barrier of entry
- Friendly and very skilled mailing list
- Credit to people
6 Research code ≠ software library

Factor 10 in time investment

- Corner cases in algorithm (numerical stability)
- Multiple platforms and library versions (Blas 😞)
- Documentation
- Making it simpler (and get less educated users)
- User and developer support (~ 100 mails/day)

Exhausting, but has impact on science and society
Research code ≠ software library

Factor 10 in time investment

Technical + scientific tradeoffs

■ Ease of install/ease of use rather than speed

■ Focus on “old science”

■ Nice publications and theorems are not a recipe for useful code

Exhausting, but has impact on science and society

G Varoquaux
Project scope

Machine learning for neuroimaging:
make using scikit-learn on neuroimaging easy

The target user base is small 😞

Examples in the docs

- Run out of the box, downloading open data
- Produce a clear figure

Data from Miyawaki 2008

Routine, simple, reproduction of papers

G Varoquaux
NeuroSynth + Neurovault: decoding as a service

Decoding results

Map  Plot

Feature loadings
To compare the decoded image against a term, click on an arrow below.

Show 10 entries

Search:

- auditory: 0.535
- sounds: 0.437
- listening: 0.437
- sensorimotor: 0.387
- somatosensory: 0.383
- execution: 0.346
- music: 0.346
- speech production: 0.344
- hand: 0.335

Intensity: 14.86

X: 0  Y: 0  Z: 48

Layers
- Right Button Press Auditory Cue
- anatomical
Machine learning for brain imaging
Statistical learning to study brain function

- Learning problems, but not only about prediction error

- Spatial regularization for linear models
  Total variation + randomization

- Validation is very hard
  All model are wrongs, and data is scarce

- Speed and parameter selection matter
  Users will not adapt

Positions available
[Abraham 2013] A. Abraham et al., Extracting brain regions from rest fMRI with Total-Variation constrained dictionary learning, Med Imag Comp Aided Intervention (2013)
http://hal.inria.fr/hal-00839984/en


http://hal.inria.fr/hal-00991743/en

http://hal.inria.fr/hal-00839984/en
References (not exhaustive)


  http://hal.inria.fr/inria-00563468/en

  http://hal.inria.fr/inria-00589201/en

- [Smith 2010] *Network Modelling Methods for fMRI*, NeuroImage

  http://hal.inria.fr/hal-00650905/en

[Varoquaux MICCAI 2010] Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling, Med Imag Comp Aided Intervention
http://hal.inria.fr/inria-00512417/en

http://hal.inria.fr/inria-00512451/en

[Varoquaux IPMI 2011] *Multi-subject dictionary learning to segment an atlas of brain spontaneous activity*, IPMI
http://hal.inria.fr/inria-00588898/en
http://hal.inria.fr/hal-00705192/en

[Varoquaux 2013] *Learning and comparing functional connectomes across subjects*, NeuroImage
http://hal.inria.fr/hal-00812911/en