Understanding multivariate pattern analysis for neuroimaging applications

Maria Giulia Preti
Dimitri Van De Ville

Medical Image Processing Lab,
Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne
Department of Radiology & Medical Informatics, Université de Genève

http://miplab.epfl.ch/

April 17, 2014
Overview

- Brain decoding: motivation and applications
  - Limitations of “regular SPM” (GLM)
  - “Hall of fame” of brain decoding

- Pattern recognition principles for brain decoding
  - Basics
  - Data processing pipeline for fMRI decoding
    - Preprocessing
    - Feature extraction
    - Feature selection
  - Cross-validation
    - Training phase
    - Evaluation phase
    - Testing phase
  - Classifier selection
whole-brain scan
2x2x2mm³, 20-30 slices
every 2-4 sec

100’000 intracranial voxels
1’000 timepoints
Limitations of the GLM approach

- Massively univariate approach
  - Hypothesis testing is applied voxel-wise
  - All voxels are treated independently, but are obviously not independent
- Textbook multivariate approaches are not practical
  - E.g., estimating spatial covariance matrix (100’000x100’000) by 1’000 timepoints will be rank-deficient
- fMRI data are complicated
  - Data-driven: let the data speak
- Group statistics are great...
  - ... but the ultimate goal is often to know how well the results generalizes (prediction)
  - Learn from the group, apply to the (unseen) individual
Forward versus backwards inference

stimuli
conditions
disease
disorder
...

encoding

generative model

data

stimuli
conditions
disease
disorder
...

decoding

discriminative model

data

phrenology

blind reading
Exploiting “subtle” spatial patterns

Brain activity recorded in the brain

Vectorization

Pattern vector

[a] Pattern vector

[b] Pattern vector

[c] Pattern vector

[d] Correct

[Cox et al., 2003; Haynes and Rees, 2006]
Brain decoding of visual processing

[Brain imaging studies with distributed patterns and subtle patterns in V1]

[Haxby et al., 2001] [Kamitani and Tong, 2005, 2006] [Haynes, Rees., 2006]

[Advanced models for receptive fields]

[Miyawaki et al., 2008]
Decoding video sequences

reverse statistical inference
(“decode” stimulus from data)

[Nishimoto et al, 2011]
Emotions in voice-sensitive cortices

How is emotional information treated in aud. cortex?

- 22 healthy right-handed subjects
- 10 actors pronounced “ne kalibam sout molem” (anger, sadness, neutrality, relief, joy)
  - normalization for mean sound intensity; rated by 24 (other) subjects
- Classifier approach on aud. cortex
  - localizer: human voices > animal sounds
  - 20 stimuli for each category (5); 5 classifiers (one-versus-others)

[Ethofer, VDV, Scherer, Vuilleumier, *Current Biology*, 2009; collaboration with LABNIC/CISA, UniGE]

Anger     Sadness     Neutral     Relief     Joy

z=0

z=+3

z=+6

“Ne kalibam sout molem”  
[Ethofer, VDV, Scherrer, Vuilleumier, Current Biology, 2009]
Decoding from the emotional brain

Results for bilateral voice-sensitive cortices

- solid: smoothed (10mm FWHM)
- dashed: non-smoothed
- dotted: spatial average

Chance level: 20%
Classifier: linear SVM

Evidence that auditory cues for emotional processing are spatially encoded; bilateral representation

Comprehension of emotional prosody is crucial for social functioning psychiatric disorders (schizophrenia, bipolar affective disorder, ...)

[Ethofer, VDV, Scherer, Vuilleumier, Current Biology, 2009]
Brain decoding

- Automatic classification of structural images
  - SVM applied to similarity measures (inner product)

[ Klöppel et al., 2008 ]
Feature vector is typically embedded in a vector space, where each feature is a dimension.
Classification as a regression problem

- Linear model:
  \[ y^T = X^T w_1 + w_0 + \text{noise} \]
  where the weight vector \( w = (w_1, w_0) \) characterizes the model

- Learning the model by least-squares fitting:
  \[ w = \arg\min_w \left\| y^T - X^T w_1 - w_0 \right\|^2 \]
  (notice problem when \( D > N \))

- Classify new data by comparing \( x'^T w \) against 0.5?
Basics of pattern recognition

- Linear model projects instance on line parallel to $w$
- Decision boundary is a hyperplane in the feature space for a linear model
- If feature dimensions correspond to voxels, then $w$ corresponds to a map

[adapted from Jaakkola] 15
Classification and decision theory

- Assume class-conditional densities $p(x|y)$ and class probabilities $P(y)$ known

- Minimizing the overall probability of error corresponds to

$$y' = \arg \max_y P(y|x')$$

$$= \arg \max_y \frac{p(x'|y)P(y)}{p(x')}$$  \hspace{1cm} \text{(Bayes’ rule)}

$$= \arg \max_y p(x'|y)P(y)$$

\text{likelihood}
Wonderful link between discriminative model (posterior) and generative model (likelihood)

\[ P(y|x) \propto p(x|y)P(y) \]

- **Discriminative model; e.g., logistic regression**

\[
\log \left( \frac{P(y = 1|x)}{P(y = 0|x)} \right) = \mathbf{w}_1^T \mathbf{x} + w_0 \text{ leads to } P(y = 1|x) = g(\mathbf{w}_1^T \mathbf{x} + w_0)
\]

- **Generative model; e.g., naïve Bayes**

\[
p(x|y) = \prod_{i=1}^{D} p(x_i|y)
\]

where the features are modeled as independent random variables
Basics of pattern recognition

- Model order selection

$$-2 \log p(x|\theta_K, y) + \text{penalty}(\theta_K)$$

- Model $\theta_K$ with $K$ parameters
- Occam’s razor: make model as simple as possible
- Akaike information criterion: penalty $= 2K$
- Bayesian information criterion: penalty $= K \log D$
FMRI processing pipeline

- From fMRI raw data to decisions

[Diagram showing the FMRI processing pipeline]

- Acquire signal
- Pre-process signal
- Extract features
- Select features
- Train classifier
- Classify

- Motion correction
- Coregistration
- Normalisation
- Parcellation

- Univariate
- Multivariate

- D
- D' < D

- BOLD response amplitude
- Time

- Mean per voxel
- Mean per region

- Activity
- Connectivity

- Classifier θ
- Decision Ω
Due to large inter-subject variability (anatomical and functional) and artifacts, (f)MRI data must be preprocessed before feature extraction.

Possible preprocessing steps:
- Slice-timing correction
- Susceptibility correction
- Motion correction
- Detrending (scanner drift)
- Physiological artifacts removal
- Normalisation to common template (e.g., MNI)
  - Exploit structural MRI data (T1)
Feature extraction

* Three key principles
  - Use engineering know-how (e.g., try to increase SNR)
  - Use domain knowledge (e.g., look at the right regions)
  - Aim at interpretability, depending on the task
    - Intra- vs intersubject decoding

* Spatial dimensions
  - Gray-matter segmentation (10’000 voxels)
  - Atlasing (100-1’000 regions)
  - Interhemispherical differences
  - Searchlight approach [Kriegeskorte et al., 2006]
  - Transform domain (e.g., PCA/ICA, wavelets)

* Time dimension
  - Exploit task paradigm
  - Incorporate knowledge about hemodynamic response
  - Temporal compression
    - Within and between timecourses
Feature extraction (temporal compression)

- **local features**: one vector per scan (e.g., intensity or derivative)
- **segmental features**: computed over a number of time points (e.g., sliding-window measures)
- **global features**: computed over whole timecourse (e.g., mean, standard deviation)
- **statistical dependencies**: between timecourses (e.g., functional connectivity based on Pearson correlation)
Feature selection

- Establish a subset of most **discriminative** features
  - Reduce dimensionality of the feature space to D’

- Feature selection is a **combinatorial search problem** for the best feature subset based on a **search strategy** and an **objective function**
  - Search strategies can be univariate (select one voxel at a time) or multivariate
    - optimal (exhaustive), suboptimal; best-N, genetic algorithms, backward/forward recursive selection...
  
- Objective functions are of three kinds:
  - **filter**: the objective function is not a discriminant function
  - **wrapper**: the objective function is the error rate of the final classifier
  - **embedded**: the objective function is directly based on the objective function of the final classifier
Simple method: point-biserial correlation

- For each of the $D$ features, measure point-biserial correlation:
  \[
  r_{pb} = \frac{\bar{X}_\omega_1 - \bar{X}_\omega_0}{\sigma_X} \sqrt{\frac{N_0 N_1}{N^2}}
  \]

- Then generate a ranking according to $|r_{pb}|$ and keep only top $D'$ features
Model learning

- **Training** set with known labels (supervised learning)
  - Model and its parameters depends on type of classifier

- Discriminative models
  - Logistic regression
  - Fisher discriminant analysis
  - Support vector machines
  - Random forest
  - Neural networks

- Generative models
  - Naïve Bayes, can assume different types of densities
  - Gaussian mixture model
  - Hidden Markov model
Classifier prediction can be correct or wrong

Accuracy statistics can be shown in a confusion matrix (summary table):

- Class 0 accuracy = $A/(A+B)$
- Class 1 accuracy = $D/(C+D)$
- Accuracy = $(A+D)/(A+B+C+D)$

Biostats ($\omega_0 =$ healthy, $\omega_1 =$ disease):

- Sensitivity (“Disease is PRESENT and I report disease”): $D/(C+D) = TP/(FN+TP)$
- Specificity: (“Disease is ABSENT and I report no disease”: $A/(A+B) = TN/(FP+TN)$

Perfect: $B=C=0$. **Be suspicious if this happens!**

Random: $A=B=C=D$. Same as flipping a coin
Testing classifiers: the ROC curve

- Increasing sensitivity can never increase specificity at the same time, and vice-versa

- Receiver Operating Characteristic (ROC) curve
  - Allow to see compromise over the operating range of a classifier
  - Can be used for classifier comparison e.g., compute the Area Under Curve (AUC) as a summary measure of performance
Proper tuning and error estimation

- Maintain strict separation throughout
  - Remember our goal: classify **unseen** data
  - This applies at all stages of the processing chain
  - Always ask yourself: **what would I do if I acquired data for one more subject after my classifier is trained?**
  - If not, main cause of over-optimistic results (and even useless in some cases)
Classifier selection

- How to choose a classifier?
  - **No Free Lunch Theorem**
    - You cannot get learning for free, every classifier has its biases
  - Problem characteristics (data structure, dimensions, ...)
  - Prior experience and personal preference
  - Cross-validation + statistical testing

- Model complexity and optimal parameters?
  - **Occam’s razor**
    - As simple as possible
  - Information-theoretic criteria
    - AIC, BIC/MDL, ... 
  - Empirical approaches
    - Cross-validation
    - Surrogate data

[Hasbie et al., 2009]
Challenges and opportunities

- High-dimensional learning problem
  - Large number of features compared to instances
  - Kernel trick
    - Basically, look at $X^T X$ instead of $XX^T$
    - “Distance” between instances depends on definition of kernel function
  - Interpretation
    - How to get “brain maps” (e.g., showing where information is processed)
    - Integration versus segregation
  - Estimate where you are in the bias-variance trade-off

- Multimodal approaches
  - Combine different classifiers (“ensemble learning”)

- Intra-subject decoding
  - Fine-grained organization of brain activation patterns (“hyperacuity”)

- Inter-subject decoding
  - Multicentric datasets
  - Diagnosis and prognosis
    - Prediction at the individual patient level
    - Beyond patient-or-not, predict clinical measures ~ surrogate functional markers
Recommended reading

- Haufe et al., *On the interpretation of weight vectors of linear models in multivariate neuroimaging*, NeuroImage, in press
**Classifier selection**

- **Is classifier A better than classifier B?**
- **Hypothesis testing framework**
  - “Do the two sets of decisions of classifiers A and B represent two different populations?”

<table>
<thead>
<tr>
<th>test sample</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✘</td>
<td>✘</td>
</tr>
<tr>
<td>B</td>
<td>✔</td>
<td>✔</td>
<td>✘</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✘</td>
<td>✔</td>
<td>✔</td>
<td>✘</td>
</tr>
</tbody>
</table>

- Careful: Samples (sets of decisions) are **dependent** (same evaluation/testing dataset), and measurement type is **categorical binary** (for a two-class problem)

⇒ Use **McNemar test**

H₀: “there is no difference between the accuracies of the two classifiers”
Selection with the McNemar test

- Compute relationships between classifier decisions

<table>
<thead>
<tr>
<th></th>
<th>B ✔</th>
<th>B ✘</th>
</tr>
</thead>
<tbody>
<tr>
<td>A ✔</td>
<td>N_{11}</td>
<td>N_{10}</td>
</tr>
<tr>
<td>A ✘</td>
<td>N_{01}</td>
<td>N_{00}</td>
</tr>
</tbody>
</table>

- If H₀ is true, we should have N₀₁ = N₁₀ = 0.5(N₀₁ + N₁₀)

- We can measure the following test-static based on the observed counts:

\[ x^2 = \frac{(|N_{01} - N_{10}| - 1)^2}{N_{01} + N_{10}} \]

- Then, compare \(x^2\) to critical value of \(\chi^2\) (1 DOF) at a level of significance \(\alpha\)
  - (requires \(N_{01} + N_{10} > 25\))

After [Dietterich, 1998]