Multivariate analyses and decoding

Kay H. Brodersen
Computational Neuroeconomics Group
University of Zurich

Machine Learning and Pattern Recognition Group
ETH Zurich
1 Introduction
Why multivariate?

Haxby et al. 2001 Science
Why multivariate?

- Multivariate approaches can reveal information jointly encoded by several voxels.

Kriegeskorte et al. 2007 NeuroImage
Multivariate approaches can exploit a sampling bias in voxelized images.

Boynton 2005 Nature Neuroscience
Mass-univariate approaches treat each voxel independently of all other voxels such that the implicit likelihood factorises over voxels:

\[ p(Y \mid X, \theta) = \prod_i p(Y_i \mid X, \theta_i) \]

Spatial dependencies between voxels are introduced after estimation, during inference, through random field theory. This allows us to make multivariate inferences over voxels (i.e., cluster-level or set-level inference).

Multivariate approaches, by contrast, relax the assumption about independence and enable inference about distributed responses without requiring focal activations or certain topological response features. They can therefore be more powerful than mass-univariate analyses.

The key challenge for all multivariate approaches is the high dimensionality of multivariate brain data.
0 Prediction or inference?

- The goal of **prediction** is to maximize the accuracy with which brain states can be decoded from fMRI data.
- The goal of **inference** is to decide between competing hypotheses about structure-function mappings in the brain. Typically: compare a model that links distributed neuronal activity to a cognitive state with a model that does not.

1 Encoding or decoding?

2 Univoxel or multivoxel?

3 Classification or regression?
**Models & terminology**

1. **Encoding or decoding?**
   - An **encoding** model (or generative model) relates context (independent variable) to brain activity (dependent variable).
   - A **decoding** model (or recognition model) relates brain activity (independent variable) to context (dependent variable).

   \[ g : X \to Y \]
   \[ h : Y \to X \]

2. **Univoxel or multivoxel?**
   - In a **univoxel** model, brain activity is the signal measured in one voxel. (Special case: mass-univariate.)
   - In a **multivoxel** model, brain activity is the signal measured in many voxels.

   \[ Y \in \mathbb{R} \]
   \[ Y \in \mathbb{R}^n, \ n \gg v \]

3. **Regression or classification?**
   - In a **regression** model, the dependent variable is continuous.
   - In a **classification** model, the dependent variable is categorical (typically binary).

   e.g., \[ Y \in \mathbb{R}^n \] or \[ X \in \mathbb{R} \]
   e.g., \[ X \in \{-1, +1\} \]
2 Classification
Classification

1. Feature extraction
   - fMRI timeseries

2. Feature selection
   - e.g., voxels

3. Classification
   - Accuracy estimate [% correct]

Trials

Voxels

Training examples

Test examples

A A B A B A A B A A A B A
Most classification algorithms are based on a **linear** model that discriminates the two classes.

If the data are not linearly separable, a **nonlinear** classifier may still be able to tell different classes apart.

here: discriminative point classifiers
We need to train and test our classifier on separate datasets. Why?

- Using the same examples for training and testing means overfitting may remain unnoticed, implying an optimistic accuracy estimate.
- Instead, what we are interested in is generalizability: the ability of our algorithm to correctly classify previously unseen examples.

An efficient splitting procedure is cross-validation.
Target questions for decoding studies

(a) Pattern discrimination (overall classification)

Accuracy [%]

100 %

50 %

Classification task

Left or right button?

Truth or lie?

Healthy or diseased?

(b) Spatial pattern localization

Accuracy rises above chance

Participant indicates decision

Inferring a representational space and extrapolation to novel classes

(c) Temporal pattern localization

Accuracy [%]

100 %

50 %

Intra-trial time

Accuracy rises above chance

Participant indicates decision

Inferring a representational space and extrapolation to novel classes

(d) Pattern characterization

Inferring a representational space and extrapolation to novel classes

Brodersen et al. 2009 The New Collection

Mitchell et al. 2008 Science
(a) Overall classification

Overall classification is about achieving maximal prediction performance.

Performance evaluation – example

- Given 100 trials, leave-10-out cross-validation, we measure performance by counting the number of correct predictions on each fold:

```
6 5 7 8 4 9 6 7 7 5 ...
```

... out of 10 test examples correct

- How likely is it be to get 64 out of 100 correct we had been guessing?

\[
p = P(N_{\text{correct}} \geq 64) = 1 - \sum_{k=1}^{64-1} \binom{100}{i} \times 0.5^k \times 0.5^{100-k}
\]

\[
= 0.00176
\]

- Thus, we have made a Binomial assumption about the Null model to show that our result is statistically significant at the 0.05 level.
The support vector machine

- Intuitively, the support vector machine finds a hyperplane that maximizes the margin between the plane and the nearest examples on either side.
- For nonlinear mappings, the kernel converts a low-dimensional nonlinear problem into a high-dimensional linear problem.

![Diagram of support vector machine](chart.png)
Deconvolved BOLD signal

→ result: one beta value per trial, phase, and voxel
(b) Spatial information mapping

**METHOD 1** Consider the entire brain, and find out which voxels are jointly discriminative

- e.g., based on a classifier with a constraint on sparseness in features

**METHOD 2** At each voxel, consider a small local environment, and compute a distance score

- e.g., based on a CCA

- e.g., based on a classifier

- e.g., based on Euclidean distances

- e.g., based on Mahalanobis distances
  - Kriegeskorte et al. 2006, 2007a, 2007b
  - Serences & Boynton 2007 *J Neuroscience*

- e.g., based on the mutual information
(b) Spatial information mapping

**Example 1** – decoding whether a subject will switch or stay

**Example 2** – decoding which option was chosen

Hampton & O’Doherty 2007 *PNAS*

Brodersen et al. 2009 *HBM*
(c) Temporal information mapping

Example – decoding which button was pressed

Soon et al. 2008 Nature Neuroscience
(c) Pattern characterization

**Example** – decoding which vowel a subject heard, and which speaker had uttered it.

Fingerprint plot (one plot per class)

Formisano et al. 2008 *Science*
Limitations

- **Constraints on experimental design**
  - When estimating trial-wise Beta values, we need longer ITIs (typically 8 – 15 s).
  - At the same time, we need many trials (typically 100+).
  - Classes should be balanced.

- **Computationally expensive**
  - e.g., fold-wise feature selection
  - e.g., permutation testing

- **Classification accuracy is a surrogate statistic**

- **Classification algorithms involve many heuristics**
3 Multivariate Bayesian decoding
Multivariate Bayesian decoding (MVB)

- Multivariate analyses in SPM are not implemented in terms of the classification schemes outlined in the previous section.
- Instead, SPM brings classification into the conventional inference framework of hierarchical models and their inversion.

- MVB can be used to address two questions:
  - **Overall classification** – using a cross-validation scheme (as seen earlier)
  - **Inference on different forms of structure-function mappings** – e.g., smooth or sparse coding (new)
Model

Encoding models
$X$ as a cause

$$X \rightarrow A = X\beta$$

$$Y = TA + G\gamma + \varepsilon$$

$g(\theta): X \rightarrow Y$

Decoding models
$X$ as a consequence

$$A \rightarrow X = A\beta$$

$$Y = TA + G\gamma + \varepsilon$$

$g(\theta): Y \rightarrow X$

$$TX = Y\beta - G\gamma\beta - \varepsilon\beta$$
Decoding models are typically ill-posed: there is an infinite number of equally likely solutions. We therefore require constraints or priors to estimate the voxel weights $\beta$.

SPM specifies several alternative coding hypotheses in terms of empirical spatial priors on voxel weights.

$$\text{cov}(\beta) = U\Sigma\eta U^T$$

- Null: $U = \emptyset$
- Spatial vectors: $U = I$
- Smooth vectors: $U(\tilde{x}_i, \tilde{x}_j) = \exp(-\frac{1}{2}(\tilde{x}_i - \tilde{x}_j)^2\sigma^{-2})$
- Singular vectors: $UDV^T = RY^T$
- Support vectors: $U = RY^T$

Friston et al. 2008 *NeuroImage*
MVB – example

- MVB can be illustrated using SPM’s attention-to-motion example dataset.
  Buechel & Friston 1999 *Cerebral Cortex*
  Friston et al. 2008 *NeuroImage*

- This dataset is based on a simple block design. Each block is a combination of some of the following three factors:
  - photic – there is some visual stimulus
  - motion – there is motion
  - attention – subjects are paying attention

- We form a design matrix by convolving box-car functions with a canonical haemodynamic response function.
MVB – example
MVB – example

- MVB-based predictions closely match the observed responses. But crucially, they don’t perfectly match them. Perfect match would indicate overfitting.
The highest model evidence is achieved by a model that recruits 4 partitions. The weights attributed to each voxel in the sphere are sparse and multimodal. This suggests sparse coding.
4 Further model-based approaches
Challenges for all decoding approaches

- **Challenge 1** – feature selection and weighting to make the ill-posed many-to-one mapping tractable

- **Challenge 2** – neurobiological interpretability of models to improve the usefulness of insights that can be gained from multivariate analysis results
Further model-based approaches (1)

- Approach 1 – identification (inferring a representational space)
  1. estimation of an encoding model
  2. nearest-neighbour classification or voting

Mitchell et al. 2008 Science
Approach 2 – reconstruction / optimal decoding

1. estimation of an encoding model
2. model inversion

Paninski et al. 2007 *Progr Brain Res*

Pillow et al. 2008 *Nature*

Miyawaki et al. 2009 *Neuron*
Further model-based approaches (3)

- Approach 3 – decoding with model-based feature construction

Brodersen et al. 2009 (under review)
Multivariate analyses can make use of information jointly encoded by several voxels and may therefore offer higher sensitivity than mass-univariate analyses.

There is some confusion about terminology in current publications. Remember the distinction between prediction vs. inference, encoding vs. decoding, univoxel vs. multivoxel, and classification vs. regression.

The main target questions in classification studies are (i) pattern discrimination, (ii) spatial information mapping, (iii) temporal information mapping, and (iv) pattern characterization.

Multivariate Bayes offers an alternative scheme that maps multivariate patterns of activity onto brain states within the conventional statistical framework.

The future is likely to see more model-based approaches.
5 Supplementary slides
The most common multivariate analysis is classification

- **Classification** is the most common type of multivariate fMRI analysis to date. By classification we mean: to decode a categorical label from multivoxel activity.

- Lautrup et al. (1994) reported the first classification scheme for functional neuroimaging data.

- Classification was then reintroduced by Haxby et al. (2001). In their study, the overall spatial pattern of activity was found to be more informative in distinguishing object categories than any brain region on its own.

Haxby et al. 2001 *Science*
Temporal unit of classification

- The temporal unit of classification specifies the amount of data that forms an individual example. Typical units are:
  - one trial → trial-by-trial classification
  - one block → block-by-block classification
  - one subject → across-subjects classification

- Choosing a temporal unit of classification reveals a trade-off:
  - smaller units mean noisier examples but a larger training set
  - larger units mean cleaner examples but a smaller training set

- The most common temporal unit of classification is an individual trial.
Temporal unit of classification

Brodersen, Hunt, Walton, Rushworth, Behrens 2009 HBM
Alternative temporal feature extraction

Interpolated raw BOLD signal

signal (a.u.)

averaged signal across all trials
subject 1
subject 2
subject 3

decision
delay
result

→ result: any desired number of sampling points per trial and voxel
Alternative temporal feature extraction

Deconvolved BOLD signal, expressed in terms of 3 basis functions

- **Step 1:** sample many HRFs from given parameter intervals
- **Step 2:** find set of 3 orthogonal basis functions that can be used to approximate the sampled functions

→ result: three values per trial, phase, and voxel

![HRF Samples](image1)

![HRF Basis Functions](image2)
Classification of methods for feature selection

- A priori structural feature selection
- A priori functional feature selection

- Fold-wise univariate feature selection
  - Scoring
  - Choosing a number of features

- Fold-wise multivariate feature selection
  - Filtering methods
  - Wrapper methods
  - Embedded methods

- Fold-wise hybrid feature selection
  - Searchlight feature selection
  - Recursive feature elimination
  - Sparse logistic regression

- Unsupervised feature-space compression
Training and testing a classifier

- **Training phase**
  - The classifier is given a set of \( n \) labelled training samples
    \[
    S_{\text{train}} = \{(x_1, y_1), \ldots, (x_n, y_n)\}
    \]

  from some data space \( X^d \times \{-1, 1\} \), where
  - \( x_i = (x_1, \ldots, x_d) \) is a \( d \)-dimensional attribute vector
  - \( y_i \in \{-1, 1\} \) is its corresponding class.

  - The goal of the learning algorithm is to find a function that adequately describes the underlying attributes/class relation.

  - For example, a *linear* learning machine finds a function \( f_{w,b}(x) = \langle w \cdot x \rangle + b \)
    which assigns a given point \( x \) to the class
    \[
    \hat{y} = \text{sgn}(f_{w,b}(x))
    \]

    such that some performance measure is maximized, for example:
    \[
    (w, b) = \arg \max_{w,b} \sum_{i=1}^{n} y_i \hat{y}_i
    \]
Test phase

The classifier is now confronted with a test set of unlabelled examples

\[ S_{test} = \{x_1, \ldots, x_k\} \]

and assigns each example \( x \) to an estimated class

\[ \hat{y} = \text{sgn}(f_{w,b}(x)) \]

We could then measure generalization performance in terms of the relative number of correctly classified test examples:

\[ \text{acc} = \frac{\sum_{i=1}^{k} 1_{\hat{y}_i = y_i}}{k} \]
Nonlinear prediction problems can be turned into linear problems by using a nonlinear projection of the data onto a high-dimensional feature space.

This technique is used by a class of prediction algorithms called **kernel machines**.

The most popular kernel method is the **support vector machine** (SVM).

- SVMs make training and testing computationally efficient.

  \[
  \begin{align*}
  & \min_{w,b} \langle w, w \rangle + C \sum_{i=1}^{n} \xi_i \\
  & \text{s.t. } \xi_i \geq 1 - y_i (\langle w, x_i \rangle + b) \quad \forall i = 1, ..., n \\
  & \xi_i \geq 0,
  \end{align*}
  \]

- We can easily reconstruct feature weights:

  \[
  w = \sum_{i=1}^{n} y_i \alpha_i x_i
  \]

- However, SVM predictions do not have a probabilistic interpretation.
Multivariate Bayes – maximization of the model evidence

\[ Q_{i+1} = L^{(i)} L^T \]
\[ I^{(m+1)} = I^{(m)} \land \mu^\eta \supseteq \mu^{(m)} \]

**M-step**

\[ L_{ij} = -\frac{1}{2} \text{tr}(P_j (WXX^T W^T - \Sigma)) - \prod_i (\mu^\lambda_i - \pi_i) \]
\[ L_{iij} = -\frac{1}{2} \text{tr}(P \Sigma P_j \Sigma) - \prod_i \]

\[ \Delta \mu^\lambda = -L_{\alpha \lambda}^{-1} L_{\lambda} \]
\[ \Sigma^\lambda = -L_{\alpha \lambda}^{-1} \]

**M-step**

 Until convergence

**E-step**

\[ q(\lambda) = N(\mu^\lambda, \Sigma^\lambda) \]

\[ \Sigma^\eta = \exp(\mu^\lambda) I^{(1)} + \exp(\mu^\lambda) I^{(2)} + K \]

\[ \mu^\eta = \Sigma^\eta L \Sigma^{-1} WX \]
\[ \mu^\beta = U \mu^\eta \]
\[ \Sigma^\beta = U (\Sigma^\eta - \Sigma^\eta L \Sigma^{-1} L \Sigma^\eta) U^T \]

**E-step**

\[ q(\beta) = N(\mu^\beta, \Sigma^\beta) \]

\[ \ln p(X \mid \theta, Y) \geq F = -\frac{1}{2} (X^T W^T \Sigma (\mu^\lambda)^{-1} W X) - \ln |\Sigma (\mu^\lambda)| - w \ln 2\pi + \ln |\Pi \Sigma^\lambda| - (\mu^\lambda - \pi)^T \Pi (\mu^\lambda - \pi) \]
MVB can be illustrated using SPM’s attention-to-motion example dataset.
Buechel & Friston 1999 Cerebral Cortex
Friston et al. 2008 Neuroimage

This dataset is based on a simple block design. Each block belongs to one of the following conditions:

- fixation – subjects see a fixation cross
- static – subjects see stationary dots
- no attention – subjects see moving dots
- attention – subjects monitor moving dots for changes in velocity

We wish to decode whether or not subjects were exposed to motion. We begin by recombining the conditions into three orthogonal conditions:

- photic – there is some form of visual stimulus
- motion – there is motion
- attention – subjects are required to pay attention
Further model-based approaches

- Approach 1 – identification (inferring a representational space)

Kay et al. 2008 *Science*
Further reading

On classification


On multivariate Bayesian decoding