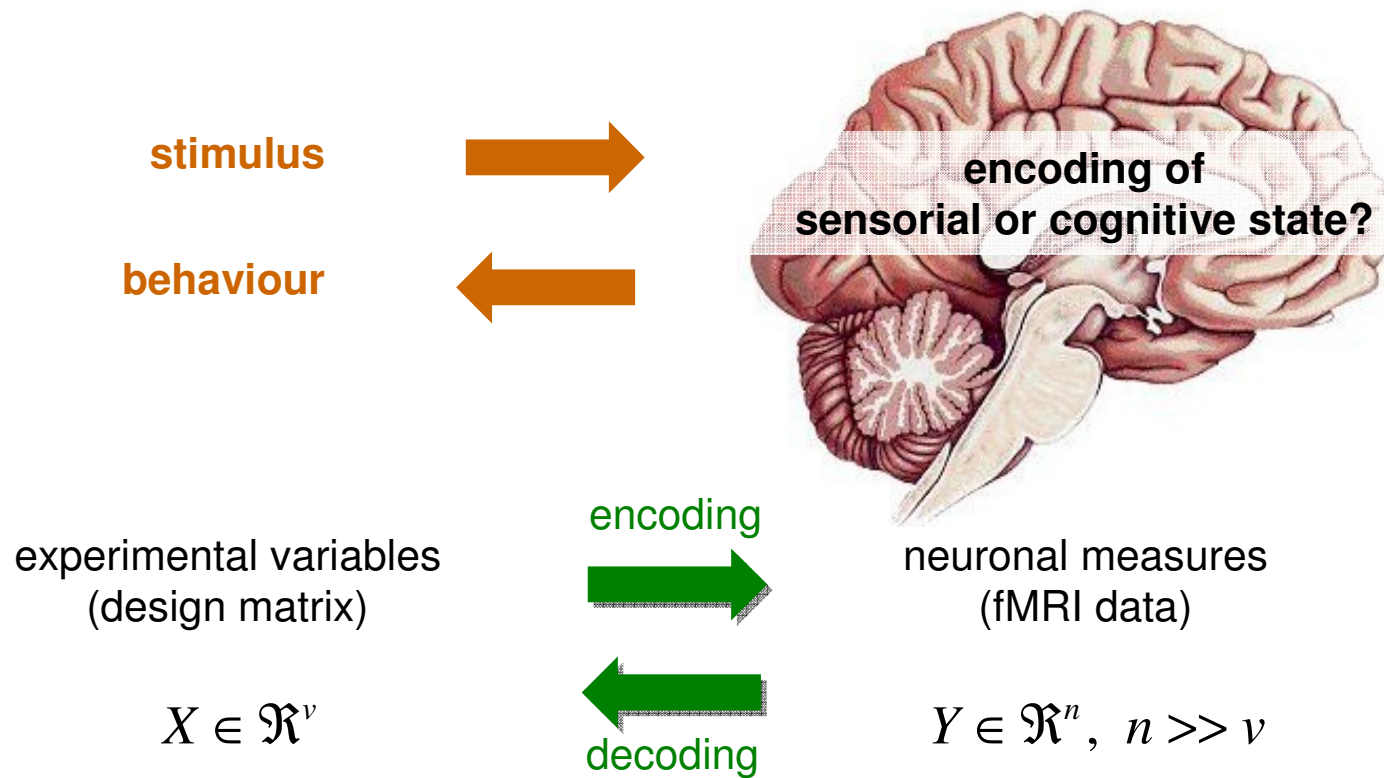


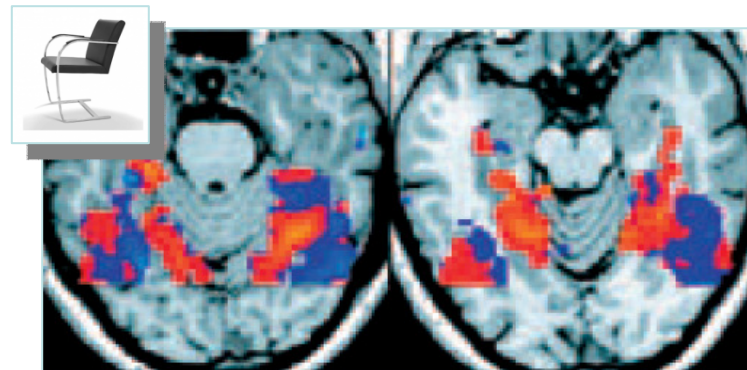
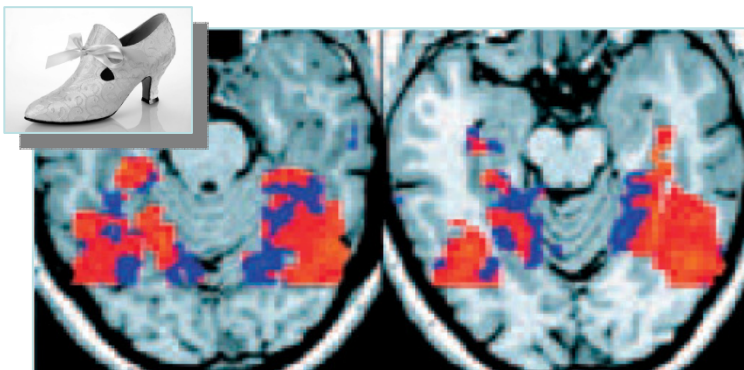
# MultiVariate Bayesian (MVB) decoding of brain images

**J. Daunizeau**

*Institute of Empirical Research in Economics, Zurich, Switzerland  
Brain and Spine Institute, Paris, France*



What if neuronal responses are distributed (over space)?



# Overview of the talk

## 1 Introduction

*1.1 Lexicon*

*1.2 “Decoding”: so what?*

*1.3 Multivariate: so what?*

*1.4 Preliminary statistical considerations*

## 2 Multivariate Bayesian decoding

*2.1 From classical encoding to Bayesian decoding*

*2.2 Hierarchical priors on patterns*

*2.3 Probabilistic inference*

## 3 Example

## 4 Summary

# Overview of the talk

## 1 Introduction

*1.1 Lexicon*

*1.2 “Decoding”: so what?*

*1.3 Multivariate: so what?*

*1.4 Preliminary statistical considerations*

## 2 Multivariate Bayesian decoding

*2.1 From classical encoding to Bayesian decoding*

*2.2 Hierarchical priors on patterns*

*2.3 Probabilistic inference*

## 3 Example

## 4 Summary

# Lexicon

the jargon to swallow

## 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).

$$X \rightarrow Y$$

$$Y \rightarrow X$$

## 2 Univariate or multivariate?

- In a **univariate** model, brain activity is the signal measured in one voxel.
- In a **multivariate** model, brain activity is the signal measured in many voxels (NB: *decoding*  $\rightarrow$  *ill-posed problem*).

$$Y \in \mathfrak{R}$$

$$Y \in \mathfrak{R}^n, \quad n \gg v$$

## 3 Regression or classification?

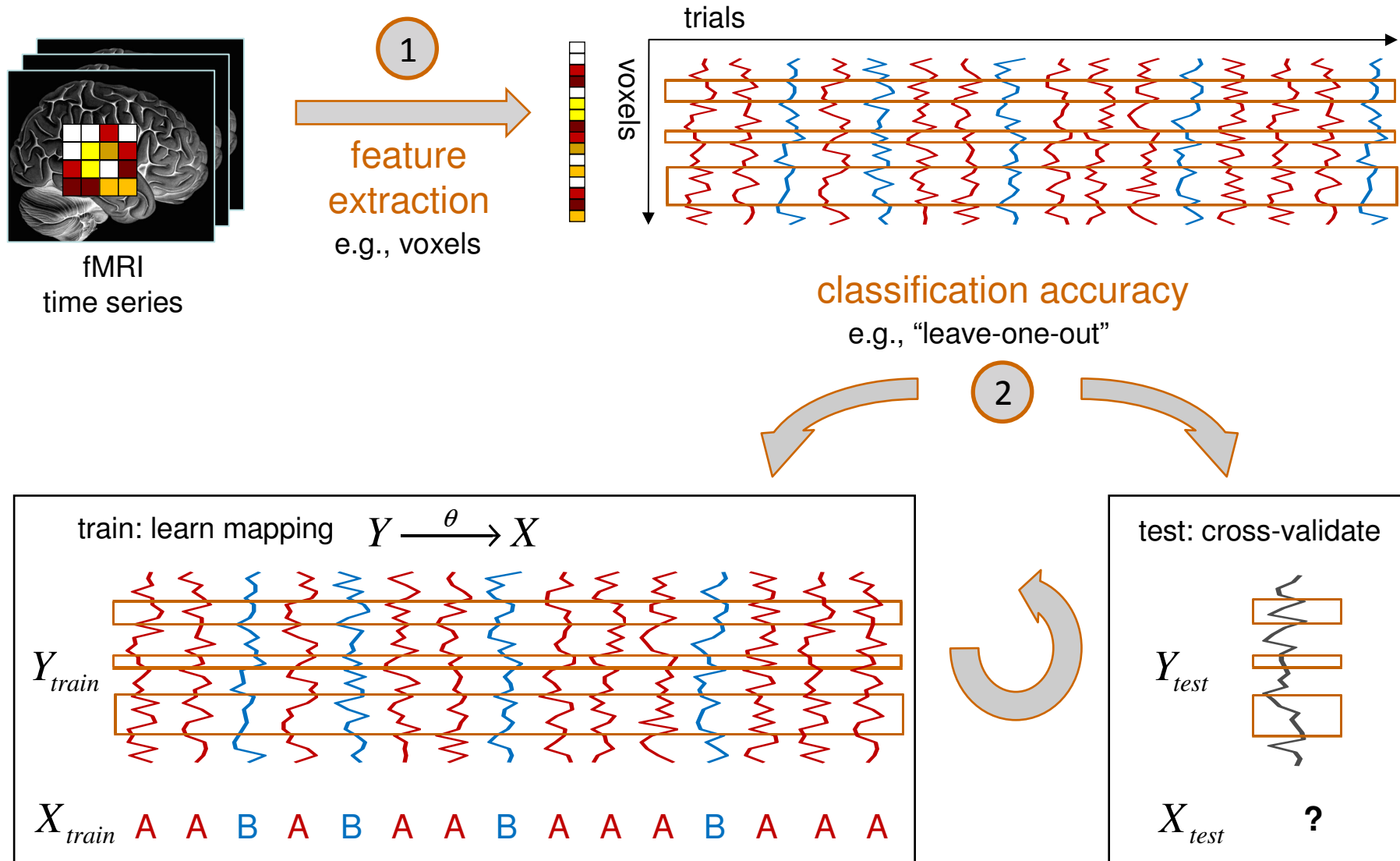
- In a **regression** model, the dependent variable is continuous.
- In a **classification** model, the dependent variable is categorical (typically binary).

$$X \in \mathfrak{R} \quad \text{or} \quad Y \in \mathfrak{R}^n$$

$$X \in \{-1, +1\}$$

# “Decoding”: so what?

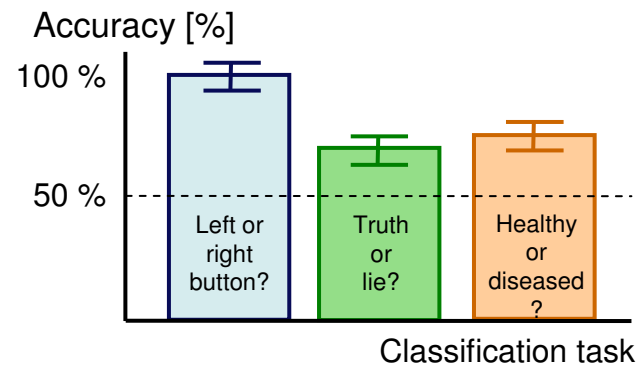
The seminal approach: classification



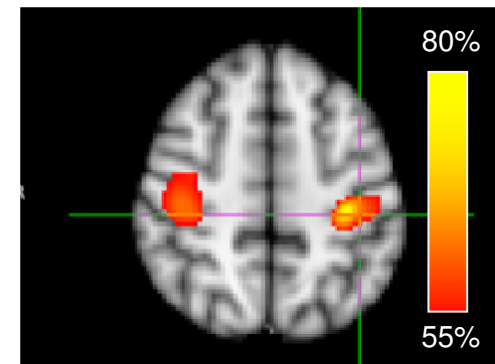
# “Decoding”: so what?

Reversing the X-Y mapping: target questions

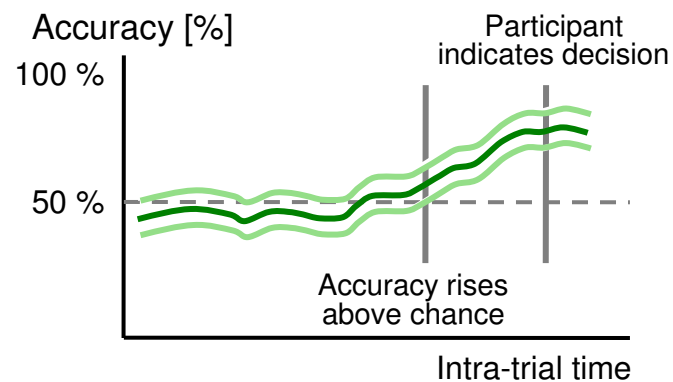
## (1) X-Y mapping overall reliability



## (2) X-Y mapping spatial deployment

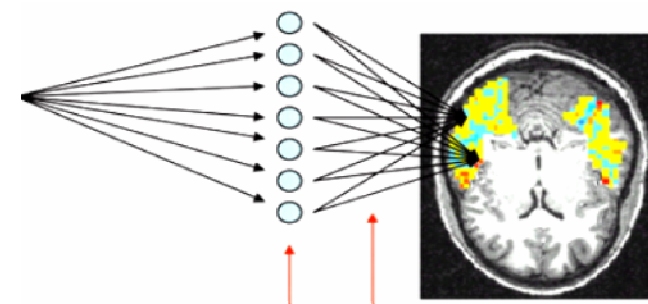


## (3) X-Y mapping temporal evolution



## (4) X-Y mapping: subtle issues

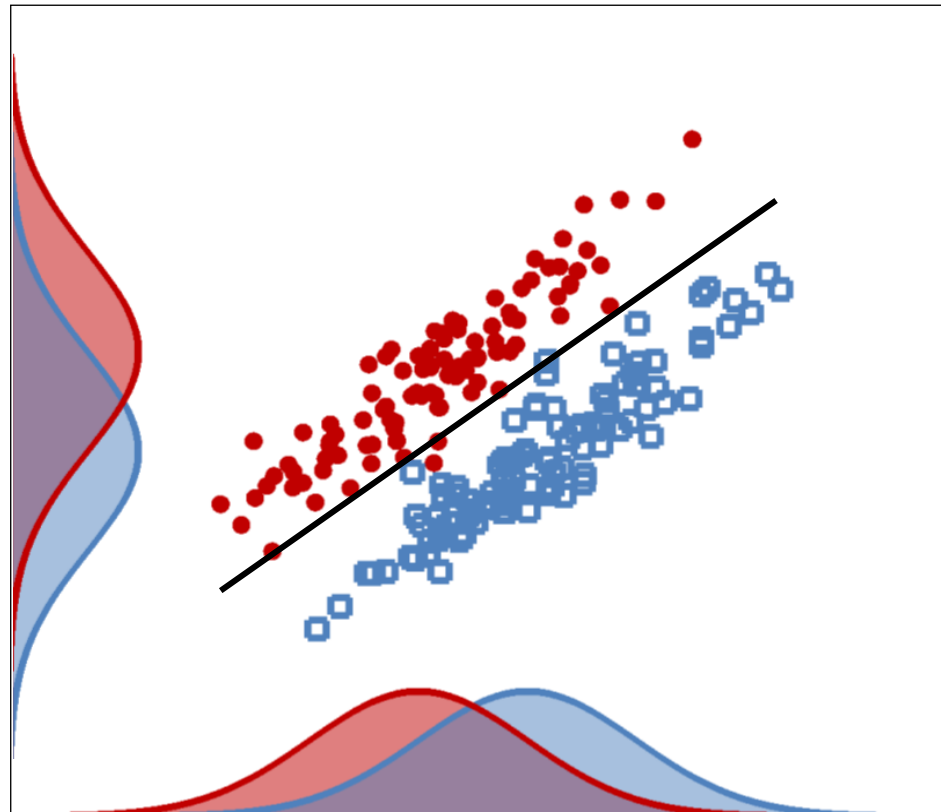
- functionally selective vs segregated representations
- degenerative (many-to-one) structure-function mappings



# Multivariate: so what?

Well, we might need it.

- Multivariate approaches can reveal information jointly encoded by several voxels. This is because the (multivariate) distance between two categories accounts for correlations among these.

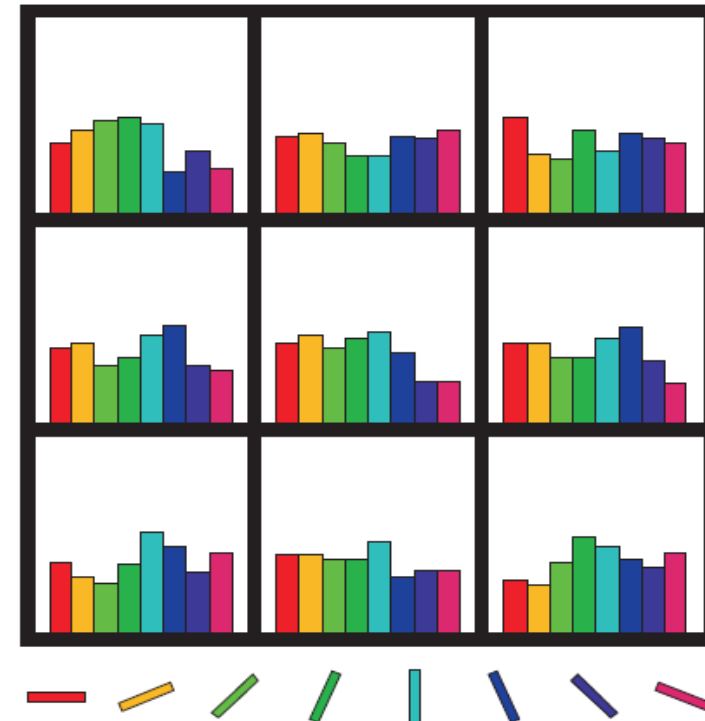
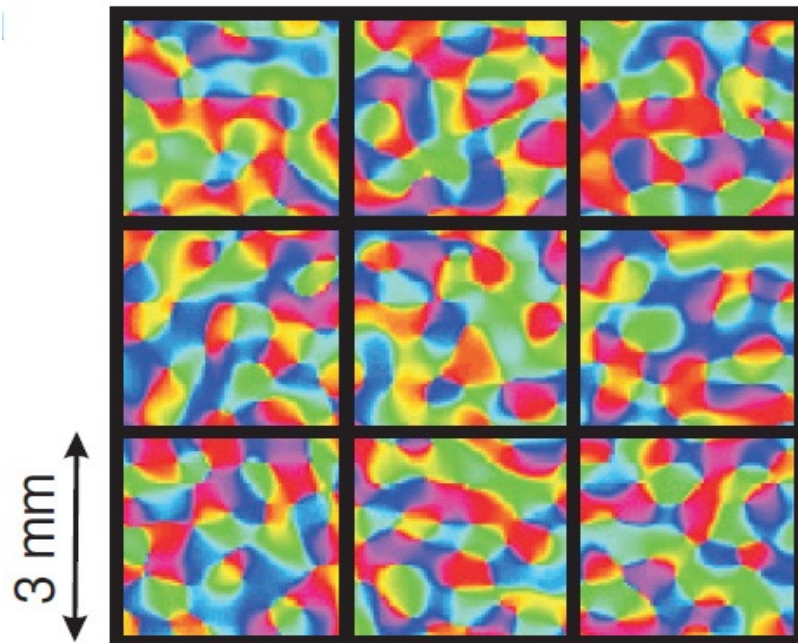




# Multivariate: so what?

Why we might need it: subvoxel processing.

- Multivariate approaches can exploit a sampling bias in voxelized images. Such subvoxel processing is unlikely to be detected by univariate methods.



Boynton 2005 *Nature Neuroscience*

# Preliminary statistical considerations

## lessons from the Neyman-Pearson lemma

- Do neuronal responses encode some sensorial or cognitive state of the subject?
- Null assumption: there is no dependency between Y and X

$$H_0 : p(Y|X) = p(Y)$$

- **Neyman-Pearson lemma**: the likelihood ratio (or Bayes factor)

$$\Lambda = \frac{p(Y|X)}{p(Y)} = \frac{p(X|Y)}{p(X)} \geq u$$

is the most powerful test of size  $\alpha = p(\Lambda \geq u | H_0)$  to test the null.

- So what? Well...

- 1 All we have to do is comparing a model that links Y to X with a model that does not.
- 2 The link can be from X to Y or from Y to X. From the point of view of inferring a link exists, its direction is not important (but...).

# Preliminary statistical considerations

## prediction and inference

- Some confusion about the roles of prediction and inference may arise from the use of classification accuracy to infer a significant relationship between  $X$  and  $Y$ .

- This is because « cross-validation » relies on the predictive density:

$$p(X_{new} | Y_{new}, X, Y) = \int p(X_{new} | Y_{new}, \theta) p(\theta | X, Y) d\theta$$

where  $\theta$  are unknown parameters of the mapping  $Y \xrightarrow{\theta} X$   
to check the « generalization error » of the inferred mapping.

- Note:

- 1 The only situation that legitimately requires us to predict a new target is when we do not know it, e.g.:
  - brain-computer interface
  - automated diagnostic classification
- 2 When used in the context of experimental neuroscience, standard classifiers provide suboptimal inference on the mapping  $Y \rightarrow X$

# Overview of the talk

## 1 Introduction

*1.1 Lexicon*

*1.2 “Decoding”: so what?*

*1.3 Multivariate: so what?*

*1.4 Preliminary statistical considerations*

## 2 Multivariate Bayesian decoding

*2.1 From classical encoding to Bayesian decoding*

*2.2 Hierarchical priors on patterns*

*2.3 Probabilistic inference*

## 3 Example

## 4 Summary

# From classical encoding to Bayesian decoding

MVB: inferring on the multivariate X-Y mapping

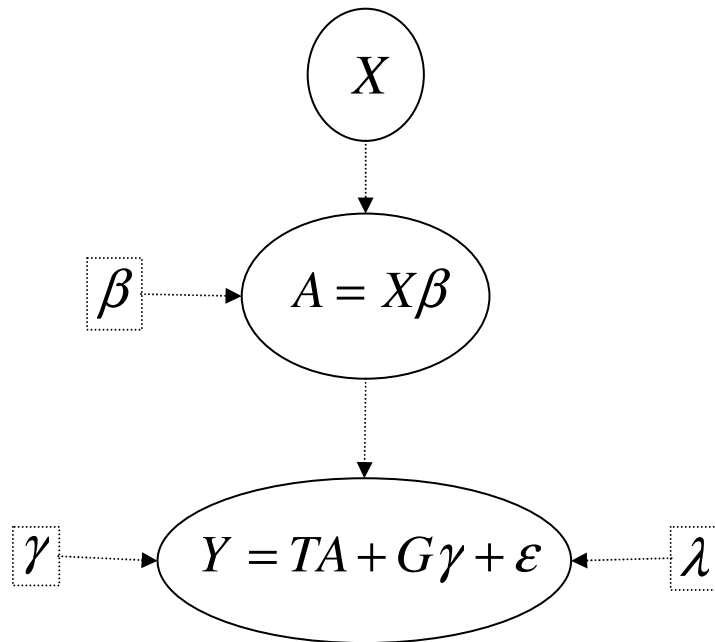
- Multivariate analyses in SPM are not implemented in terms of the classification schemes outlined in the previous section.
- Instead, SPM brings decoding into the conventional inference framework of hierarchical models and their inversion (c.f. Neyman-Pearson lemma).
- MVB can be used to address two questions:
  - **Overall significance of the X-Y mapping** (as with classical SPM or classifiers)  
... using probabilistic inference (model comparison, cross-validation)
  - **Inference on the form of the X-Y mapping** (no other alternative)
    - 1 Identify the spatial structure of the X-Y mapping (smooth, sparse, etc...)
    - 2 Disambiguate between category-specific representations that are functionally selective (with overlap) and functionally segregated (without).
    - 3 Tell whether the X-Y mapping is degenerate (many-to-one).

# From classical encoding to Bayesian decoding

reversing the standard GLM

## Encoding models

$X$  as a cause

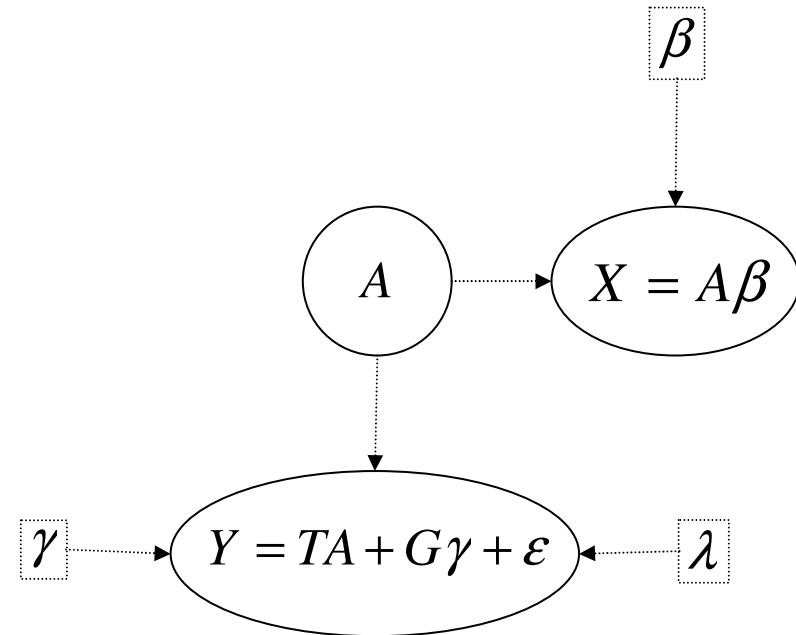


$$g(\theta) : X \rightarrow Y$$

$$Y = TX\beta + G\gamma + \varepsilon$$

## Decoding models

$X$  as a consequence



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

$$X = A\beta$$

$$TX = Y\beta - G\gamma\beta - \varepsilon\beta$$

# Hierarchical priors on patterns

spatial deployment of the X-Y mapping

- 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  $\beta$ .
- MVB specifies several alternative coding hypotheses in terms of empirical spatial priors on voxel weights.

→ project onto spatial basis function set:

$$\beta = U \eta \quad \text{patterns}$$

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

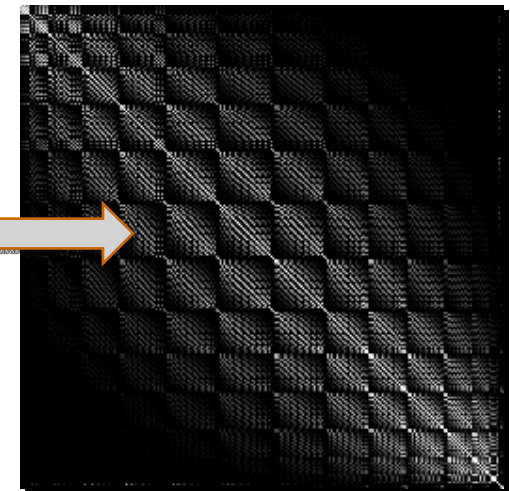
null:  $U = \emptyset$

compact vectors:  $U = I$

smooth vectors:  $U(\vec{x}_i, \vec{x}_j) = \exp(-\frac{1}{2}(\vec{x}_i - \vec{x}_j)^2 \sigma^{-2})$

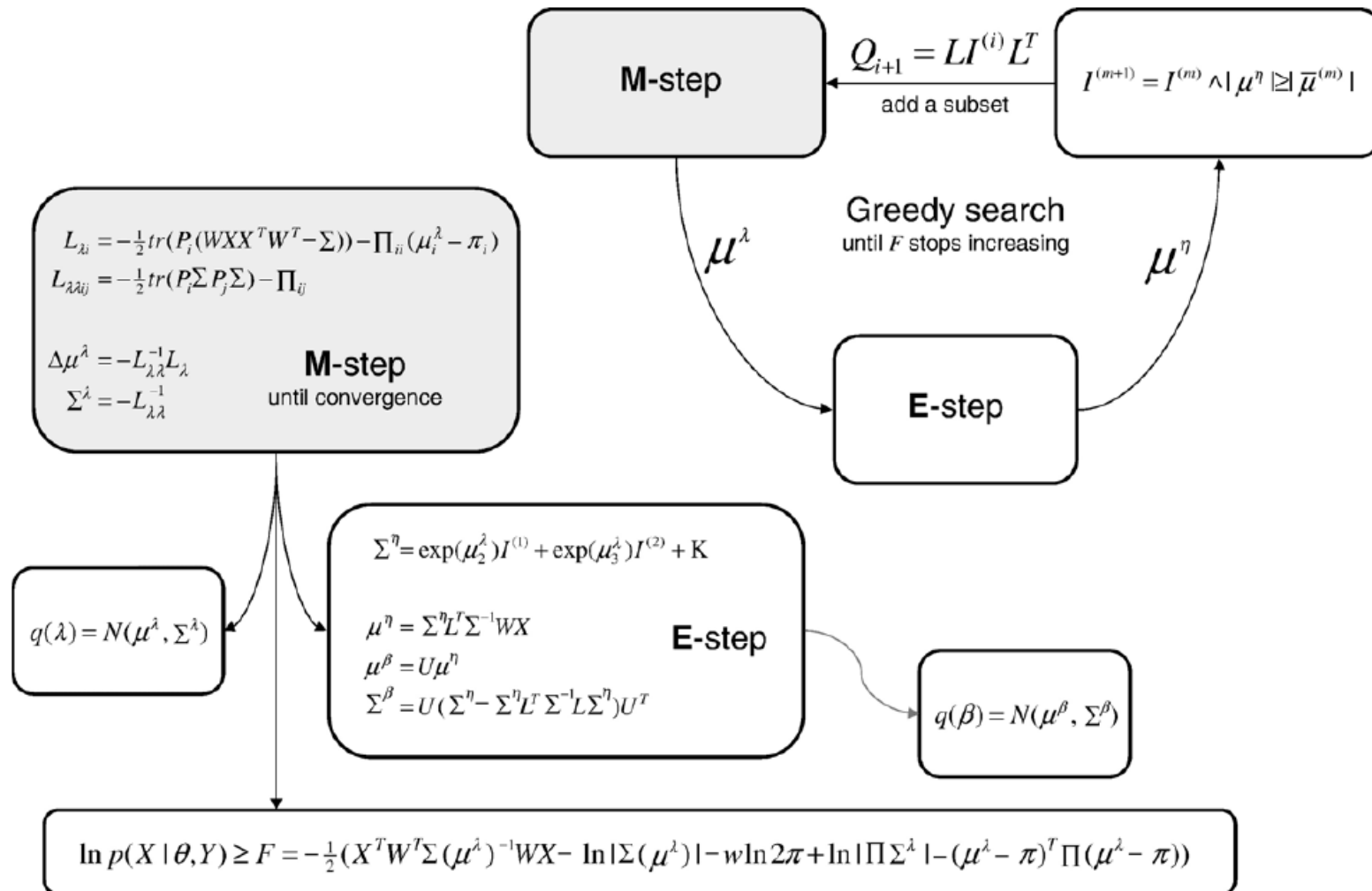
singular vectors:  $UDV^T = RY^T$

support vectors:  $U = RY^T$



# Hierarchical priors on patterns

Expectation-Maximization and the greedy search





# Probabilistic inference

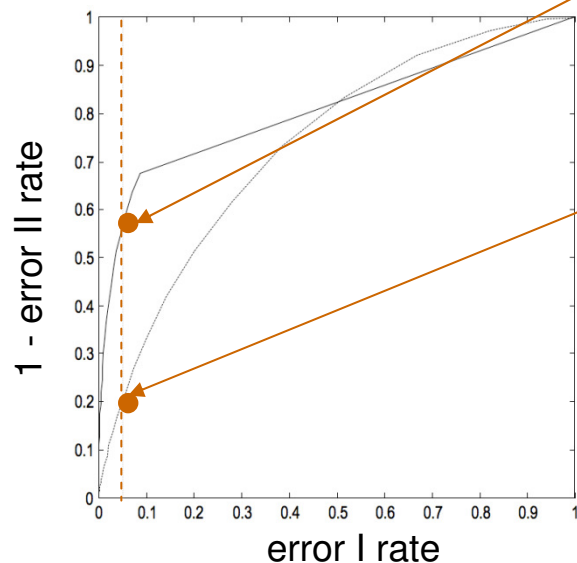
## classical significance of Bayesian model comparison

- What is the threshold  $u$  above which the log-Bayes factor *test*

$$\ln \Lambda = \ln p(X|Y, U \neq \emptyset) - \ln p(X|H_0 : U = \emptyset) \geq u$$

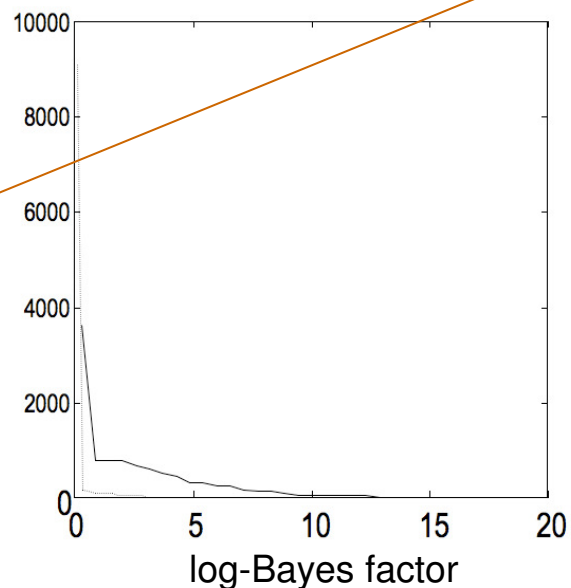
yields a type I error rate of  $p(\ln \Lambda \geq u | H_0) = 0.05$  ?

### ROC analysis



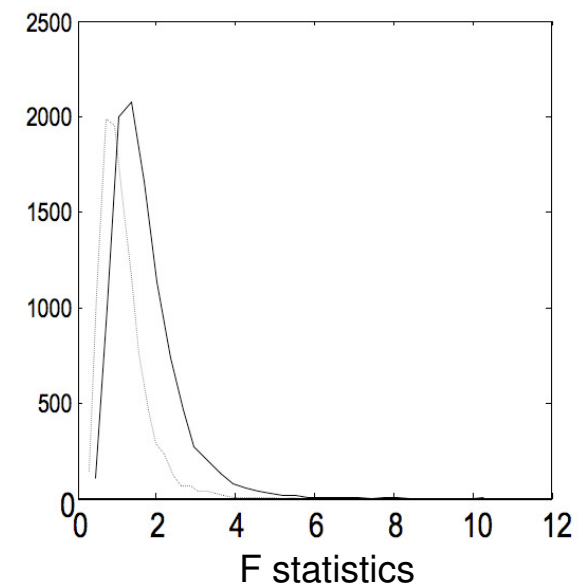
### MVB

$u=1.09$ , power=56%



### CCA

$F=2.20$ , power=20%



# Probabilistic inference

classical inference with cross-validation

- p-values from the standard leave-one-out scheme can't be used for inference (train and test data *are not* independent)
- Recall compact form for the decoding model:

$$WX = RY\beta + \zeta \quad \text{target variable}$$

$$W = RT \quad \text{weighting matrix: temporal convolution + confounds removal}$$

$$R = I - GG^{-} \quad \text{residual forming matrix: confounds removal}$$

- Use train/test k-fold data features that are linearly independent:

**train** (identify mapping)

$$\hat{\beta}_{(-k)} = \langle \beta | Y_{(-k)} \rangle$$

$$Y_{(-k)} = R_{(-k)} Y$$

$$R_{(-k)} = (I - G_{(-k)} G_{(-k)}^{-})$$

$$G_{(-k)} = \begin{bmatrix} G & I^{(k)} \end{bmatrix}$$

**test** (measure generalization error)

$$WX \stackrel{?}{=} \hat{X}_{(k)}$$

$$\hat{X}_{(k)} = R_{(k)} Y \hat{\beta}_{(-k)}$$

$$R_{(k)} = (I - G_{(k)} G_{(k)}^{-})$$

$$G_{(k)} = \begin{bmatrix} G & I - I^{(k)} \end{bmatrix}$$

# Overview of the talk

## 1 Introduction

*1.1 Lexicon*

*1.2 “Decoding”: so what?*

*1.3 Multivariate: so what?*

*1.4 Preliminary statistical considerations*

## 2 Multivariate Bayesian decoding

*2.1 From classical encoding to Bayesian decoding*

*2.2 Hierarchical priors on patterns*

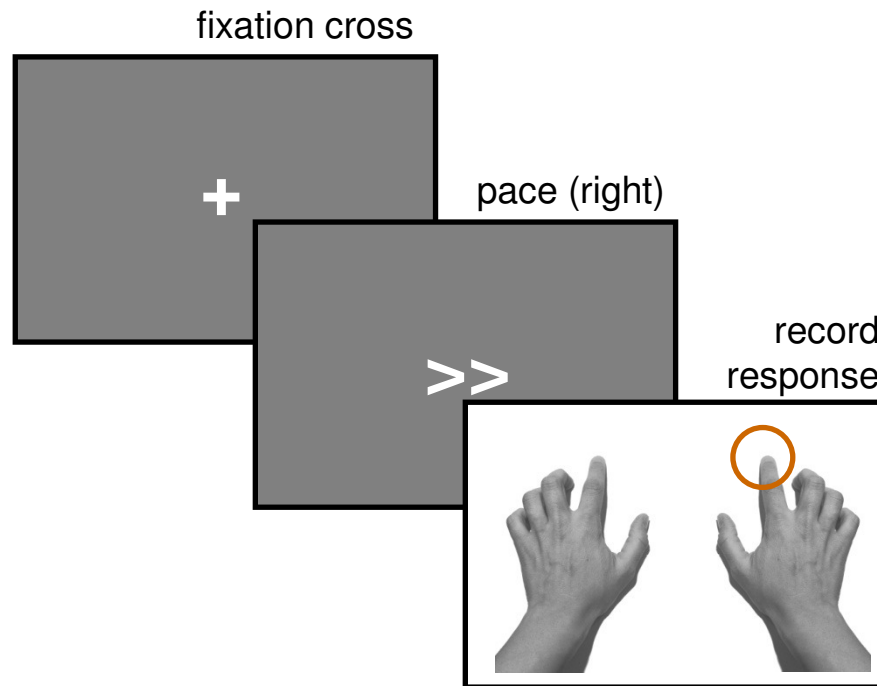
*2.3 Probabilistic inference*

## 3 Example

## 4 Summary

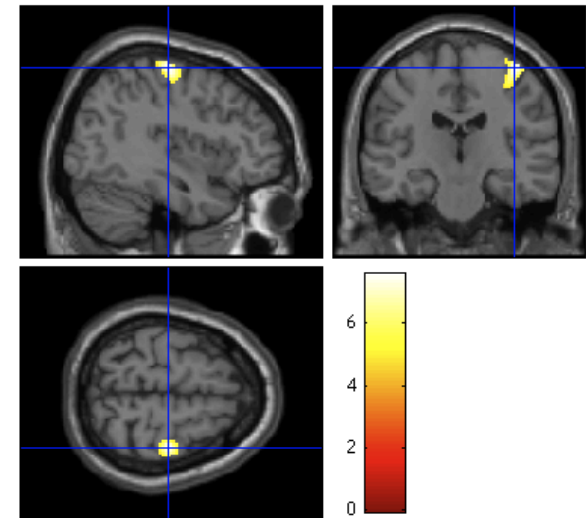
# Example

finger tapping dataset

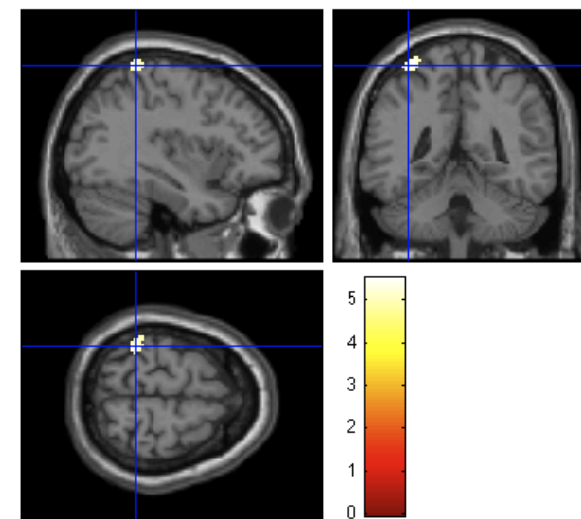


- 400 events (100 left, 100 right, 100 left & right, 100 null)
- average ITI = 2 sec
- block design (10 trials/block)
- TR = 1.3 sec

SPM{T} : left > right

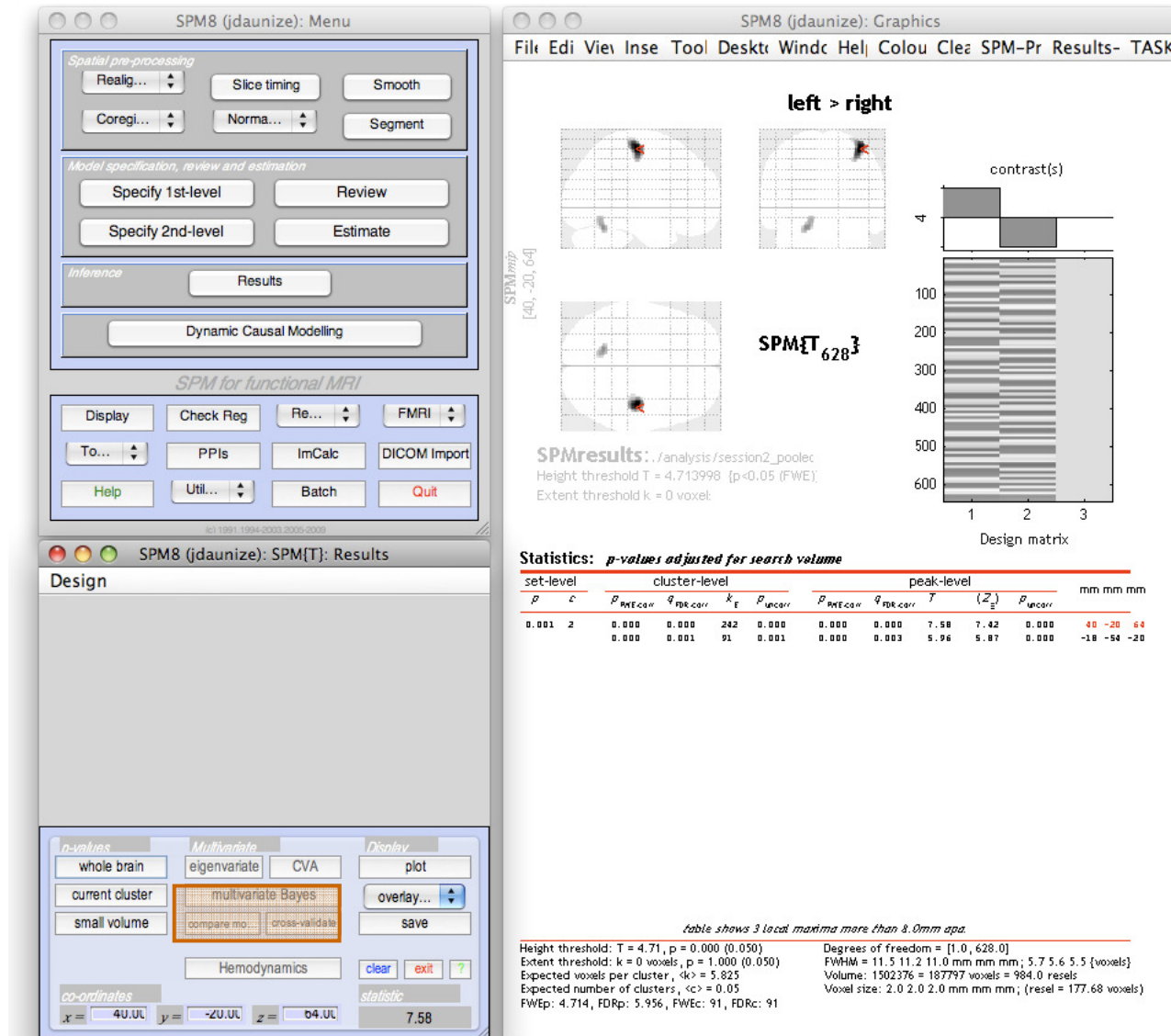


SPM{T} : right > left



# Example

## MVB in SPM: decoding within a search volume



- target:

$$TX = Xc$$

design  
matrix

contrast

- confounds:

$$G = X(I - cc^{-1})$$

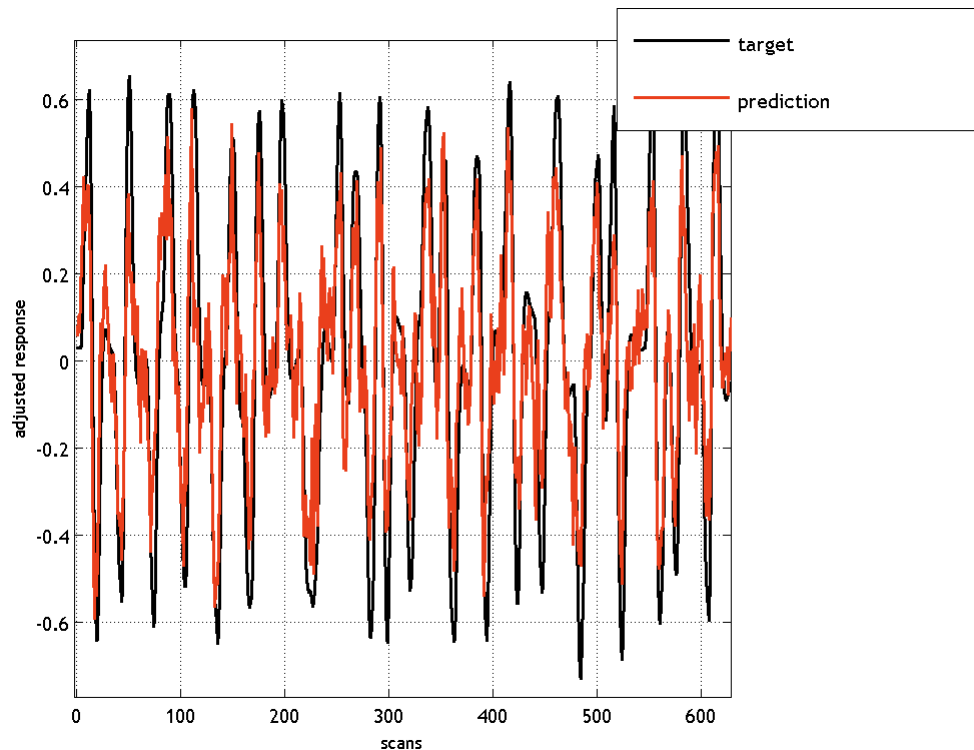
$$\Rightarrow Gc = 0$$

# Example

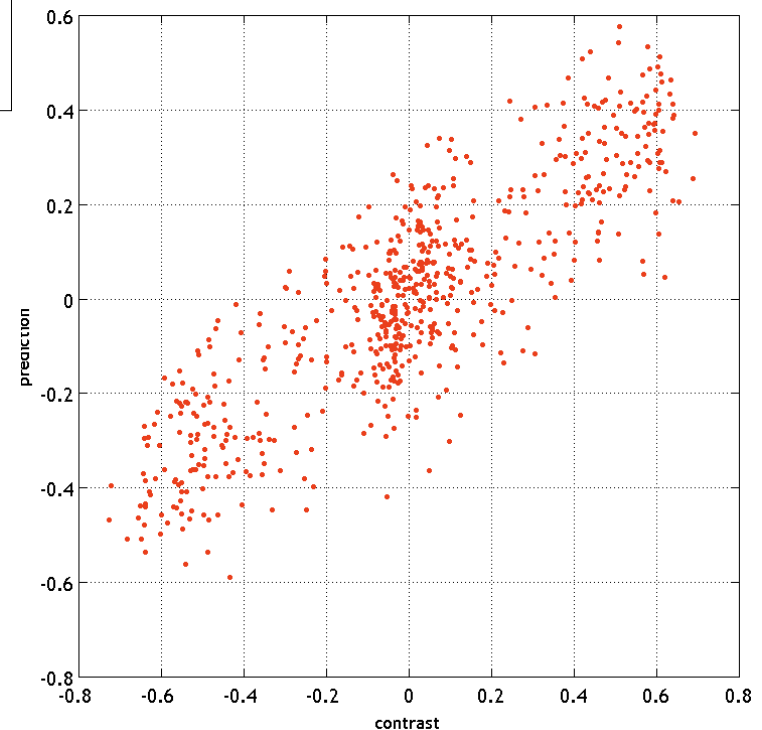
predicted responses from left & right motor cortices

- MVB-based predictions closely match the observed responses. But crucially, they don't perfectly match them. Perfect match would indicate overfitting.

target and (MVB) predicted responses over scans



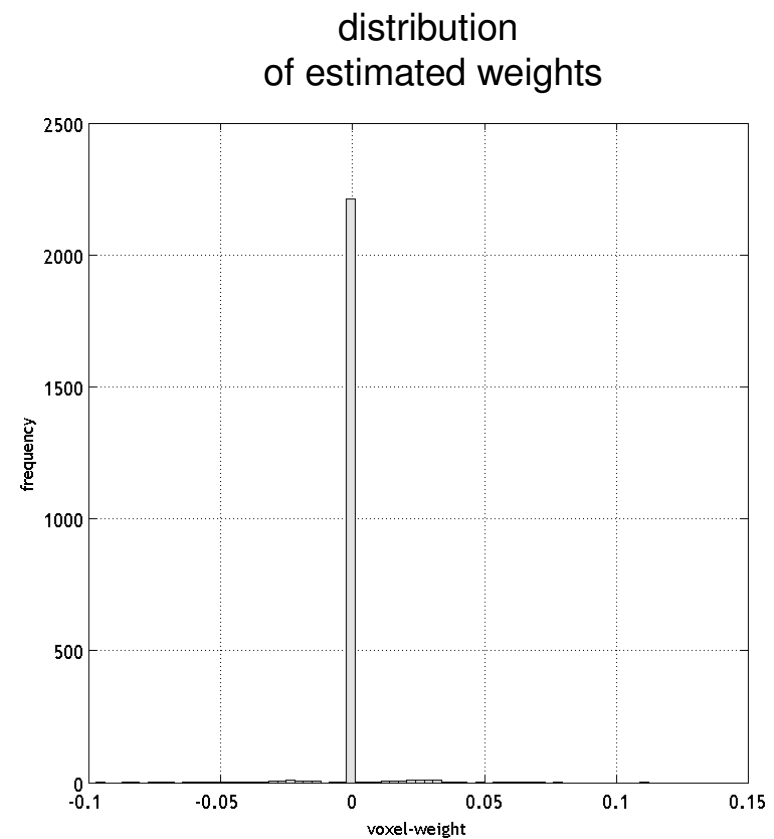
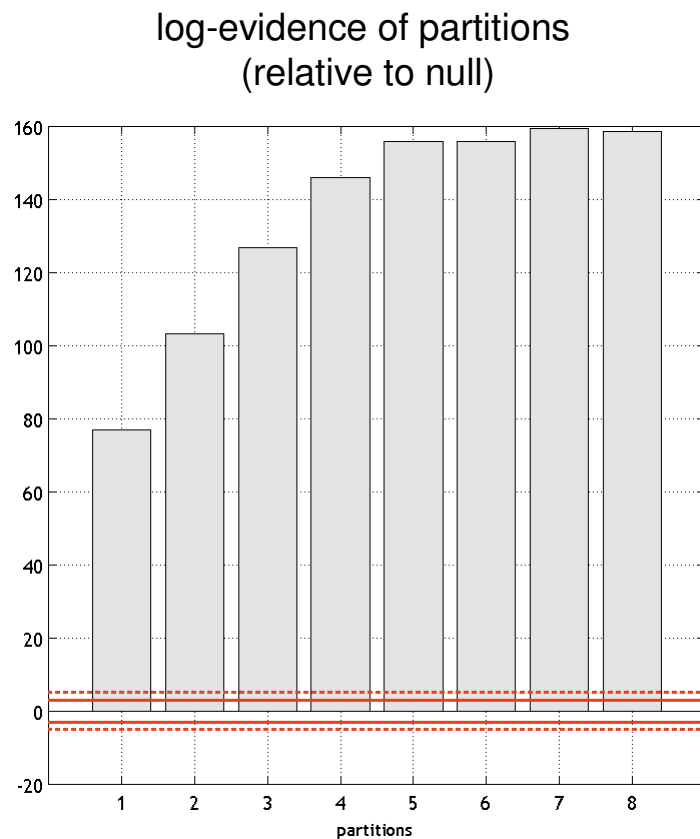
target versus (MVB) predicted responses



# Example

patterns sparsity

- The highest model evidence is achieved by a model that recruits 7 partitions. The weights attributed to each voxel in the sphere are sparse and multimodal. This suggests sparse coding.

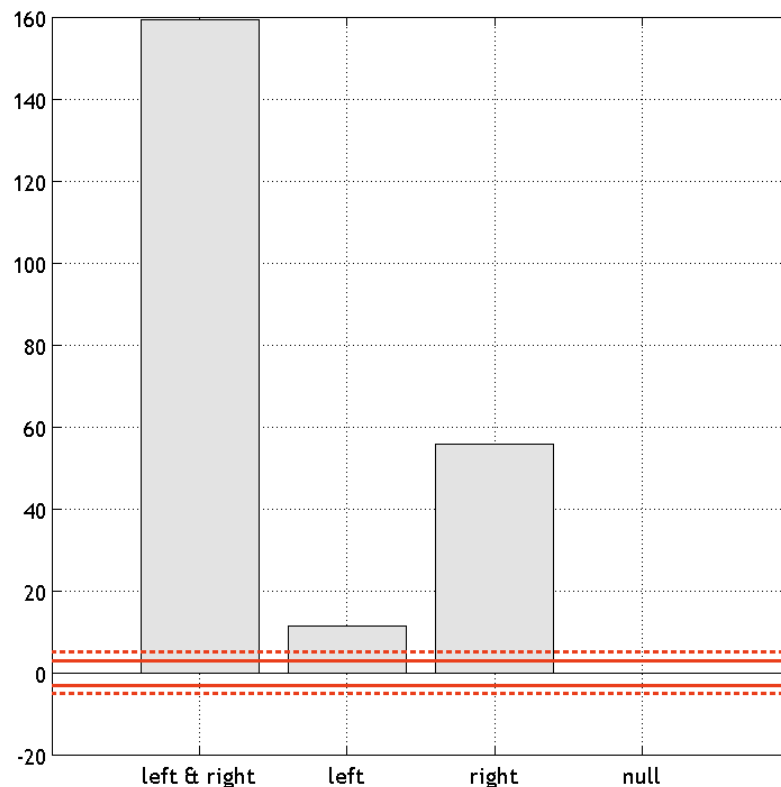


# Example

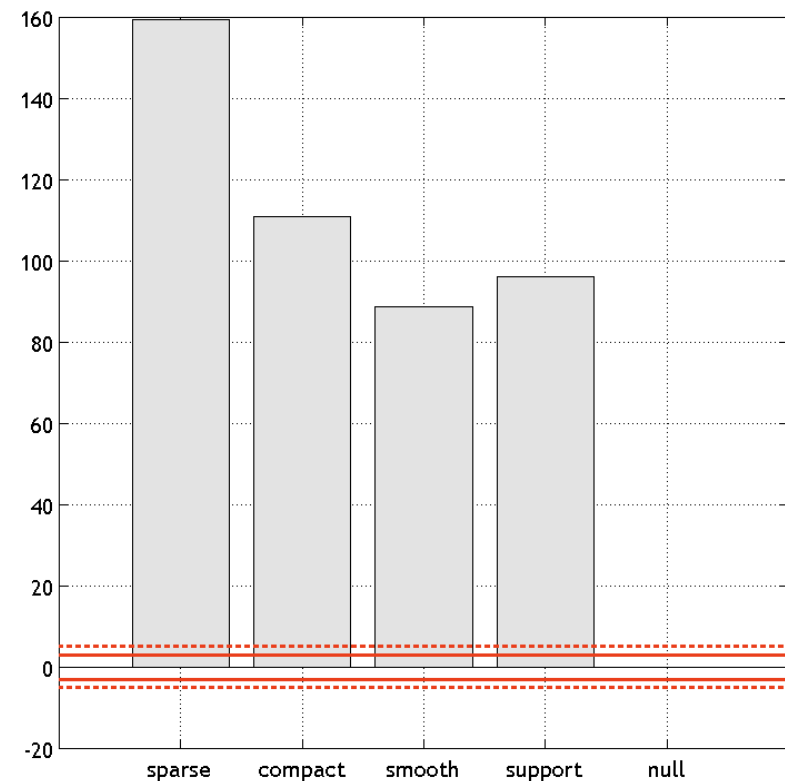
## model comparison illustration

- The best model corresponds to a sparse representation of motion ; as one would expect from functional segregation.

log-evidence of X-Y sparse mappings:  
effect of lateralization



log-evidence of X-Y bilateral mappings:  
effect of spatial deployment



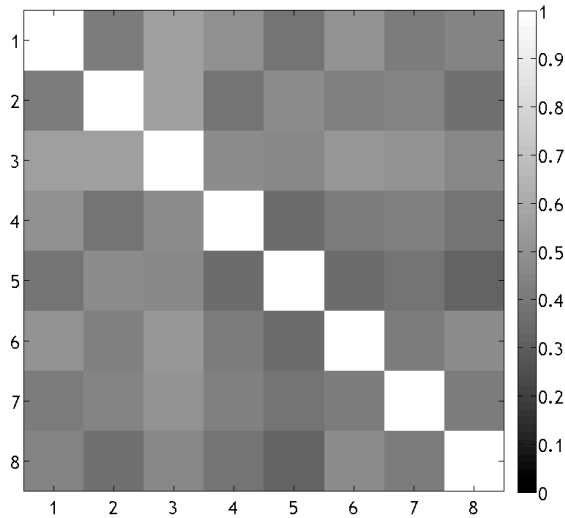


# Example

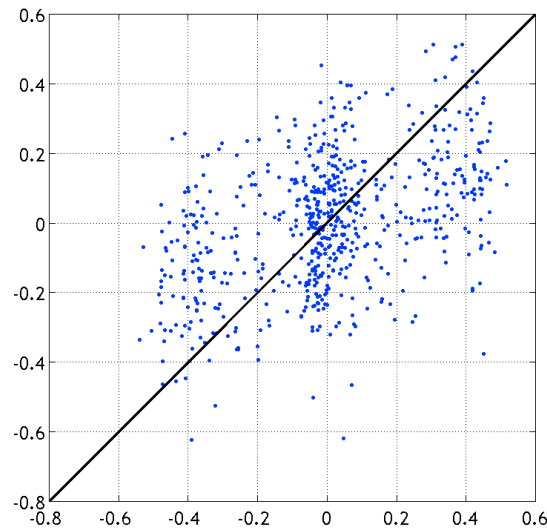
cross-validation : k-fold scheme

- $k = 8$
- $p\text{-value} < 0.0001$
- classification accuracy = 65.8%
- R-squared = 20.7%

absolute correlation among  
among k-fold feature weights

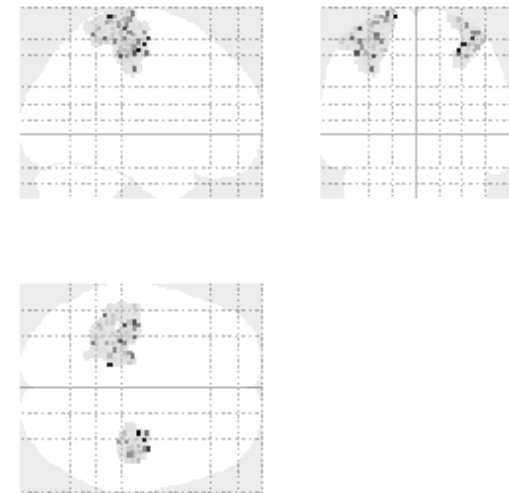


test predictions versus  
test k-fold features



maximum intensity projection:

$$\prod_k P(|\beta| > 0 | Y_{(-k)})$$



# Overview of the talk

## 1 Introduction

*1.1 Lexicon*

*1.2 “Decoding”: so what?*

*1.3 Multivariate: so what?*

*1.4 Preliminary statistical considerations*

## 2 Multivariate Bayesian decoding

*2.1 From classical encoding to Bayesian decoding*

*2.2 Hierarchical priors on patterns*

*2.3 Probabilistic inference*

## 3 Example

## 4 Summary

# Summary

- 1 Inference on the form of the X-Y mapping rests on model comparison, using the marginal likelihood of competing models. The marginal likelihood derives from the specification of a generative model prescribing the form of the joint density over observations (X,Y) and model parameters ( $\theta$ ).
- 2 Multivariate models can map from experimental variables (X) to brain responses (Y) or from Y to X. In the latter case (i.e., decoding), identifying the mapping is an ill-posed problem, which is resolved with appropriate constraints or priors on model parameters. These constraints are part of the model and can be evaluated using model comparison.
- 3 Cross-validation is not necessary for decoding brain activity but generalization error is a proxy for testing whether the observed X-Y mapping is unlikely to have occurred by chance. This can be useful when the null distribution of the likelihood ratio (i.e. Bayes factor) is not evaluated easily.