

# A functional MRI mind-reading game

Charles Zheng and Yuval Benjamini


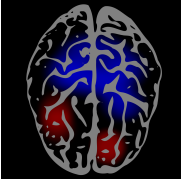

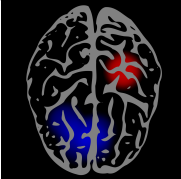
Stanford University

March 29, 2015

# Section 1

## Introduction

# Functional MRI

Stimuli	Response
	
	

# Functional MRI

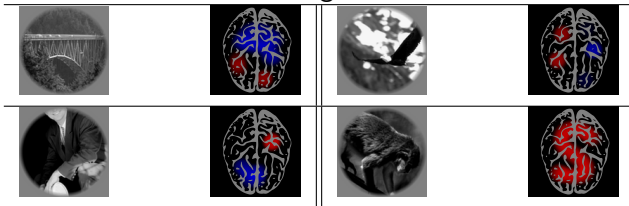
Stimuli $x$	Response $y$
$\begin{pmatrix} 1.0 \\ 0 \\ 3.0 \\ 0 \\ -1.2 \end{pmatrix}$	$\begin{pmatrix} 1.2 \\ 0 \\ -1.8 \\ -1.2 \end{pmatrix}$
$\begin{pmatrix} 0 \\ -2.2 \\ -3.1 \\ 4.5 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -1.2 \\ -1.9 \\ 0.5 \\ 0.6 \end{pmatrix}$

# Encoding vs Decoding

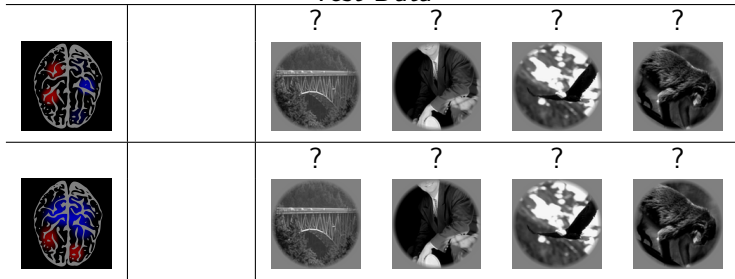
- Encoding: predict  $y$  from  $x$ .
- Decoding: reconstruct  $x$  from  $y$  (mind-reading).

# A mind-reading game: Classification

Training Data

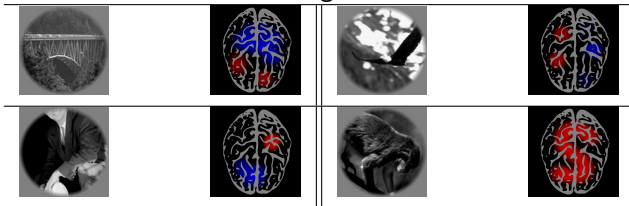


Test Data

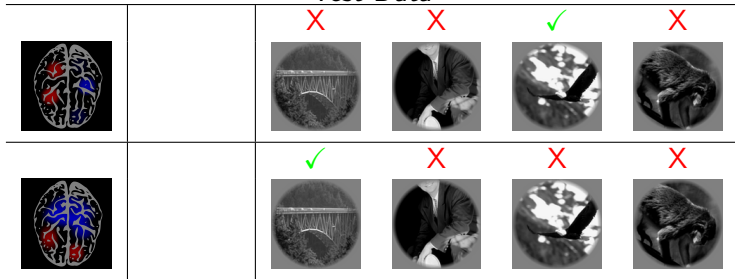


# A mind-reading game: Classification

Training Data

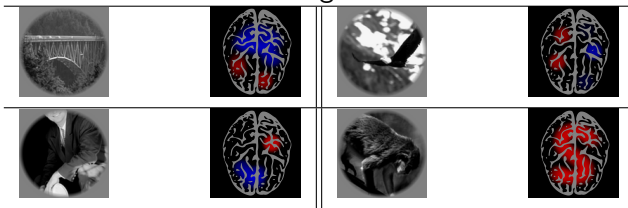


Test Data

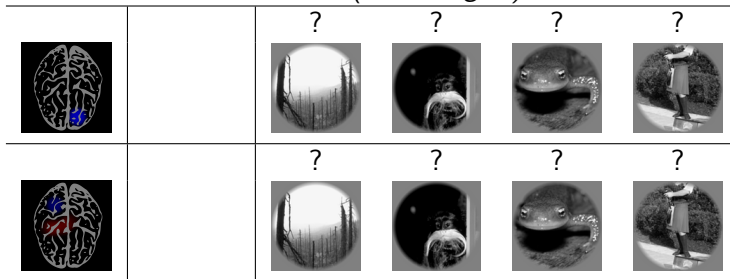


# A mind-reading game: Identification

Training Data



Test Data (*new images!*)





# Statistical formulation I

## *Training data.*

- Given training classes  $S_{\text{train}} = \{\text{train:1}, \dots, \text{train:k}\}$  where each class  $\text{train:i}$  has features  $x_{\text{train:i}}$ .
- For  $t = 1, \dots, T_{\text{train}}$ , choose class label  $z_{\text{train:t}} \in S_{\text{train}}$ ; generate

$$y_{\text{train:t}} = f(x_{z_{\text{train:t}}}) + \epsilon_t$$

where  $f$  is an unknown function, and  $\epsilon_t$  is i.i.d. from a known or unknown distribution.

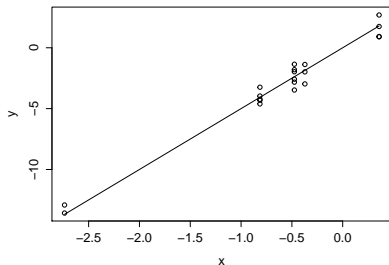
## *Test data.*

- Given test stimuli  $S_{\text{test}} = \{\text{test:1}, \dots, \text{test:l}\}$  with features  $\{x_{\text{test:1}}, \dots, x_{\text{test:l}}\}$
- Task: for  $t = 1, \dots, T_{\text{test}}$ , label  $y_{\text{test:t}}$  by stimulus  $\hat{z}_{\text{test:t}} \in S_{\text{train}}$ ; try to minimize misclassification rate

# Statistical formulation II

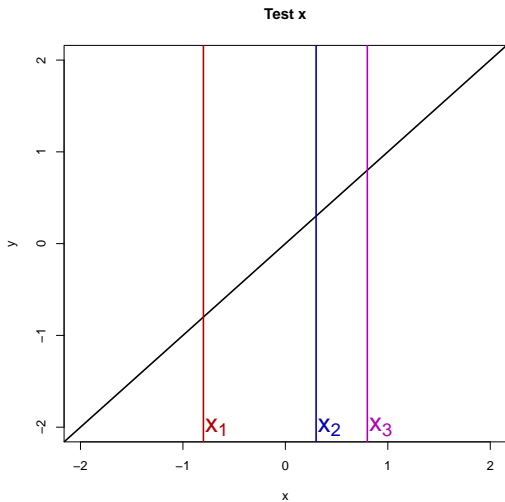
- $f$  is an unknown function
- $P$  is a known or unknown distribution over image features
- *Training data.* Draw  $x_{\text{train}:i} \sim P$  for  $i = 1 \text{ hdots}, k$ .
- *Test data.* Draw  $x_{\text{train}:i} \sim P$  for  $i = 1 \text{ hdots}, \ell$ .
- Theoretical question: Analyze average misclassification rate when classes are generated this way

# Toy example I

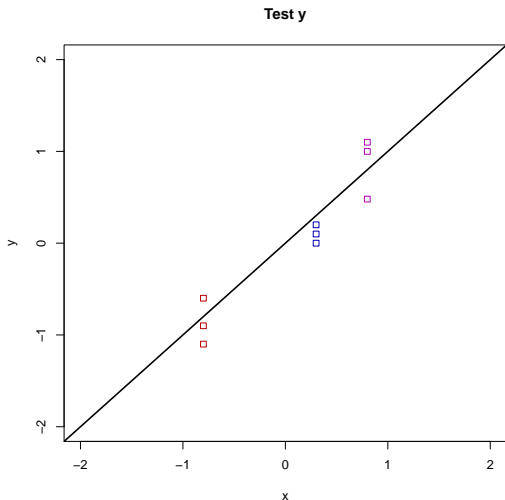


- Features  $x$  are one-dimensional real numbers, as are responses  $y$ . Parameter  $\beta$  is also a real number.
- Model is linear:  $y \sim N(x\beta, \sigma_\epsilon^2)$



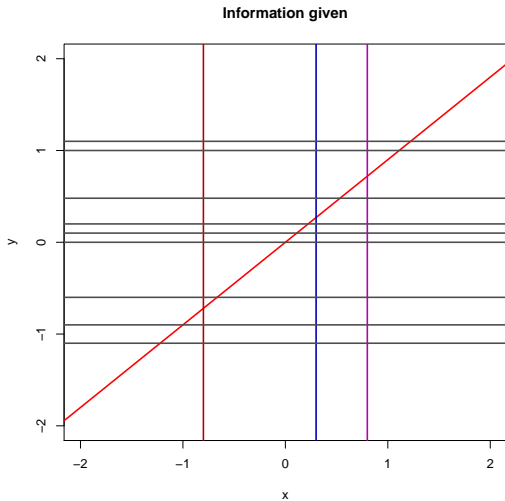


Generate features  $x_{\text{test}:1}, \dots, x_{\text{test}:\ell}$  iid  $N(0, \sigma_x^2)$ .

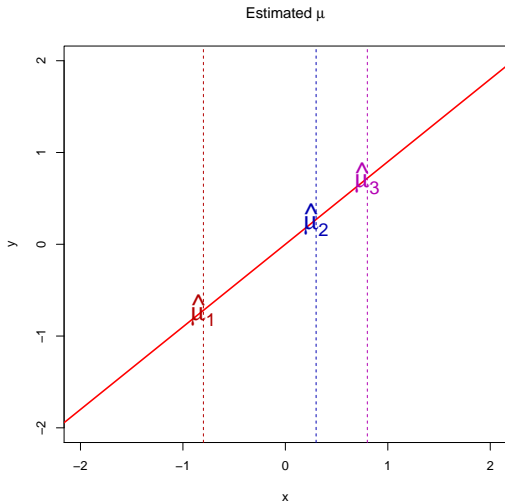


Hidden labels  $z_{\text{test}:t}$  are iid uniform from  $S_{\text{train}}$ .

Generate  $y_{\text{test}:t} \sim N(\beta x_{z_{\text{test}:t}}, \sigma_{\epsilon}^2)$

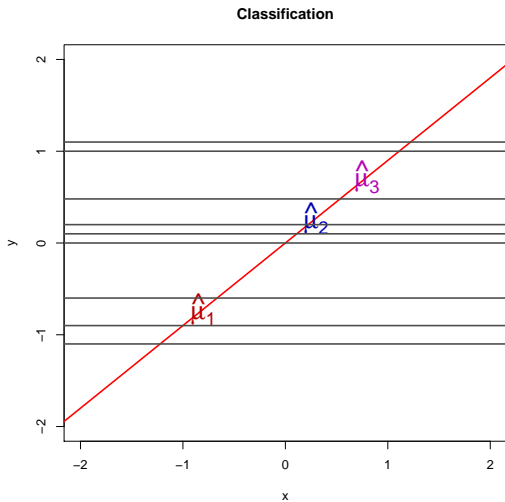


Classify  $\hat{y}_{\text{test}:t}$



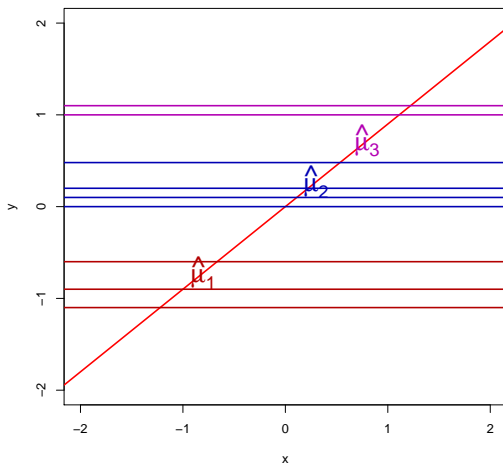
$$\hat{\mu}_{\text{test}:i} = \hat{\beta} x_{\text{test}:i}$$





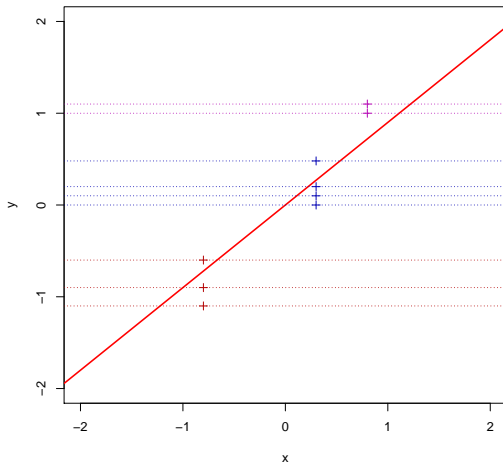
$$\hat{z}_{\text{test}:t} = \operatorname{argmin}_z \ell_{\hat{\mu}_z}(y_{\text{test}:t})$$

# Classification

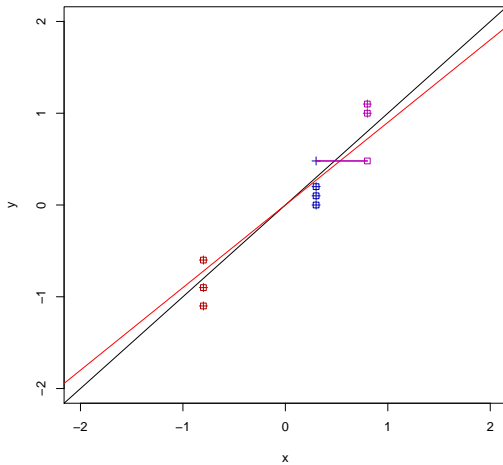


$$\hat{z}_{\text{test}:t} = \operatorname{argmin}_z (\hat{\mu}_z - y_{\text{test}:t})^2$$

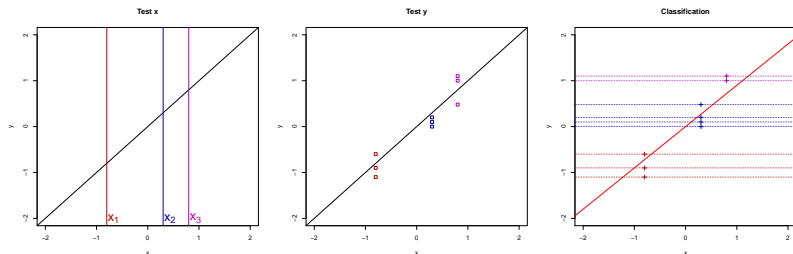
## Classification



### Misclassification



# Toy example I



- Generate features  $x_{\text{test}:1}, \dots, x_{\text{test}:\ell}$  iid  $N(0, \sigma_x^2)$ .
- Hidden labels  $z_{\text{test}:t}$  are iid uniform from  $S_{\text{train}}$ . Generate  $y_{\text{test}:t} \sim N(\beta x_{z_{\text{test}:t}}, \sigma_\epsilon^2)$
- Classify  $\hat{y}_{\text{test}:t}$  by maximum likelihood assuming  $\hat{\beta}$  is correct. Thus:

$$\hat{z}_{\text{test}:t} = \operatorname{argmin}_z (\hat{\beta} x_z - y_{\text{test}:t})^2$$

# Toy example I: Questions

- 1 We know the prediction error is minimized when  $\hat{\beta} = \beta$ . Is it also true that misclassification error in the mind-reading game is minimized when  $\hat{\beta} = \beta$ ?
- 2 Even if the answer to 1. is yes, should we estimate  $\hat{\beta}$  using the same methods as in least-squares regression?

# Toy example I: Analysis

- The expected misclassification error is the same if we take  $T_{\text{test}} = 1$ . Then let  $(x_*, y_*)$  be the feature-response pair in the test set, where

$$y_* = x_*\beta + \epsilon_*$$

- Denote the features for the incorrect classes as  $x_1, \dots, x_{\ell-1}$ .
- Let  $\delta = \hat{\beta} - \beta$ .

Ignore the possibility of ties. The response  $y_*$  is misclassified if and only if

$$\min_{i=1,\dots,\ell-1} |y_* - x_i \hat{\beta}| < |y_* - x_* \hat{\beta}|$$

equivalently

$$\cup_{i=1,\dots,\ell-1} E_i$$

where  $E_i$  is the event

$$|x_* \beta + \epsilon_* - x_i(\beta + \delta)| < |-\delta x_* + \epsilon_*|$$

with probability

$$\Pr[E_i] = \left| \Phi\left(\frac{x_*}{\sigma_x}\right) - \Phi\left(\frac{x_*(\beta - \delta) + 2\epsilon_*}{\sigma_x(\beta + \delta)}\right) \right|$$



# Toy example I: Analysis

- Use the following conditioning

$$\mathbf{E}[\text{misclassification}] = \mathbf{E}[\mathbf{E}[\Pr_{x_1, \dots, x_\ell}[\cup_i E_i] | x_* = x, \epsilon_* = \epsilon]]$$

- An exact expression for expected misclassification is therefore

$$1 - \int_{\epsilon} \left[ \int_x \left( 1 - \left| \Phi\left(\frac{x}{\sigma_x}\right) - \Phi\left(\frac{x(\beta - \delta) + 2\epsilon}{\sigma_x(\beta + \delta)}\right) \right| \right)^{\ell-1} d\Phi\left(\frac{x}{\sigma_x}\right) \right] d\Phi\left(\frac{\epsilon}{\sigma_{\epsilon}}\right)$$

- Question 1: Is this minimized at  $\hat{\beta} = \beta$ ?

Answer: yes. (Part of a proof:)

Fix  $\epsilon > 0$ . The derivative of the inner integral wrt  $\delta = 0$  is proportional to

$$\int_x (1 - \Phi(\frac{x\beta + 2\epsilon}{\sigma_x\beta}) + \Phi(\frac{x}{\sigma_x})) \phi(\frac{x\beta + 2\epsilon}{\sigma_x\beta}) (x + \frac{\epsilon}{\beta}) \phi(\frac{x}{\sigma_x}) dx$$

In turn

$$\phi\left(\frac{x\beta + 2\epsilon}{\sigma_x\beta}\right) \phi\left(\frac{x}{\sigma_x}\right) \propto \phi\left(\frac{\sqrt{2}(x + \frac{\epsilon}{\beta})}{\sigma_x}\right)$$

which is the density of a normal variate with mean  $-\epsilon/\beta$

But now note that the other terms

$$\left(1 - \Phi\left(\frac{x\beta + 2\epsilon}{\sigma_x\beta}\right) + \Phi\left(\frac{x}{\sigma_x}\right)\right) \left(x - \frac{\epsilon}{\beta}\right)$$

are symmetric about  $x = -\frac{\epsilon}{\beta}$ .

Thus by symmetry, the derivative of the inner integral  $\delta = 0$  vanishes. The same argument works for  $\epsilon < 0$ , hence the misclassification rate is stationary at  $\hat{\beta} = \beta$ .