# A principled approach to decoding

Charles Zheng and Yuval Benjamini

March 16, 2015

**Abstract**

The goal of these function MRI studies is to understand the relationship between $x^{(t)}$ and $y^{(t)}$, where $x^{(t)}$ is a vector of stimuli features and $y^{(t)}$ is a vector of brain activity features: this goal can be subdivided into the subgoal of learning an *encoding model*, which predicts the response $y$ given the simulus, and the subgoal of learning a *decoding model*, which reconstructs the stimulus given the response $y$. One could interpret both models as multivariate regression problems, with encoding fitting a model of the form $Y = f(X)+\epsilon$ and decoding fitting a model of the form $X = g(Y)+\varepsilon$. However, the regression formulation is not the only interpretation of the encoding/decoding problem. Notably, Kay *et al* treat the encoding problem as a linear model, but pose the decoding problem as one of *identification*: that is, given stimuli-response pairs $(x^{[i_1]}, y^{(1)}), \ldots, (x^{[i_j]}, y^{(j)})$ where the unobserved $x^{[i]}$ lie in a known set of stimuli $S = \{x^{[1]}, \ldots, x^{[|S|]}\}$, correctly recover the labels $i_1, \ldots, i_j$ given only the responses $y^{(1)}, \ldots, y^{(j)}$. Furthermore, Kay *et al* quantify the quality of the decoding model by the classification rate for the identification problem when $S$ is selected randomly from a larger database of images $\mathcal{S}$ (Kay 2008, Vu 2011). This approach is more suited for the goal of identifying *natural images* from fMRI responses, and has been adopted by numerous fMRI studies (Chen 2013). Such studies usually use a combination of multivariate linear or nonlinear models and feature selection to implement the decoding model. However, such studies have not explicitly motivated their decoding models based on the criterion of maxmizing correct classification for random stimuli subsets. We proposed a principled approach to decoding, wherein we formulate a decoding model which optimally maximizes the identification perfomance of the model. Our approach is based on a theoretical analysis of the identification perfomance of a linear model, resulting in an approximate measure of identification performance which can be tractably optimized in training data.

# 1 Introduction

## 1.1 Background

In functional MRI (fMRI) studies, one presents a sequence of $T$ (possibly repeated) stimuli parameterized by features $x^{(1)}, \ldots, x^{(T)}$, where each $x^{(i)}$ is a $p$-dimensional vector. The time-varying MRI image is processed to yield corresponding response profiles $y^{(1)}, \ldots, y^{(T)}$, where each $y^{(i)}$ is a vector of $V$ voxel-specific responses.

# 2 Theory

## 2.1 Classification of random images

We deal first with the *classification* approach to image identification.
*Simplest case*

The simplest model is as follows. Let $\mu_1, \ldots, \mu_N$ be $d$-dimensional mean fMRI responses for images $1, \ldots, n$, drawn iid from a normal distribution: $\mu_i \sim N(0, \Sigma_\mu)$, and suppose for now that $\mu_i$ are known to the experimenter. Let $j_1, \ldots, j_T$ be random labels drawn uniformly from $\{1, \ldots, N\}$, and let $y_t = \mu_{j_t} + \epsilon_t$ where $\epsilon_t \sim N(0, \sigma^2 I)$. Then the classification rule is to estimate

$$\hat{j}_t = \text{argmin}_{j \in \{1,\ldots,N\}} ||y_t - \mu_j||^2$$

The classification is correct in the event that $||y_t - \mu_{j_t}||^2 < ||y_t - \mu_j||^2$ for all $j \neq j_t$, and hence the average correct classification rate is

$$\text{CC} = \frac{1}{T} \sum_{i=1}^T \Pr[||y_t - \mu_i||^2 = \min_j ||y_t - \mu_j||^2 | j_t = i]$$

Due to exchangeability, we need only consider the expression for $t = 1$, and conditional on $j_1 = 1$, hence

$$\begin{aligned} \text{CC} &= \Pr[||y_1 - \mu_1||^2 < \min_{j>1} ||y_t - \mu_j||^2 | j_1 = 1] \\ &= \Pr[\forall j > 1 : \mu_j \notin B_{||\epsilon_1||}(y_1)] \\ &= \Pr[\mu_2 \notin B_{||\epsilon_1||}(y_1)]^{T-1} \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \left[ 1 - \int_{B_{||\epsilon||}(y)} p(\mu) d\mu \right]^{T-1} dP(\epsilon, y) \end{aligned}$$

where $B_r(x)$ is the euclidean ball of radius $r$ centered at $x$ and

$$p(\mu) = \frac{1}{(2\pi|\Sigma_\mu|)^{d/2}} \exp(-\frac{1}{2}\mu^T \Sigma_\mu^{-1} \mu)$$

The preceding integral is over the joint distribution over $\epsilon, y$, where $y = \mu_1 + \epsilon$. The quantities $\epsilon, y$ effectively decouple if $\sigma^2 I << \Sigma_\mu$, in which case

$$\int_{B_{||\epsilon||}(y)} p(\mu) d\mu \approx \int_{B_{||\epsilon||}(\mu_1)} p(\mu) d\mu \approx p(\mu_1) \text{Vol}(B_{||\epsilon||})$$

Hence letting $\eta = ||\epsilon||^2$, we get

$$\begin{aligned} \text{CC} &\approx \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left[1 - \text{Vol}(B_{\sqrt{\eta}}) p(\mu)\right]^{T-1} d\mu dP(\epsilon) \\ &= \int_0^\infty \int_{\mathbb{R}^d} \left[1 - \eta^{d/2} \text{Vol}(B_1) p(\mu)\right]^{T-1} p(\eta) d\eta \\ &= \int_{\mathbb{R}^d} p(\mu) \int_0^\infty p(\eta) \left[1 - p(\mu) \eta^{d/2} V_d\right]^{T-1} d\eta d\mu \end{aligned}$$

where $\eta$ has a scaled Chi-squared distribution

$$p(\eta) = \frac{1}{2^{d/2} \Gamma(d/2) \sigma^2} \left(\frac{\eta}{\sigma^2}\right)^{k/2-1} e^{-\eta/2\sigma^2}$$

and

$$V_D = \text{Vol}(B_1) = \frac{\pi^{d/2}}{\Gamma((d+2)/2)}$$

We now seek to approximate the inner integral, denoted as $I(p(\mu))$:

$$I(p) = \int_0^\infty \left[1 - p\eta^{d/2} V_d\right]^{T-1} p(\eta) d\eta$$

When $T$ is large and $\eta$ is small, we can use the exponential function to approximate the power, giving

$$I(p) \approx \int_0^\infty \exp(-p(T-1) V_d \eta^{d/2}) p(\eta) d\eta = \mathbf{E}[\exp(-p(T-1) V_d \eta^{d/2})]$$

Making the additional assumption that $d$ is large, we can use the approximation that for nonnegative random variables $Z$,

$$\mathbf{E}[\exp[-cZ^d]] = \mathbf{E}[\exp[-(Z\sqrt[d]{c})^d]] \approx \Pr\left[Z\sqrt[d]{c} < 1\right] = \Pr[Z < 1/\sqrt[d]{c}]$$

This gives

$$I(p) \approx \Pr[\eta < 1/\sqrt[d/2]{p(T-1)V_d}] = \Pr\left[\chi_d^2 < \frac{1}{\sigma^2 \sqrt[d/2]{p(T-1)V_d}}\right]$$

All in all, we have

$$CC \approx \int_{\mathbb{R}} p(\mu) \Pr\left[\chi_d^2 < \frac{1}{\sigma^2 \sqrt[d/2]{p(\mu)(T-1)V_d}}\right] d\mu$$

### 2.2 Decoding of random images

## 3 Simulations

### 3.1 Validation of formulae

(This subsection will be omitted in the submitted version.)

## 4 References

- Kay, KN., Naselaris, T., Prenger, R. J., and Gallant, J. L. "Identifying natural images from human brain activity". *Nature* (2008)
- Vu, V. Q., Ravikumar, P., Naselaris, T., Kay, K. N., and Yu, B. "Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models", *The Annals of Applied Statistics.* (2011)
- Chen, M., Han, J,. Hu, X., Jiang, Xi., Guo, L. and Liu, T. "Survey of encoding and decoding of visual stimulus via fMRI: an image analysis perspective." *Brain Imaging and Behavior.* (2014)
- Schoenmakers, S., Barth, M., Heskes, T., van Gerven, M., "Linear reconstruction of percieved images from human brain activity" *NeuroImage* (2013)