Lecture Notes
Bayesian Neurocognitive Modeling
Whole-brain encoding and decoding

Marcel van Gerven

November 28, 2013

In this lecture, we will focus on the development of whole-brain encoding and decoding models. That is, how do we optimally model the BOLD response in individual voxels and how do we invert these models to predict the most likely stimulus that caused a response?

Learning Goals

We discuss different encoding and decoding approaches. After studying these notes, you should be able to explain the assumptions made by these models and how this affects the way decoding is implemented. You should be able to reproduce the derivations for the Gaussian encoding/decoding model.

Required Reading

You should study the following papers:

- Schoenmakers S, Barth M, Heskes T, van Gerven M. Linear reconstruction of perceived images from human brain activity. Neuroimage. 2013. http://dx.doi.org/10.1016/j.neuroimage.2013.07.043

- Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL. Bayesian reconstruction of natural images from human brain activity. Neuron. 2009;63(6):90215. http://dx.doi.org/10.1016/j.neuron.2009.09.006

1 Encoding

Encoding refers to modeling of how single voxels respond to stimuli. We define the model for a voxel k as:

$$\mathbf{y}_k = \mathbf{H}^{(k)}\mathbf{X}\boldsymbol{\beta}_k + \mathbf{G}\boldsymbol{\gamma}_k + \boldsymbol{\epsilon}_k \tag{1}$$

where $\mathbf{y}_k$ is the BOLD response for voxel k, $\mathbf{H}^{(k)}$ is an $N \times N$ temporal convolution matrix, $\mathbf{X}$ is the design matrix, $\mathbf{G}$ are confounds and $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \sigma_k^2\mathbf{I}_N)$ is the residual error. We temporarily drop the dependence on k and consider the model for one voxel. We also assume that confounds are absent (or regressed out) and the HRF is absorbed in the design matrix. In that case, we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

We have

$$P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \mathcal{N}(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N)\,. \tag{2}$$

This is a standard linear regression model. Hence, the parameters of the model (regression coefficients $\boldsymbol{\beta}$ and variance σ^2) can be estimated using standard approaches.

Dealing with confounds (motion, drifts, breathing, heartbeat) can be handled by either regressing them out beforehand or by absorbing it into the design matrix (without convolving with an HRF). Dealing with the hemodynamic response function can be done in various ways. We could assume a known HRF (i.e. the canonical HRF) and just absorb it in $\mathbf{X}$. In case of an unknown HRF, could redefine $\mathbf{X}$ as the K-times concatenation of the design matrix shifted forwards in time. That is,

$$\mathbf{X}^* = \left[\mathbf{X}, \mathbf{L}^1\mathbf{X}, \ldots, \mathbf{L}^K\mathbf{X}\right]$$

where $\mathbf{L}$ is the lower shift matrix with $l_{ij} = \delta_{i,j+1}$. Another approach is to first run a general linear model in order to get trial specific beta estimates. These estimates give for each voxel and each trial an estimated voxel response. This latter approach was used in [SBHvG13].

1.1 Parameter estimation

Parameter estimation boils down to estimating

$$(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k^2) = \arg\max_{\mathbf{b},\sigma^2} \left\{P(\mathbf{y}_k \mid \mathbf{X}, \mathbf{b}, \sigma^2) P(\mathbf{b}, \sigma^2)\right\}$$

for each voxel k. This can alternatively be written as

$$\begin{aligned}(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k^2) &= \arg\min_{\mathbf{b},\sigma^2} \left\{-\log P(\mathbf{y}_k \mid \mathbf{X}, \mathbf{b}, \sigma^2) - \log P(\mathbf{b}, \sigma^2)\right\} \\ &= \arg\min_{\mathbf{b},\sigma^2} \left\{L(\mathbf{b}, \sigma^2) + R(\mathbf{b}, \sigma^2)\right\}\end{aligned}$$

Let us focus on estimation of the regression coefficients beta (the variance can be analytically computed as well or estimated from hold out data). Given a univariate gaussian prior on the regression coefficients, we obtain (see practical):

$$\begin{aligned}\hat{\boldsymbol{\beta}}_k &= \arg\min_{\mathbf{b}} \left\{L(\mathbf{b}, \hat{\sigma}_k^2) + R(\mathbf{b})\right\} \\ &= \arg\min_{\mathbf{b}} \left\{\frac{1}{2}||\mathbf{y}_k - \mathbf{X}\mathbf{b}||_2^2 + \frac{\lambda}{2}||\mathbf{b}||_2^2)\right\}\end{aligned}$$

with L_p norm $||\cdot||_p$. The solution is given by

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}_k$$

2 Decoding

Consider the multivariate Gaussian density for a d-dimensional vector $\mathbf{x}$, written as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) .$$

This expression written in terms of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is called the moment form parameterization. An alternative parameterization of the Gaussian density is the canonical form parameterization, written as:

$$p(\mathbf{x}) = \exp\left(\alpha + \boldsymbol{\eta}^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Lambda}\mathbf{x}\right)$$

where $\alpha = -\frac{1}{2}(d \log 2\pi - \log|\boldsymbol{\Lambda}| + \boldsymbol{\eta}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\eta})$ is the normalizing constant. The two parameterizations are related as follows:

$$
\begin{aligned}
\boldsymbol{\Lambda} &= \boldsymbol{\Sigma}^{-1} \\
\boldsymbol{\eta} &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\
\boldsymbol{\Sigma} &= \boldsymbol{\Lambda}^{-1} \\
\boldsymbol{\mu} &= \boldsymbol{\Lambda}^{-1}\boldsymbol{\eta}
\end{aligned}
$$

The forward model is given by the multivariate Gaussian:

$$
p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}(\mathbf{y}; \mathbf{B}^T\mathbf{x}, \boldsymbol{\Sigma})
$$

with mean $\mathbf{B}^T\mathbf{x}$ and covariance matrix $\boldsymbol{\Sigma}$. This can equivalently be written in canonical form as

$$
p(\mathbf{y} \mid \mathbf{x}) = \exp\left(-\frac{1}{2}(d \log 2\pi - \log|\boldsymbol{\Lambda}| + \boldsymbol{\eta}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\eta}) + \boldsymbol{\eta}^T\mathbf{y} - \frac{1}{2}\mathbf{y}^T\boldsymbol{\Lambda}\mathbf{y}\right) \tag{3}
$$

with $\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1}\mathbf{B}^T\mathbf{x}$, $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ and $\alpha = -\frac{1}{2}(d \log 2\pi - \log|\boldsymbol{\Lambda}| + \boldsymbol{\eta}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\eta})$

We further assume that the image prior is given by a zero-mean multivariate Gaussian with covariance matrix $\mathbf{R}$ which, in canonical form, is given by

$$
p(\mathbf{x}) = \exp\left(-\frac{1}{2}(d \log 2\pi - \log|\boldsymbol{\Lambda}_0| + \boldsymbol{\eta}_0^T \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\eta}_0) + \boldsymbol{\eta}_0^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Lambda}_0\mathbf{x}\right) \tag{4}
$$

with $\boldsymbol{\eta}_0 = \mathbf{R}^{-1}\mathbf{0} = \mathbf{0}$ and $\boldsymbol{\Lambda}_0 = \mathbf{R}^{-1}$. Hence, (4) simplifies to

$$
p(\mathbf{x}) = \exp\left(-\frac{1}{2}(d \log 2\pi - \log|\boldsymbol{\Lambda}_0|) - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Lambda}_0\mathbf{x}\right) \tag{5}
$$

Given $p(\mathbf{y} \mid \mathbf{x})$ and $p(\mathbf{x})$, we can proceed with decoding. That is, we are interested in computing the mode of the distribution $p(\mathbf{x} \mid \mathbf{y})$. Combining Eqs. (3) and (5) using Bayes' rule and ignoring the normalization term, we obtain

$$
\begin{aligned}
p(\mathbf{x} \mid \mathbf{y}) &\propto \exp\left(-\frac{1}{2}d \log 2\pi + \frac{1}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\boldsymbol{\eta}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\eta} + \boldsymbol{\eta}^T\mathbf{y} - \frac{1}{2}\mathbf{y}^T\boldsymbol{\Lambda}\mathbf{y}\right) \\
&\quad \times \exp\left(-\frac{1}{2}d \log 2\pi + \frac{1}{2}\log|\boldsymbol{\Lambda}_0| - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Lambda}_0\mathbf{x}\right) \\
&= \exp\left(-d \log 2\pi + \frac{1}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\boldsymbol{\eta}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\eta} + \boldsymbol{\eta}^T\mathbf{y} - \frac{1}{2}\mathbf{y}^T\boldsymbol{\Lambda}\mathbf{y} + \frac{1}{2}\log|\boldsymbol{\Lambda}_0| - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Lambda}_0\mathbf{x}\right)
\end{aligned}
$$

Note that all terms that do not depend on $\mathbf{x}$ act as a normalization term given a fixed value of $\mathbf{y}$. Hence, plugging in the moment form parameters, we may write the preceding as

$$
\begin{aligned}
p(\mathbf{x} \mid \mathbf{y}) &\propto \exp\left(\alpha - \frac{1}{2}\boldsymbol{\eta}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\eta} + \boldsymbol{\eta}^T\mathbf{y} - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Lambda}_0\mathbf{x}\right) \\
&= \exp\left(\alpha - \frac{1}{2}(\boldsymbol{\Sigma}^{-1}\mathbf{B}^T\mathbf{x})^T\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1}\mathbf{B}^T\mathbf{x}) + (\boldsymbol{\Sigma}^{-1}\mathbf{B}^T\mathbf{x})^T\mathbf{y} - \frac{1}{2}\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right)
\end{aligned}
$$

where $\alpha = -d \log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{R}^{-1}|$ collects terms that do not depend on $\mathbf{x}$. By making use of the symmetry of the covariance matrix $\boldsymbol{\Sigma}$ and the properties of transposition, we can write this equivalently as

$$
\begin{aligned}
p(\mathbf{x} \mid \mathbf{y}) &\propto \exp\left(\alpha - \frac{1}{2}(\mathbf{x}^T\mathbf{B}\boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1}\mathbf{B}^T\mathbf{x}) + (\mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{y})^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right) \\
&= \exp\left(\alpha + (\mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{y})^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T(\mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{B}^T + \mathbf{R}^{-1})\mathbf{x}\right)
\end{aligned}
$$

This is recognized as a multivariate Gaussian in canonical form whose moment form parameterization is given by the mean $\mathbf{m} \equiv \mathbf{QB\Sigma}^{-1}\mathbf{y}$ and covariance $\mathbf{Q} = \left(\mathbf{R}^{-1} + \mathbf{B\Sigma}^{-1}\mathbf{B}^T\right)^{-1}$. It immediately follows that

$$\hat{\mathbf{x}} = \mathbf{m} = \left(\mathbf{R}^{-1} + \mathbf{B\Sigma}^{-1}\mathbf{B}^T\right)^{-1} \mathbf{B\Sigma}^{-1}\mathbf{y} \tag{6}$$

since the mode of a Gaussian distribution is equal to its mean.

3 Towards more complex models

Naselaris et al. [NPK$^+$09] proposed a model which combines structural and semantic information with a natural image prior. Decoding is given by

$$p(\mathbf{s} \mid \mathbf{r}) \propto p(\mathbf{s}) p_1(\mathbf{r}_1 \mid \mathbf{s}) p_2(\mathbf{r}_2 \mid \mathbf{s})$$

where $\mathbf{s}$ is the stimulus, $\mathbf{r}$ is the response, $p(\mathbf{s})$ is the empirical prior, p_1 is the distribution for structural voxels and p_2 is the distribution for semantic voxels.

The probability of a response for a structural voxel is given by

$$p(r \mid \mathbf{s}) \propto \exp\left(-\frac{(r - \mathbf{h}^T\mathbf{f}(\mathbf{s}))^2}{2\sigma^2}\right)$$

with nonlinear feature representation $\mathbf{f}(\mathbf{s}) = \log(|W^T\mathbf{s}| + 1)$ where W are the filters of a Gabor wavelet pyramid.

The probability of a semantic voxel is given by

$$p(r \mid c(\mathbf{s})) = \sum_z p(r \mid z) p(z \mid c(\mathbf{s}))$$

where $c(\mathbf{s})$ is the stimulus category and z is an indicator variable which models low, medium and high BOLD responses.

The multivoxel versions of the structural and semantic encoding model were generated from the single voxel models by mapping mean predicted responses by the single voxel models to a lower-dimensional representation using principal components analysis. That is, we have

$$p(\mathbf{r} \mid \mathbf{s}) \propto \exp\left(-\frac{1}{2}(\mathbf{r}' - \hat{\mathbf{r}}(\mathbf{s})))^T \mathbf{\Lambda}^{-1}(\mathbf{r}' - \hat{\mathbf{r}}(\mathbf{s})))\right)$$

where

$$\hat{\mathbf{r}}(\mathbf{s}) = \frac{P^T\hat{\boldsymbol{\mu}}(\mathbf{s})}{||P^T\hat{\boldsymbol{\mu}}(\mathbf{s})||}$$

and

$$\mathbf{r}' = \frac{P^T\mathbf{r}}{||P^T\mathbf{r}||}$$

with $\hat{\boldsymbol{\mu}}(\mathbf{s}) = (\hat{\mu}_1(\mathbf{s}), \ldots, \hat{\mu}_N(\mathbf{s}))^T$ the predicted mean responses for N voxels and P a PCA basis.

Different kinds of priors were explored:

- A flat prior with
$$p(\mathbf{s}) \propto 1$$

- A sparse Gabor prior with
$$p(\mathbf{s}) = \int_{\mathbf{a}} p(\mathbf{s} \mid \mathbf{a}) p(\mathbf{a}) \, d\mathbf{a}$$

where $p(\mathbf{s} \mid \mathbf{a}) = 1$ if $\mathbf{s} = UG^{-1}\mathbf{a} + \boldsymbol{\mu}_s$ and zero otherwise. I.e., it is a deterministic function which maps Gabor coefficients back to the stimulus via an unwhitening matrix U, a Gabor basis G and an offset $\boldsymbol{\mu}_s$. The prior on wavelet coefficients is given by $p(\mathbf{a}) = \prod_i p(a_i)$ where

$$p(a_i) = \frac{1}{2\beta_i} \exp\left(-\left|\frac{a_i - u_i}{\beta_i}\right|\right)$$

is a Laplace distribution.

- a natural image (empirical) prior with

$$p(\mathbf{s}) = \frac{1}{C} \sum_{i=1}^{C} \delta_{\mathbf{s}^{(i)}}(\mathbf{s})$$

where $\mathbf{s}^{(i)}$ is the i-th image in a set of natural images.

For the flat prior and the Gabor prior, reconstruction entails the use of a stochastic search algorithm. For the empirical prior it entails cycling through all possible images and picking the one which maximizes $p(\mathbf{r} \mid \mathbf{s})$.

4 Further Reading

Classical work on (the inversion of) encoding models can be found in [TDH+06, KNPG08]. Various extensions are possible, such as unsupervised learning of feature representations [vGdLH10], more complex encoding models [KWR+13] or extensions to movie decoding [NVN+11]. Lately, people have focused on how representations are affected e.g. by attention or scene statistics [ÇNHG13, SNG13]. Nice overviews can be found in [KG09, NKNG11, KK11].

References

[ÇNHG13] Tolga Çukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature Publishing Group*, 16(6):763–770, April 2013.

[KG09] K. N. Kay and J. L. Gallant. I can see what you see. *Nat. Neurosci.*, 12(3):245–246, 2009.

[KK11] Nikolaus Kriegeskorte and Gabriel Kreiman. *Visual Population Codes*. Toward a Common Multivariate Framework for Cell Recording and Functional Imaging. MIT Press (MA), October 2011.

[KNPG08] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying natural images from human brain activity. *Nature*, 452:352–355, 2008.

[KWR+13] K. N. Kay, Jonathan Winawer, Ariel Rokem, Aviv Mezer, and Brian A. Wandell. A Two-Stage Cascade Model of BOLD Responses in Human Visual Cortex. *PLoS Comput Biol*, 9(5):e1003079, May 2013.

[NKNG11] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant. Encoding and decoding in fMRI. *Neuroimage*, 56:400–410, 2011.

[NPK+09] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

[NVN^{+}11] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, pages 1–6, September 2011.

[SBHvG13] S Schoenmakers, M Barth, T. Heskes, and MAJ van Gerven. Linear reconstruction of perceived images from human brain activity. *Neuroimage*, 2013.

[SNG13] Dustin E Stansbury, Thomas Naselaris, and Jack L Gallant. Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. *Neuron*, August 2013.

[TDH^{+}06] B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, Jean-Baptiste Poline, D. Lebihan, and S. Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116, 2006.

[vGdLH10] Marcel van Gerven, F. P. de Lange, and T. Heskes. Neural decoding with hierarchical generative models. *Neural Comput.*, 22(12):3127–3142, 2010.