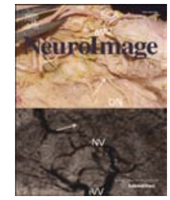
NeuroImage xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Linear reconstruction of perceived images from human brain activity

S. Schoenmakers [a,*], M. Barth [a], T. Heskes [b], M.A.J. van Gerven [a]

[a] Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
[b] Radboud University Nijmegen, Institute for Computing and Information Sciences, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Accepted 13 July 2013
Available online xxxx

Keywords:
fMRI analysis
Image reconstruction
Linear regression
Regularization

ABSTRACT

With the advent of sophisticated acquisition and analysis techniques, decoding the contents of someone's experience has become a reality. We propose a straightforward linear Gaussian approach, where decoding relies on the inversion of properly regularized encoding models, which can still be solved analytically. In order to test our approach we acquired functional magnetic resonance imaging data under a rapid event-related design in which subjects were presented with handwritten characters. Our approach is shown to yield state-of-the-art reconstructions of perceived characters as estimated from BOLD responses. This even holds for previously unseen characters. We propose that this framework serves as a baseline with which to compare more sophisticated models for which analytical inversion is infeasible.

Introduction

Neural encoding and decoding are two topics which are of key importance in contemporary cognitive neuroscience. Neural encoding refers to the representation of certain stimulus features by particular neuronal populations as reflected by measured neural responses. Conversely, neural decoding refers to the prediction of such stimulus features from measured brain activity. Encoding is a classical topic in neuroscience which has often been tackled using reverse correlation methods (Ringach and Shapley, 2004). Decoding has gained much recent popularity with the adoption of multivariate analysis methods by the cognitive neuroscience community (Haynes and Rees, 2006). While the first decoding studies focused exclusively on the prediction of discrete states such as object category (Haxby et al., 2001) or stimulus orientation (Kamitani and Tong, 2005), more recent work has focused on the prediction of increasingly complex stimulus properties, culminating in the reconstruction of the contents of perceived images (Kay et al., 2008; Miyawaki et al., 2008; Naselaris et al., 2009; Thirion et al., 2006; van Gerven et al., 2010) and even video clips (Nishimoto et al., 2011).

From the Bayesian point of view, encoding and decoding are intimately related via Bayes' rule where the probability $p(\mathrm{x}|\mathrm{y})$ of a stimulus x given a response y is expressed as the product of a likelihood term $p(\mathrm{y}|\mathrm{x})$ and aprior $p(\mathrm{x})$, up to some normalizing constant (Friston et al., 2008; Naselaris et al., 2010). The likelihood implements a forward model expressing how certain stimulus features are encoded by neural populations, as reflected by the measured response. The prior specifies how likely each stimulus is before observing any data. Stimulus reconstruction is then tantamount to inverse inference in a generative model. This approach has been advocated before. (Thirion et al., 2006) assumed that each voxel has a Gaussian receptive field which allows inversion of the generative model. (Naselaris et al., 2009), in contrast, used a complex forward model and did not perform the inversion explicitly. Instead they used an empirical prior which assigns a uniform probability to images in a predefined set and zero probability to all other images. This essentially allows the decoding to be performed by the forward model only, without the explicit need for inverse inference.

In this paper we present a general framework for decoding that expands on the ideas put forward in the aforementioned papers. Specifically, similar to (Naselaris et al., 2009), we assume that the forward model is given by the representation of an image in terms of a set of features, followed by a regularized linear regression. We then derive the formulas which, in conjunction with a suitable image prior, allow explicit decoding of the images as in (Thirion et al., 2006). The ideas presented in this paper extend earlier work on the decoding of discrete (binary) inputs to continuous (grey-scale) images (van Gerven et al., 2011) and improve on results presented in (van Gerven and Heskes, 2012). We focus on the reconstruction of multiple handwritten characters that have been presented to subjects using a rapid event-related design. We develop a linear Gaussian approach, analyze properties of the encoding models obtained in combination with different regularization approaches, and show that decodingperformance is remarkably good in this context. The simplicity of our framework makes it an ideal benchmark method with which to compare more sophisticated encoding and decoding methods.

Materials and methods

In this section, we will first explain the Gaussian decoding model and describe how parameters of the model are estimated in the presence of

* Corresponding author at: Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Donders Centre for Cognition, P. O. Box 9104, 6500 HE Nijmegen, The Netherlands. Fax: +31 24 36 52728.
E-mail address: s.schoenmakers@donders.ru.nl (S. Schoenmakers).

1053-8119/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.neuroimage.2013.07.043

different regularization methods. Subsequently, we present the functional magnetic resonance imaging (fMRI) experiment which has been conducted in order to validate our approach. Finally, we describe the analyses which have been performed using our approach, based on acquired fMRI data.

Gaussian decoding

Let (x,y) denote a stimulus–response pair, say, an image $\mathrm{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$, characterized by its pixel values x_i, and the associated measured response vector $\mathrm{y} = \left(y_1, \ldots, y_q\right)^\top \in \mathbb{R}^q$. Without loss of generality, both the stimulus and the response are assumed to be standardized to have zero mean and unit standard deviation. In this paper we are interested in decoding the most probable image x from the BOLD response y:

$$\hat{\mathrm{x}} = \arg\max_{\mathrm{x}}\{p(\mathrm{x}|\mathrm{y})\}. \quad (1)$$

In previous work, we have shown how this problem can be solved in a discriminative way using a partial least squares approach (van Gerven and Heskes, 2010). Here, we focus on the generative setting, where we wish to use the equivalent formulation:

$$\hat{\mathrm{x}} = \arg\max_{\mathrm{x}}\{p(\mathrm{y}|\mathrm{x})p(\mathrm{x})\}. \quad (2)$$

In order to compute this maximum a posteriori (MAP) estimate, we require an image prior $p(\mathrm{x})$ and a forward model $p(\mathrm{y}|\mathrm{x})$. In Naselaris et al. (2009), this problem was solved by assuming an empirical prior that assigned uniform probability to any of n possible images and zero probability to the remaining images. The decoding problem could thus be solved by identifying that image which gave the largest likelihood. Here, in contrast, we solve the decoding problem without relying on a restricted subset of possible images. Our approach is related to the work presented in Thirion et al. (2006), but we make weaker assumptions on the form of the forward model and the image prior. Particularly, we assume that the forward model is given by a regularized linear Gaussian model and the image prior is given by a multivariate Gaussian.

We assume that the forward (encoding) model is given by a multiple-output linear regression model, such that

$$\mathrm{y} = \mathrm{B}^\top \mathrm{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0; \Sigma), \quad (3)$$

with regression coefficients $\mathrm{B} = (\mathrm{b}_1, \ldots, \mathrm{b}_q)$ and covariance matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)$. It follows that the forward model can be written as a multivariate Gaussian

$$\begin{aligned} p(\mathrm{y}|\mathrm{x}) &= \mathcal{N}\left(\mathrm{y}; \mathrm{B}^\top \mathrm{x}, \Sigma\right) \\ &\propto \exp\left(-\frac{1}{2}\mathrm{y}^\top \Sigma^{-1}\mathrm{y} + \left(\mathrm{B}\Sigma^{-1}\mathrm{y}\right)^\top \mathrm{x} - \frac{1}{2}\mathrm{x}^\top \mathrm{B}\Sigma^{-1}\mathrm{B}^\top \mathrm{x}\right), \end{aligned} \quad (4)$$

where (4) is its canonical form representation. We further assume that the image prior is given by a zero-mean multivariate Gaussian of the form:

$$p(\mathrm{x}) \propto \exp\left(-\frac{1}{2}\mathrm{x}^\top \mathrm{R}^{-1}\mathrm{x}\right), \quad (5)$$

with covariance matrix R.

Given $p(\mathrm{y}|\mathrm{x})$ and $p(\mathrm{x})$, we can proceed with decoding. That is, we are interested in computing the mode of the distribution $p(\mathrm{x}|\mathrm{y})$. Dropping terms in Eq. (4) not depending on x, this yields

$$p(\mathrm{x}|\mathrm{y}) \propto \exp\left(\left(\mathrm{B}\Sigma^{-1}\mathrm{y}\right)^\top \mathrm{x} - \frac{1}{2}\mathrm{x}^\top\left(\mathrm{R}^{-1} + \mathrm{B}\Sigma^{-1}\mathrm{B}^\top\right)\mathrm{x}\right). \quad (6)$$

This is recognized as a multivariate Gaussian in canonical form with mean $\mathrm{m} \equiv \mathrm{QB}\Sigma^{-1}\mathrm{y}$ and covariance $\mathrm{Q} = (\mathrm{R}^{-1} + \mathrm{B}\Sigma^{-1}\mathrm{B}^\top)^{-1}$. It immediately follows that

$$\hat{\mathrm{x}} = \mathrm{m} = \left(\mathrm{R}^{-1} + \mathrm{B}\Sigma^{-1}\mathrm{B}^\top\right)^{-1}\mathrm{B}\Sigma^{-1}\mathrm{y}, \quad (7)$$

since the mode of a Gaussian distribution is given by its mean. Eq. (7) is a standard result obtained in Bayesian linear regression (Bishop, 2006). Note further that the covariance matrix Q captures the posterior variance of the image reconstructions.

For large images, computing (7) may be prohibitively expensive since it requires inversion of a $p \times p$ covariance matrix, where p is the number of pixels. In that case, we can make use of the matrix inversion lemma to obtain

$$\hat{\mathrm{x}} = \left(\mathrm{R} - \mathrm{RB}\left(\Sigma + \mathrm{B}^\top \mathrm{RB}\right)^{-1}\mathrm{B}^\top \mathrm{R}\right)\mathrm{B}\Sigma^{-1}\mathrm{y}. \quad (8)$$

This requires the inversion of a $q \times q$ matrix, where q is the number of voxels. Which formulation is most convenient depends on the problem at hand.

Parameter estimation

In order to be able to use our model for decoding, we first need to estimate the parameters of the prior and the forward model. We assume that training data $D = \{\mathrm{X}, \mathrm{Y}\}$ has been collected, where X is an $N \times p$ matrix, such that x_{ij} denotes the value of pixel j for the i-th image, and Y is an $N \times q$ matrix, such that y_{ij} denotes the response of voxel j to the i-th image. Furthermore, we assume that an independent set of images Z has been collected, which will be used to estimate the image prior. We use notation m^i and m_j to denote the i-th row and j-th column of a matrix M, respectively.

The parameters of the image prior are estimated from an independent large set of images $\{\mathrm{z}^n\}_{n=1}^M$, which are standardized to have zero mean and unit variance. In the linear Gaussian case, the required covariance matrix for the prior is given by

$$\mathrm{R} = \frac{1}{N-1}\sum_n \mathrm{z}^n (\mathrm{z}^n)^\top. \quad (9)$$

For the forward model, it is easy to see that the parameters for each of the responses can be estimated independently due to the diagonality of Σ. That is, for each response k, we need to solve an independent linear regression problem.Since we are dealing with the small N, large p case, regression coefficients need to be properly regularized. Let $\left(\hat{\mathrm{b}}_k, \hat{\sigma}_k^2\right)$ denote the estimates of the vector of regression coefficients and variance for voxel k. This estimate takes the form[1]

$$\left(\hat{\mathrm{b}}_k, \hat{\sigma}_k^2\right) = \arg\min_{\mathrm{b}, \sigma^2}\left\{\frac{1}{2N\sigma^2}\|\mathrm{y}_k - \mathrm{Xb}\|_2^2 + R_{\lambda,\alpha,\mathrm{G}}(\mathrm{b})\right\} \quad (10)$$

where

$$R_{\lambda,\alpha,\mathrm{G}}(\mathrm{b}) = \lambda\left(\alpha\|\mathrm{b}\|_1 + (1-\alpha)\frac{1}{2}\mathrm{b}^\top \mathrm{Gb}\right) \quad (11)$$

is a regularization term which, following Grosenick et al. (2013), we refer to as the graph-constrained elastic net (graphnet for short) regularizer.

The graphnet regularizer contains three parameters that can be set to obtain different models: λ, α and G. The regularization parameter λ determines the amount of regularization. The mixing parameter α determines the relative contribution of the ℓ_1 regularization term, which

[1] We divide by N to make the regularization strength for a fixed λ independent of N.

Please cite this article as: Schoenmakers, S., et al., Linear reconstruction of perceived images from human brain activity, NeuroImage (2013), http://dx.doi.org/10.1016/j.neuroimage.2013.07.043