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

Sanne Schoenmakers [a,*], Markus Barth [a], Tom Heskes [b], Marcel 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 a prior $p(\mathrm{x})$, 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 decoding performance 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 Elsevier Inc. All rights reserved.
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 $\mathrm{y} = (y_1, \ldots, y_q)^\top \in \mathbb{R}^q$. 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:

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

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:

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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 $\left(\widehat{\mathrm{b}}_k, \widehat{\sigma}_k^2\right)$ denote the estimates of the vector of regression coefficients and variance for voxel k. This estimate takes the form[1]

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

where

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

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.

induces sparseness, and the ℓ_2 regularization term, which induces shrinkage. Different kinds of regularization are achieved using different choices of α and the coupling matrix G. If we set $\alpha = 1$ we obtain the lasso (ℓ_1) regularizer (Tibshirani, 1996). If we set $\alpha = 0$ and $\mathrm{G} = \mathrm{I}_p$, where I_p is the $p \times p$ identity matrix, we obtain the ridge (ℓ_2) regularizer (Hoerl and Kennard, 1970). If we set $0 < \alpha < 1$ and $\mathrm{G} = \mathrm{I}_p$, we obtain the elastic net regularizer (Carroll et al., 2009; Zou and Hastie, 2005). If we set $0 \leq \alpha < 1$ and use non-diagonal G, we obtain the graphnet regularizer, which induces a coupling between features. If, in the latter case, $\alpha = 0$, only the ridge term remains. In that case, we use graphridge to refer to the resulting regularizer.

In case of a non-diagonal coupling matrix, we assume G to be the graph Laplacian L, which is a matrix with $l_{ij} = -1$ for each $i \neq j$ that are defined to be neighboring image pixels and l_{ii} equal to the number of neighbors of node i (Grosenick et al., 2013). Note that the graphnet regularizer can be interpreted in probabilistic terms since $\log p(\mathrm{b}) \propto -R_{\lambda,\alpha,\mathrm{G}}(\mathrm{b})$. Hence, the prior on the regression coefficients is given by

$$p(\mathrm{b}) \propto \prod_i \exp(-\lambda\alpha|b_i|) \prod_j \exp\left(-\lambda(1-\alpha)\frac{1}{2}\sum_{i\sim j} b_i G_{ij} b_j\right) \tag{12}$$

which is a convex combination of a global Laplacian density and a local Markov Random Field prior. Hence, the graphnet regularizer expresses our prior beliefs about the model coefficients being globally sparse yet locally structured (Grosenick et al., 2013).

In order to estimate the regularized regression coefficients b_k, we need to solve the following minimization problem:

$$\widehat{\mathrm{b}}_k = \arg\min_{\mathrm{b}} \left\{\frac{1}{2N}\|\mathrm{y}_k - \mathrm{Xb}\|_2^2 + R_{\lambda,\alpha,\mathrm{G}}(\mathrm{b})\right\}. \tag{13}$$

We use different strategies depending on the used regularizer. For ridge and graphridge regression we can simultaneously estimate regression coefficients for all voxels k in closed form using

$$\widehat{\mathrm{B}} = \left(\mathrm{X}^\top\mathrm{X} + \widetilde{\mathrm{G}}\right)^{-1}\mathrm{X}^\top\mathrm{Y}, \tag{14}$$

with $\widetilde{\mathrm{G}} = N\lambda\mathrm{G}$. Alternatively, we can make use of a kernel formulation, which replaces Eq. (14) by

$$\widehat{\mathrm{B}} = \widetilde{\mathrm{G}}^{-1}\mathrm{X}^\top\left(\mathrm{X}\widetilde{\mathrm{G}}^{-1}\mathrm{X}^\top + \mathrm{I}\right)^{-1}\mathrm{Y}, \tag{15}$$

with $N \times N$ kernel matrix $\mathrm{K} = \mathrm{XX}^\top$, requiring inversion of an $N \times N$ matrix rather than a $p \times p$ matrix (Hastie et al., 2008). See Appendix A for a derivation.

For lasso, elastic net and graphnet regression we minimize Eq. (13) using a slight generalization of an efficient coordinate descent algorithm (Friedman et al., 2010), applied to each voxel k independently. In order to estimate λ we use a nested five-fold cross-validation and choose for each voxel k that model which minimizes the residual variance $v_k = \mathrm{var}(\mathrm{Xb}_k - \mathrm{y}_k)$ on hold-out data. For ridge regression, we sample λ in the range $(10^5, 10^{-5})$ on a log scale. In other cases, we sample different values of λ by starting at $\lambda_{\max}$ at which point the first variable enters the model (see Appendix B) and continuing until $\lambda \leq 0.05 \cdot \lambda_{\max}$. After an optimal value of λ was selected, parameters were re-estimated using all training data. Based on a preliminary analysis which considered data for Subject 3 only, the parameter α was set to 0.005 for the elastic net regularizer and to 0.05 for the graphnet regularizer. For smaller values of α, the graphnet model became cumbersome to estimate due to slow convergence. The parameter $\widehat{\sigma}_k^2$ introduced in Eq. (10) was taken to be the residual variance v_k on the training data, computed for the optimal model selected during nested cross-validation.

fMRI experiment

Participants

Three healthy native Dutch-speaking participants took part in the study. All participants gave written consent according to the institutional guidelines set forth by the local ethics committee (CMO region Arnhem–Nijmegen, The Netherlands) before the experiment. The participants were not paid for participation.

Stimuli

The stimuli consisted of grayscale handwritten characters on a black background (van der Maaten, 2009). The character database consists of 40,000 handwritten characters by 250 writers. The images in the database were rescaled and centered so they fill the canvas. Six characters were selected: B, R, A, I, N, and S. For each character, 60 unique instances were centrally presented during the experiment. The size of the images was 9 × 9° degrees of visual angle (56 × 56 pixels). A central white square served as a fixation point (0.2° of visual angle). The images were shown as flickering stimuli (200 ms ON, 200 ms OFF) for one second, followed by three seconds of black background. The fixation point was present at the center of the screen throughout the whole experiment. A total of 360 different characters were shown and this was repeated once, giving a total of 720 presented stimuli. Stimuli were repeated to get a better estimate of the BOLD response to individual character instances(see FMRI data preprocessing section).

Procedure

Participants were asked to focus on the fixation point and to respond with a button press when the fixation point changed color from dark gray to light blue in order to keep their vigilance. The fixation point changed color once every six stimuli on average. Changes were presented randomly but evenly over the full length of the experiment and counterbalanced over characters. The characters were shown in pseudo-random order where instances of all six letters were reshuffled in order to prevent long repetitions of the same letter. The experiment lasted for 50 min. with a self-paced rest period in the middle. After the experiment, a structural scan was made. Subsequently, or in a next session, a functional localizer for the visual cortex was employed. The stimulus shown in the functional localizer was a rotating checkerboard wedge for polar retinotopy, which was presented in four blocks of five minutes.

FMRI data acquisition

Imaging was conducted at the Donders Institute for Brain, Cognition and Behaviour (Nijmegen, The Netherlands). The functional images were collected with a Siemens Trio 3 T MRI system (Siemens, Erlangen Germany) with an EPI sequence using a 32 channel head coil (TR = 1.74 s, TE = 30 ms, GRAPPA acceleration factor 3, 83° flip angle, 30 slices in ascending order, voxel size 2 × 2 × 2 mm). Head movement was restricted with foam cushions and a tight strip of tape over the forehead. After functional imaging, a structural scan was acquired using an MPRAGE sequence (TR = 2.3 s, TE = 3.03 ms, voxel size 1 × 1 × 1 mm, 192 sagittal slices, FoV = 256 mm). In a separate session, the functional localizer data was acquired, again using an EPI sequence (TR = 2 s, TE = 30 ms, 83° flip angle, 33 slices in ascending order, voxel size 2 × 2 × 2 mm, FoV = 192 mm). During acquisition an eye tracker was used to verify if participants were fixating their gaze.

FMRI data preprocessing

With the use of SPM8 software (Wellcome Department of Imaging Neuroscience, University College London, UK), the functional volumes were reconstructed, realigned to the first scan of the session and slice time corrected. Participants moved less than 0.5 mm across the sessions. For each unique stimulus, which was presented twice to the subject, the response of each voxel to a stimulus was computed using a general linear model (GLM). The design matrix of the GLM was given

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

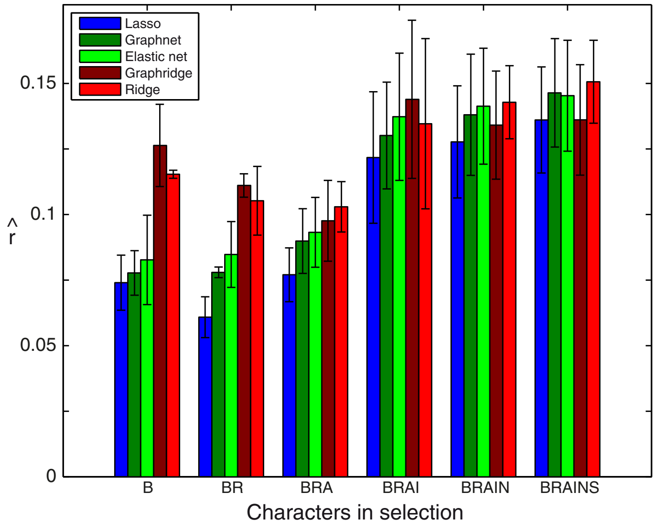

Fig. 1. Encoding performance quantified in terms of summed explained variance for models that employ different regularizers and varying input data, averaged over all participants. Error bars indicate standard error of the mean.

by one regressor encoding the two stimulus repetitions, one regressor encoding all other stimuli, as well as nuisance regressors that encoded movement parameters and drift terms, similar to the approach presented in Mumford et al. (2011). The design matrix was convolved with the canonical hemodynamic response function. The voxel response for each stimulus was given by the beta estimate which was normalized for each voxel. Freesurfer software was used together with functional localizer data in order to isolate voxels belonging to visual area V1 using well-established methods (DeYoe et al., 1996; Engel et al., 1997; Sereno et al., 1995).

Empirical validation

In order to validate our approach, we used the acquired fMRI data to learn encoding and decoding models. In order to examine how regularities in the input data influence encoding and decoding results, we trained models for different subsets of input data. Six selections were chosen: one character (B), two characters (B, R), three characters (B, R, A), four characters (B, R, A, I), five characters, (B, R, A, I, N) and six characters (B, R, A, I, N, S). Data were randomly split into training data and test data. We used 80% of the data (48 exemplars per letter class) for training our models and 20% of the data for testing our models (12 exemplars per letter class). Our goal was to compare encoding performance and decoding performance computed for the test data while different regularization approaches were used to estimate the encoding models from the training data. Specifically, we compared ridge, lasso, elastic net, graphnet and graphridge regression.

Encoding

To compare encoding performance between models we calculated the explained variance. Explained variance reflects how well the model predicts the real data. For each voxel k, explained variance was calculated in accordance to Michel et al. (2011):

$$\hat{r}_k = \left(\mathrm{var}(y_k) - \mathrm{var}(y_k - \hat{y}_k)\right) / \mathrm{var}(y_k), \qquad (16)$$

where y_k is the actual response for voxel k and $\hat{y}_k$ the estimate thereof, based on Eq. (3). To facilitate comparison, voxels were sorted according to explained variance and we only used 150 voxels with highest explained variance per model to determine encoding performance. We use $\hat{r}$ to refer to the average explained variance per model. A binomial test comparing the explained variance of all sorted voxels between models for which $\hat{r}_k > 0$ served to show which regularizer performed best.

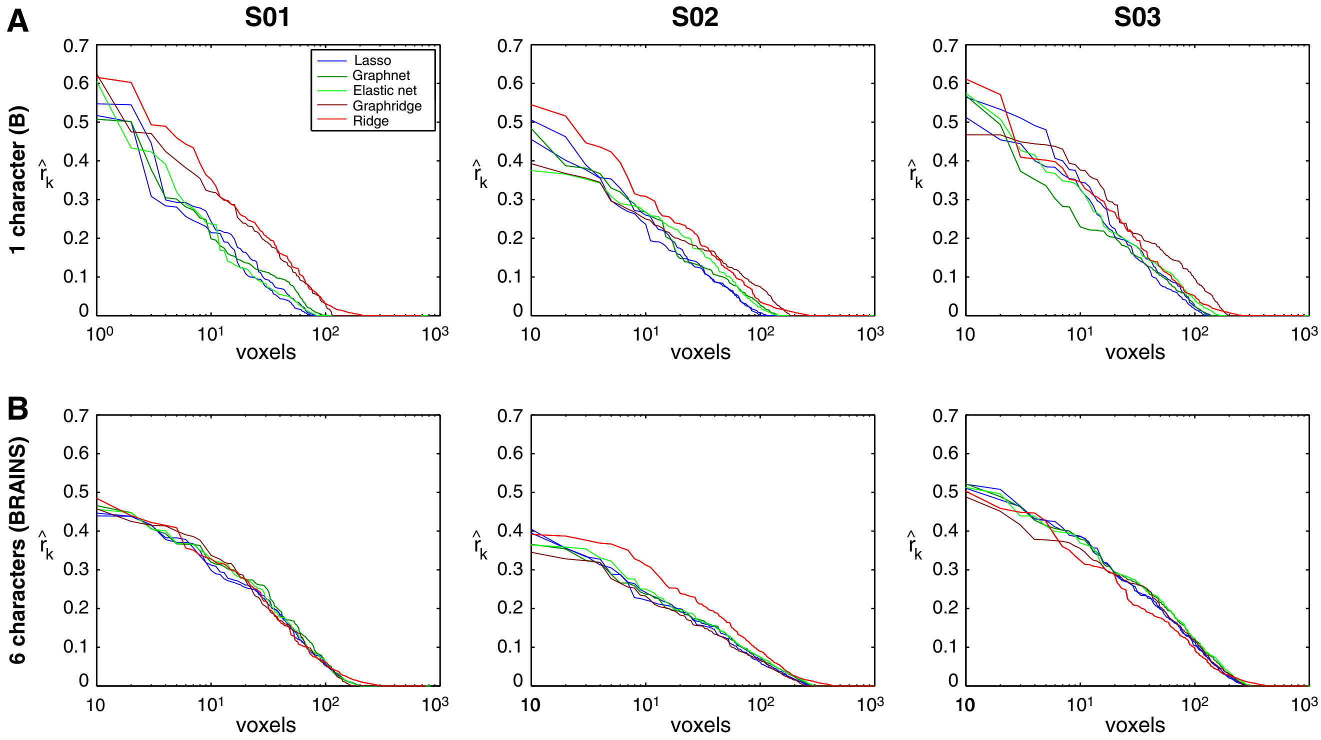

Fig. 2. Voxels sorted according to explained variance for the three participants on a logarithmic scale. Panels A and B show the sorted explained variance obtained using different regularizers for input data consisting of one character or all six characters, respectively.

Explained variance was mapped back to the primary visual cortical surface in order to determine which voxel responses were predicted best. Model parameters were visualized by taking the vector of regression coefficients b_k and reshaping it to a 56 × 56 pixel image, which we refer to as the (linear) filter for voxel k. Such a filter shows to which pixels the voxel is responsive. Also, the filters provide insight in the sparseness and smoothness of parameter estimates under different regularization schemes.

Decoding

For the decoding analysis, we estimated a Gaussian image prior based on 700 images per character which had not been used in the experimental run but came from the same handwritten characters database. Subsequently, we used the mode of the posterior density, computed using Eq. (7), to produce image reconstructions. For decoding, only those voxels were used whose explained variance exceeded zero. This was implemented by setting all filters of the remaining voxels to zero such that they exerted no influence on the reconstruction. Reconstruction quality was measured in terms of the correlation between an original and its reconstruction.

A binomial test comparing the correlations obtained for different models served to show which regularizer performed best. To quantify whether an observed mean correlation for the twelve reconstructed images per character was significantly better than chance-level performance, we estimated a p-value based on a permutation test which compared the observed mean correlation with the mean correlation computed for random reconstructions. These random reconstructions were generated by sampling from the image prior. The rationale for this significance test is that, if the BOLD responses convey information, then the informed reconstructions should be closer to the true images than reconstructions obtained by sampling from the prior.

Finally, correlation matrices were estimated visualizing the correlation between all original stimuli and all reconstructions. These matrices were sorted to show the rank of the reconstruction that belonged to the original relative to all other reconstructions. Ranks below the diagonal indicate that the reconstructions matched with their originals compared to random guessing.

Results

In the following, we discuss results obtained for models that employ different input data or different regularizers. We separately describe the outcomes of the encoding analyses and decoding analyses that have been performed.

Encoding analysis

Fig. 1 depicts the summed explained variance for all sets of characters, averaged over subjects. A trend can be observed that encoding performance increases for all regularizers when the size of the input data increases. For the ridge and graphridge regularizers, the increase is not quite as dramatic as for the other three regularizers, as they already perform quite well given limited input data.

Fig. 2 shows the explained variance for all voxels on a logarithmic scale, sorted from highest to lowest explained variance, for all three participants. Panel 2A shows results obtained when using instances

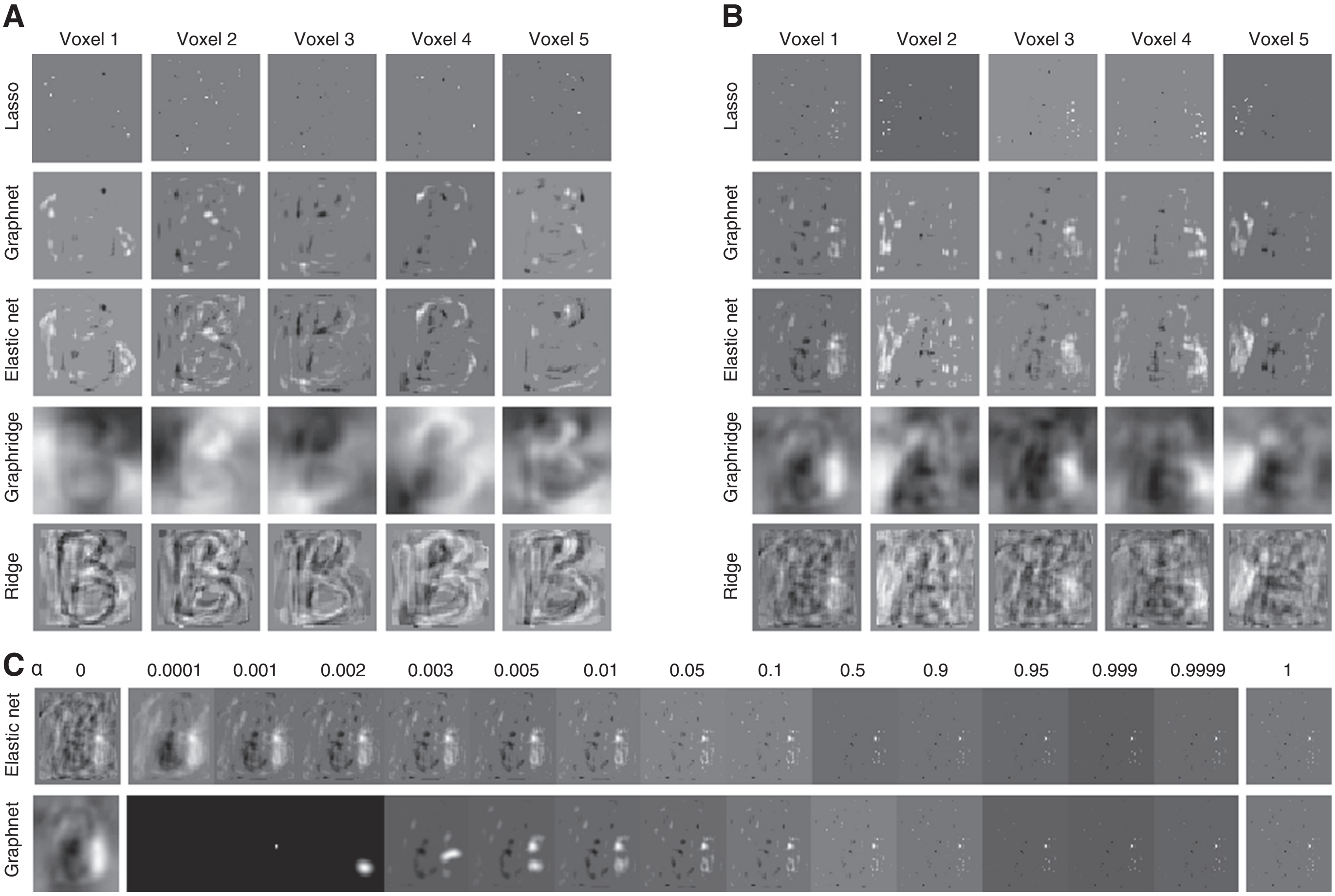

Fig. 3. Examples of filters estimated for voxels with high explained variance from Subject 3. Filters are individually scaled to emphasize the contrast between high and low values. (A) Filters for the small dataset. (B) Filters for the large dataset. (C) Demonstration of how the filter for one of the voxels changes as a function of the mixing parameter α when models were trained on the large dataset.

belonging to one character as input data. Panel 2B depicts results when using instances of all six characters as input data. For convenience, in the remainder, we refer to these datasets as the small dataset and large dataset, respectively. For the small dataset, fewer voxels with above-zero explained variance remained compared to the large dataset. At the same time, maximal explained variance was consistently higher for the small dataset as compared with the large dataset. For the small dataset, ridge and graphridge regularizers outperformed the other regularizers, since explained variance was consistently higher. Furthermore, for these two regularizers, more voxels contributed to the model, as can be seen in the tail of the figures. For the large dataset, differences are less obvious. Still, significance tests for the three participants show that the ridge regularizer significantly outperforms lasso, graphnet and elastic net regularizers for the large dataset as well ($p < 10^{-4}$ for S01, S02 and S03, Bonferroni corrected for number of comparisons). Furthermore, graphnet and elastic net regularizers score significantly higher on explained variance than lasso and graphridge regularizers ($p < 10^{-4}$ for all participants, Bonferroni corrected). The superior performance of some regularizers on the large dataset seems to be driven by the high number of voxels in the tail that still add some explained variance to the model.

Fig. 3 depicts examples of voxel-wise linear filters b_k obtained for Subject 3. Fig. 3A shows filters obtained with the small dataset whereas Fig. 3B shows filters obtained with the large dataset. The sparseness of filters estimated by the lasso, elastic net and graphnet regularizers is clearly visible, in contrast to filters obtained with graphridge and ridge regularizers. Furthermore, graphnet and graphridge regularizers lead to filters that smooth regression coefficients between neighboring pixels. Moreover, the filters for small input data clearly reflect the structure present in the input data. The filters for the large dataset seem to be less tuned to a single character though characteristics of the input data are still visible. Fig. 3C depicts for one voxel how the filters change as a function of the mixing parameter α when using the elastic net regularizer and the graphnet regularizer. Note that the other regularizers are included here as special cases with $\alpha = 0$ or $\alpha = 1$. Clearly, the tradeoff between sparseness and smoothness is strongly dependent on the choice of α.

Fig. 4 shows the projection of the explained variance on an inflated brain for Subject 3 when training on either the small or large dataset.

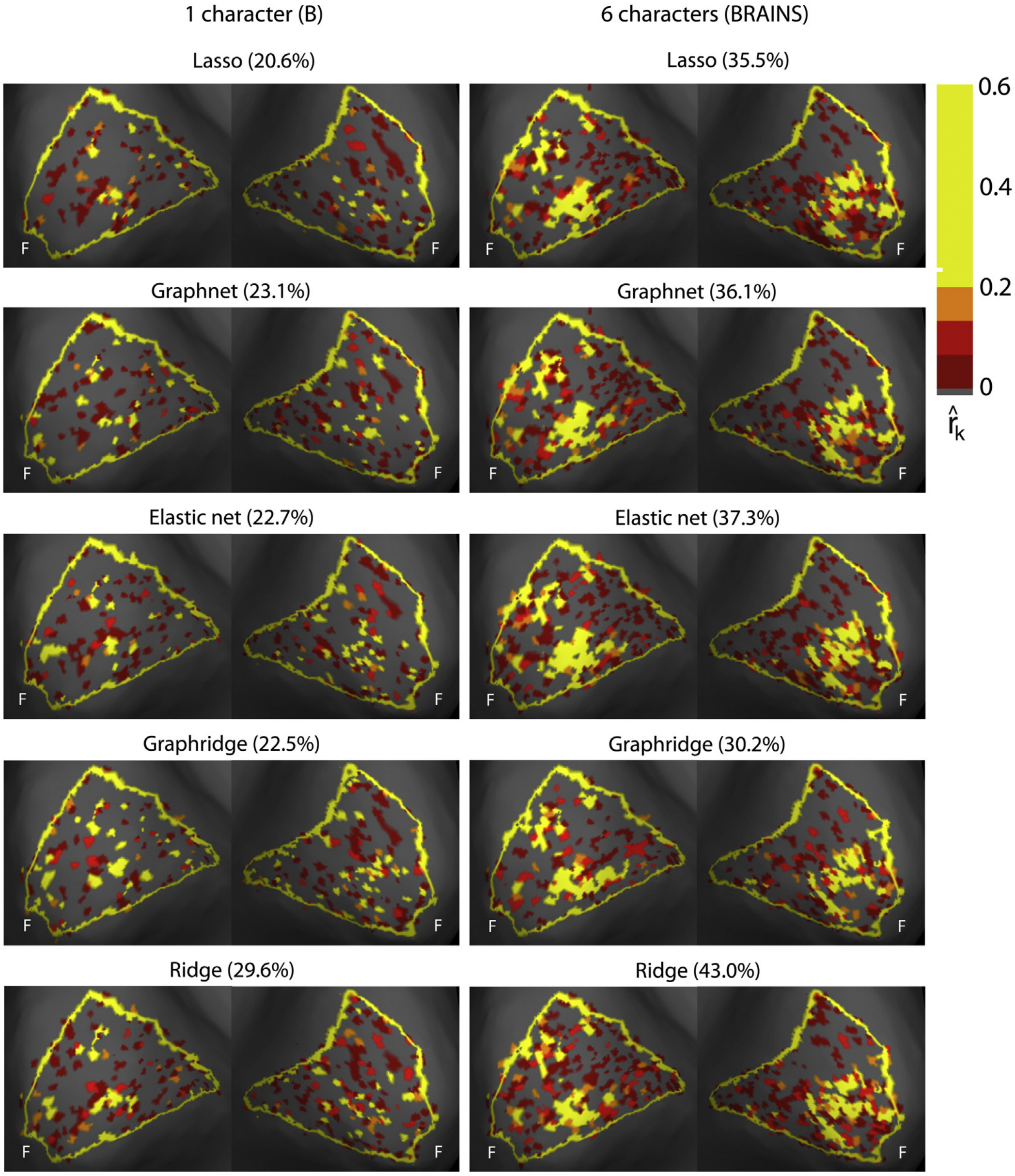

Fig. 4. Explained variance per voxel, plotted on the inflated visual cortex of the left and right hemisphere for Subject 3. The yellow border outlines visual area V1. The letter 'F' indicates the location of the fovea. Numbers between parentheses show the percentage of selected voxels whose explained variance was above zero, averaged over subjects.

V1 is indicated by a yellow contour and plotted for left and right hemispheres. Evidently, a smaller number of voxels is included when models are trained on the small dataset compared to when models are trained on the large dataset. Strongest contributions seem to come from foveal rather than peripheral voxels. This observation becomes more pronounced when a large dataset is used. Finally, note that voxels with high explained variance tend to cluster together.

Decoding analysis

Fig. 5 shows the average correlation ρ between the originals and their reconstruction for all models for different sizes of input data averaged over all participants. Overall, the graphnet regularizer seems to perform best, but the differences in reconstruction quality of the different regularizers are negligible. When the dataset increases in size, the graphridge regularizer performs less well compared to the other regularizers. When training on all characters, the elastic net, graphnet and lasso regularizers were shown to outperform the graphridge regularizer in terms of decoding performance ($p < 10^{-4}$, Bonferroni corrected for number of comparisons). The ridge regularizer also performed less well than other regularizers although differences were marginally insignificant.

Fig. 6A depicts all reconstructions for participant S03 for the small dataset that contains only presentations of the character ‘B’. The obtained reconstructions are all unique and share certain characteristics of their corresponding original images. Also, reconstructions 7, 11 and 12 seem to contain two superimposed characters. This might be due to the fact that the BOLD response was estimated using two representations of the same character. An alternative explanation is that the reconstructions represent two consecutively presented characters which both modulate the BOLD response due to the sluggishness of the hemodynamic response function. Nevertheless, reconstructions are of high quality in general. Optimal reconstruction performance for the small dataset was achieved by graphnet regression (cf. Fig. 5). Fig. 6B depicts reconstructions of different letters when models were trained on the large dataset containing all characters. All regularizers produce good reconstructions of the originals. These results demonstrate that instances belonging to different letter classes are easily distinguished.

The question remains to what extent reconstructions rely on the contribution by the likelihood versus the prior. In order to address this question, we also estimated reconstructions using either the likelihood or the prior. These reconstructions correspond to the maximum likelihood estimate (MLE) given by the mode of Eq. (4) and to the mode of Eq. (5), respectively. Fig. 6C shows reconstructions based on the likelihood, based on the prior, and based on both. A comparison of the decoding performance shows that reconstruction quality heavily depends on both the information conveyed by the likelihood as well as the constraints imposed by the image prior.

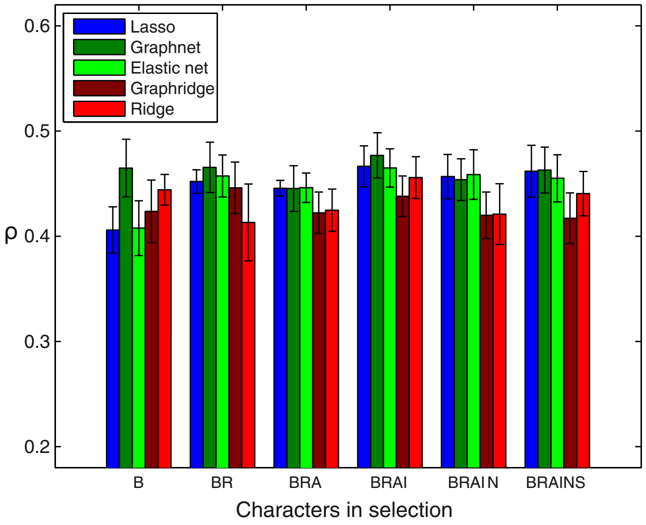

Fig. 5. Decoding performance quantified in terms of the average correlation between original and reconstructed images for all regularizers and for different sizes of the input data, averaged over participants. Error bars indicate standard error of the mean.

In order to examine the quality of individual reconstructions, Fig. 7A shows the correlation matrices for all regularizers for the large dataset containing all six characters. Each correlation matrix shows the correlation coefficient between all originals and all reconstructions. The block diagonal structure of the correlation matrices reflects the fact that reconstructions tend to look like within-class exemplars. Note also that some of the letter classes are more easily confused, notably ‘B’ versus ‘S’ and ‘R’ versus ‘A’.

Fig. 7B shows the sorted correlation matrices. That is, for each original the reconstructions are sorted according to their rank. The rank of the correct reconstruction is indicated in dark red. Clearly, reconstructions outperform random reconstructions. When comparing reconstructions obtained with the large dataset with samples drawn from the image prior we found that these were significantly better than chance ($p < 10^{-4}$, Bonferroni corrected for number of comparisons). We also compared reconstructions using our approach with those obtained using an empirical prior as employed by Naselaris et al. (2009). This procedure amounts to selecting that image in the image prior which has maximal likelihood given the observed BOLD response. The rank of the correct reconstruction when using this approach is indicated in dark blue. These results show that the empirical prior approach is outperformed by the explicit inversion scheme derived in this paper.

The final question we address is how well we can reconstruct images belonging to an image category which has not been observed during training. That is, how well do we generalize to previously unseen image categories? In order to address this question, the graphnet model was trained six times on BOLD data associated with five out of six letter classes for Subject 3. Subsequently, these models were used to reconstruct letters belonging to the sixth remaining letter class. During reconstruction, we either used a prior for the five letter classes that were presented during training or a prior which also incorporated the sixth letter class present during testing. Note that both cases only used BOLD data acquired for five letter classes. Fig. 8 shows reconstructions obtained using either the five-letter or the six-letter prior. Decoding performance averaged over letter subsets is $\rho = 0.40$ (± 0.02 SEM) for the five-letter prior versus $\rho = 0.46$ (± 0.02 SEM) for the six-letter prior.

Discussion

We introduced a linear Gaussian framework for reconstructing perceived images from measured neural responses. Results show that high-quality reconstructions can be obtained by inverting properly regularized encoding models. Reconstructions relied on the use of encoding models that explain much of the variance in BOLD responses acquired under a rapid event-related design. While ridge regression performed best in terms of encoding performance, graphnet regression performed best in terms of decoding performance. Estimated filters that model how visual stimulation leads to observed BOLD response were shown to rely heavily on the employed regularizers (cf. Fig. 5). Decoding relied on computing a MAP estimate as given by the mean of a multivariate Gaussian which incorporates both prior and likelihood terms.

The high quality of the reconstructions was both driven by information contained in the estimated responses as well as by the employed Gaussian image prior (see Fig. 6C). Our comparison between models trained using one letter class up to all six letter classes present in the data showed that encoding performance tends to increase for larger datasets. In contrast, decoding performance was shown to be quite stable for different subsets of the input data. Interestingly, decoding performance remained good even when reconstructions were made for

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

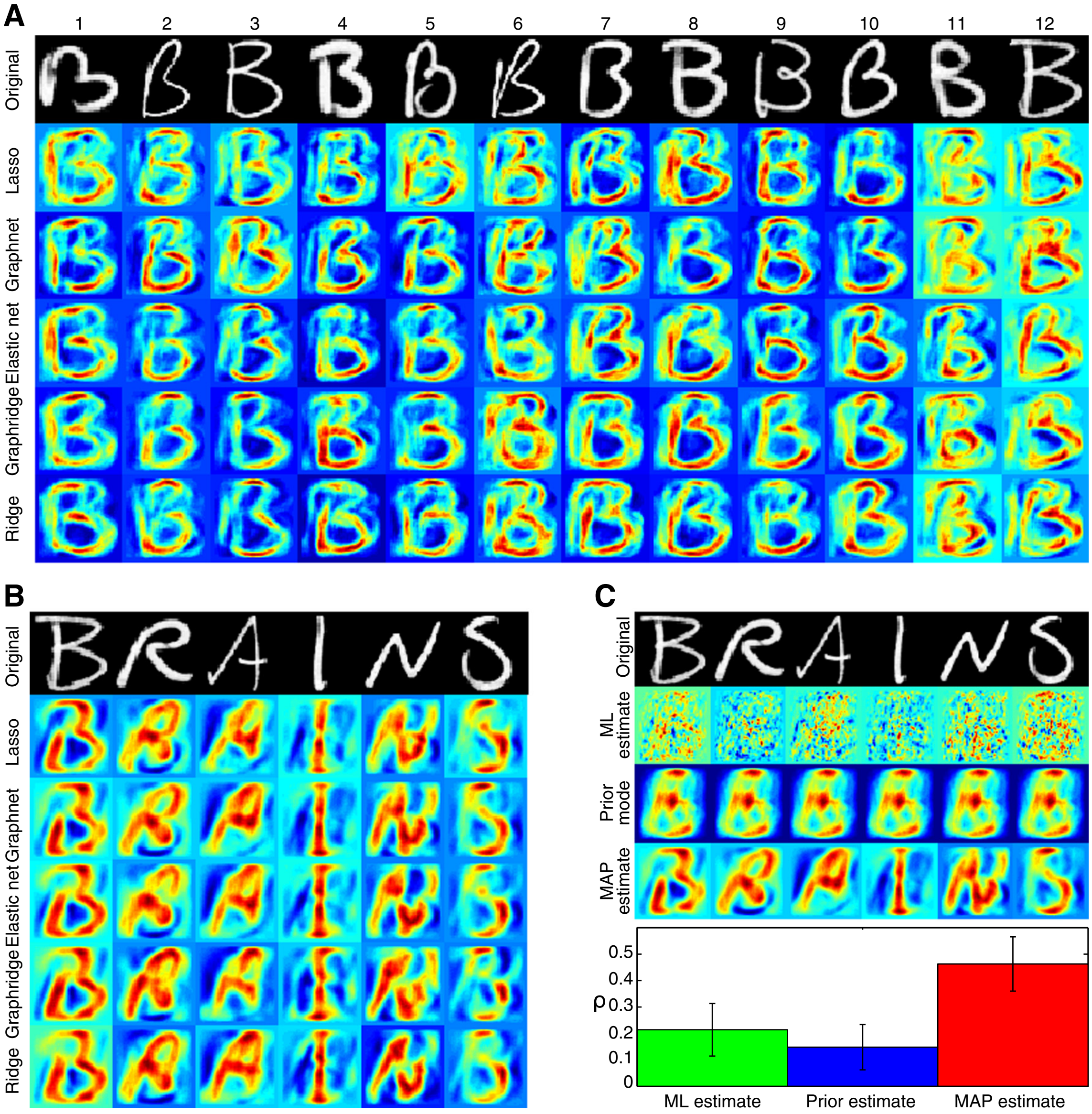

Fig. 6. Reconstructions produced by the Gaussian decoding approach. (A) All reconstructions from the test set for models trained on the small dataset containing only presentations of the character 'B' for participant S03. (B) A sample of reconstructions obtained with models trained on the large dataset. (C) Maximum likelihood estimates obtained when using the graphnet regularizer versus samples obtained using the prior. For the MLE, a small amount of regularization was used in order to prevent numerical problems. The MAP estimates that integrate likelihood and prior are shown as well. The bar graph shows decoding performance averaged over participants. Error bars indicate standard error of the mean.

previously unseen letters, especially when the prior was extended to take the new letter class into account (cf. Fig. 8). This suggests that generic decoders may be trained on arbitrary input data and tailored to the specifics of a particular dataset by adjusting the prior.

An important observation is that the learned filters do not only reflect the receptive field of individual voxels but also statistical regularities that are present in the input data. This holds more strongly for smaller datasets (see Fig. 5). This behavior can be understood by realizing that, even if a voxel responds selectively to one location in the visual field, other locations in the visual field could be active simultaneously due to regularities in the input data. For example, the letter 'I' will tend to activate the vertical midline. For this reason, locations in the visual field that do not fall within a voxel's receptive field but are correlated with locations that do fall within the receptive field are still able to predict voxel responses. The emergence of filters driven by input statistics will hold for any dataset whose features are not statistically independent, including natural images (Simoncelli and Olshausen, 2001). This implies that learned filters should be interpreted with care.

In this work, we compared several different regularizers in terms of encoding and decoding performance. All regularizers yielded high-quality reconstructions. Overall, the graphnet regularizer tended to perform best in terms of decoding performance, even though it performed less well in terms of encoding, especially for a small amount of input data (cf. Figs. 1 and 5). In general, encoding performance and decoding performance did not show a direct linear relationship. This could be due to the fact that decoding performance not only depends on the precision with which BOLD responses are predicted but also on the properties of the filters b_k estimated by the different models. For example, the joint constraint of sparseness and smoothness imposed by the graphnet regularizer induces a strong inductive bias. This inductive bias could result in filters that impose stronger or more independent constraints on the reconstructions.

We have shown in this study that our framework allows high-quality reconstructions. The question remains how reconstruction quality could be further improved. In terms of data acquisition, we used a rapid event-related design where each unique stimulus was shown

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

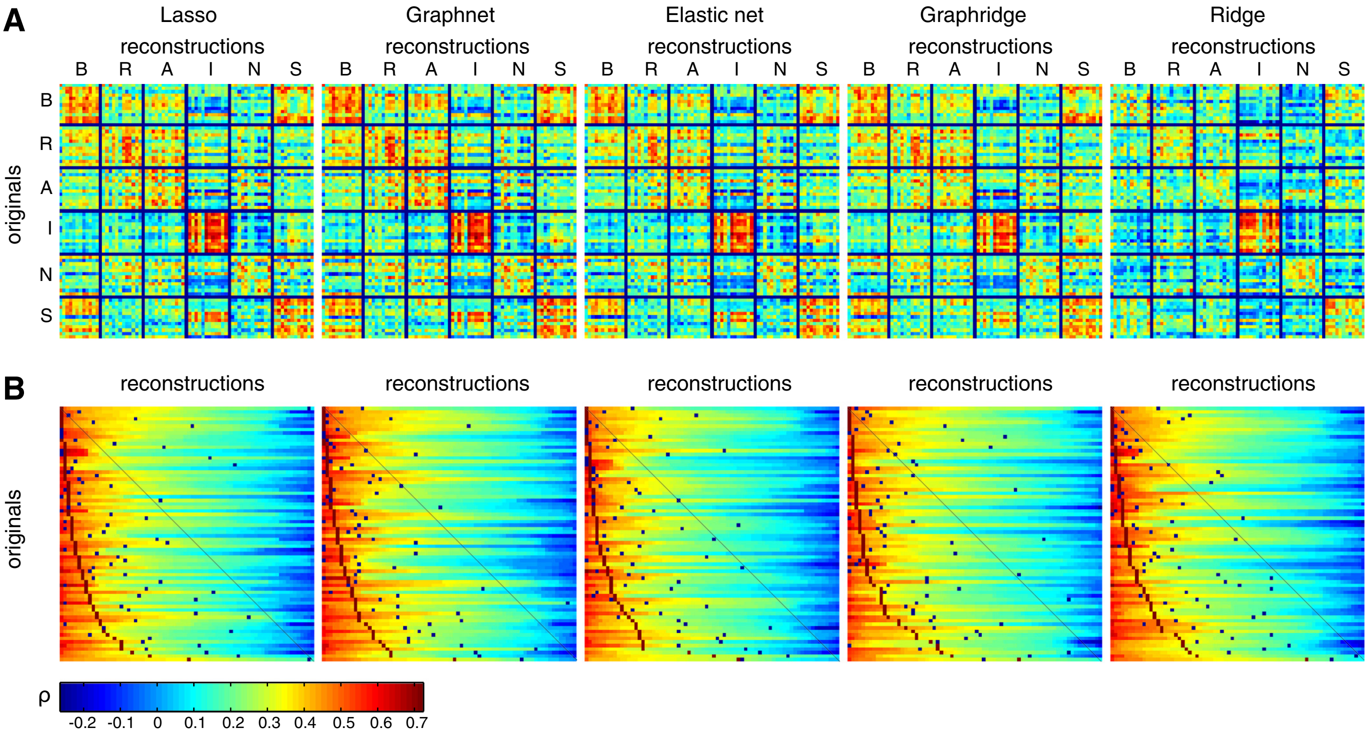

Fig. 7. Reconstruction quality of individual exemplars. (A) Correlation matrices for all regularizers for the large dataset containing all six characters. Entries of the correlation matrix indicate the correlation between an original and reconstructed image. Dark blue lines are used to separate the letter classes. (B) Correlation matrix with rows sorted according to the correlation between the original and its corresponding reconstruction. For each row, the correlations between an original and all reconstructions were again sorted. The rank of the correct reconstruction when using our explicit inversion scheme is indicated in dark red. The rank of the correct reconstruction when using an empirical prior is indicated in dark blue. The diagonal indicates the rank which is expected based on chance.

twice. BOLD responses acquired at a magnetic field strength of 3 T were quantified in terms of beta weights as estimated using the general linear model (Mumford et al., 2011). Other experimental designs and other approaches to BOLD deconvolution might lead to better decoding performance. Also, acquisition at higher field strengths, allowing imaging of BOLD responses at the level of cortical columns or layers, is expected to yield considerably improved decoding results (Polimeni et al., 2010).

In terms of the employed linear decoding approach, various modifications could further improve reconstruction performance. First, parameters $\widehat{\sigma}_k^2$, which model the variance of the BOLD response, have been derived from training data, which may lead to over-optimistic estimates. Estimation of the variance parameters from test data using a proper nested cross-validation, though costly, might lead to improved reconstructions.

Second, the regularization approaches could be optimized even further, either by using a different coupling matrix, employing adaptive shrinkage, or using more robust loss functions than the squared loss function used here Grosenick et al. (2013).

Third, in our current approach, predictions were based on pixel intensities. Linear transformations of these intensities might provide better basis functions and could lead to better reconstructions while still allowing a closed-form solution. That is, we can trivially include any

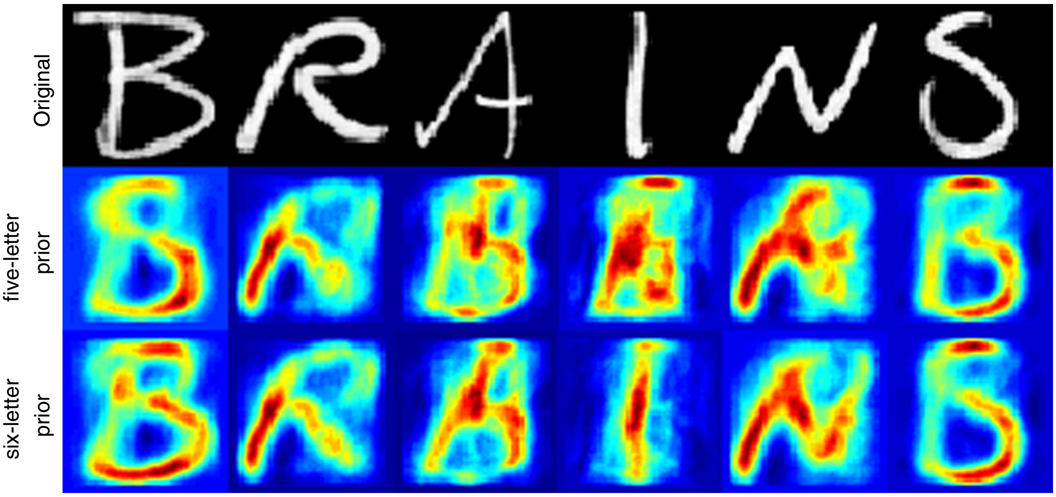

Fig. 8. Generalization to new letter classes. Each letter is predicted using models trained on BOLD data for the remaining letter classes. Extending the prior to include the new letter class is shown to substantially improve the reconstructions.

desired linear transformation Ux of the inputs by replacing B with $U\widetilde{B}$ and/or replacing R^{-1} with $U\widetilde{R}^{-1}U^\top$ in Eq. (7). Note, however, that such an approach would need to be accompanied by restrictions on the linear transformations since otherwise no expressive power would be gained. Restrictions can take the form of allowing only a restricted number of basis functions or by regularizing the parameters of the linear transformation.

Finally, the current approach could be improved by using richer image priors that still afford the analytical approach put forward in this paper. For example, the prior could be given by a mixture of Gaussians. Such a model could be estimated using an expectation maximization (EM) algorithm (Dempster et al., 1977). In case the image categories are known beforehand, mixture components can be estimated independently without resorting to an EM approach. The mixture model could also be made dependent on semantic information as in Naselaris et al. (2009). That is, we could use a discriminative approach to predict the mixture weights, effectively adapting the image prior based on semantic information.

Summarizing, results show that good reconstructions can be obtained by inverting properly regularized encoding models in the linear Gaussian setting. Results show that the graphnet-regularized linear Gaussian model performs best in terms of decoding performance and at the same time learns smooth yet localized linear filters. When speed is of the essence, the kernel formulation of the ridge-regularized linear Gaussian model may be the preferred choice. The question remains how these models compare to more complex decoding approaches that rely on non-linear transformations (Kay et al., 2008; van Gerven et al., 2010; Vu et al., 2011). In order to address this question, source code implementing our approach is available upon request. We propose that the outlined analytical approach serves as a baseline against which to compare other approaches.

Acknowledgments

This research was funded by the Netherlands Organization for Scientific Research (NWO-EW) under grant number 612.001.013.

Appendix A. Analytical expression for the regression coefficients

Here, we derive the analytical expression for the regression coefficients (14), as well as the kernel formulation (15), in case $\alpha = 0$. Consider a coefficient vector b and responses y. We wish to compute $\widehat{b} = \arg\min_b E(b)$ with objective function

$$E(b) = \arg\min_b \left\{ \frac{1}{2N} \|y - Xb\|_2^2 + \frac{\lambda}{2} b^\top G b \right\}.$$

The gradient of the objective function takes the form

$$\nabla E(b) = \frac{1}{N} \sum_n \left(b^\top x^n - y^n \right) (x^n)^\top + \frac{\lambda}{2} \left(G + G^\top \right) b.$$

Setting to zero, we obtain

$$\begin{aligned} 0 &= -\sum_n y^n (x^n)^\top + b^\top \sum_n x^n (x^n)^\top \\ &\quad + \frac{N\lambda}{2} \left(G + G^\top \right) b \\ &= -X^\top y + X^\top X b + \widetilde{G} b \end{aligned}$$

with symmetric $\widetilde{G} = \frac{N\lambda}{2}(G + G^\top)$. We write

$$X^\top X b + \widetilde{G} b = \left(X^\top X + \widetilde{G} \right) b = X^\top y. \tag{A.1}$$

Solving for b, we obtain the standard analytical solution

$$b = \left(X^\top X + \widetilde{G} \right)^{-1} X^\top y.$$

Alternatively, we can write $\widetilde{b} \equiv \widetilde{G} b = X^\top (y - Xb) = X^\top \beta$ with $\beta \equiv y - Xb$. Hence, $\widetilde{b}$ can be written as a linear combination of the training samples. By substituting $\widetilde{b}$ with this dual representation into Eq. (A.1), we obtain

$$\begin{aligned} \left(X^\top X \widetilde{G}^{-1} + I \right) \widetilde{b} &= X^\top y \Rightarrow \\ \left(X^\top X \widetilde{G}^{-1} + I \right) X^\top \beta &= X^\top y \Rightarrow \\ \beta &= y - X \widetilde{G}^{-1} X^\top \beta \Rightarrow \\ \left(X \widetilde{G}^{-1} X^\top + I \right) \beta &= y \Rightarrow \\ \beta &= \left(X \widetilde{G}^{-1} X^\top + I \right)^{-1} y \Rightarrow \\ \widetilde{b} &= X^\top \left(X \widetilde{G}^{-1} X^\top + I \right)^{-1} y \Rightarrow \\ b &= \widetilde{G}^{-1} X^\top \left(X \widetilde{G}^{-1} X^\top + I \right)^{-1} y, \end{aligned}$$

which is the kernel formulation of Eq. (15). Note that this formulation requires $\widetilde{G}$ to be invertible. This does not hold in case of graphridge regression in conjunction with the graph Laplacian. In that case, a small diagonal term can be added to the graph Laplacian for stability.

Appendix B. Regularization path

As the regularization path, we take a uniform interval on the log scale from $\lambda_{\max}$ to $\lambda_{\min} \equiv 10^{-4}\lambda_{\max}$. The parameter $\lambda_{\max}$ is defined to be that value of λ for which one of the variables enters the model. It holds that

$$\lambda_{\max} = \frac{1}{\alpha N} \max_i \left| (Xy)_i \right|. \tag{B.1}$$

Proof. Define the objective function

$$\begin{aligned} E(b) &= L(b) + R_{\lambda,G}(b) \\ &= L(b) + \lambda \left(\alpha \sum_i |b_i| + (1 - \alpha) \frac{1}{2} \sum_{i,j} b_i G_{ij} b_j \right), \end{aligned}$$

with $L(b) \equiv \| y - X^\top b \|_2^2 / 2N$. A variable is included whenever the solution $b_i = 0$ becomes unstable. Now, consider changing b_i away from zero. Since variables b_j with $i \neq j$ are fixed at zero, we can restrict ourselves to study the dependency of $E(b)$ on those terms that have elements in common with b_i and are non-zero:

$$E(b_i) \equiv L\left(b_i^*(b_i)\right) + \lambda \left(\alpha |b_i| + (1 - \alpha) \frac{1}{2} G_{ii} b_i^2 \right) + C, \tag{B.2}$$

where $E(b_i) \equiv E(b_i^*(b_i))$ with $b_i^*(b_i)$ the zero vector whose i-th element is replaced by b_i. Note that this expression is equivalent to the expression we obtain using an elastic net regularizer.

The solution $b_i = 0$ becomes unstable if it holds that $E(b_i) < E(0)$ for some infinitesimally small change in b_i. A first-order Taylor expansion for b_i close to 0 yields:

$$E(b_i) \equiv E(0) + g_i b_i + \lambda \left(\alpha |b_i| + (1 - \alpha) G_{ii} b_i^2 \right), \tag{B.3}$$

where here and in the following we ignore higher order terms and we defined $g_i \equiv \frac{\partial L(b)}{\partial b_i}\Big|_{b=0}$ for ease of notation. A variable b_i thus enters the model at

$$\begin{aligned} \lambda_i &\equiv \max_{b_i} \left[-\frac{g_i b_i}{\alpha |b_i| + (1 - \alpha) G_{ii} b_i^2} \right] \\ &= \max_{b_i} \left[-\frac{g_i \operatorname{sgn}(b_i)}{\alpha + (1 - \alpha) G_{ii} |b_i|} \right]. \end{aligned} \tag{B.4}$$

Since the numerator is independent of the magnitude of b_i and since, assuming positive G_{ii}, $|b_i|$ only reduces the magnitude of the quantity

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

between brackets, the optimal value of b_i is an infinitesimally small value whose sign is such that the quantity between brackets becomes positive. Since we can ignore the second term in the denominator as $b_i \rightarrow 0$, we obtain

$$\lambda_i = \frac{1}{\alpha}\left|\frac{\partial L(\mathrm{b})}{\partial b_i}\bigg|_{\mathrm{b}=0}\right|. \tag{B.5}$$

It follows that

$$\lambda_{\max} = \frac{1}{\alpha}\max_i\left|\frac{\partial L(\mathrm{b})}{\partial b_i}\bigg|_{\mathrm{b}=0}\right| = \frac{1}{\alpha N}\max_i\left|(\mathrm{Xy})_i\right|.$$

References

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer Verlag.

Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R., 2009. Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44, 112–122.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B 39, 1–38.

DeYoe, E.A., Carman, G.J., Bandettini, P., Glickman, S., Wieser, J., Cox, R., Miller, D., Neitz, J., 1996. Mapping striate and extrastriate visual areas in human cerebral cortex. Proc. Natl. Acad. Sci. U. S. A. 93, 2382–2386.

Engel, S.A., Glover, G.H., Wandell, B.A., 1997. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cereb. Cortex 7, 181–192.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22.

Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2008. Bayesian decoding of brain images. NeuroImage 39, 181–205.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with GraphNet. NeuroImage 72, 304–321.

Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition. Springer, New York, NY.

Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7, 523–534.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8, 679–685.

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

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B., 2011. Total variation regularization for fMRI-based prediction of behavior. IEEE Trans. Med. Imaging 30, 1328–1340.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H.C., Sadato, N., Kamitani, Y., 2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60, 915–929.

Mumford, J.A., Turner, B.O., Ashby, F.G., Poldrack, R.A., 2011. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. NeuroImage 1–36.

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., Gallant, J.L., 2009. Bayesian reconstruction of natural images from human brain activity. Neuron 63, 902–915.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2010. Encoding and decoding in fMRI. NeuroImage 56, 400–410.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 1–6.

Polimeni, J.R., Fischl, B., Greve, D.N., Wald, L.L., 2010. Laminar analysis of 7T BOLD using an imposed spatial activation pattern in human V1. NeuroImage 52, 1334–1346.

Ringach, D., Shapley, R., 2004. Reverse correlation in neurophysiology. Cogn. Sci. 28, 147–166.

Sereno, M.I., Dale, A.M., Reppas, J.B., Kwong, K.K., Belliveau, J.W., Brady, T.J., Rosen, B.R., Tootell, R.B., 1995. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science 268, 889–893.

Simoncelli, E.P., Olshausen, B.A., 2001. Natural image statistics and neural representation. Ann. Rev. Neurosci. 24, 1193–1216.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.B., Lebihan, D., Dehaene, S., 2006. Inverse retinotopy: inferring the visual content of images from brain activation patterns. NeuroImage 33, 1104–1116.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B 58, 267–288.

van der Maaten, L., 2009. A new benchmark dataset for handwritten character recognition. Technical Report. Tilburg University, Tilburg, The Netherlands.

van Gerven, M.A.J., Heskes, T., 2010. Sparse orthonormalized partial least squares. Benelux Conference on Artificial Intelligence.

van Gerven, M.A.J., Heskes, T., 2012. A linear Gaussian framework for decoding of perceived images. 2011 International Workshop on Pattern Recognition in Neuroimaging (PRNI), pp. 1–4.

van Gerven, M.A.J., de Lange, F.P., Heskes, T., 2010. Neural decoding with hierarchical generative models. Neural Comput. 22, 3127–3142.

van Gerven, M.A.J., Maris, E., Heskes, T., 2011. A Markov random field approach to neural encoding and decoding. International Conference on Artificial Neural Networks.

Vu, V.Q., Ravikumar, P., Naselaris, T., Kay, K.N., 2011. Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models. Ann. Appl. Stat. 5, 1159–1182.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B 67, 301–320.