A Linear Gaussian Framework for Decoding of Perceived Images

Marcel A. J. van Gerven
Radboud University Nijmegen
Donders Institute for Brain, Cognition and Behaviour
Nijmegen, The Netherlands
Email: m.vangerven@donders.ru.nl

Tom Heskes
Radboud University Nijmegen
Institute for Computing and Information Sciences
Nijmegen, The Netherlands
Email: tomh@cs.ru.nl

***Abstract*—With the advent of sophisticated acquisition and analysis techniques, decoding the contents of someone's experience has become a reality. We propose a simple linear Gaussian framework where decoding relies on the inversion of properly regularized encoding models. We show that this approach yields state-of-the-art decoding performance on an fMRI dataset.**

***Keywords*-Bayesian decoding, perception, fMRI analysis**

I. 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 [12]. Decoding has gained much recent popularity with the adoption of multivariate analysis methods by the cognitive neuroscience community [5]. While the first decoding studies have focused exclusively on the prediction of discrete states such as object category [4] or stimulus orientation [8], more recent work has focused on the prediction of increasingly complex stimulus properties, culminating in the reconstruction of the contents of visual scenes [9], [11], [13].

From the Bayesian point of view, encoding and decoding are intimately connected via Bayes rule where the probability $p(\mathbf{x} \mid \mathbf{y})$ of a stimulus $\mathbf{x}$ given a response $\mathbf{y}$ is expressed as the product of an encoding distribution $p(\mathbf{y} \mid \mathbf{x})$ and a prior $p(\mathbf{x})$, up to some normalizing constant [2], [10]. The encoding distribution is assumed to embody 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. [13] assume that each voxel has a Gaussian receptive field which allows inversion of the generative model. Naselaris et al. [11], in contrast, use a complex forward model and do not perform the inversion explicitly. Instead they use a natural image 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 need for inverse inference.

In this paper we present a framework for decoding that expands on the ideas put forward in the aforementioned papers. Specifically, similar to [11], 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 [13]. The ideas presented in this paper extend earlier work [15] on the decoding of discrete (binary) inputs to continuous (grey-scale) images. We specialize here to the linear Gaussian setting and show that decoding performance is quite good in this context.

II. Methods

Let $(\mathbf{x}, \mathbf{y})$ denote a stimulus-response pair, say, an image $\mathbf{x} = (x_1, \ldots, x_p)^\top$, characterized by its pixel values x_i, and the associated measured response vector $\mathbf{y} = (y_1, \ldots, y_q)^\top$, in our case, BOLD responses in early visual cortex . Without loss of generality, we can assume that the response is centered and the stimuli are standardized. That is, $\mathrm{E}(\mathbf{y}) = \mathrm{E}(\mathbf{x}) = \mathbf{0}$ and $\mathrm{E}(x_i^2) = 1$ for $i = 1, 2, \ldots, p$. The stimulus can be either discrete or continuous and the response is typically continuous, e.g., the BOLD response for multiple voxels. In this paper we are interested in decoding the most probable image from BOLD response:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \{p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x})\} \; . \tag{1}$$

To do so, we require a forward model $p(\mathbf{y} \mid \mathbf{x})$ and an image prior $p(\mathbf{x})$. In [11], 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 selecting that image which gave the largest likelihood. Here, in contrast, we solve the decoding problem without relying on a restricted subset of possible images by making use of a generic image prior. This approach is related to the work presented in [13], although here we make less strong assumptions on the forward model and the image prior.

978-0-7695-4765-7/12 $26.00 © 2012 IEEE
DOI 10.1109/PRNI.2012.10

A. Encoding and decoding

We assume that the forward (encoding) model is given by a linear regression model. Furthermore, the response is assumed to depend on a set of image features $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_r(\mathbf{x}))^\top$ with $\mathrm{E}(\phi_j(\mathbf{x})) = 0$ for $1 \leq j \leq r$. That is, individual responses y_k are conditionally independent and given by a linear function of $\boldsymbol{\phi}(\mathbf{x})$ with additive Gaussian noise, such that $y_k = \boldsymbol{\beta}_k^\top \boldsymbol{\phi}(\mathbf{x}) + \epsilon_k$ where $\boldsymbol{\beta}_k$ is a vector of regression coefficients for response k and ϵ_k is a zero mean Gaussian random variable with variance σ_k^2. Hence, we have

$$p(\mathbf{y} \mid \mathbf{x}) \quad = \quad \mathcal{N}\left(\mathbf{y}; \mathbf{B}^\top \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\Sigma}\right) \tag{2}$$

where $\mathbf{B} \equiv (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)$. Without loss of generality we assume the image prior takes the form of a Gibbs distribution $p(\mathbf{x}) \propto \exp\left(-J(\mathbf{x})\right)$ with parameters that we leave as yet unspecified.

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 $p(\mathbf{x} \mid \mathbf{y})$, given by Eq. (1). Conditional on $\mathbf{y}$, we can drop the terms in (2) not depending on $\mathbf{x}$ and, after some rewriting, obtain

$$p(\mathbf{x} \mid \mathbf{y}) \propto \exp\left((\mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{y})^\top \boldsymbol{\phi}(\mathbf{x}) - \frac{1}{2}\boldsymbol{\phi}(\mathbf{x})^\top \mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{B}^\top \boldsymbol{\phi}(\mathbf{x}) - J(\mathbf{x})\right) \tag{3}$$

The question is for which $\boldsymbol{\phi}$ and J we can estimate the required parameters and solve (3) in a reasonable amount of time. In this paper, we show that the solution can be obtained in closed form in the linear Gaussian case. While this linear case might seem too restrictive, we show that, using additional assumptions on $\mathbf{B}$, the obtained reconstructions are quite good, as shown on a real-world dataset.

B. Linear Gaussian case

In the linear Gaussian case, we assume that the image prior is given by a multivariate Gaussian over image features $\boldsymbol{\psi}(\mathbf{x})$ of the form:

$$p(\mathbf{x}) \quad \propto \quad \exp\left(\left(\mathbf{R}^{-1}\boldsymbol{\mu}\right)^\top \boldsymbol{\psi}(\mathbf{x}) - \frac{1}{2}\boldsymbol{\psi}(\mathbf{x})^\top \mathbf{R}^{-1}\boldsymbol{\psi}(\mathbf{x})\right)$$

with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$. Furthermore, we assume that the features are given by linear transformations. That is, $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{U}^\top \mathbf{x}$ and $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{V}^\top \mathbf{x}$. In that case, since $\mathrm{E}(\mathbf{V}^\top \mathbf{x}) = \mathbf{V}^\top \mathrm{E}(\mathbf{x}) = \mathbf{0}$, the prior becomes

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{V}\mathbf{R}^{-1}\mathbf{V}^\top \mathbf{x}\right). \tag{4}$$

By plugging (4) into (3), and using notation $\tilde{\mathbf{B}} = \mathbf{U}\mathbf{B}$ and $\tilde{\mathbf{R}}^{-1} = \mathbf{V}\mathbf{R}^{-1}\mathbf{V}^\top$, we arrive at $p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}(\mathbf{x}; \mathbf{m}, \mathbf{Q})$ where $\mathbf{m} \equiv \mathbf{Q}\tilde{\mathbf{B}}\boldsymbol{\Sigma}^{-1}\mathbf{y}$ and $\mathbf{Q} \equiv \left(\tilde{\mathbf{R}}^{-1} + \tilde{\mathbf{B}}\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{B}}^\top\right)^{-1}$. It immediately follows that

$$\mathbf{x}^* = \mathbf{m} = \left(\tilde{\mathbf{R}}^{-1} + \tilde{\mathbf{B}}\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{B}}^\top\right)^{-1} \tilde{\mathbf{B}}\boldsymbol{\Sigma}^{-1}\mathbf{y} \tag{5}$$

since the mode of a Gaussian distribution is given by its mean. For large images, computing (5) may be prohibitively expensive since it requires the inversion of an $r \times r$ covariance matrix, where r is the number of features. Fortunately, using the matrix inversion lemma, we obtain

$$\mathbf{x}^* = \left(\tilde{\mathbf{R}} - \tilde{\mathbf{R}}\tilde{\mathbf{B}}\left(\boldsymbol{\Sigma} + \tilde{\mathbf{B}}^\top \tilde{\mathbf{R}}\tilde{\mathbf{B}}\right)^{-1}\tilde{\mathbf{B}}^\top \tilde{\mathbf{R}}\right)\tilde{\mathbf{B}}\boldsymbol{\Sigma}^{-1}\mathbf{y}\,, \tag{6}$$

which requires the inversion of a $q \times q$ matrix, where q is the number of voxels. Matrices $\mathbf{U}$ and $\mathbf{V}$ could be chosen as fixed transformations, such as differences of Gaussians, or learned from data. Here, however, we further reduce to the special case where $\mathbf{U} = \mathbf{V} = \mathbf{I}$. Hence, we focus on learning a forward model directly from image pixels using regularized linear regression while assuming a Gaussian image prior and the optimal solution is given by Eq. (5).

C. Parameter estimation

In order to estimate the parameters of the forward model, we use a regularized linear regression approach. That is, for each response k, we need to solve a minimization problem of the form

$$(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k) = \arg\min_{\boldsymbol{\beta}_k, \sigma_k^2} \left\{ \frac{||\mathbf{y}_k - \boldsymbol{\Phi}^\top \boldsymbol{\beta}_k||_2^2}{2N\sigma_k^2} + R_{\nu,\boldsymbol{\Lambda}}(\boldsymbol{\beta}_k) \right\} \tag{7}$$

where $\boldsymbol{\Phi} = (\boldsymbol{\phi}(\mathbf{x}^1), \ldots, \boldsymbol{\phi}(\mathbf{x}^N))$ and $\mathbf{y}_k = (y_k^1, \ldots, y_k^N)^\top$ denote the image features and BOLD response for N samples, respectively, and

$$R_{\nu,\Lambda}(\boldsymbol{\beta}_k) = -\log p(\boldsymbol{\beta}_k) = \nu||\boldsymbol{\beta}_k||_1 + \frac{1}{2}\boldsymbol{\beta}_k' \boldsymbol{\Lambda}\boldsymbol{\beta}_k$$

is the graph-constrained elastic net (GraphNet) regularizer [3], which generalizes the elastic net regularizer [16]. The factor $1/N$ is used to make the regularization invariant to the number of training samples. The parameter ν determines the amount of ℓ_1 regularization and $\boldsymbol{\Lambda}$ is a matrix which induces shrinkage where non-zero off-diagonal terms induce a coupling between parameters. With $\boldsymbol{\Lambda} = \mathbf{0}$ we have zero contribution of the ℓ_2 norm and end up with lasso regression. With diagonal $\boldsymbol{\Lambda}$ and $\nu = \mathbf{0}$ we have zero contribution of the ℓ_1 norm and end up with ridge regression. Minimization of Eq. (6) with respect to $\boldsymbol{\beta}_k$ boils down to computing

$$\hat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}_k} \left\{ \frac{1}{2N}||\mathbf{y}_k - \boldsymbol{\Phi}^\top \boldsymbol{\beta}_k||_2^2 + R_{\nu,\boldsymbol{\Lambda}}(\boldsymbol{\beta}_k) \right\}$$

which can be achieved using an efficient coordinate gradient descent algorithm [1]. Let $\boldsymbol{\Phi}_{\mathcal{A}}$ denote $\boldsymbol{\Phi}$ restricted to the columns in $\mathcal{A}$. Minimization of Eq. (6) with respect to the variance yields

$$\hat{\sigma}_k^2 = ||\mathbf{y}_k - \boldsymbol{\Phi}^\top \hat{\boldsymbol{\beta}}_k||_2^2 / (N - df(\hat{\boldsymbol{\beta}}_k))$$

with $df(\hat{\boldsymbol{\beta}}_k) = \mathrm{Tr}\left(\boldsymbol{\Phi}_{\mathcal{A}}\left(\boldsymbol{\Phi}_{\mathcal{A}}^{\top}\boldsymbol{\Phi}_{\mathcal{A}} + \boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Phi}_{\mathcal{A}}^{\top}\right)$ the degrees of freedom for the GraphNet model, where $\mathcal{A}$ is defined by the indices of the non-zero elements of $\hat{\boldsymbol{\beta}}_k$.

The parameters of the image prior can either be estimated from the training data or using an independent (and potentially much larger) set of images $\{\mathbf{z}^n\}_{n=1}^{M}$. In the linear Gaussian case, the required covariance matrix is readily estimated as $\mathbf{R} = \mathbf{V}^{\top}\left(\frac{1}{N-1}\sum_n \mathbf{z}^n(\mathbf{z}^n)^{\top}\right)\mathbf{V}$.

D. Experimental validation

We tested the regularized linear Gaussian model on data which has previously been used in the same context [14]. Briefly, we collected blood-oxygenation-level dependent (BOLD) data for 100 trials in one participant. In each trial a handwritten 6 or 9 was visually presented to the subject. The character remained visible for 12.5 seconds and flickered at a rate of 6 Hz on a black background. In order to ensure sustained attention during the entire scanning session, the subject's task was to maintain fixation to a fixation dot and to detect a brief change in color from red to green and back occurring once and randomly within a trial. BOLD data were obtained by means of a Siemens 3T MRI system using a 32-channel coil for signal reception. We used a single-shot gradient EPI sequence with a TR of 2500 ms, TE of 30 ms, and isotropic voxel size of $2 \times 2 \times 2$ mm. Functional images were acquired in 42 axial slices in ascending order. A high-resolution anatomical image was acquired using an MP-RAGE sequence. Functional data were preprocessed and analyzed within the framework of SPM8 (Statistical Parametric Mapping, www.fil.ion.ucl.ac.uk/spm). Functional data were motion-corrected, coregistered with the anatomical scan, detrended and high-pass-filtered. The volumes acquired 10 to 15 seconds after trial onset were averaged in order to obtain an estimate of the steady-state response in individual voxels. As input to our reconstruction model, we used 1185 voxels in visual area V1 as determined by a rotating wedge functional localizer.

We tested encoding and decoding performance by training on 40 sixes and 40 nines and testing on the remaining 10 sixes and 10 nines. The Gaussian image prior was learned by computing the covariance matrix for an independent set of data containing 1000 handwritten sixes and 1000 handwritten nines. Encoding performance is quantified in terms of explained variance whereas decoding performance is evaluated by comparing the decoded stimuli with the original input. In our analyses, we compared forward models which made use of a ridge, elastic net or GraphNet regularizer. For the ridge regularizer, we used $\nu = 0$ and $\boldsymbol{\Lambda} = \lambda\mathbf{I}$ with fixed $\lambda = 10^{-3}$. For the elastic net regularizer, we traversed the regularization path while keeping $\alpha = \lambda/(\lambda + \nu)$ fixed at 0.5. The optimal parameters λ and ν were then determined using a five-fold nested cross-validation. For the GraphNet regularizer, we used $\boldsymbol{\Lambda} = \lambda(\mathbf{D} - \mathbf{A})$ where $\mathbf{D}$ is the diagonal matrix of node degrees and $\mathbf{A}$ the pixel adjacency matrix. We used $\lambda = 10$ to induce a small but non-negligible correlation between neighbouring pixels. Again, we traversed the regularization path while varying ν and keeping $\boldsymbol{\Lambda}$ fixed. Optimal parameters were again chosen using a nested five-fold cross-validation.

III. Results

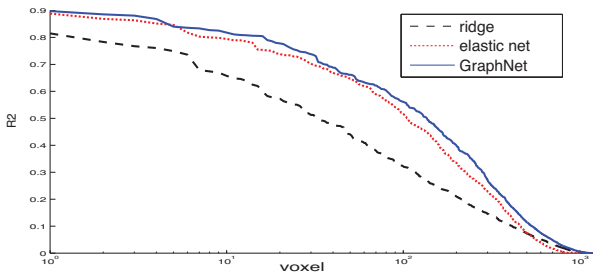

Figure 1. Explained variance per voxel for the three encoding models. Explained variance for a BOLD response $\mathbf{y}_k$ given the predicted response $\hat{\mathbf{y}}_k$ was given by $(\mathrm{var}(\mathbf{y}_k) - \mathrm{var}(\mathbf{y}_k - \hat{\mathbf{y}}_k))/\mathrm{var}(\mathbf{y}_k)$. The reported R2 value is the explained variance averaged over all voxels.

We tested encoding and decoding performance for the ridge, elastic net and GraphNet models. Figure 1 depicts the explained variance per voxel where, for all three models, we independently ordered the explained variance over voxels in descending order. The elastic net and GraphNet models clearly outperform the ridge model, with the GraphNet model improving somewhat on the elastic net model.

In order to examine what kind of filters $\tilde{\mathbf{B}}$ were learned by the three models, Fig. 2 depicts the filters corresponding to the ten voxels which showed the largest explained variance using the GraphNet model. For the ridge model, the filters show little discernible structure. For the elastic net model, the filters are sparse and show some clustering. For the GraphNet model, filters are less sparse and show stronger clustering as those of the elastic net model. For some filters, these clusters start to resemble the handwritten digit classes.

Finally, we examined the quality of the reconstructions which could be obtained using our framework. Figure 3 shows the original stimuli and their reconstructions based on the inversion of the three alternative encoding models. All models lead to reasonable reconstructions although the ridge model is clearly outperformed by both the elastic net and GraphNet models. Note that characteristic properties of individual letters can be discerned in the reconstructions, meaning that it is not only stimulus category on which the reconstructions are based.

IV. Conclusion

We introduced a straightforward framework for the decoding of images based on measured neural responses only. Results show that acceptable reconstructions can be obtained by inverting regularized encoding models in the linear Gaussian setting. While our framework generalizes to the case

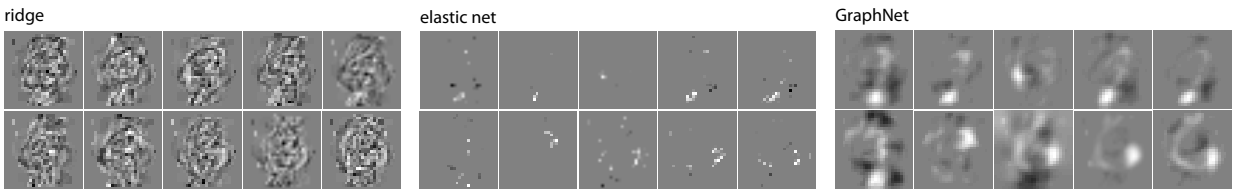

Figure 2. Filters learned for ten voxels using the three encoding models.

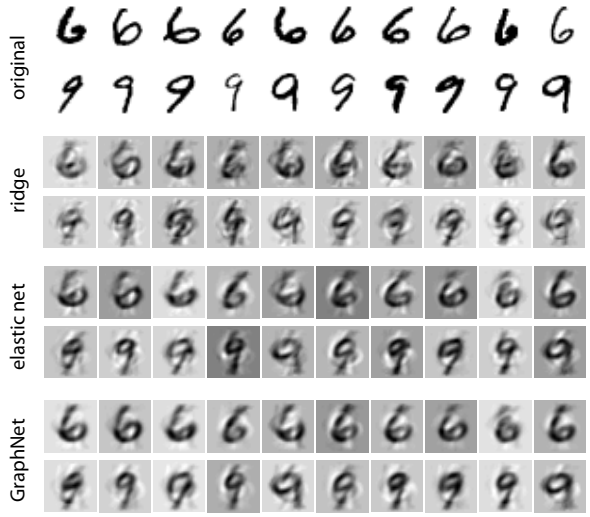

Figure 3. Reconstructions of original stimuli using three different decoding models. For clarity, the inverted images are shown.

where features can be linear transformations of the images, here, we showed that, by working directly on the image pixels and given suitably regularized models, acceptable reconstructions could be obtained. The question remains whether additional gains can be obtained by working in a different image basis. The best results were obtained by the elastic net and GraphNet models, with the latter showing some improvement over the former. Note, however, that these models have been estimated for fixed $\mathbf{\Lambda}$. A more extensive grid search over both values of ν and $\mathbf{\Lambda}$ could further improve results in both models. Finally, note that results may also be improved by using alternative regularized estimators which are either special cases of the GraphNet regularizer (e.g. the smooth Lasso [6]) or by using more general structured sparsity-inducing norms [7].

REFERENCES

[1] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

[2] K. Friston, C. Chu, J. Mourão-Miranda, O. Hulme, G. Rees, W. Penny, and J. Ashburner. Bayesian decoding of brain images. *Neuroimage*, 39:181–205, 2008.

[3] L. Grosenick, B. Klingenberg, B. Knutson, and J. Taylor. A family of interpretable multivariate models for regression and classification of whole-brain fMRI data. Technical report, arXiv:1110.4139v1, 2011.

[4] J.V. Haxby, M.I. Gobbini, M.L. Furey, A. Ishai, J.L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2425–2430, 2001.

[5] J.-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.*, 7:523–534, 2006.

[6] M. Hebiri and S. Van de Geer. The smooth-lasso and other l1 + l2-penalized methods. *Electron J Stat*, 5:1184–1226, 2011.

[7] R. Jenatton, J.-Y. Audibert, and F. R. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

[8] Y. Kamitani and F. Tong. Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.*, 8(5):679–685, 2005.

[9] Y. Miyawaki, H. Uchida, O. Yamashita, M. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani. Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders. *Neuron*, 60(5):915–929, 2008.

[10] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant. Encoding and decoding in fMRI. *Neuroimage*, 2010.

[11] 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.

[12] D. Ringach and R. Shapley. Reverse Correlation in Neurophysiology. *Cogn. Sci.*, 28:147–166, 2004.

[13] 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.

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

[15] M. A. J. van Gerven, E. Maris, and T. Heskes. A Markov Random Field Approach to Neural Encoding and Decoding. In *ICANN 2011*, 2011.

[16] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B*, 67(2):301–320, 2005.