Unsupervised Learning of Features for Bayesian Decoding in Functional Magnetic Resonance Imaging

Umut Güçlü U.GUCLU@DONDERS.RU.NL
Marcel van Gerven M.VANGERVEN@DONDERS.RU.NL
Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

Keywords: Bayesian inference, functional magnetic resonance imaging, independent component analysis, neural decoding, unsupervised feature learning

Abstract

Neural decoding is concerned with inferring certain aspects of a stimulus from brain activity. With the recent advent of functional magnetic resonance imaging (fMRI), it has become possible to create a literal picture of a visual stimulus from the human brain. Most conventional decoders are based either on the input space or on a hand-designed feature space. An alternative to hand-designing a feature space is unsupervised feature learning, which has seen much success in computer vision. Here, we present a new decoder, which combines Bayesian inversion of voxel-based encoding models with unsupervised feature learning (independent component analysis). We validated our decoder by reconstructing images of handwritten digits from human brain activity measured using fMRI, with state-of-the-art accuracy. Our results show that the feature space has a substantial effect on the accuracy of the reconstructions, and independent component analysis provides an effective means to learn feature spaces for neural decoding in fMRI.

1. Introduction

Neural decoding is concerned with inferring certain aspects of a stimulus from stimulus-evoked brain activity. Functional magnetic resonance imaging (fMRI) measures the activity of many separate voxels (i.e. volumetric pixels) in the brain by detecting the associated changes in the blood-oxygen-level-dependent (BOLD) haemodynamic responses. The spatial resolution afforded by fMRI has made it possible to take advantage of the information contained in distributed patterns of activity evoked by a stimulus in order to classify (Haxby et al., 2001; Kamitani & Tong, 2005; van Gerven et al., 2010a), identify (Mitchell et al., 2008; Kay et al., 2008) or reconstruct (Thirion et al., 2006; Miyawaki et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011; van Gerven et al., 2010b; van Gerven & Heskes, 2012) the original stimulus.

In the context of neural decoding, reconstruction refers to creating a literal picture of a stimulus from the brain. That is, given an encoding relationship that characterizes how a stimulus or some stimulus features are represented by brain activity, reconstruction is the process of determining the inverse of the encoding relationship (i.e. the decoding relationship) in order to reproduce the stimulus.

Inverting a neural response function is non-trivial because of the stochastic dynamics of neural processes (Brown et al., 2004). Therefore, the encoding relationship is described by a stochastic model (Dayan & Abbott, 2001). Furthermore, prior information about the stimulus is often incorporated in the process of reconstruction, which can also be described by a stochastic model, in order to capture the statistical properties of the environment (Dayan & Abbott, 2001).

Bayesian decoding combines the encoding relationship (i.e. the likelihood) and the prior information (i.e. the prior) using Bayes' theorem in order to describe the decoding relationship (i.e. the posterior). The conventional approach to Bayesian decoding is to characterize how certain "hand-designed" features of a stimulus (e.g. Gabor features) are represented by brain activity.

Appearing in *Proceedings of BENELEARN 2013*.

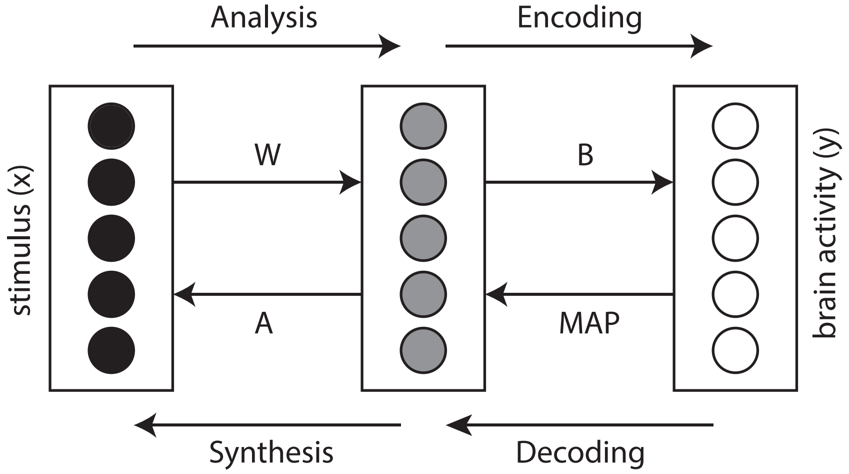

Figure 1. Our framework for reconstructing images from stimulus-evoked BOLD haemodynamic responses. The matrix of linear feature detectors (W) is the parameter of the statistical generative model. W is learned from unlabeled images. The matrix of linear features (A) is the inverse of W. The matrix of regression coefficients (B) is the parameter of the voxel-based encoding models. B is learned from stimulus-response pairs such that the images are analyzed in order to find their latent independent components. The latent independent components are represented by the gray circles in the figure. Reconstruction of a stimulus from stimulus-evoked BOLD haemodynamic responses is the process of maximum a posteriori (MAP) estimation to obtain point estimates of the latent independent components of the stimulus followed by image synthesis to reproduce the stimulus.

However, hand-designing features for complex stimuli and adapting hand-designed features of a particular set of stimuli with certain characteristics to another set of stimuli with different characteristics can be difficult. Unsupervised feature learning is an alternative to the conventional approach, which can mitigate the limitations of hand-designing features and has seen much success in computer vision (Bengio et al., 2012).

Furthermore, while it has been shown that prior information has a substantial effect on reconstruction accuracy (Naselaris et al., 2009), determining a suitable prior that can be used in Bayesian inference has been a challenging goal such that generic priors and empirical priors have often been used (Thirion et al., 2006; Naselaris et al., 2009; Nishimoto et al., 2011; van Gerven & Heskes, 2012). Another advantage of unsupervised feature learning is that a statistical generative model can be used as a prior in Bayesian inference (Hyvärinen et al., 2009). Unsupervised feature learning has already been used in the context of neural decoding. For example, van Gerven et al., (2010b) reconstructed handwritten digits using deep belief networks.

Here, we introduce a new, more straightforward approach to unsupervised feature learning for neural decoding that mitigates the limitations of hand-designing features and gives a proper prior that can be used in Bayesian inference. Our framework combines unsupervised feature learning with Bayesian decoding for reconstructing images from stimulus-evoked BOLD haemodynamic responses (Figure 1).

In particular, we use independent component analysis (ICA) to define a statistical generative model that describes how images are generated as linear transformations of their latent independent components (Hyvärinen, 2010) and linear regression to define voxel-based encoding models that characterize how latent independent components of images are represented by BOLD haemodynamic responses. That is, combining the analysis-synthesis loop and the encoding-decoding loop, reconstruction is defined as the process of Bayesian inference from the voxel-based encoding models followed by image synthesis from the statistical generative model.

In order to learn useful linear features from unlabeled data, to be used in linear regression, we have to impose constraints on the statistical generative model. Two approaches that are typically used is to impose a bottleneck to learn an under-complete representation (van Gerven & Heskes, 2010) and constrain the representation to be sparse (Olshausen & Field, 1996). The statistical generative model defined using ICA discovers interesting structure in the data by learning under-complete non-Gaussian (i.e. sparse) representations.

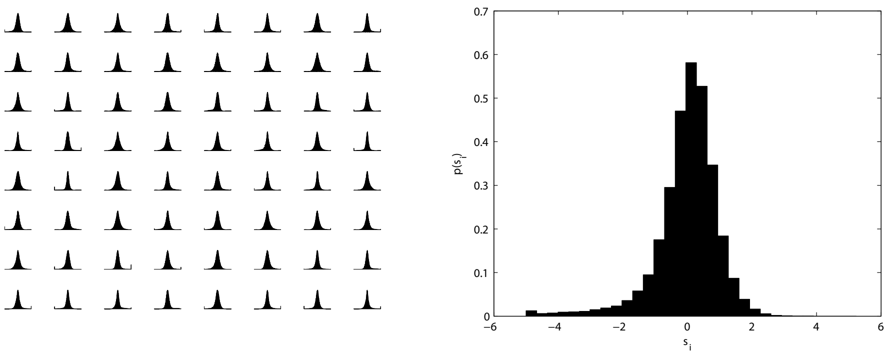

Figure 2. Left panel shows the distributions of 64 latent independent components estimated from grayscale images of handwritten digits. The x-axes represent s_i and the y-axes represent $p(s_i)$. The right panel shows the distribution of one the component in more detail. Note that the distributions are indeed peaked at zero and have high kurtosis.

In the following sections, we first present the derivation of our framework. We then validate our framework by reconstructing grayscale images of handwritten digits from stimulus-evoked BOLD haemodynamic responses. We finally show the effect of unsupervised feature learning on the reconstruction accuracy.

2. Methods

Let $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$ be a stimulus-response pair where $\mathbf{x}$ is a vector of pixel gray-scale values in an image, and $\mathbf{y}$ is a vector of multiple voxel activities evoked by $\mathbf{x}$. Furthermore, let $\phi : \mathbb{R}^p \to \mathbb{R}^m$ be an invertible linear transformation between the stimulus space and a feature space.

Without loss of generality, we assume that both $\phi(\mathbf{x})$ and $\mathbf{y}$ are normalized to have zero mean and unit variance.

We are interested in the problem of reconstructing $\mathbf{x}$ from $\mathbf{y}$:

$$\hat{\mathbf{x}} = \phi^{-1}\left(\arg\max_{\phi(\mathbf{x})} \{p\left(\phi\left(\mathbf{x}\right)|\mathbf{y}\right)\}\right) \tag{1}$$

where $\hat{\mathbf{x}}$ is a reconstruction of $\mathbf{x}$, and $p(\phi(\mathbf{x})|\mathbf{y})$ is a decoding distribution. We can equivalently formulate the problem of reconstructing $\mathbf{x}$ from $\mathbf{y}$ using Bayes' theorem:

$$\hat{\mathbf{x}} = \phi^{-1}\left(\arg\max_{\phi(\mathbf{x})} \{p\left(\mathbf{y}|\phi\left(\mathbf{x}\right)\right) p\left(\phi\left(\mathbf{x}\right)\right)\}\right) \tag{2}$$

where $p(\mathbf{y}|\phi(\mathbf{x}))$ is an encoding distribution, and $p(\phi(\mathbf{x}))$ is a prior. Therefore, in order to solve the problem of reconstructing $\mathbf{x}$ from $\mathbf{y}$, we need to define ϕ, $p(\phi(\mathbf{x}))$ and $p(\mathbf{y}|\phi(\mathbf{x}))$.

2.1. Unsupervised Feature Learning

We start by defining a statistical generative model of images. Assuming that an image is generated by a linear superposition of some features, we use ICA to define the statistical generative model of images by a linear transformation of the latent independent components of the image:

$$\mathbf{z} = \mathbf{A}\mathbf{s} \tag{3}$$

where $\mathbf{z} \in \mathbb{R}^p$ is a vector of pixel gray-scale values in an image, $\mathbf{A} \in \mathbb{R}^{p \times m}$ is a matrix of linear features, and $\mathbf{s} \in \mathbb{R}^m$ is a vector of the latent independent components of $\mathbf{z}$ such that $m \leq p$. In order to compute s_i as a linear function of $\mathbf{z}$, we invert the linear system defined by $\mathbf{A}$:

$$s_i = \mathbf{W}\mathbf{z} \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{m \times p}$ is a matrix of linear feature detectors such that $\mathbf{W} = \mathbf{A}^{-1}$. Furthermore, we make the simplifying assumption that s_i have unit variance in order to make s_i unique, up to a multiplicative sign.

We then define $p(s_i)$, assuming that the s_i are non-Gaussian and sparseness is the most dominant type of non-Gaussianity in the images that we are considering (Fig. 2), we use a distribution that is peaked at zero and has high kurtosis to define $p(s_i)$. In particular, we use the logistic distribution:

$$p(s_i) = \text{logistic}\left(0, \frac{\sqrt{3}}{\pi}\right) \tag{5}$$

We then factorize $p(\mathbf{s})$ as the prior on individual s_i:

$$p(\mathbf{s}) = \prod_{i=1}^{m} p(s_i) \tag{6}$$

We can now represent the invertible linear transformation from the input space to the feature space by $\mathbf{W}$ (i.e. $\phi(\mathbf{x}) = \mathbf{s} = \mathbf{W}\mathbf{x}$) and use $p(\mathbf{s})$ as the prior in Bayesian inference (i.e. $p(\phi(\mathbf{x})) = p(\mathbf{s})$).

2.2. Encoding and Decoding

We continue by defining voxel-based encoding models. We use multiple linear regression to define the voxel-based encoding models by a weighted sum of the linear feature detector outputs for responses $1 \leq i \leq q$:

$$y_i = \boldsymbol{\beta}_i^\top \mathbf{s} + \varepsilon_i \tag{7}$$

where $\boldsymbol{\beta}_i \in \mathbb{R}^m$ are vectors of regression coefficients and $\varepsilon_i \in \mathbb{R}$ are Gaussian noise such that $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. We then use the multivariate Gaussian distribution to define the encoding distribution:

$$p\left(\mathbf{y} | \phi\left(\mathbf{x}\right)\right) = \mathcal{N}\left(\mathbf{B}^\top \mathbf{s}, \boldsymbol{\Sigma}\right) \tag{8}$$

where $\mathbf{B} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_q) \in \mathbb{R}^{m \times q}$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, ..., \sigma_q^2) \in \mathbb{R}^{q \times q}$.

Combining the prior and the encoding distribution using Bayes' theorem results in the decoding distribution:

$$p(\mathbf{s}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{s})p(\mathbf{s}) \tag{9}$$

Having defined the invertible linear transformation from the input space to the feature space and the decoding distribution, we can finally solve the problem of reconstructing $\mathbf{x}$ from $\mathbf{y}$ using maximum a posteriori (MAP) estimation to obtain a point estimate of $\mathbf{s}$ (i.e. $\hat{\mathbf{s}}_{\mathrm{MAP}} = \arg\max_{\mathbf{s}}\{p(\mathbf{y}|\mathbf{s})p(\mathbf{s})\}$) and synthesizing $\hat{\mathbf{x}}$ from the statistical generative model of images (i.e. $\hat{\mathbf{x}} = \mathbf{A}\hat{\mathbf{s}}_{\mathrm{MAP}}$).

2.3. Experimental Validation

For unsupervised feature learning, we used the entire MNIST database of handwritten digits, without the labels (LeCun et al., 1998). That is, the training set consisted of 70000 unlabeled grayscale images of 28 × 28 pixels in 10 categories (i.e. handwritten zeros through handwritten nines). We preprocessed the images by centering, PCA whitening and dimensionality reduction such that we retained the least possible number of principal components that account for 90% of the variability in the images (i.e. the first 64 principal components). We estimated the parameters of the statistical generative model (Fig. 3) using the FastICA algorithm (Hyvärinen, 1999).

For encoding and decoding, we used the dataset originally published in van Gerven et al. (2010a) and van Gerven et al. (2010b). Briefly, it consisted of estimated peak fMRI responses to grayscale images of handwritten sixes and handwritten nines. The training set consisted of 80 stimulus-response pairs, and the test set consisted of the remaining 20 stimulus-response pairs. Both the training set and the test set had equal number of stimuli from each of the two categories (i.e. handwritten sixes and handwritten nines). We preprocessed the images as in unsupervised feature learning. We trained the voxel-based encoding models using kernel ridge regression and performed hyperparameter optimization using grid search with a nested

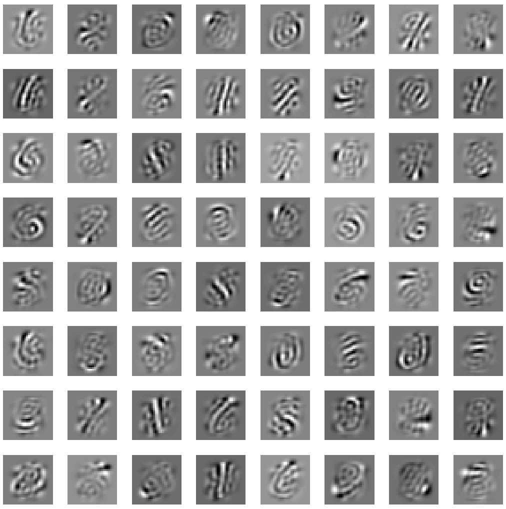

Figure 3. 64 linear feature detectors learned from the MNIST database using the FastICA algorithm.

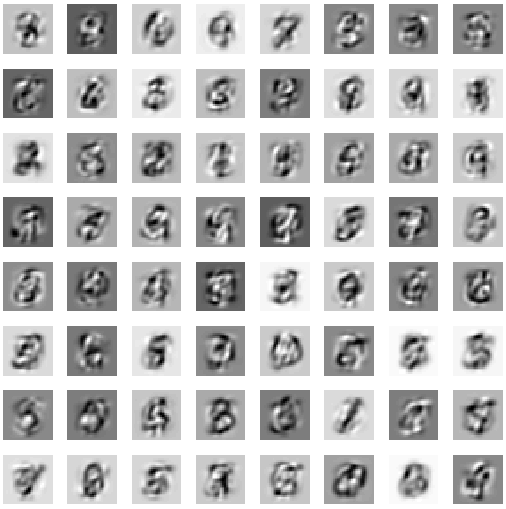

Figure 4. 64 images synthesized from the statistical generative model by sampling linear feature detector outputs from the estimated distributions of the latent independent components.

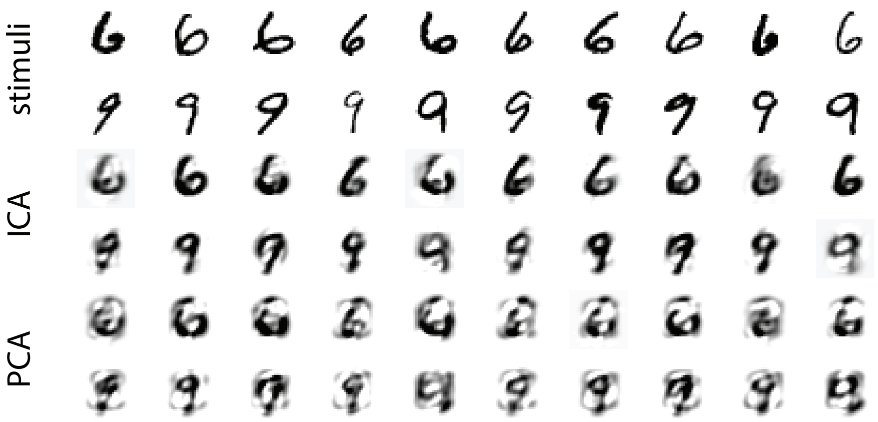

Figure 5. Stimuli, ICA reconstructions and PCA reconstructions.

leave-one-out cross validation on the training set. We computed the MAP estimate of the linear feature detector outputs using the minFunc implementation of the limited-memory BFGS algorithm (Schmidt, 2005). For comparative purposes, we used another framework based on preprocessed images (i.e. PCA features), with $\phi(\mathbf{x}) = \mathbf{x}$ and $p(\phi(\mathbf{x})) = \mathcal{N}(0, \mathbf{I}_m)$.

3. Results

We first examined the linear feature detectors (Fig. 3) and synthesized 64 images from the statistical generative model (Fig. 4) by sampling linear feature detector outputs from the estimated distributions of the latent independent components (Figure 2). Visual inspection shows that the linear feature detectors are tuned for meaningful features that resemble pen-strokes, which is consistent with the results in the literature (Ranzato et al., 2006; Lee et al., 2007). That is, the statistical generative model describes images as a linear transformation of pen-strokes. Furthermore, the synthesized images resemble handwritten digits, which suggests that the statistical generative model captures image statistics of handwritten digits.

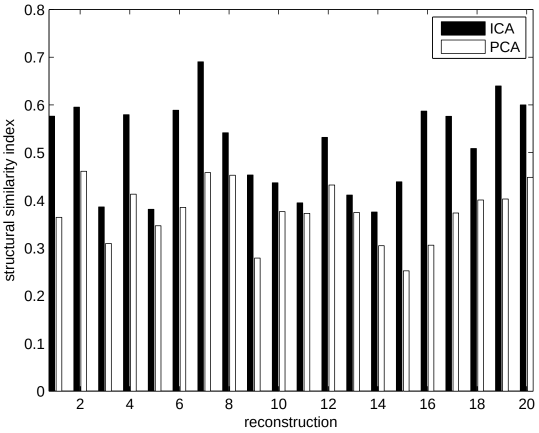

Figure 6. Reconstruction accuracy quantified by computing the structural similarity index.

We then quantified the encoding performance by computing explained variance per voxel. The difference between the mean explained variance of the two frameworks was not significant (p > 0.05), with both of the two frameworks having a state-of-the art encoding performance.

We finally evaluated the reconstruction accuracy. Visual inspection shows that our framework has more accurate reconstructions (Fig. 5). We quantified the reconstruction accuracy by computing the structural similarity index (Wang et al., 2004) per reconstruction (Fig. 6). The difference between the mean structural similarity indices of the two frameworks was significant (p < 0.05), with our framework having a higher structural similarity index for each of the 20 reconstructions.

4. Conclusion

Here, we introduced a new framework that combines unsupervised feature learning and Bayesian decoding. We validated our framework by accurately reconstructing grayscale images of handwritten sixes and handwritten nines from stimulus-evoked BOLD haemodynamic responses.

The significant improvement in the reconstruction accuracy, but not the encoding performance, demonstrated the importance of prior information in Bayesian decoding. Using the statistical generative model defined using ICA as a prior in Bayesian decoding results in significantly better reconstructions since ICA captures the image statistics of grayscale handwritten digits better than PCA.

Our framework can be extended beyond grayscale images of handwritten characters, both within the visual modality (e.g. any combination of larger images, natural images, color images, stereo images, temporal sequences of images) and across modalities (e.g. the auditory modality). Furthermore, our statistical generative model can be extended into multiple layers to learn hierarchical features of the stimuli. It remains an open question whether we can accurately reconstruct such complex stimuli from stimulus-evoked BOLD haemodynamic responses.

In conclusion, our results show that the features have a significant effect on the reconstruction accuracy. We also demonstrated that independent component analysis captures the image statistics of grayscale handwritten digits and provides an effective means for unsupervised learning of features for Bayesian decoding in fMRI that can mitigate the limitations of hand-designing features.

Acknowledgments

This work was partially supported by the Academy Assistants Programme of the Royal Netherlands Academy of Arts and Sciences. The authors would like to thank the reviewers for their useful comments.

References

Bengio, Y., Courville, A. C., & Vincent, P. (2012). Unsupervised feature learning and deep learning: a review and new perspectives. *arXiv:1206.5538.*

Brown, E. N., Kass, R. E., & Mitra, P. P. (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, *7*, 456–61.

Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* The MIT Press.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–30.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*, 626–634.

Hyvärinen, A. (2010). Statistical models of natural images and cortical visual representation. *Topics in Cognitive Science*, *2*, 251–264.

Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural Image Statistics.* Springer London.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*, 679–85.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*, 352–5.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*, 2278–2324.

Lee, H., Ekanadham, C., & Ng, A. (2007). Sparse deep belief net model for visual area V2. *Advances in Neural Information Processing Systems.*

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*, 1191–5.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., 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*, 915–29.

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*, 902–15.

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. *Current Biology*, *21*, 1641–6.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–9.

Ranzato, M. A., Poultney, C., Chopra, S., & LeCun, Y. (2006). Efficient learning of sparse representations with an energy-based model. *Advances in Neural Information Processing Systems.*

Schmidt, M. (2005). minfunc - unconstrained differentiable multivariate optimization in matlab. www.di.ens.fr/˜schmidt/Software/minFunc.html.

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*, 1104–16.

van Gerven, M., Cseke, B., de Lange, F. P., & Heskes, T. (2010a). Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*, *50*, 150–61.

van Gerven, M., de Lange, F. P., & Heskes, T. (2010b). Neural decoding with hierarchical generative models. *Neural Computation*, *22*, 3127–42.

van Gerven, M., & Heskes, T. (2010). Sparse orthonormalized partial least squares. *Benelux Conference on Artificial Intelligence.*

van Gerven, M., & Heskes, T. (2012). A linear Gaussian framework for decoding of perceived images. *2012 Second International Workshop on Pattern Recognition in NeuroImaging* (pp. 1–4).

Wang, Z., Bovik, A. C., Sheikh, H. R., & P Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*, 600–612.