ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 39 (2008) 181 – 205

Bayesian decoding of brain images

Karl Friston,[a,*] Carlton Chu,[a] Janaina Mourão-Miranda,[b] Oliver Hulme,[a] Geraint Rees,[a] Will Penny,[a] and John Ashburner[a]

[a]*Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen Square, London WC1N 3BG, UK*
[b]*Biostatics Department, Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, UK*

Received 9 July 2007; revised 7 August 2007; accepted 12 August 2007
Available online 24 August 2007

This paper introduces a multivariate Bayesian (MVB) scheme to decode or recognise brain states from neuroimages. It resolves the ill-posed many-to-one mapping, from voxel values or data features to a target variable, using a parametric empirical or hierarchical Bayesian model. This model is inverted using standard variational techniques, in this case expectation maximisation, to furnish the model evidence and the conditional density of the model's parameters. This allows one to compare different models or hypotheses about the mapping from functional or structural anatomy to perceptual and behavioural consequences (or their deficits). We frame this approach in terms of decoding measured brain states to predict or classify outcomes using the rhetoric established in pattern classification of neuroimaging data. However, the aim of MVB is not to predict (because the outcomes are known) but to enable inference on different models of structure–function mappings; such as distributed and sparse representations. This allows one to optimise the model itself and produce predictions that outperform standard pattern classification approaches, like support vector machines.

Technically, the model inversion and inference uses the same empirical Bayesian procedures developed for ill-posed inverse problems (e.g., source reconstruction in EEG). However, the MVB scheme used here extends this approach to include a greedy search for sparse solutions. It reduces the problem to the same form used in Gaussian process modelling, which affords a generic and efficient scheme for model optimisation and evaluating model evidence. We illustrate MVB using simulated and real data, with a special focus on model comparison; where models can differ in the form of the mapping (i.e., neuronal representation) within one region, or in the (combination of) regions per se.

Keywords: Parametric empirical Bayes; Expectation maximisation; Gaussian process; Automatic relevance determination; Relevance vector machines; Classification; Multivariate; Support vector machines; Classification

Introduction

The purpose of this paper is to describe an empirical Bayesian approach to the multivariate analysis of imaging data that brings pattern classification and prediction approaches into the conventional inference framework of hierarchical models and their inversion. The past years have seen a resurgence of interest in the multivariate analysis of functional and structural brain images. These approaches have been used to infer the deployment of distributed representations and their perceptual or behavioural correlates. In this paper, we try to identify the key issues entailed by these approaches and use these issues to motivate a better approach to estimating and making inferences about distributed neuronal representations.

This paper comprises three sections. In the first, we review the development of multivariate analyses with a special focus on three important distinctions; the difference between mass-univariate and multivariate models, the difference between generative and recognition models and the distinction between inference and prediction. The second section uses the conclusions of the first section to motivate a simple hierarchical model of the mapping from observed brain responses to a measure of what those responses encode. This model allows one to compare different forms of encoding, using conventional model comparison. In the final section, we apply the multivariate Bayesian model of the second section to real fMRI data and ask where and how visual motion is encoded. We also show that the ensuing model outperforms simple classification devices like linear discrimination and support vector machines. We conclude with a discussion of generalisations; for example, nonlinear models and the comparison of multiple conditions to disambiguate between functional selectivity and segregation in the cortex.

Multivariate models and classification

Mappings and models

In this section, we review multivariate approaches and look at the distinction between inference and prediction. This section is written in a tutorial style in an attempt to highlight some of the

* Corresponding author. Fax: +44 207 813 1445.
E-mail address: k.friston@fil.ion.ucl.ac.uk (K. Friston).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.08.013

basic concepts underlying inference on structure–function mappings in the brain. We try to link the various approaches that have been adopted in neuroimaging and identify the exact nature of inference these approaches support.

The question addressed in most applications of multivariate analysis is whether distributed neuronal responses encode some sensorial or cognitive state of the subject (for a review, see Haynes and Rees, 2006). Universally, this entails some form of model comparison, in which one compares a model that links neuronal activity to a presumed cognitive state with a model that does not. The link can be from the neuronal measure or response variable, $Y\in\Re^n$ to an experimental or explanatory variable, $X\in\Re^v$, or the other way around. From the point of view of inferring a link exists, its direction is not important; however, the form of the model may depend on the direction. This becomes important when one wants to compare different models (as we will see below). In current fMRI analysis, inference on models or functions that map $g{:}X\rightarrow Y$ include conventional mass-univariate models as employed by statistical parametric and posterior probability mapping (that use classical and Bayesian inference respectively; Friston et al., 2002) or classical multivariate models such as canonical variate analysis. The converse mapping from $h{:}Y\rightarrow X$ is used by classification schemes, such as linear discriminant analysis and support vector machines. Typically $Y\in\Re^n$ has many more elements or dimensions than $X\in\Re^v$ (i.e., $n>v$). For example, $X\in\Re$ could be a scalar or label indicating whether motion is present in the visual field and $Y\in\Re^n$ could be the fMRI signal from a thousand voxels, in a visual cortical area. Similarly, X could be a label indicating whether a subject has Alzheimer's disease and Y could be the grey matter density over the entire brain. In what follows, we review some of the basics of inference that are needed to understand the relationship between model comparison and classification.

Marginal likelihoods and statistical dependencies

We can reduce the problem of linking observed brain responses to their causes (in the case of perception) or consequences (in the case of behaviour) to establishing the existence of some mapping, $g{:}X\rightarrow Y$; in other words, inferring that there is some statistical dependency between the experimental variable and measured response (or the other way around). If this mapping exists, we can infer that brain states cause or are caused by X. This means we can formulate our question in terms of a null hypothesis H_0 that there is no dependency, in which case the measurements are equally likely, whether or not we know the experimental variable; $p(Y|X)=p(Y)$. The Neyman–Pearson lemma states that the likelihood ratio test

$$\Lambda = \frac{p(Y|X)}{p(Y)} \geq u \tag{1}$$

is the most powerful test of size $\alpha=p(\Lambda\geq u|H_0)$ for testing this hypothesis. Generally, the null distribution of the likelihood ratio statistic $p(\Lambda|H_0)$ is determined non-parametrically or under parametric assumptions (e.g., a t-test). The likelihood ratio, $\Lambda(Y)$ underlies most statistical inference and model comparison and is the basis of nearly all classical statistics; ranging from Wilk's Lambda in canonical correlation analysis to the F ratio in analysis of variance.

In Bayesian inference, the likelihood ratio is known as a Bayes factor (Kass and Raftery, 1995) that compares models of Y with and without X. Usually, in Bayesian model comparison, one uses the log-likelihoods directly to quantify the relative likelihood of two models

$$\ln \Lambda = \ln p(Y|X) - \ln p(Y) \geq u \tag{2}$$

where u is generally three (see Penny et al., 2004). This means the first model is at least $20\approx\Lambda=\exp(3)$ times more likely than the second, assuming both models are equally likely a priori. Another way of expressing this is to say that one model is $0.95\approx\Lambda/(\Lambda+1)$ more likely than the other, given the data. We will use both classical and Bayesian inference in this paper.

Evaluating the marginal likelihood

To evaluate the likelihood ratio, we need to evaluate the likelihood under the null hypothesis and under some mapping. To do this, we need to posit a probabilistic model of the mapping $g(\theta)$: $X\rightarrow Y$ and integrate out the dependence on the unknown parameters of the mapping, θ. This gives the marginal likelihood (also known as the integrated likelihood or evidence)

$$p(Y|X) = \int p(Y,\theta|X)d\theta \tag{3}$$

This marginalisation requires the joint density $p(Y,\theta|X)=p(Y|\theta,X)p(\theta)$ that is usually specified in terms of a likelihood, $p(Y|\theta,X)$ and a prior, $p(\theta)$. In general, the integral above cannot be evaluated analytically. This problem can be finessed by converting a difficult integration problem into an easy optimisation problem; by optimizing a [free energy] bound on the evidence with respect to an arbitrary density $q(\theta)$

$$F = \int q(\theta)\ln\frac{p(Y,\theta|X)}{q(\theta)}d\theta = \ln p(Y|X) - D(q(\theta)||p(\theta|Y,X)) \tag{4}$$

When this bound is maximised, the Kullback–Leibler divergence $D(q\|p(\theta|Y,X))$ is minimised and $q(\theta)\approx p(\theta|Y,X)$ becomes an approximate conditional or posterior density on the parameters. Coincidentally, the free energy[1] becomes the log-evidence, $F\approx\ln p(Y|X)$. All estimation and inference schemes based on parameterised density functions can be formulated in this way; from complicated extended Kalman filters for dynamic systems to the simple estimate of a sample mean. The only difference among these schemes is the form assumed for $q(\theta)$ and how easy it is to maximise the free energy by optimising its sufficient statistics (e.g., conditional mean and covariance) of $q(\theta)$. Because the bound, $F(q(\theta))$ is a function of a function, the optimisation rests on the method of variations (Feynman, 1972); this is why the above approach is know as variational learning (for a comprehensive discussion, see Beal, 1998). This may seem an abstract way to motivate the specification of a model; however, it is a useful perspective because it highlights the difference between the role of $q(\theta)$ in inference and prediction (see below).

The free energy bound on the log-evidence plays a central role in what is to follow; in that it quantifies how good a model is, in relation to another model. The free energy can be expressed in terms of accuracy and complexity terms (Penny et al., 2004), such that the best model represents the optimum trade-of between fit and

[1] The free energy in this paper is the negative free energy is statistical physics. This means the free energy and log-evidence have the same sign.

parsimony. This trade-off is known as Occam's razor, which is sometimes formulated as a minimum description length principle (MDL). MDL is closely connected to probability theory and statistics through the correspondence between codes and probability distributions. This has led some to view MDL as equivalent to Bayesian inference, for particular classes of model: in MDL, the code length of the model and the code length of model and data together, correspond to the prior probability and marginal likelihood respectively in the Bayesian framework (see MacKay, 2003; Grunwald et al., 2005).

In summary, inference can be reduced to model comparison, which rests on the marginal likelihood of each model. To evaluate the marginal likelihood it is necessary to specify the parametric form of the joint density entailed by the model. Integrating out dependency on the parameters of this model rests on optimising a bound on the marginal likelihood with respect to a density, $q(\theta)$. Optimisation makes $q(\theta)$ the conditional density on the unknown parameters (i.e., an implicit estimation). The implication is that parameter estimation is a necessary and integral part of model comparison. The key thing to take from this treatment is that inference about how the brain represents things reduces to model comparison. This comparison is based on the marginal likelihood or evidence for competing models of how neurophysiological variables map to observed responses, or vice versa. Next, we look at the most prevalent model in neuroimaging.

The general linear model and canonical correlation analysis

The simplest model is a linear mapping under Gaussian assumptions about random effects; i.e., $\varepsilon \sim N(0,\Sigma)$

$$Y = X\beta + \varepsilon \Rightarrow p(Y|\theta, X) = N(X\beta, \Sigma(\lambda)) \tag{5}$$

where $N(\mu,\Sigma)$ denotes a normal or Gaussian density with mean μ and covariance Σ and the unknown parameters, $\theta=\{\beta,\lambda\}$ control the first and second moments (i.e., mean and covariance) of the likelihood respectively. This is the general linear model, which is the cornerstone for neuroimaging data analysis. We will restrict our discussion to linear models because they can be extended easily to cover nonlinear mappings; these extensions use nonlinear projections onto a high-dimensional feature space of the experimental data (e.g., Büchel et al., 1998) or the images, using kernel methods. Kernel methods are a class of algorithms for pattern analysis, whose best known example is the support vector machine (SVM). Kernel methods transform data into a high-dimensional feature space, where a linear model is applied. This converts a difficult low-dimensional nonlinear problem into an easy high-dimensional linear problem.

Under the general linear model (GLM), it is easy to show (see Friston, 2007) that the log-likelihood ratio is simply the mutual information between X and Y

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= \ln\Lambda \end{aligned} \tag{6}$$

where $H(Y)=-\int p(Y)\ln p(Y)dY$ is the entropy or expected surprise. In other words, $\ln\Lambda$ reflects the reduction in surprise about observed data that is afforded by seeing the explanatory variables. Crucially, this is exactly the same reduction in surprise about the explanatory variables, given the data. This symmetry, i.e., $I(X,Y)=I(Y,X)$, means that we can swap the explanatory and response variables in a general linear model with impunity. This is one perspective on why the inference scheme for GLMs, namely canonical *correlation* analysis (CCA) does not distinguish between explanatory and response variables.

Canonical correlation analysis (CCA), also known as canonical variate analysis (CVA), computes the likelihood ratio using generalised eigenvalues solutions of Y^TY explained and not explained by X. In this context, Λ is known as Wilk's Lambda and is a composition of generalised eigenvalues (also know as canonical values). Canonical correlation analysis is fundamental to inference on general linear models and subsumes simpler variants like MANCOVA, Hotellings T^2 test, partial least squares, linear discriminant analysis and other *ad hoc* schemes. One might ask, if CCA provides the optimal inference (by the Neyman–Pearson Lemma) for GLMs, why is it not used in conventional analyses of fMRI data with the GLM? In fact, conventional mass-univariate analyses do use a special case of CCA, namely ANCOVA.

Multivariate vs. mass-univariate

The mass-univariate approach to identifying the mapping $g(\theta)$: $X \rightarrow Y$ is probably the most common in neuroimaging, as exemplified by statistical parametric mapping (SPM). These approaches treat each data element (i.e., voxel) as conditionally independent of all other voxels such that the implicit likelihood factorises over voxels, indexed by i

$$p(Y|X,\theta) = \prod_i p(Y_i|\theta_i, X) \tag{7}$$

In the classification literature, this would be called a naive Bayes classifier (also known as Idiot's Bayes) because the underlying probability model rests on conditionally independent data features. In SPM, the spatial dependencies among voxels are introduced after estimation during inference, through random field theory. Random field theory provides a model for the prevalence of topological features in the SPM under the null hypothesis, such as the number of peaks above some threshold. This allows one to make multivariate inferences over voxels (e.g., set-level inference; Friston et al., 1996). The advantage of topological inference is that random field theory provides a very efficient model for spatial dependences that is based on the fact that images are continuous; other multivariate models ignore this. The disadvantage of random field theory is that the p-value is not based on a likelihood ratio and is therefore suboptimal by the Neyman–Pearson lemma. However, SPM is not usually used to make multivariate inference because it is used predominantly to find regionally specific effects.

Multivariate models relax the naive independence assumption and enable inference about distributed responses.[2] The first multivariate models of imaging data (scaled sub-profile model: Moeller et al., 1987) appeared in the nineteen eighties and focused on disambiguating global and regionally specific effects using principal component analysis. Principal component analysis also featured in early data-led multivariate analyses of Alzheimer's disease (e.g., Grady et al., 1990). The first canonical correlation analysis of functional imaging data addressed the mapping between resting regional cerebral activity and the expression of

[2] Although it is difficult to generalise, multivariate inference is usually more powerful than mass-univariate topological inference because the latter depends on focal responses that survive some threshold (and induce a topological feature).

symptoms in schizophrenia. This analysis showed that distinct brain systems correlated with distinct sub-syndromes of schizophrenia (Friston et al., 1992a).

In Friston et al. (1995), we generalised canonical correlation analysis to cover all the voxels in the brain: The problem addressed in that paper was that CCA requires the number of observations to be substantially larger than the dimensionality of the data features (i.e., number of voxels) or experimental variables. Clearly, in imaging, the number of voxels exceeds the number of scans. This means that one cannot estimate the marginal likelihood because there are insufficient degrees of freedom to estimate the covariance parameters, $\lambda \subset \theta$. This problem can be finessed by invoking priors $p(\theta)$ or constraints on the parameters. In Friston et al. (1995), the parameters were constrained to a low-dimensional subspace, spanned by the major singular vectors, U of the data. This effectively re-parameterised the model in terms of a smaller number of parameters; $\tilde{\beta} = \beta U \Leftrightarrow \beta = \tilde{\beta} U^T$. Major singular vectors (i.e., eigenimages) span the greatest variance seen in the data and are identified easily using singular value decomposition (SVD).

This dimension reduction furnishes a constrained linear model

$$\begin{aligned} \tilde{Y} &= X\tilde{\beta} + \tilde{\varepsilon} \\ \tilde{Y} &= YU \quad \tilde{\beta} = \beta U \quad \tilde{\varepsilon} = \varepsilon U \end{aligned} \tag{8}$$

which can be treated in the usual way. We will revisit the use of singular vectors in the context of multivariate Bayesian models below and contrast them with the use of support vectors.

Worsley et al. (1998) used canonical variates analysis (CVA) of the estimated effects of predictors from a multivariate linear model. The advantage of this, over previous methods, was that temporal correlations could be incorporated into the model, making it suitable for fMRI data. CCA has re-appeared in the neuroimaging literature over the years (e.g., Friman et al., 2001). An interesting application of CCA was presented in Nandy and Cordes (2003) where the analysis was repeated over small regions of the brain, thereby eschewing the dimensionality problem. The same idea of using a multivariate 'searchlight' has been exploited recently (Kriegeskorte et al., 2006). These authors used a Mahalanobis distance statistic that is closely related to Hotellings T^2 statistic (a special case of Wilk's Lambda that obtains when X is univariate).

The key point here is that constraints on the dimensionality of $Y \in \Re^n$ or, equivalently, priors on the parameters, become essential when dealing with high-dimensional feature spaces, which are typical in imaging. The same theme emerges when we look at pattern classifiers in imaging.

Generative, recognition and classification models

In the recent neuroimaging literature one often comes across the phrase: 'novel multivariate pattern classifiers'. This section tries to argue that multivariate models and pattern classification should not be conflated and that neither are novel. Critically, it is the multivariate mapping from brain measurements to their consequences that characterise recent advances; classification *per se* is somewhat incidental.

The first formal classification scheme for functional neuroimaging was reported in Lautrup et al. (1994). These authors used nonlinear neural network classifiers to classify images of cerebral blood flow according to the experimental conditions (i.e., causes), under which the images were acquired. In this application, constraints on the mapping from the high-dimensional feature (voxel) space to target class were imposed through massive weight sharing. Classifiers have played a prominent role in structural neuroimaging (e.g., Herndon et al., 1996) and are now an integral part of computational anatomy and segmentation schemes (e.g., Ashburner and Friston, 2005). However, classification schemes received little attention from the functional neuroimaging community until they were re-introduced in the context of mind-reading (Carlson et al., 2003; Cox and Savoy, 2003; Hanson et al., 2004; Haynes and Rees, 2005; Norman et al., 2006; Martinez-Ramon et al., 2006).

So far, we have limited the discussion to parameterised mappings $g(\theta): X \rightarrow Y$ from experimental labels to data features. In a probabilistic setting, these can be considered as generative functions or models of experimental causes that produce observed data. Indeed experimental neuroscience rests on comparing generative models that embody competing hypotheses about how data are caused. However, one can also parameterise the inverse mapping from data to causes; $h(\theta): Y \rightarrow X$, to provide a function of the data that recognises what caused them. These are called recognition models. What is the relationship between recognition models and prediction in classification schemes? In classification, one wants to predict or classify a new observation Y_{new} using a recognition model whose parameters have been estimated using training data and classification pairs. Classification is based on the predictive density

$$p(X_{\text{new}} | Y_{\text{new}}, X, Y) = \int p(X_{\text{new}} | \theta, Y_{\text{new}}) q(\theta) d\theta \tag{9}$$

where $q(\theta) = p(\theta | X, Y)$ is the conditional density. Classification, or more generally prediction, is fundamentally different from inference on the model or mapping *per se*: In prediction, one uses $q(\theta)$ to make an inference about an unknown label, X_{new}, in terms of the predictive density, $p(X_{\text{new}} | Y_{\text{new}}, X, Y)$. In experimental neuroscience, this label is known and inference is on the mapping itself; e.g., $h(\theta): Y \rightarrow X$. In short, one uses $q(\theta)$ to evaluate the marginal likelihood, $p(X|Y)$, as opposed to the predictive density. In other words, the predictive density is not used to address whether the prediction is possible or whether there is a better predictor, these questions require inference on models; prediction requires only inference on the target, given a model.

The only situation that legitimately requires us to predict what caused a new observation is when we do not know that cause. An important example is brain computer interfacing, where a subject is trying to communicate through measured brain activity. Other examples include automated diagnostic classification or the classification of tissue type in computational anatomy mentioned above. In summary, the predictive density plays no role in testing hypotheses about the mapping between causes and data features; these inferences are based on the marginal likelihood of the model.

Support vector machines

Many classification schemes (e.g., support vector machines) do not even try to estimate the predictive density; they simply optimise the parameters of the recognition function to maximise accuracy. These schemes can be thought of as using point estimates of θ, which ignore uncertainty about the parameters inherent in $q(\theta)$. We will refer to these as point classifiers, noting that probabilistic formulations are usually available (e.g., variational relevance vector machines; see Bishop and Tipping, 2000). Support vector machines (Vapnik, 1999) are point classifiers that

use a recognition model; for example, a linear SVM assumes the mapping $h(\theta):Y\rightarrow X$

$$\begin{aligned} X &= K(Y)\tilde{\beta} + \varepsilon \\ K(Y) &= YY^T \\ \beta &= Y^T\tilde{\beta} \end{aligned} \tag{10}$$

Here $K(Y)$ is called a kernel function, which, in this case, is simply the inner product of the data features (i.e., images). The important thing to note here is that the parameters, $\beta=Y^T\tilde{\beta}$ of the implicit recognition model are constrained to lie in a subspace spanned by the images. This is formally related to the constraint used in CCA for images; $\beta=\tilde{\beta}U^T$. In other words, both constrained CCA and SVM require the parameters to be a mixture of data features. The key difference is that constrained CCA imposes sparsity by using a small number of informed basis sets (i.e., singular vectors), whereas SVM selects a small number of original images (i.e., support vectors). These support vectors define the maximum margin hyperplane separating two classes of observations encoded in $X\in[-1,1]$. These classifiers are also known as maximum margin classifiers. We introduce the linear SVM because it will be used in comparative evaluations later.

Gaussian process models

Support vector machines (for $X\in[-1,1]$) and regression (for continuous targets; $X\in\Re$) are extremely effective prediction schemes, in a high-dimensional setting. However, from a Bayesian perspective, they rest on a rather *ad hoc* form of recognition model (their motivation is based on statistical learning theory and structural risk minimisation; Vapnik, 1999). Over the same period that support vector approaches were developed, Gaussian process modelling (Ripley, 1994; Rasmussen, 1996; Kim and Ghahramani, 2006) has emerged as an alternative and generic approach to prediction (for an introduction, see MacKay, 1997): The basic idea behind Gaussian process modelling is to replace priors $p(\theta)$ on the parameters of the mapping, $h(\theta):Y\rightarrow X$ with a prior on the space of mappings; $p(h(Y))$, where the mappings or functions themselves can be very complex and highly nonlinear. This is perfectly sufficient for prediction and model comparison because the predictive density $p(X_{\text{new}}|Y_{\text{new}},X,Y)$ and marginal likelihood $p(X|Y)$ are not functions of the parameters. The simplest form of prior is a Gaussian process prior, which leads to a Gaussian likelihood; $p(X|Y,\lambda)=N(0,\Sigma(Y,\lambda))$. This is specified by a Gaussian covariance, $\Sigma(Y,\lambda)$, whose elements are the covariance between the values of the function or prediction, $h(Y)$ at the two points in feature space. The covariance $\Sigma(Y,\lambda)$ is optimised, given training data, in terms of covariance function hyperparameters, λ. This optimisation provides a nice link with classical covariance component estimation and techniques like restricted maximum likelihood (ReML) hyperparameter estimation (Harville, 1977).

We will use this approach below; however, our covariance functions are constrained by simple linear mappings, of different sorts, between features and targets. After $\Sigma(Y,\lambda)$ has been optimised with respect to the free energy bound above, it can be used to evaluate the marginal likelihood and infer on the model it encodes. Typically, in Gaussian process modelling, one uses maximum likelihood or *a posteriori* point estimates of the hyperparameters to approximate the marginal likelihood; here, we marginalise over the hyperparameters using their conditional density to get more accurate estimates (see also MacKay, 1999, who discusses related issues under the evidence framework used below).

Inference vs. prediction

Some confusion about the roles of prediction and inference may arise from the use of classification performance to infer a significant relationship between data features and perceptual or behavioural states. There is a fundamental reason why some classification schemes have to use their classification performance to make this sort of inference: This is because point classifiers are not probabilistic models, which means their evidence is not defined: recall that a model is necessary to specify a form for the joint density of the data and unknown model parameters. Integrating out the dependency on the parameters provides the marginal likelihood that is necessary for inference about that model. In short, model inversion optimises the conditional density of the parameters to maximise the marginal likelihood. In contradistinction, point classification schemes optimise the parameters to maximise accuracy. This is problematic in two ways.

First, point classification schemes do not furnish a measure of the marginal likelihood and cannot be used for inference. This means that the model evidence has to be evaluated indirectly through cross-validation: Cross-validation (sometimes called rotation–estimation), involves partitioning the data into subsets such that the analysis is performed on one (*training*) subset, while the other (*test*) data are retained to confirm and validate the initial analysis.[3] A significant mapping can be inferred if the performance on the test subset exceeds chance levels. However, by the Neyman–Pearson lemma, this inference is suboptimal because it does not conform to a likelihood ratio test on the implicit recognition model. Having said this, cross-validation can be very useful for classical inference when the null distribution of the likelihood ratio statistic is unavailable; for instance when it is analytically intractable or it is computationally prohibitive to compute using sampling techniques (see also Lukic et al., 2002). In this context, classification can be used as surrogate statistic because the null distribution of predictive performance can be derived easily (e.g., a binomial distribution for chance classification into two classes). We will use cross-validation *p*-values for classical inference below.

The second problem for classifiers is that the marginal likelihood depends on both accuracy and model complexity (see Penny et al., 2004). However, many classification schemes do not minimise complexity explicitly. This shortcoming can be ameliorated in two ways. The first is to minimise complexity through the use of formal constraints (cf. the sparsity assumptions implicit in SVM). The second is to optimise the recognition model parameters (e.g., the parameter C in SVM, which controls the width of the maximum margin hyperplane) with respect to generalisation error (i.e., the classification error on test data). However, to evaluate the generalisation error one needs to know the classes and therefore there is no need for classification. In summary, classification *per se* appears to play an incidental role in answering key questions about structure–function relationships in brain imaging, so why have they excited so much interest?

[3] k-fold cross validation involves randomly partitioning the data into k partitions, training the classifier on all but one and evaluating classification performance on that partition. This procedure is repeated for all k partitions.

Encoding and decoding models

When one looks closely at pattern recognition or classification schemes in functional neuroimaging they have been used as generative models, not recognition models; they have been used to test models of how physical brain states generate percepts, behaviours or deficits. For example, studies looking for perceptual correlates in visual cortex are not trying to recognise the causes of physiological activations; they are modelling the perceptual products of neuronal activity. Perhaps an even clearer example comes from recent developments in computational anatomy, where multivariate data-mining methods have been used to study lesion-deficit mappings. Here, the imaging data are used as a surrogate marker of the lesion and resulting behavioural deficits are modelled using Bayesian networks (Herskovits and Gerring, 2003).

In short, the key difference between conventional multivariate analyses and so-called classification schemes does not rest on classification; the distinction rests on whether X causes Y, e.g., stimulus motion causes activation in V5; or whether Y causes X, e.g., activation of V5 causes a percept of motion. Both are addressed by inference on models but, in the latter case, the experimental variable X is a consequence not a cause. Put simply, the important distinction is whether the experimental variable is a cause or consequence. If it is a cause then the appropriate generative model is $g(\theta):X\rightarrow Y$; this could be called an *encoding* model in the sense that the brain responses are encoding the experimental factors that caused them. Conversely, if X is a consequence, we still have a generative model but the causal direction has switched to give, $g(\theta):Y\rightarrow X$. These have been called *decoding* models in the sense that they model the decoding of neuronal activity that causes a percept, behaviour or deficit (Hasson et al., 2004; Kamitani and Tong, 2006; Thirion et al., 2006). In some situations, the distinction is subtle but important. For example, using the presence of visual motion as a cause in an encoding model implies that X is a known deterministic quantity. However, using the presence of motion as a surrogate for motion perception means that X becomes a response or dependent variable reflecting the unknown perceptual state of the subject.

The importance of the distinction between encoding and decoding models is that we can disentangle inference from prediction and focus on the problem of inverting ill-posed decoding models of the form, $g(\theta):Y\rightarrow X$. Happily, there is a large literature on these ill-posed problems; perhaps the most familiar in neuroimaging is the source reconstruction problem in electro-encephalography (EEG). In this context, one has to estimate up to ten thousand model parameters (dipole-activities) causing observed responses in a small number of channels. Formally, this is like estimating the parameters coupling activity in thousands of voxels to a small number of experimental or target variables. In the next section, we will use exactly the same hierarchical linear models and their variational inversion used in source reconstruction (e.g., Phillips et al., 2005; Mattout et al., 2006) to decode functional brain images. Critically, this modelling perspective exposes the dependence of decoding models on prior assumptions about the parameters and their spatial disposition. These priors enter the EEG inverse problem in terms of spatial constraints on the sources (e.g., point sources in equivalent current dipole models *vs*. distributed solutions with smoothness constraints). The inversion scheme used below allows one to compare models that differ only in terms of their priors, using Bayesian model selection. This allows one to compare models of distributed or sparse coding that are specified in terms of spatial priors.

Summary

In summary, we have seen that:

- Inference on the mapping between neuronal activity and its causes or consequences rests on model comparison, using the marginal likelihood of competing models. The marginal likelihood requires the specification of a generative model prescribing the form of the joint density over observations and model parameters. This model may be explicit (e.g., a general linear model) or implicit (e.g., a Gaussian process model). Model inversion corresponds to optimising the conditional density of the model parameters to maximise the marginal likelihood (or some bound), which is then used for model comparison.
- Multivariate models can map from the causes of brain responses (encoding models; $g(\theta):X\rightarrow Y$) or from brain activity to its consequences (decoding models; $g(\theta):X\rightarrow Y$). In the latter case there is a curse of dimensionality, which is resolved with appropriate constraints or priors on model parameters. These constraints are part of the model and can be evaluated using model comparison in the usual way.
- Prediction (e.g., classification) and cross-validation schemes are not necessary for decoding brain activity but can provide surrogates for inference. This can be useful when the null distribution of the model likelihood ratio (i.e., Bayes factor) is not evaluated easily.

The next section describes a decoding model for imaging data sequences that can be inverted efficiently to give the marginal likelihood, which allows one to compare different priors on the model parameters.

A Bayesian decoding model

In this section, we describe a multivariate decoding model that uses exactly the same design matrices of experimental variables X and neuronal responses Y used in conventional analyses. Furthermore, the inversion scheme uses standard techniques that can be applied to any model with additive noise. It should be noted that the inversion of these models conforms to the free energy optimisation approach described above but is very simple and can be reduced to a classical covariance component estimation (for details, see Friston et al., 2007).

Hierarchical models

We want a simple model of how measured neuronal responses predict perceptual or behavioural outcomes (or their surrogates). Consider a linear mapping $X=A\beta$ between a scalar target variable, $X\in\Re$ and underlying neuronal activity in n voxels; $A\in\Re^n$; where X corresponds to a scan-specific measure of perceptual, cognitive or behavioural state induced by distributed activity A. Imagine that we obtain noisy measurements $Y\in\Re^{s\times n}$ of $A\in\Re^{s\times n}$ in s scans and n voxels (e.g., 128 scans and 1024 voxels from the lateral occipital cortex). Let $Y=TA+G\gamma+\varepsilon$ be observed signal, with noise, $\varepsilon\in\Re^{s\times n}$ and additive confounds, $G\in\Re^{s\times g}$ scaled by unknown parameters,

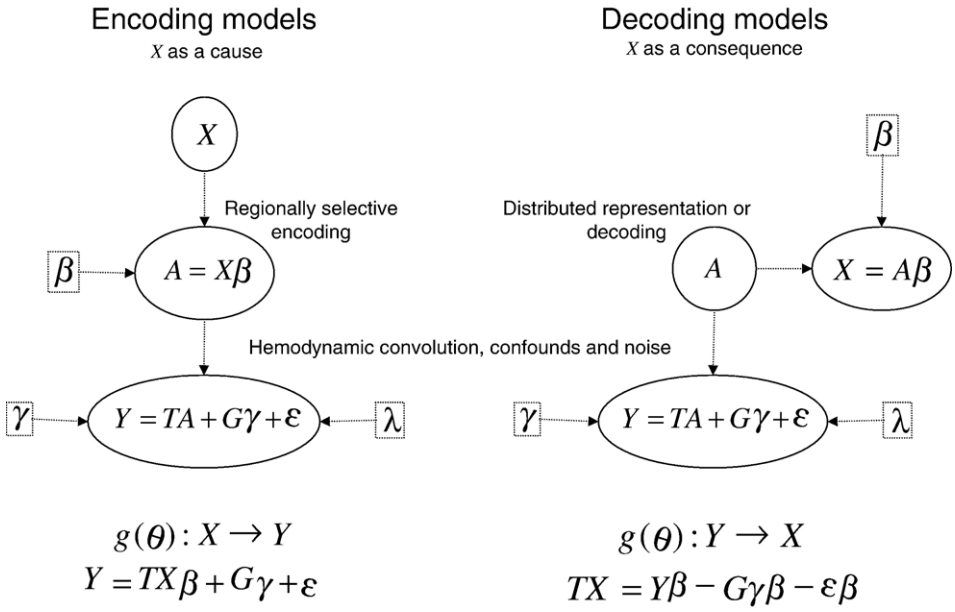


Fig. 1. Schematic highlighting the differences between encoding and decoding models, which couple measured brain responses to their causes (encoding) or consequences (decoding). The arrows denote conditional dependences. The variables are described in the main text.

$\gamma \in \Re^g$. Here, any effects of hemodynamics are modelled with the temporal convolution matrix $T \in \Re^{s \times s}$ embedding a hemodynamic response function. Clearly, for structural and PET data, this convolution is unnecessary and $T=I$ is simply the identity matrix.

Under the assumptions above, we can specify the following likelihood model under Gaussian assumptions about the noise (for a schematic summary, see Fig. 1)

$$X = A\beta \Rightarrow TX = TA\beta = Y\beta - G\gamma\beta - \varepsilon\beta \tag{11}$$

In this model, β are the unknown parameters of the mapping we want to infer; we will call these parameters *voxel weights*. Under the simplifying assumption that the temporal convolution matrix is known and is roughly the same for all voxels, this likelihood model is a weighted general linear model with serially correlated errors. Note that TX corresponds to a stimulus (or behavioural) function that has been convolved with a hemodynamic response function; this vector has the form of a regressor in the design matrix, $\mathbf{X}$, of conventional encoding models. In our implementation, we use $TX=\mathbf{X}c$ where the contrast weight vector, c, specifies the contrast to be decoded. Conversely, the confounds are the remaining effects; $G=\mathbf{X}(I-cc^{-})$, which ensures that $Gc=0$. Effectively, this partitions the conventional design matrix of explanatory variables into a target variable and confounds, where the target variable comes to play the role of a response variable that has to be predicted.

We can simplify this model by projecting the target and predictor variables onto the null space of the confounds to give a model for weighted target vectors

$$\begin{aligned} WX &= RY\beta + \varsigma \\ W &= RT \\ R &= orth(I - GG^{-})^T \end{aligned} \tag{12}$$

Here R is a residual forming matrix that removes confounds from the model. W is a weighting matrix that combines the residual forming and temporal convolution matrices to give a convolved target variable, with confounds removed. The fluctuations $\varsigma = -R\varepsilon\beta \in \Re^s$ are a vector of unknown random effects that retain their multivariate Gaussian distribution, where $\mathrm{cov}(\varsigma)=\Sigma^{\varsigma}=\exp(\lambda^{\varsigma})RVR^{\mathrm{T}}$. Here, λ^{ς} is some unknown covariance parameter or hyperparameter and V represents serial correlations or non-sphericity before projection.[4] The nice thing about decoding models is that we do not have to worry about spatial dependencies among the measurement noise (i.e., smoothness in images). This is because the random effects are a linear mixture of noise *over voxels*.

Empirical priors

There is a special aspect of decoding models that operate on large numbers of voxels (i.e., when the number of voxels exceeds the number of scans); they are ill-posed in the sense that there are an infinite number of equally likely solutions. In this instance, estimating the voxel weights $\beta \in \Re^n$ requires constraints or priors. This is implemented easily by invoking a second level in the model

$$\begin{aligned} WX &= RY\beta + \varsigma \\ \beta &= U\eta \\ \mathrm{cov}(\varsigma) &= \Sigma^{\varsigma}(\lambda) = \exp(\lambda^{\varsigma})RVR^T \\ \mathrm{cov}(\eta) &= \Sigma^{\eta}(\lambda) = \exp(\lambda_1^{\eta})I^{(1)} + \ldots + \exp(\lambda_m^{\eta})I^{(m)} \end{aligned} \tag{13}$$

Here, the columns of $U \in \Re^{n \times u}$ contain spatial patterns or vectors and η are unknown pattern weights. These weights are treated as second-level random effects with covariance, $\mathrm{cov}(\eta)=\Sigma^{\eta}$, which induces empirical priors on the voxel weights; $p(\beta)=N(0, U\Sigma^{\eta}U^T)$. This is a convenient way to specify empirical priors because it separates the specification of prior spatial covariance into patterns

[4] In our implementation, we use the ReML estimates of serial correlations from a conventional encoding formulation of the model. This provides a very efficient estimate because there are generally large numbers of voxels (for more details, see Friston et al., 2007).

encoded by U and the variances in the leading diagonal matrix, Σ^{η}. In this model, $\Sigma^{\eta}(\lambda)$ is a mixture of covariance components arising from a nested set of pattern weights, $s^{(1)} \supset s^{(2)} \supset s^{(3)} \supset \ldots$ where each subset has the same variance. The ith subset $s^{(i)}$ is encoded by a leading diagonal matrix, $I^{(i)}$, containing dummy or switch variables indicating which patterns or columns of U belong to that subset. The construction of this nested set means that the variance; $\exp(\lambda_1^{\eta})+\ldots+\exp(\lambda_i^{\eta})$ of a pattern weight in $s^{(i)}$ is always greater than a pattern weight in its superset, $s^{(i-1)}$.

There are many priors that one could specify with this model, one common prior, used implicitly in fMRI, is that spatial patterns contribute sparsely to the decoding. In other words, a few voxels (or patterns) have large values of β, while most have small values. This is the underlying rationale for support vector machines that presuppose only a few data features (support vectors) are needed for classification. Relevance vector machines make this prior explicit, by framing the elimination of redundant vectors in terms of empirical priors on the parameters. Relevance vector machines are a special case of automatic relevance determination, which is itself a special case of variational Bayes. In fact, these special cases can be expressed formally in terms of conventional expectation maximisation (EM; Dempster et al., 1977), which, for linear models, is formally related to restricted maximum likelihood (ReML; Harville, 1977). See Friston et al. (2007) and references therein (e.g., Mackay and Takeuchi, 1996; Tipping, 2001). In this paper, optimisation is formulated in terms of expectation maximisation.

The model above allows us to compare a wide range of spatial models for decoding. Sparsity is accommodated by having more than one subset; where most subsets have small variance and some have large variance. Crucially, we can control what is sparse. If $U=I$ is the identity matrix, the spatial vectors encode single voxels and we have the opportunity to model sparse representations over anatomical regions. This deployment would be consistent with functional segregation. Furthermore, we could assume that this segregation is spatially coherent (for a theoretical motivation in terms of neuronal computation, see Friston et al., 1992a,b); this would entail using smooth vectors with local support. Conversely, we may assume representations are distributed sparsely over patterns (i.e., one of a small number of patterns is expressed at any one time). These patterns could be the principal modes of covariation in the data. This would correspond to making U the major singular vectors of the data, as in the constrained CCA of the previous section. Finally, these patterns may simply be the patterns expressed from moment to moment. In other words, $U=Y^T$; this is the model used in [linear] support vector machines and regression; in fact, these images may contain confounds, which speak to the use of adjusted images $U=RY^T$. Fig. 2 lists the various models considered in this paper and the corresponding spatial patterns in U. Models with spatial and smooth vectors imply anatomically sparse representations. Conversely, models with singular or support vectors imply the representation is distributed over patterns (which may be sparse in pattern space but not sparse anatomically, in voxel space). The key thing about the hierarchal decoding model above is that it can accommodate different hypotheses about spatial coding. These hypotheses can be compared using Bayesian model comparison; provided we can evaluate the marginal likelihood of each model. In the next section, we describe this evaluation.

Evaluating the marginal likelihood

In what follows, we describe a simple inversion of the model in Eq. (13) using conventional EM, under sparse priors on the parameters. This can be regarded as a generalisation of classification schemes used currently for fMRI, in which the nature of the priors becomes explicit. This inversion uses standard techniques and furnishes the log-evidence or marginal likelihood of the model itself and the conditional density of the voxel weights or decoding parameters. The former can be used to infer on mappings between brain states and their consequences, using model comparison. The latter can be used to construct posterior probability maps showing which voxels contribute to the decoding, for any particular model.

For a more general and technical discussion of the following, see Friston et al. (2007). In brief, we use a fixed-form variational approximation to the approximating posterior under the Laplace approximation and the mean field approximation; $q(\theta)=q(\beta)q(\lambda)$. The Laplace approximation means $q(\beta)=N(\mu^{\beta},\Sigma^{\beta})$ and $q(\lambda)=N(\mu^{\lambda},\Sigma^{\lambda})$ are Gaussian and are defined by their conditional means and covariances. Under these assumptions, the variational scheme

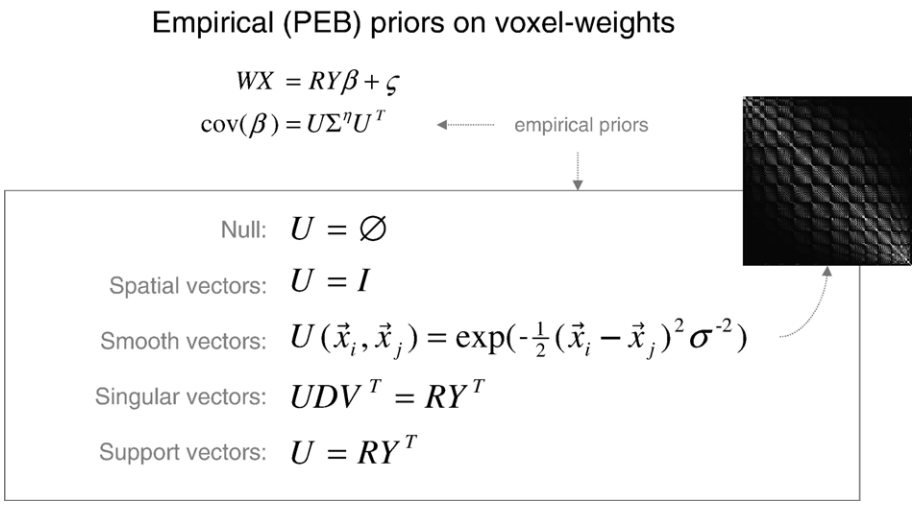


Fig. 2. Taxonomy of different decoding models that are defined by spatial patterns or vectors encoding empirical priors on voxel weights linking brain activity to perceptual or behaviour variables.

reduces to EM. Furthermore, because we can eliminate the parameters β from the generative model (by substituting the second level of Eq. (13) into the first), we only need the M-step to estimate $q(\lambda)=N(\mu^\lambda,\Sigma^\lambda)$ for model comparison and indeed prediction (cf. Gaussian process modelling). This M-step is formally related to ReML.[5]

Bayesian inversion with EM

The inversion of Eq. (13) is straightforward because it is a simple hierarchical linear model. Inversion proceeds in two stages: first, hyperparameters encoding the covariances of the error and the empirical prior covariance are estimated in an M-step. After convergence, the conditional moments of the hyperparameters are used to evaluate the conditional moments of the parameters in an E-step and the log-evidence for model comparison. Because we are dealing with a linear model there is no need to iterate the two steps; it is sufficient to iterate the M-step. For simplicity, we will assume that the pattern sets encoded by $I^{(1)},\ldots,I^{(m)}$ are given and deal with their optimisation later.

First, we simplify the model further by eliminating the parameters through substitution

$$\begin{aligned} WX &= L\eta + \varsigma \\ \mathrm{cov}(WX) &= \Sigma(\lambda) = \exp(\lambda_1)Q_1 + \ldots + \exp(\lambda_{m+1})Q_{m+1} \\ \lambda &= \{\lambda^\varsigma, \lambda_1^\eta, \ldots, \lambda_m^\eta\} \\ Q &= \{RVR^T, LI^{(1)}L^T, \ldots, LI^{(m)}L^T\} \end{aligned} \tag{14}$$

where $L=RYU$ maps the second-level random effects to the weighted target variable. In this form, the only unknown quantities are the hyperparameters, λ controlling the covariance $\Sigma(\lambda)$ of the weighted target variable. This means we have reduced the problem to optimising the hyperparameters of $\Sigma(\lambda)$; this is exactly the form used in Gaussian process modelling.

This covariance includes the covariance of the observation noise and covariances induced by the second level of the model. $w=\mathrm{rank}(W)$ corresponds to the degrees of freedom left after removing the effects of confounds. The log-evidence, $\ln p(X|Y)$ is approximated with the free energy (see Eq. (4)):

$$\begin{aligned} F = -\frac{1}{2}\Big(&X^TW^T\Sigma(\mu^\lambda)^{-1}WX - \ln|\Sigma(\mu^\lambda)| - w\ln 2\pi + \ln|\Pi\Sigma^\lambda| \\ &- (\mu^\lambda-\pi)^T\Pi(\mu^\lambda-\pi)\Big) \end{aligned} \tag{15}$$

The first two terms reflect the accuracy of the model and the last two its complexity ($w\ln 2\pi$ is a constant). This approximation requires only the prior $p(\lambda)=N(\pi,\Pi^{-1})$ and posterior $q(\lambda)=N(\mu^\lambda,\Sigma^\lambda)$ densities of the hyperparameters. In our work, we set the prior expectation and covariance to $\pi_i=-32$ and $\Pi=I/256$, respectively. This is a relatively uninformative hyperprior with a small expectation. A hyperprior variance of 256 means that a scale parameter $\exp(\lambda_i)$ can vary by many orders of magnitude; for example, a value of $1=\exp(0)$ is two prior standard deviations from the prior mean of $1.26\times10^{-14}=\exp(-32)$.

[5] This scheme shares formal aspects with relevance vector machines and automatic relevance determination (*e.g.*, Tipping, 2001); however, the hyperparameters control covariance components as opposed to precision components. This allows for flexible models through linear mixtures of covariance components and renders it an extension of classical covariance estimation (Harville, 1977).

Note that the free energy also depends on the conditional uncertainty about the hyperparameters encoded in Σ^λ. The conditional moments of the hyperparameters are given by iterating

The M-step

$$\begin{aligned} L_{\lambda i} &= -\frac{1}{2}tr\big(P_i\big(WXX^TW^T - \Sigma(\mu^\lambda)\big)\big) - \Pi_{ii}\big(\mu_i^\lambda - \pi_i\big) \\ L_{\lambda\lambda ij} &= -\frac{1}{2}tr\big(P_i\Sigma P_j\Sigma\big) - \Pi_{ij} \\ \Delta\mu^\lambda &= -L_{\lambda\lambda}^{-1}L_\lambda \\ \Sigma^\lambda &= -L_{\lambda\lambda}^{-1} \end{aligned} \tag{16}$$

until convergence. This is effectively a Fisher-scoring scheme that optimises the free energy bound with respect to the hyperparameters. It usually takes between four and sixteen iterations (less than a second for a hundred images). $P_i=-\exp(\mu_i^\lambda)\Sigma^{-1}Q_i\Sigma^{-1}$ is the derivative of the precision $\Sigma(\mu^\lambda)^{-1}$, with respect to the ith hyperparameter, evaluated at its conditional expectation. Critically, the computational complexity $O(s^3m)$ of this scheme does not scale with the number of voxels or patterns, but the number of pattern subsets, m. This reflects one of the key advantages of hyperparameterising the covariances (as opposed to precisions); namely, that one can model mixtures of covariances, induced hierarchically, at the lowest (observation) level of the hierarchy.

Given the conditional expectations of the covariance hyperparameters from the M-step, the conditional mean and expectation of the parameters obtain analytically from

The E-step

$$\begin{aligned} \mu^\eta &= \mathrm{M}WX \\ \mu^\beta &= U\mu^\eta \\ \Sigma^\beta &= U(\Sigma^\eta(\mu^\lambda) - \mathrm{M}L\Sigma^\eta(\mu^\lambda))U^T \\ \mathrm{M} &= \Sigma^\eta(\mu^\lambda)L^T\Sigma(\mu^\lambda)^{-1} \end{aligned} \tag{17}$$

Where M is a maximum *a posteriori* projector matrix. This may look unfamiliar to some readers who work with linear models, because we have used the matrix inversion lemma to suppress large matrices. This remarkably simple EM scheme solves the difficult problem of inference on massively ill-posed models in a very efficient fashion; we use this scheme for source reconstruction in ill-posed EEG and MEG problems (Mattout et al., 2006). However, the current problem requires us to address a further issue, namely the optimisation of the partition (i.e., number and composition of the subsets) encoded in, $I^{(i)}$. This bring us to the final component of Bayesian decoding

A greedy search on pattern sets

Many schemes that seek a sparse solution, such as relevance vector regression (Bishop and Tipping, 2000), use a top-down strategy and start with a separate precision hyperparameter for each pattern or vector. By estimating the conditional precision of each pattern weight, redundant or irrelevant patterns can be eliminated successively until a sparse solution emerges. Clearly, this can entail estimating an enormous number of hyperparameters. We take an alternative bottom-up approach, which generalises minimum norm solutions. We start with the minimum norm assumption that all pattern weights have the same variance $I^{(1)}=I$ and use the conditional expectations of the pattern weights to create a new

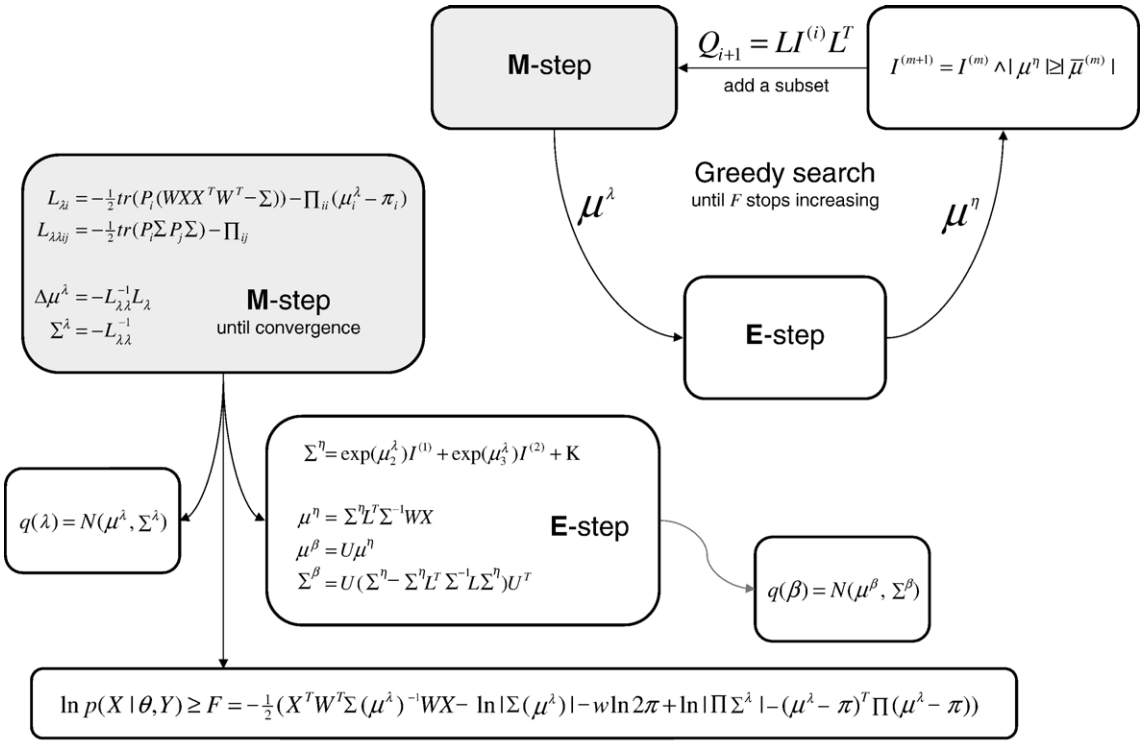


Fig. 3. The EM schemes and its embedding within a greedy search for the optimum set of patterns that maximises the free energy bound on log-evidence. The variables are defined in the main text.

subset; with the highest [absolute] values. We then repeat the EM using two subsets. The subset of patterns with high weights is split again to create a new subset and the procedure repeated until the log-evidence stops increasing (or the mth partition contains a subset with just one pattern). This can be expressed formally as

$$I^{(m+1)} = I^{(m)} \wedge |\mu^\eta| \geq |\overline{\mu}^{(m)}| \tag{18}$$

where $\overline{\mu}^{(m)}$ is the median of the conditional pattern weights of the mth subset. The "and" operator $\wedge$ ensures that the new set is a subset of the previous set. The result is a succession of smaller subsets, each containing patterns with a higher covariances and weights, which is necessarily sparse. Clearly, if the underlying weights are not sparse the search will terminate with a small number of subsets and the solution will not be sparse. This optimisation of subsets corresponds to a greedy search: a greedy algorithm uses the meta-heuristic of making the locally optimum choice with the hope of finding the global optimum. Greedy algorithms produce good solutions on some problems, but not all. Most problems for which they work well have optimal substructure, which is satisfied in this case, at least heuristically. This is because the problem of finding a subset of patterns with high variance can be reduced to finding a bipartition that contains a subset. This is assured, provided we always select a subset with the highest pattern weights. The result of the greedy search is a sparse solution over patterns; where those patterns can be anatomically sparse or distributed. See Fig. 3 for a schematic summary of the scheme.

In principle,[6] adding a subset will either increase the free energy or leave it unchanged. This is because each new subset must, by construction, have a variance that is greater than or equal to its superset. Once the optimal set size is attained, any further subsets will have a vanishingly small variance scale-parameter and the corresponding hyperparameter will tend to its prior expectation; $\mu_i^\lambda \to \pi_i$. In this instance, the curvature approaches the prior precision, $L_{\lambda\lambda ii} \to -\Pi_{ii}$ (see Eq. (16)). This means the conditional covariance approaches the prior covariance, which provides an upper bound. It can be seen from Eq. (15) that the free energy is unchanged under these conditions and the subset is effectively switched off. This is an example of automatic model selection discussed in Friston et al. (2007).

Unlike SVM and related automatic relevance determination (ARD) procedures, Bayesian decoding does not eliminate irrelevant patterns. All the patterns are retained during the optimisation, although some subsets can be switched off as mentioned above. There is no need to eliminate patterns because the computational complexity grows with the log of the number of data features; $O(s^3\ln(n))$. This is because m subsets cover 2^m patterns. This means typically, the greedy search takes a few seconds, even for thousands of voxels.

Summary

In summary:

- We can formulate a MVB decoding model that maps many data features to a target variable, as a simple hierarchal model; known as a parametric empirical Bayes model (PEB; Efron and Morris, 1973; Kass and Steffey, 1989). The hierarchical structure induces empirical priors on the data features (i.e., voxels) which we can prescribe in terms of patterns over features. Each

[6] Ignoring problems of local minima.

pattern is assigned to a subset of patterns, whose pattern weights (unknown parameters of the mapping) have the same variance.
- Each prescription of patterns (i.e., partition) constitutes a hypothesis about the nature of the mapping between voxels and the target variable (i.e., the neuronal representation or cause). One can select among competing hypotheses using model selection based on the model evidence. This evidence can be evaluated quickly using standard variational techniques; formulated as covariance component estimation using EM.
- The partition can be optimised using a greedy search that starts with a classical minimum norm solution and iterates the EM scheme with successive bipartitions of the subset with the largest pattern weights. The free energy or log-evidence of successive partitions or models increases until the optimum set size is reached.

This concludes the specification of the model and its inversion. In the next section, we turn to applications and illustrate the nature of inference entailed by Bayesian decoding.

Illustrative analyses

This section illustrates Bayesian decoding using synthetic and real data. We start with a simple example to show how the greedy search works. This uses simulated data generated by anatomically sparse representations. We then analyse these data to show how the log-evidence (or its free energy bound) can be used to compare models of anatomically sparse and distributed coding. We will analyse three sets of synthetic data (sparse, distributed and null) with three models (spatial, singular and null) and ensure that the inversion scheme identifies the correct model in all cases. A null model is one in which there are no patterns and no mapping. The simulations conclude with a comparative evaluation of MVB with a conventional linear discriminant analysis. The focus here is on the increased power of hierarchical models, over classical models that do not employ empirical priors.

We then apply the same models to real data obtained during a study of attention to visual motion. The emphasis here is on model comparison both in terms of different empirical priors (spatial, smooth, singular and support) and different brain regions. Finally, we cross-validate the results of decoding visual motion (i.e., presence or absence) from single scans using a leave-one-out protocol. We show that the Bayesian classification out-performs a SVM applied to the same problem. We use this analysis to motivate a cross-validation *p*-value for MVB models, for which the null distribution of the likelihood ratio is not readily available.

Simulations

In all simulations, we used the same error variance, fMRI data and confounds used in the empirical analyses below. Using these features (i.e., voxel-wise fMRI time-series) and assumed pattern weights we were able to generate target variables and analyse the data knowing the true values. The data features comprised 583 voxel values from 360 scans, with 26 confounds (see Fig. 4 and below for a detailed description). We first removed the confounds from the data features to give, *RY*. Synthetic target variables were then constructed by taking a weighted average of the voxel time-series and adding noise. The voxel weights were generated using one of the models described in the previous section and depicted in Fig. 2.

Bayesian decoding

First, we generated data under a sparse spatial model using the first 128 scans and 256 voxels. Here the voxel weights were sampled from a normal distribution and raised to the fifth power, to make them sparsely distributed. Random variables were added to the ensuing target variable after they were scaled to give a signal to noise of four; i.e., the standard deviation of signal was four times the noise. Because the signal and random effects at each voxel are mixed with the same weights (see Eq. (11)), the implicit signal to noise at each voxel (on average) is also four. The resulting target variable and error and are shown in Fig. 4 (left panel). The upper-left panel in Fig. 5 shows the voxel weights, whose sparsity is self-

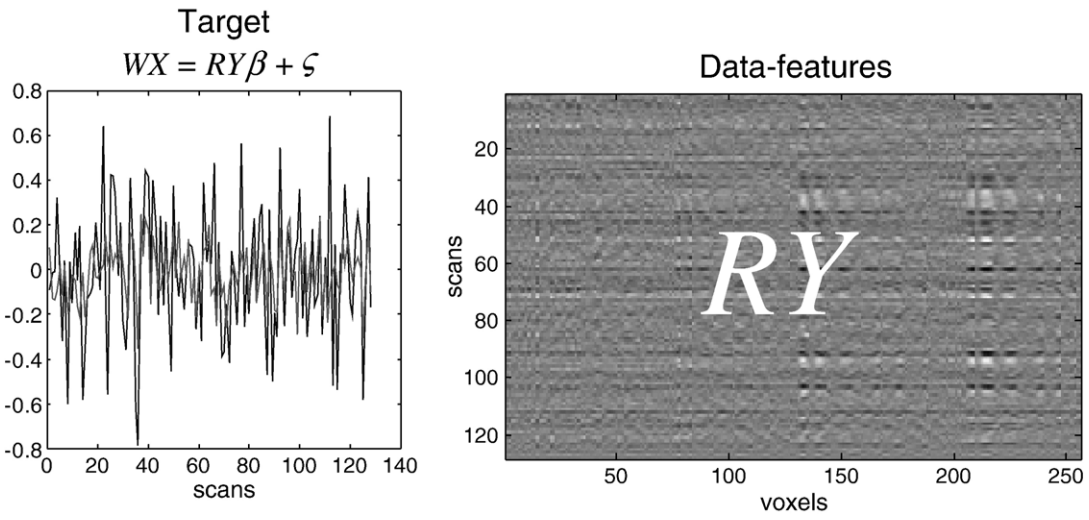


Fig. 4. Right: Data features that were mixed to generate the target variable in the simulations. These are a subset (128 scans and 256 voxels) of the voxel data from the analysis of real fMRI data reported in Fig. 8. Left: Target variable (solid line) and noise (broken line) for the simulation demonstrating the nature of the greedy search (reported in Fig. 5).

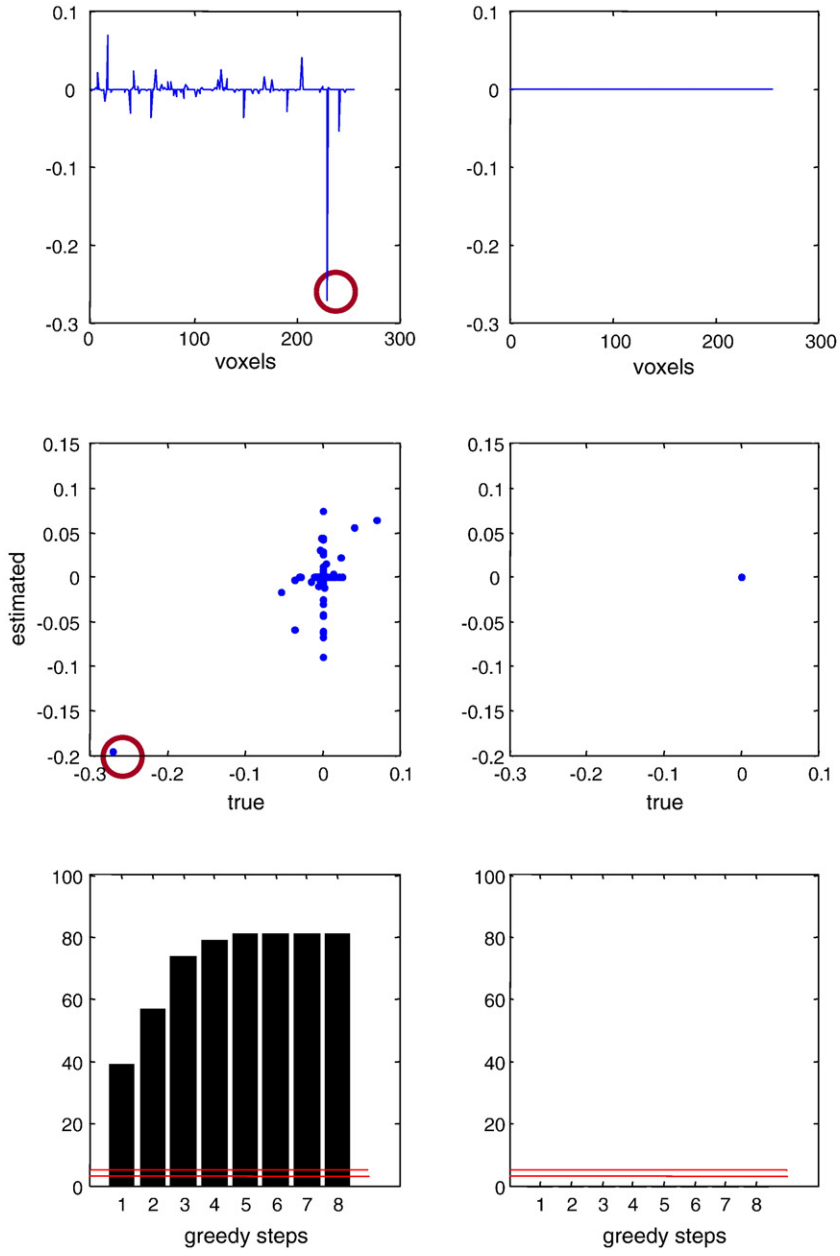


Fig. 5. Left panels: results for a greedy search for the optimum set of spatial patterns using targets generated from sparse voxel weights. The true weights are shown on the top and the estimated weights are plotted against the true weight in the centre. The lower panel shows the log-evidence, relative to a null model with no patterns, as a function of the number of greedy steps (i.e., the size of the set). Right panels: the same format as the upper row but showing an analysis of null targets, formed by setting the voxels-weights to zero; using exactly the same noise terms. The predicted voxel weights come from the first $m=1$ model. The red circles highlight the voxel with the largest voxel weight.

evident. The lower-left panel shows the free energy bound on the log-evidence as a function of greedy steps (i.e., number of pattern subsets). We have subtracted the log-evidence for the null model so that any value greater than three can be considered as strong evidence for sparse coding.[7] It can be seen that the log-evidence increases systematically with the size of the partition, until it peaks after about five subsets. The conditional expectation of the voxel weights for the partition with the greatest free energy is shown in the left middle panel, plotted against the true value. Although the agreement is not exact, the MVB scheme has identified voxels with truly large weights (circled).

We repeated exactly the same analysis but set the weights to zero to simulate a null model. In this instance the log-evidence optimised by the greedy search never increased above the null model and we would infer there was no mapping. Even if we take the optimum set ($m=1$) from the greedy search on the spatial

[7] Strong evidence requires the Bayes factor to be between 20 and 150, or the differential log-evidence to be between $3 \approx \ln(20)$ and $5 \approx \ln(150)$ (Penny et al., 2004). This corresponds to a posterior probability for the better model between $p=0.95 \approx 20/21$ and $p=0.99 \approx 150/151$, under flat priors on the models.

model, the estimated weights are appropriately small (see right panels in Fig. 5).

Model comparison

In the next simulations, we generated target variables using the model in Eq. (13) and different patterns. In all cases we selected the pattern weights, η as above from a normal distribution and raised them to the fifth power. We generated three target variables corresponding to a null model; $U=\varnothing$, a sparse spatial model; $U=I$ and a distributed singular model; $UDV=RY$, where this equality signifies a singular value decomposition of the adjusted data into orthonormal vectors. We then inverted each of the three models using the three target variables. The free energies of the resulting nine analyses are shown in Fig. 6. It can be seen that the decoding scheme correctly identified the true model in all cases; in the sense that the greatest free energy was obtained with the model that generated each target variable. It should be noted that when the signal to noise was decreased, we often observed that the sparse spatial model was favoured over the distributed singular model even when data were generated using the latter. This may be because the singular vectors of the data used were themselves sparse over voxels. However, we never observed the sparse model to be better than a null model when decoding null data. In the final simulations, we look more closely at the sensitivity and specificity conferred by the empirical priors implicit in hierarchical models.

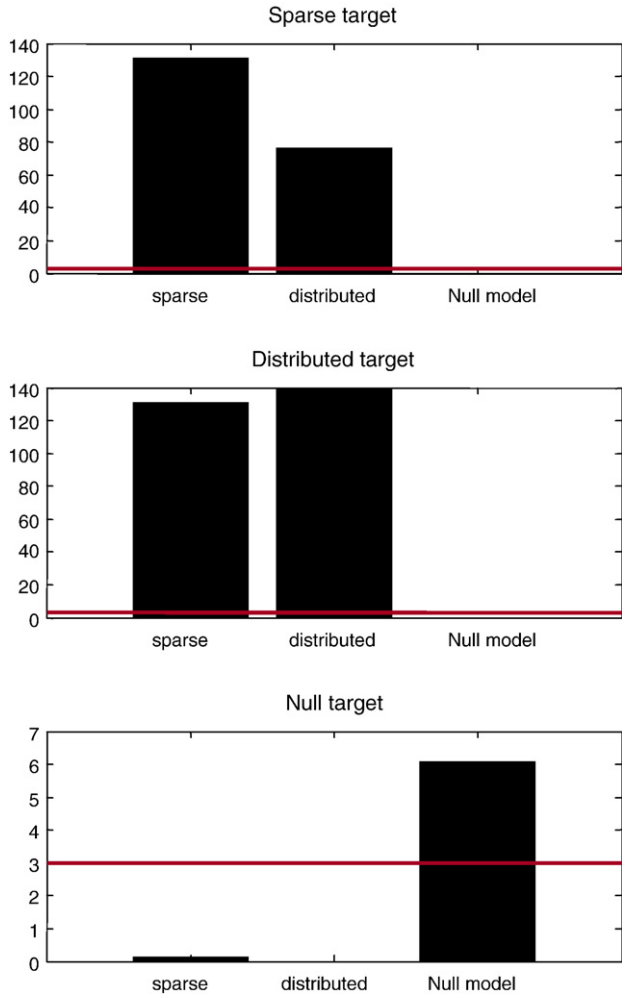


Fig. 6. An illustration of model comparison using three models (spatial, singular and null) applied to three synthetic target variables that were generated by the same three models. Each row corresponds to the log-evidences (normalised to their minimum) for each target. These results show that model comparison allows one to identify the form of the model that generated the data. The horizontal line is set at three (i.e., a difference in log-evidence that would be regarded as strong evidence in favour of a model).

Hierarchical vs. non-hierarchical models

To compare empirical Bayesian with classical models, we repeated the first set of simulations using a sparse model but reduced the number of scans to 64, the number of voxels to 32 and reduced the signal to noise to a half. Reducing the number of voxels to less than the number of scans enabled us to use conventional CCA to infer on the coupling between voxel activity and the simulated target variables. Recall that CCA uses exactly the same linear model as MVB but there are no empirical priors (i.e., the voxel weights are treated as fixed effects). Because the target variable is univariate, this CCA model is the same as an analysis of covariance (AnCova), which is exactly the same as a linear discriminate function analysis. The likelihood ratio statistic for AnCova is, after transformation, the F-statistic. We generated target variables as above and evaluated the log-likelihood ratio, $\ln \Lambda$ using the free energy of sparse and null models, optimised using MVB. For each realisation, we also computed the F-statistic using a standard CCA. We repeated this ten thousand times for both sparse and null targets. This allowed us to plot the proportion of sparse targets identified by both statistics as a function of their threshold; this is the sensitivity. Conversely, the proportion of null targets identified falsely at each threshold gives a measure of specificity. Plotting one against the other gives receiver-operator curves for the two statistics.

The results of these simulations are shown in Fig. 7. It is immediately obvious that $\ln\Lambda$ based on MVB is much more sensitive for all acceptable levels of specificity. This is not surprising because the data were generated in a way that the MVB scheme could model. What is remarkable is the quantitative improvement in sensitivity or power: The classical analysis shows about 20% sensitivity at 5% false-positive rate. The threshold for this rate was, $F=2.20$, which agrees well with the $p=0.05$ threshold; $F=2.181$ based on its null distribution under Gaussian assumptions. At this level of specificity, the MVB scheme exhibited about 56% sensitivity. Interestingly, the threshold for this specificity was, $\ln \Lambda=1.09$. In other words, the optimum threshold for classical inference on the Bayes factor would require positive (but not strong) evidence in favour of the alternative hypothesis. However, unfortunately there are no analytic results for this threshold because there are no analytic results for the null distribution of the MVB log-likelihood ratio (unlike the F-statistic).

This means that although we can always select the best model, we cannot use $\ln\Lambda$ to assign a p-value to test the null hypothesis of independence between the data features and target. However, we can use the selected model for cross-validation and use the ensuing predictions to get a classical p-value. This is useful because we can then make classical inferences about over-determined models that would elude conventional statistics. We will illustrate this in the final section.

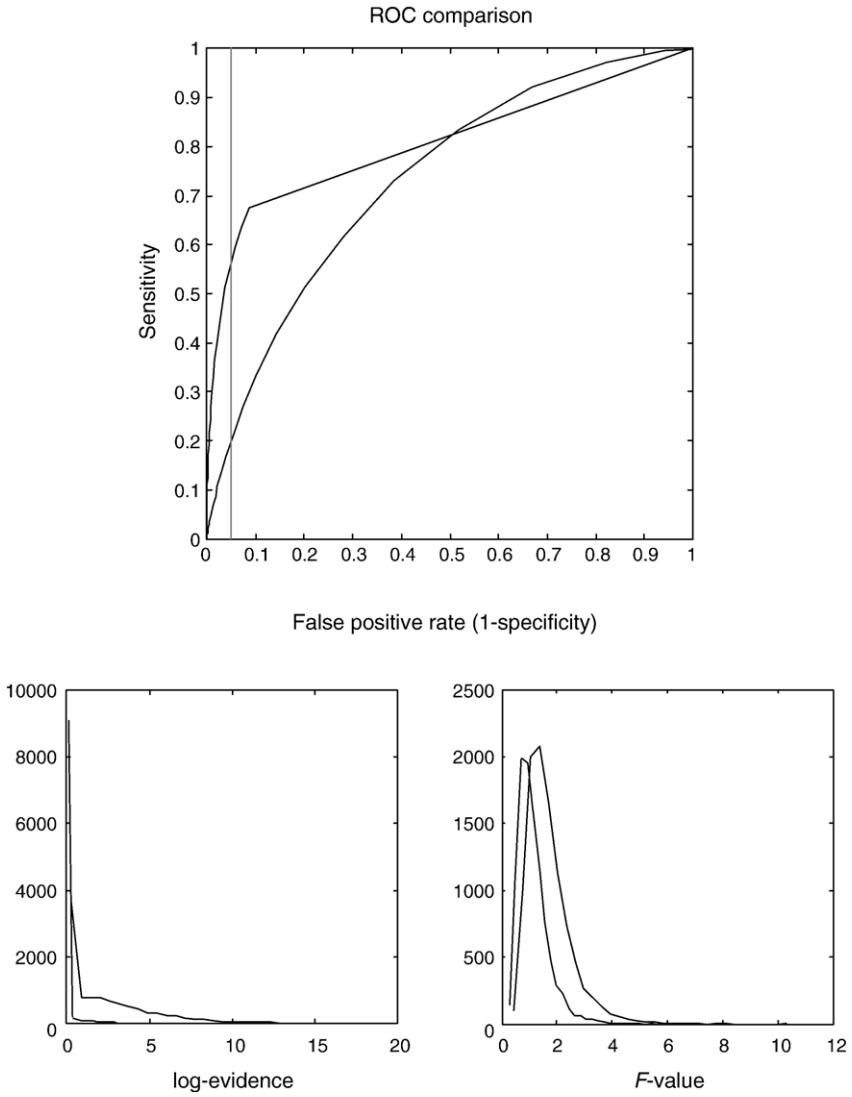


Fig. 7. Upper panel: Receiver operator curves summarising a power analysis for the Bayesian decoding scheme and an analysis based on a conventional linear model (CCA). These curves depict the sensitivity as a function of false-positive rate for various thresholds applied to likelihood ratio statistics. These statistics were the log-evidences difference for the MVB (i.e., log-Bayes factor; solid line) scheme and the *F*-statistic for the general linear model (dotted line). The vertical line marks a false-positive rate of 0.05. This rate obtained with a threshold of 1.09 for the log-Bayes factor and 2.20 for the *F*-statistic. The corresponding power was 56.4% or MVB and 19.6% for CCA. Lower panels: Distribution of the statistics (MVB; left and CCA; right) over ten thousand realisations for the null targets (dotted lines) and a target generated with a signal to noise of one half (solid lines).

It should be noted that these simulations were performed for comparative purposes only. As mentioned in the previous section, it is not possible to use CCA when the number of voxels exceeds the number of scans, nor is it possible to compare CCA models specified in terms of different spatial priors, because there are none. Clearly, in the simulations above we knew what caused the target variable. In the next section, we apply the analyses above to real data where the model and their parameters are unknown.

Empirical demonstrations

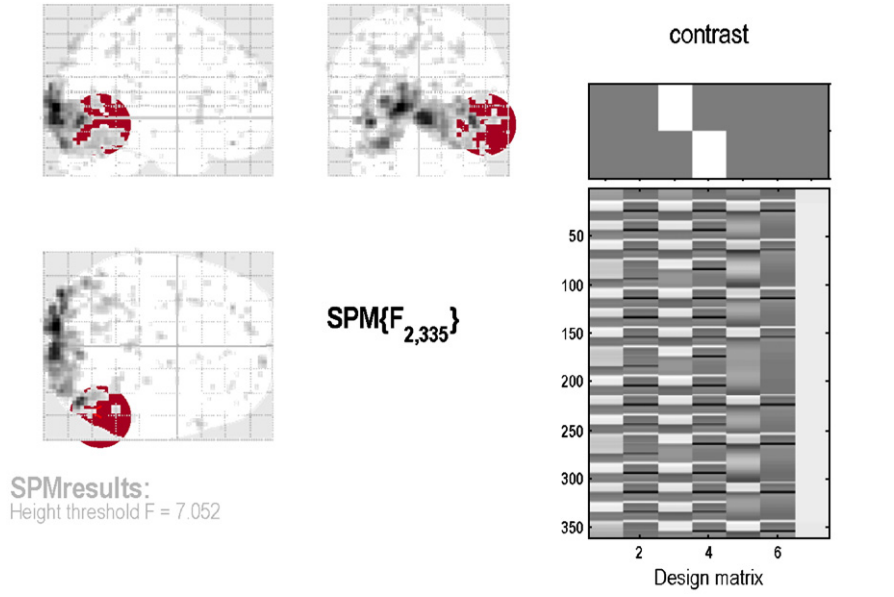
In this section, we apply the analysis above to real data obtained during a study of attention to visual motion. We have deliberately used a standard data set, which is available from http://www.fil.ion.ucl.ac.uk/spm, so that readers can reproduce the analyses below. These data have been used previously to illustrate various developments in data analysis. In many decoding and classification analyses, one generally uses high-resolution unsmoothed data and small volumes of interest. However, the principles of inference are exactly the same for any imaging data and we will illustrate the sorts of questions that can be addressed using this standard smoothed data set.

fMRI data

Subjects were studied with fMRI under identical stimulus conditions (visual motion subtended by radially moving dots) under different attentional tasks (detection of velocity changes). The data were acquired from normal subjects at 2 T using a Magnetom VISION (Siemens, Erlangen) whole-body MRI system, equipped with a head volume coil. Contiguous multi-slice T2*-weighted fMRI images were obtained with a gradient echo-planar

sequence (TE = 40 ms, TR = 3.22 s, matrix size = 64 × 64 × 32, voxel size 3 × 3 × 3 mm). The subjects had four consecutive hundred-scan sessions comprising a series of ten-scan blocks under five different conditions D F A F N F A F N S. The first condition (D) was a dummy condition to allow for magnetic saturation effects. F (Fixation) corresponds to a low-level baseline where the subjects viewed a fixation point at the centre of a screen. In condition A (Attention), subjects viewed 250 dots moving radially from the centre at 4.7° per second and were asked to detect changes in radial velocity. In condition N (No attention) the subjects were asked simply to view the moving dots. In condition S (Stationary), subjects viewed stationary dots. The order of A and N was swapped for the last two sessions. In all conditions subjects fixated the centre of the screen. In a pre-scanning session the subjects were given five trials with five speed changes (reducing to 1%). During scanning there were no speed changes. No overt response was required in any condition. Data from the first subject are used here.

Fig. 8 shows the results of a conventional analysis using a linear convolution model formed by convolving box-car stimulus functions with a canonical hemodynamic response function and its temporal derivative. The stimulus functions encoded the presence of photic stimulation (first two columns of the design matrix on the upper right), visual motion (second two columns) and attention (last two columns). The design matrix shows only the first constant term of a series of drift terms (a discrete cosine set) modelling slow fluctuations in signal as confounds. The SPM shown in the upper panel uses the *F*-Statistic to test for motion; the corresponding contrast weights are shown above the design matrix. The red circle depicts a 16-mm radius spherical volume of interest, encompassing 583 voxels in early striate and extrastriate cortex (deliberately chosen to include V5/MT complex). The table (lower panel) shows classical *p*-values testing for the contrast after adjustment using random field theory for the spherical search volume. These voxels survived an uncorrected threshold of $p < 0.001$. We will attempt to decode motion from all the grey matter voxels in this spherical region, using MVB. This may seem a trivial problem; however, this design was optimised to detect the effects of attention on motion-related responses, not motion *per se*. Decoding motion is actually quite a challenge because there were



Statistics: ***search volume: 16.0mm sphere at [48,-63,0]***

set-level		cluster-level			voxel-level					mm mm mm		
p	*c*	$p_{corrected}$	k_E	$p_{uncorrected}$	$p_{FWE-corr}$	$p_{FDR-corr}$	*F*	$(Z_{=})$	$p_{uncorrected}$			
0.006	4		39		0.000	0.000	40.66	Inf	0.000	39	-72	0
					0.000	0.000	40.09	Inf	0.000	42	-75	-3
			3		0.002	0.000	13.35	4.55	0.000	39	-63	-12
			1		0.124	0.003	8.67	3.52	0.000	51	-66	-3
			1		0.436	0.010	7.36	3.18	0.001	48	-63	0

table shows 16 local maxima more than 4.0mm apart

Height threshold: F = 7.05, p = 0.001 (0.513) {p<0.001 (unc.)}
Extent threshold: k = 0 voxels, p = 1.000 (0.513)
Expected voxels per cluster, <k> = 1.212
Expected number of clusters, <c> = 0.72
Expected false discovery rate, <= 0.01

Degrees of freedom = [2.0, 335.0]
FWHM = 7.5 7.4 7.1 mm mm mm; 2.5 2.5 2.4 {voxels};
Volume: 15741; 583 voxels; 43.3 resels
Voxel size: 3.0 3.0 3.0 mm mm mm; (resel = 14.69 voxels)

Fig. 8. Results of a conventional encoding analysis of the visual motion study. The upper left panel shows the maximum intensity projection of the SPM, thresholded at $p < 0.001$ (uncorrected). The upper right panel shows the design matrix and contrast used to construct the SPM. The table lists maxima in the SPM, using random field theory, to give adjusted *p*-values for the number, size and height of subsets in the excursion set. The red circle depicts the 16 mm spherical volume of interest used to adjust the *p*-values and employed for decoding in subsequent figures. This volume was centred at 48, −63, 0 mm and contained 583 grey matter voxels.

only four epochs of stationary stimuli (note that the effects of photic stimulation are treated as confounds in the decoding model).

Before looking at Bayesian decoding, it is worthwhile noting that multivariate inference using random field theory suggests the mutual information between the voxel time courses and motion is significant. This can be inferred from the set-level inference with $p<0.006$ (left-hand column of the table, Fig. 8). This is based on the observed number of peaks surviving a $p<0.001$ threshold in the volume of interest. Here we expected 0.72 peaks (see table footnote, Fig. 8) but observed four. Under the Poisson clumping heuristic; this number of 'rare events' is very unlikely to have occurred by chance (for more details, see Friston et al., 1996).

Bayesian decoding

Fig. 9 shows the results of a spatial MVB decoding of the first (canonical) motion regressor. The upper left panel shows the free

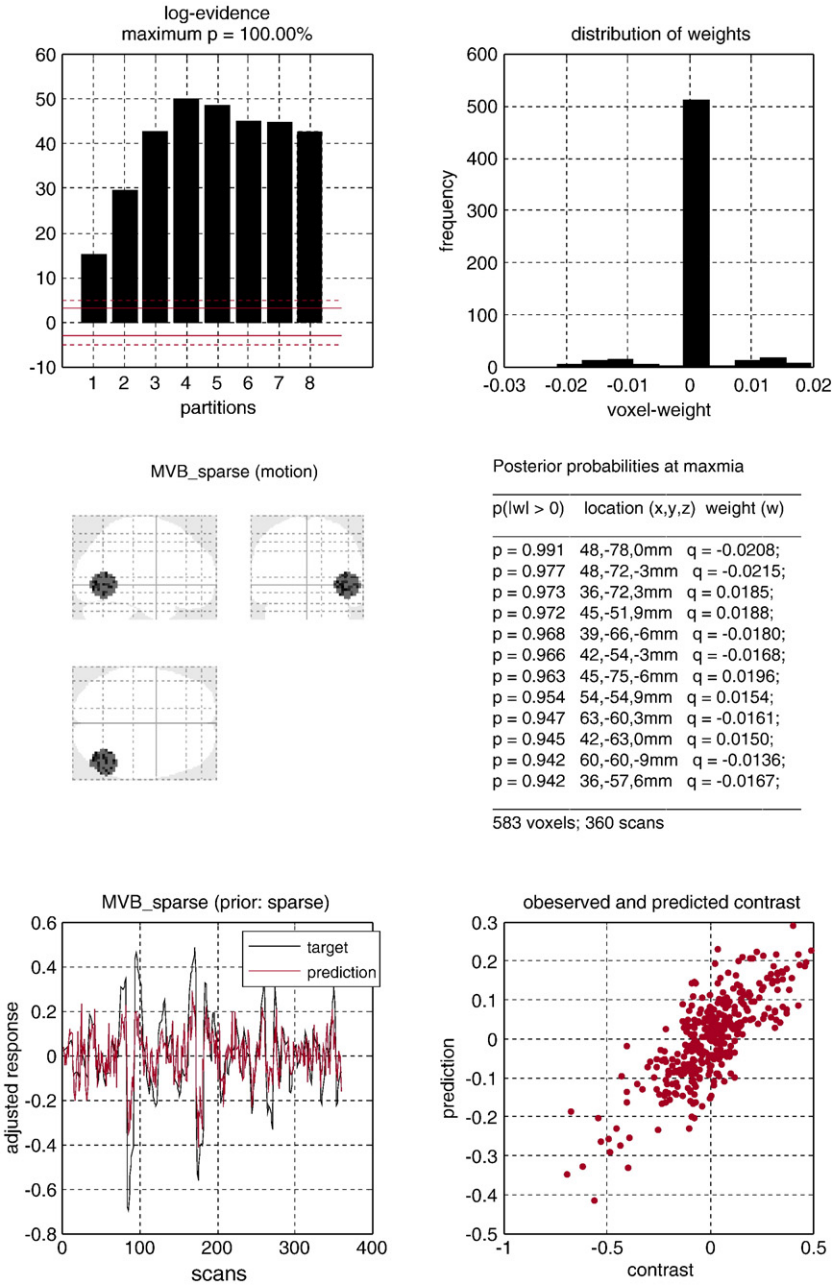


Fig. 9. Results of an MVB analysis using the voxels highlighted in the previous figure. The upper left panel shows the free energy bound on the log-evidence, relative to a null model, as a function of set size; the red lines depict the threshold for strong and very strong evidence for each model (i.e., an extra subset of patterns). The upper right panel shows the distribution of voxel weights for the optimum model and discloses their sparse distribution. The middle panels show the conditional estimates of the voxels-weights as a maximum intensity projection and in Tabular format (reporting the sixteen most significant voxels spaced at least 4mm apart). The lower panels show the observed and predicted target as a function of scan number and plotted against each other.

energy approximation to the log-evidence for each of eight greedy steps, having subtracted the log-evidence for the corresponding null model. As in the previous section, any log-evidence difference of three or more can be considered strong evidence in favour of the model. It can be seen that the log-evidence peaks with four subsets, giving an anatomically sparse deployment of voxel weights (upper right panel). This sparsity is evidenced by the heavy tails of the distribution, lending it a multimodal form. These weights (positive values only) are shown as a maximum intensity projection and in tabular format in the middle row. The table also provides the posterior probability that the voxel weight is greater or less than zero (for peaks that are at least 4mm apart). Note that these probabilities are conditioned on the model as well as the data. That is, under the sparse model with spatial vectors, the probability that the first voxel has a weight greater than zero, given the target variable, is 99.1%. Note that the free energy decreases after four subsets. Strictly speaking this should not happen because the free energy can only increase or stay the same with extra components. However, in this case, the EM scheme has clearly converged on a local maximum, when there are too many subsets. This is not an issue in practice, because one would still select the best model, which hopefully is a global maximum.

The bottom row shows the target variable and its prediction as a function of scan and by plotting them against each other. This weighted target variable is simply $WX = R\mathbf{X}c$ from Eq. (12), where the contrast, c, selects the first motion regressor from the design matrix, $\mathbf{X}$, in Fig. 8. Note that the motion target has a complicated form because all other effects (photic stimulation, attention and confounds) have been explained away. The agreement is not surprising because one could produce a perfect prediction given 583 voxels and only 360 scans. The key thing is that the match is *not* perfect but is optimised under the empirical priors prescribed by the model. To illustrate the specificity of this analysis, we repeated exactly the same analysis but randomised the target variable by convolving a time-series of independent Gaussian variables with the hemodynamic response function used to create the real target variables. Fig. 10 shows the results of this analysis, using the same format as the previous figure. The prediction is now, properly, much poorer, because there is no mapping between the neuronal activity and target. Critically, this can be inferred from the optimised log-evidence (upper left panel), which fails to provide strong evidence over the null model.

Model comparison

To illustrate model comparison we repeated the analysis above using five different models; null, spatial, smooth, singular and support. The smooth patterns were based on a Gaussian smoothing kernel with a standard deviation of 4 mm. The singular vectors were selected automatically so that they explained 95% of the variance in the data features. After optimising the log-evidence with the greedy search, the model evidences were compared directly. Fig. 11a shows that the best model was a spatial model and therefore indicates that the representation of motion is anatomically sparse; as one would predict under functional segregation. Interestingly, of the non-null models, the smooth vectors were the worst. This suggests that, even though the fMRI data were smoothed, the underlying representation of motion is not dispersed spatially; again this would be expected under patchy functional segregation (see Zeki, 1990). One would imagine that, in the absence of smoothing, the model with smooth patterns would fare even worse. Although the informed singular vectors outperform the image-based support vectors, there is no strong evidence for the former model relative to the latter. This simple example illustrates the sorts of questions that can be addressed using MVB.

In the previous example, we compared models that differed in the patterns encoding the form of the empirical priors. Clearly models can also differ in terms of which data features we chose as predictors. In imaging, this translates into comparing the predictive ability of different brain regions (or combinations of brain regions) using model comparison. As a simple example, we selected a 16-mm spherical volume of interest in the prefrontal region and repeated the MVB analysis. The log-evidence for both regions and the null model are provided in Fig. 11b. These show that canonical motion can be readily decoded from both regions but, if one wanted to ask which region afforded the better model, then clearly the visual region supervenes. We have deliberately chosen a rather trivial question to illustrate model comparison; however, it is easy to imagine interesting questions that can be formulated using MVB. For example, using combinations of regions it is possible to compare models with two regions (say right and left hemispheric homologues), one or the other or neither and infer on the lateralisation of representations. This allows one to ask specific questions about the nature of distributed codes in the brain and how they are integrated functionally.

Cross-validation

This section concludes with a comparison with a standard classifier and cross-validation of the MVB decoding. Point classifiers like SVM cannot be assessed in terms of model evidence because the requisite probability densities are not formally parameterised. However, we can assess the MVB model using cross-validation by evaluating the predictive density using $q(\theta)$. In this example, we used the occipital volume of interest of 16-mm radius above, encompassing 538 voxels. We addressed cross-validity by trying to classify each scan as a motion or non-motion scan; this entailed thresholding the target variable, WX around its median to produce a list of class labels (one or minus one). The median threshold ensures that there are an equal number of targets in each class. Because the target variable has been convolved with a hemodynamic response function, these labels reflect the amount of motion under the support of this function (i.e., within the preceding few seconds).

We used a leave-one-out strategy by designating one scan in the time series as a test scan and estimating the pattern weights using the remaining training scans. We used a MVB model with sparse spatial patterns. The pattern weights were then used to form a prediction, which was thresholded around its median and compared with the target class. This was repeated 360 times (for every scan) and the significance of the classification assessed against the null hypotheses of chance performance, using the binomial distribution. The MVB classification performed at 64.4%, (232 out of 360; $p = 1.25 \times 10^{-8}$), which appears extremely significant (but see below). For comparison purposes, we trained a standard SVM classifier[8] with exactly the same training and test samples, using the adjusted imaging data RY as data features. To

[8] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

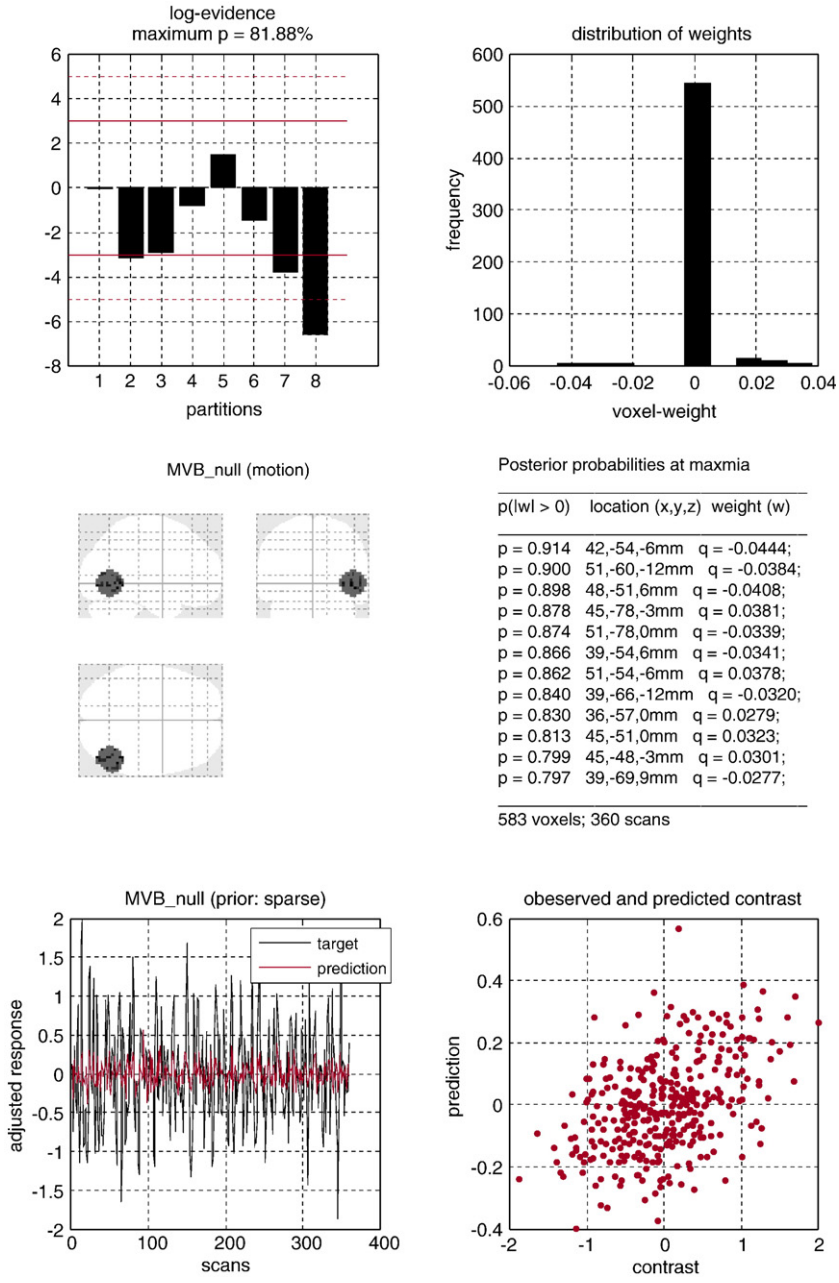


Fig. 10. This figure uses the same format as the previous figure but shows the results of an analysis applied to a randomised target in which any coupling between data features and targets was destroyed.

ensure we optimised the SVM we repeated the leave-one-out scheme for nine levels of the hyperparameter $C=10^0$, …, 10^{-8}. The best performance of the SVM (with $C=10^{-3}$) was well above chance level, correctly classifying 61.4% (221 out of 360; $p=5.58\times10^{-6}$) of scans; however, its performance was poorer in relation to the Bayesian classification (64.4%). See Fig. 12.

The voxel weights for the SVM and MVB classifiers are shown in Fig. 13 (upper panels). The difference is immediately obvious; the MVB profile shows the anatomical sparsity implicit in characterisations above; whereas the SVM weights are not sparse. However, if we plot the MVB weights against the cubed SVM weights we see that, with one exception, when the MVB found a large positive or negative weight, so did the SVM.

One may ask why the MVB decoding model was better than the SVM, given both sought sparse solutions and the SVM was explicitly optimising classification performance (while the decoding scheme optimised free energy). The most parsimonious answer is that the SVM is using a suboptimal model. The results of model comparison in Fig. 11a suggest that visual motion has a sparse anatomical representation and this is the model used by Bayesian decoding. Conversely, the SVM is obliged to use a sparse mixture of non-sparse patterns and is consequently poorer at classifying.

This intuition was confirmed by repeating the MVB classification but using support vectors. As can be seen in Fig. 12, the classification performance fell from 64.4% to 63.0%. This is still

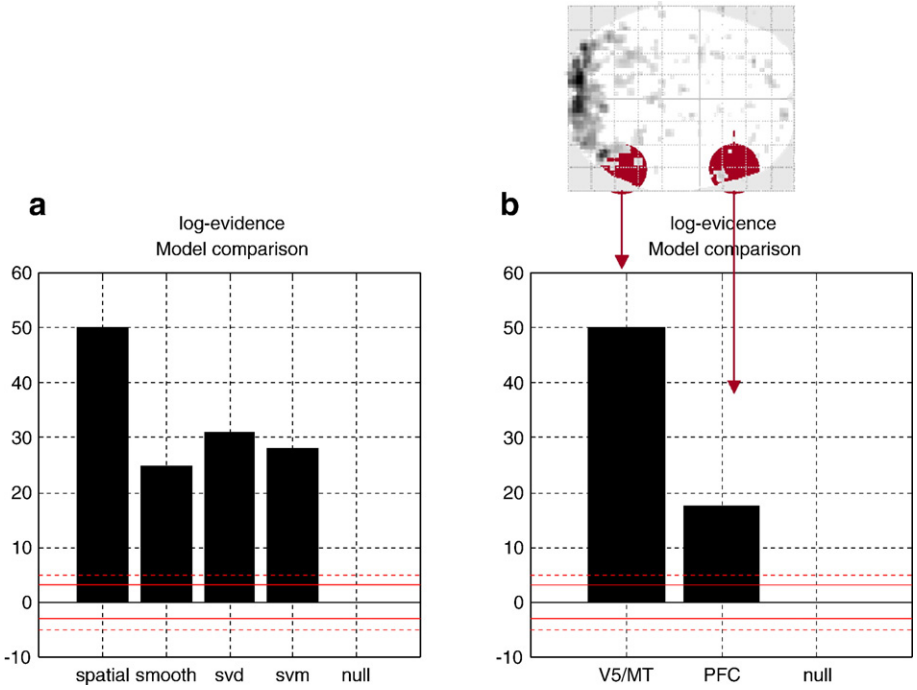


Fig. 11. Bayesian model comparison: Left: (a) Log-evidences for five models of the same target and voxels. These models differ in terms of the spatial patterns that might be used to encode motion. Right: (b) log-evidences for three models of the same target but now using different voxels. These voxels came from two 16mm spherical volumes of interest in the visual and prefrontal regions, depicted by the red circles on the maximum intensity projection.

better than the SVM, with a difference of 1.66%. However, this difference is less than the standard deviation of the underlying binomial distribution; $2.63\% = \sqrt{360\frac{1}{2}(1-\frac{1}{2})}$. One might anticipate that a finer-tuned optimisation of the SVM hyperparameters would equate the performance of the SVM and MVB, under support vectors. Note that MVB optimises its own hyperparameters automatically (without needing the true test class).

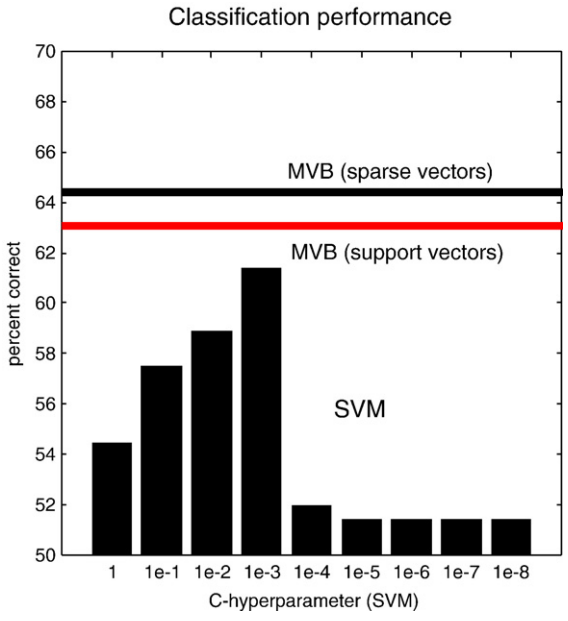


Fig. 12. Classification performance of the MVB scheme (horizontal lines) and SVM (bars) over different levels of the SVM hyperparameter, C. Performance is shown in terms of percentage correct classification of motion scans using the occipital volume of interest in the previous figures. The MVB models here employed sparse (top line) and support (lower line) vectors, where as the SVM used, by definition, support vectors.

Classical inference with cross-validation

One might ask whether the p-values from the leave-one-out scheme could be used for inference? Unfortunately they cannot, because the removal of confounds and serial correlations render them invalid. In other words, the binomial test assumes that the training and test data features are independent and this assumption is violated when we use data features that have been adjusted for confounds; or when the data are serially correlated as in fMRI. In what follows, we describe a cross-validation scheme that resolves this problem and produces p-values for any MVB model.

The solution rests on using weighted samples of data features for training that are linearly independent (i.e., orthogonal) of the test data. This is achieved by removing serial correlations and eliminating the test data before estimating the voxel weights. These weights are then applied to de-correlated features, with the training data eliminated. Critically, elimination proceeds by treating unwanted data as confounds, as opposed to simply discarding portions of the data. More precisely, consider a k-fold scheme in which the test subset is encoded by indicator variables in the leading diagonal matrix, $I^{(k)}$. The model is optimised using a residual forming matrix, $R^{-k}=(I-G_{-k}G_{-k}^{-})S$; where (i) the confound matrix $G_{-k}=[SG,I^{(k)}]$ includes any effects due to the de-correlated test data and (ii) $S=V^{-1/2}$ is a de-correlation matrix that renders the errors spherical. The ensuing weights μ_k^β are then applied to test-features, $Y_k=R_kY$. These test features are formed

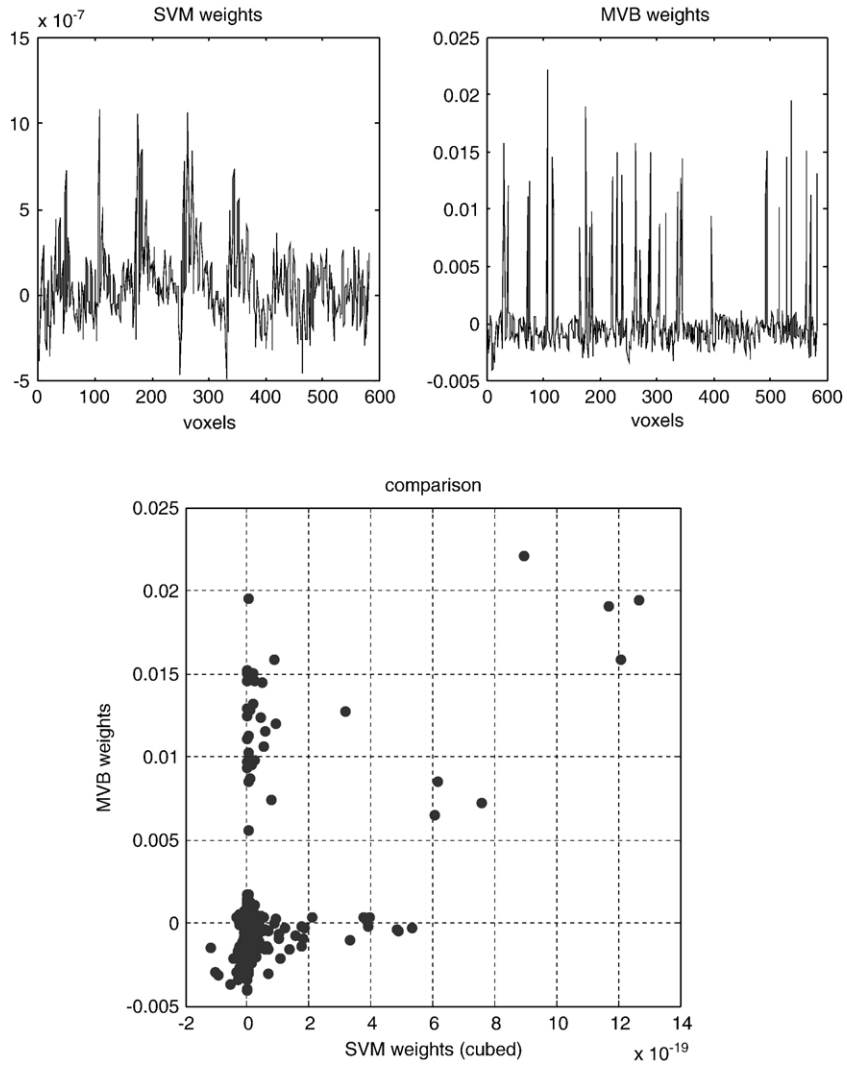


Fig. 13. Comparative analysis if the encoding of visual motion using MVB and SVM. Upper panels: voxel weights from SVM (left) and MVB (right) showing the sparsity over voxels of the latter, relative to the former. Lower panel: The same weights plotted against each other, showing that large values in one correspond to large values in the other.

using a residual forming matrix, $R_k=(I-G_k\bar{G_k})S$ with confounds $G_k=[SG,(I-I^{(k)})]$ that allow the effects due to the de-correlated training data to be explained away. The result is a cross-validation prediction, $X_k=Y_k\,\mu_k^{\beta}$, that accounts properly for serial correlations and confounds by ensuring that the cross-validation weights cannot be influenced by the test data and the prediction is conditionally independent (to first order) of the training data. Furthermore, it ensures, under the null hypothesis, that the test and training data are linearly independent; i.e.,

$$\left\langle Y_{-k}Y_k^T\right\rangle = R_{-k}\left\langle YY^T\right\rangle R_k = R_{-k}VR_k = 0 \tag{19}$$

The predictions can now be added to give a cross-validation prediction for the entire sequence of weighted targets; i.e., $\hat{X}=S^{-1}(X_1+\ldots+X_K)$. Under the null hypothesis, there can be no correlation between WX and $\hat{X}$. Furthermore, because the random effects in this model are mixtures of random effects over features (i.e., voxels), we know by central limit theorem that they are normally distributed. This means the most powerful test of the null hypothesis is a simple t-test; testing for dependency between the observed and predicted targets, in the presence of confounds. This can be tested using the simple model

$$\hat{X} = [TX, G]\beta + \varepsilon \tag{20}$$

under normal parametric assumptions about the serially correlated errors with a test of the null hypothesis that $\beta_1=0$. The astute reader may notice that we have come full-circle; in that this test is exactly the same as testing a contrast, i.e., $c^T\beta=0$, under the original encoding model; $\hat{X}=\mathbf{X}\beta+\varepsilon$. The only difference is that we have replaced the original voxel values with a summary of the activity over voxels, using a cross-validation procedure.

To ensure the assumptions above are not violated in a practical setting, we applied the procedure using a two-fold cross-validation scheme to the empirical data above, using weighted targets formed by convolving random vectors with a hemodynamic response function. We repeated this a thousand times and accumulated the p-values from the t-test on the model in Eq. (21). Fig. 14 shows the results of this analysis in terms of a Q-Q plot (i.e., cumulative frequency of ranked p-values). An

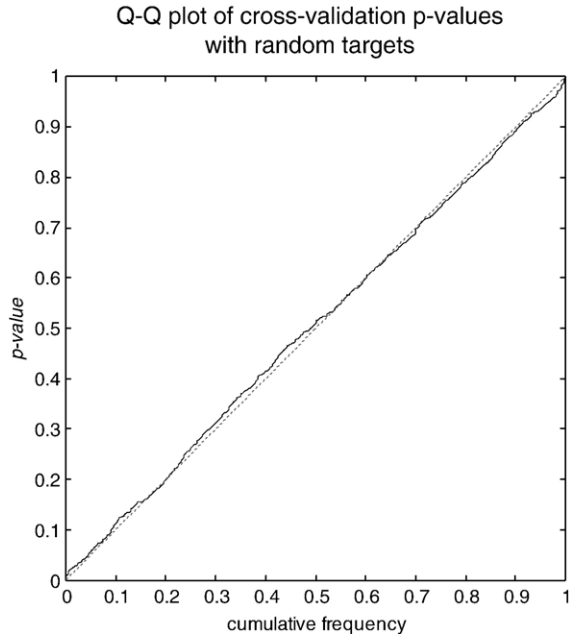


Fig. 14. Results of a simulation study to ensure the exact nature of cross-validation *p*-values. This *Q*-*Q* plot shows the cumulative frequency of ranked *p*-values from one thousand, two-fold cross-validation tests using the data features from the fMRI study (V5/MT volume of interest) and a randomised weighted target. Ideally the *Q*-*Q* plot should be a straight line passing though the origin. The number of p-values falling below $p=0.05$ was 0.056, suggesting a reasonably exact test.

exact and valid test should produce a straight line though the origin; happily this is what we observed. This is important because it gives a reasonably powerful test that enables classical inference about multivariate models for which no conventional results exist.

Fig. 15 shows the results of the cross-validation prediction for the 16-mm occipital volume of interest. This prediction should be compared with that in Fig. 9 that was obtained without the cross-validation constraint. It can be seen that the accuracy of this prediction is unlikely to have occurred by chance; the corresponding cross-validation *p*-value was $p=0.000046$ and was extremely significant.

Summary

In summary, this section has demonstrated the nature of inference with MVB. We anticipate that most analyses could use both Bayesian and classical inference. First, [Bayesian] model comparison would be used to identify the best qualitative form of model for any structure–function relationship, using the log-evidence over models. Having established the best model the cross-validation *p*-value can be used for a quantitative [classical] inference that any dependencies between observed brain measures and their consequences are unlikely to have occurred by chance.

Discussion

This paper has described a multivariate Bayesian (MVB) scheme to decode neuroimages. This scheme resolves the ill-posed many-to-one mapping, from voxels or data features to a target variable, using a parametric empirical Bayesian model with covariance hyperpriors. This model is inverted using expectation maximisation to furnish the model evidence and the conditional density of the parameters of each model. This allows one to compare different models or hypotheses about the mapping from functional or structural anatomy to perceptual and behavioural consequences (or their deficits). The primary aim of MVB is not to predict or classify these consequences but to enable inference on different models of structure–function

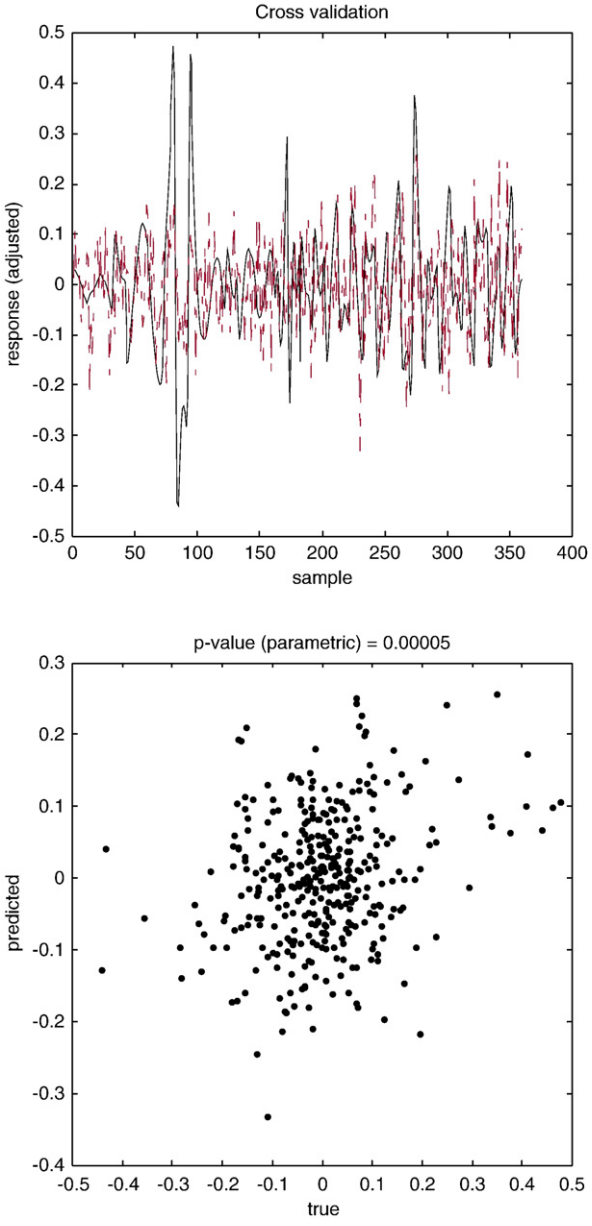


Fig. 15. Results of a two-fold cross-validation using the motion target and the 360 scans of 583-voxel features from the visual volume of interest. Upper panel: The target variable (solid line) and its prediction (broken line) from the cross-validation. Lower panel: the same data plotted against each other. These results can be compared with the lower panels in Fig. 9. The difference here is that the prediction of one half of the time-series is based on data from the other.

mappings; such as the distinction between distributed and sparse representations. This allows one to optimise the model itself and produce predictions that can outperform standard pattern classification approaches.

MVB is a model comparison approach that is well suited for testing specific hypotheses about structure–function mappings; e.g., is the representation of objects sparse or distributed in the visual cortex? The outputs of MVB are the log-evidences for the models tested, which allows inference about spatial coding and the conditional density of the voxel weights. In addition, one can also derive a cross-validation *p*-value for MVB models. On the other hand, classifier approaches like support vector machines (SVM) optimise the parameters of a discriminating function to maximise classification accuracy. They are useful when one wants to make predictions about new examples, especially when there is no prior hypothesis available or the model assumptions can not be guaranteed (e.g., classifying patients *vs.* controls, predicting treatment response, predicting subject decisions about novel stimuli, etc.). The outputs of this approach are the classification accuracy and the voxel weights.

Model inversion and inference in MVB uses exactly the same empirical Bayesian procedures developed for ill-posed inverse problems (e.g., source reconstruction in EEG). However, the MVB scheme extends this approach to include an efficient greedy search for sparse solutions. In contradistinction to top-down strategies, employed by things like automatic relevance determination, MVB uses a computationally expedient bottom-up search for the optimum partition; i.e., number and composition of pattern weight subsets with the same variance. It should be noted that there is a distinction between the abstraction of model comparison as the computation of a metric (e.g., likelihood ratio) and a search algorithm (e.g., greedy search). One reason to make this distinction lies in the need to consider when a particular search algorithm is appropriate; for example, a greedy search may be suitable for our purposes, in optimising linear models, yet may fail for nonlinear multivariate associations (e.g., the exclusive- or (XOR) association that eluded early neural-network solution algorithms).

We have illustrated MVB using simulated and real data, with a special focus on model comparison; where models can differ in the form of the mapping (i.e., neuronal representation) within one region, or in terms of the regions themselves. These demonstrations concluded with a procedure to compute exact *p*-values for classical inference on the model selected, using cross-validation. We organise the rest of the discussion around some obvious questions; many of which have been posed by colleagues after discussing the material above.

Can one use Bayesian decoding with event-related fMRI paradigms?

Yes. In fact the scheme can be applied to any data and design that can be formulated as a conventional linear model. This includes convolution models for fMRI studies with efficient design. Unlike classification schemes, the model does not map to classes or labels, but to continuous, real-valued target variables; therefore, one is not forced to assign each scan to a class. Furthermore, the target variable is convolved by a hemodynamic response so that the delay and dispersion inherent in fMRI measures of neuronal activity becomes an explicit part of the decoding.

Does the scheme assume the same hemodynamic response function in all voxels?

Not if variations in voxel-specific hemodynamic response functions are included as confounds. For example, to a first-order approximation, one can model differences in hemodynamic latency with the temporal derivative of the target variable (and any other confounds). This derivative enters the projection matrix, R, which effectively removes variations in latency in both the target variable and voxel time series. Similar arguments apply to other variations in the response function. In practice, decoding uses exactly the same model specification as encoding. The target variable is specified with contrast weights in the usual way but they are used to subtract the corresponding column (or mixture of columns) from the design matrix. This reduced design matrix now becomes the confound matrix in a decoding model and will contain regressors necessary for explaining away differences in hemodynamic responses (provided a suitable basis set was specified for the conventional analysis).

Can MVB be applied to structural images or contrasts in multi-subject studies?

Yes. As mentioned above, it can be used in any context that lends itself to conventional modelling. This includes the analysis of grey matter segments in voxel-based morphometry. This means it is possible to infer on structure–function mappings explicitly; for example one can use grey matter segments from a group to predict neuropsychological deficit, diagnosis or response to treatment. In fact, this application was one of the primary motivations for this work (see below).

Can the scheme cope with serial correlations in the errors?

Yes. These can be accommodated by adding extra error covariance components modelling any non-sphericity. The associated hyperparameter will be estimated in the EM scheme along with the others. In the current implementation, this is not necessary because we use the serial correlations V from a conventional analysis using ReML.

Is a greedy search for a sparse solution appropriate if the neuronal representation is not sparse (i.e., if it is distributed)?

Yes. A sparse solution in the space of pattern weights does not mean the solution is anatomically sparse because the patterns can be sparse or distributed. Both the hyperpriors and greedy search can accommodate sparse solutions in the space of pattern weights. This sparsity is a useful constraint on the many-to-one nature of the decoding problem; it means the scheme will seek an optimum sparse solution for any set of patterns that are specified. However, the patterns that model the anatomical deployment of neuronal activity may or may not be sparse. This means one can infer a representation is distributed by comparing two models with sparse (e.g., spatial vectors) and non-sparse (e.g., support vectors) patterns.

Will the greedy search find significant subsets when there is no mapping?

No. The free energy bound that is optimised by both the greedy search, and each iteration of the EM scheme, embodies both

accuracy and complexity. This means that adding a hyperparameter to the model will only increase the bound, if the increase in accuracy (i.e., fit) more than compensates for increased model complexity. This is why the log-evidence did not exceed the null model during the greedy search, using the null data in the simulations (see Fig. 5).

Why does Bayesian inference not use cross-validation like classification schemes?

Because it does not need to; classification schemes are generally obliged to use cross-validation or generalisation error to assess how good they are because they do not furnish inference on the mapping they are trying to model. Making inferences with classification still rests on model comparison but does so only at the last step, where one compares classification performance with a null model of chance classification (e.g., using a binomial distribution). From the perspective of model comparison, classification performance is a surrogate for the likelihood ratio. We exploited this approach to compute a *p*-value for classical inference using cross-validation.

Can the Bayesian decoding model be used to classify?

Generally, if a model is to be applied to classification problems, where the class labels are discrete, one usually uses logistic or multinomial regression, where the log-likelihood ratio is a linear function of some parameters. These models are based on binomial/multinomial distributions, as opposed to the Gaussian densities used in MVB. However, the continuous target variables, assumed by MVB, can be thresholded to give distinct classes or labels (for an example, see the comparison between MVB and SVM above). Having said this, the objective of Bayesian decoding is more general and is not simply to establish a statistical dependency between neuronal representations and a perceptual or behavioural consequences; it is concerned with comparing different models of that mapping. This is not possible with simple classification because classification schemes use only one model (the model that has been optimised with respect to generalisation error). Classification is therefore a special application of the more general MVB framework presented here.

We conclude with a brief review of extensions of the linear model presented here. These include nonlinear models and extensions to cover multiple target variables.

Nonlinear models

As mentioned above, we envisage applying this sort of analysis to look at structure–function relationships in the brain, using structural images (e.g., grey matter segments from multiple subjects). An important application here is lesion-deficit analysis, where one wants to understand how damage to different brain areas conspires to provide a behavioural deficit. A critical aspect of this mapping is that is may be nonlinear. In other words, the production of a deficit following damage to one region depends on the integrity of another (as in degenerate structure–function mappings). We have emphasised the necessary role of multivariate models in this context previously, when qualifying the use of voxel-based morphometry (Friston and Ashburner, 2004). There have been some exciting developments in this context; using directed Bayesian graphs (see Herskovits and Gerring, 2003). In the context of our parametric model, nonlinearities are easy to include, through the use of polynomial expansions. For example,[9]

$$WX = R\begin{bmatrix} y_1U & y_1U \otimes y_1U \\ \vdots & \vdots \\ y_sU & y_sU \otimes y_sU \end{bmatrix}\begin{bmatrix}\eta_1 \\ \eta_2\end{bmatrix} + \varsigma \tag{21}$$

can be treated as in exactly the same way as the first-order model above to provide the log-evidence and conditional estimates of the first and second-order pattern weights; η_1 and η_2. The first-order weights play exactly the same role as previously; however, the second-order weights model interactions between patterns in causing the target. An important example of this is predicting a psychological deficit by damage to two regions that have a degenerative (many-to-one) structure–function relationship (see Price and Friston, 2002). Under second-order degeneracy, a deficit would not be evident in damage to either region alone and would require a non-zero second-order weight on bilateral regional damage to predict a deficit. In principle, one could establish second-order degeneracy by comparing the second-order model above to its reduced first-order form.

$$WX = R\begin{bmatrix} y_1U \\ \vdots \\ y_sU \end{bmatrix}\eta_1 + \varsigma \tag{22}$$

which is exactly the same as Eq. (13). We will exploit nonlinear MVB in future work on multi-lesion deficit analyses of structural scans.

Comparing different representations

This paper has dealt with the simple case, where X is univariate (e.g., a subspace of a fuller design, specified with one-dimensional contrast). The more general case of multivariate decoding entails exactly the same formulation but with vectorised variables. An important example of this would be models for two contrasts or targets (e.g., house and face perception). A model for two perceptual targets X_1 and X_2 is

$$\begin{bmatrix} W_1X_1 \\ W_2X_2 \end{bmatrix} = \begin{bmatrix} R_1Y & -R_1Y \\ R_2Y & R_2Y \end{bmatrix}\begin{bmatrix} U\eta^{(+)} \\ U\eta^{(-)} \end{bmatrix} + \begin{bmatrix}\varsigma_1 \\ \varsigma_2\end{bmatrix} \tag{23}$$

This model has the same form as Eq. (13) but has been arranged so that the pattern weights $\eta^{(+)}$ map activity in patterns to both targets, whereas $\eta^{(-)}$ map differential activity to the target. This means that $\eta^{(+)}$ are weights that mediate overlapping representations and $\eta^{(-)}$ determine which patterns or voxels predict the targets uniquely. Note that the errors are uncorrelated because they are mixture of orthogonal voxel weights. By comparing this full model with a reduced model

$$\begin{bmatrix} W_1X_1 \\ W_2X_2 \end{bmatrix} = \begin{bmatrix} -R_1Y \\ R_2Y \end{bmatrix}U\eta^{(-)} + \begin{bmatrix}\varsigma_1 \\ \varsigma_2\end{bmatrix}\varsigma \tag{24}$$

one should be able to test for common or overlapping representations and disambiguate between category-specific representations

[9] The symbol ⊗ means Kronecker tensor product and is equivalent taking all the products of elements in two vectors (or matrices).

that are functionally selective (with overlap) and functionally segregated (without). We will explore this in future work.

Software note

The Bayesian decoding scheme, described in this paper, will be available in the next release of SPM5 (http://www.fil.ion.ucl.ac.uk/spm). It is accessed through the results panel (multivariate Bayes) after displaying a contrast as an SPM. It is assumed that the target variable is the compound or contrast of regressors specified by the contrast weights of the SPM. If an *F*-contrast is specified, the first component is used for decoding. The volume of interest is specified in the usual way (sphere, box or mask) and the greedy search is initiated for the model (spatial, smooth, singular or sparse) requested. After the model has been optimised or selected its cross-validation *p*-value can be accessed using a two-fold scheme illustrated in the main text. The results are displayed using the same format used (Figs. 9, 10 and 15) above.

Acknowledgments

The Wellcome Trust funded this work. This work was undertaken under the auspices of the Brain Network Recovery Group (http://www.brainnrg.org), sponsored by the James S. McDonnell Foundation.

References

Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26 (3), 839–851.

Beal, M.J., 1998. Variational algorithms for approximate Bayesian inference; PhD Thesis: http://www.cse.buffalo.edu/faculty/mbeal/thesis/, p. 58.

Bishop, C.M., Tipping, M.E., 2000. Variational relevance vector machines. In: Boutilier, C., Goldszmidt, M. (Eds.), Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, pp. 46–53.

Büchel, C., Holmes, A.P., Rees, G., Friston, K.J., 1998. Characterizing stimulus-response functions using nonlinear regressors in parametric fMRI experiments. NeuroImage 8 (2), 140–148.

Carlson, T.A., Schrater, P., He, S., 2003. Patterns of activity in the categorical representations of objects. J. Cogn. Neurosci. 15 (5), 704–717.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19 (2 Pt. 1), 261–270.

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

Efron, B., Morris, C., 1973. Stein's estimation rule and its competitors—An empirical Bayes approach. J. Am. Stat. Assoc. 68, 117–130.

Feynman, R.P., 1972. Statistical Mechanics. Benjamin, Reading MA, USA.

Friman, O., Cedefamn, J., Lundberg, P., Borga, M., Knutsson, H., 2001. Detection of neural activity in functional MRI using canonical correlation analysis. Magn. Reson. Med. 45 (2), 323–330.

Friston, K.J., 2007. Linear models and inference. In: Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., Penny, W.D. (Eds.), Statistical Parametric Mapping. Academic Press, London, pp. 589–591.

Friston, K.J., Ashburner, J., 2004. Generative and recognition models for neuroanatomy. NeuroImage 23 (1), 21–24.

Friston, K.J., Liddle, P.F., Frith, C.D., Hirsch, S.R., Frackowiak, R.S., 1992a. The left medial temporal region and schizophrenia. A PET study. Brain 115 (Pt 2), 367–382.

Friston, K.J., Frith, C.D., Passingham, R.E., Dolan, R.J., Liddle, P.F., Frackowiak, R.S., 1992b. Entropy and cortical activity: information theory and PET findings. Cereb. Cortex 2 (3), 259–267.

Friston, K.J., Frith, C.D., Frackowiak, R.S., Turner, R., 1995. Characterizing dynamic brain responses with fMRI: a multivariate approach. NeuroImage 2 (2), 166–172.

Friston, K.J., Holmes, A., Poline, J.B., Price, C.J., Frith, C.D., 1996. Detecting activations in PET and fMRI: levels of inference and power. NeuroImage 4, 223–235.

Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: theory. NeuroImage 16, 465–483.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. NeuroImage 34 (1), 220–234.

Grady, C.L., Haxby, J.V., Schapiro, M.B., Gonzalez-Aviles, A., Kumar, A., Ball, M.J., Heston, L., Rapoport, S.I., 1990. Subgroups in dementia of the Alzheimer type identified using positron emission tomography. J. Neuropsychiatry Clin. Neurosci. 2 (4), 373–384.

Grunwald, P., Pitt, M.A., Myung, I.J. (Eds.), April 2005. Advances in Minimum Description Length: Theory and Applications. M.I.T. Press (MIT Press).

Hanson, S.J., Matsuka, T., Haxby, J.V., 2004. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? NeuroImage 23 (1), 156–166.

Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. 72, 320–338.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science 303 (5664), 1634–1640.

Haynes, J.D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. 8 (5), 686–691.

Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nature reviews. Neuroscience 7 (7), 523–534.

Herndon, R.C., Lancaster, J.L., Toga, A.W., Fox, P.T., 1996. Quantification of white matter and gray matter volumes from T1 parametric images using fuzzy classifiers. J. Magn. Reson. Imaging 6 (3), 425–435.

Herskovits, E.H., Gerring, J.P., 2003. Application of a data-mining method based on Bayesian networks to lesion-deficit analysis. NeuroImage 19 (4), 1664–1673.

Kamitani, Y., Tong, F., 2006. Decoding seen and attended motion directions from activity in the human visual cortex. Curr. Biol. 16 (11), 1096–1102.

Kass, RE., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.

Kass, R.E., Steffey, D., 1989. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). J. Am. Stat. Assoc. 407, 717–726.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103 (10), 3863–3868.

Kim, H.-C., Ghahramani, Z., 2006. Bayesian Gaussian process classification with the EM-EP algorithm. IEEE Trans. Pattern Anal. Mach. Intell. 28 (12), 1948–1959.

Lautrup, B., Hansen, L.K., Law, I., Mørch, N., Svarer, C., Strother, S.C., 1994. Massive weight sharing: a cure for extremely ill-posed problems. In: Hermann, H.J., Wolf, D.E., Poppel, E.P. (Eds.), Proceedings of Workshop on Supercomputing in Brain Research: From Tomography to Neural Networks. HLRZ, KFA Julich, Germany, pp. 137–148.

Lukic, A.S., Wernick, M.N., Strother, S.C., 2002. An evaluation of methods for detecting brain activations from functional neuroimages. Artif. Intell. Med. 25, 69–88.

MacKay, D.J.C., 1997. Introduction to Gaussian processes. http://www.inference.phy.cam.ac.uk/mackay/gpB.pdf.

MacKay, D.J.C., 1999. Comparison of approximate methods for handling hyperparameters. Neural Comput. 11, 1035–1068.

MacKay, D.J.C., 2003. Information Theory, Inference, and Learning Algorithms. CUP, Cambridge, UK.

Mackay, D.J.C., Takeuchi, R., 1996. Interpolation models with multiple hyperparameters. In: Skilling, J., Sibisi, S. (Eds.), Maximum Entropy and Bayesian Methods. Kluwer, pp. 249–257.

Martinez-Ramon, M., Koltchinskii, V., Heileman, G.L., Posse, S., 2006. fMRI pattern classification using neuroanatomically constrained boosting. NeuroImage 31, 1129–1141.

Mattout, J., Phillips, C., Penny, W.D., Rugg, M.D., Friston, K.J., 2006. MEG source localization under multiple constraints: an extended Bayesian framework. NeuroImage 30, 753–767.

Moeller, J.R., Strother, S.C., Sidtis, J.J., Rottenberg, D.A., 1987. Scaled subprofile model: a statistical approach to the analysis of functional patterns in positron emission tomographic data. J. Cereb. Blood Flow Metab. 7, 649–658.

Nandy, R.R., Cordes, D., 2003. Novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data. Magn. Reson. Med. 50 (2), 354–365.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10 (9), 424–430.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. NeuroImage 22, 1157–1172.

Phillips, C., Mattout, J., Rugg, M.D., Maquet, P., Friston, K.J., 2005. An empirical Bayesian solution to the source reconstruction problem in EEG. NeuroImage 24, 997–1011.

Price, C.J., Friston, K.J., 2002. Degeneracy and cognitive anatomy. Trends Cogn. Sci. 6, 416–421.

Rasmussen, C.E., 1996. Evaluation of gaussian processes and other methods for non-linear regression. PhD thesis, Dept. of Computer Science, Univ. of Toronto, 1996. Available from http://www.cs.utoronto.ca/~carl/.

Ripley, B.D., 1994. Flexible non-linear approaches to classification. In: Cherkassy, V., Friedman, J.H., Wechsler, H. (Eds.), From Statistics to Neural Networks. Springer, pp. 105–126.

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 (4), 1104–1116.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.

Vapnik, V.N., 1999. The Nature of Statistical Learning Theory. Springer-Verlag0-387-98780-0.

Worsley, K.J., Poline, J.B., Friston, K.J., Evans, A.C., 1998. Characterizing the response of PET and fMRI data using multivariate linear models (MLM). NeuroImage 6, 305–319.

Zeki, S., 1990. The motion pathways of the visual cortex. In: Blakemore, C. (Ed.), Vision: Coding and Efficiency. Cambridge University Press, UK, pp. 321–345.