THE ROYAL
SOCIETY

Nonlinear PCA: characterizing interactions between modes of brain activity

Karl Friston[*], **Jacquie Phillips, Dave Chawla and Christian Büchel**

The Wellcome Department of Cognitive Neurology, Institute of Neurology, Queen Square, London WC1N 3BG, UK

This paper presents a nonlinear principal component analysis (PCA) that identifies underlying sources causing the expression of spatial modes or patterns of activity in neuroimaging time-series. The critical aspect of this technique is that, in relation to conventional PCA, the sources can interact to produce (second-order) spatial modes that represent the modulation of one (first-order) spatial mode by another. This nonlinear PCA uses a simple neural network architecture that embodies a specific form for the nonlinear mixing of sources that cause observed data. This form is motivated by a second-order approximation to any general nonlinear mixing and emphasizes interactions among pairs of sources. By introducing these nonlinearities principal components obtain with a unique rotation and scaling that does not depend on the biologically implausible constraints adopted by conventional PCA.

The technique is illustrated by application to functional (positron emission tomography and functional magnetic resonance imaging) imaging data where the ensuing first- and second-order modes can be interpreted in terms of distributed brain systems. The interactions among sources render the expression of any one mode context-sensitive, where that context is established by the expression of other modes. The examples considered include interactions between cognitive states and time (i.e. adaptation or plasticity in PET data) and among functionally specialized brain systems (using a fMRI study of colour and motion processing).

Keywords: functional neuroimaging; PCA; interactions; spatial modes; nonlinear unmixing; sources

1. INTRODUCTION

This paper introduces a new technique that falls under the heading of nonlinear principal component analysis (PCA), in the characterization of functional neuroimaging time-series. This technique identifies the underlying dynamics that determine the expression of spatial modes or patterns of brain activity where, in contradistinction to conventional PCA, the underlying causes can interact to produce second-order spatial modes. These second-order modes represent the modulation of one distributed brain system by another and provide for a parsimonious characterization of multivariate time-series that embody nonlinear interactions.

(a) *Eigenimage analysis*

In Friston *et al.* (1993) we introduced voxel-based PCA of functional neuroimaging time-series to characterize distributed brain systems implicated in sensorimotor, perceptual or cognitive processes. These distributed systems are identified with principal components or eigenimages that correspond to spatial modes of coherent brain activity. This approach represents one of the simplest multivariate characterizations of functional neuroimaging time-series and falls into the class of exploratory analyses. Principal component or eigenimage analysis generally uses singular value decomposition (SVD) to identify a set of orthogonal spatial modes that capture the greatest amount of variance, expressed over time. As such, the ensuing modes embody the most prominent aspects of the variance–covariance structure of a given time-series. Noting that the covariances among brain regions is equivalent to functional connectivity renders eigenimage analysis particularly interesting because it was among the first ways of addressing functional integration (i.e. connectivity) in the human brain. Subsequently eigenimage analysis has been elaborated in a number of ways. Notable among these are the application of canonical variate analysis (CVA; Friston *et al.* 1996*a*), multidimensional scaling (Friston *et al.* 1996*b*) and partial least squares (PLS; McIntosh *et al.* 1996). Canonical variate analysis was introduced in the context of ManCova (multiple analysis of covariance) and uses the generalized eigenvector solution to maximize the variance that can be explained by some explanatory variables relative to error. CVA can be thought of as an extension of eigenimage analysis that refers explicitly to some explanatory variables and allows for statistical inference. Partial least squares is another name for SVD and can be thought of as an eigenimage analysis of the cross covariances between two sets of data. It was first applied to neurophysiological data from different parts of the brain (the right and left hemispheres; Friston 1995) and has been developed to look at the relationships between imaging time-series and explanatory variables pertaining to experimental design and behaviour (McIntosh *et al.* 1996).

[*] Author for correspondence (k.friston@fil.ion.ucl.ac.uk).

Eigenimage analysis has been applied widely to positron emission tomographic (PET) and subsequently functional magnetic resonance imaging (fMRI) data. It is generally used as an exploratory device to characterize the main contributions to coherent brain activity. These variance components may, or may not, be related to experimental design and recently spontaneous or endogenous correlations have been observed in the motor system without experimental manipulation (Biswal *et al.* 1995). Despite its exploratory power eigenimage analysis is fundamentally limited for two reasons. First, it offers only a linear decomposition of any set of neurophysiological measurements, and second, the particular set of eigenimages or spatial modes obtained are uniquely determined by constraints that are biologically implausible. These aspects of PCA represent inherent limitations on the interpretability and usefulness of eigenimage analysis of biological time-series.

In this paper we introduce a new technique that resolves both the problems of linearity and non-uniqueness of the modes. This technique is a special case of nonlinear PCA that is motivated by a second-order approximation to any general interactions among a small number of 'sources' or 'causes' of variance in multivariate time-series. In brief, this technique identifies a small number of sources or components that explain the most variance in the observed data while allowing for high-order interactions among these sources. The sources are identified subject to, and only to, the constraint that they are orthogonal or uncorrelated. By virtue of the nonlinear interactions the sources are uniquely identified eschewing the need to refer to unnatural constraints.

(b) *The importance of nonlinear PCA in characterizing distributed brain systems*

As noted above, the two main limitations of conventional eigenimage analysis are that the decomposition of any observed time-series is in terms of linearly separable components characterized by their spatial modes and scores. Second, that the spatial modes are somewhat arbitrarily constrained to be orthogonal and account, successively, for the largest amount of variance. In general, the identification of independent components (independent component analysis or ICA) is only possible to within some permutation and scaling. PCA relaxes the requirement of independence and replaces it with orthogonality, introducing the further problem that there is no unique rotation of the principal components. In PCA, a unique rotation and permutation is obtained by requiring successive modes to account for the greatest amount of variance that remains once higher components have been removed. Scaling is constrained by ensuring the spatial modes have unit sum of squares.

From a biological perspective, the linearity constraint is a rather severe one. Because the decomposition is linear it precludes interactions among the causes of spatial modes. This is a highly unnatural restriction on the activity expressed by distributed brain systems, where one expects to see substantial interactions that render the expression of one mode sensitive to the expression of others. There are numerous examples of nonlinear interactions and modulatory effects that shape the context-sensitive nature of neuronal dynamics and brain activity. Perhaps the most compelling example, at a systems level, is attentional modulation. Consider two distributed brain systems, one subserving the processing of dynamic visual stimuli and the other responsible for a particular attentional set. In the context of attending to visual motion, the neuronal responses in the visual system, elicited by motion stimuli, will depend on whether the subject is attending to this attribute or not. Attentional status will be reflected in the activity of some attentional mode and therefore the expression of the visual processing mode will be a function of the expression of the attentional mode. It is more than likely that the implicit interaction between the visual and attentional modes will result not only in the degree to which the modes are expressed, but in their form or relative regionally specific contributions. For example, activity in visual area V5, that has been implicated in the processing of visual motion (Zeki 1990), may be enhanced (relative to say V2 or V1) whenever the appropriate attentional mode is being expressed (Treue & Maunsell 1996; Büchel *et al.* 1998). This context-sensitive expression of spatial modes can be modelled conceptually in terms of first- and second-order effects. In this example, there are two sources or causes of distributed neuronal responses, namely the presence of visual motion in the visual field and attention. Both these causes are expressed in terms of activity in their respective spatial modes and the interactions between these two causes would correspond to a second-order effect that was expressed in V5. These second-order effects can be thought of as changes in the first mode that depend on the expression of activity in the second mode or equivalently modulation of one mode that is sensitive to the context engendered by the other.

The example considered in this paper is based on a fMRI study of visual processing that was designed to address the interaction between colour and motion processing. We had expected to demonstrate that a 'colour' mode and 'motion' mode would interact to produce a second-order mode reflecting (i) reciprocal interactions between extrastriate areas functionally specialized for colour and motion, (ii) interactions in lower visual areas mediated by convergent backwards efferents, or (iii) interactions in the pulvinar mediated by corticothalamic loops). Two out of three of these predictions were seen (see § 3(b)).

In summary, to properly model the context-sensitive nature of distributed but coherent brain responses, it may be necessary to address interactions among spatial modes that allow for the modulation of one mode by another. These modulatory effects are second- or high-order in nature and correspond to an interaction among the underlying causes in determining a particular pattern of cortical responses. This is the principle motivation for the development and use of nonlinear forms of PCA.

This paper is divided into two sections. The first section reviews the theoretical background to nonlinear PCA, first in general terms and then the specific implementation proposed here. This section includes the theoretical motivation behind the particular form of the decomposition employed, how sources and modes are identified and how the ensuing modes can be interpreted. The second section is an illustrative application of nonlinear PCA to the original PET study used to illustrate eigenimage analysis (Friston *et al.* 1993) and a fMRI study of visual motion and colour processing that exemplifies the modulation of one brain system by another.

2. THEORETICAL BACKGROUND

(a) *Nonlinear PCA*

Nonlinear PCA (e.g. Kramer 1991; Softky & Kammen 1991; Karhunen & Joutsensalo 1994) is a natural extension of PCA in a sense that it aims to identify a small number of underlying components or sources causing a multivariate data set that best explain the observed variance–covariance structure. Nonlinear PCA is itself a variant of related nonlinear approaches to structural analysis that have been pioneered over the last few decades. These include nonlinear factor analysis (e.g. McDonald 1984) and nonlinear partial least squares (e.g. Wold 1992).

Imagine that we had *m* observations of an *n*-variate. For example *m* scans, where each scan comprised *n* voxels. We can represent these data as *m* points in an *n*-dimensional space. In conventional PCA, the first principal component corresponds to the direction of a line running along the principal axis of the resulting cloud of points. The direction of this line is specified by an *n*-vector that corresponds to the eigenimage and the projections of any one point on to this line give the expression of this component for the observation (i.e. point) in question. There are two equivalent perspectives on the role that the principal axis serves. The first is that the projection of the *m* points on to this line has the maximum dispersion or variance. In other words, the component scores of the first principal component has the most variance. The other perspective is that the average distance of any point from this line is minimized in relation to all possible lines. In other words, the first principal component is the pattern over the *n* voxels that minimizes the unexplained variance in the data. Nonlinear PCA adopts exactly the same principles but allows for curvilinear lines. In brief, a curve is fitted to the data in *n*-space such that the average distance of the data points from this principal curve is minimized (figure 1). This heuristic description highlights the intimate relationship between nonlinear PCA and the identification of principal curves or surfaces (Dong & McAvoy 1996). In the case of linear PCA, the principal axes are determined analytically using the eigenvector solution of the $n \times n$ covariance matrix. In nonlinear PCA there is no closed-form solution and iterative techniques are generally employed. These iterative approaches are usually best framed in terms of simple neural networks using gradient ascent or descent on the weights of the connections within the network. One attractive architecture (Kramer 1991) that has been used in this context is based on the notion of 'bottle-neck nodes'. These architectures have five layers with a mirror symmetry about the middle or third layer. The first and fifth layers represent inputs and outputs. The middle layer has, typically, a very small number, J, of nodes. The intermediate layers two and four have larger numbers of nodes or neurons than the input or output layers and employ some nonlinear activation function. The network is trained to reproduce its input at the outputs. This simple training forces the network to learn a nonlinear function of the inputs that best predicts the inputs themselves, subject to the constraint that it can be expressed as a function of small number, J, of sources (activities of the 'bottle-neck nodes' in the middle layer). The transformation from the input to the middle layer represents a projection of the data on to the J principal surfaces of the data and the nonlinear transformation from the middle layer to the output layer defines the form of these surfaces. There are some extremely interesting issues pertaining to the use of these architectures in identifying principal components of a nonlinear sort, but we will not pursue them here. In this paper, we take a somewhat different approach that embodies some explicit constraints on the form of the nonlinearities that may cause biological data and develop an alternative architecture that retains the two basic principles of (i) using 'bottle-neck nodes', and (ii) training the

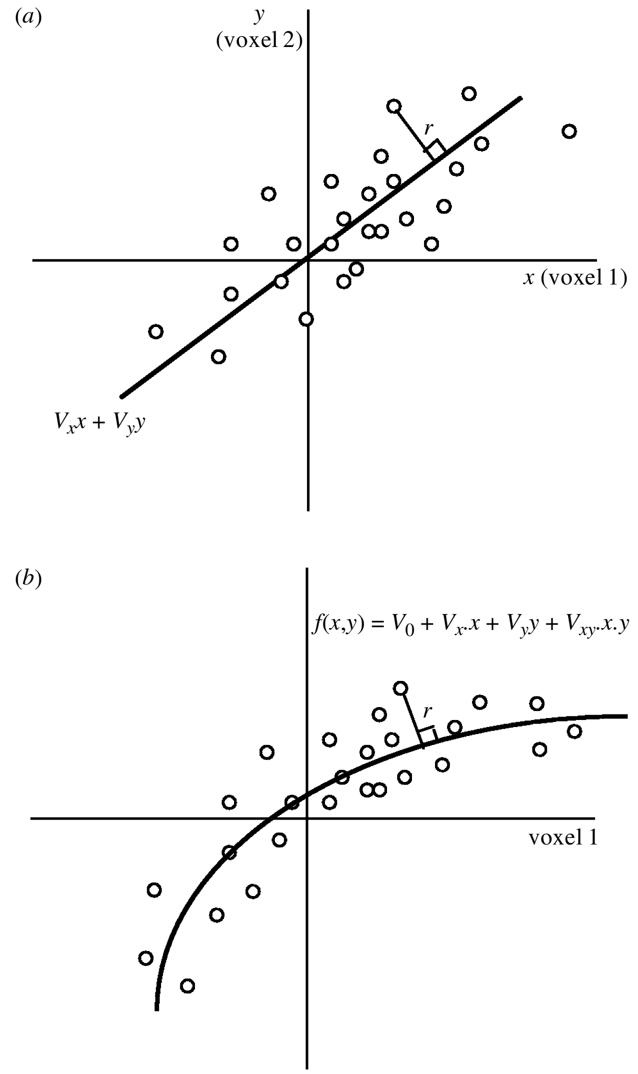

Figure 1. Schematic illustrating the idea behind nonlinear PCA and principal curves. In this simple example there are only two voxels and a series of images corresponding to points with values $\{x,y\}$ in the plots. A conventional PCA finds the principal line or axis given by a linear function of x and y that minimizes the average squared distance (r) of each point from that line (*a*). Nonlinear PCA is exactly the same but in this instance the axis is a curve given by some nonlinear function of x and y (*b*).

network so that the output best predicts the input in a least-squares sense.

(b) *Second-order PCA*

In this subsection we will introduce a variant of nonlinear PCA that uses a specific form for the assumed nonlinear mixing of sources to produce the observed responses in data. This form is predicated on interactions among sources in the genesis of multivariate time-series. In what follows we shall assume that an n-variate observation is caused by a small number of J underlying sources and interactions among these sources. Generally the observation of the ith variate (e.g. at the ith voxel) will be some nonlinear function of the underlying sources

$$y_i(t) = f_i(\boldsymbol{s}(t)), \qquad (1)$$

where $\boldsymbol{y}(t) = [y_1(t), \dots y_n(t)]$ is an n-vector function of time. Similarly for $\boldsymbol{s}(t) = [s_1(t), \dots s_J(t)]$.

A second-order approximation of the Taylor expansion of equation (1) about some expected value $\bar{\boldsymbol{s}}(t)$ for the sources is given by

$$y_i(t) \approx f_i(\bar{\boldsymbol{s}}) + \sum_j \frac{\partial f_i}{\partial u_j} u_j + \sum_{j,k} \frac{\partial^2 f_i}{\partial u_j \partial u_k} u_j u_k, \qquad (2)$$

where $\boldsymbol{u}(t) = (\boldsymbol{s}(t) - \bar{\boldsymbol{s}}(t))$ is an alternative representation of the sources. Now incorporating all n observations (i.e. voxels) equation (2) can be expressed, in matrix form, in terms of zeroth-, first- and second-order modes represented by the n-vectors $\boldsymbol{V}^0$, $\boldsymbol{V}^1$ and $\boldsymbol{V}^2$,

$$\boldsymbol{y}(t) \approx \boldsymbol{V}^0 + \sum_j u_j \boldsymbol{V}_j^1 + \sum_{j,k} u_j u_k \boldsymbol{V}_{jk}^2,$$

where

$$\boldsymbol{V}^0 = [f_1(\bar{\boldsymbol{s}}), \dots f_n(\bar{\boldsymbol{s}})], \quad \boldsymbol{V}_j^1 = \left[\frac{\partial f_1}{\partial u_j}, \dots \frac{\partial f_n}{\partial u_j}\right],$$

$$\boldsymbol{V}_{jk}^2 = \left[\frac{\partial^2 f_1}{\partial u_j \partial u_k}, \dots \frac{\partial^2 f_n}{\partial u_j \partial u_k}\right]. \qquad (3)$$

The elements of the vectors $\boldsymbol{V}^1$ correspond to the first order partial derivatives in equation (2) and the elements of the vectors $\boldsymbol{V}^2$ correspond to the second-order derivatives. $\boldsymbol{V}^1$ and $\boldsymbol{V}^2$ have the natural interpretation of first- and second-order modes, respectively. In other words the jth source is expressed in terms of the spatial mode $\boldsymbol{V}_j^1$ and the interaction between the jth and kth modes expressed as a spatial mode $\boldsymbol{V}_{jk}^2$. Equation (3) can be considered a special case of a more general equation that embodies higher-order terms:

$$\boldsymbol{y}(t) \approx \boldsymbol{V}^0 + \sum_j u_j \boldsymbol{V}_j^1 + \sum_{j,k} \sigma(u_j u_k) \boldsymbol{V}_{jk}^2, \qquad (4)$$

$\sigma(\cdot)$ is some sigmoid or squashing function that allows for slightly more general forms of interactions among sources and ensures a unique scaling for the sources $\boldsymbol{u}(t)$.

The above gives a suitable form for a nonlinear decomposition or PCA of a multivariate data set $\boldsymbol{y}(t)$. To identify the values of $\boldsymbol{u}(t)$ and the spatial modes it is necessary to assume a constraint of orthogonality for the sources. This is a natural constraint in a sense that the underlying causes of any biological data should be independent if they represent true causes, and as such will be orthogonal. Notice that the assumption of orthogonality is a weaker assumption than independence and it is this assumption that defines the algorithm described here as a nonlinear PCA. Had we assumed independence then the problem would become that of nonlinear independent component analysis (Common 1994) which would demand a different approach.

(c) *Neuronal architecture and identification of nonlinear components*

In this section the neuronal architecture and gradient descent scheme used to identify sources and their modes are described. Note that equation (4) can be treated as a general linear model and as such, if we knew the sources $\boldsymbol{u}(t)$, the modes could be estimated by minimizing the residuals trace$\{\boldsymbol{R}\}$ in a least squares sense, where

$$\boldsymbol{R} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{V}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{V}}), \qquad (5)$$

and

$$\hat{\boldsymbol{V}} = [\hat{\boldsymbol{V}}; \hat{\boldsymbol{V}}^1; \hat{\boldsymbol{V}}^2] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

$$\boldsymbol{X} = [\mathbf{1},\ u_1, \dots u_k,\ \sigma(u_1 u_2), \dots \sigma(u_k u_k)].$$

Here $\mathbf{1}$ is a column of ones and $\boldsymbol{I}$, below, is the identity matrix. $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}$ is simply the least-squares estimator of $\boldsymbol{V}$ given the inputs $\boldsymbol{y}$ and estimated sources (and their interactions) in $\boldsymbol{X}$. The problem therefore, reduces to identifying the variates $\hat{\boldsymbol{u}}(t)$ corresponding to estimates of the sources, that minimize the norm of the residuals trace$\{\boldsymbol{R}\}$. By noting the existence of some vector $\boldsymbol{G}_i$ where $\boldsymbol{V}^0\boldsymbol{G}_i = 0$, $\boldsymbol{V}_i^1\boldsymbol{G}_i = 1$ for $i = j$ and 0 otherwise and $\boldsymbol{V}_{jk}^2\boldsymbol{G}_i = 0$ then post-multiplying equation (4) throughout by $\boldsymbol{G}_i$ gives $\boldsymbol{u}_i(t) = \boldsymbol{y}(t) \times \boldsymbol{G}_i$. This means that there must be a linear combination of the inputs that gives the ith source. One simply has to find the linear combination of inputs that minimizes trace$\{\boldsymbol{R}\}$ for a given input $\boldsymbol{y}(t)$, subject to the constraint that the sources are orthogonal.

These observations lead to the following simple neural network: the network has three layers comprising input, middle and output layers. The input and output layers have n nodes and linear activation functions and can be imagined as lying next to each other (figure 2). The middle layer comprises a small ($J < n$) number of first-order nodes with linear activation functions that receive inputs from all the input nodes. In addition the middle layer includes $p = n(n-1)/2$ second-order nodes that receive lateral inputs from the first-order nodes. Each second-order node receives two inputs that are multiplied and subject to the nonlinear function $\sigma(\cdot)$ to provide their output. The network is trained on the feedforward connection strengths from the input layer to the first-order nodes of the middle layer. The weights to the ith first-order node are effectively estimates of $\boldsymbol{G}_i$. The connections from all middle-layer nodes to the outputs are determined using the least-squares estimators of the modes given the current estimate of the sources and the inputs according to equation (5). Anti-Hebbian lateral connections (Foldiak 1993) among the first-order nodes in the middle layer ensure that the sources $\hat{\boldsymbol{u}}(t)$ are orthogonal. These $J \times J$ lateral connection strengths $\boldsymbol{L}$ are

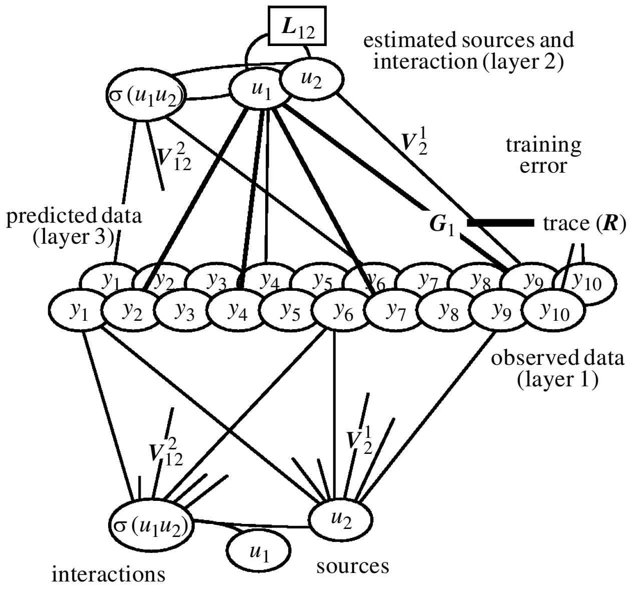

Figure 2. The neural net architecture used to estimate sources and modes. The lower half of the schematic represents the real world with its sources and interactions (here only two sources and the subsequent interaction are shown). These sources and interactions ($\boldsymbol{u}$) cause signals ($\boldsymbol{y}$) in the input layer (layer 1) that here comprises ten voxels or channels. The signals caused by the sources are weighted by the voxel-specific elements of the corresponding first- or second-order spatial modes ($\boldsymbol{V}^1$ and $\boldsymbol{V}^2$). Feedforward connections ($\boldsymbol{G}$) from the input layer to layer 2 provide an estimate of the sources ($\boldsymbol{u}$) in layer 2. This estimation obtains by changing $\boldsymbol{G}$ to minimize the sum of squared residuals (trace$\{\boldsymbol{R}\}$) or differences between the observed signals and those predicted by the activity in layer 3. The activity in layer 3 results from backwards connections from the estimated source and interaction nodes in layer 2. These backward connections are the estimates of the spatial modes ($\boldsymbol{V}^1$ and $\boldsymbol{V}^2$) and are determined using least-squares given the input ($\boldsymbol{y}$) and the current estimate of the sources ($\boldsymbol{u}$). Lateral decorrelating or anti-Hebbian connections $\boldsymbol{L}$ between the first-order modes ensure orthogonality of the source estimates. Note that in the absence of any interaction the solution would correspond to a conventional PCA where $\boldsymbol{G} = \mathrm{pinv}(\boldsymbol{V}^1)$.

determined at each iteration to render the off-diagonal elements of Cov$\{\hat{\boldsymbol{u}}\}$ zero:

$$\boldsymbol{L} = \boldsymbol{I} - \lambda^{-1}\Lambda^{1/2}E^T. \tag{6}$$

λ is a leading diagonal matrix whose elements correspond to the variance of the sources in the absence of decorrelating lateral connections (i.e. $\lambda = \mathrm{diag}\{\hat{\boldsymbol{u}}^{*T}\hat{\boldsymbol{u}}^*\}$ where $\hat{\boldsymbol{u}}^* = \boldsymbol{y} \times \boldsymbol{G}$). Λ and E are the eigenvalues and eigenvectors of $\hat{\boldsymbol{u}}^{*T}\hat{\boldsymbol{u}}^*$. Estimates of the sources are given by

$$\hat{\boldsymbol{u}} = \boldsymbol{y}\boldsymbol{G} + \boldsymbol{u}\boldsymbol{L} = \boldsymbol{y} \times \boldsymbol{G}(\boldsymbol{I} - \boldsymbol{L})^{-1}. \tag{7}$$

Note that substituting equation (6) into equation (7) gives Cov$\{\hat{\boldsymbol{u}}\} \propto \hat{\boldsymbol{u}}^T\hat{\boldsymbol{u}} = \lambda$, thereby ensuring orthogonality of the estimated sources. Implementing changes in the lateral connections in this way enforces orthogonality of the projection effected by the feed-forward connections at each iteration. $\boldsymbol{L}$ imposes this constraint because the effective feed-forward connections are $\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{L})^{-1}$. Essentially the architecture is finding the rotation and scaling, of some projection on to a low-dimensional orthogonal subspace that enables the interactions, among the projected inputs that ensue, to predict as much of the original inputs as possible. $\hat{\boldsymbol{u}}$ enters into equation (5) with $\boldsymbol{y}$ to compute trace$\{\boldsymbol{R}\}$. The learning rule for the estimates of $\boldsymbol{G}_i$ corresponds to gradient descent on the trace of the residuals. In practice, we use a Nelder–Mead simplex search as implemented in MATLAB (MathWorks Inc., Natick, MA, USA) that minimizes trace$\{\boldsymbol{R}\}$, which is simply a function of the input $\boldsymbol{y}$ and the feed-forward connections strengths $\boldsymbol{G}$. This neural network appears to be robust and usually converges within a few tens of iterations to give estimates of the underlying sources $\hat{\boldsymbol{u}}(t)$ and least-square estimates of corresponding spatial modes $\boldsymbol{V}$. Figure 2 is a schematic that tries to convey the simplicity of the architecture. Here the third or output layer has been reflected back on to the inputs to emphasize the symmetry between the assumed structure of causes in the world and the architecture used to identify them.

(d) *Interpreting the estimates*

Generally, in PCA, in the absence of any variance maximization–minimization criterion, there is no unique rotation of the spatial modes (i.e. any linear combination of the modes, that conforms to an orthogonal rotation, is as equally good as any other). The incorporation of the interaction term into the decomposition implicit in equation (3) ensures a unique rotation and, furthermore, incorporating the sigmoid function in equation (4) ensures that the scaling is uniquely determined. The latter follows from the fact that there will be some optimum squashing of each interaction term to best predict the observed data. The reason that unique solutions obtain in this form of nonlinear PCA is that we have assumed a very specific form for the high-order interactions among sources in causing the data. This is based on the pairwise interactions between sources as modelled by the second term in equation (4). The interpretation of the sources and their modes is relatively straightforward.

The observed multivariate data can be explained by a small number of J sources whose expressions are given by $\hat{\boldsymbol{u}}$. The values of $\hat{\boldsymbol{u}}_j$ scale the contribution of the first-order spatial mode $\boldsymbol{V}^1_j$ in a way that is directly analogous to conventional PCA, where $\hat{\boldsymbol{u}}_j$ would be the jth component score. In addition there are second-order effects that represent interactions between pairs of sources $\sigma(\hat{\boldsymbol{u}}_j\ \hat{\boldsymbol{u}}_k)$. These interactions are expressed in second-order modes corresponding to $\boldsymbol{V}^2_{jk}$. Each second-order mode will have a variance component that may or may not be orthogonal to the first-order modes. Although the sources are orthogonal there is no explicit requirement for the modes to be so. The variance accounted for by each source and interaction is given by

$$|u_j| \cdot |\boldsymbol{V}^1_j| \text{ and } |\sigma(u_ju_k)| \cdot |\boldsymbol{V}^2_{jk}|, \tag{8}$$

and can be used to rank the relative contributions of each source or interaction. $|\cdot|$ denotes the vector norm (i.e. sum of squares).

The nonlinear PCA proposed here therefore decomposes a multivariate data set into first- and second-order components that can be ascribed to a small number of underlying sources. The number of sources will be generally greater than the minimum number acquired to account for the rank of the multivariate data. For example, with just three voxels or channels, two sources would be sufficient if the third dimension was explained by the interaction between these two sources. Three sources would be sufficient to account for a six-dimensional data set and so on. Clearly in the analysis of functional neuroimaging time-series an initial dimension reduction is required before nonlinear PCA can be applied. For example, taking the first 36 spatial modes of a fMRI imaging time-series, using conventional singular value decomposition, would, in principle, require only eight underlying sources. In this sense, nonlinear PCA represents a parsimonious characterization of the data that captures nonlinear interactions among spatial modes or distributed brain systems in a comprehensive but intuitive fashion. These and other issues will be demonstrated in the next section, which uses real data to illustrate the technique.

3. ILLUSTRATIVE EXAMPLES

In this section we will use a multisubject PET study of verbal fluency and a fMRI case study of visual processing to illustrate the use of nonlinear PCA and some aspects of functional anatomy that can be addressed with this technique. The PET study is used to show that a considerable amount of variance can be accounted for by interactions among causes of the data and is presented for comparison with the original linear PCA characterization in Friston *et al.* (1993). In brief, we will show that the two experimental factors (task and time) combine to express themselves in a second-order mode that reflects time-dependent adaptation of task-related responses. This effect can be construed as large-scale neurophysiological plasticity attributable to strategic changes in cognitive processing during intrinsic, relative to extrinsic, generation of words. The second example, using fMRI, deals more explicitly with the modulation of one brain system by another. In particular the interactions between specialized cortical systems that may be mediated by corticothalamic loops.

(a) *A PET study of verbal fluency*

(i) *Data acquisition, experimental design and preprocessing*

The data were obtained from five subjects scanned 12 times (every 8 min) while performing one of two verbal tasks. Scans were obtained with a CTI PETcamera (model 953B, CTI, Knoxville, TN, USA). ^{15}O was administered intravenously as radiolabelled water infused over 2 min. Total counts per voxel during the build-up phase of radioactivity served as an estimate of regional cerebral blood flow (rCBF). Subjects performed two tasks in alternation. One task involved repeating a letter presented aurally, at one per two seconds (word shadowing). The other was a paced verbal fluency task, where the subjects responded with a word that began with the letter presented (intrinsic word generation). The data were realigned, stereotactically normalized and smoothed with a 16 mm Gaussian kernel (Friston *et al.* 1996*c*). The data were subject to a conventional SPM analysis using multiple linear regression with 12 condition-specific effects, five subject effects and global activity as described in Friston *et al.* (1995*a*). Parameter estimates, representing condition-specific effects averaged over subjects, were selected from voxels that exceeded a threshold of $p < 0.05$ in the ensuing SPM$\{F\}$and subject to nonlinear PCA as described below.

(ii) *Nonlinear PCA*

The data were reduced to an eight-dimensional subspace using SVD and entered into the nonlinear PCA using two sources. The ability of these two sources, and their interaction, to explain the observed regional activity is illustrated in figure 3*a*. Here an arbitrary voxel (that showing the highest *F*-value in the conventional SPM analysis) was selected from the left inferior frontal gyrus (Brodmann Area 47). The observed condition-specific activity, over 12 scans, is shown in black and that predicted by the two sources is shown in white. The relative amount of variance accounted for by the two sources and their interaction is shown in the middle panel. It can be seen that 88% of the total variance, over all voxels included in the analysis, can be explained by two sources. The second-order mode accounts of 2.2% of this (after removing that which can be modelled by the first-order effects) and would have been distributed over other modes in a conventional PCA. Figure 3*b* shows this distribution indicating that the fifth and sixth eigenimages, in a conventional PCA, largely comprise the interaction between the two modes identified by nonlinear PCA.

The first- and second-order modes are seen in figure 4, along with their expression over the 12 scans. It is immediately apparent that the first mode reflects task-related effects paralleling the alternation between word generation and word shadowing. This profile of brain regions is typical of verbal fluency paradigms that isolate the intrinsic generation of semantic representations, encoding and retrieval processes required to compare the current output with previous words and the maintenance of an appropriate cognitive set. The key regions involved include the thalamus, dorsolateral prefrontal cortex, anterior cingulate, temporal cortices and cerebellum. The second mode represents the other experimental factor, namely time or order effects. A nonlinear effect is evident with increases in activity in the cerebellar, thalamic and left basotemporal regions. More interesting is the second-order mode that, by implication, reflects an interaction between task-related responses and time, i.e. time-dependent increases in physiological responses elicited by cognitive operations that distinguish between the two tasks employed. This physiological adaptation in most pronounced in Broca's Area (Brodmann Area 44 in the left prefrontal cortex and the right lateral thalamic regions). Broca's Area is traditionally associated with speech production and appears to undergo a profound change in its relative activation during word shadowing and generation after the first pair of scans that, presumably, reflects an underlying change in cognitive architecture or set.

(b) *A fMRI study of colour and motion processing*

(i) *Data acquisition, experimental design and preprocessing*

The experiment was performed on a 2 Tesla Magnetom VISION (Siemens, Erlangen, Germany) whole body MRI system equipped with a head volume coil. Contiguous multislice T$_2^*$-weighted fMRI images (TE = 40 ms; 64 mm × 64 mm × 48 mm 3 mm × 3 mm × 3 mm voxels)

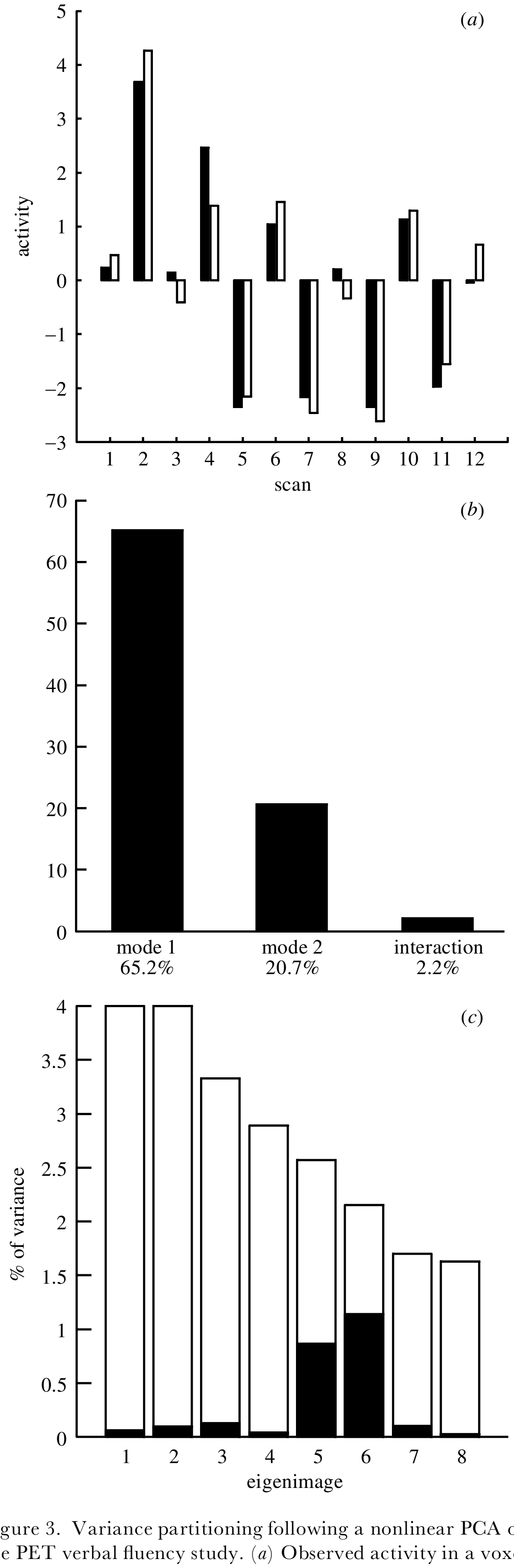

Figure 3. Variance partitioning following a nonlinear PCA of the PET verbal fluency study. (*a*) Observed activity in a voxel in the left inferior frontal gyrus (filled bars) and that predicted on the basis of two sources and their interaction (open bars) estimated with nonlinear PCA. Activity is in units corresponding to regional cerebral perfusion in ml dl^{-1} min^{-1}. (*b*) Variance, over all voxels included in the analysis, accounted for by the two sources (or modes) and their interaction. Total = 88%. (*c*) Distribution of variance accounted for by second-order or interaction effects over the conventional eigenimages obtained in the initial SVD dimension reduction.

were obtained with echoplanar imaging using an axial slice orientation. The effective repetition time was 4.8 s. A young right-handed subject was scanned under four different conditions, in six scan epochs, intercalated with a low level (visual fixation) baseline condition. The four conditions were repeated eight times in a pseudorandom order giving 384 scans in total or 32 stimulation–baseline epoch pairs. During all stimulation conditions the subject looked at dots back-projected on a screen by an LCD video projector. The four experimental conditions comprised the presentation of (i) radially moving dots, and (ii) stationary dots, using (i) luminance contrast and (ii) chromatic contrast in a two-by-two factorial design. Luminance contrast was established using isochromatic stimuli (red dots on a red background or green dots on a green background). Hue contrast was obtained by using red (or green) dots on a green (or red) background and establishing isoluminance with flicker photometry. In the two movement conditions the dots moved radially from the centre of the screen, at $8^\circ\,\text{s}^{-1}$, to the periphery where they vanished. This creates the impression of optical flow. By using these stimuli we hoped to excite activity in visual motion systems and those specialized for colour processing. Any interaction between these systems would be expressed in terms of motion-sensitive responses that depended on the hue or luminance contrast subtending that motion.

The time-series were realigned, corrected for movement-related effects and spatially normalized into the standard space of Talairach & Tournoux (1988) using the subject's co-registered structural T_1 scan and nonlinear deformations (Friston *et al.* 1996*c*). The data were spatially smoothed with a 6 mm isotropic Gaussian kernel. As in the PETexample, voxels were selected that showed significant condition-specific effects according to a conventional SPM analysis (Friston *et al.* 1995*b*; Worsley & Friston 1995). This analysis used a multiple linear regression and condition-specific box car regressors convolved with a haemodynamic response function. In this instance, the number of voxels was exceeding large and we used a higher threshold than in the PET analysis ($p = 0.001$) and included only those voxels that were posterior to the posterior commissure.

(ii) *Nonlinear PCA*

The data were again reduced to an eight-dimensional subspace using SVD and entered into the nonlinear PCA using two sources. The functional attribution of these sources was established by looking at the expression of the corresponding first-order modes over the four conditions. The expression of epoch-related responses over all 32 stimulation–baseline epoch pairs are shown in terms of the four conditions in figure 5. This expression is simply the score on the first principal component over all 32 epoch-related responses for each source. The first mode is

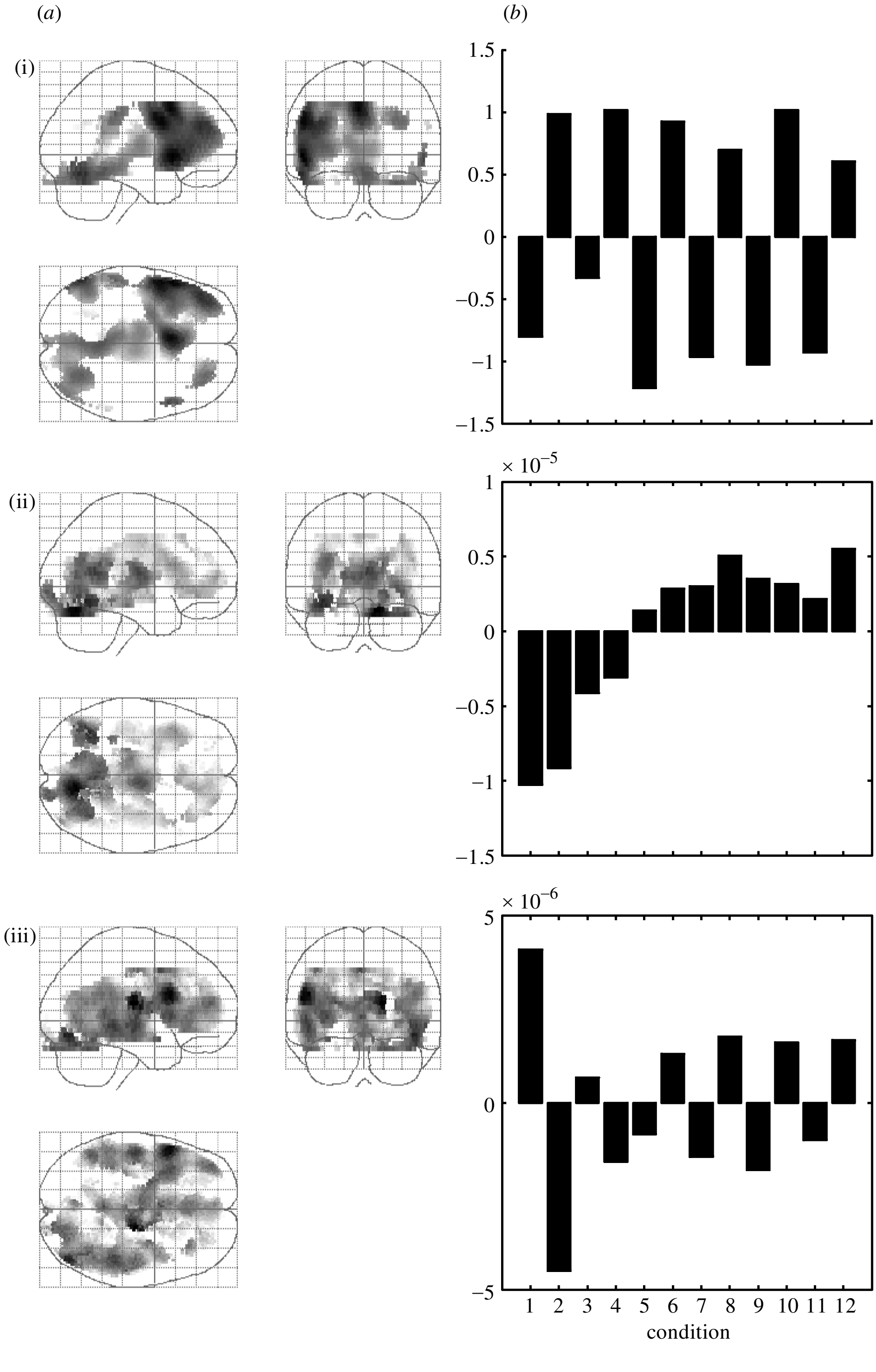

Figure 4. Maximum intensity projections and expression of the first- and second-order spatial modes of the PET verbal fluency study. (i) Spatial mode 1, (ii) spatial mode 2, and (iii) second-order mode. The maximum intensity projections (*a*) are of the positive values of each mode and are displayed in standard format. The three orthogonal brain views are from the right, the back and the top of the brain. The projections have been scaled to the maximum intensity of each mode. The time-dependent expression of these modes are in terms of the 12 scans (*b*). The units are adimensional and their absolute values are not important (the variance they account for is determined by the scaling of the spatial modes which, in turn, is dictated by the sigmoid squashing function, see figure 3).

clearly a motion-sensitive mode but one that embodies some colour preference in the sense that the motion-dependent responses of this system are accentuated in the presence of colour cues. This was not quite what we had anticipated; the first-order effect contains what would functionally be called an interaction between motion and colour processing. The second source appears to be concerned exclusively with colour processing in the sense that its expression is uniformly higher under colour stimuli relative to isochromatic stimuli in a way that does

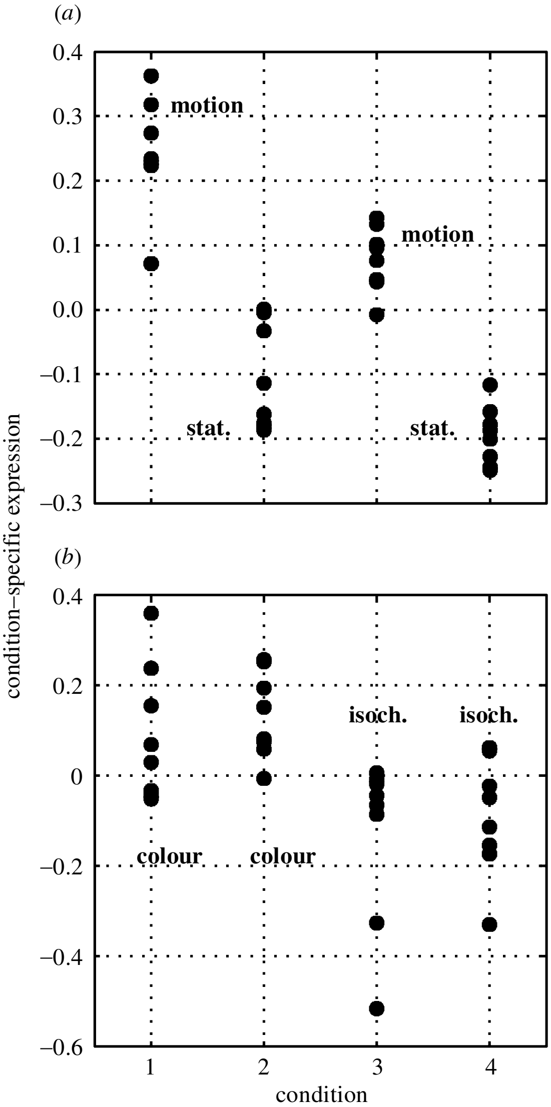

Figure 5. Condition-specific expression of the two first-order modes ensuing from the visual processing fMRI study. These data represent the degree to which the first principal component of epoch-related waveforms over the 32 photic stimulation–baseline pairs was expressed. These condition-specific responses are plotted in terms of the four conditions for the two modes. motion, motion present; stat., stationary dots; colour, isoluminant, chromatic contrast stimuli; isoch., isochromatic, luminance contrast stimuli. (*a*) First-order mode 1, (*b*) first-order mode 2.

not depend on motion. The corresponding anatomical profile is seen in figure 6 (maximum intensity projections in figure 6*a* and thresholded axial sections in figure 6*b*). The first-order mode, which shows both motion and colour-related responses, shows high loadings in bilateral motion-sensitive complex V5 (Brodmann Areas 19 and 37 at the occipto-temporal junction) and areas traditionally associated with colour processing (V4—the lingual gyrus, Brodmann Area 19 ventromedially). The second first-order mode is most prominent in the hippocampus, parahippocampal and related lingual cortices on both sides. The two more lateral blobs subsume the tails of the caudate nuclei (figure 6*b*(ii)). This system is not one normally associated with colour processing but it should be noted that some of the main effect of colour has been explained by the first mode that includes V4. In summary, the two first-order modes comprise (i) an extrastriate cortical system including V5 and V4 that responds to motion, and preferentially so when motion is supported by colour cues; and (ii) a (para)hippocampal–lingual system that is concerned exclusively with colour processing, above and beyond that accounted for by the first system. The critical question is where do these modes interact?

The interaction between the extrastriate and (para) hippocampal–lingual systems conforms to the second-order mode in the lower panels. This mode highlights the pulvinar of the thalamus and V5 bilaterally. This is a pleasing result in that it clearly implicates the thalamus in the integration of extrastriate and (para)hippocampal systems. This integration being mediated by recurrent (sub)corticothalamic connections. It is also a result that would not have obtained from a conventional SPM analysis. Indeed we looked for an interaction between motion and colour processing and did not see any such effect in the pulvinar. The reason that the nonlinear PCA was able to find this interaction was that there were no constraints on the sources underlying the interaction. In a conventional SPM analysis the sources are explicitly assumed to be colour and motion in the visual field, whereas the two interacting modes identified by the nonlinear PCA were caused by complicated admixtures of colour and motion. This result is presented to illustrate the potential usefulness of nonlinear PCA, not to make any statistical inferences about reproducible functional architectures. The exploratory analysis based on this case study could now be used to motivate hypothesis-led analyses of other subjects.

4. CONCLUSION

In this paper we have described a specific form of nonlinear PCA that is predicated on the interaction between underlying sources in modulating spatial modes of brain activity. Its theoretical motivation stems directly from a second-order approximation to the Taylor expansion of any nonlinear function of sources that can cause multivariate observations. A simple, three-layer neuronal network architecture is sufficient to identify or estimate the underlying causes and associated first- and second-order spatial modes. The first-order modes correspond to conventional eigenimages or principal components and the second-order modes describe the patterns of brain activity that result from interactions among these sources. The ensuing decomposition into first- and second-order components represents an exploratory analysis of the data that eschews some of the shortcomings of conventional PCA. In particular, nonlinear PCA allows for the context-sensitive expression of spatial modes through second-order modes that can be interpreted as the anatomical substrate of integration or modulation. The highly constrained form of nonlinear PCA presented above has an intuitive interpretation in terms of pairwise interactions among underlying sources and by virtue of this represents a useful and parsimonious characterization of functional neuroimaging time-series.

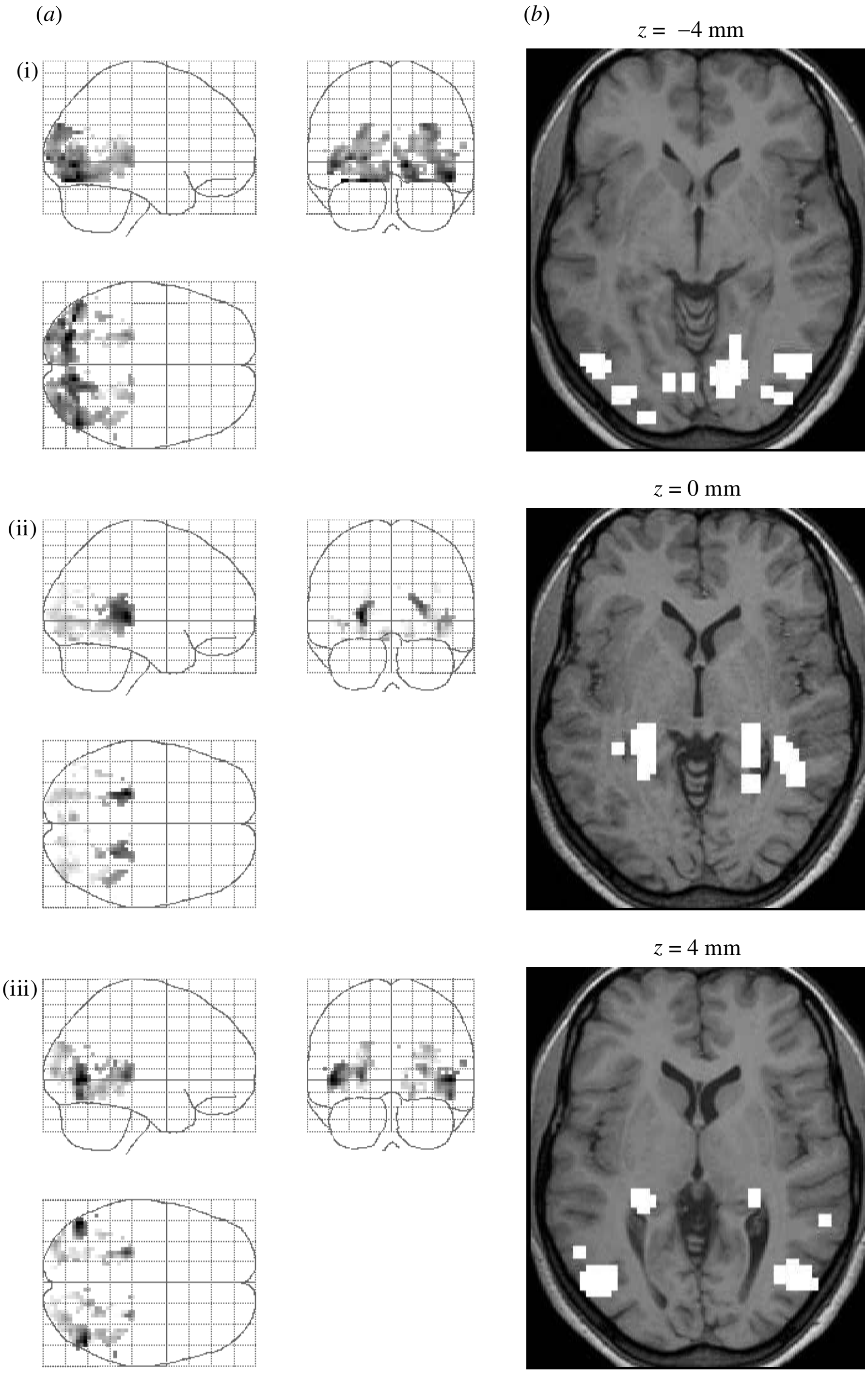

Figure 6. Maximum intensity projections and axial (transverse) sections of the first- and second-order spatial modes of the fMRI photic stimulation study. The maximum intensity projections (*a*) adhere to the same format as in figure 4. The axial slices have been selected to include the maxima of the corresponding spatial modes. (i) Spatial mode 1, (ii) spatial mode 2, and (iii) second-order mode. In this display format the modes have been thresholded at 1.64 of each mode's standard deviation over all voxels (white areas). The resulting excursion set has been superimposed onto a structural T_1-weighted MRI image conforming to the same anatomical space (*b*).

In this paper, we have chosen to illustrate the technique using the interaction between modes associated with colour and motion processing. Nonlinear PCA could of course be used in any situation where one expects the activity of a distributed brain system to be modulated by the expression of another system. Many examples come to mind that may, or may not, be grounded in cognitive science or neuroscience models. For example: Are the modes implicated in the visual processing of word forms and graphemes modulated by semantic modes in more anterior temporal and parietal cortices? Although nonlinear PCA is an exploratory device, and is implicitly data-led, careful experimental design can be used to control the expression of various spatial modes that one wishes to characterize. As a general point it is likely that the more powerful designs will be factorial in nature, allowing the expression of one mode, associated with one experimental factor, to be assessed under different levels of the expression of a second mode elicited by a second experimental cognitive or sensory factor. In some instances, factorial designs are not always easy to implement (e.g. in selective attention because it is difficult to attend selectively to a particular attribute when it is not present in the visual field). However, many multifactorial experiments, designed to look at language processing and memory, may lend themselves nicely to characterization using the techniques described in this paper.

(a) *Extensions and limitations*

The limitations of the nonlinear PCA proposed above are embodied in the constraints on the form of the decomposition assumed. The most obvious constraint is that it only allows for second-order interactions among sources or causes of the data, whereas higher-order interactions may prevail. It would, of course, be easy to extend the neural net architecture to include third- or higher-order nodes and this may be justified in some data analytic situations. In neuroimaging, however, the time-series one deals with are usually quite short and noisy and simply identifying second-order effects can be quite ambitious. The second limitation is that the number of sources has to be prespecified. Again this may be a drawback in terms of system identification and independent component analysis in general. However, in neuroimaging one has experimental control over the number of factors (i.e. sources) that are likely to cause neurophysiological changes and specifying the number of sources is a much more tenable, in terms of justifiable restrictions on the casual model assumed for the data.

Another important consideration is that, in the special application of nonlinear PCA to functional imaging data, an initial dimension reduction using SVD is required. This is because there are many more voxels than observations. It is well known that systematic errors can creep into applying SVD to simple nonlinear dependencies and that these depend on the rate of convergence of the Taylor series associated with equation (1). In this paper, the SVD is done first and then the nonlinear analysis is performed. It is always possible that the SVD has not established the right bases for the subsequent analysis and that some bias will ensue. The result will be that apparent modulations of the first-order modes will not be correct. These issues represent areas of future work and could be addressed using 'toy' nonlinear systems and synthetic imaging data and by examining the sensitivity of the second-order modes to the degree of SVD dimension reduction.

A more biological consideration relates to the mechanism of the interaction. In the examples presented above, we have assumed that the interaction is expressed at a neuronal level, in terms of modulation of neuronal responses and dynamics themselves. It should be borne in mind that interactions can also be expressed at the level of the haemodynamic response to [non-interacting] neuronal responses (e.g. Friston *et al.* 1998). This should be considered where there is a high degree of anatomical convergence between first-order modes that evidence a strong interaction.

By virtue of the iterative scheme used for learning in the neural net there is always the problem of local minima and the associated dependency on starting estimates. In our applications, we start with the conventional PCA solution and therefore ensure that the ensuing modes and interactions always account for more variance than the corresponding linear PCA. In this sense there is a unique solution (for any given gradient descent scheme) and this is the nearest to the solution where the second-order effects are zero.

Perhaps the most interesting limitation of the technique presented in this paper is buried in the assumption that there exists a linear combination of the inputs that gives the expression of the sources. This depends on the assumption (see §2) that first- and second-order modes are not collinear. As long as they are not collinear there is always a set of feed-forward connection strengths that span the subspace of one first-order mode that is orthogonal to all other modes (first and second order). What are the implications of collinearity between a first- and second-order mode? Collinearity means that the expression of a first-order mode is itself sensitive to the expression of another mode (i.e. the first- and second-order modes are the same thing). The possibility of this speaks to two fundamentally different context-sensitive effects. The first is when the interaction between two modes or causes is expressed as a second-order mode with a distribution that is distinct from both first-order modes. This is the situation considered in this paper and can be addressed using nonlinear PCA as described above. Second, the interaction may be expressed solely in terms of the expression of one of the two first-order modes. Here there is no second-order mode only a contextual expression of first-order modes. This second form of context-sensitivity requires a different sort of approach (nonlinear ICA) and is interesting because it may represents a true contextual effect with which the brain has to contend in everyday sensory processing.

This work was funded by the Wellcome Trust. We would also like to thank Gary Green and an anonymous reviewer for help in presenting this work.

REFERENCES

Biswal, B., Yetkin, F. Z., Haughton, V. M. & Hyde, J. S. 1995 Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Res. Med.* **34**, 537–541.

Büchel, C., Josephs, O., Rees, G., Turner, R., Frith, C. D. & Friston, K. J. 1998 The functional anatomy of attention to visual motion: a fMRI study. *Brain* **121**, 1281–1294.

Common, P. 1994 Independent component analysis, a new concept? *Signal Processing* **36**, 287–314.

Dong, D. & McAvoy, T. J. 1996 Nonlinear principal component analysis—based on principal curves and neural networks. *Comp. Chem. Engng* **20**, 65–78.

Foldiak, P. 1990 Forming sparse representations by local anti-Hebbian learning. *Biol. Cybern.* **64**, 165–170.

Friston, K. J. 1995 Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapp.* **2**, 56–78.

Friston, K. J., Frith, C., Liddle, P. & Frackowiak, R. S. J. 1993 Functional connectivity: the principal component analysis of large data sets. *J. Cerebr. Blood-Flow Metab.* **13**, 5–14.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C. D. & Frackowiak, R. S. J. 1995*a* Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189–210.

Friston, K. J., Holmes, A. P., Poline, J.-B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S. J. & Turner, R. 1995*b* Analysis of fMRI time-series revisited. *NeuroImage* **2**, 45–53.

Friston, K. J., Poline, J.-B., Holmes, A. P., Frith, C. D. & Frackowiak, R. S. J. 1996*a* A multivariate analysis of PET activation studies. *Hum. Brain Mapp.* **4**, 140–151.

Friston, K. J., Frith, C. D., Fletcher, P., Liddle, P. F. & Frackowiak, R. S. J. 1996*b* Functional topography: multidimensional scaling and functional connectivity in the brain. *Cerebr. Cortex* **6**, 156–164.

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D. & Frackowiak, R. S. J. 1996*c* Spatial registration and normalisation of images. *Hum. Brain Mapp.* **3**, 165–189.

Friston, K. J., Josephs, O., Rees, G. & Turner, R. 1998 Nonlinear event-related responses in fMRI. *Magn. Reson. Med.* **39**, 41–52.

Karhunen, J. & Joutsensalo, J. 1994 Representation and separation of signals using nonlinear PCA type learning. *Neural Networks* **7**, 113–127.

Krame, M. A. 1991 Nonlinear principal component analysis using auto-associative neural networks. *AIChE J.* **37**, 233–243.

McDonald, R. P. 1984 Confirmatory models for nonlinear structural analysis. In *Data analysis and informatics,* III (*Versailles, 1983*), pp. 425–432. Amsterdam: North-Holland.

McIntosh, A. R., Bookstien, F. L., Haxby, J. V. & Grady, C. L. 1996 Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage* **3**, 143–157.

Softky, W. & Kammen, D. 1991 Correlations in high dimensional or asymmetric data sets: Hebbian neuronal processing. *Neural Networks* **4**, 337–347.

Talairach, J. & Tournoux, P. 1988 *A coplanar stereotaxic atlas of a human brain.* Stuttgart, Germany: Thieme.

Taleb, A. & Jutten, C. 1997 Nonlinear source separation: the post-nonlinear mixtures. In *ESANN* '97 Bruges, April 1997, pp. 279–284.

Treue, S. & Maunsell, H. R. 1996 Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382**, 539–541.

Wold, S. 1992 Nonlinear partial least squares modelling. *Chemometrics Int. Lab. Syst.* **14**, 71–84.

Worsley, K. J. & Friston, K. J. 1995 Analysis of fMRI time-series revisited—again. *NeuroImage* **2**, 173–181.

Zeki, S. 1990 The motion pathways of the visual cortex. In *Vision: coding and efficiency* (ed. C. Blakemore), pp. 321–345. Cambridge University Press.