Revealing Interactions Among Brain Systems With Nonlinear PCA

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

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

Abstract: In this work, we present a nonlinear principal component analysis (PCA) that identifies underlying sources causing the expression of spatial modes or patterns of activity in neuroimaging time series *where these sources can interact to produce second-order modes.* This nonlinear PCA uses a neural network architecture that embodies a specific form for the mixing of sources that is based on a second-order approximation to any general nonlinear mixing. The modes obtained have a unique rotation and scaling that does not depend on the biologically implausible constraints adopted by conventional PCA. Interactions among sources render the expression of any mode or brain system sensitive to the expression of others. The example considers interactions among functionally specialized brain systems (using a fMRI study of colour and motion processing). *Hum. Brain Mapping 8:92–97, 1999.* © 1999 Wiley-Liss, Inc.

Key words: functional neuroimaging; fMRI; eigenimages; PCA spatial modes; nonlinear unmixing

INTRODUCTION

This report introduces nonlinear principal component analysis (PCA) in the characterization of neuroimaging time-series. This technique identifies the underlying dynamics that cause 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 imaging time-series that embodies nonlinear interactions.

Eigenimage analysis and nonlinearities

In Friston et al. [1993], we introduced voxel-based PCA of neuroimaging time-series to characterize distributed brain systems in terms of principal components or eigenimages that correspond to spatial modes of coherent brain activity. Principal component or eigenimage analysis generally uses singular value decomposition (SVD) to identify a set of orthogonal modes that capture the greatest amount of variance. As such they embody the most prominent aspects of the variance-covariance structure of a given time-series. Subsequently, eigenimage analysis has been elaborated in a number of ways. Notable among these are canonical variate analysis [CVA: Friston et al., 1996a], multidimensional scaling [Friston et al., 1996b], and partial least-squares [PLS: McIntosh et al., 1996]. Despite its exploratory power, eigenimage analysis is fundamentally limited because the particular modes obtained are uniquely determined by constraints that are biologically implausible. This represents an inherent limitation on the interpretability and usefulness of eigenimage analysis.

The two main limitations of conventional eigenimage analysis are: (1) the decomposition of any observed time-series is in terms of linearly separable compo-

*Correspondence to: Karl J. Friston, The Wellcome Dept. of Cognitive Neurology, Institute of Neurology, Queen Square, UK WC1N 3BG. email k.friston@fil.ion.ucl.ac.uk

Received for publication 11 February 1999; Accepted 16 June 1999.

nents, and (2) the spatial modes are somewhat arbitrarily constrained to be orthogonal and account, successively, for the largest amount of variance. From a biological perspective, the linearity constraint is a severe one because it precludes interactions among brain systems. This is a highly unnatural restriction on brain activity, where one expects to see substantial interactions that render the expression of one mode sensitive to the expression of others.

The example considered here 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: (1) reciprocal interactions between extrastriate areas functionally specialized for colour and motion, (2) interactions in lower visual areas mediated by convergent backwards efferents, or (3) interactions in the pulvinar mediated by corticothalamic loops.

Theoretical background

Nonlinear PCA [e.g., Kramer 1991; Softky and Kammen 1991; Karhunen and Joutsensalo, 1994; see also Wold, 1992] aims to identify a small number of underlying components or sources that best explain the observed variance-covariance structure of some data. This section describes a nonlinear PCA that uses a specific form for the nonlinear mixing of sources that emphasises interactions among sources in causing observed responses. Assume that an n-variate observation is caused by a small number of J underlying sources and interactions among these sources. Generally, the ith variate (e.g., observation at the ith voxel) will be some nonlinear function of the sources

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

where $y_i(t) = [y_1(t), \ldots y_n(t)]$ is an n-vector function of time. Similarly for $\mathbf{s}(t) = [s_1(t), \ldots . s_J(t)]$. The second-order approximation of the Taylor expansion of Eq(1) about some expected value $\bar{s}(t)$ for the sources is

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

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

$$\mathbf{y}(t) \approx V^0 + \sum_j u_j V_j^1 + \sum_{j,k} u_j u_k V_{jk}^2$$

where

$$V^0 = [f_1(\bar{\mathbf{s}}), \ldots f_n(\bar{\mathbf{s}})], \quad V_j^1 = \left[\frac{\partial f_1}{\partial u_j}, \ldots \frac{\partial f_n}{\partial u_j}\right], \qquad (3)$$

$$V_{jk}^2 = \left[\frac{\partial^2 f_1}{\partial u_j \partial u_k}, \ldots \frac{\partial^2 f_n}{\partial u_j \partial u_k}\right]$$

V^1 and V^2 are the first- and second-order modes, respectively. In other words, the jth source is expressed in terms of the spatial mode V_j^1 and the interaction between the jth and kth modes is expressed as the spatial mode V_{jk}^2. Eq(3) is a special case of

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

where σ(.) is some sigmoid or squashing function that ensures a unique scaling for the sources $\mathbf{u}(t)$. Eq(4) is a general linear model and as such, if we knew the sources $\mathbf{u}(t)$, the modes could be estimated by minimising the residuals trace{R} in a least squares sense, where:

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

and

$$\hat{V} = [\hat{V}^0;\ \hat{V}^1;\ \hat{V}^2] = (X^T X)^{-1} X^T \mathbf{y}$$

$$X = [1, u_1, \ldots u_K,\ \sigma(u_1 u_1), \sigma(u_1 u_2), \ldots\ \sigma(u_K u_K)]$$

Here, 1 is a column of ones and I is the identity matrix. The problem, therefore, reduces to identifying the variates $\hat{\mathbf{u}}(t)$, corresponding to estimates of the sources, that minimise the norm of the residuals trace{R}. One simply has to find the linear combination of inputs $u_i(t) = \mathbf{y}(t).\ \mathbf{G}_i$ that minimizes trace{R} for a given input $\mathbf{y}(t)$, subject to the constraint that the estimated sources are orthogonal.

This leads to the following simple neural network 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 (Fig. 1). 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 $p =$

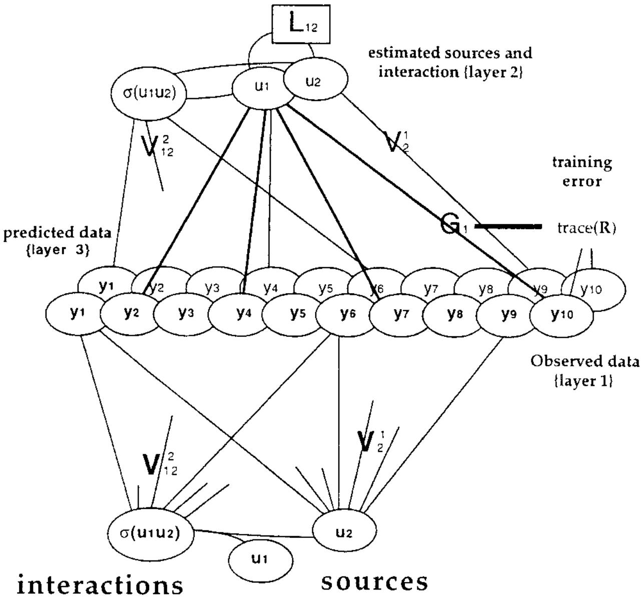

Figure 1.
Neural net architecture used to estimate sources and modes. The lower half of the schematic represents the real world with its sources and interactions (only two sources and the subsequent interaction are shown). These sources and interactions (u) cause signals (y) in the input layer (layer 1) that here comprises 10 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 ($\mathbf{V}^1$ and $\mathbf{V}^2$). Feed-forward connections (G) from the input layer to layer 2 provide an estimate of the sources (u) in layer 2. This estimation obtains by changing G to minimise the sum of squared residuals (trace{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 (V^1 and V^2) and are determined using least-squares given the input (y) and the current estimate of the sources (u). Lateral decorrelating or anti-Hebbian connections 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 G = pinv($\mathbf{V}^1$).

$n(n-1)/2$ second-order nodes that receive lateral connections from the first-order nodes. Each second-order node has two inputs [that are multiplied] and an activation function σ(.). The network is trained on $\mathbf{G}_i$, the feed-forward connection strengths from the input layer to the first-order nodes of the middle layer. The connections from 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 Eq(5). Lateral connections among the first-order nodes in the middle layer ensure that the sources $\hat{\mathbf{u}}(t)$ are orthogonal. These $J \times J$ lateral connection strengths L are determined at each iteration to ensure off-diagonal elements of Cov($\hat{u}$) are zero:

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

$\lambda = \text{diag}\{\hat{\mathbf{u}}^{*T}\hat{\mathbf{u}}^{*}\}$ where $\hat{\mathbf{u}}^{*} = \mathbf{y.G}$). Λ and E are the eigenvalues and eigenvectors of $\hat{u}^{*T}\hat{u}^{*}$. Estimates of the sources are given by

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

Note that substituting Eq(6) into Eq(7) gives Cov$\{\hat{\mathbf{u}}\} \propto \hat{\mathbf{u}}^T\hat{\mathbf{u}} = \lambda$, thereby ensuring orthogonality of the estimated sources. $\hat{\mathbf{u}}$ enters into Eq(5) with $\mathbf{y}$ to compute trace(R). We use a Nelder-Mead simplex search as implemented in MATLAB (MathWorks, Natick, MA) to find $\mathbf{G}$ that minimizes trace(R). This neural network appears to be robust and usually converges within a few tens of iterations to give estimates of the underlying sources $\hat{\mathbf{u}}(t)$ and least-square estimates of corresponding spatial modes $\hat{V}$.

The values of $\hat{u}_j$ scale the contribution of the first order spatial mode V_j^1 in a way that is directly analogous to conventional PCA, where $\hat{u}_j$ would be the jth component score. In nonlinear PCA, there are now second-order effects that represent interactions between pairs of sources $\sigma(\hat{u}_j\hat{u}_k)$. These interactions are expressed in second-order modes corresponding to V_{jk}^2. Each second-order mode will have a variance component that may or may not be orthogonal to the first-order modes. The variance accounted for by each source and interaction is given by:

$$|u_j|.|V_j^1| \quad \text{and} \quad |\sigma(u_j u_k)|.|V_{jk}^2| \tag{8}$$

and can be used to rank the relative contributions of each source or interaction. |.| denotes sum of squares.

fMRI study of colour and motion processing

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 × 64 × 48 3 × 3 × 3 mm voxels) were obtained with echoplanar imaging using an axial slice orientation. The effective repetition time was 4.8 sec. A young righthanded 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 pseudo-random 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 radially moving dots and stationary dots, using luminance contrast and 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° per second, to the periphery where they vanished. This creates the impression of optical flow. By using these stimuli, we hoped to excite activity in a visual motion system and regions 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 using the subject's co-registered structural T_1 scan [Friston et al., 1996c]. The data were spatially smoothed with a 6 mm isotropic Gaussian kernel. Voxels were selected that showed significant condition-specific effects according to a conventional SPM analysis [Worsley and Friston, 1995] using condition-specific box car regressors convolved with a hemodynamic response function. We included only those voxels that were posterior to the posterior commissure.

Nonlinear PCA

The data were reduced to an eight-dimensional subspace using SVD and entered into a nonlinear PCA using two sources. The functional attribution of the resulting 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 is shown in terms of the four conditions in Figure 2. This expression is simply the score on the first principal component over all 32 epoch-related responses for each source. The first mode is clearly motion-sensitive 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

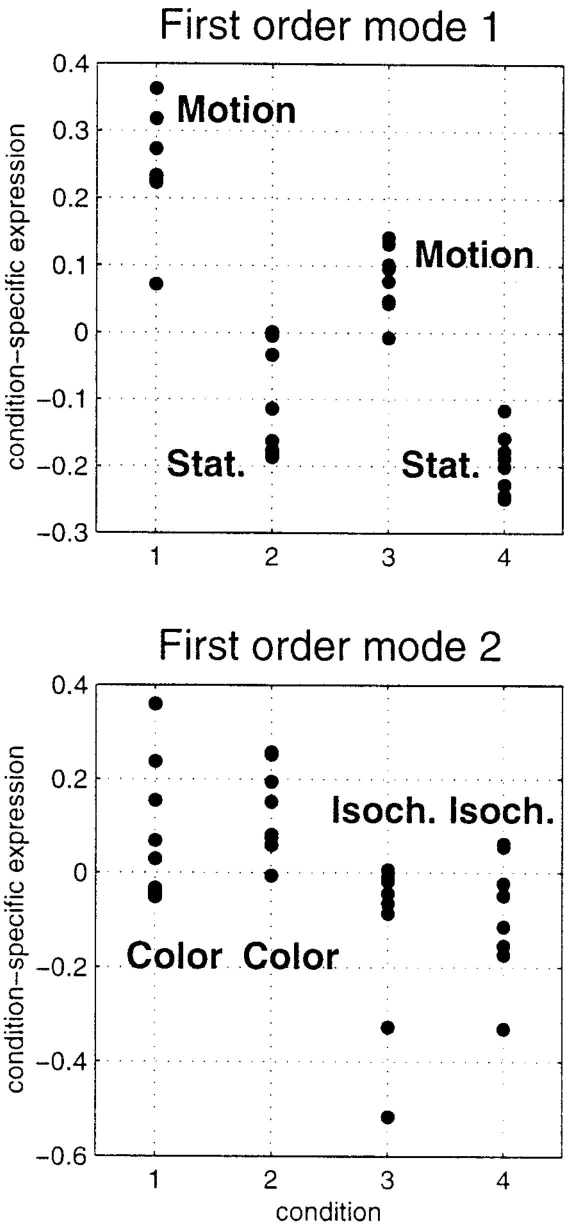

Figure 2.
Condition-specific expression of the two first-order modes ensuing from the visual processing fMRI study. These data represent the degree to which the first principal component of epoch-related waveforms over the 32 photic stimulation/baseline pairs was expressed. These condition-specific responses are plotted in terms of the four conditions for the two modes. Motion—motion present; Stat—stationary dots; Colour—isoluminant, chromatic contrast stimuli, Isochr—isochromatic, luminance contrast stimuli.

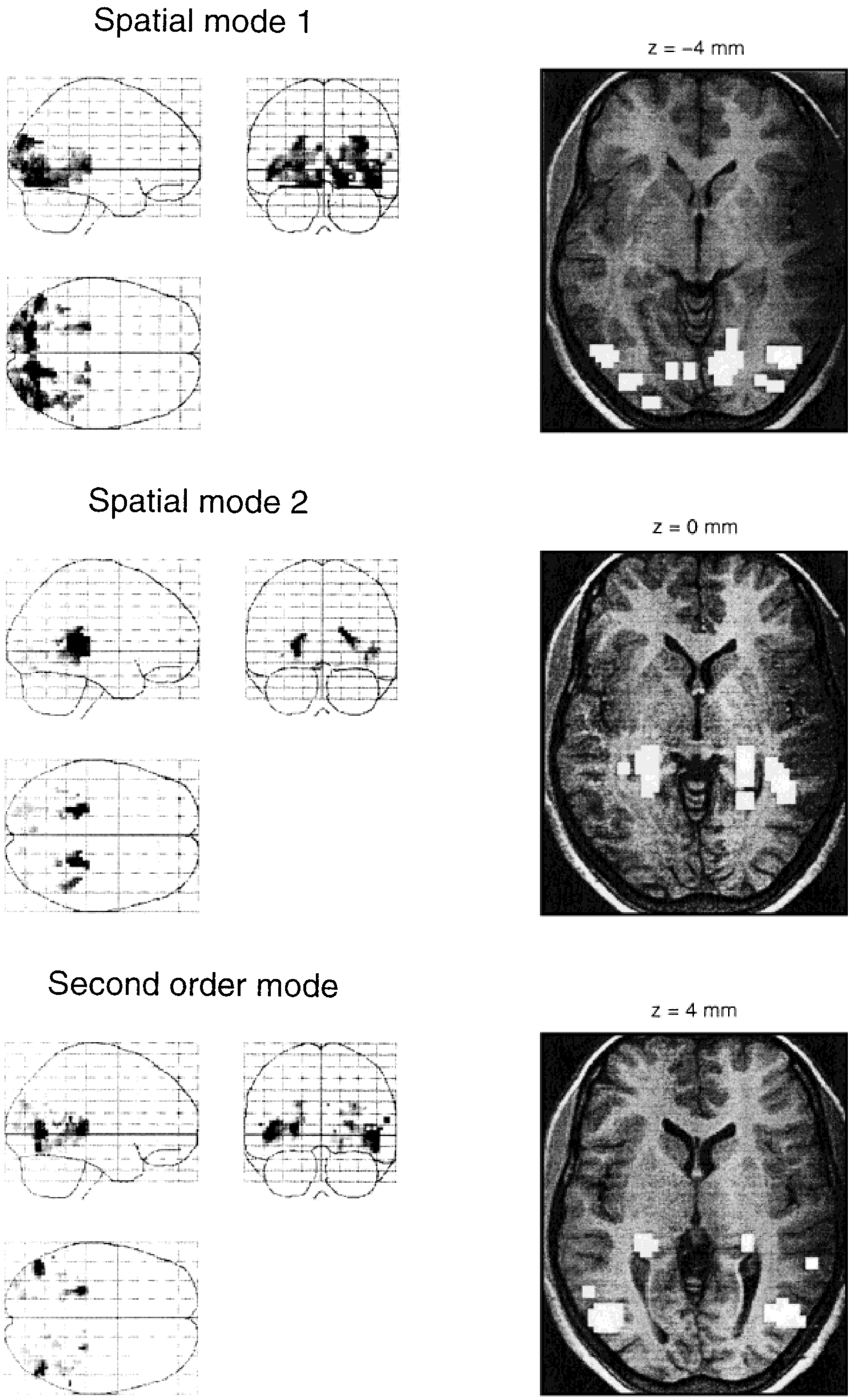

Figure 3.
Maximum intensity projections and axial (transverse) sections of the first- and second-order spatial modes of the fMRI photic stimulation study. The maximum intensity projections (left column) are in standard SPM format. The axial slices have been selected to include the maxima of the corresponding spatial modes. 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 T1 weighted MRI image conforming to the same anatomical space.

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 not depend on motion. The corresponding anatomical profile is seen in Figure 3 (maximum intensity projections on the left and thresholded axial sections on the right). 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 in a region that might correspond to V4α. The two more lateral blobs subsume the tails of the caudate nuclei (right middle panel). 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: (1) an extrastriate cortical system including V5 and V4 that responds to motion, and preferentially so when motion is supported by colour cues, and (2) 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]cortico-thalamic 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.

ACKNOWLEDGMENTS

This work was funded by the Wellcome Trust.

REFERENCES

Friston KJ, Frith C, Liddle P, Frackowiak RSJ. 1993. Functional connectivity: The principal component analysis of large data sets. J Cereb Blood Flow Metab 13:5–14.

Friston KJ, Poline J-B, Holmes AP, Frith CD, Frackowiak RSJ. 1996a. A multivariate analysis of PET activation studies. Human Brain Mapp 4:140–151.

Friston KJ, Frith CD, Fletcher P, Liddle PF, Frackowiak RSJ. 1996b. Functional topography: multidimensional scaling and functional connectivity in the brain. Cerebral Cortex 6:156–164.

Friston KJ, Ashburner J, Frith CD, Poline J-B, Heather JD, Frackowiak RSJ. 1996c. Spatial registration and normalisation of images. Hum Brain Mapp 3:165–189.

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

Kramer MA. 1991. Nonlinear principal component analysis using auto-associative neural networks. AIChE J 37:233–243.

McIntosh AR, Bookstien FL, Haxby JV, Grady CL. 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.

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

Worsley KJ, Friston KJ. 1995. Analysis of fMRI time-series revisited—again. NeuroImage 2:173–181.