◆ Human Brain Mapping 24:109–129(2005) ◆

Functional Principal Component Analysis of fMRI Data

Roberto Viviani, * Georg Grön, and Manfred Spitzer

Department of Psychiatry III, University of Ulm, Ulm, Germany

Abstract: We describe a principal component analysis (PCA) method for functional magnetic resonance imaging (fMRI) data based on functional data analysis, an advanced nonparametric approach. The data delivered by the fMRI scans are viewed as continuous functions of time sampled at the interscan interval and subject to observational noise, and are used accordingly to estimate an image in which smooth functions replace the voxels. The techniques of functional data analysis are used to carry out PCA directly on these functions. We show that functional PCA is more effective than is its ordinary counterpart in recovering the signal of interest, even if limited or no prior knowledge of the form of hemodynamic function or the structure of the experimental design is specified. We discuss the rationale and advantages of the proposed approach relative to other exploratory methods, such as clustering or independent component analysis, as well as the differences from methods based on expanded design matrices. *Hum Brain Mapp 24:109–129, 2005.* © 2004 Wiley-Liss, Inc.

Key words: principal component analysis (PCA); functional data analysis; independent component analysis (ICA); multivariate linear models (MLM); explorative methods

INTRODUCTION

Principal component analysis (PCA) is a technique to individuate important modes of variation in high-dimensional data as a set of orthogonal directions in space [Jolliffe, 1986]. Usually, only the few directions where most of the variation occurs are considered to be of interest. When PCA is used as an explorative technique, these directions are adopted as a new coordinate system to reveal the underlying structure of the data. The uses of PCA are not limited to exploration. When the data are composed of a set of interrelated variables, it is often useful to transform them with PCA to avoid problems arising from collinearity in multivariate regression settings. Because of the orthogonality of the directions of variation, in the new coordinate system the transformed data are uncorrelated, while retaining most of their variance. In fact, common techniques to carry out PCA formulate the problem as one of identifying the eigenvalues (amount of variation) and eigenvectors (direction of variation) of the covariance matrix of the data. We focus, however, on PCA as an explorative technique to visualize the variance introduced into the data by alternating experimental conditions without using much information from the experimental setting or making specific assumptions on the form of the regressor. As in other explorative techniques, the objective is that of providing an initial assessment that will give the data a chance "to speak for themselves" before an appropriate model is chosen. For example, it might be desirable to analyze data from a sample of patients without making assumptions on the form of the blood oxygenation level-dependent (BOLD) response, to verify that possible differences in the activations are genuine and do not depend on altered BOLD responses in the clinical condition [Callicott and Weinberger, 1999]. Another possible application of explorative techniques consists of the analysis of functional imaging data when there is uncertainty as to the duration of a mental state induced by an experimental stimulus, for example an emotional reaction, or as to the moment of its occurrence. In general, recourse to explorative techniques

*Correspondence to: Roberto Viviani, Department of Psychiatry III, University of Ulm, Leimgrubenweg 12, 89075 Ulm, Germany. E-mail: roberto.viviani@medizin.uni-ulm.de

Received for publication 17 December 2002; Accepted 21 June 2004
DOI: 10.1002/hbm.20074
Published online in Wiley InterScience (www.interscience.wiley.com).

may be justified by questions about differences in the form and shape of the hemodynamic response [Friston et al., 1995a], or when the shape of the hemodynamic response is itself the object of enquiry.

When applied to magnetic resonance images, ordinary PCA runs into serious difficulties because of the extremely high number of dimensions in the data relative to the number of observations. Even if the geometric properties of PCA remain valid and numerical techniques deliver stable results, the covariance matrix on which the analysis is carried out is sometimes a poor estimate of the real population covariance.

Because of these difficulties, ordinary PCA is often limited to regions of interest identified previously by the experimental model [*F*-masking; Friston et al., 1993; Friston, 1997]. This means that the PCA is carried out mostly in a subspace where the experimental regressors dominate, which will be much smaller than the space of the original data. Although this is a meaningful way of employing ordinary PCA, it also limits its scope in exploratory analyses, because the signal of interest has already been identified by other means. Another approach to constrain the outcome of PCA is partial least squares [McIntosh et al., 1996], in which the directions in space are selected that both maximize the variance of the data and correlate with the predictors of the design matrix. As it has been noted, the results are not invariant under changes of the design matrix [Petersson et al., 1999], which demonstrates that as in *F*-masking, information about the experimental setting is used directly by the method.

The approach adopted here stems from a field of statistics known as functional data analysis [Ramsay and Silverman, 1997], which has its roots in much earlier studies on growth curves [Rao, 1958] and in nonparametric regression [Eubank, 1988]. Functional data analysis exploits the fact that functions defined on a specific domain form an inner product vector space, and in most circumstances can be treated algebraically like vectors. Counterparts of conventional multivariate statistical methods are carried out in functional space rather than in the space spanned by vectors of individual observations.

Consequently, the data delivered by the functional magnetic resonance imaging (fMRI) scans will be here considered as continuous functions of time sampled at the interscan interval and subject to observational noise. These functions may be estimated by fitting a set of basis functions to each voxel time series. Collectively, the functions replace the voxels of a series of images with a single "functional image." In functional PCA, the eigenanalysis is carried out directly on these functions.

As a consequence, the eigenanalysis takes place in the space spanned by the basis functions set. The decisive advantage of this approach consists in the possibility of specifying a set of assumptions in the choice of the basis function set and in the error functional minimized by the fit. These assumptions will be weaker than the specification of a predefined hemodynamic function and a set of events or conditions as in *F*-masking, thus preserving the exploratory character of the procedure; however, the assumptions might remain stringent enough to overcome the difficulties of ordinary PCA. At a minimum, these assumptions will include the continuity and some degree of smoothness of the estimated functions, but in principle any set of differential equations may be used to specify a restriction on the functional estimation space. In the following, we will sometimes adopt periodic boundary conditions as a further restriction on the estimated functions to exploit knowledge of the periodicity of the experimental design.

As a multivariate technique, functional PCA differs from other multivariate approaches such as multivariate linear models (MLM) [Worsley et al., 1997] in the way the abovementioned assumptions are harnessed to constrain the estimate. This point will be expanded upon further in the final discussion.

As an explorative tool, functional PCA also has to contend against other advanced explorative techniques such as independent component analysis (ICA) [McKeown et al., 1998] and clustering [Wismüller et al., 2002]. We comment briefly on the reasons that justify the use of functional PCA in addition or as an alternative to such methods. As a preliminary remark, we note that in an experimental setting the controlled application of the experimental conditions or treatments constitutes a source of systematic variance in the data. A technique that selects components by capturing variance therefore seems a plausible complement to techniques such as ICA that recover a signal based on statistical independence and the extent of its departure from Gaussian normality [Hyvärinen and Oja, 2000; Lee et al., 1999]. In contrast, we make no distributional assumption on the form taken by the systematic variance introduced by the experimental manipulations.

Because the aims of functional PCA are exploratory, we will not be concerned with inferential issues such as producing a parametric image of significance values.

SUBJECTS AND METHODS

The fMRI data were obtained from three healthy young volunteers with a 1.5-Tesla Magnetom Vision (Siemens, Erlangen, Germany) whole-body MRI system equipped with a head volume coil after obtaining written consent. For block-design data (working memory and episodic memory encoding), images were obtained using echo-planar imaging (EPI) in axial orientation (T2*-weighted, TR/TE = 3,980/66 ms). Image size was 64 × 64 pixels (3.6 × 3.6 mm). For each volume, 32 slices were acquired. Slice thickness was 3.0 mm with a gap of 0.6 mm; thus, voxel size was isotropic. After discarding the first 6 volumes to allow for equilibration effects, 96 volumes were acquired. For further details on the tasks, see Walter et al. [2003a] and Grön et al. [2001]. For event-related data (TR/TE = 2,496/66 ms), in-planar resolution of the axial images was the same as above; 22 slices were acquired for each volume, and slice thickness was again 3.6 mm. During the event-related paradigm, the subject had to press a response button with his right index finger after a visual cue. After discarding the first 6 volumes,

within the time series of the remaining 102 volumes there were nine repetitions with a mean interscan interval (ISI) of 28.1 s.

All code was developed on MATLAB 6.1 R12 (MathWorks, Natick, MA) installed on a Pentium PC running Windows 2000 (Microsoft, Redmond, WA). Our implementation is written as a statistical parametric mapping (SPM) toolbox (SPM99; Wellcome Department of Cognitive Neurology, London; online at http://www.fil.ion.ucl.ac.uk). We therefore made use of the routines in that package for realignment, stereotactic normalization, segmentation, smoothing, application of a high-pass filter, and visualization of data [Ashburner and Friston, 1997], as well as for the comparison analyses using SPMs obtained from an experimental regressor [Holmes et al., 1997]. For the functional data analysis methods, our code builds on the library developed by Ramsay and Silverman [2001]. To display slices of principal component images, software developed by Brett [2000] was used.

Estimation of the Functional Image

Usually, the first step of any functional data analysis consists in subjecting the images to some preprocessing such as realignment, stereotactic normalization, and smoothing [Ashburner and Friston, 1997]. Unless low-frequency signals are of specific interest, a high-pass temporal filter should also be applied. In the case studies presented here, we applied standard normalization procedures and spatially smoothed the data with a Gaussian kernel of 8 mm. To reduce the influence of sources of undesired variance, a gray matter mask was applied to the image. To prepare the mask, the normalized image was segmented into gray, white matter, and cerebrospinal fluid (CSF). The mask was obtained by thresholding the parametric image of gray matter at about 0.1. Finally, the average of each voxel time series is calculated and subtracted from each voxel of the series.

The preprocessed N scans constitute the $N \times M$ matrix of data $\mathbf{Y}$, in which each row $\mathbf{d}_1', \mathbf{d}_2', \ldots \mathbf{d}_N'$ is one volume, and each column $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M$ is a time series composed of one of the M voxels sampled in each scan. The estimated functional image consists of M image functions $f_1, f_2, \ldots, f_M$, each obtained from independently fitting the basis function set to the columns of $\mathbf{Y}$. In the fit, therefore, the predictor variable is an index representing the time point at which the scan was taken. The response variable is the signal recorded by the scanner at a specific voxel, after the preprocessing steps described above.

The ways in which the predictors are obtained may be classified into two broad groups. When the experimental design consists of a periodic repetition of a sequence of conditions (as is commonly the case in a block design, for example, or in an event-related design where the stimuli are presented at sufficiently large intervals), the time indices of each point of the voxel time series may be folded around the period of the repetition. In this case, the range of the predictor variable is not the time span of the whole series, but that of the period itself. A variation of this scheme, applicable to relatively rare events, consists in defining a time window that is applied repeatedly in correspondence of each event, then discarding the other scans. This first group of "periodic" schemes is characterized by the fact that knowledge about the occurrence of an experimentally controlled condition is incorporated in the fitting process.

The second type of scheme for the generation of the predictor vector applies when the events or conditions are not known, occur irregularly or so close to each other that the resulting hemodynamic response cannot be expected to be repeated periodically. In this "nonperiodic" case, the predictor variable ranges over the time points of the scans. Finally, in all schemes the indices of the voxel time series are shifted to take interslice acquisition time into account, unless the volumes were preprocessed previously to correct for the interslice acquisition delays.

Formally, these schemes define a function $t(n)$, mapping the elements of scan sequence $n = 1, 2, \ldots, N$ to the elements of a predictor vector $\mathbf{t}$. If we denote with a and b the minimum and maximum of the elements of $\mathbf{t}$, the closed interval $[a, b]$ represents the domain over which the estimated functions will be defined. (More precisely, if the predictor vector has been shifted to account for the interslice acquisition time and is not periodic, not all voxel time series will start and end at the same time-point. In this case, the functions will be estimated on the interval containing the points shared by all intervals. In the interest of clarity, we will ignore this largely implementational detail in the rest of the discussion). Within the interval $[a, b]$, P will be the number of distinct points at which a scan is taken. For periodic schemes, $P < N$ if repeated observations are taken at the same point of the period; in all other cases, $P = N$.

We can now specify our model as follows:

$$y_{nm} = f_m(t(n)) + \epsilon_{nm}, \quad n = 1,2, \ldots N; m = 1,2, \ldots M \tag{1}$$

where the value y_{nm} of the mth voxel in the nth volume is given by a function f_m sampled at the time points of the predictor vector, and the error term ϵ_{nm} is subject to the usual assumptions of being distributed independently and identically with zero mean and constant finite variance. To ensure their tractability by the mathematical machinery that follows, we will also need to make some preliminary minimal assumptions on all f_m of little practical consequence, such as being continuously differentiable with square integrable second derivative.

Each function is estimated separately in each voxel time series; hence, to simplify notation, the subscript m indexing the voxel will not be specified when the context allows it. To estimate each f as a continuous function of t, $t \in [a, b]$, we will express it as a linear combination of a complete set of K suitable basis functions φ_k: $f(t) = \sum_{k=1}^{K} c_k \varphi_k(t)$, $k = 1, 2, \ldots, K$, so that the outcome of the estimate will be, as in a standard linear model, the set of coefficients c_k. The best-known basis set is probably the Fourier series $f(t) = c_0 + \sum_{k=1}^{K} (c_{2k-1} \sin k\omega t + c_{2k} \cos k\omega t)$, with period $2\pi/\omega$. The advantage of this basis set is that for uniformly spaced arguments, the basis functions are

orthogonal, and k has an explicit interpretation as the index of the increasing frequency of the basis functions. This basis set, however, tends to impose a uniform curvature on the estimated f across its domain. For this reason, most case studies presented here make use of regression splines, which are functions obtained by joining segments of polynomials smoothly at points called knots. The smoothness of the joints is ensured by imposing equality of a number of derivatives (usually equal to the degree of the polynomial) at the interior knots. Several schemes have been devised to represent splines conveniently [see de Boor, 1978]; the implementation adopted here makes use of B-splines. In principle, other basis sets can be used to model f, but not all can be used subsequently for PCA with the same ease.

The function f is estimated independently at each voxel time series by minimizing the penalized functional

$$\mathrm{SSPE}_\lambda = \sum_{n=1}^{N} [y_n - f(t(n))]^2 + \lambda \int_a^b [Lf(t)]^2 dt$$
$$= \mathrm{SSE} + \lambda \cdot \mathrm{PEN}(f). \quad (2)$$

This functional is the sum of two terms: the first term SSE is the ordinary sum of square residuals; the second term PEN(f) is a penalty regulating the amount of smoothness of the estimated function f, defined by the linear differential operator L. The positive smoothness parameter λ regulates the trade-off between minimizing the square error of the fit and the penalty term [for details on the statistical properties of this type of estimators, see Eubank, 1988].

By far the most common penalized functional, which was also adopted here, imposes a smoothness constraint by penalizing the square integral of the second derivative of the estimated function:

$$\mathrm{PEN}_2(f) = \int_a^b [f''(t)]^2 dt. \quad (3)$$

In the periodic case with repeated observations, and especially when B-splines are used, it is usually necessary to constrain the solution further by excluding functions not satisfying the condition

$$f^{(u)}(a) = f^{(u)}(b), \qquad u = 0, 1, \ldots l, \quad (4)$$

where $f^{(u)}$ is the uth derivative of the function f evaluated at the boundaries a, b of the function domain, and l is the degree of the spline polynomials. This condition specifies that the estimated function at a is the smooth continuation of the function at b. If this condition is not specified, the spline fit tends to be underconstrained in proximity of the boundaries, and the estimate to have high variance in those regions. Given that the BOLD signal generated by neural activity in the gray matter is small, this variance will inevitably dominate in the PCA. After de Boor [1978], satisfaction of this constraint was approximated in our implementation by duplicating the data beyond the two boundaries and fitting the B-spline on the enlarged interval.

A remarkable property of the penalized functional in equation (2) is that because we have only a finite number of observations, its unique minimizer can be shown to lie in a finite-dimensional functional subspace, even if its penalty term is specified more generally in a space of infinite dimensions [Poggio and Girosi, 1990; Schoenberg, 1964; Wahba, 1990]. More precisely, the penalized functional SSPE_λ may be expressed as a linear combination of only a finite number K of basis functions φ_k *where* K is the minimum number of basis functions that suffice to interpolate a set of P observations occurring at the points of $[a, b]$ when the signal was sampled, and as many linear coefficients c_k:

$$f(t) = \sum_{k=1}^{K} c_k \varphi_k(t). \quad (5)$$

Given this finite set of basis function, the penalized functional may be rewritten in vector notation

$$\mathrm{SSPE}_\lambda = \|\mathbf{y} - \mathbf{Fc}\|^2 + \lambda \mathbf{c}'\mathbf{Pc}, \quad (6)$$

where $\mathbf{c}$ is the vector of the coefficients of the basis expansion to be estimated, $\mathbf{F} = \{\varphi_k(t(n))\}$ is the $N \times K$ matrix of the basis functions evaluated at the time points of $\mathbf{t}$, and $\mathbf{P}$ is the $K \times K$ "penalty matrix":

$$\mathbf{P} = \left\{ \int_a^b L\varphi_q(t) L\varphi_r(t) dt \right\}, \quad q,r = 1, 2, \ldots K. \quad (7)$$

Differentiating SSPE_λ with respect to $\mathbf{c}$ and setting to zero as in ordinary regression gives the estimate of the coefficients

$$\mathbf{c} = (\mathbf{F}'\mathbf{F} + \lambda\mathbf{P})^{-1}\mathbf{F}'\mathbf{y}. \quad (8)$$

Quite efficient algorithms exist for inverting the matrix in equation (8), because the compact support of many basis functions sets (and notably that of B-splines) implies that $\mathbf{F}'\mathbf{F}$ is banded. For many basis function sets, the matrix $\mathbf{P}$ will also possess properties that facilitate its computation. In the general case, $\mathbf{P}$ may be computed by discretizing the integrals at sufficiently small intervals [Press et al., 1988]. For further details, see de Boor [1978], Green and Silverman [1994], or Ramsay and Silverman [1997].

Principal Component Analysis

Given a vector of interrelated random variables $\mathbf{x}_i$, $\Sigma_i \mathbf{x}_i = \mathbf{0}$, $i = 1, 2, \ldots, Z$, a step-by-step procedure to carry out ordinary PCA is described below. To obtain the first component, find the direction in space such that the projections

of the occurrences of the random vector $\mathbf{x}_i$ on it achieve maximal variation. Representing this direction as a vector of unitary length $\mathbf{g}$, we have

$$h_i = \langle \mathbf{x}_i, \mathbf{g} \rangle, \tag{9}$$

where we will choose $\mathbf{g}$ so as to maximize $Z^{-1}\Sigma_i h_i^2 = Z^{-1}\mathbf{g}'\mathbf{X}'\mathbf{X}\mathbf{g}$ subject to the constraint $\|\mathbf{g}\|^2 = 1$. The matrix $\mathbf{X}$ is formed by the random vectors $\mathbf{x}_i'$ taken row-wise. For the subsequent components, we repeat this step, subject to the additional constraint that each new component be orthogonal to all previous components. We cannot determine the components uniquely, because the same direction in space is represented by two vectors having the same elements but opposite signs. Standard linear algebra arguments show that these directions can be found as the solutions of the eigenequation

$$Z^{-1}\mathbf{X}'\mathbf{X}\mathbf{g} = \gamma\mathbf{g} \tag{10}$$

where the eigenvalue γ is the variation of the random vector projected on $\mathbf{g}$ [Anderson, 1984].

As mentioned above, one leading theme in functional data analysis exploits of the fact that functions defined on an interval $[a, b]$ form an inner product vector space [Kolmogorov and Fomin, 1968]. Multiplication by a scalar and addition of functions are defined as is usually the case for scalars and functions, and in the inner product between two functions f, g defined on $[a, b]$ summation is replaced by integration:

$$\langle f, g \rangle = \int_a^b f(x)g(x)dx, \tag{11}$$

implying that

$$\|f\|^2 = \langle f, f \rangle = \int_a^b [f(x)]^2 dx. \tag{12}$$

With these definitions in mind, it is easy to check that all properties defining a vector space are satisfied. The basis functions that are used to estimate the functional image are in fact the equivalent in functional space of a set of Euclidean coordinate vectors spanning a discrete estimation space.

To define functional PCA, we only need to replace the inner product and norm of vectors of equation (9) with their equivalent in functional space. In the fMRI setting, the realizations of the random variable are replaced by the image functions f_m that have been estimated from the voxel time series. The first principal component is therefore the eigenfunction g maximizing the score

$$h_m = \langle f_m, g \rangle = \int_a^b f_m(t)g(t)dt \tag{13}$$

over the image functions f_m, $m = 1, 2, \ldots, M$. The constraint on the length of the eigenfunction becomes

$$\|g\|^2 = \int_a^b [g(t)]^2 dt = 1 \tag{14}$$

and the orthogonality constraint on any two eigenfunctions g_q, g_r is expressed by

$$\langle g_q, g_r \rangle = \int_a^b g_q(t)g_r(t)dt = 0. \tag{15}$$

Like the directions found by its ordinary counterpart, functional PCA produces a set of functions that, when added and subtracted to the mean image function, indicate the modes of variation in the functional image. As in discrete PCA, it will be convenient to center the functions to avoid the first component mainly to represent average activation levels. The eigenfunctions then capture the covariance structure of a symmetric bivariate function ν, which may be defined in analogy with its discrete counterpart:

$$\nu(y,x) = M^{-1} \sum_{m=1}^{M} (f_m(y) - \bar{f})(f_m(x) - \bar{f}), \qquad x, y \in [a,b], \tag{16}$$

where $\bar{f} = M^{-1}\Sigma_{m=1}^{M} f_m$.

There are two approaches to carry out PCA in functional space. The older approach is to evaluate the functions on a fine grid, carry out ordinary PCA on the obtained values, and estimate the principal functions by fitting the basis set to the discrete components thus obtained [Rao, 1958]. More recently, however, it was shown that functional PCA can be carried out on finite matrices, finding the eigenvectors and eigenvalues of a positive definite matrix as in the discrete case. Again, this is because, being the outcome of the estimation process, the image functions in question are linear combinations of only a finite number of basis functions. Because the eigenfunctions belong to the same space spanned by the image functions, they will be expressed by a linear combination of the same basis set. Define the $K \times K$ matrix $\mathbf{U}$:

$$\mathbf{U} = \left\{ \int_a^b \varphi_q(t)\varphi_r(t)dt \right\} \qquad q, r = 1, 2, \ldots K. \tag{17}$$

This matrix is a particular case of the matrix **P** in equation (7), and can be computed in the same way. In particular, note that for an orthonormal basis set, $\mathbf{U} = \mathbf{I}$. Then it is possible to express the eigenequation of principal component analysis in terms of discrete matrices:

$$M^{-1}\tilde{\mathbf{C}}'\tilde{\mathbf{C}}\mathbf{U}\mathbf{g} = \gamma\mathbf{g} \tag{18}$$

where $\tilde{\mathbf{C}}$ is the $M \times K$ matrix of the coefficients of the basis expansion of the centered image functions. Solving equation (18) for γ and **g** returns the eigenvalues and the coefficients of the eigenfunctions, respectively. For details, see Ramsay and Silverman [1997].

Once the eigenfunctions have been determined, the principal component scores of all image functions give a set of parametric images, providing a graphic representation of the areas of the brain that load on each eigenfunction. For an eigenfunction g,

$$h_m = \langle f_m, g\rangle = \tilde{\mathbf{c}}'_m\mathbf{U}\mathbf{g}, \quad m = 1, 2, \ldots M, \tag{19}$$

where the tilde indicates the centered function coefficients. At most min(M, K) eigenfunctions and functional component score images can be computed, but usually only the first few are of interest, as is the case with data contaminated by noise.

Selection of the Smoothness Parameter

The specification of a smoothness constraint is crucial for the success of the method, and is justified by the notion that the signal recorded by the scanner is contaminated by noise. Another reason to impose a smoothness constraint in nonperiodic settings is that if the estimated function is allowed to interpolate the observations exactly, there will be little difference between functional and ordinary PCA. It is also clear, however, that the amount of imposed smoothness cannot be arbitrarily high. If the estimated functions are excessively constrained, details of the signal of interest or even the signal itself may not be captured by the fit. For this reason, it is important to make a reasonably good choice of the smoothness parameter λ. Usually, however, we have no direct information to assist us in choosing its value. In the following, we will consider three approaches for the selection of the smoothness parameter. In so doing, we will also clarify the theoretical justifications for preferring the functional to the discrete approach in estimating the structure of the covariance of the data, and characterize the advantages of functional PCA with respect of other explorative approaches.

Uniform setting of the smoothness parameter

The first approach is to use visual judgment, trying different values of λ until one is found that seems to reflect the course of signal well, and apply it uniformly to the whole functional image. This approach should not be dismissed out of hand on the grounds of its subjective character. In many cases, it produces entirely acceptable results. Furthermore, its analysis leads to a more precise formulation of the way in which our smoothness assumptions bear on the outcome of the PCA. Theoretically at least, this approach is therefore of fundamental importance.

We will consider the case of interpolating splines, when the matrix **F** is full rank and hence invertible, and the data have already been centered. For a given value of λ, be $\mathbf{H}_\lambda$ the hat matrix from the penalized regression of equation (8):

$$\mathbf{H}_\lambda = \mathbf{F}(\mathbf{F}'\mathbf{F} + \lambda\mathbf{P})^{-1}\mathbf{F}'. \tag{20}$$

This matrix defines the estimation space of the regression; in fact, it is a linear operator that when applied to the predictor transforms it into the estimate lying in the subspace spanned by the columns of **F**. Clearly, the subspace of $\mathbf{H}_\lambda$ could be spanned by another basis, and we will express it in terms of the coefficients of its own eigenfunctions [Hastie et al., 2001]. Being positive definite, $\mathbf{H}_\lambda$ has a positive real eigendecomposition:

$$\mathbf{H}_\lambda = \mathbf{Q}\mathbf{W}\mathbf{Q}', \tag{21}$$

where the columns of **Q** define the eigenfunctions of $\mathbf{H}_\lambda$, and **W** is a diagonal matrix containing the eigenvalues. The columns of **Q** appropriately scaled according to their eigenvalues constitute the alternative "Demmler-Reinsch" basis for the estimation space [Demmler and Reinsch, 1975]. It is important here to mention three things about this basis set. First, whereas the curves of a usual basis set have norm one but are not necessarily orthogonal, the curves of the Demmler-Reinsch basis are orthogonal, but their norm is equal to the respective eigenvalue. Second, when the penalty is PEN_2, the eigenvalues order the eigenvectors according to their smoothness, i.e., smoother curves have larger eigenvalues. Expressed in this form, splines show their affinity with Fourier series, because they are orthogonal and indexed according to the "number of oscillations" [Eubank, 1988]. Third, although the relative order of eigenvalues and eigenvectors remain unaltered, changes of λ affect the absolute size of the eigenvalues. In fact, it can be shown that $\mathbf{W} = (\mathbf{I} + \lambda\mathbf{B})^{-1}$, where **B** is the diagonal matrix of the eigendecomposition $(\mathbf{F}\mathbf{P}^{-1}\mathbf{F}')^{-1} = \mathbf{Q}\mathbf{B}\mathbf{Q}'$ (see the Appendix for details).

Rewriting equation (20) in terms of **Q**, λ, and **B**, we have

$$\mathbf{H}_\lambda = \mathbf{Q}(\mathbf{Q}'\mathbf{Q} + \lambda\mathbf{B})^{-1}\mathbf{Q}', \tag{22}$$

a ridge regression [Draper and Smith, 1998; Hoerl and Kennard, 1970] in which the estimate is shrunk differentially according to the value of the diagonal elements of **B**. Remembering that these latter are also an index of the number of oscillations of the corresponding eigenfunction, we see that the penalty term acts as a form of shrinkage toward smoother curves, rather than toward smaller coefficients as in ordinary ridge regression. The amount of shrinkage is determined by λ through its indirect effect on the size of the

trace of **W**. Note that those functions that satisfy $Lf = 0$ in equation (2), and therefore incur no penalty, are not shrunk at all.

The connection with ridge regression has several important implications.

First, as in ridge regression, the direction of shrinkage is interpretable as a Bayesian prior on the distribution of the estimate [Kimeldorf and Wahba, 1970]. The differential operator L therefore specifies a prior belief on the result of the estimate, which favors smooth function estimates in the sense specified by the penalty term. This prior is appropriate for the recovery of a hemodynamic response function, because typically data in an fMRI experiment are collected at higher frequencies than are those of the reconstructed BOLD response [Turner et al., 1997].

Second, the estimate of the covariance resulting from the penalized fits differs from the maximum likelihood estimate in that it too has been subjected to shrinkage toward lower frequencies. Lower-frequency signals will therefore tend to be overrepresented in the first components of the eigendecomposition, even if they explain proportionally less variance in the original data.

Third, because of the bias toward lower frequencies, the application of a high-pass filter may be used to investigate the presence of specific signals in the data, because the roughness penalty will preferentially extract the frequencies immediately below the filter threshold and place it among the first components (if these frequencies have enough support in the data). The standard approach of applying a high-pass filter with a threshold at twice the period of the experimental paradigm to remove signals of no interest such as a low frequency drift before analysis [Turner et al., 1997] is thus all that is required to investigate the presence of a signal of interest generated by the alternation of the experimental conditions.

This is in contrast with other explorative techniques, which generate a number of "components" with little information about how to select from them. In the case of clustering [Wismüller et al., 2002], for example, the number of voxels assigned to a time-course cluster according to a nearest-neighbor scheme cannot be used to select the signal of interest, because there is generally no reason to assume that the signal of interest is present in any but a small number of voxels. Similarly, in ICA [McKeown et al., 1998], it is unclear how the component that represents the signal of interest should be selected.

Automatic selection by generalized cross-validation

In the second approach, an attempt is made to assign the value of λ automatically from the data themselves using generalized cross-validation [for a justification of the procedure, see Green and Silverman, 1994; Wahba, 1990]. The aim is to set a value of λ that reflects an estimate of what in the data is signal, discarding higher-frequency variance. Importantly, this approach allows setting λ to a different value in each voxel series.

The generalized cross-validation score is given by

$$\mathrm{GCV}_\lambda = N \frac{\|(\mathbf{I} - \mathbf{H}_\lambda)\mathbf{y}\|^2}{[\mathrm{trace}(\mathbf{I} - \mathbf{H}_\lambda)]^2} = N \frac{\mathrm{SSPE}_\lambda}{(\mathrm{df\ residuals})^2}, \quad (23)$$

where trace(·) is the matrix trace. An intuitive understanding of how this score works can perhaps be gained if it is viewed as a corrected ratio between the sum of square errors, and the square of a value representing the degrees of freedom of the subspace of the residuals. The optimal λ is given by the minimum cross-validation score, by which the least square errors sum is obtained with the largest number of degrees of freedom in error space. Unlike direct cross-validation, GCV_λ does not require resampling the data for its computation. The score is calculated from a series of fits on a range of λ values, after which the value of λ may be chosen according to the lowest score. For more exact determination of λ, this procedure may be iterated restricting the range of λ to the lowest two scores.

An interesting aspect of the general cross-validation score is that the related ratio $\mathrm{SSPE}_\lambda/\mathrm{trace}(\mathbf{I} - \mathbf{H}_\lambda)$ for the selected λ may be interpreted as an estimate of the variance of the error term ϵ of our model, under the assumption that the signal is smooth [Green and Silverman, 1994]. In fact, under this assumption, generalized cross-validation represents a theoretically motivated attempt to separate genuine signal from noise [Carew et al., 2003; Craven and Wahba, 1979]. Because the smoothed estimates of the activation functions translate into a smooth covariance matrix, generalized cross-validation eventually determines what part of the lower frequencies should be retained in the covariance matrix to be later recovered by the eigendecomposition, and what part should be discarded.

The justification for estimating a separate smoothing parameter for each voxel series is that we do not expect the proportion of noise and genuine signal to be uniform over the fMRI image. The variance of the voxels of the ventricles is usually larger, and through spatial smoothing or imprecise segmentation, some of this variance may spill over to the gray matter. Equally important is variance originating from aliased signal, whose spatial distribution is also unlikely to be uniform. In keeping with this expectation, we observed that both the GCV_λ minimum and the related estimate of the error variance vary across the image.

One possibly problematic aspect of generalized cross-validation is that, even when calculated from spatially smoothed data, the GCV_λ minimum does not always vary smoothly across adjacent voxels. This might be because, as Wahba [1990] notes, the theory justifying generalized cross-validation is an asymptotic one, and therefore departures from a good estimate of the smoothness parameter may occur with a finite number of observations. Wahba [1990] reported that even when the observations are in the hundreds, two or three obviously wrong estimates are produced every thousand repetitions of the sampling. Because in the fMRI setting the number of the repeated applications of generalized cross-validation is in the order of tens of thousands, we should expect the occurrence of quite a few wrong smoothness parameters in the estimate. Nevertheless, the

results presented below demonstrate the practical effectiveness of generalized cross-validation when it is desirable not to impose explicit assumptions on the range of frequencies of the signal of interest. Confidence in the effectiveness of generalized cross-validation is also bolstered by the results of Carew et al. [2003], who successfully applied generalized cross-validation to fMRI datasets with the purpose of correcting for the autocorrelation of the signal.

A practical problem of generalized cross-validation is that it can require rather lengthy computations. For the interactive exploration of the data, an acceptable procedure might consist of carrying out generalized cross-validation once, taking note of value of λ for the estimated functions with the highest score for the component of interest, and then applying this value uniformly to other similar datasets, relying on the shrinkage effect of the penalized regression to emphasize the signal of the corresponding frequency across the whole dataset.

Cross-validation has been used here on the data to calculate a different smoothing parameter for each image function, but there are other interesting approaches that need to be explored. One possibility is that of fitting the data with an interpolating function, incorporate the penalty in the PCA, and cross-validate the result at this final stage [Ramsay and Silverman, 1997]. Another approach is that of estimating the smoothness parameter separately for each component [Rice and Silverman, 1991].

Linear smoothers from restricted basis expansions

When the recording sessions are long and nonperiodic predictor vectors are specified, generalized cross-validation becomes impractical, and indeed even just fitting an interpolating *B*-spline to the whole time series may require memory resources that exceed the capability of the computer. In this case, it is possible to regularize the functional image by specifying a smaller basis function set, i.e., one in which $K < P$. Because the knots of the *B*-splines are spaced regularly, the remaining degrees of freedom are distributed evenly across the function domain, resulting in a uniformly smoother estimate. It is then not necessary to specify any penalty term, and the fitting process minimizes the ordinary sum of error squares:

$$\mathrm{SSE} = \|\mathbf{y} - \mathbf{F}\mathbf{c}\|^2, \tag{24}$$

leading to the familiar estimate

$$\mathbf{c} = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{y}. \tag{25}$$

The hat matrix is in this case an ordinary projection matrix, and there is no shrinkage of the estimated coefficients. Clearly, for low enough values of *K*, the fitted function will not be able to capture the signal of interest even when this latter is present. The advantage of this approach is that its computation is fast, and that of the PCA that follows even more so. If a Fourier basis set is used, it is easy to bind the choice of the number of basis functions to an a priori specification of the dominant frequency of the signal of interest.

RESULTS

First Case Study: Working Memory

In the first case study, functional PCA was applied to realigned, stereotactically normalized, smoothed, and high-pass filtered fMRI from a working memory block design ("two-back") [Cohen et al., 1994]. After Turner et al. [1997], the filter threshold was set to twice the period of the block. In the first half of the block, subjects were required to press a button indicating if the presented stimulus was identical to a predefined letter. In the second half of the block, the same decision was taken regarding the sameness of the present stimulus (a letter again) and the stimulus two presentations earlier. The image functions were estimated with *B*-splines using a periodic scheme for the generation of the predictors, and periodic constraints on the estimated image functions. The number of base functions was the smallest to interpolate the data. The amount of smoothing was determined through generalized cross-validation.

Figure 1 displays the results of the functional eigenanalysis (top), compared to those of the ordinary version of PCA (bottom). The data on which ordinary PCA was carried out were subjected to the same preprocessing (including application of a high-pass filter) as that in the functional version of the algorithm. For a fair comparison, in the calculation of the ordinary version of PCA, the raw data were also folded around the period of the block and averaged before being used to form the discrete variance matrix (for details on the procedure applied to carry out ordinary PCA, see the Appendix). The covariance function of the fit (equation [16]) is displayed by drawing the values taken by the function in shades of gray (top left). Here, the two phases of the block of the experimental paradigm are clearly visible as a pattern of rough squares of higher covariance, shifted to the right and downward because of the BOLD delay. These two areas of higher covariance arise because, in the centered functions from the voxels affected by the paradigm, the off and on phases of the block create correlated deviations relative to the average zero signal. Note that the main diagonal has two drops in intensity corresponding to the point where the signal inverts at the boundaries of the phases, when it crosses zero. For comparison, the discrete covariance matrix of the centered and filtered data is displayed at the bottom left. It is very difficult here to discern the pattern of variance/covariance induced by the paradigm. In the discrete covariance matrix, there are high-frequency signals that obscure the biphasic structure of the experimental block. On the diagonal, four data points give rise to outliers, in contrast to the much more regular distribution of the variance in the functional counterpart.

The scree plots of the eigenanalyses, displayed under the covariance structures, indicate the existence of a very clear first component in the functional version (top), whereas the discrete PCA produces at least two components that might

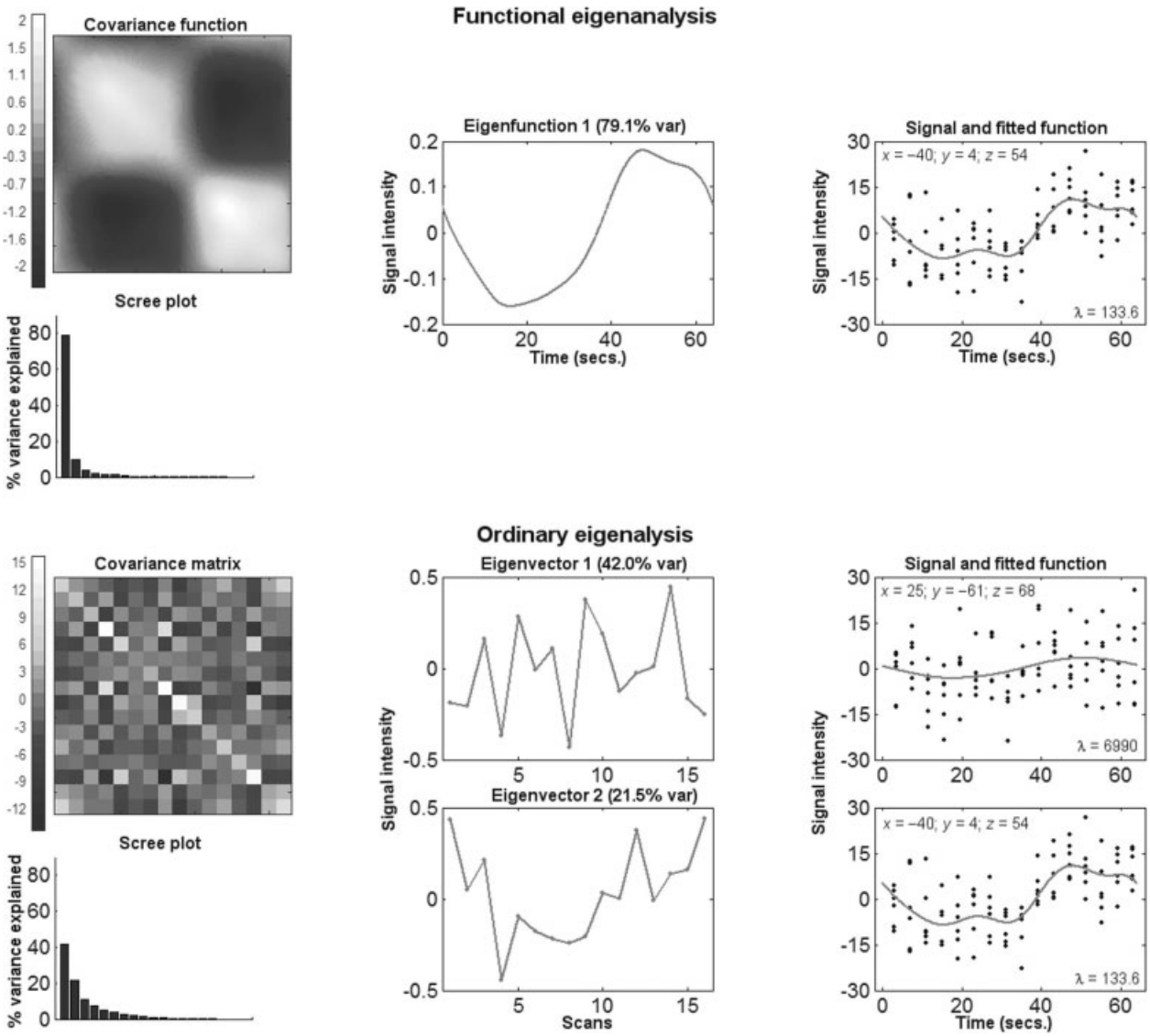

Figure 1.

Top: Functional eigenanalysis of the data from the working memory paradigm. **Bottom**: Ordinary principal component analysis of the same dataset. **Left**: Covariance function and matrices from which the eigenanalyses are carried out, below which the respective scree plots are displayed. **Top center**: The first eigenfunction for the memory-encoding paradigm. For display, the direction was chosen that makes the coefficients positive relative to the expected pattern of activation. **Bottom center**: The first and second eigenvectors, displayed using the same convention. **Top right**: Signal (black dots) and fitted function (gray) from the voxel with the highest first functional component score, located in BA6. Here and in the following plots of the signal and the fitted function, the Talairach coordinates of the voxel are displayed in the upper left corner of the plot. When calculated by generalized cross-validation and varying in each voxel, the smoothing coefficient λ of the fit is displayed on the right lower corner of the plot. **Bottom right**: Signal and fitted functions from the voxels with the highest scores on the first and second components from ordinary PCA; the second voxel is here the same as that selected through the first functional component. In this and all subsequent case studies, the data were subjected to the same preprocessing steps before application of both methods. The block design had a periodicity of 64 s. A high-pass filter of 128 s was applied to the series of 96 scans acquired with an interscan interval of 4 s, and subsequently smoothed with a Gaussian kernel of 8 mm. A mask selecting gray matter was applied to select voxels of interest.

refer to signal in the data (bottom). Figure 1 also displays the first eigenfunction and the first two eigenvectors resulting from the respective eigenanalyses. The first eigenfunction shows deviations from the average signal corresponding to the two phases of the reconstructed BOLD response function, delayed of about 5 s relative to the boundaries of the block (top center). This eigenfunction captures the pattern of variation clearly visible in the covariance function. In contrast, the first eigenvector from the ordinary PCA shows little relation to the expected BOLD response; only the second eigenvector presents some resemblance to the expected signal (bottom center).

On the right of Figure 1 the signal and fitted functions from the voxels with the highest PCA are displayed. Because they correspond to the maximum of the component scores, these voxels contain the highest variation in the direction detected by the respective principal component. The course of the signal in these voxels therefore has the highest correlation with the respective eigenfunction or eigenvector. This means that the course of the signal deviates from zero and has the same overall shape as the respective eigenfunction or eigenvector.

In functional PCA (top), the eigenanalysis is carried out on all fitted functions. As a result, the displayed fitted function (rather than the raw signal) has the highest correlation with the eigenfunction in the voxel with the highest component score. The extent to which the fitted function follows the variation of the raw signal (black dots in the plot on the right) depends on the smoothing coefficient, which is determined here automatically in each voxel using generalized cross-validation. If the smoothing coefficient is zero, the fit attempts to follow the signal as closely as possible; if it is very large, the fit approximates a straight line irrespective of the variation of the signal over time. The smoothing coefficient found with generalized cross-validation is small enough for the fit to reproduce most of the variation in the signal. In this voxel, therefore, the eigenfunction reproduces the fitted function and the course of the signal well.

In case of ordinary PCA (bottom), the eigenanalysis takes place over the averaged raw signal. As a result, the average row signal (rather than the fitted function) has the highest correlation with the respective eigenvector in the voxel with the highest component score. The fitted function here fails to reproduce the eigenvector or the fitted signal well. This is due to the quite large smoothing coefficient selected by generalized cross-validation. The voxel with the highest score on the second ordinary principal component is the same as the voxel with the highest score on the first functional principal component, where, as we have seen, the penalized fit follows the signal quite faithfully. The second ordinary eigenvector does not, however, and contains much high-frequency variation that is not prominent in the raw signal or the fitted function. The first functional component and the second ordinary component locate the peak of activation in Brodmann's area (BA) 6 [Talairach and Tournoux 1988], a part of prefrontal cortex involved in working memory. We could not give a functional interpretation of the activation located by the first ordinary component.

Functional PCA gave good results also when applied to the same data without folding them around the period of the block (Fig. 2). The task became more difficult because no knowledge about the phase and number of the blocks was used. As before, the functions were fitted using *B*-splines with the amount of smoothing determined by generalized cross-validation. The number of basis functions was chosen to enable the interpolation of the data. The covariance function displays a more complex patterned structure corresponding to the six blocks of the experimental design (top left), which are clearly visible in the first eigenfunction (top center). This pattern is much more difficult to recognize in the maximum likelihood estimate of the covariance matrix of the raw data (bottom left), where some single data points look like outliers. Consequently, it is difficult to identify the activation induced by the blocks in the eigenvectors produced by ordinary PCA (bottom center). In contrast, the blocks are clearly retrieved by the first functional component, which comes close to replicating the experimental regressor produced by the SPM package [Friston et al., 1995b]. The voxel with the highest functional component score (top right) is situated as in the previous analysis in BA6. As in the previous analysis with folded data in Figure 1, one can see that raw signal, fitted function, and eigenfunction in the voxel with the highest first component score are in good agreement with each other and with the experimental regressor (top center and right). We therefore conclude that the variation identified by the first functional component well represents the systematic variation that must have been introduced by the alternation of the conditions of the experimental setting (the blocks of the paradigm). By contrast, the picture resulting from the comparison of raw signal, fitted function, and eigenvectors resulting from the first two components of ordinary PCA (bottom center and right) is much more difficult to interpret. Even if not immediately apparent at the naked eye, there is necessarily some correlation between raw signal and the respective eigenvector. Generalized cross-validation discarded all variations in these voxels, however, so that the fitted function is a flat line (i.e., there is no signal left). This seems appropriate for our purposes, because the course of the signal in these voxels seems to have little in common with the experimental regressor.

Figure 3 displays the comparison between the "beta image" produced by the experimental regressor using the SPM package (top row) and the images composed of the first functional component scores (second and third rows from top). The beta image contains the coefficients of a regressor obtained by convoluting a boxcar function representing the conditions of the block with a standard hemodynamic function (Fig. 2, center). The beta image from the experimental regressor and the functional component score images are very similar to each other. Moreover, the transverse slices demonstrate parts of the frontal-parietal network usually observed during two-back tasks in fMRI [Cohen et al., 1994;

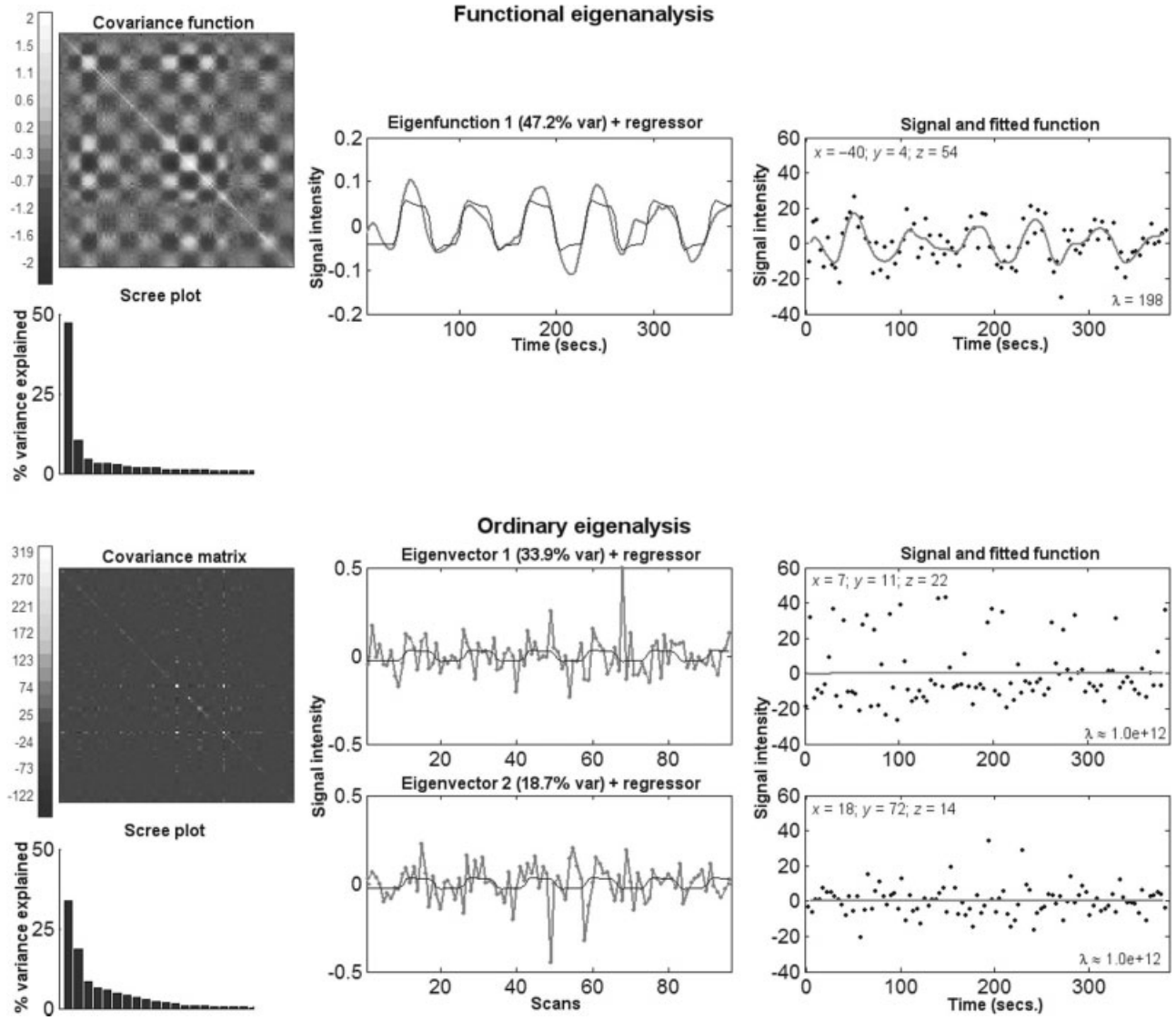

Figure 2.

Eigenanalysis of the working memory data of Figure 1 without folding around the period of the block. As in Figure 1, the covariance function and matrix and the scree plots from the respective eigenanalyses are on the **left**, the first eigenfunction and the first two eigenvectors (gray) are in the **center**. Eigenfunction and eigenvectors have been displayed with the experimental regressor created by SPM by convolving the blocks of the paradigm with a canonical BOLD response (black). The signal and fitted function charts on the right are chosen according to the highest component score of the respective PCA. The voxel individuated by functional PCA is the same as that in the analysis with folded data. The two voxels from ordinary PCA are situated in the frontal lobe. In both cases the smoothing coefficient is the highest allowed by the implementation.

Walter et al., 2003a,b] and constituting parts of the phonological loop as described by Baddeley [1992]. Activation in BA6 is particularly prominent here, with a lack of diffused activation in the functional component score image from the data that were not folded around the period of the block (third row from top). This is a consequence of generalized cross-validation, which discarded most variance in these voxels. Ordinary PCA (not folded, bottom row) was unable to demonstrate relevant foci of activation. Folded ordinary PCA did (fourth row); however, it was necessary to select the second component after visual inspection.

Second Case Study: Episodic Memory Encoding

The second case study evaluates data from a study in which functional PCA was carried out on fMRI data from a block design for a memory encoding paradigm [Grön et al., 2001]. In the first part of the block, two abstract geometric

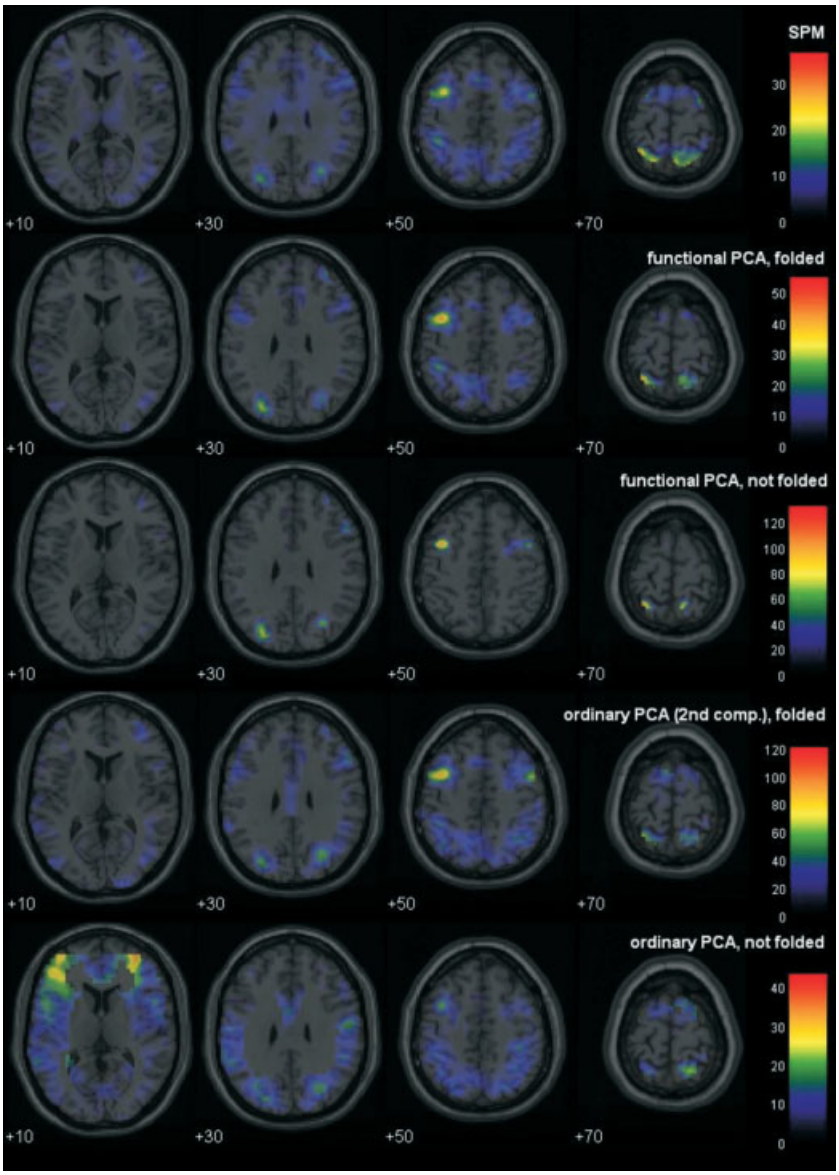

Figure 3.
Working memory paradigm. **Top row:** Slices from the beta image from the regression obtained with the SPM package. **Second row:** Slices from the first functional principal component scores image using a periodic scheme for the generation of the predictors after folding the data around the period of the experimental block. **Third row**: slices from the first functional principal component scores image using a nonperiodic scheme for the generation of the predictors. **Fourth row:** Second principal component scores image from an ordinary PCA on the data folded around the period of the block. **Bottom row:** First principal component scores image from an ordinary PCA on the not folded data. The component score images derive from the component direction that best matches the beta image of the SPM linear model.

patterns were presented, each for 6 s. In the second part of the block, the subject viewed red concentric circles for 20 s. The subject was required to intentionally memorize the stimuli for later recall. Each block was presented five times. As in the first case study, the data were first realigned, stereotactically normalized, smoothed, and high-pass filtered. As in the previous case study, the image functions were estimated with *B*-splines using a periodic scheme for the generation of the predictors and periodic constraints on the estimated image functions after folding the data around the period of the block. Enough base functions were used to allow interpolation of the sampling points, whereas the smoothness parameter was estimated by generalized cross-validation (Fig. 4).

In this case study, we find again that the covariance function from the smoothed fit reveals the biphasic structure of the block more clearly than does the covariance matrix of the raw data. Unlike the previous case, the first eigenvector from ordinary PCA displays a good resemblance to the expected activation, although it tends to detect the variance in the middle of the block to the detriment of the boundaries. In contrast, variance at the boundaries is captured by the second eigenvector. As a result, the eigenfunction is again a much better reconstruction of the signal in the area of interest (top right). The comparison of the component score images reveals that this difference is enough to lead to some minor mismatch in the detection of the activated areas.

To explore the robustness of the method, we changed the basis set to carry out the functional PCA on the same data without folding them around the period of the block. We used a nonperiodic scheme to generate the predictors and a Fourier series as basis function set. The use of a Fourier series implies the periodicity of the function modeled but no

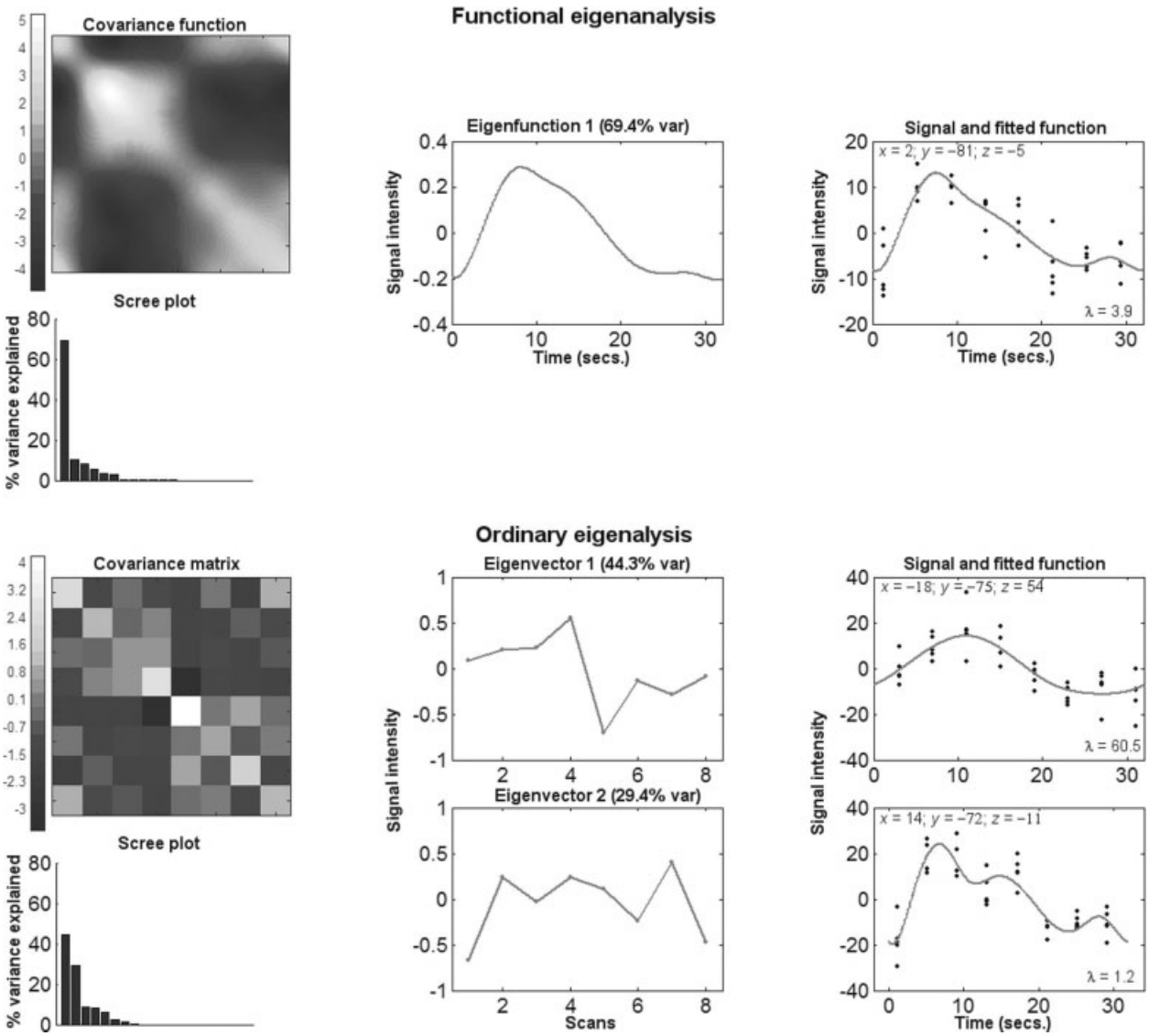

Figure 4.

Top: Functional eigenanalysis of the data from the episodic memory paradigm. **Bottom**: Ordinary principal component analysis of the same dataset. For a detailed explanation, see the legend to Figure 1. The signal and fitted function relative to functional PCA is situated in BA18. The voxels for ordinary PCA are those with the highest scores on the respective eigenvector. The block design had a periodicity of 32 s. A high-pass filter of 64 s was applied to the series of 40 scans acquired with an interscan interval of 4 s, and subsequently smoothed with a Gaussian kernel of 8 mm. A mask selecting gray matter was applied to select voxels of interest.

knowledge of the phase of the block. Unlike the previous analyses, we used a fixed smoothness parameter to obtain a quick computation of the functional fits (Fig. 5).

As in the working memory case study, the structure of the block emerges more clearly from the functional version of the eigenanalysis. The first eigenfunction replicates closely the experimental regressor (top center). Although dominated by an outlier, the first eigenvector manages to capture some of the variation induced by experimental paradigm. In contrast, the second eigenvector retrieves variance in voxels where it seems to be constituted mostly of noise (bottom right).

The analysis of the component score images (Fig. 6) reveals that although activation of the functional PCA corresponds closely to that retrieved by SPM, ordinary PCA leads sometimes to incorrectly active areas and sometimes misses important activations. SPM analysis (top) demonstrates a subset of clusters of active voxels in regions expected to be recruited during episodic learning of visual material [Desgranges et al., 1998; Gabrieli, 1998]: lingual gyrus, inferior occipital gyrus, inferior and superior parietal lobule, and anterior cingulated gyrus. The same clusters of activation were identified by functional PCA (folded and not folded;

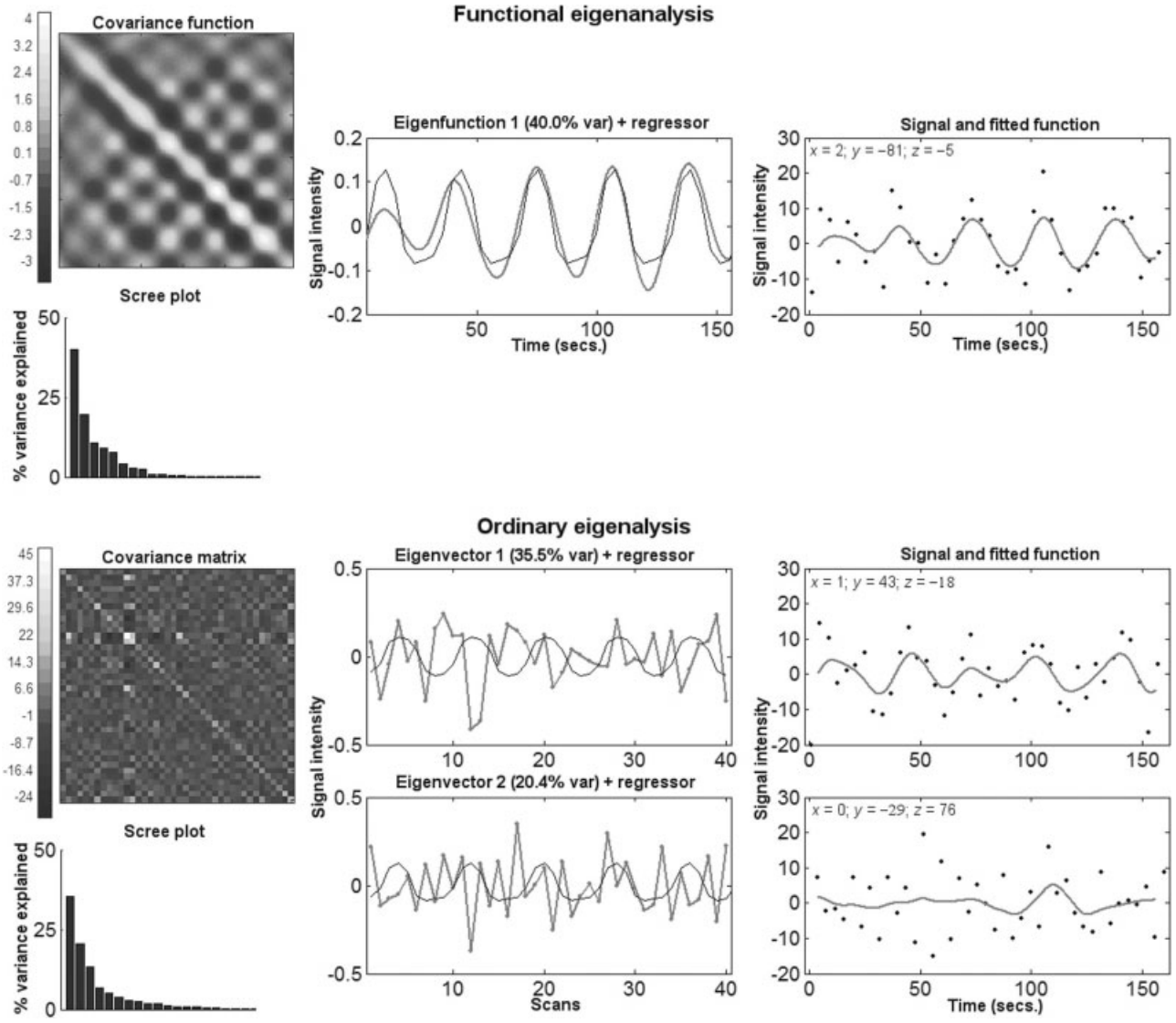

Figure 5.

Eigenanalysis of the episodic memory data without folding around the period of the block, and using a Fourier basis set with a fixed smoothness coefficient of 120. The signal and fitted function for functional PCA is situated in the same position as in the preceding figure. The voxels for traditional PCA are those with the highest scores on the respective eigenvector. For an explanation of the figure, see the legend to Figure 1.

second and third row). In contrast, ordinary PCA (folded, fourth row) fails to locate the lingual gyrus activation. Not folded ordinary PCA (bottom row) demonstrates various additional regions that were not expected to be relevant in episodic learning, especially insular cortex bilaterally and residual posterior white matter at the level $z = 35$.

Third Case Study: Finger Tapping

To investigate the effectiveness of functional PCA in a different setting, we applied it to an event-related design in which the subject was required to press a button with her right index finger upon presentation of a visual stimulus. Because in this experiment the stimuli do not occur regularly, a periodic scheme for the regressor with data folding around the period of the stimulus is not applicable. In this case study, instead of comparing folded and not folded data we investigate the role of generalized cross-validation by comparing it to the setting in which smoothing is carried out by reducing the number of the bases and applying a predefined smoothness coefficient.

The signal is strong enough to lead to its individuation in both the functional and ordinary versions of PCA. The slight superior performance of functional PCA with generalized cross-validation, however, can still be appreciated in the

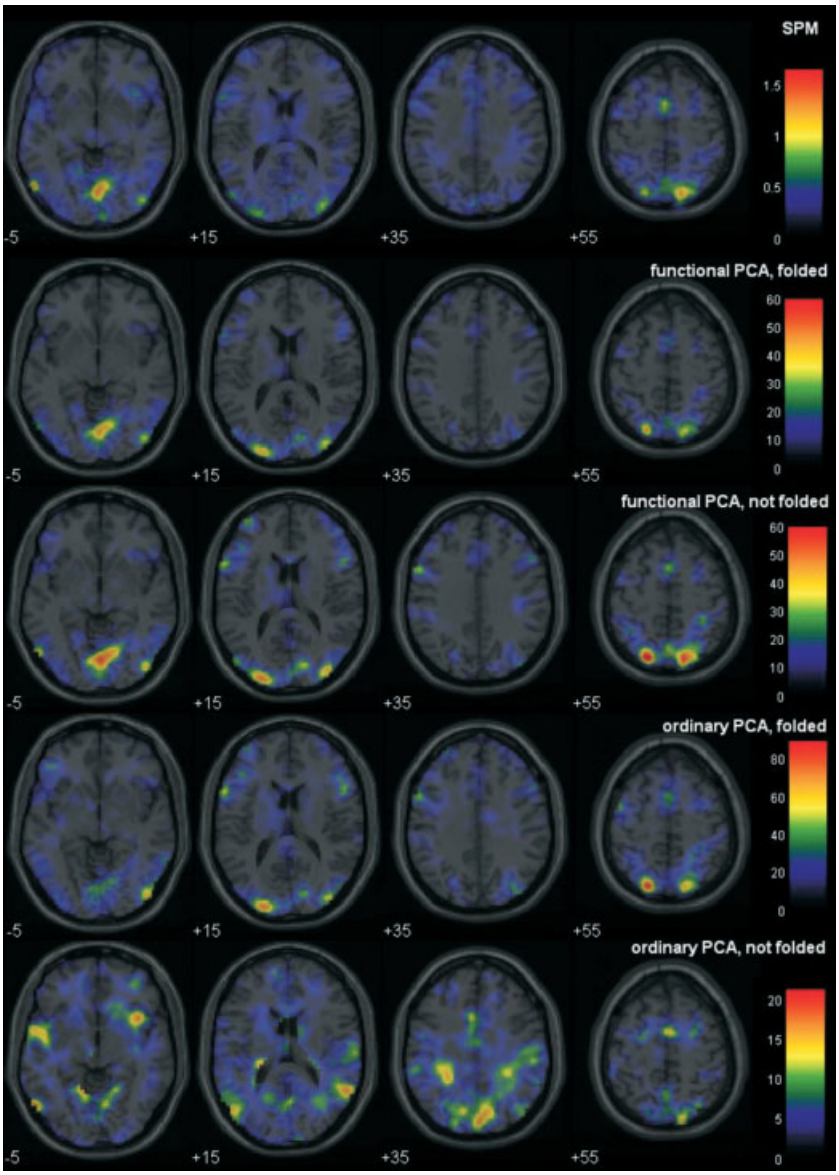

Figure 6.
Episodic memory paradigm. **Top row**: Slices from the contrast image from the regression obtained with the SPM package. **Second row**: Slices from the first functional principal component scores image using a periodic scheme for the generation of the predictors and folding of the data. **Third row**: Slices from the first functional principal component scores image using a nonperiodic scheme for generation of the predictors. **Fourth row** and **bottom row**: Slices from the first ordinary component scores images of the equivalent traditional PCA eigenanalyses.

shape of the scree plot. Although ordinary PCA retrieves several components that seem to explain some portion of variance, generalized cross-validation eliminates any plausible candidate component after the first (Fig. 7). This is particularly remarkable in the case of the second component of ordinary PCA, where all variance has been discarded by generalized cross-validation.

For a large dataset like this one, computation of the generalized cross-validation smoothing parameter for a fit with interpolating splines is computationally intensive (about 3.5 hr on a Pentium PC). In contrast, fitting the same data using *B*-splines and half the number of bases that would have been required to interpolate the data took about 3.5 min. Although the covariance function here includes variance that partially obscures the pattern of the experimental events, the first eigenfunction still displays a very good similarity to the expected BOLD response (Fig. 8).

Figure 9 shows the comparison between the beta image obtained with the SPM package and the component score images obtained with different methods. For the tapping task, the reference analysis (top) applying the general linear model demonstrates a network of motor and premotor areas (contralateral precentral gyrus, left lateral middle frontal gyrus, and dorsal superior frontal gyrus; BA6). This pattern could be repeated with minor variations in all forms of PCA (second to bottom rows). Although all images display fairly concordant results, the areas of activation in the component score image deriving from the application of generalized cross-validation are isolated, instead of being surrounded by a penumbra of weakly activated voxels. These weakly acti-

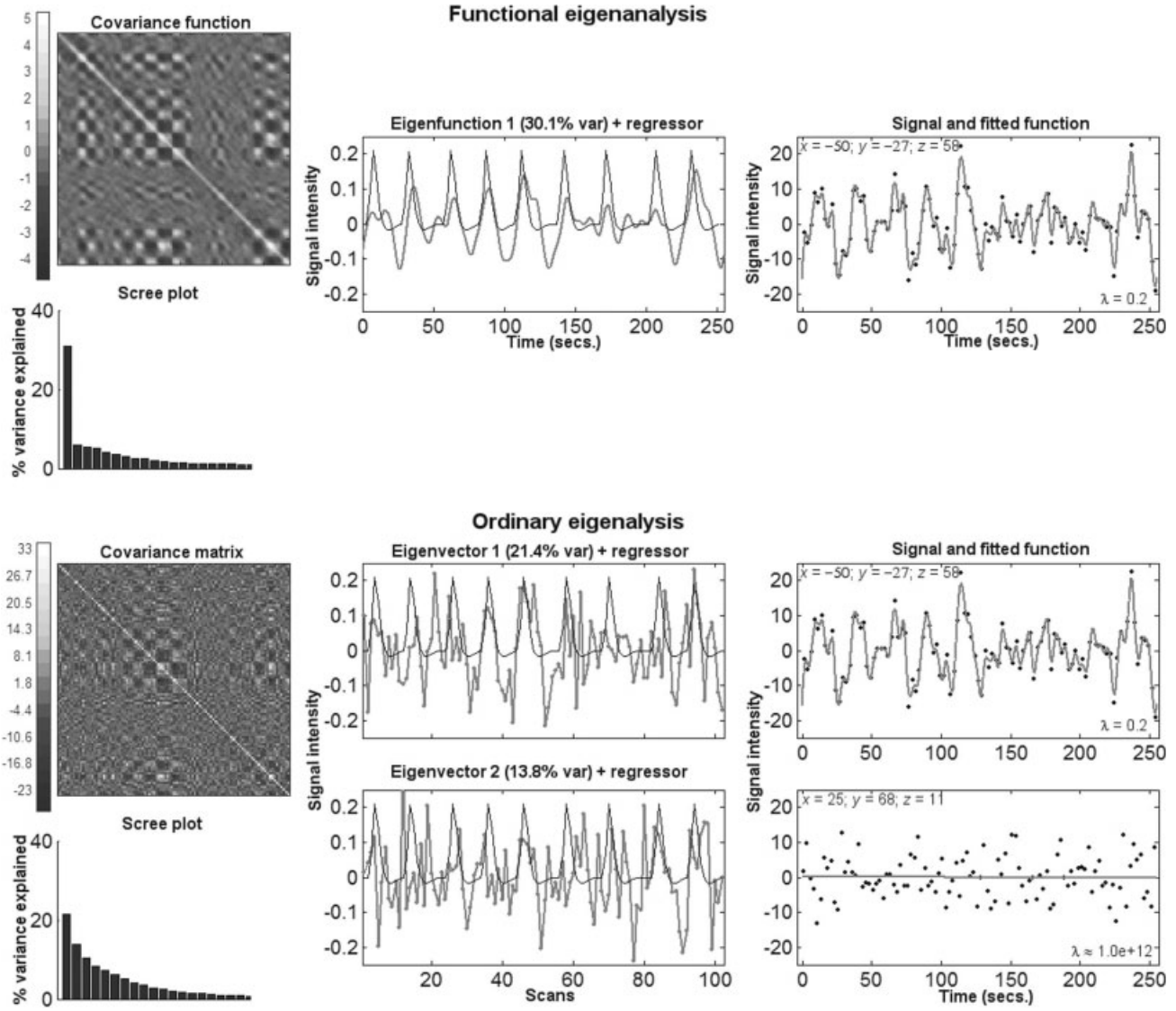

Figure 7.

Top: Functional eigenanalysis of the data from the finger-tapping paradigm. **Bottom**: Ordinary principal component analysis of the same dataset. For detailed explanation, see the legend to Figure 1. The signal and fitted function for functional PCA and the first principal component of ordinary PCA are taken from the same voxel in the precentral gyrus. The signal and fitted function for the second ordinary principal component is from the voxel with the highest score, The smoothing coefficient was the largest allowed by the implementation. A high-pass filter of 60 s was applied to the series of 102 scans acquired with an interscan interval of 2.49 s, and subsequently smoothed with a Gaussian kernel of 8 mm. A mask selecting gray matter was applied to select voxels of interest.

vated voxels are apparent especially in the beta image from SPM and in the component score image from ordinary PCA.

The elimination of diffuse activation in the functional component images is an effect of smoothing, which discards variance at high frequencies, and is most pronounced when generalized cross-validation is used. We have noted above (see comments regarding Fig. 1, 2, 7) that the fitted functions in the voxels with maximum component scores in the traditional PCA are so smooth that most or all variance is discarded by the penalized fit. This is actually a more general phenomenon induced by generalized cross-validation in these datasets, affecting not only the voxels with a high ordinary component score, but also most voxels outside the areas loading on the main functional component. Figure 10 displays a comparison of selected slices of the component scores images where the areas of activation are prominent, and corresponding slices in images composed of the smoothing coefficient estimated by generalized cross-validation. The high values of the smoothing coefficient in the areas where there is no activation in the component score

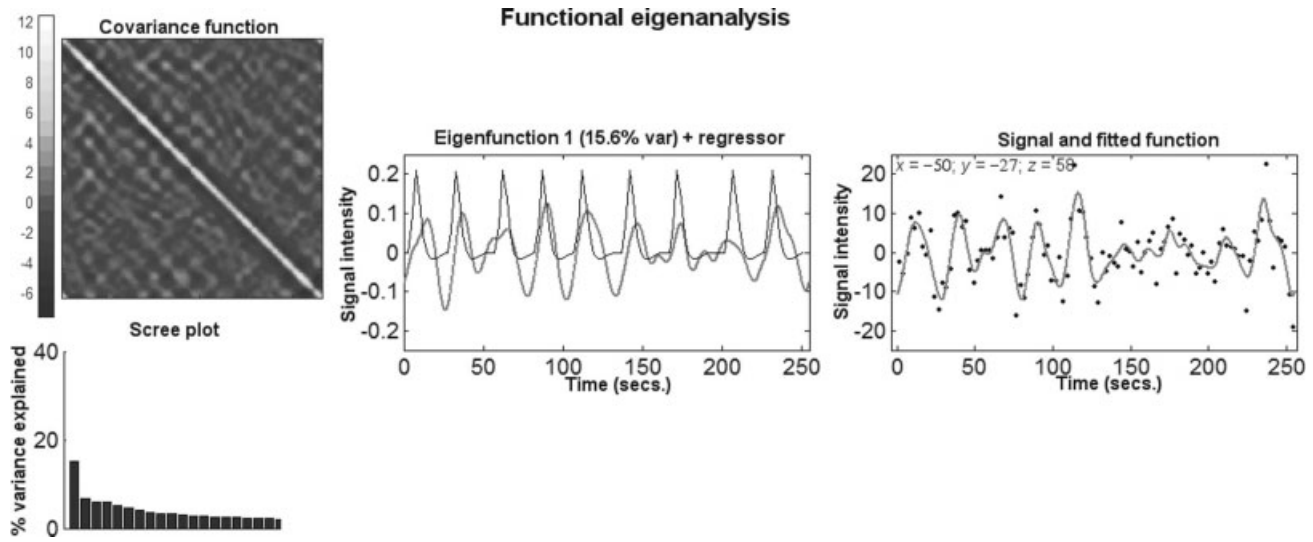

Figure 8.

Left: Covariance function and scree plot resulting from eigenanalysis of the finger-tapping experiment fitted with half the number of bases necessary to interpolate the data together with a fixed smoothing coefficient of 12. **Center**: The first eigenfunction (gray) is displayed together with the regressor obtained from the SPM package (black). **Right**: Signal (black dots) and fitted function (in gray) for the voxel in the dorsolateral prefrontal cortex also displayed in the previous figure.

images mean that in these voxels the variance has been discarded by generalized cross-validation. The component score is therefore about zero in all these voxels.

Although Figure 10 is limited necessarily to displaying selected slices, a rough idea of the extent of this phenomenon is given from the scree plots of the functional PCAs from the generalized cross-validation fits (Fig. 2 and 7). These scree plots show that apart from the first component, little variance has remained in the data.

DISCUSSION

There are two concurring reasons why functional PCA is effective in recovering the signal generated from the experimental manipulations. The first is that, as discussed in the section on smoothing amount selection, in the typical fMRI experimental setting the frequency of the BOLD signal is low relative to the sampling rate of the scans [Turner et al., 1997]. This means that the signal at adjacent time points is likely to be correlated. In the presence of noise, the functions from which the data are sampled may therefore be better estimated by adding a roughness penalty and thus biasing the estimate toward smooth curves [Eubank, 1988]. The resulting covariance function is also comparatively smoother. The second reason is that, as in all experimental settings, the alternation of the experimental conditions is a source of systematic variance in the data. It therefore seems justified to use a PCA method, which selects components based on the amount of variance, to detect it. Taken together, these two considerations provide the rationale for looking for the signal arising from the experimental manipulations in the first components individuated by functional PCA on the penalized fits.

Comparison With Ordinary PCA

The case studies presented here confirm these theoretical suggestions. Systematic variance introduced by alternation of experimental conditions is always retrieved by the first functional component. In contrast, the success of traditional PCA in retrieving this variance is much less constant.

Although these results demonstrate that functional PCA was usually more effective than its ordinary counterpart in retrieving the variance associated with the experimental conditions, one may wonder if ordinary PCA retrieved some genuine signal that the functional version was missing. It is interesting, in this respect, that often the signal retrieved by ordinary PCA is discarded in a fit carried out with generalized cross-validation. This observation casts doubts on the nature of the signal recovered by ordinary PCA.

The effects of generalized cross-validation are apparent in the improvements in the interpretability of the scree plots. Our case studies indicate, however, that the use of other forms of smoothing (such as using a fixed smoothness parameters or reducing the number of basis functions) is also beneficial. Because of their reduced computational requirements, these techniques are probably indicated as a first quick explorative approach to the data. As an objective method, generalized cross-validation removes possible interpretation uncertainties generated by a manual selection of the smoothness coefficient in a final analysis.

Comparison With ICA

Because the assumptions behind functional PCA and ICA are quite different, it is helpful to review them to contrast these two approaches and determine the respective appro-

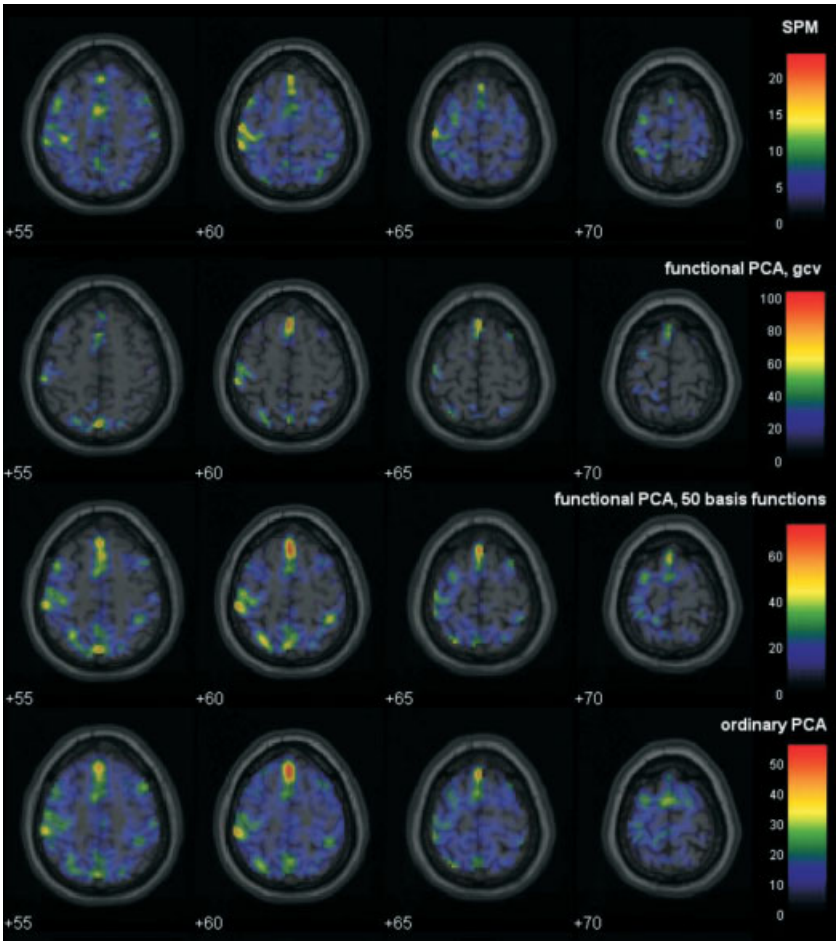

Figure 9.
Finger-tapping paradigm. **Top row**: Slices from the beta image from the regression obtained with the SPM package. **Second row**: Slices from the first functional principal component scores image, obtained with generalized cross-validation. **Third row**: Slices from the first functional principal component scores image, with smoothing obtained by fitting half the number of bases necessary to interpolate the data, and a fixed smoothing coefficient of 12. **Bottom row**: Slices from the first ordinary component score image.

priate domains of application. ICA assumes the components to be independent [Bell and Sejnowski, 1995; McKeown et al., 1998] and non-Gaussian [Hyvärinen and Oja, 2000]. This is a powerful and attractive set of assumptions that make ICA very competitive for exploratory tasks; however, ICA treats the observed signals as a set of random variables without considering the dependency of adjacent time points. In contrast, functional PCA exploits this dependency by imposing a smoothing constraint that can be fixed or estimated through generalized cross-validation. The resulting shrinkage of the data helps bringing the signal of interest to the fore in the first few components.

If the task is rather one of detecting signal unrelated to the experimental paradigm or the assumption of smoothness of the signal no longer applies, the rationale for the adoption of functional PCA is no longer present. In this case, it may be helpful to use different approaches and compare the results in light of the different underlying assumptions.

Comparison With Other Methods Based on the Design Matrix

We address the issue of the difference between functional PCA and other forms of model selection in multivariate methods, such as multivariate linear models (MLM) [Worsley et al., 1997] or in parametric mapping with an extended design matrix [Kherif et al., 2002]. The first fundamental difference is that although these methods address the issue of inference on the effects of the experimental paradigm, functional PCA is purely explorative. By specification, our method is required not to use information from the experimental design matrix. On the contrary, use of such information is appropriate and necessary in MLM, a form of canonical correlation analysis, and in SPM.

A further general difference between functional PCA and MLM is in the estimate of the covariance structure of the data (we use the word structure because in the functional case it is a bivariate function, not a matrix). This point can be appreciated easily if one compares the graphic renderings of both covariance structures in the preceding figures. Because of the different character of the covariance structures, there is a significant difference between the functional approach and any multivariate method that makes use of the maximum likelihood estimate of the covariance matrix of the data. Whether the use of penalized covariance structures may also be of advantage in functional MRI when applying multivariate methods in general is an issue that lies beyond

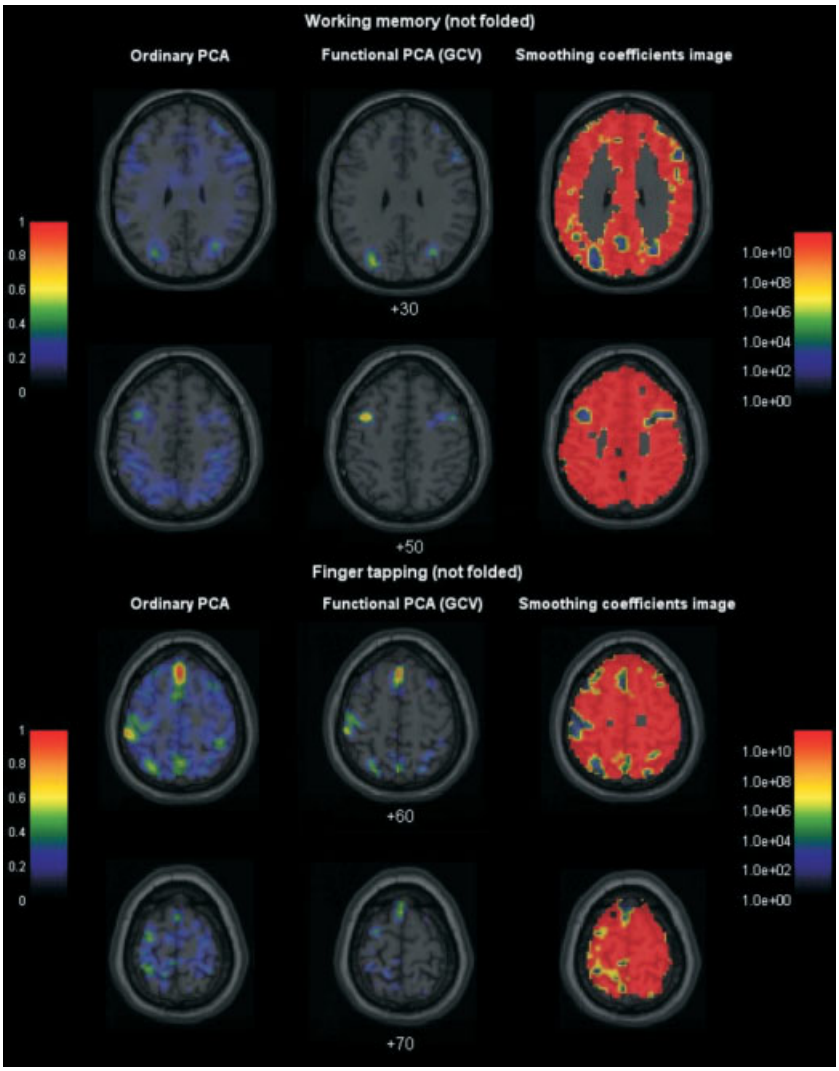

Figure 10.
Selected slices of the beta images from ordinary (**left**) and functional PCA (**center**) from the working memory (**top two rows**) and finger-tapping paradigm (**bottom two rows**), calculated without folding the data around the period of the block. To clarify the reasons for the different distribution of activation in these two methods, the corresponding slices from the images composed of the smoothing coefficients estimated through generalized cross-validation are displayed on the **right**. The gray patches near the midline in these coefficient images are due to the segmentation process, which has classified these voxels to the white matter compartment. To facilitate comparison, the beta and component score values have been normalized to the unit scale. To obtain adequate contrast in the display, the smoothing coefficient images are in logarithmic scale.

the focus of the present work. At least when the dependent variable is categorical (classification tasks), there are studies that suggest that using the functional approach is advantageous [Hastie et al., 1995].

In the literature stemming from the SPM approach, it has been proposed that uncertainties in the form of the BOLD function may be overcome by forming an expanded design matrix that includes, for example, derivatives of the experimental regressor [Kherif et al., 2002]. To avoid overfitting the data, one needs to carry out model selection on the expanded design matrix [Kherif et al., 2002]. There is here a point of similarity with our approach because in both cases, the detection of a signal is accomplished by means of model selection (in functional PCA implemented through the shrinkage given by smoothing and generalized cross-validation) instead of using precise information from the experimental paradigm. However, the type and direction of model selection differ in several respects in functional PCA and in expanded design matrix approaches.

Firstly, in functional PCA shrinkage is carried out on the data, not on the design matrix as in Kherif et al. [2002]. Although the motivation for the adoption of PCA and ridge methods in statistical applications is collinearity of the predictors, this typically happens in the context of observed predictor variables [Hastie et al., 2001], not in the context of experimental variables as in most fMRI design matrices. In a multivariate approach to fMRI data, it is the dependent variables that are highly multidimensional and contaminated by noise, so that shrinking of the dependent variables seems justified. Furthermore, because it makes very little or no use of the experimental paradigm, functional PCA casts a wider net than do methods based on an expanded design matrix. For example, the expanded design matrix approach would be difficult to execute in a situation when the onset of the response is not known at all.

Secondly, in the functional approach the direction of the shrinkage is determined by a notion of frequency (coefficients of base function with many "oscillations" are penal-

ized by the fit). By contrast, discrete multivariate methods have no notion of frequency. This direction of shrinkage is motivated by extensive experience in the field of nonparametric curve estimation under the assumption that the signal is smooth in the presence of measurement noise. In the comparative illustrations of the covariance structures, one can see that this direction of shrinkage results in qualitative differences in the estimate of the covariance structure and hence on any multivariate estimate that depends on the covariance structure.

REFERENCES

Anderson TW (1984): An introduction to multivariate statistical analysis. Second ed. New York: John Wiley and Sons. 675 pp.

Ashburner J, Friston KJ (1997): Spatial transformation of images. In: Frackowiak RSJ, Friston KJ, Frith CD, Dolan RJ, Mazziotta JC, editors. Human brain function. London: Academic Press. p 43–58.

Baddeley AD (1992): Working memory. Science 255:556–559.

Bell AJ, Sejnowski TJ (1995): An information-maximization approach to blind separation and blind deconvolution. Neural Comput 7:1129–1159.

Brett M (2000): Slice display software. Online at http://www.mrc-cbu.cam.ac.uk/Imaging/display_slices.html (accession date 2 November 2002).

Callicott JH, Weinberger DR (1999): Functional MRI in psychiatry. In: Moonen CTW, Bandettini PA, editors. Functional MRI. Berlin: Springer. p 501-512.

Carew JD, Wahba G, Xie X, Nordheim EV, Meyerand ME (2003): Optimal spline smoothing of fMRI time series by generalized cross-validation. Neuroimage 18:950–961.

Cohen JD, Forman SD, Braver TS, Casey BJ, Servan-Schreiber D, Noll DC (1994): Activation of prefrontal cortex in a nonspatial working memory task with functional MRI. Hum Brain Mapp 1:293–304.

Craven P, Wahba G (1979): Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math 31:377–403.

de Boor C (1978): A practical guide to splines. Berlin: Springer. 392 pp.

Demmler A, Reinsch C (1975): Oscillation matrices with spline smoothing. Numer Math 24:375–382.

Desgranges B, Baron JC, Eustache F (1998): The functional neuroanatomy of episodic memory: the role of the frontal lobes, the hippocampal formation, and other areas. Neuroimage 8:198–213.

Draper NR, Smith H (1998): Applied regression analysis, Third ed. New York: John Wiley and Sons. 672 p.

Eubank RL (1988): Spline smoothing and nonparametric regression. New York: Marcel Dekker. 438 p.

Friston KJ (1997): Characterising distributed functional systems. In: Frackowiak RSJ, Friston KJ, Frith CD, Dolan RJ, Mazziotta JC, editors. Human brain function. London: Academic Press. p 107–126.

Friston KJ, Frith CD, Liddle PF, Frackowiak RSJ (1993): Functional connectivity: the principal-component analysis of large (PET) data sets. J Cereb Blood Flow Metab 13:5–14.

Friston KJ, Frith CD, Frackowiak RSJ, Turner R (1995a): Characterizing dynamic brain responses with fMRI: a multivariate approach. Neuroimage 2:166–172.

Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ (1995b): Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 2:189–210.

Gabrieli JD (1998): Cognitive neuroscience of human memory. Ann Rev Psychol 49:87–115.

Green PJ, Silverman BW (1994): Nonparametric regression and generalized linear models: a roughness penalty approach. London: Chapman and Hall. 182 pp.

Grön G, Bittner D, Schmitz B, Wunderlich AP, Tomczak R, Riepe MW (2001): Hippocampal activations during repetitive learning and recall of geometric patterns. Learn Mem 8:336–345.

Hastie TJ, Buja A, Tibshirani RJ (1995): Penalized discriminant analysis. Ann Statist 23:73–102.

Hastie TJ, Tibshirani RJ, Friedman J (2001): The elements of statistical learning. Data mining, inference, and prediction. New York: Springer. 533 p.

Hoerl AE, Kennard RW (1970): Ridge regression: biased estimation for non-orthogonal problems. Technometrics 12:55–67.

Holmes A, Poline JB, Friston KJ (1997): Characterising brain images with the general linear model. In: Frackowiak RSJ, Friston KJ, Frith CD, Dolan RJ, Mazziotta JC, editors. Human brain function. London: Academic Press. p 141–159.

Hyvärinen A, Oja E (2000): Independent component analysis: algorithms and applications. Neural Netw 13:411–430.

Jolliffe IT (1986): Principal component analysis. Heidelberg: Springer. 271 pp.

Kherif F, Poline JB, Flandin G, Benali H, Simon O, Dehaene S, Worsley KJ (2002): Multivariate model specification for fMRI data. Neuroimage 16:1068–1083.

Kimeldorf GS, Wahba G (1970): A correspondence between bayesian estimation on stochastic processes and smoothing by splines. Ann Math Statist 41:495–502.

Kolmogorov AN, Fomin SV (1968): Introductory real analysis [English translation]. Englewood Cliffs, NJ: Prentice-Hall. 403 pp.

Lee TW, Girolami M, Bell AJ, Sejnowski TJ (1999): A unifying information-theoretic framework for independent component analysis. Neural Comput 10:2103–2144.

McKeown MJ, Makeig S, Brown GG, Jung T-P, Kindermann SS, Bell AJ, Sejnowski TJ (1998): Analysis of fMRI data by blind separation into independent spatial components. Hum Brain Mapp 6:160–188.

McIntosh AR, Bookstein FL, Haxby JV, Grady CL (1996): Spatial pattern analysis of functional brain images using partial least squares. Neuroimage 3:143–157.

Petersson KM, Nichols TE, Poline J-B, Holmes AP (1999): Statistical limitations in functional neuroimaging I. Non-inferential methods and statistical models. Philos Trans R Soc Lond B Biol Sci 354:1239–1260.

Poggio T, Girosi F (1990): Networks for approximation and learning. Proc IEEE 78:1481–1497.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1988): Numerical recipes in C. Cambridge: Cambridge University Press. 735 pp.

Ramsay JO, Silverman BW (1997): Functional data analysis. Berlin: Springer. 328 pp.

Ramsay JO, Silverman BW (2001): Functional data analysis software, MATLAB edition. Online at http://www.psych.mcgill.ca/faculty/ramsay/software.html (accession date 2 November 2002).

Rao RC (1958): Some statistical methods for comparison of growth curves. Biometrics 14:1–17.

Rice JA, Silverman BW (1991): Estimating the mean and covariance structure nonparametrically when the data are curves. J R Stat Soc B 53:233–243.

Schoenberg I (1964): Spline functions and the problem of graduation. Proc Natl Acad Sci USA 52:947–950.

Talairach J, Tournoux P (1988): Co-planar stereotaxic atlas of the human brain. Stuttgart: Thieme. 122 pp.

Turner R, Howseman A, Rees G, Josephs O (1997): Functional imaging with magnetic resonance. In: Frackowiak RSJ, Friston KJ, Frith CD, Dolan RJ, Mazziotta JC, editors. Human brain function. London: Academic Press. p 467–486.

Wahba G (1990): Spline models for observational data. Philadelphia: Society for Industrial and Applied Mathematics. 169 pp.

Walter H, Wunderlich AP, Blankenhorn M, Schäfer S, Tomczak R, Spitzer M, Grön G (2003a): No hypofrontality, but absence of prefrontal lateralization comparing verbal and spatial working memory in schizophrenia. Schizophr Res 61:175–184.

Walter H, Bretschneider V, Grön G, Zurowski B, Wunderlich AP, Tomczak R, Spitzer M (2003b): Evidence for quantitative domain dominance for verbal and spatial working memory in frontal and parietal cortex. Cortex 39:897–911.

Wismüller A, Lange O, Dersch D, Leinsinger G, Hahn K, Pütz B, Auer D (2002): Cluster analysis of biomedical image time-series. Int J Comp Vision 46:102–128.

Worsley KJ, Poline JB, Friston KJ, Evans AC (1997): Characterizing the response of PET and fMRI data using multivariate linear models. Neuroimage 6:305–319.

APPENDIX

Ordinary PCA on fMRI Data

Remembering that the $N \times M$ matrix of data $\mathbf{Y}$ is composed by the rows $\mathbf{d}_1', \mathbf{d}_2', \ldots \mathbf{d}_N'$, each being one volume, and by the columns $\mathbf{y}_1', \mathbf{y}_2', \ldots \mathbf{y}_M'$, each being a time series composed of one of the M voxels sampled in each scan, the first step to carry out ordinary PCA is to double-center the data, thus obtaining the centered data matrix $\tilde{\mathbf{Y}}$. The double-centering is appropriate because the data are not composed of "observations" and "variables," but rather of variation in time and space coordinates [Jolliffe, 1986]. It also is the direct counterpart of the procedure here adopted for the functional version of the algorithm. The eigenvectors and the eigenvalues may be obtained by carrying out the singular-value decomposition of the $N \times N$ covariance matrix $M^{-1}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'$, which retrieves the solutions of the eigenequation

$$M^{-1}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'\mathbf{g} = \gamma\mathbf{g} \tag{26}$$

in $\mathbf{g}$ and γ subject to the constraint $\|\mathbf{g}\|^2 = 1$ (compare to equation [10]). The matrix $M^{-1}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}'$ is the maximum likelihood estimate of the real covariance matrix [Anderson, 1984].

Derivation of the Demmler-Reinsch Form of Spline Regression

We will consider here the case in which the matrix $\mathbf{F}$ is square and of full rank. This will generally obtain when the minimum number of splines is used that can interpolate the data exactly (and would do so, if no penalty term were used). The main difficulty consists in reparametrizing the hat matrix taking into account the penalty matrix $\mathbf{P}$ and the smoothness parameter λ separately. To achieve this, we first rearrange the expression for the hat matrix:

$$\mathbf{H}_\lambda = \mathbf{F}(\mathbf{F}'\mathbf{F} + \lambda\mathbf{P})^{-1}\mathbf{F}' = [(\mathbf{F}')^{-1}(\mathbf{F}'\mathbf{F} + \lambda\mathbf{P})(\mathbf{F})^{-1}]^{-1}$$
$$= [(\mathbf{F}')^{-1}\mathbf{F}'\mathbf{F}\mathbf{F}^{-1} + \lambda(\mathbf{F}')^{-1}\mathbf{P}\mathbf{F}^{-1}]^{-1} = [\mathbf{I} + \lambda(\mathbf{F}')^{-1}\mathbf{P}\mathbf{F}^{-1}]^{-1}.$$

If we eigendecompose the quadratic form $(\mathbf{F}')^{-1}\mathbf{P}\mathbf{F}^{-1} = \mathbf{Q}\mathbf{B}\mathbf{Q}'$, and replace in the above, then, remembering that $\mathbf{Q}\mathbf{Q}' = \mathbf{Q}'\mathbf{Q} = \mathbf{I}$,

$$\mathbf{H}_\lambda = (\mathbf{I} + \lambda\mathbf{Q}\mathbf{B}\mathbf{Q}')^{-1} = \mathbf{Q}(\mathbf{I} + \lambda\mathbf{B})^{-1}\mathbf{Q}' = \mathbf{Q}\mathbf{W}\mathbf{Q}',$$

where we let $(\mathbf{I} + \lambda\mathbf{B})^{-1} = \mathbf{W}$. Because $(\mathbf{I} + \lambda\mathbf{B})^{-1}$ is diagonal, we conclude that $\mathbf{Q}\mathbf{W}\mathbf{Q}'$ is also the eigendecomposition of $\mathbf{H}_\lambda$, and therefore that $\mathbf{H}_\lambda$ and $(\mathbf{F}')^{-1}\mathbf{P}\mathbf{F}^{-1}$ share the same eigenvectors $\mathbf{Q}$. From the equation above, we have

$$\mathbf{H}_\lambda = \mathbf{Q}(\mathbf{I} + \lambda\mathbf{B})^{-1}\,\mathbf{Q}' = \mathbf{Q}(\mathbf{Q}'\mathbf{Q} + \lambda\mathbf{B})^{-1}\,\mathbf{Q}'.$$

The proof of the common ordering of eigenvalues and number of oscillations of the eigenfunctions can be found in Demmler and Reinsch [1975].