THE ROYAL SOCIETY

Statistical limitations in functional neuroimaging I. Non-inferential methods and statistical models

Karl Magnus Petersson[1*], **Thomas E. Nichols**[2,3], **Jean-Baptiste Poline**[4,5] **and Andrew P. Holmes**[5,6]

[1]*Cognitive Neurophysiology R2-01, Department of Clinical Neuroscience, Karolinska Institute, Karolinska Hospital, S-171 76 Stockholm, Sweden (karlmp@neuro.ks.se)*

[2]*Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA (nicholst@stat.cmu.edu)*

[3]*Center for the Neural Basis of Cognition, Carnegie Mellon University and University of Pittsburgh, 4400 Fifth Avenue, Pittsburgh, PA 15213, USA*

[4]*Commissariat l'Energie Atomique, Direction de la Recherche Medicale, Service Hospitalier Frederic Joliot, 4 Pl General Leclerc, 91406 Orsay, France (poline@shfj.cea.fr)*

[5]*Wellcome Department of Cognitive Neurology, Functional Imaging Laboratory, 12 Queen Square, London WC1N 3BG, UK (andrew@fil.ion.ucl.ac.uk)*

[6]*Robertson Centre for Biostatistics, Department of Statistics, University of Glasgow, Boyd Orr Building, University Avenue, Glasgow G12 8QQ, UK (andrew@stats.gla.ac.uk)*

CONTENTS PAGE

Functional neuroimaging (FNI) provides experimental access to the intact living brain making it possible to study higher cognitive functions in humans. In this review and in a companion paper in this issue, we discuss some common methods used to analyse FNI data. The emphasis in both papers is on assumptions and limitations of the methods reviewed. There are several methods available to analyse FNI data indicating that none is optimal for all purposes. In order to make optimal use of the methods available it is important to know the limits of applicability. For the interpretation of FNI results it is also important to take into account the assumptions, approximations and inherent limitations of the methods used. This paper gives a brief overview over some non-inferential descriptive methods and common statistical models used in FNI. Issues relating to the complex problem of model selection are discussed. In general, proper model selection is a necessary prerequisite for the validity of the subsequent statistical inference. The non-inferential section describes methods that, combined with inspection of parameter estimates and other simple measures, can aid in the process of model selection and verification of assumptions. The section on statistical models covers approaches to global normalization and some aspects of univariate, multivariate, and Bayesian models. Finally, approaches to functional connectivity and effective connectivity are discussed. In the companion paper we review issues related to signal detection and statistical inference.

Keywords: functional neuroimaging methods; PET; fMRI; statistical models; non-inferential methods; model selection

[*] Author for correspondence.

1. INTRODUCTION

During the last two decades a body of well-described theories and empirically validated methods has been developed, providing a framework for investigating functional neuroimaging (FNI) data and making neuroscientific inferences. This provides a background for the development of new analytical tools and experimental methods. The methods for analysing FNI data are in rapid development, illustrating the need for descriptive tools, validated statistical models, and methods for effective statistical inference. Descriptive and exploratory methods are used to characterize the nature of the signals present in the data, including possible unsuspected effects. Inferential methods are used to test hypotheses and determine confidence intervals addressing the reproducibility or predictability of the observed effects.

Progress in science is dependent on the long-term consistency and convergence of empirical results. Discussion and evaluation of the methods used in a scientific field is of central importance in this process (e.g. Aguirre *et al.* 1998*a*; Clark & Carson 1993; Ford 1995; Ford *et al.* 1991; Frackowiak *et al.* 1996; Friston 1993, 1995; Halber *et al.* 1997; Holmes *et al.* 1998; McColl *et al.* 1994; Petersson 1998; Roland & Gulyas 1996; Strother *et al.* 1995*b*; Taylor *et al.* 1993; Van Horn *et al.* 1995; Worsley *et al.* 1993). In line with this perspective, several investigations of different approaches to the statistical analysis of FNI data (Arndt *et al.* 1995; Grabowski *et al.* 1996; Holmes 1994; McColl *et al.* 1994; Xiong *et al.* 1996), as well as different statistical models (Holmes 1994; McColl *et al.* 1994; Senda *et al.* 1998; Woods 1996), have been reported. Reliability across different variables has also been studied, for example, across laboratories (Ojemann *et al.* 1998; Poline *et al.* 1996; Senda *et al.* 1998), groups, sample sizes, and imaging modalities (Andreasen *et al.* 1995; Grabowski *et al.* 1996; Ojemann *et al.* 1998; Strother *et al.* 1997). In addition, studies of statistical power (Friston *et al.* 1994*b*, 1996*a*; Van Horn *et al.* 1998) and activation pattern reproducibility (Strother *et al.* 1997) have recently been reported.

The field of FNI methodology has developed into a mature but still evolving area of knowledge. The scope of this overview is limited to the common methods used when analysing data from positron emission tomography (PET) or functional magnetic resonance imaging (fMRI), concentrating on regional cerebral blood flow (rCBF) PET and blood oxygenation level dependent (BOLD) fMRI. In this paper we give an overview of some non-inferential methods (principal components analysis (PCA), independent components analysis (ICA), and the scaled subprofile model (SSM)) that in conjunction with inspection of parameter estimates and other simple measures can be used in the process of model selection and verification of assumptions in an informal sense. Next, we cover some models for global effects and approaches to global normalization. Univariate (the general linear model (GLM), fixed versus random effects), multivariate (MANCOVA, canonical variates analysis (CVA), partial least squares (PLS), multivariate linear models (MLM)), and Bayesian models, including some specific issues relating to fMRI (temporal autocorrelation, models of the haemodynamic response) will also be discussed. The last section of this paper describes approaches to functional connectivity (covariance analysis, partial correlation coefficients, covariance fields) and network analysis of effective connectivity (structural equations modelling). In the companion paper (Petersson *et al.*, following paper) we give an overview and discuss some issues relating to signal detection and methods for statistical inference used in FNI. The emphasis of both papers is on assumptions and limitations. The descriptions of the methods reviewed are necessarily brief and for more details the reader is referred to the appropriate literature.

Several methods for FNI data analysis have been proposed indicating that none is optimal for all purposes. In this context, it is important to know the limitations inherent in the different approaches enabling the optimal use of available methods. It should be noted that focusing on assumptions and limitations of the methods reviewed is different from the claim that these methods should or should not be used. On the contrary, this indicates the limits of applicability and usefulness inherent in any given method. In order to interpret accurately FNI results it is of importance to know the assumptions, approximations and inherent limitations of the methods used. When the assumptions and limitations are taken into account, the different methods and approaches reviewed in this paper (and its companion paper) generally serve their purposes well. The benefits and examples of their applicability are well described in the original literature and are not repeated here.

Statistical models make explicit assumptions about data. Both the explicit and implicit assumptions about data need to be critically examined (Lange 1997). The methods used in FNI differ in assumptions made regarding the data and in the approximations used in the statistical analysis. What are of importance in this context are not the assumptions or approximations themselves but how well these are fulfilled by empirical data. Of crucial importance, in the case where these assumptions or approximations are not fully met, is the robustness of the methods used. This notion emphasizes the importance of theoretical and empirical investigations of its robustness in addition to empirical validation and explicit characterization of the inherent limitations of a given method.

The classic strategy for data analysis starts with data exploration and model selection, fitting of a statistical model, assessing the goodness-of-fit and investigating diagnostics for violations of assumptions. If the model does not fit or assumptions are seriously violated then the model selection starts anew. When an appropriate model has been selected and assumptions are not seriously violated, valid statistical inferences can be made. In the case of ill-fitting models or violated assumptions the ensuing inference may be statistically invalid. However, it should be stressed that model selection is a complex process and it is difficult to fully account for the interaction between model selection and statistical inference unless model selection and statistical inference are performed on independent sets of data.

In general, to study a phenomenon of interest, the investigator chooses an experimental design and primary FNI data are collected. The primary data are commonly pre-processed (e.g. realigned, anatomically normalized, and low-pass filtered), a statistical model and a test statistic chosen, model parameters estimated, and statistical inference obtained, taking into account multiple non-independent comparisons and possible temporal autocorrelation (figure 1). Most FNI methods are based on voxel (volume element) data, even though some use regions of interest (ROI) data. Most often, the ROI approach presupposes prior regionally specific hypotheses and brain regions outside the chosen ROIs are not investigated potentially leading to undetected effects. The voxel-by-voxel approach, pioneered by Fox *et al.* (Fox & Mintun 1989; Fox *et al.* 1988) and Friston *et al.* (1990, 1991), was proposed as a less arbitrary alternative. Several standard approaches used in image processing and signal detection (e.g. optimal filtering theory, linear and nonlinear systems theory) are naturally applied to voxel data. Voxel approaches also preserve more of the inherent resolution of the imaging system and make it possible to investigate brain functions without a regional specific hypothesis. However, small structures may naturally be viewed as ROIs and with sufficient prior information regionally specific hypothesis can be formulated and an ROI approach is natural. Here the prior information is used to restrict the search volume and correspondingly increase the sensitivity of the hypothesis testing.

2. NON-INFERENTIAL METHODS FOR SIGNAL CHARACTERIZATION

Non-inferential or exploratory methods are used to characterize the nature of the signal present in data in a manner that does not strongly depend on a particular choice of model for the data. Exploratory methods can play an important role both before and after statistical inference. Before inference, they can aid the process of model building and model selection by pointing to sources of variability that might not have been expected. After inference, they can similarly serve as a check that the model has adequately accounted for most of the systematic variability in the data. In this section we review three non-inferential methods: PCA, ICA and the SSM. We close by suggesting some basic exploratory approaches that can be applied to almost any inferential model.

(a) ***Principal components analysis***

PCA provides a means to identify spatio-temporal patterns in a data-driven manner. In general, PCA is a way of summarizing the sample variance–covariance structure of multivariate data. As will be seen later, PCA plays a key role in multivariate statistical techniques, used as an exploratory tool to guide model building, to identify patterns, and to estimate the approximate dimensionality of the data, and may be used as a dimensionality reduction device. Here we will concentrate on its descriptive use. In addition, Friston *et al.* (1993) interpret PCA results as information on functional connectivity, defined as the observed correlation over time between different brain areas. This is distinguished from effective connectivity, which is defined as the direct influence one neural system has on another (see also § 3(e) on functional connectivity and network analysis).

The definition of PCA is as follows: consider each brain image to be a single multivariate observation, with each voxel in the brain being an element in a long row vector. Stacking the rows will create a data matrix, X, where each column represents a voxel and each row a scan of centred (mean corrected) data. The voxel-by-voxel sample variance–covariance matrix, proportional to XX', expresses the first-order relationship between each pair of voxels. A PCA of XX' can be accomplished by a singular value decomposition (SVD) of X (Jolliffe 1986). The SVD produces three objects: principal components (PCs), which are spatial patterns; component scores, which are temporal or scan-order patterns; and singular values (eigenvalues), which express the relative variability accounted for by each PC.

PCs are defined in a natural order: the first component is the single image or brain pattern (also called eigen- or singular image or vector) that explains the most variability across all images. Each subsequent component explains the most variability, subject to the constraint that it is orthogonal to all the previous components (Jolliffe 1986). The component scores indicate the temporal pattern corresponding to each PC, and the singular values allow for a qualitative assessment of the importance of each PC.

The SVD actually produces eigenvalues and eigenvectors of both XX' and $X'X$; the former is proportional to the sample voxel-by-voxel covariance matrix, the latter is similar to the scan-by-scan covariance matrix. The voxel-by-voxel eigenvectors are the PCs, and the scan-by-scan eigenvectors are the component scores. The eigenvalues are the same for both and, when scaled to sum to 1, are the proportion of variability that each eigenvector or PC accounts for. There is a symmetry between PCs and component scores: the first component is the pattern over voxels (i.e. brain pattern) that accounts for the most variability across all scans. The first component score is the pattern over scans (i.e. temporal pattern) that accounts for the most variability across all voxels. In fact, the components and corresponding scores should be viewed as spatio-temporal objects, and interpreted conjointly.

PCA may be used after inference to check for unexpected or unaccounted patterns in the data. If the PCA produces a small number of PCs whose components correspond to the experimental paradigm, and these PCs account for a large part of the observed variability, then one can be fairly confident that the experiment incurred the majority of the variability in the data, and that this variability is appropriately modelled. If, on the other hand, there are spatio-temporally structured components that explain a large amount of variability but do not correspond to the experimental paradigm, then there may be important additional sources of variance not included in the statistical model, potentially biasing both the model coefficients and variance estimates making subsequent inference invalid (cf. Petersson *et al.*, following paper). Typical examples of such sources are temporal effects (e.g. trends in levels of anxiety or attention, increasing boredom or discomfort, and effects of practice), evidenced by components similar to temporal trends or PCs consisting of a few strong weightings on just the first or last replications of an experimental condition. In addition, it is possible to investigate the standardized residual images with a PCA (Ford 1995), to check if there are any

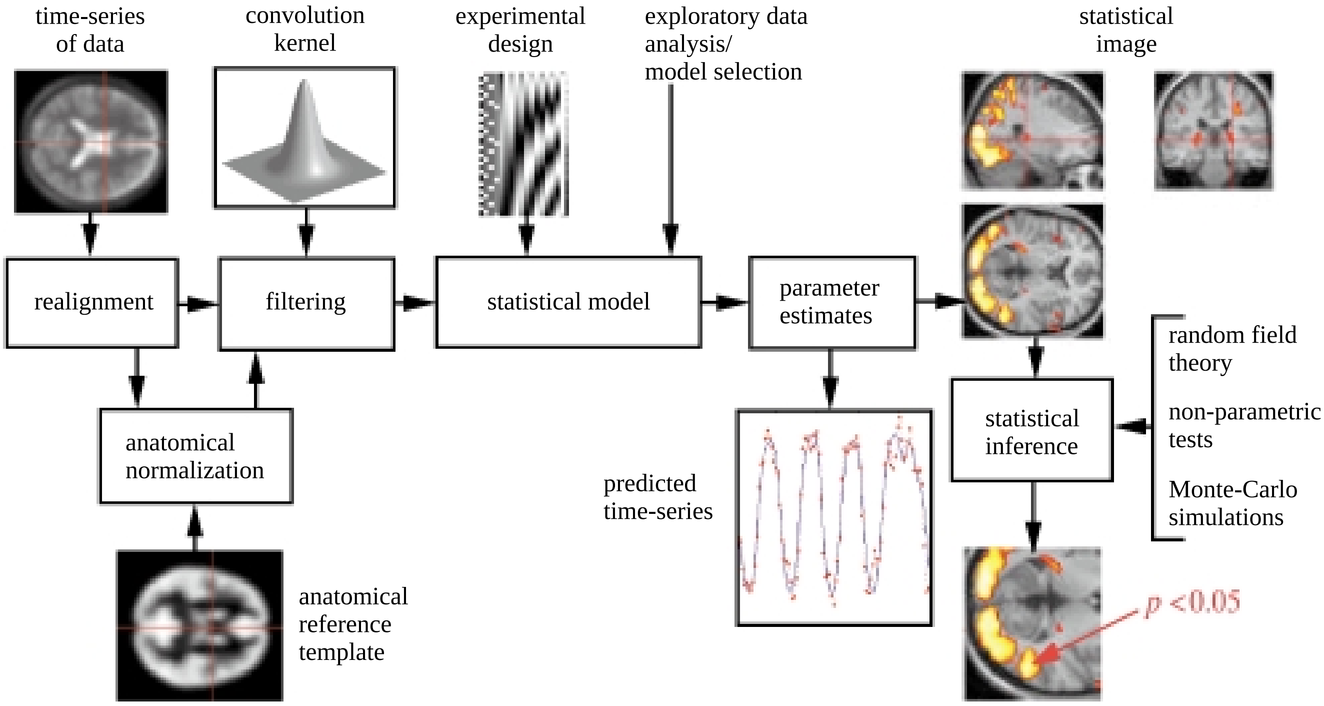

Figure 1. The classic strategy for data analysis starts with data exploration and model selection, fitting of a statistical model, assessing the goodness-of-fit and investigating diagnostics for violations of assumptions. In general, proper model selection is a necessary prerequisite for the validity of the subsequent statistical inference. Model selection is a complex process and it is difficult to account for the interaction between model selection and statistical inference unless model selection and statistical inference are performed on independent sets of data. In FNI, the primary data are commonly pre-processed, e.g. realigned, anatomically normalized, and filtered. A statistical model is chosen, determined by the experimental design and the previous model selection procedures. A test statistic is chosen and model parameters are estimated. Finally, statistical inference, taking into account multiple non-independent comparisons and possible temporal autocorrelation, represented as a statistic image is obtained.

structured components left in the residual. If structured components are found, then an assumption of independence may be violated (Petersson *et al.*, following paper).

The result of a PCA may be sensitive to outliers, and is crucially dependent on the type of pre-processing performed on data. For example, if the image data are normalized by dividing by the standard deviation image, then the PCs are based on the correlation matrix rather than the covariance matrix and the influence of high variance regions will be reduced possibly giving a different result (Chatfield & Collins 1980; Jolliffe 1986). Another example is how the PCA of Friston *et al.* (1993) only examines voxels that survive an *F*-test of significant experimental variance, the result of so-called *F*-masking. This implies that variability not sufficiently well accounted for by the statistical model will not be considered in the PCA, and this changes the interpretation of the results. Rather than partition all the observed variability, the PCs only partition the variability among those voxels in which the model explains a sufficient part (as determined by the *F*-threshold) of the observed variation. We note, though, that the *F*-masking may not always have a great impact on the results (Strother *et al.* 1995*a*). In addition, since the PCA (Friston *et al.* 1993) is typically based on data after adjustments for nuisance effects, the results will be dependent on the particular nuisance effects modelled, their model fit, and the proper removal of these.

There is nothing inherent in PCA to guarantee a straightforward interpretation of the PCs. For example, it is possible that a given PC represents a mixture of effects, and understanding spatio-temporal patterns that are constrained to be orthogonal may be challenging unless the observed variability has a natural elliptic structure. This is particularly the case with fMRI data, where there are multitudes of effects. While the first component always has the intuitive explanation of expressing the greatest amount of variability in the data, each of the subsequent components must be interpreted as expressing the greatest amount of variability orthogonal to the previous components. The easiest way to conceptualize this is to think of the PCA as a sequential process. After the first component is identified, the data are regressed on it and any patterns it can account for are removed. Then the next component is the pattern that expresses the greatest amount of variability in this modified data set. The same process is repeated for each subsequent component.

(b) *Independent components analysis*

The ICA is another non-inferential method closely related to PCA. The ICA approach was originally proposed by Common (Common 1994) and it has recently been applied to fMRI data (McKeown *et al.* 1998). For an introduction to ICA see McKeown *et al.* (1998), and for theoretical details see Bell & Sejnowski (1995); here we describe ICA as it has been applied to fMRI. The application of ICA to fMRI data so far has focused on the spatial aspects of data, which may be called ‘spatial ICA’. Like PCA, it is also possible to analyse the temporal aspects of data with 'temporal ICA'. In contrast to PCA, the spatial patterns or component maps (CM) generated by ICA are constrained to be not just orthogonal but statistically independent and spatially sparse, that is, with only a few voxels having large values in each CM. Further, while PCA constrains the component scores (temporal patterns) to be orthogonal, ICA puts no constraint on the temporal components.

ICA is motivated through information theory (Ash 1965; Cover & Thomas 1991) with the objective of maximizing the joint entropy of suitably transformed CMs (McKeown *et al.* 1998). Maximizing the joint entropy is equivalent to minimizing the mutual information, which produces independent components. Independent components have zero mutual information and the transformation of the CMs biases the method in favour of spatially sparse CMs. It must be stressed that the independence at hand is not linear independence, a geometric property, but statistical independence, a distributional property. Statistical independence is a notion of separability. Two random variables or random vectors are independent if, and only if, their joint distribution can be written as a product of the marginal distributions: $p(x,y) = p(x)p(y)$.

ICAs principle strength is that the interpretation of the temporal profiles may be easier compared with the temporal profiles generated by PCA, since they are not forced to be orthogonal. As a result, several temporal patterns can be found which are very similar but not identical to the main task component. In contrast, PCA would not identify transiently task-related temporal components since they would tend not to be orthogonal to the task. The main weakness is that the spatial modes are required to be independent. For example, brain areas activated by task performance must be spatially independent of the distributions of areas affected by artefacts (McKeown *et al.* 1998), though this seems not to be a problem in the data analysed by McKeown *et al.* (1998). However, it has been argued that functional integration among brain regions implies that large-scale cognitive neural networks may have substantial anatomic overlap, and that ICA precludes nonlinear interaction between spatial modes subserving context-sensitive modulation of one area by another (Friston 1998*b*). This issue could be addressed by performing a 'temporal ICA', where the temporal profiles are independent and the spatial profiles are unconstrained. This requires a dimensionality reduction step so that the spatial dimensionality of the data is reduced below that of time.

(c) *The scale subprofile model approach*

The SSM was introduced by Moeller *et al.* (1987) to study subject-by-region interactions while addressing the issue of confounding between the global and regional signal. SSM is a non-inferential multivariate method initially described for ROI analyses of IP-fluorodeoxy-D-glucose (FDG) data and has subsequently been extended to ^{15}O water activation data (Strother *et al.* 1995*a*). SSM works with log-transformed data and is a type of factor ANOVA, that is, a two-way ANOVA with factor analysis of the residuals. SSM decomposes variability into three different components: a multiplicative global scale factor (GSF), a group mean image or group mean profile (GMP), and components and scores akin to PCA. These components and scores describe the covariance structure of the remaining variability (the components are named group invariance subprofiles, and the scores subprofile scaling factors (SSF)). These effects are estimated on the log-transformed data by a serial process. For purposes of explanation we assume that there is only one scan per subject, though this is not a limitation of SSM (Moeller *et al.* 1987; Strother *et al.* 1995*a*). First the SSF are estimated from a PCA of centred data (i.e. after voxel and subject effects have been removed). The SSF consists of the first k eigenvectors of the subject by subject sample covariance matrix. A multivariate test of dimensionality is used to determine k. The GSF are then estimated as the global mean activity orthogonalized with respect to the SSFs. The SSFs represent the principal subject covariance patterns across the volume, after removal of the main effects of subject and location, and thus give an estimate of the vector space of subject-by-location interactions. This approach is intended to prevent the GSF from being confounded with individual differences. The group effect (GMP) is then estimated from the log-transformed data adjusted for the global effect; the GMP with the SSF are used to estimate the group invariance subprofile (cf. Moeller & Strother 1991; Strother *et al.* 1995*a*).

Originally the authors presented evidence that their method for estimating global effects allows for more accurate modelling and more specific estimates of regional activity (rA) (Moeller *et al.* 1987). The imaging work (Strother *et al.* 1995*a*) suggested that SSM was less susceptible to artefactual decreases than other methods. SSM stands out for its consideration of intra-subject and intersubject variation as well as the thoughtful estimation of global effects. The latter is important since the method of intensity normalization can have a strong impact on the analysis of spatial covariance structure (Ford 1986). Working with ROIs, Moeller *et al.* (1987) presented evidence that their method for estimating global effects allows for more accurate modelling, and more specific estimates, of rA. Theoretical comparisons are not possible, since a statistical model does not back the SSM.

(d) *Inspection of functional neuroimagimg data*

We close this non-inferential section by noting some basic descriptive approaches that can be employed with almost any statistical model of data. Simple devices such as viewing the parameter estimate images can be most valuable in assessing model fit and increasing understanding of the properties of data. For example, in the case of FNI activation studies, these images would be the parameter estimates of the difference between two conditions, or in the case of a covariate regressor, the slope estimates of rA against the covariate. Examining the images of the parameter and its estimated standard error can indicate possible problematic features in data. For example, a parameter image and corresponding t-statistic image that are starkly different may indicate an unexpected pattern in the standard error image. Such features may be artefacts due to a failure in some preprocessing step, but other possibilities exist. For a given statistical model, it may be useful to examine the different components of variation represented by, for example, subject, condition or various interactions. In this context, useful images to inspect are the root of the mean sum of squares related to these components. For example, Senda *et al.* (1998) illustrate the usefulness of these procedures examining the mean sum of squares images in order to investigate the effect of different anatomical normalization procedures and different statistical models on the results of the data analysis.

In conclusion, non-inferential methods can be most useful in describing properties or characteristics of particular data sets and also in outlier detection. In particular, simple methods can be of great value, like direct inspection of the mean difference, variance, sums of squares and residual images, or after transformation using more sophisticated methods like PCA or SSM. The nature of FNI data often makes the systematic diagnosing of model fit and verification of assumptions challenging. However, many problems can be identified by simple methods and close examination of the image data at all steps, from initial data collection through model fitting. Unexplained structure in the normalized residual image may be investigated by PCA. Recent elaboration of the PCA concept in terms of subspace transformations like projection pursuit (Huber 1985), or other related approaches like ICA, may turn out to be valuable exploratory tools and are naturally applied to FNI data.

3. STATISTICAL MODELS FOR FUNCTIONAL NEUROIMAGING

As a prelude to statistical inference, non-inferential descriptive methods can aid in the process of model building and model selection. The importance of reasonable model fit and verification of assumptions cannot be overestimated since this is of central importance for the validity of the subsequent statistical inference (cf. Petersson *et al.*, following paper). In general, statistical inference requires a sufficiently well-fitting statistical model and several statistical models have been described and applied to FNI data. In this section, models for global effects and approaches to global normalization are described. In particular, the importance of correctly modelling a fluctuating baseline, potential confounding between global and regional effects, and limitations of the commonly used models in the case of several different ranges of global effects will be considered. Different issues relating to univariate statistical inference (e.g. sources of random variation, fixed versus random effects and scope of inference), Bayesian, and multivariate models (e.g. dimensionality reduction, the need to characterize the multivariate signal when detected) are also discussed. Last, some issues specific to fMRI (e.g. temporal autocorrelation, models of the haemodynamic response) are covered.

(a) *Baseline fluctuations and global normalization*

FNI experiments usually test hypotheses regarding regionally specific changes in neuronal activity. These changes are, in the case of PET, indirectly indicated by changes in rCBF or regional cerebral counts (rCC), and by changes in regional susceptibility in the case of BOLD fMRI. (Below regional activity will represent rCBF, rCC, or regional fMRI BOLD time-series, depending on the imaging modality.) For practical and other reasons, the imaging systems are commonly used in a non-quantitative mode. Therefore the focus is on relative regional changes which are related to a baseline. This can be problematic since, for example, variability in global factors often induces baseline fluctuations. Different measures of global activity (gA) have been used to account for some of the baseline variability. An often used simple estimator of gA is the intracerebral average of rA. Global activity defined in this way varies between subjects and over time. The variability depends on several variables (Frackowiak *et al.* 1997; McColl *et al.* 1994), for example, physiological (e.g. changes in pCO_2 levels and circulatory system changes), factors relating to the measurement procedures (e.g. differences in injected radioactive dose) and the imaging system (e.g. between-run variability in fMRI gain). Global changes are therefore difficult to interpret without quantification.

When there is a lack of absolute quantification and the experimentally induced regional changes are assessed relative to a baseline, changes in this baseline are often considered a nuisance effect. Since baseline fluctuations may be large, potentially hiding the effects of interest, it is necessary to account for or remove this variability in some appropriate manner. The notion of baseline variability as a nuisance effect implicitly assumes that the scan-to-scan baseline fluctuations are independent of the experimental manipulations. In order to properly account for baseline variability there are two issues. First, how to measure or estimate the baseline fluctuations, and second, how these measurements are used to explicitly model or remove the variability in baseline activity. Measurements of global effects, and consequently global normalization, is predicated on the assumption that the variability in global effects adequately represent the baseline fluctuations and that the experimentally induced regional changes are superimposed on this according to some model.

(i) *Different approaches to global normalization*

Several approaches to account for global changes have been proposed and compared. For example, proportional scaling (Fox & Raichle 1984; Kanno *et al.* 1996), log-linear regression models (Herholz *et al.* 1993), histogram–rank equalization (Arndt *et al.* 1996), *Z*-score transformation of data (McIntosh *et al.* 1996*b*), or modelled as a nuisance covariate in the GLM (Friston *et al.* 1990). Both the ANCOVA (Friston 1995; Friston *et al.* 1990; Ramsay *et al.* 1993) and the proportional scaling (Kanno *et al.* 1996) approaches have been empirically validated for PET data.

The relation between rCBF and global cerebral blood flow (gCBF) is most likely nonlinear. However, over small ranges it can be expected that the relation is well approximated as linear. For normal subjects and small ranges of gCBF, the incorporation of the gCBF as a covariate in a linear model affords a reasonably good model of the relationship between rCBF and gCBF (Frackowiak *et al.* 1997). The additive ANCOVA model was proposed under the assumption that changes in gCBF and the experimentally induced changes in rCBF are well approximated as independent (Ramsay *et al.* 1993). It was pointed out that the results of this approach may be problematic to interpret if changes in gCBF are correlated with experimentally induced changes in rCBF (Ramsay *et al.* 1993; see also Aguirre *et al.* 1998*b*; Andersson 1997) or if the gCBF estimation is biased. This is also the case for proportional scaling. Recent fMRI studies indicate that there may be complex interactions between induced vasodilation (CO_2 modulation, Bandettini & Wong 1997; acetazolamide modulation, Bruhn *et al.* 1994) and activation-related BOLD signal changes. For example, Bandettini & Wong (1997) observed that the amplitude of activation-induced signal changes was damped during hypercapnia. However, there are some indications that the interaction between activation-related signal changes and vasodilation induced by hypercapnia can be limited (Corfield *et al.* 1998) in certain cases.

The variability in global cerebral counts (gCC) is often larger than in gCBF. Even if gCBF is relatively constant, subject differences in head fraction and variability in the introduced radioactive dose cause variability in gCC. In the case of count data, rCC is proportional to gCC when rCBF is constant. If it is expected that the variability in, for example, head fraction or introduced radioactive dose is dominating, proportional scaling is a reasonable approach. It has been suggested that the regional variance may increase with increased radioactive dose reflecting the Poisson origin of data (Frackowiak *et al.* 1997). In the case of a large range or variability in gCC, a proportional model may therefore be preferred, in particular since this approach has variance-stabilizing properties (Holmes 1994). The *Z*-score transformation also has variance-stabilizing properties (McIntosh *et al.* 1996*b*). The use of variance-stabilizing transforms is a well-known approach in statistics (Bickel & Docksum 1977) and estimation of variance-stabilizing transforms with applications to metabolic PET data have recently been described (Moeller & Strother 1991; Ruttimann *et al.* 1998). The empirical comparisons performed so far have yielded little differences between the proposed approaches to global normalization, neither for PET data (Arndt *et al.* 1996; Frackowiak *et al.* 1997; Holmes 1994; McIntosh *et al.* 1996*b*) nor for fMRI data (Aguirre *et al.* 1998*b*). More specifically, in most FNI studies using normal subjects, the results are similar using either proportional scaling or the ANCOVA approach. (Provided the effect of gA in the ANCOVA approach is allowed to vary between subjects, i.e. a subject-specific ANCOVA model of gA. This allows for a subject-by-gA interaction. However, if the effect of gA is included as a single effect in the GLM, with a single global slope used for all subjects, there may be differences in results between the ANCOVA and proportional scaling approach.)

(ii) *Confounding of global and regional effects*

Another important issue in the context of global normalization is whether global and regional effects are confounded; that is, global and regional effects can be more or less dependent. If the gA is estimated as the intracerebral spatial average of rA there is a risk that the variability in gA may be significantly confounded with behavioural manipulations or experimentally induced effects. Since the gA is the average of rA, and rA is hypothesized to correlate with experimental manipulations, the gA may be correlated with changes in experimental conditions, unless all regional increases are matched by regional decreases. This seems unlikely, implying that gA estimated in this way is more or less confounded with the experimental paradigm. Whether this confounding is appreciable to the point of affecting results is an empirical matter. However, there are indications that this sometimes can be the case. Suppose, for example, that state A activates several brain regions compared to a reference state B. Then, in addition to components reflecting baseline fluctuations in the two states, the estimate of gA in A will contain a component reflecting regional activations in relation to B. This implies that voxels will tend to get relatively overcorrected in A but not in B, biasing the distribution of voxel values, and increasing the risk of artefactual deactivations or decreased sensitivity for detecting activations. Consistent with this are observations of a greater than expected proportion of negative *Z*-score voxel values (Andersson 1997; Strother *et al.* 1995*a*, 1996). This problem should be less important if closely matched activation and reference states are used. However, with increasing activation differences between states, this may become a significant problem, reiterating the need for carefully designed experiments that includes active reference conditions.

It has been suggested that when the global signal is significantly confounded with the experimental paradigm, it may be preferable in some situations to omit global normalization entirely and examine absolute changes (Aguirre *et al.* 1998*b*). An alternative strategy is to use a more robust measure of gA, that is estimating gA independent of task-induced changes in rA to give a more accurate estimate of baseline fluctuations. One possibility is to estimate the background activity from white matter or by examining brain regions known to be relatively unaffected by the experimental paradigm (Frackowiak *et al.* 1997). An iterative solution to the latter suggestion has been proposed that successively eliminates voxels that indicate experimental effect from the set used to compute gA (Andersson 1997). These suggestions should work sufficiently well in the case of relative moderate or localized activations but may become problematic in the case of large spatially extended activations that are not matched in the reference condition.

(iii) *Global normalization in group studies and pharmacological neuroimaging*

Since the relationship between rA and gA is most likely nonlinear it may be inappropriate to use simple additive or proportional models to account for global effects in situations in which the gA varies over large or different ranges. For example, in pharmacological neuroimaging there may be several separated ranges of gA, corresponding to several drug-altered states, each demanding a different linear approximation. Comparing these states, either directly or via their interaction on a task–baseline comparison, effectively extrapolates the data beyond the different gA ranges observed to compare with a gA attained in neither condition. The validity of such comparisons is dependent on the appropriateness of the extrapolation, which in turn is determined by the accuracy of the local gA model over the larger gA range implied by the comparison. Most of the simple models described so far may not be comprehensive enough to approximate the actual behaviour of gA over large ranges. Using an inappropriate model may imply that the results of different comparisons simply highlight areas where the gA model is inappropriate; that is, real effects are confounded by apparent effects that are related to an ill-fitting model. In addition, if the drug greatly affects the gA, then it is likely that the global value and drug concentration covariates are highly correlated. Under such circumstances the effects are highly parallel, and if an ANCOVA approach is used, estimates of the global and drug effects will be unstable. Similar problems arise with the proportional scaling approach.

Other examples when simple models of global effects may be problematic are when patients with gross structural or functional abnormalities (significantly affecting the gA) are compared with normal controls, or, when ictal scans are compared with inter-ictal scans in epileptic patients. In the ictal scans there may be intense focal activity in the seizure area, contributing to the global mean measure of gA. Normalizing for the global effect by simple standard approaches, in effect interpreting changes relative to the global measure, will then tend to underestimate the magnitude of the ictal activation and other areas which have not changed their level of activity will appear deactivated, biasing the results. One approach here has been to use reference regions to estimate the baseline flow and normalize in relation to this measure. However, the results of such an approach are sensitive to the choice of reference region (McCrory & Ford 1991).

In summary, several different approaches to global normalization have been proposed and empirical comparisons indicate that there is little difference between them. Which model is most appropriate depends on the sources of global variability. For example, if the major part of the variability is due to differences in the amount of injected activity (PET), or drifts in gain (fMRI), then a proportional model may be appropriate. If, on the other hand, the variability is due to physiological variability (e.g. heart rate, pCO_2) the choice is more difficult. With proportional scaling, the noise is scaled, but is unaffected by the additive ANCOVA model. This has implications for the validity of subsequent statistical analysis. The gA can be confounded with the experimental induced effects if gA is estimated as a simple spatial average of rA. Hence, it may be necessary to use more robust methods to estimate the global effects independent of experimentally induced regional effects. Furthermore, when large or several different ranges of gA are expected, it may be necessary to first empirically study the behaviour of gA over the relevant ranges, in order to develop a more comprehensive model of gA. This is necessary in order to base comparisons on appropriately modelled gA. More comprehensive models of global effects may also be necessary to account for baseline fluctuations in fMRI data.

(b) *Univariate statistical models*

A single observation in FNI is generally a volume or image of voxels. When individual observations are not scalars, multivariate strategies are often the natural approach to statistical modelling. However, for unique reasons, the most common FNI analysis strategy is essentially a (massively) univariate approach, where identical univariate models are fitted at each voxel, so-called image regression. There are two main reasons for this. First, there are usually far more voxels than observations, which prevents the standard use of an arbitrary spatial covariance structure. (The covariance structure may be estimated, but the estimate is singular, precluding most multivariate analyses, which require the determinant of the estimated variance–covariance matrix.) Second, multivariate techniques characterize image volumes (observations) as a whole, and not as individual elements. Hence they do not directly address regionally specific questions, and commonly do not allow for any statistical inference at specific voxels. In this section, we review some common univariate statistical models used in FNI; the process of assigning significance to the results of the model, statistical inference, is reserved for the companion paper (Petersson *et al.*, following paper).

Univariate modelling is a well-developed area, whose methods form the core toolbox of statistics (e.g. Bickel & Docksum 1977; Edgington 1995; Good 1994; Winer *et al.* 1991). The enterprise consists of estimating the relationship between known effects (e.g. condition, time, subject, performance) and the data, then using the estimated effects to eliminate systematic variability from the data, leaving only residual variability, which is used to estimate the variance parameter(s) in the model. There is a trade-off between including all conceivable explanatory variables (effects), and parsimony, using the fewest number of effects to form an adequate model. Including too many effects can make the fit of the model too specific to the data at hand, degrading the generalizability of the results, while not including an effect that is present in the data generally will inflate the residual variability and hence bias the estimate of the model variance.

Suppose there is a parsimonious or true model, such that the residuals are independently identically distributed. Using an over-parameterized model reduces the power, since each additional effect modelled consumes degrees of freedom (d.f.), while the additional variability explained is limited. This reduction in power may be negligible when there are many d.f., but when there are few d.f. (e.g. in PET or random effects models), additional effects will, by reducing d.f., increase uncertainty in the estimate of the model variance. Hence with fewer than about 20 d.f., adding an effect to the model will reduce the certainty of the variance estimate and the significance of a given effect (unless including the additional effect significantly reduces the residual variability). Using instead a reduced model has two consequences. First, the unmodelled effects will appear in the residuals in a structured way, introducing dependencies among the residuals. This implies that the assumptions are not fulfilled, making standard inference invalid. Practically, the unmodelled effects usually, but not always, inflate the variance estimate. Second, there are more d.f. available for estimating the variance, implying increased confidence in the estimated variance, and therefore greater significance for a given change, with increased effect at lower d.f. So, on the one hand the variance increases, which would suggest decreased power, but on the other hand the d.f. increases, suggesting increased power. In general, all that can be said is that the inference is invalid, reiterating the importance of appropriate model selection. However, at high d.f. (e.g. fMRI and fixed effects models) the additional d.f. are unlikely to have much impact, and one might expect inference to be conservative in general. Note that the 'true' model may still be over-fitted (under-fitted) in that parameters may be included (excluded) just because there is evidence for (against) them in this particular data set, such that the model is larger (smaller) than need be when applied to a subsequent data set, degrading generalizability of the results.

Another consideration is the sources of random variation. The most basic models consider only one source, the residual error variation. These models are called fixed effects models as all the model effects are considered fixed (but unknown). A hierarchical mixed or random effects model has multiple sources of random variation. An effect is considered to be a random effect if it is the realization of a stochastic variable, that is, its values can be considered random draws from a population. For example, subjects that are studied are typically considered as randomly drawn from some population. In contrast, the levels of parametric manipulation of an experimental task are systematic, and not randomly drawn from a greater population. Random effects models will be further discussed below.

The general linear model (GLM) is a framework that encompasses all basic univariate fixed effects models, e.g. ANOVA, ANCOVA, and multiple regression models. (Note that while we abbreviate general linear models as GLM, it should be recognized that statisticians currently use GLM to refer to generalized linear models, a regression framework for non-normal, possibly discrete data (see McCullagh & Nelder 1989). Fortunately the GLM is a special case of the generalized linear model, so the ambiguity is not a source of problems.) In the GLM framework n observations (from a single voxel) are represented as a column vector of length n, $\boldsymbol{Y}$; the p effects and predictor variables are represented as p column vectors also of length n, forming an $n \times p$ matrix X called the design matrix. The fixed unknown parameters are represented as a column vector $\boldsymbol{\beta}$ of length p; the residual random error is written as the column vector $\boldsymbol{\epsilon}$ of length n. With the assumption of mean zero, independent and identically distributed error of magnitude σ^2, the concise representation of the GLM is

$$E(\boldsymbol{Y}) = X\boldsymbol{\beta} \text{ and } \mathrm{var}(\boldsymbol{Y}) = \sigma^2 I,$$

where I is the $n \times n$ identity matrix. Note that we have made no specific distributional assumptions; the usual normality assumption is only needed for statistical inference. Using only the general assumptions above, according to the Gauss–Markov theorem (Bickel & Docksum 1977), the linear unbiased estimates of $\boldsymbol{\beta}$ and σ^2 that are best in terms of minimizing squared estimation error are given by

$$\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{Y},$$
$$s^2 = 1/(n-p)(\boldsymbol{Y} - X\boldsymbol{b})'(\boldsymbol{Y} - X\boldsymbol{b}),$$

where $\boldsymbol{b}$ and s^2 are the estimate of the true unknown $\boldsymbol{\beta}$ and σ^2, respectively. The form of $\boldsymbol{b}$ can be found from algebraic manipulation of $\boldsymbol{Y}=X\boldsymbol{\beta}$. Note that $\boldsymbol{Y} - X\boldsymbol{b}$ is the residuals, so that the form of s^2 is just the mean squared residuals (the $n-p$ reflecting the dimensionality of the residuals that are left after fitting p effects). Tests of linear combinations of the parameters can be made under the normality assumption, which gives

$$\boldsymbol{Cb} \sim N(\boldsymbol{C\beta}, \boldsymbol{C}(X'X)^{-1}\boldsymbol{C}'),$$

where $\boldsymbol{C}$ is a row vector of length p, often called a contrast (see further Frackowiak *et al.* 1997). As an example of how this framework accounts for the basic models, consider the ubiquitous t-test for comparing the sample A with B. In this case, the X-matrix will consist of two columns composed of ones and zeros; for observations belonging to group A, the corresponding elements in the first column of X will be one, the elements of the second column zero; for observations belonging to group B, the corresponding elements in the first column of X will be zero, the second column one. Then applying the machinery above with $\boldsymbol{C} = [-1\ 1]$ will effect a t-test. Note that we had to make no special accommodations for unequal group sizes. Note, also, if we had three groups instead of two, we simply add another column to X and create contrasts $\boldsymbol{C}$ that will express tests of interest; this is equivalent to estimating contrasts for a fixed effect ANOVA. When the X-matrix consists of a column of ones and column containing a continuous covariate, a linear regression is affected, and the test for zero slope ($\boldsymbol{Cb}, \boldsymbol{C} = [0\ 1]$) is equivalent to a test of a zero correlation coefficient (Snedecor & Cochran 1967).

One application when the homogeneity of variance assumption may not be valid is the case of pharmacological neuroimaging. The pre- and the post-drug scans may show differences in terms of variance. For example, the neurophysiological response may be more stable than the post-drug response, implying that the variance appropriate for assessing the post-drug scans is underestimated. The GLM cannot directly account for heterogeneity of variance and including the pre- and post-drug scans into a comprehensive analysis introduces bias violating the common assumption of homogeneous variance. Similar problems may arise in FNI studies of patients. Instead other approaches have to be used.

(i) *Random versus fixed effects models—the scope of inference*

Fixed effects model parameters are linear combinations of the data (see equations above), and so can be estimated very efficiently. In contrast, random effect parameters are generally not linear functions of the data, and requiring nonlinear, iterative methods of estimation (e.g. expectation maximization, Laird & Ware (1982)). There is one notable special case where the standard GLM can be used with a random effects model, that is, when there is only one random effect, the design balanced, and the model is separable into individual subject models. In this case the data can be analysed in two stages. For the first stage, each subject is analysed individually, creating images of parameters of interest, one for each subject. For the second stage, the parameter images are analysed, with a one-sample t-test or, if there are two groups, with a two-sample t-test (Holmes & Friston 1998).

The scope of inference of fixed effects is for those values of the fixed effects, that is, if a subject effect is regarded as fixed then the inferences are for the cohort of subjects studied. Random effects models are used to make inferences about the population sampled. The only variance modelled in a fixed effects model of repeated observations on a group of subjects is the within-subject, within-condition variability (the variability from scan-to-scan of the same condition within an individual). This includes measurement variability, confounded with other physiological, physical and cognitive effects. The random effects model additionally accounts for the between-subject variability. This implies that a fixed effect analysis can declare a significant effect in a set of subjects, while a random effects analysis can declare an effect significant for the population sampled. This is especially important for group comparisons.

While it might seem that one would only want to perform population inference, there are several critical issues to consider. First, the ability to randomly sample a broad population of human subjects is often practically impossible. When non-random samples are used, the random effects models will have questionable validity. The lack of correct random sampling is an argument for the use of fixed effects models. It has been argued that in the context of non-random sampling, fixed effect (non-parametric) models should be used in conjunction with non-statistical generalization, as a natural part of science (Edgington 1995). Note that even if the non-random sample can be viewed as a 'representative' sample from some population of interest, defining the correct population, if it exists, seems difficult. Hence the results from random effects models of non-random samples are inherently difficult to interpret. Second, if it can be argued that a population is randomly sampled, random effects models have the additional assumption of normality of the population sampled. This is often an assumption that is difficult to verify, in particular for a broad population of human subjects. Last, the second-level analysis of the random effects model (Holmes *et al.* 1998) has degrees of freedom determined by the number of subjects. This implies that the number of subjects included in a random effects study is central for sensitivity and statistical power. This has implications for the use of random field theory methods for statistical inference which are best at large d.f. (cf. Petersson *et al.* 1999; Worsley *et al.* 1996). In the situation of low d.f., the use of the smooth random field approach to statistical inference can be problematic and lack sensitivity for voxel level inference. However, alternative approaches are available, for example, the use of variance smoothing and pseudo *t*-tests in combination with non-parametric inference (Holmes *et al.* 1996); this is further discussed in Petersson *et al.* (following paper).

It should be noted that group comparisons are not served well by fixed effects models, since a between-group difference must be compared with a measure of between-group variation, that is, a between-subject estimate of subject variability within group. A fixed effects model with repeated measures only accounts for within-subject variability and measurement error. For fMRI the within-subject variability is most certainly smaller than the between-subject variability, and hence significance of between-group differences will be overestimated. It follows that group comparisons aiming at inference about population differences are crucially dependent on the use of random effects models.

(ii) *Statistical challenges for the analysis of fMRI data*

In this subsection, three principal statistical challenges introduced by fMRI will be addressed. First, the problem to identify a function of time that predicts the haemodynamic response, the haemodynamic response function (HRF). Second, methods to account for the temporal autocorrelation of fMRI data, and third, modelling of slow variations drift in fMRI data. The first issue relates to the fact that the BOLD response is delayed and dispersed in time, the second that the residual error of fMRI time-series are dependent (Aguirre *et al.* 1997; Boynton *et al.* 1996; Friston *et al.* 1994*a*, 1995; Purdon & Weisskoff 1998; Weisskoff *et al.* 1993; Worsley & Friston 1995; Zarahn *et al.* 1997), and the last issue regards slow changes in the fMRI signal whose sources are unknown, but whose occurrence is common.

The BOLD contrast is the result of an interaction between neuronal activity, oxygen extraction, blood flow and blood volume (Buxton *et al.* 1998; Ogawa *et al.* 1998; Vazquez & Noll 1998). While the exact relationship between the quantity of interest, the neuronal activity, and the BOLD response is unknown, the qualitative character of the HRF is well known. Given a discrete on–off stimulus, the BOLD response is delayed and dispersed in time; from time of stimulus start, there is approximately a 2–3 s delay until an appreciable response (some laboratories report an immediate, small-magnitude negative response (Hu *et al.* 1997)), a maximum is reached after approximately 4–7 s, and after stimulus cessation there is a delay before the signal falls and it generally falls to negative (the 'undershoot') before returning to baseline. The exact character, especially the delay until response, is variable both between subjects, and across the brain within a subject (Aguirre *et al.* 1998*c*).

The earliest approaches used *t*-tests and simple tests on the correlation coefficient (Bandettini *et al.* 1993) to determine significant changes. *F*-test methods comparing the power at the paradigm frequency with the average power have also been used (Bullmore *et al.* 1996; Lange & Zeger 1997), with the advantage that they are insensitive to the exact phase or delay of the response. However, they suffer when the haemodynamic response departs significantly from a sinusoidal response, since they are equivalent to regressing the data on a sine and cosine with frequency identical to the paradigm. Recently, a χ^2 random field approach for detecting sinusoidal signals has been described (Worsley 1999; Worsley *et al.* 1997*a*).

Many recent approaches have focused on treating the BOLD response as a linear time-invariant (LTI) system (Jain 1989; Oppenheim & Schafer 1989). The linearity implies that the predicted haemodynamic response is the convolution of a fixed impulse response function (IRF or HRF) with the waveform of the experimental paradigm. The time invariance implies that the IRF is stationary, that is, independent of time and prior responses. This approach is attractive as the IRF completely characterizes the system (Oppenheim & Schafer 1989). Usually the IRF is assumed to be known and the statistical modelling reduces to estimating the amplitude of the response. We begin by giving an overview of models for the IRF and then review the evidence for the BOLD response as an LTI system.

Use of the *t*-tests corresponds to a Dirac-delta HRF. Early work used functions corresponding to common probability distributions, though there is no theoretical motivation for this, just a qualitative similarity between the distributions and the HRF. Friston *et al.* (1994*a*) first used a Poisson probability mass function; others have used the slightly more flexible gamma density (Cohen 1997; Lange & Zeger 1997). Friston *et al.* (1995) later used the Gaussian density not so much as an IRF but as part of a filtering approach motivated by the matched filter theorem (Rosenfeld & Kak 1982). Others have also used the Gaussian kernel (Rajapakse *et al.* 1998). An alternative approach is to specify not a single IRF but a family or basis of IRFs; this will potentially enhance the fit of a model, though possibly at the expense of the interpretability of the fit. A third approach is to first estimate the IRF on an independent data set.

(iii) *Models of the haemodynamic response function and linear time invariance*

It is generally accepted that the transformation from the experimentally induced signal via the neural response and the BOLD response generated to the fMRI signal (i.e. input→neural activity→BOLD response→fMRI measurement) is not a perfect LTI (Boynton *et al.* 1996; Dale & Buckner 1997; Friston *et al.* 1998; Rosen *et al.* 1998; Vazquez & Noll 1998). It is an open question whether it is well approximated as an LTI system or not. The central issues are identifying the conditions under which is it well approximated as linear and time invariant, and characterizing the forms and sources of the nonlinearities. The primary visual system seems satisfactorily approximated as an LTI, over a range of stimulus duration (5–22.5 s) and contrasts (0–100%) (Boynton *et al.* 1996). A recent investigation of this question, also in the primary visual system (Vazquez & Noll 1998), found evidence indicating that stimuli shorter than 4 s and less than 40% contrast yielded significant nonlinear responses. So, the transformation from neural activity to the BOLD response may (under suitable conditions) be well approximated as an LTI system in primary sensory (visual) areas, when driven by simple sensory input. Whether this generalizes to other brain regions and experimental paradigms engaging higher cognitive function is an open question (for an interesting example see Buckner *et al.* (1998*a*,*b*)). It is presently hypothesized that higher cognitive functions are subserved by nonlinear network interactions (Amit 1989; Arbib 1995; Koch & Davis 1994; Rumelhart & McClelland 1986; Schuster 1991). In addition, learning, memory and the capacity of the brain to adapt in a non-stationary environment indicate that the transformation input→neural activity is neither linear nor time invariant. This represents challenging problems for the analysis of fMRI data.

Statistical modelling under the assumption of linearity is straightforward, as the predicted haemodynamic response may be used as a covariate in a multiple regression. A distinction must be made between the assumption of linearity of the BOLD response and linearity of the statistical models, which simply means the data is model as an (unknown) linear combination of (known) predictors. For example, Friston and colleagues (Friston & Buechel 1998; Friston *et al.* 1998) model nonlinearities in the HRF with a linear model using a truncated Volterra series expansion. Another example is that the variable delay of the HRF, while not expressible in a statistical linear model, can be approximated for small delays by including the temporal derivative of the HRF in the model (Friston *et al.* 1998).

In general, arbitrary HRF models that are not linear in their parameters must be fitted with nonlinear methods. Recently models that do not make the assumption that the BOLD response is an LTI system have been described (Frank *et al.* 1998; Genovese 1997; Genovese & Sweeney 1998; Lange & Zeger 1997). These approaches can yield richer information on the form of the HRF. Frank *et al.* (1998) parameterize the response amplitude as a proportion of baseline to demonstrate a nonlinear model. Genovese also uses proportional response amplitudes, but elaborately parameterizes the shape of the HRF with eight parameters. In addition, recent work has been focusing on directly modelling blood flow, blood volume and blood oxygenation (Buxton & Frank 1997; Buxton *et al.* 1998; Davis *et al.* 1994); these approaches may improve the understanding of the BOLD response and hold the potential for quantification in fMRI.

(iv) *Temporal autocorrelation and dependent residuals*

The development of the described GLM above was based on the assumption that the residual errors were mean zero, had constant variance and were independent. It has been widely observed that fMRI time-series display temporal autocorrelation (Aguirre *et al.* 1997; Boynton *et al.* 1996; Friston *et al.* 1994*a*, 1995; Purdon & Weisskoff 1998; Weisskoff *et al.* 1993; Zarahn *et al.* 1997). While autocorrelation does not bias estimates of the effects (e.g. mean difference), it does bias estimates of variability, thus affecting the significance of effects and altering the false positive rates. Here, we review two general approaches to coping with correlated errors, and how they have been applied in the literature.

There are two linear model approaches to autocorrelation, generalized least squares (GLS) and ordinary least squares (OLS) with adjustment for correlated errors. GLS in essence de-correlates, or 'whitens', the data and then applies OLS. The second approach uses OLS on the correlated data (in violation of its assumptions) and then approximates the null distributions of test statistics by adjusting the conventional degrees of freedom to the so-called effective degrees of freedom. The GLS approach is the optimal linear estimator (minimum variance in the class of unbiased estimators), and corresponds to the maximum likelihood estimator for Gaussian data, but is sensitive to the correct form of the covariance matrix. The latter approach is not optimal, but it is more robust to mispecifications of the covariance structure (Worsley & Friston 1995; Worsley *et al.* 1997*c*). However, there are special cases when the OLS-then-correct approach is optimal, for example, when the regressors are sinusoidal (Worsley & Friston 1995). Time-series methods provide alternative means to account for autocorrelation (Chatfield 1996; Wei 1990), though these fall into the former GLS category as they tend to use whitening to provide unbiased variance estimates.

To our knowledge, the GLS approach has not been used directly, but rather through the machinery of time-series methods. Bullmore *et al.* (1996) performed a thorough statistical analysis of a collection of voxels and found that a first-order autoregressive model (AR(1)) seemed appropriate. Locascio *et al.* (1997) applied general autoregressive moving average (ARMA) modelling at a voxel-by-voxel basis. Prudon & Weisskoff (1998) suggest the use of an AR(1) plus white noise model.

The OLS model with adjusted null distribution was first presented to the FNI community by Worsley & Friston (Friston *et al.* 1995; Worsley & Friston 1995). In these works, a temporal smoothing filter was applied referring to the matched filter theorem (Rosenfeld & Kak 1982). It was then assumed that the intrinsic autocorrelation was negligible relative to the temporal smoothing kernel, such that the form of the covariance matrix can be approximated by that determined by the temporal smoothing kernel. The temporal smoothing of fMRI time-series acts as a low-pass filter, discarding high-frequency effects. An alternative approach is to empirically model the intrinsic temporal autocorrelation, for example, using a $1/f$ form in the square root of the power spectrum, estimated from null data and pooled across the brain (Aguirre *et al.* 1997; Zarahn *et al.* 1997). It was found that the convolution with an empirically determined HRF, and an assumed $1/f$ intrinsic autocorrelation, produced the most accurate false-positive rates with smoothed data (Aguirre *et al.* 1997; Zarahn *et al.* 1997). Significantly, the use of the $1/f$ intrinsic autocorrelation function with no HRF convolution did not control the false-positive rate (using the local maximum statistic), a result explained by spatially inhomogeneous autocorrelation, that is, the degree of autocorrelation varied across the brain. This inhomogeneity has also been noted by others (Purdon *et al.* 1998). Finally, the temporal smoothing may not be beneficial for event-related fMRI data. Since the BOLD HRF has high-frequency components, temporal smoothing discards the high-frequency components of the signal, and this portion of the signal may convey specific localizing information (Paradis *et al.* 1998).

Finally, the source of slow variations in fMRI time-series is not well understood, but represents a large source of variability (Genovese 1997). The likely causes include slow subject motion artefacts, biorhythms (respiratory and cardiac pulsation artefacts) and physiological deformation of the brain. These drifts have been modelled in a variety of ways, including linear slopes (Fitzgerald 1996), exponentials (Vazquez & Noll 1998), and discrete cosine bases (Holmes *et al.* 1997). The last approach essentially implements a high-pass filter as a part of the model. There have also been attempts to directly correct physiological effects, such as respiration and the cardiac cycle (Hu *et al.* 1995).

We note that most approaches described use the same model at every voxel. For independent data this amounts to using the same regressors at every voxel, but for temporally correlated data it additionally means using the same autocorrelation function model at every voxel. An alternative is to use model selection techniques and consider ARMA models of arbitrary order at each voxel (Locascio *et al.* 1997). Locascio *et al.* (1997) found that some voxels required higher order ARMA terms, while others passed a test for white noise without whitening. New approaches may leave behind the computational convenience of image regression in exchange for more comprehensive models improving the quality of the statistical modelling.

(c) *Bayesian models for functional neuroimaging data*

Bayesian methods can be used to incorporate prior knowledge in a systematic fashion (Bretthorst 1990*a*,*b*; Descombes *et al.* 1998; Frank *et al.* 1998; Genovese 1997; Genovese & Sweeney 1998; Holmes & Ford 1993). In general, Bayesian methods can be used in two different ways; either for Bayesian estimation, or as a method for stochastic regularization (cf. Petersson *et al.*, following paper). Bayesian estimation is naturally biased towards the prior information used (Billingsley 1995) and can be regarded as enforcing soft constraints on parameter values. However, the Bayesian approach provides a coherent framework for statistical analysis (Box & Tiao 1992; Lee 1997), of which the classical approach can be viewed as a special case. The bulk of statistical tools are labelled frequentist, and their justification is built on repeated sampling from a theoretical population. The observed data are a realization of a random process and the parameters of interest are fixed and unknown quantities. In contrast, the Bayesian approach regards the parameters as random variables as well. Before an experiment is performed, the parameters have an *a priori* distribution, called the prior. After the experiment is performed the prior is updated to give the posterior distribution, which reflects the information gained from the data. Questions about the parameters are addressed via the posterior distribution: for example, the probability that the response amplitude is greater than zero is simply the integral of the posterior from zero to infinity.

While regarding unknown, unobservable quantities (the parameters) as random may seem intuitively reasonable, it is an area of contention in statistics. The main criticism is the potential subjective element inherent in assigning prior distributions. Despite this issue, and the ubiquity of classical statistical tools, Bayesian methods are gaining popularity. This growth is due in large part to increased access to the necessary computing power. Bayesian methods can also have a greater intuitive appeal than frequentist methods. For example, the frequentist 95% confidence interval is often misinterpreted as the range in which the parameter lies with probability 0.95; in fact, this is the correct interpretation for the Bayesian credible interval. The correct interpretation for a confidence interval is that with repeated experiments with identical conditions, 95% of the confidence intervals created will contain the true, fixed, unknown parameter.

Bayesian methods are also understood as a formalized version of regularized optimization, or penalized likelihood, where the standard maximum-likelihood approach is biased towards favourable parameter values. From this approach, one can understand the Bayesian and frequentist perspectives together. With increasing number of observations, the prior becomes less and less important, and hence the Bayesian results will converge to the likelihood-only frequentist result. One can also think of increasing the spread of the prior until it is flat, at which point the Bayesian machinery will in general produce the frequentist result.

Frank *et al.* (1998) present a good introduction to the general Bayesian framework for fMRI, calling it a 'probabilistic analysis', considering illustrative examples. A key point of this work is that if one considers more general models than permitted by regression, a much richer variety of questions can be answered. For example, a haemodynamic delay with fixed HRF could be directly modelled. For prior distributions they constrain themselves to so called non-informative priors, which are invariant under natural transformations of the parameter (Jaynes 1968).

Genovese (1997; Genovese & Sweeney 1998) presents a Bayesian model which considers a very general HRF. Defined as a sum of polynomial bells, the HRF smoothly rises to, and falls from, a flat plateau. In addition, the shape and response amplitude is separately parameterized. For example, the shape is described by at least four parameters: the delay until initial rise; time of plateau start; delay until fall from plateau; and time of return to baseline (delays are relative to stimulus onset and cessation, respectively). Further, the model includes baseline drifts modelled with cubic smoothing splines. This represents an alternative to the use of high-pass filters to reduce the influence of low-frequency noise. In contrast to Frank *et al.* (1998), this work also uses informative priors. For example, the response amplitude prior is the sum of an impulse at zero amplitude and a gamma density, reflecting the expectation that most voxels are not activated (zero amplitude) by the stimulus, and that, if activated, the expected range of positive activation is likely to be *ca.* 1–5% (as described by the parameters of the gamma density). While both these works allude to the need to account for autocorrelation in the time-series, neither has addressed this in the models presented.

Finally, it should be pointed out that there are alternatives to the Bayesian approach for incorporating prior knowledge. For example, pre-processing, various regularization approaches and the use of appropriate subspaces spanned by sets of basis functions may be viewed as ways of incorporating prior knowledge. For example, a set of temporal basis functions can offer some flexibility in the modelling of a (partially) unknown response form (e.g. the HRF) and at the same time incorporate prior knowledge of the response form. Fundamentally, the whole model-building process and the choice of a statistical model reflect prior knowledge.

(d) *Multivariate statistical models*

FNI data are inherently multivariate and multivariate approaches are natural alternatives for data analysis. Three different multivariate approaches (PCA–MANCOVA–CVA, PLS, and MLM) will be reviewed in this section. Issues discussed include linear dimensionality reduction, the need to characterize the multivariate response, some non-parametric approaches to multivariate data, the generalized *S*-test and dimensionality estimation. Since it is difficult to clearly separate the modelling aspects from the inferential aspects in the multivariate approaches described, some of the inferential aspects are naturally commented upon in this section. A more general picture of statistical inference is given in the companion paper (Petersson *et al.*, following paper).

As already noted, the high dimensionality of FNI data relative to the number of observations often excludes straightforward applications of standard multivariate statistics, since the estimated covariance structure of the data will be singular. In the case of a few pre-selected ROIs and large enough number of measurements, standard multivariate statistics are applicable. However, several multivariate voxel approaches have been adapted to the context of FNI. These approaches often aim at characterizing the overall distributed pattern of experimentally induced changes in brain activity regardless of location. In general, this precludes statistical inference about regional specific effects; that is, the regional structure of the experimental effect can only be interpreted descriptively. This lack of localizing power in multivariate approaches implies that univariate techniques are complementary to the multivariate. Some of the suggested multivariate techniques (for an overview, see Worsley 1997*b*) are closely related to CVA (Chatfield & Collins 1980). CVA is a standard multivariate technique for selecting the linear compound of the multivariate response, which demonstrates the greatest inconsistency between the experimentally induced effect and the null hypothesis (Chatfield & Collins 1980). The CVA can be used to characterize the signal present in data when the null hypothesis has been rejected. The CVA therefore shares some features with the descriptive non-inferential methods. For example, it has been pointed out that CVA informally may be viewed as a PCA that accounts for error effects (Friston *et al.* 1996*b*).

(i) *Optimal linear dimensionality reduction, MANCOVA and CVA*

One of the earliest voxel-based multivariate strategies for PET data analysis suggested that the problem with the high dimensionality of FNI data may be handled by optimal linear dimensionality reduction (Friston *et al.* 1993, 1996*b*). In this approach, dimensionality reduction is achieved by applying a PCA to the adjusted PET data (mean corrected data adjusted for confounding effects, e.g. global and block effects) and keeping only the most important eigenimages (e.g. the PCs with eigenvalues greater than unity). This dimensionality reduction represents the optimal linear approximation in a mean square sense. Another way of looking at this is to consider the linear transformation as representing that projection of data on a subspace of given dimensionality which preserves the maximum amount of observed variability.

The outcome of the PCA is sensitive to the pre-processing performed on the primary data (e.g. image smoothing, fitting and adjusting for confounding effects or standardization of voxel-per-voxel variance). For example, unless the components of interest are first orthogonalized to components of no interest, part of the components of no interest may be left in the adjusted data. In addition, the result of the PCA may be sensitive to outliers in the data, if the particular experimental effects are relatively weak. The dimensionality reducing PCA also alters the multivariate distribution of data. If the observations are temporally uncorrelated, as in PET, then the inference procedure described by Friston *et al.* (1996*b*) is still valid. However, if the observations are temporally correlated, this is no longer the case (Worsley *et al.* 1997*c*).

The result of the PCA is a set of pairs, each pair consisting of a time component (component scores or temporal pattern) and a space component (eigenimage or spatial pattern, cf. § 2(a)). Taking the N first pairs reduces the dimensionality of the data to N. The component scores represents the temporal expression over observed time points or scans of the corresponding eigenimage. In effect, a transformed data set is created; it has the same number of observations as the original data set but now each observation only consists of N elements, instead of the number of voxels (i.e. an N-dimensional time-series is generated). This multivariate data set is then modelled with a multivariate GLM (MANCOVA). Statistical inference is achieved using the Wilks' Λ-statistic, and a χ^2-distributional approximation for the log-transformed Wilks' statistic (Chatfield & Collins 1980). This is valid to the extent that the residuals of the multivariate GLM are distributed identically and independently multivariate normal. The Wilks' test is comparable to an F-test in the univariate case, that is, the Wilks' statistic is a test of whether the model as a whole explains a significant part of the observed variability. In other words, the overall significance of the multivariate GLM, allowing for the covariates of no interest, is assessed. The Wilks' statistic may in this way also be used as a test of specific effects in a way analogous to the use of an F-statistic under the extra sum of squares principle (Draper & Smith 1981).

When the null hypothesis is rejected this indicates that there is an experimentally induced signal present in data, but without giving a detailed characterization of the signal. It is therefore necessary to characterize the discrepancies between the experimental effects and the null hypothesis. Friston *et al.* (1996*b*) suggest the use of a CVA to this aim. The multivariate experimental effect is characterized as so-called canonical variates (orthogonal or uncorrelated by construction, Chatfield & Collins (1980)). Since this characterization is performed on transformed data (i.e. transformed into the space spanned by the N eigenimages) it is necessary to transform the characterized effects back into image space, where the effects are represented as canonical images (i.e. canonical variate→canonical image). The effects described by the canonical images need to be interpreted post hoc, commonly guided by the form of the temporal response profile of the corresponding canonical variate. Any local structure in the canonical image has to be interpreted descriptively. As with PCA, there is nothing inherent in the CVA approach to guarantee a straightforward interpretation of the results, unless the signal is naturally divided into several orthogonal components. In general, it is possible that a canonical variate–image represents a mixture of effects. In principle, characterization of the experimentally induced response necessitates an estimate of its dimensionality; that is, the number of canonical variates required to describe the experimental effect. With each canonical variate there is associated a canonical value that may be interpreted as a variance ratio between the experimental effect as expressed by the corresponding canonical variate in relation to its expression in the residual. Friston *et al.* (1996*b*) suggest that the canonical value can be interpreted heuristically as an F-value and used as an indication of the more important effects and their corresponding canonical images.

An interesting application of the PCA–MANCOVA–CVA approach is the possibility to investigate different models for data. If the models can be organized in a hierarchical tree-structure, the Wilks' statistic may be used for model selection purposes in a way analogous with a sequential hierarchical F-test (Holmes 1994). This allows for a comprehensive model selection procedure, relatively independent of any assumptions on the spatial covariance structure of the PET data (Friston *et al.* 1996*b*).

(ii) *PLS approach and bootstrapping*

A different technique to analyse spatial patterns is the PLS approach (Jöreskog & Wold 1982; Wold 1985), described and applied to PET data by McIntosh *et al.* (1996*a*). Hypotheses are often tested in the form of specific contrasts related to the experimental design, and are used in PLS to study multivariate effects induced by experimental manipulation. Briefly, the so-called cross-correlation matrix between a matrix of contrasts (e.g. representing differences between states) and the data matrix is generated. The PLS approach avoids dimensionality reduction by directly studying the cross-correlation matrix. An SVD is then performed on the cross-correlation matrix generating singular images, the corresponding singular values, and effect profiles. Subject scores are generated by the inner product of a given singular image and the subject's brain images. The subject score reflects the degree to which the singular image is expressed in a given brain image, that is, how parallel the singular and the brain image are. The effect profiles reflect the degree a given singular image is related to (or expresses) the contrasts of interest. Similar to the canonical images, the singular images have to be interpreted post hoc, often guided by the form of the effect profiles or the variation of subject scores over states (McIntosh *et al.* 1996*a*). As with PCs or canonical images, a singular image may represent a mixture of structured effects.

The PLS approach entails statistical test procedures that requires careful interpretation. The singular images and their effect profiles are assessed in an indirect way using a permutation test. Specifically, the explanatory significance represented by the matrix of contrasts in relation to the subject scores is assessed. This means that the subject scores are regressed on the matrix of contrasts and a test statistic, R^2 (representing the proportion of the overall observed variance explained by the model) is used to assess the overall significance of the effects represented in the matrix of contrasts. The significance of the R^2-statistic is assessed using a permutation test (Edgington 1995; Good 1994). The R^2-test is generally not a test on any specific effect represented by the singular images or the subject scores, but tests whether the matrix of contrasts as a whole explains a significant part of the observed variability in the subject scores. However, since the scores are derived from the singular image and the contrasts are the same as those used in the original analysis, this variant of the PLS approach may in an informal sense be viewed as assessing the link between the two elements. The permuting and redoing of the PLS for each permutation in effect assesses how many other pairings of contrasts and brain images produces as good a link as the original pairing, measured as the proportion of the overall variance in the subject scores explained by the matrix of contrasts. An alternative to this PLS scheme has been proposed (Grady *et al.* 1998), in which the singular values are sequentially assessed by permutation tests. It should be noted that permutation tests are valid only if the data fulfil the assumption of exchangeability under the null hypothesis (cf. the non-parametric section in Petersson *et al*. (following paper)).

As with the PCA–MANCOVA–CVA approach, it is difficult to use the significance assessment as a basis for more detailed empirical conclusions in any direct sense, for example, interpreting specific effects characterized by the contrasts used in the analysis or in terms of regional specificity. Instead, the singular images are often arbitrarily thresholded and the local structure in these descriptively interpreted. It has been suggested that this aspect of the PLS approach can be made more rigorous by generating confidence intervals corresponding to the voxel values of a given singular image, using non-parametric bootstrap estimates (Efron & Tibshirani 1986). Under the assumption of independent and identically distributed observations, the bootstrap procedure is asymptotically exact (for a precise definition of exactness, see Petersson *et al.* (following paper)). This implies that for a sufficiently large sample size the bootstrap estimates can be considered approximations of the exact *p*-value or confidence interval.

The PLS approach is, like PCA, sensitive to the type of pre-processing performed on the data. The singular images depend on what effects are incorporated in the matrix of contrasts and what pre-processing has been performed on data. This implies that the singular images depend on the context in which they are computed. Even if there is a hierarchical or nested relationship between different matrices of contrasts (e.g. if other contrasts representing subject or repetition effects are included or not), there may not always be a simple relation between the corresponding sets (or subsets) of singular images (McIntosh *et al.* 1996*a*). McIntosh *et al.* (1996*a*) suggest that the results from different matrices of contrasts may be compared to see if results are stable or not. If the result is judged unstable, then it is suggested that these effects are adjusted for during pre-processing.

The description above has focused on the use of PLS to study the cross-correlation between a matrix of contrasts and PET data. The approach has been extended to study the cross-correlation between a matrix of behavioural covariates and PET data, as well as the cross-correlation between different PET data sets (McIntosh *et al.* 1998*a*,*b*). It should be noted that the PLS approach (as described) is generally not invariant under arbitrary linear transformations of the matrix of contrasts. For example, if the matrix of contrasts is scaled differently the results will be different. The same may be the case if different predictors in the same predictor space are chosen (Worsley *et al.* 1997*c*). As a solution to these problems, Worsley *et al.* (1997*c*) suggest an orthonormalized PLS approach. (Note that PLS with orthonormal matrices of contrasts is a special case of the orthonormalized PLS approach.) Finally, it may also be noted that it is not straightforward to extend the use of non-parametric tests to temporally correlated fMRI data (cf. the section on non-parametric inference in Petersson *et al.* (following paper)).

(iii) *General MLMs and the generalized S-test*

Worsley *et al.* (1995) described a test for comparing distributed non-focal differences between two states (cf. the omnibus tests section in Petersson *et al.* (following paper)). The approach described did not attempt to characterize the experimental response. Recently, this approach was generalized to arbitrary GLMs and a strategy for characterizing the departure from the null hypothesis, that is, the experimentally induced response was described (Worsley *et al.* 1997*c*). This general MLM approach has similarities with the MANCOVA–CVA approach. The MLM approach uses a test for distributed change in combination with a PCA of the normalized effects. This approach can also be used to analyse temporally correlated data and is thus applicable to fMRI data.

In brief, given a GLM, the voxel *F*-statistic image is generated and the *F*-statistic averaged over all voxels in the search volume, representing a generalization of the *S*-statistic described by Worsley *et al.* (1995). In this way the MLM approach avoids the need for explicit dimensionality reduction. Rather than filter the data with a pre-whitening filter in combination with maximum-likelihood estimation, this approach uses the more robust OLS estimation procedure (cf. §3(b)(iv)). The generalized *S*-statistic is approximately *F*-distributed with estimable effective degrees of freedom (d.f. > 10 is required for the approximation to be sufficiently accurate). The *S*-statistic is used to assess the null hypothesis that no signal explained by the model is present in the search volume. If the null hypothesis is rejected, indicating that experimentally induced changes are present in the search volume, the spatio-temporal response needs to be characterized in greater detail. Note that the *F*-statistic image may alternatively be submitted to a univariate approach and searched for activations using the unified *p*-value for *F*-fields (Worsley *et al.* 1996).

As indicated above, CVA is a technique for finding the linear compounds of multivariate response which demonstrates the greatest inconsistency between the experimentally induced effect and the null hypothesis (Chatfield & Collins 1980). In this context, Worsley *et al.* (1997*c*) suggest that a PCA of the normalized effects may be used to find the linear combination of predictors that optimally describes the distributed response. Specifically, a PCA is performed on the mean sums of squares and cross-products matrix, and the dimensionality of the multivariate response is estimated. The dimensionality is estimated using a test similar to the Lawley–Hotelling trace, sequentially testing the significance of the PCs. The sequential testing procedure presupposes that the signal is large enough to be detected. If the signal is weak it may remain undetected, or perhaps the signal PCs may become contaminated by lower order sample PCs (noise). Simulations may be used to investigate these issues further (K. Worsley, personal communication). Furthermore, the MLM approach assumes that the Gaussian random field theory is applicable and that the spatial autocorrelation function can be accurately approximated by a Gaussian point spread function (cf. the section on random field theory in Petersson *et al.* (following paper)). The MLM approach is validated on simulated data, indicating that the method works well when data fulfil the assumptions made. There remains the question of how robust the method is to departures from these assumptions.

(e) *Functional connectivity and network analysis*

The statistical models so far described have been used to investigate the relationship between the experimental paradigm and the changes induced in brain activity. These approaches study changes in regional activity and how these changes covary with specific external experimental manipulations (i.e. changes over which experimental control is exerted). From the perspectives of theoretical modelling (Amit 1989; Hertz *et al.* 1991; Rumelhart & McClelland 1986), cognitive psychology (Horgan & Tienson 1996; Macdonald & Macdonald 1995), and cognitive neuroscience (Mesulam 1990, 1998), as well as from lesion observations (Squire 1992; Zola-Morgan & Squire 1993) and FNI data (Friston 1994; Friston *et al.* 1993; Gonzalez-Lima & McIntosh 1994; Horwitz *et al.* 1984), it has been suggested that higher cognitive functions are the result of the network interactions between different brain regions. This indicates that the understanding of different brain functions may benefit from analysing the interactions between brain regions. Based on the idea that brain regions that constitute components of a functional network will have activities that are correlated this is often done by studying the covariance pattern observed in FNI data. The previously described models and approaches, which are used to analyse the relative changes in regional activity, can be used to identify the components of such large-scale cognitive networks.

Functional and effective connectivity was originally defined in the context of electrophysiology (Aertsen *et al.* 1989; Aertsen & Preissl 1991) and these concepts were introduced into FNI with a modified connotation by Friston (1994). Functional connectivity was defined as the observed correlations over time between different brain areas, independent of the sources of these correlations, and effective connectivity refers explicitly to the influence that one neural system exerts over another (Friston 1994). In this section, we will discuss some issues related to covariance sources, different covariance approaches (partial correlation coefficients and covariance fields) and network analysis of effective connectivity (structural equations modelling).

(i) *Covariance sources*

Two different ways to estimate the covariances within a cognitive state have been described: over time within subject (Buechel & Friston 1997), and over subjects (Horwitz *et al.* 1995). The basic hypothesis is that the intrinsic variability in the neural response of a cognitive state will emulate the relevant functional interactions and that these interactions will be reflected in the covariance structure. It should also be noted that the sources of within-state interregional covariances are beyond experimental control.

Several sources of interregional covariances have been proposed (Horwitz *et al.* 1992*b*) and the actual sources of the observed covariances are largely unknown. However, several of the proposed sources of observed covariances may give rise to spurious correlations that are necessarily confounded with correlations arising because of effective connectivity (e.g. adaptation, fatigue, attentional drift, anomalous task response, however, cf. Horwitz *et al.* (1992*b*)). Obvious potential confounders in the study of interregional covariances are global effects. Variability in any global signal will introduce correlations that most often are of no interest. This reiterates the importance of adequate global normalization. It was suggested that the problem of global effects may be discounted by using partial correlation coefficients (Horwitz *et al.* 1984) or proportional scaling, both of which yield similar results (Horwitz & Rapoport 1988). However, it has been indicated that accounting for global effects with simple approaches such as ANCOVA, proportional scaling or the use of partial correlation coefficients may not always yield appropriate results. Ford (1986) pointed out that these approaches could yield highly biased results introducing spurious correlations. In this context, it was argued that the spurious correlations most often are small and their presence can (to some extent) be tested for (Horwitz & Rapoport 1988). However, the presence of spurious correlations will bias the results unless properly removed or accounted for.

If the covariances are estimated over subjects, it is necessary to assume that the subjects implement a sufficiently similar functional organization. In this case, speaking informally, the covariance structure may reflect an average common functional organization. However, the functional organization can vary substantially between subjects, that is, the covariance structure of a subject may or may not be related to a common functional organization (Friston 1995). In PET studies the number of intra-subject observations is limited (restricted by radiation exposure), so, in order to increase sensitivity data is often pooled over subjects. With fMRI, it is possible to study functional and effective connectivity in single subjects (Buechel & Friston 1997). One advantage of performing several single subject studies is that this can give an indication of the generalizability of the results. Within- and between-subjects studies are complementary, and both group and single-subject analyses may be confounded by the factors described above.

(ii) *Partial correlations and covariance fields*

Several different approaches have been proposed to study interregional covariances. Early attempts studied the matrix of partial correlation coefficients between pairs of pre-selected ROIs (Horwitz *et al.* 1984). Subsequent approaches concentrated on selecting a reference region (Horwitz *et al.* 1992*a*) or a reference voxel (Horwitz *et al.* 1995), and then studying the correlations with the rest of the brain or a set of pre-selected regions. In both cases, the multiple comparisons problem needs to be addressed (cf. the multiple comparisons section in Petersson *et al.* (following paper)). For the reference region or voxel approach known results for *t*-fields may be applied (Worsley 1994; Worsley *et al.* 1996, 1998). In the case of voxel-by-voxel correlations, a recently developed theory for so-called autocorrelation fields may be used to handle the multiple comparisons problem (Cao & Worsley 1999; Worsley *et al.* 1998). These new theoretical advances in random field theory also include cross-correlation and homologous correlation fields (Cao & Worsley 1999).

(iii) *Structural equation modelling*

To characterize effective connectivity in FNI data a network approach based on structural equation modelling

(SEM) (Bollen 1989; Hayduk 1987) was suggested by McIntosh & Gonzalez-Lima (1994). SEM provides the opportunity to investigate functional–anatomical models subserving different cognitive functions in terms of which regions are involved and how they interact in a given network model. SEM commonly assumes multivariate normally distributed data. In order to characterize a functional network a specific functional–anatomical model is used in conjunction with SEM to model the observed covariance structure between the regions included in the model. The functional–anatomical model is specified by selecting the network components (ROIs or voxels) and the connections between the components based on theoretical or empirical considerations. Different constraints on the connections may also be specified (cf. McIntosh & Gonzalez-Lima 1994). The interregional covariances are computed and finally the connection strengths or path coefficients are estimated within condition. Differences between conditions or groups can be estimated using a stacked models approach (Bollen 1989; Hayduk 1987; McIntosh & Gonzalez-Lima 1994).

SEM commonly uses a linear system of equations to describe the interrelation between regional activities in the functional–anatomical model with the connection coefficients as free parameters. Nonlinear extensions have been described (Kenny & Judd 1984) and applied to fMRI data (Buechel & Friston 1997). The connection strengths are estimated in an optimization process. This procedure recreates the observed covariance between regions as closely as possible by finding optimal values of the path coefficients. There are several optimization algorithms available. Commonly, the optimization process uses estimated starting values in combination with an iterative maximum-likelihood estimation procedure. For example, the standard implementation in the LISREL program (Jöreskog & Sörbom 1996; see also Boomsma 1985) uses instrumental variables and a two-stage least square approach in combination with the Davidon–Fletcher–Power algorithm and line search (other alternatives are available, cf. Jöreskog & Sörbom (1996)). With reasonably well-fitting models, the initial estimates are often close enough to the final maximum-likelihood estimate for the optimization algorithms to quickly converge to this estimate. It should be noted that when the estimates depend nonlinearly on the model parameters there is no guarantee that the global optimum will be reached with deterministic gradient descent algorithms or non-exhaustive search procedures. Alternatively, a simulated annealing approach to optimization can be used (Geman & Geman 1984; Kirkpatrick *et al.* 1983) even though practical annealing schedules generally only generate good sub-optimal solutions.

The results of SEM analysis are potentially difficult to interpret for several reasons. There is no guarantee that the connections modelled actually reflect direct effective connections—it is possible that they are mediated through areas or connections not included in the model. Similarly, observed changes in the weights between states or groups may reflect common input from regions not modelled. This touches on the general problem of model selection, that is, the problem of matching model and data complexity. In the case of SEM, model selection may be performed in a data-driven mode, guided by goodness-of-fit values, modification indices or using a hierarchical model-building approach when subsets of weights are estimated recursively. The data-driven approach is vulnerable to over-fitting, since sample specific characteristics may be modelled, which may limit the generalizability of the results. For example, outliers or noise may be modelled increasing the risk that the model becomes over-fitted. Alternatively, model selection may be theory driven, running the risk of investigating incomplete models where there may be regions or links missing in the model. These aspects illustrate the model selection problem.

The results of a stacked models comparison can be difficult to interpret, unless reasonable goodness-of-fit can be achieved with a given model in all states or groups investigated. For example, using an under-parameterized model to test differences between states or groups in a stacked approach may yield results due to an ill-fitting model (in one of the states or groups). The effect of using under-parameterized models (i.e. omission of one network component, connection, or feedback loop) has been investigated in a relatively simple model (McIntosh & Gonzalez-Lima 1994). This simulation study indicates that the results from analysing moderately reduced models can be fairly stable and the modification indices can to some extent provide indications of such omissions. These and other issues may be of interest for further investigation in more complex models, for example, the implications of introducing different constraints on the weights, the residuals, as well as not taking nonlinear effects into account. More severely under-parameterized models may also be studied. The aspects so far investigated relate to model selection. The inferential consequences of ill-fitting models are also relevant to investigate, when a stacked models comparison is attempted.

Alternative approaches to effective connectivity have been proposed. For example, McIntosh & Gonzalez-Lima (1994) describe a simple model for studying the effects of experimental manipulation on both regional activity and interregional covariance. This approach has since become known as psychophysiological interactions (Friston *et al.* 1997). Other approaches suggest the use of nonlinear techniques (Friston & Buechel 1998; see also Friston *et al.* 1998) in the form of truncated Volterra series expansion (Priestley 1988), or variable parameter regression (Buechel & Friston 1998; cf. Chatfield 1996) in conjunction with Kalman filtering (Chatfield 1996; Wei 1990). Note that when these network approaches or SEM are applied to fMRI data it is necessary to take the temporal autocorrelation into account. Finally, it has been suggested that the interface between FNI, network analysis and large-scale neural modelling may add to our understanding of human cognition (Friston 1998*a*; Horwitz 1998; Horwitz *et al.* 1999).

4. CONCLUSION

FNI methods provide experimental access to the living human brain and have been rapidly developing during the last two decades. A framework of well-described theories and empirically validated methods are available providing a background for the development of new analytical tools. The FNI methods used differ in assumptions and these need to be examined in order to indicate the boundaries of optimal use. Central to this is, on the one hand, how well empirical data fulfil the assumptions and the approximations made, and on the other, the robustness of the methods used. This notion emphasizes the importance of empirical validation, investigation of robustness, and the explicit characterization of the inherent limitations of a given method. In this paper we have focused on assumptions and inherent limitations of the methods reviewed. This indicates the limits of applicability and defines constraints on the interpretations of results obtained. When these assumptions and limitations are taken into account, the different methods and approaches described generally serve their purposes well.

We have discussed some aspects of the complex problem of model selection. Model identification is of particular importance for future developments in the analysis of fMRI data. In general, proper model selection is a necessary prerequisite for the validity of the subsequent statistical inference, which depends on the use of sufficiently well-fitting models. Assessing model fit and verification of assumptions are challenging tasks and effective tools for assessing the goodness-of-fit of models and diagnostics for violations of assumptions are generally lacking in FNI. However, there are several non-inferential descriptive methods that, combined with inspection of parameter estimates and other simple measures, can help in the process of model selection, outlier detection, and verification of assumptions. In addition, multivariate methods can be used to perform model selection in a comprehensive way.

It is of importance for the interpretation of FNI results to take into account the assumptions, approximations and inherent limitations of the methods used. There are several areas in need of attention. One is the characterization of the baseline and its fluctuations relative to which regional activations are measured and effective methods for estimating global effects independent of experimentally induced changes. In addition, comprehensive models of gA are lacking, which are necessary when large or several different ranges of activity are expected. Another issue is the need for HRF models that effectively incorporate potential regional specificity of the haemodynamic response as well as its variability over subjects and time. There may also be a need for more comprehensive statistical modelling allowing for regionally specific models moving beyond the image regression approach. We have also indicated the possibility of using a general Bayesian framework. This allows for the systematic incorporation of prior knowledge, informed modelling, and more flexible models than permitted by regression and a richer variety of questions can potentially be answered. Last, it is important to acknowledge the importance of random effects models when the scope of inference is to the whole population sampled. Group comparisons are not served well by fixed effects models when population inference is attempted, instead random effects models are necessary.

Finally, one of the great challenges to the field of FNI is the development of effective methods to study higher cognitive functions, subserved by nonlinear network interactions and non-stationary dynamics, in greater detail, using the full potential of the methods available in terms of spatio-temporal resolution.

K.M.P. was supported by grants from the Swedish Medical Research Council (8276) and the Karolinska Institute. A.P.H was supported by the Wellcome Trust for part of this work. The authors are grateful for many fruitful discussions concerning methodological issues with Peter Fransson, Karl J. Friston, Anthony R. McIntosh, John Ollinger and Keith Worsley.

REFERENCES

Aertsen, A. M. H. & Preissl, H. 1991 Dynamics of activity and connectivity in physiological neuronal networks. In *Nonlinear dynamics and neuronal networks* (ed. H. G. Schuster), pp. 281–302. New York: VHC Publishers Inc.

Aertsen, A. M. H., Gerstein, G. L., Habib, M. K. & Palm, G. 1989 Dynamics of neuronal firing correlation: modulation of 'effective connectivity'. *J. Neurophysiol.* **61**, 900–917.

Aguirre, G. K., Zarahn, E. & D'Esposito, M. 1997 Empirical analyses of BOLD fMRI statistics. II. Spatially smoothed data collected under null hypothesis and experimental conditions. *NeuroImage* **5**, 199–212.

Aguirre, G. K., Zarahn, E. & D'Esposito, M. 1998*a* A critique of the use of the Kolmogorov–Smirnov (KS) statistic for the analysis of BOLD fMRI data. *Magn. Reson. Med.* **39**, 500–505.

Aguirre, G. K., Zarahn, E. & D'Esposito, M. 1998*b* The inferential impact of global signal covariates in functional neuroimaging analyses. *NeuroImage* **8**, 302–306.

Aguirre, G. K., Zarahn, E. & D'Esposito, M. 1998*c* The variability of human, BOLD hemodynamic responses. *NeuroImage* **8**, 360–369.

Amit, D. J. 1989 *Modeling brain function: the world of attractor neural networks.* Cambridge University Press.

Andersson, J. L. R. 1997 How to estimate global activity independent of changes in local activity. *NeuroImage* **6**, 237–244.

Andreasen, N. C., Arndt, S. A., Cizadlo, T., O'Leary, D. S., Watkins, G. L., Boles Ponto, L. L. & Hichawa, R. D. 1995 Sample size and statistical power in [O^{15}]H$_2$O studies of human cognition. *J. Cerebr. Blood-Flow Metab.* **16**, 804–816.

Arbib, M. A. 1995 *The handbook of brain theory and neural networks.* Cambridge, MA: MIT Press.

Arndt, S. A., Cizadlo, T., Andreasen, N. C., Zeien, G., Harris, G., O'Leary, D. S., Watkins, G. L., Boles Ponto, L. L. & Hichawa, R. D. 1995 A comparison of approaches to the statistical analysis of [O^{15}]H$_2$O PET cognitive activation studies. *J. Neuropsychiat. Clin. Neurosci.* **7**, 155–168.

Arndt, S. A., Cizadlo, T., O'Leary, D. S., Gold, S. & Andreasen, N. C. 1996 Normalizing counts and cerebral blood flow intensity in functional imaging studies of the human brain. *NeuroImage* **3**, 175–184.

Ash, R. B. 1965 *Information theory.* New York: Dover Publications.

Bandettini, P. A. & Wong, E. C. 1997 A hypercapnia-based normalization method for improved spatial localization of human brain activation with fMRI. *NMR Biomed.* **10**, 197–203.

Bandettini, P. A., Wong, E. C., Hinks, R. S., Tikofsky, R. S. & Hyde, J. S. 1993 Time course EPI of human brain function during task activation. *Magn. Reson. Med.* **25**, 390–397.

Bell, A. J. & Sejnowski, T. J. 1995 An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7**, 1129–1159.

Bickel, P. J. & Docksum, K. A. 1977 *Mathematical statistics: basic ideas and selected topics.* Oakland, CA: Holden-Day.

Billingsley, P. 1995 *Probability and measure*, 3rd edn. New York: Wiley.

Bollen, K. A. 1989 *Structural equations with latent variables.* New York: Wiley.

Boomsma, A. 1985 Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometry* **50**, 229–242.

Box, G. E. P. & Tiao, G. C. 1992 *Bayesian inference in statistical analysis.* New York: Wiley.

Boynton, G. M., Engel, S. A., Glover, G. H. & Heeger, D. J. 1996 Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.* **16**, 4207–4221.

Bretthorst, G. L. 1990*a* Bayesian analysis. I. Parameter estimation using quadrature NMR models. *J. Magn. Reson.* **88**, 533–551.

Bretthorst, G. L. 1990*b* Bayesian analysis. II. Signal detection and model selection. *J. Magn. Reson.* **88**, 552–570.

Bruhn, H., Kleinschmidt, A., Boecker, H., Merboldt, K.-D., Hänicke, W. & Frahm, J. 1994 The effects of acetazolamide on regional cerebral blood oxygenation at rest and under stimulation as assessed by MRI. *J. Cerebr. Blood-Flow Metab.* **14**, 742–748.

Buckner, R. L., Koutstaal, W., Schacter, D. L., Dale, A. M., Rotte, M. & Rosen, B. R. 1998*a* Functional–anatomic study of episodic retrieval. II. Selective averaging of event-related fMRI trials to test the retrieval success hypothesis. *NeuroImage* **7**, 163–175.

Buckner, R. L., Koutstaal, W., Schacter, D. L., Wagner, A. D. & Rosen, B. R. 1998*b* Functional–anatomic study of episodic retrieval using fMRI. I. Retrieval effort versus retrieval success. *NeuroImage* **7**, 151–162.

Buechel, C. & Friston, K. J. 1997 Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. *Cerebr. Cortex* **7**, 768–778.

Buechel, C. & Friston, K. J. 1998 Dynamic changes in effective connectivity characterized by variable parameter regression and Kalman filtering. *Hum. Brain Mapp.* **6**, 403–408.

Bullmore, E., Brammer, M., Williams, S. C. R., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R. & Sham, P. 1996 Statistical methods of estimation and inference for functional MR image analysis. *Magn. Reson. Med.* **35**, 261–277.

Buxton, R. B. & Frank, L. R. 1997 A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *J. Cerebr. Blood-Flow Metab.* **17**, 64–72.

Buxton, R. B., Wong, E. C. & Frank, L. R. 1998 Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magn. Reson. Med.* **39**, 855–864.

Cao, J. & Worsley, K. J. 1999 The geometry of correlation fields with an application to functional connectivity of the brain. *Ann. Appl. Prob.* (In the press.)

Chatfield, C. 1996 *The analysis of time series: an introduction*. London: Chapman & Hall.

Chatfield, C. & Collins, A. J. 1980 *Introduction to multivariate analysis*. London: Chapman & Hall.

Clark, C. & Carson, R. 1993 Letter to the editor: analysis of covariance in statistical parametric mapping. *J. Cerebr. Blood-Flow Metab.* **13**, 1038.

Cohen, M. S. 1997 Parametric analysis of fMRI data using linear systems methods. *NeuroImage* **6**, 93–103.

Common, P. 1994 Independent component analysis: a new concept? *Signal Processing* **36**, 11–20.

Corfield, D. R., Murphy, K., Josephs, O., Adams, L. & Turner, R. 1998 Modulation by hypercapnia of activation-related BOLD signal changes in the visual cortex. *NeuroImage* **7**, S259.

Cover, T. M. & Thomas, J. A. (eds) 1991 *Elements of information theory.* New York: Wiley.

Dale, A. M. & Buckner, R. L. 1997 Selective averaging of rapidly presented individual trials using fMRI. *Hum. Brain Mapp.* **5**, 329–340.

Davis, T. L., Weisskoff, R. M., Kwong, K. K., Boxerman, J. L. & Rosen, B. R. 1994 Temporal aspects of fMRI task activation: dynamic modeling of oxygen delivery. Society for Magnetic Resonance, 2nd Annual Meeting, August 6–12 1994, San Francisco, CA, USA, p. 69.

Descombes, X., Kruggel, F. & von Cramon, D. Y. 1998 fMRI signal restoration using a spatiotemporal Markov random field preserving transitions. *NeuroImage* **8**, 340–349.

Draper, N. R. & Smith, H. 1981 *Applied regression analysis*, 2nd edn. New York: Wiley.

Edgington, E. S. 1995 *Randomization tests*, 3rd edn (revised and expanded). New York: Marcel Dekker.

Efron, B. & Tibshirani, R. 1986 Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statist. Sci.* **1**, 54–77.

Fitzgerald, M. 1996 Registration and estimation of functional magnetic resonance images. PhD thesis, University of Carnegie Mellon, PA, USA.

Ford, I. 1986 Confounded correlations: statistical limitations in the analysis of interregional relationships of cerebral metabolic activity. *J. Cerebr. Blood-Flow Metab.* **6**, 385–388.

Ford, I. 1995 Commentary and opinion. III. Some nonontological and functionally unconnected views on current issues in the analysis of PET datasets. *J. Cerebr. Blood-Flow Metab.* **15**, 371–377.

Ford, I., McColl, J. H., McCormack, A. G. & McCroy, S. J. 1991 Statistical issues in the analysis of neuroimages. *J. Cerebr. Blood-Flow Metab.* **11**, A89–A95.

Fox, P. T. & Mintun, M. A. 1989 Non-invasive functional brain mapping by change distribution analysis of averaged PET images of $H_2^{15}O$ tissue activity. *J. Nucl. Med.* **30**, 141–149.

Fox, P. T. & Raichle, M. E. 1984 Stimulus rate dependence of regional cerebral blood flow in human striate cortex demonstrated with positron emission tomography. *J. Neurophysiol.* **51**, 1109–1121.

Fox, P. T., Mintun, M. A., Reiman, E. M. & Raichle, M. E. 1988 Enhanced detection of focal brain responses using intersubject averaging and change-distribution analysis of subtracted PET images. *J. Cerebr. Blood-Flow Metab.* **8**, 642–653.

Frackowiak, R. S. J., Zeki, S., Poline, J.-B. & Friston, K. J. 1996 A critique of a new analysis proposed for functional neuroimaging. *Eur. J. Neurosci.* **8**, 2229–2231.

Frackowiak, R. S. J., Friston, K. J., Frith, C. D., Dolan, R. J. & Mazziotta, J. C. 1997 *Human brain function*. San Diego, CA: Academic Press.

Frank, L. R., Buxton, R. B. & Wong, E. C. 1998 Probabilistic analysis of functional magnetic resonance imaging data. *Magn. Reson. Med.* **39**, 132–148.

Friston, K. J. 1993 Letter to the editor: author's reply. *J. Cerebr. Blood-Flow Metab.* **13**, 1038–1040.

Friston, K. 1994 Functional and effective connectivity: a synthesis. *Hum. Brain Mapp.* **2**, 56–78.

Friston, K. J. 1995 Commentary and opinion. II. Statistical parametric mapping—ontology and current issues. *J. Cerebr. Blood-Flow Metab.* **15**, 361–370.

Friston, K. J. 1998*a* Imaging neuroscience: principles or maps? *Proc. Natl Acad. Sci. USA* **95**, 796–802.

Friston, K. J. 1998*b* Modes or models: a critique of independent component analysis for fMRI. *Trends Cogn. Sci.* **2**, 373–375.

Friston, K. J. & Buechel, C. 1998 A nonlinear analysis of functional brain architectures. *NeuroImage* **7**, S776.

Friston, K. J., Frith, C. D., Liddle, P. F., Dolan, R. J., Lammertsma, A. A. & Frackowiak, R. S. 1990 The relationship between global and local changes in PET scans. *J. Cerebr. Blood-Flow Metab.* **10**, 458–466.

Friston, K. J., Frith, C. D., Liddle, P. F. & Frackowiak, R. S. J. 1991 Comparing functional (PET) images: the assessment of significant change. *J. Cerebr. Blood-Flow Metab.* **11**, 690–699.

Friston, K. J., Frith, C. D., Liddle, P. F. & Frackowiak, R. S. 1993 Functional connectivity: the principal-component analysis of large (PET) data sets. *J. Cerebr. Blood-Flow Metab.* **13**, 4–14.

Friston, K. J., Jezzard, P. & Turner, R. 1994*a* Analysis of functional MRI time-series. *Hum. Brain Mapp.* **1**, 210–220.

Friston, K. J., Worsely, K. J., Frackowiak, R. S. J., Mazziotta, J. C. & Evans, A. C. 1994*b* Assessing the significance of focal activations using their spatial extent. *Hum. Brain Mapp.* **1**, 214–220.

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

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

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

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E. & Dolan, R. J. 1997 Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage* **6**, 218–229.

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

Geman, S. & Geman, D. 1984 Stochastic relaxation, Gibbs distribution, and Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intell.* **6**, 721–741.

Genovese, C. R. 1997 *Statistical inference in functional magnetic resonance imaging*, technical report 674. Pittsburgh: Department of Statistics, Carnegie Mellon University.

Genovese, C. R. & Sweeney, J. S. 1998 Functional connectivity in the cortical regions subserving eye movements (with discussion). In *Case studies in Bayesian statistics* (ed. R. B. Kass, B. P. Carlin, A. L. Carriquiry, C. Catsonis, A. Gelman, I. Verdinelli & M. West), pp. 59–132. New York: Springer.

Gonzalez-Lima, F. & McIntosh, A. R. 1994 Neural network interactions related to auditory learning analyzed with structural equation modeling. *Hum. Brain Mapp.* **2**, 23–44.

Good, P. 1994 *Permutation tests*. New York: Springer.

Grabowski, T. J., Frank, R. J., Brown, C. K., Damasio, H., Boles Ponto, L. L., Watkins, G. L. & Hichwa, R. D. 1996 Reliability of PET activation across statistical methods, subject groups and sample sizes. *Hum. Brain Mapp.* **4**, 23–46.

Grady, C. L., McIntosh, A. R., Rajah, M. N. & Craik, F. I. M. 1998 Neural correlates of the episodic encoding of pictures and words. *Proc. Natl Acad. Sci. USA* **95**, 2703–2708.

Halber, M., Herholz, K., Wienhard, K., Pawlik, G. & Heiss, W.-D. 1997 Performance of a randomization test for single-subject O^{15}-water PET activation studies. *J. Cerebr. Blood-Flow Metab.* **17**, 1033–1039.

Hayduk, L. A. 1987 *Structural equation modeling with LISREL: essentials and advances*. Baltimore, MD: Johns Hopkins University Press.

Herholz, K., Kessler, J., Slansky, I., Mielke, R. & Heiss, W. D. 1993 A model for separation of regional from global metabolic activation during continuous visual recognition in Alzheimer's disease. In *Quantification of brain function: tracer kinetics and image analysis in brain PET* (ed. K. Uemura, N. A. Lassen, T. Jones & I. Kanno), pp. 555–560. Amsterdam: Excerpta Medica.

Hertz, J., Krogh, A. & Palmer, R. G. 1991 *Introduction to the theory of neural computation*. San Diego, CA: Addison-Wesley.

Holmes, A. P. 1994 Statistical issues in functional brain mapping. PhD thesis, University of Glasgow.

Holmes, A. P. & Ford, I. 1993 A Bayesian approach to significance testing for statistic images from PET. In *Quantification of brain function: tracer kinetics and image analysis in brain PET* (ed. K. Uemura, N. Lassen, T. Jones & I. Kanno), pp. 521–531. Amsterdam: Excerpta Medica.

Holmes, A. P. & Friston, K. J. 1998 Generalizability, random effects and population inference. *NeuroImage* **7**, S754.

Holmes, A. P., Blair, R. C., Watson, J. D. G. & Ford, I. 1996 Non-parametric analysis of statistic images from functional mapping experiments. *J. Cerebr. Blood-Flow Metab.* **16**, 7–22.

Holmes, A. P., Josephs, O., Buchel, C. & Friston, K. J. 1997 Statistical modelling of low-frequency confounds in fMRI. *NeuroImage* **5**, S480.

Holmes, A. P., Watson, J. D. G. & Nichols, T. E. 1998 Holmes and Watson on 'Sherlock'. *J. Cerebr. Blood-Flow Metab.* **18**, 697.

Horgan, T. & Tienson, J. 1996 *Connectionism and the philosophy of psychology.* Cambridge, MA: MIT Press.

Horwitz, B. 1998 Using functional brain imaging to understand human cognition. *Complexity* **3**, 39–52.

Horwitz, B. & Rapoport, S. I. 1988 Partial correlation coefficients approximate the real intrasubject correlation pattern in the analysis of interregional relations of cerebral metabolic activity. *J. Nucl. Med.* **29**, 392–399.

Horwitz, B., Duara, R. & Rapoport, S. I. 1984 Intercorrelations of glucose metabolic rates between brain regions: application to healthy males in a state of reduced sensory input. *J. Cerebr. Blood-Flow Metab.* **4**, 484–499.

Horwitz, B., Grady, C. L., Haxby, J. V., Ungerleider, L. G., Schapiro, M. B., Mishkin, M. & Rapoport, S. I. 1992*a* Functional associations among human posterior extrastriate brain regions during object and spatial vision. *J. Cogn. Neurosci.* **4**, 311–322.

Horwitz, B., Soncrant, J. V. & Haxby, J. V. 1992*b* Covariance analysis of functional interactions in the brain using metabolic and blood flow data. In *Advances in metabolic mapping techniques for brain imaging of behavioral and learning functions* (ed. F. Gonzalez-Lima, T. Finkenstaedt & H. Scheich), pp. 189–217. Dordrecht: Kluwer.

Horwitz, B., McIntosh, A. R., Haxby, J. V. & Grady, C. L. 1995 Network analysis of brain cognitive function using metabolic and blood flow data. *Behav. Brain Res.* **66**, 187–193.

Horwitz, B., Tagamets, M.-A. & McIntosh, A. R. 1999 Neural modeling, functional brain imaging, and cognition. *Trends Cogn. Sci.* **3**, 91–98.

Hu, X., Le, T. H., Parrish, T. & Erhard, P. 1995 Retrospective estimation and correction of physiological fluctuation in functional MRI. *Magn. Reson. Med.* **34**, 201–212.

Hu, X., Le, T. H. & Ugurbil, K. 1997 Evaluation of the early response in fMRI in individual subjects using short stimulus duration. *Magn. Reson. Med.* **37**, 877–884.

Huber, P. J. 1985 Projection pursuit. *Ann. Statist.* **13**, 435–475.

Jain, A. K. 1989 *Fundamentals of digital image processing.* Englewood Cliffs, NJ: Prentice-Hall.

Jaynes, E. T. 1968 Prior probabilities. *IEEE Trans. Syst. Sci. Cybernetics SSC* **4**, 227–241.

Jolliffe, I. T. 1986 *Principal component analysis*. New York: Springer.

Jöreskog, K. & Sörbom, D. 1996 *LISREL 8 user's reference guide.* Chicago, IL: Scientific Software International.

Jöreskog, K. G. & Wold, H. 1982 The ML and PLS techniques for modeling with latent variables: historical and comparative aspects. In *Systems under indirect observation: causality, structure, prediction* (ed. H. Wold & K. G. Jöreskog), pp. 263–270. Amsterdam: North-Holland.

Kanno, I., Hatazawa, J., Shimosegawa, E., Ishii, K. & Fujita, H. 1996 Proportionality of reaction CBF to baseline CBF with neural activation and deactivation. In *Quantification of brain function using PET* (ed. R. Myers, V. Cunningham, D. Bailey & T. Jones), pp. 362–362. San Diego, CA: Academic Press.

Kenny, D. A. & Judd, C. M. 1984 Estimating the nonlinear and interactive effects of latent variables. *Psychol. Bull.* **96**, 201–210.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. 1983 Optimization by simulated annealing. *Science* **220**, 671–680.

Koch, C. & Davis, J. L. 1994 *Large-scale neuronal theories of the brain*. Cambridge, MA: MIT Press.

Laird, N. M. & Ware, J. H. 1982 Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Lange, N. 1997 Empirical and substantive models, the Bayesian paradigm, and meta-analysis in functional brain imaging. *Hum. Brain Mapp.* **5**, 259–263.

Lange, N. & Zeger, S. L. 1997 Non-linear Fourier time-series analysis for human brain mapping by functional magnetic resonance imaging. *J. R. Statist. Soc. Appl. Statist.* **46**, 1–29.

Lee, P. M. 1997 *Bayesian statistics: an introduction*, 2nd edn. New York: Wiley.

Locascio, J. J., Jennings, P. J., Moore, C. I. & Corkin, S. 1997 Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Hum. Brain Mapp.* **5**, 168–193.

McColl, J. H., Holmes, A. P. & Ford, I. 1994 Statistical methods in neuroimaging with particular application to emission tomography. *Statist. Meth. Med. Res.* **3**, 63–86.

McCrory, S. J. & Ford, I. 1991 Multivariate analysis of SPECT images with illustrations in Alzheimer's disease. *Statist. Med.* **10**, 1711–1718.

McCullagh, P. & Nelder, J. A. 1989 *Generalized linear models*, 2nd edn. London: Chapman & Hall.

Macdonald, C. & Macdonald, G. (eds) 1995 *Connectionism: debates on psychological explanation*. Oxford: Blackwell.

McIntosh, A. R. & Gonzalez-Lima, F. 1994 Structural equation modeling and its application to network analysis in functional brain imaging. *Hum. Brain Mapp.* **2**, 2–22.

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

McIntosh, A. R., Grady, C. L., Haxby, J. V., Maisog, J. M., Horwitz, B. & Clark, C. M. 1996*b* Within-subject transformation of PET regional cerebral blood flow data: ANCOVA, ratio, and *Z*-score adjustment on empirical data. *Hum. Brain Mapp.* **4**, 93–102.

McIntosh, A. R., Cabeza, R. & Lobaugh, N. J. 1998*a* Explaining the activation of occipital cortex by an auditory stimulus through analysis of neural interactions. *J. Neurophysiol.* **80**, 2790–2796.

McIntosh, A. R., Lobaugh, N. J., Cabeza, R., Bookstein, F. L. & Houle, S. 1998*b* Convergence of neural systems processing stimulus associations and coordinating motor responses. *Cerebr. Cortex* **8**, 648–659.

McKeown, M. J., Makeig, S., Brown, G. B., Jung, T.-P., Kindermann, S. S., Bell, A. J. & Sejnowski, T. J. 1998 Analysis of fMRI data by blind separation into independent spatial components. *Hum. Brain Mapp.* **6**, 160–188.

Mesulam, M. M. 1990 Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Ann. Neurol.* **28**, 597–613.

Mesulam, M. M. 1998 From sensation to cognition. *Brain* **121**, 1013–1052.

Moeller, J. R. & Strother, S. C. 1991 A regional covariance approach to the analysis of functional patterns in positron emission tomographic data. *J. Cerebr. Blood-Flow Metab.* **11**, A121–A135.

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

Ogawa, S., Menon, R. S., Kim, S.-G. & Ugurbil, K. 1998 On the characteristics of functional magnetic resonance imaging of the brain. *A. Rev. Biophys. Biomol. Struct.* **27**, 447–474.

Ojemann, J. G., Buckner, R. L., Akbudak, E., Snyder, A. Z., Ollinger, J. M., McKinstry, R. C., Rosen, B. R., Petersen, S. E., Raichle, M. E. & Conturo, T. E. 1998 Functional MRI studies of word-stem completion: reliability across laboratories and comparison to blood flow imaging with PET. *Hum. Brain Mapp.* **6**, 203–215.

Oppenheim, A. V. & Schafer, R. W. 1989 *Discrete-time signal processing*. Englewood Cliffs, NJ: Prentice-Hall.

Paradis, A.-L., Van de Morrtele, P.-F., Le Bihan, D. & Poline, J.-B. 1998 Do high temporal frequencies of the event-related fMRI response have a more specific spatial localization? *NeuroImage* **7**, S617.

Petersson, K. M. 1998 Comments on a Monte Carlo approach to the analysis of functional neuroimaging data. *NeuroImage* **8**, 108–112.

Poline, J.-B., Vandenberghe, R., Holmes, A. P., Friston, K. J. & Frackowiak, R. S. J. 1996 Reproducibility of PET activation studies: lessons from a multi-center European experiment. *NeuroImage* **4**, 34–54.

Priestley, M. B. 1988 *Non-linear and non-stationary time series analysis*. New York: Academic Press.

Purdon, P. L. & Weisskoff, R. M. 1998 Effect of temporal autocorrelation due to physiological noise stimulus paradigm on voxel-level false-positive rates in fMRI. *Hum. Brain Mapp.* **6**, 239–249.

Purdon, P. L., Solo, V., Brown, E., Buckner, R., Rotte, M. & Weisskoff, R. M. 1998 fMRI noise variability across subjects and trials: insights for noise estimation methods. *NeuroImage* **7**, S617.

Rajapakse, J. C., Kruggel, F., Maisog, J. M. & von Cramon, D. Y. 1998 Modeling hemodynamic response for analysis of functional MRI time-series. *Hum. Brain Mapp.* **6**, 283–300.

Ramsay, S. C., Murphy, K., Shea, S. A., Friston, K. J., Lammertsma, A. A., Clark, J. C., Adams, L., Guz, A. & Frackowiak, R. S. 1993 Changes in global cerebral blood flow in humans: effect on regional cerebral blood flow during a neural activation task. *J. Physiol.* **471**, 521–534.

Roland, P. E. & Gulyas, B. 1996 Assumptions and validations of statistical tests for functional neuroimaging. *Eur. J. Neurosci.* **8**, 2232–2235.

Rosen, B. R., Buckner, R. L. & Dale, A. M. 1998 Event-related functional MRI: past, present, and future. *Proc. Natl Acad. Sci. USA* **95**, 773–780.

Rosenfeld, A. & Kak, A. C. (eds) 1982 *Digital picture processing*. Orlando, FL: Academic Press.

Rumelhart, D. E. & McClelland, J. L. 1986 *Parallel distributed processing: explorations in the microstructures of cognition*, vols 1 and 2. Cambridge, MA: MIT Press.

Ruttimann, U. E., Rio, D., Rawlings, R. R., Andreasen, P. & Hommer, D. M. 1998 PET analysis using a variance stabilizing transform. In *Quantitative functional brain imaging with positron emission tomography* (ed. R. E. Carson, M. E. Daube-Witherspoon & P. Herscovitch), pp. 217–222. San Diego, CA: Academic Press.

Schuster, H. G. 1991 *Nonlinear dynamics and neuronal networks*. New York: VHC Publishers Inc.

Senda, M., Ishii, K., Oda, K., Sadato, N., Kawashima, R., Sugiura, M., Kanno, I., Ardekani, B., Minoshima, S. & Tatsumi, I. 1998 Influence of ANOVA design and anatomical standardization on statistical mapping for PET activation. *NeuroImage* **8**, 283–301.

Snedecor, G. W. & Cochran, W. G. 1967 *Statistical methods,* 6th edn. Ames, IA: Iowa State University Press.

Squire, L. R. 1992 Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychol. Rev.* **99**, 195–231.

Strother, S. C., Anderson, J. R., Schaper, K. A., Siditis, J. J., Liow, J.-S., Woods, R. P. & Rottenberg, D. A. 1995*a* Principal component analysis and the scaled subprofile model compared to intersubject averaging and statistical parametric mapping. I. Functional connectivity of the human motor system studied with [^{15}O]water PET. *J. Cerebr. Blood-Flow Metab.* **15**, 738–753.

Strother, S. C., Kanno, I. & Rottenberg, D. A. 1995*b* Commentary and opinion. I. Principal component analysis, variance partitioning, and 'functional connectivity'. *J. Cerebr. Blood-Flow Metab.* **15**, 353–360.

Strother, S. C., Siditis, J. J., Anderson, J. R., Hansen, L. K., Schaper, K. & Rottenberg, D. A. 1996 [^{15}O]water PET: more noise than signal? In *Quantification of brain function using PET* (ed. R. Myers, V. Cunningham, D. Bailey & T. Jones), pp. 378–383. San Diego, CA: Academic Press.

Strother, S. C., Lange, N., Anderson, J. R., Schaper, K. A., Rehm, K., Hansen, L. K. & Rottenberg, D. A. 1997 Activation pattern reproducibility: measuring the effect of group size and data analysis models. *Hum. Brain Mapp.* **5**, 312–316.

Taylor, S. F., Minoshima, S. & Koeppe, R. A. 1993 Letter to the editor: instability of localization of cerebral blood flow activation foci with parametric maps. *J. Cerebr. Blood-Flow Metab.* **13**, 1040–1041.

Van Horn, J. D., McIntosh, A. R. & Maisog, J. M. 1995 Letter to the editor: complications in the use of the SPM χ^2 statistic. *J. Cerebr. Blood-Flow Metab.* **15**, 895–896.

Van Horn, J. D., Ellmore, T. M., Esposito, G. & Berman, K. F. 1998 Mapping voxel-based statistical power on parametric images. *NeuroImage* **7**, 97–107.

Vazquez, A. L. & Noll, D. C. 1998 Nonlinear aspects of the BOLD response in functional MRI. *NeuroImage* **7**, 108–118.

Wei, W. W. S. 1990 *Time series analysis: univariate and multivariate methods*. Redwood City, CA: Addison-Wesley.

Weisskoff, R. M., Baker, J., Belliveau, J., Davis, T. L., Kwong, K. K., Cohen, M. S. & Rosen, B. R. 1993 Power spectrum analysis of functionally weighted MR data: what's in the noise? *Proc. Soc. Magn. Reson. Med.* **1**, 7.

Winer, B. J., Brown, D. R. & Michels, K. M. 1991 *Statistical principles in experimental design*, 3rd edn. New York: McGraw-Hill.

Wold, H. 1985 Partial least squares. In *Encyclopedia of statistical sciences* (ed. S. Kotz & N. L. Johnson), pp. 581–591. New York: Wiley.

Woods, R. P. 1996 Modeling for intergroup comparisons of imaging data. *NeuroImage* **4**, S84–S94.

Worsley, K. J. 1994 Local maxima and the expected Euler characteristic of excursion sets of χ^2, *F* and *t* fields. *Adv. Appl. Prob.* **26**, 13–42.

Worsley, K. J. 1997 An overview and some new developments in the statistical analysis of PET and fMRI data. *Hum. Brain Mapp.* **5**, 254–258.

Worsley, K. J. 1999 Testing for signals with unknown location and scale in a χ^2 random field, with an application to fMRI. *Adv. Appl. Prob.* (In the press.)

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

Worsley, K. J., Evans, A. C., Marrett, S. & Neelin, P. 1993 Letter to the editor: authors reply. *J. Cerebr. Blood-Flow Metab.* **13**, 1041–1042.

Worsley, K. J., Poline, J. B., Vandal, A. C. & Friston, K. J. 1995 Tests for distributed nonfocal brain activations. *NeuroImage* **2**, 183–194.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J. & Evans, A. C. 1996 A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp.* **4**, 58–73.

Worsley, K. J., Wolforth, M. & Evans, A. C. 1997*a* Scale space searches for a periodic signal in fMRI data with spatially varying hemodynamic response. (Submitted.)

Worsley, K. J., Poline, J.-B., Friston, K. J. & Evans, A. C. 1997*b* Characterizing the response of PET and fMRI data using multivariate linear models. *NeuroImage* **6**, 305–319.

Worsley, K. J., Cao, J., Paus, T., Petrides, M. & Evans, A. C. 1998 Applications of random field theory to functional connectivity. *Hum. Brain Mapp.* **6**, 364–367.

Xiong, J., Gao, J.-H., Lancaster, J. L. & Fox, P. T. 1996 Assessment and optimization of functional MRI analyses. *Hum. Brain Mapp.* **4**, 153–167.

Zarahn, E., Aguirre, G. K. & D'Esposito, M. 1997 Empirical analyses of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null hypothesis conditions. *NeuroImage* **5**, 179–197.

Zola-Morgan, S. & Squire, L. R. 1993 Neuroanatomy of memory. *A. Rev. Neurosci.* **16**, 547–563.