Statistical Analysis of fMRI Time-Series: A Critical Evaluation of the GLM Approach

Martin M. Monti[♯]

Princeton University, Dept. of Psychology

April 18, 2006

NUMBER OF PAGES: 44

ABBREVIATED TITLE: Statistical Analysis of fMRI Time-Series.

Acknowledgements: I'd like to thank Damien Rice, Muse, Radiohead and Dolorean for the musical support throughout the redaction of this manuscript.

[♯] Correspondence address:

Martin M. Monti

Princeton University

Department of Psychology – Green Hall,

Princeton, NJ 08544.

Tel. (609)258-5679

email: mmonti@princeton.edu

Abstract

Functional Magnetic Resonance Imaging (fMRI) is currently one of the most widely used tools to study *in vivo* the neural underpinnings of human cognition. Analysis of fMRI data relies on a General Linear Model (GLM) approach to separate noise from the actual signal corvarying with some experimental task of interest. Validity of inferences drawn from such approach to data-analysis is secondary to the satisfaction of condition imposed by the statistical model. In the present paper we review the GLM approach to fMRI time-series analysis by considering the degree by which such data abides by the hypothesis of the model and by presenting the methodologies that have been put forward in order to correct for assumptions' infringement.

KEYWORDS: Functional Magnetic Resonance Imaging (fMRI), Blood Oxygenation Level-Dependent (BOLD), General Linear Model (GLM), Ordinary Least Squares (OLS), Autocorrelation, Heteroscedasticity, Multicollinearity, Fixed Effects, Random Effects, Conjunction Analysis.

Contents

0.1 Introduction

In the past decade the study of human cognition has largely benefited from technological innovations in the field of nuclear magnetic resonance (NMR). Its application to study *in vivo* the functional architecture of the brain levers on the dynamic change of oxyhemoglobin (Hb) – or rather venous deoxyhemoglobine (dHb) – levels in the capillaries. In a high magnetic field (e.g. 3 Tesla) paramagnetic dHb produces local decreases in the field intensities (i.e. local field inhomogeneities). Increased capillary oxygenation conversely diminishes signal disruption by decreasing the levels of capillary dHb. Functional magnetic resonance imaging (fMRI) thus detects fluctuations in a blood oxygenation level-dependent (BOLD) magnetic signal ([1, 2, 3]). Underlying such technique is the assumption that metabolic changes reflect neural activity with some regard to spatial extent and intensity of firing, an idea originally postulated by Roy and Sherrington [4]. At present the exact system by which this "automatic mechanism" couples metabolic activity and neural processing is still unclear. In support of the coupling hypothesis though Logothetis [5] and Logothetis *et al.* [6] have reported a strong correlation between local field potentials (LFPs) recorded with intracortical electrodes and the simultaneously acquired BOLD signal, suggesting synaptic potentials and dendritic processing as the main components underlying the BOLD effect. Yet, such coupling has been called into question by discordant evidence (e.g. [7]). Despite the growing evidence in favor of a coupling of BOLD signal and LFPs the issue doesn't seem to be at present univocally resolved (see [8, 9, 10] for a review of the main issues and experimental findings).

The reminder of the paper will be concerned with statistical methods for separating noise from systematic fluctuations of the BOLD signal in correlation to a stimulation pattern. First we will briefly introduce the General Linear Model (GLM) statistical framework for analysis of a single subject data. Following we will analyze the degree by which

fMRI data actually conform to each of the model's hypotheses and review the main approaches used to overcome assumption infringements. We will then turn our attention to how data-sets from multiple runs/sessions of a single subject and from multiple subjects are combined together reviewing problems and merits of the major approaches in terms of their validity and inferential scope.[1] Finally we will briefly discuss some other more general statistical issues of fMRI data analysis.

0.2 Single Subject Analysis (I): The General Linear Model Approach

An fMRI data set, can be seen as a set of cuboid elements (i.e. voxels) of variable dimension, each of which has an associated time series of as many time-points as volumes acquired per session. The aim of the statistical analysis is to determine which voxels activate and deactivate (as measured by the BOLD signal) in correlation with some specific task of interest.

The first step in fMRI data analysis is typically a series of "pre-processing" transformations applied to the aim of "conditioning" the data, possibly increasing robustness of the following statistical analysis and adjusting for several artifacts introduced at data acquisition. Each transformation can be applied independently as a function of the specific needs or requirements of the experimental design used. The most typical steps include reconstructing the images in "brain-space" from the frequency domain in which the machine encodes the data (so called *k-space*), adjusting for acquisition-specific artifacts (e.g. slice timing realignment), subject motion, and often temporal and spatial smoothing. (See [11] for a thorough analysis of each step.) Following pre-processing data analysis is carried

[1]For clarity purposes the rest of the manuscript will refer to a "run" (also a "scan") as one continuous stream of data acquisition (i.e. from "scanner on" to "scanner off"). Each subject usually undergoes multiple runs, with brief interruptions in between, and typically remaining inside the bore of the machine. The set of these multiple runs is referred to as a "session". A standard fMRI data-set for a complete experiment usually comprises one (or more) session *per* each of about a dozen (or more) subject.

in two general steps: first-level analysis, typically a time series analysis of data relative to one subject's run and second-level analysis, in which results from multiple runs and multiple subjects are combined together.[2]

In the GLM framework, single subject fMRI data is analyzed by fitting at each voxel independently (i.e. univariate approach) a linear combination of independent variables, plus an error term. Two main reasons, mostly unique to functional neuro-imaging, underlie the choice of adopting a "massive univariate" approach rather than a multivariate one. First, there usually are more voxels than observations, whereas multivariate approaches need more observations (i.e. time-points) than dimension of the response variable (i.e. voxels). Second, multivariate techniques would characterize image volumes as a whole, thus not supporting statistical inference about regionally (i.e. voxel clusters) specific effects. The voxel-wise GLM is expressed as:

$$Y = \beta X + \varepsilon \tag{1}$$

where Y is a column vector of N rows (the number of collected time-points) representing the time-series BOLD signal associated to a single voxel. X represents the design matrix with N rows $\times$ p columns, each representing a regressor (i.e. an explanatory variable). Of interest are the columns representing manipulations or experimental conditions, although the matrix often may include regressors of *non-interest*, modelling the mean signal (i.e. the intercept), trends (typically linear and quadratic) and other design specific confounds. β is a column vector with p rows representing the unknown parameter associated to each regressor. Finally, ε is also a column vector, with N rows, representing the estimation error (or residuals) defined as $Y - \hat{\beta}X$.

[2]First-level analysis can be carried out either on one single subject's run, or on the concatenation of all the runs from a session. Accordingly the second level will thus either be a combination of multiple runs from multiple subjects or a combination of all the subjects' sessions (i.e. the runs' concatenation).

The end result of the mass-univariate GLM is to create a statistical parametric map (SPM) in "brain-space" of voxels responding systematically (significantly according to some pre-determined criterion) to one or more effects (of interest) modelled in the X matrix. Estimation of the unknown (i.e. β) is usually accomplished with an Ordinary Least Squares (OLS) [3] approach which computes the β-estimates that minimizes $\sum_{t=1}^{N} \hat{\varepsilon}^2$ (i.e. the squared difference between the observed signal Y and the estimated signal $\hat{Y}$, given the matrix). In this approach the estimator and its variance are computed as follows:[4]

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2}$$

$$var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \tag{3}$$

It can be shown that if the following Gauss-Markov ($\mathcal{GM}$) assumptions relating to the properties of the error term ε and the matrix X of explanatory variables hold then (2) will yield the best (i.e. minimum variance), linear and unbiased estimator (BLUE)[5] within the class of linear and unbiased estimators of the parameter. The $\mathcal{GM}$ assumptions require that:

(A1) ε is independently and identically distributed (i.i.d.) $\sim N(0, \sigma^2 I)$

(A2) The effects of the X matrix are independent of error (i.e. $E(\varepsilon, X) = 0$), non-stochastic (i.e. deterministic) and known.

(A3) No regressor is a linear transformation of one (or more) regressors.

[3] Generalized Least Squares (GLS) approaches are also used, especially with consideration to temporal autocorrelation issue, see below.

[4] It should be noted that in equation (2) the estimator $\hat{\beta}$ should result simply by YX^{-1}. Yet X is almost never a square matrix, thus it can't be inverted (i.e. X^{-1} can't be computed). To obviate the Moore-Penrose "pseudo-inverse" is typically used and $X^{-1} \sim (X^T X)^{-1} X^T$. Substituting the pseudo-inverse for X^{-1} we obtain expression (2).

[5] In addition to the unbiasedness and maximum efficiency properties, the BLUE is asymptotically normal $\sim N(\beta, \sigma^2(X'X))$ which is a requirement for subsequent parametric tests that subsume such property.

It should be noted that (A1) is in fact a three-part assumption requiring that the residuals are not serially correlated (A1a), that $E(\varepsilon) = 0$ (A1b), and a scalar variance, equal at all observation (i.e. equality of all the on-diagonal elements of the variance-covariance matrix) (A1c).

We will now discuss the degree by which fMRI data abides by these assumptions, consider the bearings of infringements of any of the $\mathcal{GM}$ assumptions (e.g. non-validity of statistical inferences) and the main methods currently available to adjust for them.

0.3 Single Subject Analysis (II): The $\mathcal{GM}$ Assumptions & fMRI Time-Series

Validity of statistical models and inferences crucially depend upon the extent to which the data actually satisfies the model's assumptions. When these are not upheld inferences can be biased and even rendered invalid.

To assess the validity of the GLM approach to fMRI time-series analysis we now consider the degree by which these conform to the $\mathcal{GM}$ assumptions and the tools that have been proposed to overcome failure of meeting any of the hypotheses.

0.3.1 Autocorrelation

The assumption of i.i.d. residuals (A1a) is necessary for the GLM model to allow for valid inferences. Yet, BOLD time-series suffer from many sources of serial correlation of residuals, the most typical being hardware related low-frequency drifts, oscillatory noises related to respiration and cardiac pulsation, and residual movement artifacts not accounted for by rigid body registration (see for example [12, 13] and [14]••, for reports of temporal autocorrelation sources in fMRI data). The presence of serial correlation does not directly affect unbiasedness of the $\hat{\beta}$, rather it affects its variance, computed by using

the residuals. Biased variance of an estimator will lead to biased T-values when assessing its statistical significance. Underestimates of the variance will lead to greater T-statistics and as a consequence overestimate significance (inflating type I errors) possibly leading to extremely liberal (i.e. invalid) inferences[6]. Viceversa, overestimating the variance will lead to smaller T-values, decreasing statistical power. Different approaches have been suggested to solve this problem.

Temporal Smoothing ("pre-coloring"). Friston *et al.* [15]•• and Worseley and Friston [16]•• suggest an extension of the GLM to accommodate serial correlation via "temporal smoothing". Their proposal is to reframe (1) as:

$$Y = \beta X + \Sigma\varepsilon \tag{4}$$

where Σ represents some process hidden in the residual characterizing the serial correlation, and ε represents a "well-behaved" error term $\sim N(0, \sigma^2 I)$. By then imposing a linear transformation S to (4) the idea is to "swamp" the endogenous – unknown – correlation structure with some exogenously imposed, therefore known, correlation structure S, obtaining:

$$SY = \beta SX + S\Sigma\varepsilon \tag{5}$$

If Σ were known then the best approach would be to clean the data (i.e. "whiten") by setting $S = \Sigma^{-1}$ thus compensating for the Σ processes. Being this process typically unknown the problem relies in "pre-coloring" the data by imposing an S correlation structure that may minimize the unknown Σ process. The proposed approach by Friston, Worsely and colleagues ([15]••, [16]••, [17]••) is to minimize the estimators' variance bias by setting the linear transformation matrix S to be a smoothing kernel similar to the

[6]The effective df_{model} are less than those expected in the independent case.

hemodynamic response function (typically a Gaussian or Poissonian distribution), which according to Friston *et al.* [15]•• maximizes the signal-to-noise ration (SNR). The assumption underlying this method is that the S Gaussian/Poissonian transformation is robust enough so that $S\Sigma S^T \sim SS^T$, effectively "swamping" the unknown endogenous serial correlation. If this assumption holds then (2), (3) can be re-written as:

$$\hat{\beta} = (X^{*T}X^*)^{-1}X^{*T}SY$$
$$var(\hat{\beta}) = \sigma^2(X^{*T}X^*)^{-1}X^{*T}SS^T(X^{*T}X^*)^{-1}$$

with $X^* = SX$ and where the "colored" noise can now be assumed to be i.i.d and $\sim N(0, \sigma^2 SS^T)$ (see [16]•• for a complete derivation of the adjusted estimates). These estimates of β do not retain maximal efficiency, which varies as a function of how effective the "pre-coloring" filter is in swamping the endogenous correlation, but do retain unbiasedness. A different approach to choosing the bias-minimizing transformation S is presented in Carew *et al.* [18]. The authors suggest using a "smoothing spline" approach.[7] The basic idea is to fit a spline at each voxel and then use the spline smoothing matrix as the linear transformation S in (5). The spline model for a time-series y_i is:

$$y_i = f(t_i) + \varepsilon_i$$

where y_i is the time-series associated to voxel i, $f(\bullet)$ is a smoothing function, ε is $\sim N(0, \sigma^2 I)$ and t_i for $i = 1...n$ (with N being the total number of TRs) are equally spaced volume acquisition time-points. The authors then show that the optimal smoothing value (thus the optimal S matrix) is given by application of a generalized cross-validation (GCV)

[7] A spline function is formed by adjoining polynomials together at fixed points called *knots*, to the aim of approximating a given curve with a series of intra-knots components.

method to the spline model (see [18] for the proof). The authors also provide an empirical validation of their method by showing the increased efficiency (i.e. reduced variance) of the $\hat{\beta}$s computed with the spline method as compared to both the Friston *et al.* [15]•• and Worsley and Friston [16]•• alike approach and a no-smoothing "baseline" analysis.

Pre-Whitening. In opposition to the "pre-coloring" approach Bullmore and colleagues [19]•• suggest a "pre-whitening" approach. This technique consists in removing an estimate of the autocorrelation to then fit the GLM model and find via OLS the BLUEs. To accomplish this [19]•• make use of a two step procedure. First, a GLM is fit to the data under i.i.d. error assumption, given the residuals the structure of their correlation is modelled with a simple Auto-Regressive model of order 1 (AR(1)) in which the error at each time-point is assumed to be a combination of the error at the previous time-point with some "fresh" error. Following, the raw data is "pre-whitened" by removing the estimated residual structure. Finally the voxel-wise GLM is fitted on the transformed (i.e. "whitened") signal. The intuition underlying such approach is that if a good data driven estimate of the residual autocorrelation structure can be computed and removed, then the i.i.d. assumption (A1a) will hold and the post-whitening $\hat{\beta}$(s) are BLUE(s). As a parallel to (5) the pre-whitening approach can be expressed as:

$$K^{-1}Y = \beta K^{-1}X + K^{-1}\Sigma\varepsilon$$

where $K \approx \Sigma$. Thus, instead of the convolution of a temporal smoothing matrix S as in Friston *et al.* [15]•• a de-convolution or "pre-whitening" matrix obtained by data driven estimation of the Σ structure of serial correlation is applied. Unlike the "pre-coloring" approach, inclusion of the K^{-1} deconvolution matrix will allow computed estimators to be the most efficient linear estimator of the unknown parameter β, within the set of its

unbiased estimators. Thus, equations (2) and (3) can be re-written as:

$$\hat{\beta}_{GLS} = (X^T K^{-1} X)^{-1} X^T K^{-1} Y$$
$$var(\hat{\beta}_{GLS}) = \sigma^2 (X^T K^{-1} X)^{-1}$$

If the estimation of the process Σ is exactly characterized by KK^T then in the transformed data:

$$KK^T = K^{-1}\Sigma(K^{-1})^T = K^{-1}(KK^T)(K^{-1})^T = I$$

thus the error variance is equal to $\sigma^2 I$ again.

Purdon *et al.* [20] note that additional white noise is produced by the scanner as a function of the rate of acquisition (i.e. TR). Their experimental data for instance indicated a negative relationship between the TR and false discovery rates for a Student's T-test, a Fourier F-Test and a Kolmogorov-Smirnov non-parametric test in fMRI data acquired under the null hypothesis. For this reason the authors propose a modification of the Bullmore *et al.* [19]•• approach by including, in addition to the AR(1) model, also a white noise model dependant on acquisition rate. Purdon *et al.* [21] compare the AR(1)+WN method with the standard SPM-GLM approach (i.e. [14]••, [15]•• and [16]••) finding the former superior to the second in both synthetic and human data in terms of better modelling of the noise. The sub-optimality of the latter is imputed to the use of the hemodynamic response as the smoothing kernel, which assumes the primary source of fluctuations in fMRI data to be real activations that can induce fluctuations that appear like response activations. In opposition Purdon and colleagues [21] use an empirically driven model of noise which incorporates different noise sources: low-frequency physiological noise (for

which the AR(1) is used) and scanner background noise (for which they use a drift term - the WN model). Their complete model thus characterizes the BOLD signal as being explained, at each voxel, by three components: (i) the drift (scanner background noise), (ii) the physiological signal, which is the actual signal (the convoluted X matrix) and (iii) the noise, which includes a well behaved ε and serial correlation process (modelled by an AR(1)) due to physiologic noise.

Another slightly different solution to serial correlation has been proposed by Locascio *et al.* [22] who assume the noise to be of (at least) two forms: "autoregressive" (AR) and/or "moving average" (MA). The authors therefore propose the use of an ARMA model which should ameliorate the approximation of the actual serial correlation structure by sequentially testing (voxel-wise) the significance of several orders of AR models and MA components. Besides the double characterization of possible noise, one important contribution of [22] is to allow for local (spatial) variation in specifying voxel-wise serial correlation. By iterative procedure each voxel is independently fit with an increasing number of ARMA components until they pass a test of residuals "whiteness".[8] This approach has two main advantages over the Bullmore *et al.* [19]•• (and Purdon *et al.* [20]) approach. First, it responds to the criticisms of inadequacy of an AR(1) model to capture the correlation structure (see [17]••) by allowing for higher order autoregressive models. Second, it is much more flexible by allowing each voxel to have a variable number of AR and MA components according to its specific time series profile, as opposed to the a uniform AR(1) model applied across the whole volume.

As compared with temporal smoothing, pre-whitening has the advantage of being more efficient (see [23]) across different experimental designs (e.g. Boxcar, randomized

[8]The software implementation of the ARMA model described in [22] only allows for up to third order AR and MA models. From a theoretical standpoint many more could be used, although each additional model affects the subsequent analyses by degrading the number of degrees of freedom, thus the power of the ARMA-whitened model.

Event Related) — albeit under the non trivial assumption of satisfactory characterization of the endogenous correlation structure — but is more prone to bias in case of non-optimal serial correlation modelling. Indeed, Friston *et al.* [17]•• report a bias of 24% in parameter estimation when an AR(1) model (as proposed in [19]••) is (mis-)used to estimate a serial correlation structure that is actually approximated by an AR(16). The effect of such a bias, in [17]••, is to reduce by about 10% the T-values, a substantial loss of power. Pre-coloring on the other side may risk attenuating part of the high frequency features in the data that may be experimentally determined ([24]) and/or may convey specific localizing information ([25]). Overall, though, as noted by Bullmore *et al.* [26] there still may be insufficient data to definitively judge the adequacy of either approach.

Explicit Noise Modelling. Finally, a promising alternative approach to the data driven modelling of serial correlation structures has been proposed by Lund *et al.* [27]••. The authors approach serial correlation as a source of information regarding excluded variables (that thus show up as serial correlation), rather than as a simple nuisance (as in the temporal smoothing and, partially, the pre-whitening approaches) and attempt at "suggesting a unified theory of physiological noise in fMRI". In their Nuisance Variable Regression (NVR) approach Lund and colleagues explicitly model specific factors known to be inducive of autocorrelation: hardware instability-related low-frequency drift, residual movement effects and aliased physiological noise (i.e. cardiac pulsation and respiration). Specifications of these effects are then used as regressors in the GLM X matrix. The authors also test empirically the performance of their NVR approach; two results are noteworthy. First, the physiological effects resulted significant in predicted brain regions, thus validating the used specifications (see [27]•• for a review of evidence on the specification of the three noise components). The cardiac induced noise, for instance, was found dominant near the major arteries of the brain (e.g. CW and MCA). Second, the NVR resulted

superior to both the AR(1) and the simple high-pass filtering approaches for dealing with serial correlation. A possible criticism to this approach is the fact of relying uniquely on specification of the three sources of noise described, while other un-modelled factors could be present, thus not guaranteeing an error term as required by $\mathcal{GM}$ A1a. The robustness of this methodology may thus be questionable and may require integration with supplementary noise-cleaning to capture remaining sources of serial correlation. On a positive note though this approach holds great potential especially in perspective of the ever growing understanding and characterization of noise sources.

0.3.2 Heteroscedasticity

Assumption (A1c), requires $var(\varepsilon) = \sigma^2 I$, thus that the variance of residuals, assumed to be a scalar, is constant across the different observations (i.e. time-points), and that their paired covariances (i.e. the off-diagonal elements of the variance-covariance matrix) are all equal to zero. Violation of such assumption is referred to as heteroscedasticity. When this assumption is not upheld the estimator $\hat{\beta}$ is still unbiased, but no longer efficient (it is no longer the minimum variance estimator). As for the case of autocorrelation, if the variance is biased subsequent paramteric testing will yield incorrect statistics.

In the fMRI literature heteroscedasticity hasn't received much attention. One of the few exception is Luo and Nichols [28]• who created a tool to assess whether a given fMRI data-set does conform to the $\mathcal{GM}$ assumptions. The authors mention the possibility of heteroscedasticity in fMRI data, for example due to a dependency of the variances on the response, or because of other factors such as time or physical ordering. For these reasons they include a specific diagnostic test in their package to assess whether the hypothesis is upheld (see [28]•, p. 1016). As a testimony to the little attention this assumption has received a SCIENCE@DIRECT search revealed 6 citations of this work. One represented an actual use of the tool for data diagnostics, whereas all the other were presentations of

novel statistical techniques that didn't address the homoscedasticity assumption in any detail.

0.3.3 Multicollinearity

$\mathcal{GM}$ (A3) requires that none of the explanatory variables (i.e. columns of the X matrix) is perfectly correlated with any other explanatory variable, or any linear combination of. The problem with violation of such assumption is that the X matrix is no longer invertible, thus the BLUE estimator can no longer be computed. Use of the Moore-Penrose pseudo-inverse (see footnote 4) allows for avoiding non-invertibility of the X matrix but still leaves with a problem of degrees of multicollinearity, from perfect (when all columns are perfectly correlated – in which case also the pseudo-inverse can't be computed) to mild (when only one or few columns are linearly correlated). According to the extent of multicollinearity the efficiency of the estimate will be reduced thus vitiating parametric testing by degrading power.

As for the homoscedasticity assumption the multicollinearity issue didn't find so far much space in the fMRI methodology literature. There may be at least two reasons for this: first, the standard use of the pseudo-inverse method and second, the fact that this problem is typically dealt with and prevented at creation of the experimental design (i.e. the X matrix).

One interesting point though is raised in relation to the X matrix specification by Petersson *et al.* [29] who stress the importance of its appropriate specification. The model specification, according to the authors, faces two connected trade-offs related to the cases of over- and under-specification. On one side inclusion of a maximum number of effects in the model would be desirable to increase fit, yet at the cost of reducing power by consuming one $d.f.$ for each additional effect, while the marginal increase of explained variance decreases with each additional factor. Further, over-modelling of the signal may

degrade the generalizability of the results. On the other hand though, exclusion of regressors from the model may have the effect of inflating the error variance, reducing power, and possibly introducing serial dependencies in the error term, thus infringing assumption (A1a) of the $\mathcal{GM}$ theorem. It should be noted though, that exclusion of effects from the model has also the (positive) consequence of increasing power via increase of the $d.f._{model}$ (one per each excluded variable). (See [29], pp.1246-7 for a complete discussion on the point).

0.3.4 Linearity

The GLM approach also assumes the effects to add linearly to compose the response measurements. Boynton *et al.* [13] tested the GLM by parametrically varying a visual stimulus' duration and contrast, and investigating the additivity of the noise in V1. Their conclusion was that although deviations from linearity were measurable, these were not strong enough to reject the GLM. Support for the use of a linear approach is also offered by Cohen [30] who showed that the amplitude of the responses to parametric variations of the stimuli well fit a piecewise linear approximation. Despite this initial evidence, it has been shown that there are at least two sources of non linearities in the BOLD signal. One relating to the vascular response – especially the vasoelastic properties of the blood vessel (see [31]) and the other relating to non-linearities at the neuronal level due to adaptive behavior (see, for example [10]).

Evidence for the presence of nonlinearities has been offered by Vazquez and Noll [32] who compared the linearity of the response with parametric increases of visual stimuli duration from 1 to 8 seconds. Results pointed out that while stimulations greater than 4 seconds were approximated well enough by a linear model, shorter displays lead to great discrepancies between the actual and the expected response. A similar result was reported by Robson *et al.* [33]•• where the authors used as stimuli trains of sounds of dif-

ferent – parametrically varying – durations (from 100ms to 25.5s). Under the assumption of linearity of the response it should be possible to predict the amplitude of the response at a given duration via a (linear) combination of the amplitude response at some other (shorter) stimulus duration. Consistently with the results in [32] it was possible to predict amplitudes only when using, as the amplitude predictor, the response amplitude of a trial with duration greater than 6 seconds. Using shorter stimuli as predictors resulted in massive overestimation of the response amplitude as a positive function of the time difference between the predictor and the predicted condition (see [33]•• Fig. 4, p. 191 for a dramatic depiction of these results). The authors thus suggest including in the model an adaptive component that may discount the response amplitude for short stimulations specified as:

$$E(t) = (1 - A) + Ae^{-t\alpha} \tag{6}$$

Equation (6) essentially represents a scaling factor to be applied to the amplitudes of short latency stimuli in order to correct for the "transient" non-linearity. Applying this transformation to the amplitudes allowed the authors to reduce the discrepancy (for example) in the prediction of the longest latency (25.5 sec) from the shortest (100ms) from an average of 11.09% signal change to .88%.[9]

Friston *et al.* [34]•• also used parametric variations in the rate of word presentation to assess the presence of nonlinear BOLD effects. The significant departure from linearity was interpreted in terms of a hemodynamic 'refractoriness', according to which a prior stimulus interacts with a following (temporally contiguous) stimulus by modulating its response amplitude. To solve for the presence of significant departures from linearity Friston and colleagues [34]•• proposed to generalize the standard GLM approach ([15]••,

[9]The values of parameters A and α were computed empirically by minimizing the discrepancy between predicted and actual signal.

[16]••) using Volterra series to "linearize the problem" by characterizing the non-linear component of the response. The signal y(t) is then characterized as:

$$y(t) = g^0 + \sum_{i=1}^{P} g_i^1 x_i(t) + \sum_{i=1}^{P}\sum_{j=1}^{P} g_{ij}^2 x_i(t) \cdot x_j(t) + e(t) \tag{7}$$

Equation (7) is a GLM with the response $y(t)$ predicted by the explanatory variables $x_i(t)$ and $x_i(t) \cdot x_j(t)$, and parameters g^0, g^1, g^2 representing the scaling factor of a series of P basis functions approximating the zeroth, first and second Volterra smoothing kernels $h^0, h^1(\tau_1), h^2(\tau_1, \tau_2)$ (see [34, p. 42] for the full derivation):

$$h^0 = g^0$$
$$h^1(\tau_1) = \sum_{i=1}^{P} g_i^1 b_i(\tau_i)$$
$$h^2(\tau_1, \tau_2) = \sum_{i=1}^{P}\sum_{j=1}^{P} g_{ij}^2 b_i(\tau_i) \cdot b_j(\tau_2)$$

In equation (7) the second term represents the change in output for a change in input. The third term is the part of the model that includes the interactions of the response at one point in time on the response amplitude at a contiguous time. Resolving for the parameters g^0, g^1, g^2 then allows for computation of the Volterra smoothing kernels, testing for significance of nonlinear effects and testing on the hemodynamic response at each voxel.

One criticism to this approach noted in Calvisi *et al.* [35] and Friston *et al.* [36]•• is that while data driven computation of Volterra series parameters may allow for a better mapping of the input to the output, it does so in a black-box fashion without being informative on what are the processes generating the nonlinearities. In response to these

criticisms Friston and coworkers [36]•• present evidence for the nonlinearities expressed in the Balloon model of hemodynamic signal transduction (see [31]) being compatible with a second order Volterra characterization, thus adding biological plausibility to the model.

A different approach has been proposed by Wager *et al.* [37] who report substantial nonlinearities in the magnitude, peak delay and dispersion of the (hemodynamic) response for a stimulus presentation rate of 1s. Noting the consistency of such nonlinearities across the brain they suggest empirically deriving the functional form of each of these characteristics of the response as a function of stimulus history. The authors chose to approximate the nonlinearities with the biexponential model:

$$y = Ae^{-\alpha x} + Be^{-\beta x}$$

By fitting the parameters A and α, B and β, the scaling and exponent of two exponential curves, the authors empirically characterize the nonlinear changes in BOLD magnitude, onset time and peak delay. The idea is to first run an experiment from which to derive the fixed parameters estimates and then use the nonlinear characterizations as scaling factors for individual responses – according to the history of stimulation up to each response – in following experiments.

Overall though, as noted by Wager and colleagues [37] nonlinear effects are largely ignored in the neuroscientific and psychological studies using BOLD fMRI. The authors suggest that at least three reasons may explain why consideration of nonlinearities is minimal in this literature. First, the linear approximation of the BOLD signal becomes unfit only in a restricted range of stimuli spacing (i.e. $<\sim 5s$). Second, the bulk of the work has been devoted to determining canonical responses to a single stimulus rather than to exploring interactions among multiple stimuli. Finally, most proposed solutions

(e.g. Volterra series, see [34]•• and [17]••) require fitting of a large number of parameters which may cause severe degradation of power.

0.4 Multiple Subjects Analysis

Once individual SPMs for some effect of interest have been computed for multiple sessions (from a single subject) and multiple subjects the second step of the analysis is to aggregate them to allow for more general inferences to assess whether the effects found in the single-subject analysis are common and stable between or across groups of interest ([38]••, [39]••). Prior to running aggregate analysis though individual data needs to be transformed into some "standard" three-dimensional space, typically either in Talairach ([40]) or MNI152 ([41]) space. This transformation allows for alignment of corresponding cerebral structures across subjects with differing brain anatomy. The normalization procedure of data is all but uncontroversial, especially in relation to its effectiveness (see [42] for a brief review of the inherent problems of normalization), although a discussion of the issue extends beyond the scope of this review.

Following computations of individual SPMs the question is then how to combine these maps to make inferences pertaining to the group of sampled individuals and, desirably — and more interestingly — to the population from which the sample is drawn. This assessment is the very purpose of multi-subject analysis. Several approaches have been proposed to multi-subject fMRI data analysis, each with its merits and pitfalls; most importantly it should be noted that different methods may bear on the type of conclusion that can be made (e.g. validity with respect to sample, validity with respect to sampled population).

0.4.1 Fixed Effects.

The most simple and straightforward way to analyze a multi-subject data is to concatenate the time-series obtained from each subject and run a GLM ([43]). This approach is a so-called "fixed effects" analysis (FFX). There are two main differences between the single-subject GLM and the multi-subject one. First, the former includes in the Y time-series a concatenation of the data obtained from each session of a given subject. Contrarily the multi-subject GLM includes in Y a concatenation of all runs from all subjects. To accommodate for this a second difference is usually introduced in the X matrix which now contains one regressor column for each effect *per* each subject. That is, given an "ON" condition (for example), the X matrix will include for each subject one "ON" condition regressors, each specifying when the given subject was undergoing such condition. This is to say that the design matrix is specified as a subject-separable GLM (which also allows for creating contrasts comparing – or isolating – activity of one or more given subject). The relevant question is thus what can be inferred, validly, from such analysis. As Holmes and Friston nicely put, a classical statistical hypothesis test proceeds by comparing the difference between the observed and hypothesised effect against the "yardstick" of variance ([38]•• p. S745). Thus, the scope of an inference can be assessed by considering the used yardstick. FFX make use *scan-to-scan* variability, which, according to Friston and colleagues [44]•• includes physiological task-related (e.g. adaptation, learning and strategic changes in cognitive or sensory-motor processing) and task non-related variability (e.g. changes in global perfusion secondary to vasopressin (AHD) secretion in the supine position), and non-physiological noise (e.g. gradient instabilities). In other words FFX analyses only make use of within-subject variability (σ_w^2) when computing significance testing. In the interpretation of Penny and Holmes [45]•• FFX analyses represent the population variance as being a sole function of within-subject variability divided by

the product of subjects (N) and number of sessions per subject (p)[10]. In their example, the effect size for a given subject i at a specific session j is equal to the subject mean effect plus some session specific error. Thus, formally:

$$d_{ij} = d_i + e_{ij} \tag{8}$$

(Note that e_{ij} is σ_w^2 for subject i, and is assumed equal across subjects.) Thus the parameter estimate and its variance for a given subject i are given by:

$$\hat{d}_i = \frac{1}{p}\sum_{j=1}^{p} d_{ij} \tag{9}$$

$$Var(\hat{d}_i) = \frac{\sigma_w^2}{p} \tag{10}$$

Similarly, the population effect in this approach is given by the simple aggregation of all the individuals' effects (i.e. (8)). Thus, aggregating the subjects' estimates (d_i) the population estimate (and its variance) would thus be:

$$\hat{d}_{pop} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{11}$$

$$Var(\hat{d}_{pop}) = \frac{1}{N}Var(\hat{d}_i) = \frac{\sigma_w^2}{Np} \tag{12}$$

The crucial point relies in the fact that the estimate of the population effect d_{pop} in the FFX approach is a function of only the scan-to-scan (i.e. within-subject) variability σ_w^2, as (12) makes clear. In this sense, running the GLM with the design matrix differentiated by subject, except for allowing for inspection or comparison of individual subjects, is equivalent to concatenating all data and using a same regressor for all subjects as in an

[10]In fact [45]•• makes use of n to define the number of sessions per subject, here p is used to avoid confusions due to the unfortunate use of N and n to indicate different things.

experiment with one "super-subject" whom underwent $N \times p$ sessions.

It should be clear though that the inferences drawn from a FFX analysis are not invalid, rather they are only valid in reference of the used yardstick. Inference is thus supported at the level of the sample analyzed but not in reference to the population from which this sample is drawn (given that there is no consideration of "sampling variability"). As noted by Friston and collaborators [44]•• a FFX approach makes the assumptions that each subject makes the same contribution to the observed activation thus discounting random variation from subject to subject (see the data presented in [45]••, Figure 2 for a dramatic example of *subject-to-subject* variability). This type of analysis can thus be seen as relevant in "single case" studies ([45]••), but seems unacceptable for the standard fMRI experiment and its (desired) inferential scope.

0.4.2 Random Effects.

To overcome the limited scope of inference-making from FFX analyses a different yardstick should be used. As Beckmann and colleagues observe, in order to generate results that may support inferences about the population, it is necessary to account for the fact that individual subjects themselves are sampled from the population and thus random quantities with associated variances ([46], p. 1053). In other words, the yardstick must also account for the subject-to-subject (i.e. between-subject, σ_b^2) variability. Indeed, Friston *et al.* [44]•• note that there are several reasons for assuming that such variation is present in fMRI data and that that this can be due to any (and any interaction) of several factors such as general subject differences in neural or hemodynamic response to stimulation, and/or differing underlying anatomy. Further, any of the above-mentioned within-subject variations may be of different magnitude across subjects and, finally, many non-physiological noise sources could affect the way in which a BOLD effect (even assuming this was actually the same across several subjects) could give rise to different data (e.g. radio-frequency

and gradient instabilities, re-calibration of the scanner, repositioning effects or differential shimming effects). Finally, the authors also remark that the subject-to-subject effects are typically much greater than scan-to-scan effects.

A "random-effects" analysis (RFX) accounts for such variations by including in the error term also sampling variability (i.e. departures of the subject effect from the population effect). Keeping with the explanation in [45]••, a RFX analysis replaces the FFX assumption (8) with:

$$d_{ij} = d_i + e_{ij} \tag{13}$$

$$d_i = d_{pop} + z_i \tag{14}$$

Thus, at the level of the individual session the within-subject variability is equivalent across the two approaches (compare (8) to (13)), but the RFX also considers the effect of a single subject d_i as having a $E(d_{pop})$ plus some variability z_i. Note that the variability is σ_b^2 (i.e. across-subject variability), therefore the effect of a single subject $d_i \sim N(d_{pop}, \sigma_b^2)$. Thus, the RFX parallels to (11) and (12) would be:

$$\hat{d}_{pop} = \frac{1}{Np} \sum_{i=1}^{N} \sum_{j=1}^{p} d_{ij} \tag{15}$$

$$Var(\hat{d}_{pop}) = \frac{\sigma_w^2}{Np} + \frac{\sigma_b^2}{N} \tag{16}$$

It is immediately clear from (16) that in RFX analyses the yardstick used for statistical testing encompasses both the "within" and "across" sources of variability and thus supports inferences at the population level.

0.4.3 Conjunction Analysis.

The intuition underlying this approach was first developed by Price and Friston [47]•• who, critic of the standard cognitive subtraction technique (e.g. Task - Baseline), suggest testing for brain activations by "triangulating" on brain areas which should satisfy two conditions: (i) be jointly active across different subtractions (e.g. (Task A - Baseline A) and (Task B - Baseline B)), although (ii) not significantly different across the different subtractions. The basic idea is to create different task-baseline pairs, each isolating the process of interest (typically along with some "accessory" activations elicited by the specific task pairs, to be discarded at the second stage). Conjunction analysis can thus be thought of as testing for the activating effects of a given process of interest in a set of different contexts (i.e. task-baseline pairs) which retains activations equally robust across contexts ("context-independent" activations) while eliminating activations of significantly different intensity across contexts ("context-dependent" activations).

The implementation of this conjunction simply requires standard creation of sum SPMs (i.e. sum of all voxels active in all the different task-baseline pairs) and then elimination of all voxels that show significant difference across subtractions ("interaction masking" procedure, see [48]). In factorial analysis jargon, a conjunction is equivalent to searching for a main effect (e.g. the common underlying psychological-neural process), in absence of specific task $\times$ psychological-neural process interactions. If a given activation does abide by both (i) and (ii) above, then it must reflect some common mechanism to all task $-$ baseline pairs (thus a main effect), if instead it abides by (i) but not by (ii) then it must be reflecting some process relating to the specific task $-$ baseline subtraction and/or its interaction with the main effect.

As noted in Friston *et al.* [44]•• FFX models are very sensitive analyses, having substantially more $d.f.$ and being the scan-to-scan error typically smaller (thus typically

$t(FFX) > t(RFX)$) although they have the drawback of circumscribed validity to sample-specific inferences. On the other hand the price for the generalizability of results provided by a RFX approach is potentially very small T-statistics (given that subject-to-subject variance is usually much greater than scan-to-scan variation), and thus a very conservative analysis. Further, Friston and Holmes [49] also note that it may require many subjects to reliably assess the subject-to-subject variability. To try to retain the sensitivity of FFX but still have a methodology supporting population-valid inferences Friston *et al.* [44]•• suggest a development of the conjunction analysis proposed by Price and Friston [47]•• that, despite employing FFX models can still be used to make population inferences about qualitative responses (e.g. activated vs. non activated) in terms of confidence intervals for the proportion of population showing the effect. In this paper the authors suggest to first analyze each subject individually with a FFX conjunction approach and then applying meta-analytic approach to assess reliability of the activations across subjects. The population-wise inference is thus predicated upon the use of confidence intervals for the proportion of population that is likely to show the effect(s) identified in the single-subject FFX analysis. The first-level analysis is essentially analogous to that proposed in [47]•• where the voxels of interest are those that jointly survive multiple (and different) subtractions, although without being significantly different across them. The only innovation at this point is the use of a common significance threshold value for the multiple subtractions computed according to Worsley and Friston [50]. In this approach the theory of random fields (TRF) is applied to 3D brain images to compute the probability of finding a conjunction anywhere in the brain and thus setting an inferior bound on the probabilities required for an conjunction to be considered as non casual. The multi-subject analysis approach is instead novel, and pivots on the construction of confidence intervals which should contain, for a given (TRF-corrected) α level, the proportion γ of subjects in the population showing the effect. Specifically, the authors first represent $P(n)$ as the proba-

bility of a certain number n of individuals showing an effect, and define it in terms of the specificity α, power and the probability γ of a random subject showing the effect. Then, in relation to some critical proportion of population γ_c one wishes to test the presence of the effect for, Friston and colleagues find an upper bound for $P(n|\gamma < \gamma_c)$ function of the corrected-α level (i.e. not the family-wise α) and γ. By imposing these two measures, the upper bound in fact yields a p-value corresponding to an inference on how typical the effect is (where γ_c sets the degree of "typicality") given the specificity (i.e. α). Equal or smaller values will not support rejection of the null hypothesis that the proportion of population showing the effect is less or equal to γ_c. Further, by imposing this p-value to be equal to the family-wise (desired) false discovery rate α_c (which thus is the specificity at the population level), the critical proportion of subjects showing the effect γ_c can be expressed as a sole function of the family-wise and local (i.e. the specificity used for individual testing) α levels. The conjunction approach can therefore be interpreted as saying that with specificity of $(1 - \alpha_c)$ a proportion of the population greater than γ_c shows the effect. The authors note that this is identical to saying that this is a $100(1 - \alpha_c)\%$ confidence interval for the unknown parameter γ. This analysis thus allows to conclude that a sampled subjects activated a given region (using a FFX model) and that with some desired confidence α_c at least a certain proportion γ_c of the population would have shown this effect. It should be noted that these inferences can be constructed as statements that describe the typicality of the effect, without though assuming that it is present in all subjects.

Finally, in the discussion section, the authors make two important remarks. First, comparing RFX to conjunction analysis they note that whereas the former is a quantitative statement, positing for the average effect to be greater than zero, the latter is only a confidence region over a categorical effect (i.e. activate versus non-activate) that only requires that a (substantial) proportion of the subjects to show the effect. Second, the authors also

note that this approach is not meant to replace RFX approaches, to which in some cases there is no alternative, but rather to enable researches to make the best of small data-sets (e.g. pilot data, case studies).

The idea of conjunction analysis has not been without resistance. Caplan and Moo [51] for instance oppose the merits that [47]•• ascribe to the cognitive conjunction approach, especially the fact that it solves the problem of setting an appropriate baseline task and the "insertion" problem (see [51, 47] for a review of both issues). Indeed, they note that the baseline problems (e.g. implicit processing) apply just as forcefully to the scenario reported in [47]••, and that the statement that "conjunction looks for commonality in activation differences between two or more pairs of tasks that share *only* the component of interest"([47]••, p261) uncovers the fact that conjunction analysis still does rely on the appropriate setting of a baseline, not differently from the standard subtraction method. The second merit of conjunction analysis according to [47]•• is that it eliminates the need to assume that the cognitive process of interest added at task (as compared to baseline) does not interact with other components present in both the baseline and the task (so called "pure insertion" hypothesis). The need for assuming that a new cognitive component (the effect of interest) can be inserted without affecting other components is important for the subtraction results to be interpretable. With conjunction though, elimination of interaction effects (the interaction masking) eliminates the need to assume pure insertion. Yet, as Caplan and Moo [51] point out, for conjunction analysis to get rid of interaction effects, interactions resulting from different subtractions must not overlap, else the corresponding region will be mistaken for a "context-independent" one when in fact it is responding to (more) specific contents. However, just as for the baseline problem, conjunction analysis seems to assume that the problem will not be relevant, given the unlikeliness of two interaction areas to be overlapping, rather than solve it, as the technique proponents argue. In addition to the two above criticisms, the authors also note that while on one side

the conjunction analysis is weak, not solving the problems it was set to overcome, on the other it may impose unwanted restrictions on the identification of regions associated with a specific cognitive process. Specifically, should one of the task-baseline pairs engage the regions responsive to a given cognitive process with greater intensity than the other ones, the conjunction approach will deem them as interactions and eliminate them; potentially excluding areas of interest.

On different grounds also Nichols *et al.* [48] opposes the procedure proposed by [47]•• and [44]••, noting that the second step of the interaction masking procedure relies on the null result of failing to reject an interaction effect at a given voxel. Using statistical tests to define regions where there is *no* interaction is "accepting the null hypothesis", yet lack of evidence is not evidence of lack. Further, an even more severe problem with the Friston and colleagues approach, relates to the statistical testing of joint significance of the effects (see [44]•• and [50]). As described by Nichols and colleagues [48] the statistical procedure is to first assess the individual significance of effects (this would be the first level analysis) and then make use of a *minimum statistic* procedure to assess the joint reliability of the results. The minimum statistic test is predicated on the following logic: given a voxel i associated with one t-value per each tested effect (say effect A and effect B), the two effects may for example not reach significance – individually – yet the fact that both are greater than zero suggests a real effect could be present. To test whether this is the case a test on the minimum t-statistic can be made, under the assumption that if there is no effect then both effects will be drawn from a null t-distribution. Here is the logic pitfall: the null hypothesis in Friston *et al.* [44]•• is then that *neither* effect A *nor* effect B are significant, which, in logical terms is a conjunction of two negations. Now, if a conjunction is defined as a voxel for which both statements A and B are true, its negation (and thus the null hypothesis) should be represented by any state in which at least one (or both) statements is not true (i.e. $(\neg A) \vee (\neg B)$ with $\vee$ interpreted inclusively). Yet the

null hypothesis in [44]•• is a conjunction of two negations, implying that the alternate hypothesis being tested is in fact an "inclusive OR" rather than an "AND". Nichols and colleagues [48] remark that the test is still valid, simply this *global null hypothesis*, when rejected, only allows for the inference that at least one subject shows the effect (i.e. an inclusive OR statement). Finally, the authors conduct a literature search and report that within a sample of 42 abstracts submitted (and accepted) to Organization for Human Brain Mapping (OHBM) reporting the word "conjunction" or "conjoin" 23 erroneously concluded a strict conjunction of effects (e.g. $A \wedge B$) upon using the [44]•• procedure, six were unclear and 4 correctly inferred a conjunction of at least one or more effects.

0.4.4 Mixed-Effects & Summary Statistics Hierarchical Approach

A different approach to group analysis is first proposed by Holmes and Friston [38]•• who envision a mix-effects hierarchical model of data processing that may take into account all the sources of variability (i.e. within and across). Their intuition is to first fit the GLM at the individual level and write out the results as images (SPMs), and then with a simple t-test assess these activations across multiple-subjects, thus subsuming the RFX idea that the effect in a given subject is a function of the true population effect plus an error.

Penny and Holmes [45]•• develop further the idea of doing a RFX analysis *via* this hierarchical summary-statistics approach, which they conceive more explicitly as a two-levels model based on the sample mean $\bar{d}_i$ rather than on the individual subject $\times$ scan effect d_{ij}. The model would thus be:

$$\bar{d}_i = d_i + e_i \tag{17}$$

$$d_i = d_{pop} + z_i \tag{18}$$

At the first level (equation (17)) the variation of the sample mean for each subject around

the true mean for each subject is considered (the within-subject variation component), whereas at the second level (equation (18)) the variation of the true subject means about the population mean is considered (the across-subjects variation component). The population-wise validity of this hierarchical summary-statistic approach lies in the fact that the sample mean, which contains the within variability element, is brought forward from the first level to the second. The estimate of the population mean ($\hat{d}_{pop}$) therefore contains contributions from both the within and across components of variance. This claim is readily apparent by substituting (17) in (18):

$$\bar{d}_i = d_{pop} + z_i + e_{ij}$$

Beckmann and coworkers [46] independently develop a very similar approach that, in consideration of the magnitude of typical fMRI data sets capitalize on the summary statistic hierarchical approach and build a "single complete mixed-effects" model, equivalent to the hierarchical two level analysis (see [46], section $II.C$ for proof of the equivalence). The authors assume a standard GLM approach for analysis of subject i, specified as in (1) In their specification the two level model is represented by:

$$Y = X\beta + \varepsilon \tag{19}$$

$$\beta = X_G\beta_G + \eta \tag{20}$$

where X_G is the group-level design matrix (e.g. separating groups or conditions per each subject) β_G is the vector of group-level parameters, and η specifies the residual of the group activation (parameter) scores (i.e. $Cov(\eta) = Cov(\beta_G)$), with $E(\eta) = 0$, $Cov(\eta) = V_G$, and $Cov(\epsilon) = V$, which is the block-diagonal form of the first level covariance matrices V_i. Thus, (19) and (20) can be re-written in the following compact form:

$$Y = XX_G\beta_G + \gamma \tag{21}$$

with

$$\gamma = X\eta + \varepsilon$$

where $E(\gamma) = 0$ and $Cov(\gamma) = W = XV_GX^T + V$. Application of a GLS approach yields a "one-step" BLUE of the group-level parameter vector and its variance as follows (see [46], p.1054, for the full derivation):

$$\hat{\beta}_G = (X_G^T X^T W^{-1} X X_G)^{-1} X_G^T X^T W^{-1} Y$$
$$var(\hat{\beta}_G) = (X_G^T X^T W^{-1} X X_G)^{-1}$$

It should be noted that the main purpose of this one-step model is to be able to efficiently test for general hypothesis for a mixed-effects group analysis only making use of first-level results (i.e. parameter and variance estimates), without the need to "revisit" the actual fMRI time-series data. It should be noted though that all first (and second) level assumptions (e.g. $\mathcal{GM}$ Assumptions) still need to hold for (21) to yield a BLUE.

0.4.5 ANOVA Approach.

The intuition of "carrying-over" results (i.e. first level estimates and variances) into a second level analysis was developed in a slightly different way also by Bosch [52] who suggests an analysis of variance (ANOVA) approach to multi-subject data following an initial voxel-wise t-test. The idea is to include in the group analysis the individually computed SPMs in the group analysis rather than the actual individual time-series. In the simplest

example, a 2 conditions experiment, each voxel of each subject's run is associated with a t-statistic resulting from the comparison of the two conditions. The aggregated data thus has the form of a distribution of N (total number of subjects) statistics per each voxel, each obtained at hypothesis testing for individual subjects. Assessment of significance of this distribution (i.e. whether it is reliably different from zero) requires a simple t-test. Further, [52] notes that by making use of z-scores, the voxel-wise t-test would simply equal the average z-score multiplied by the square root of N, given that the mean variance of the z-score under the null hypothesis is known and equal to 1 (and mean is 0). Simple comparison of the t-statistic with the expected t-values under the null hypothesis will yield a map of voxels significantly active, across subjects. Other standard tools (e.g. contrast analysis) can then be used to test specific hypothesis regarding each voxel's aggregated distribution.

0.4.6 Variance Smoothing.

Finally, Worsley *et al.* [39] propose a "variance smoothing" method (see [53]), in response to the RFX approach problem of lack of $d.f.$. The authors propose an "intermediate" model that varies in between the two extremes of a FFX and a RFX analysis according to the amount of smoothing, which is a function of the actual $d.f.$, thus of the design and of the number of sessions per subject and subjects. In this approach a model much alike (18) is first built, where the effect of a given subject at run j is a function of a linear specification matrix z weighted by the parameters γ plus an error $\eta_j \sim N(0, \sigma_j^2 + \sigma_{random}^2)$ where σ_j^2 represents the standard deviation of the effect E_j in run j from the average value across runs (thus scan-to-scan error), and σ_{random}^2 represents the random effects variance (the subject-to-subject variability). It is intuitive that setting $\sigma_{random}^2 = 0$ makes the model a FFX analysis. The main problem is that usually the $d.f.$ are low, and reliable estimates of σ_{random}^2 are difficult to obtain, and will usually have great variability. A way to reduce

such variability is to smooth the observed variance, although as noted by [53] this has the drawback of assuming constant variance over the brain (i.e. spatial stationarity), which does not seem to be the case (e.g. differences across gray and white matter). So, Worsley and colleagues [39] replace this assumption with that of local stationarity of the RFX variance to FFX variance ratio. The idea is to first run a FFX and a RFX analysis, then take the ratio of the variance estimates (where $\hat{\sigma}^2_{random}$ would be a "bad" estimate of the true between variance due to the low $d.f.$) smooth them and then multiply the smoothed ratio by the FFX variance (thus effectively using it as a template for a better estimate of the between variance). Formally, their idea is to create a better estimate of the subject-to-subject variance as follows:

$$\tilde{\sigma}^2_{random} = smooth(\frac{\hat{\sigma}^2_{random}}{\sigma^2_{fixed}}) \times \sigma^2_{fixed}.$$

This estimate of the between variance can be used to get a better estimator of the parameter vector γ and its variance. It is immediate that the two extreme smoothing solutions, no-smoothing and infinite-smoothing, render the analyses a pure RFX or a pure FFX, respectively. The question then relies to what is the appropriate degree of smoothing to be applied. [39] argue, on empirical bases (i.e. simulations) that the best smoothing kernel is the one that yields at least $d.f. = 100$ for the regularized variance ratio (see [39], APPENDIX C for the computation of the $d.f.$ of the smoothed ratio, which in their simulation appears to be equivalent to $15mm$ FWHM smoothing).

0.5 Conclusion

In the past fifteen years fMRI has assurged as one of the primary tools for *in vivo* imaging of human cognition. Yet, the typical GLM approach to data analysis – currently the most used analysis strategy – does not seem excessively fit to fMRI time-series given

the fact that these do not conform to most of the requirements of the model. At both a first "single-session/subject" and second "multi-subject" levels several assumptions of the model simply do not hold. Nonetheless the intuitive nature of the GLM approach has fostered a (quite successful) effort to resist these shortcomings by using several "corrective strategies" (e.g. data pre-whitening or pre-coloring for autocorrelation). Two final considerations should be made. First, one should wonder whether this "curing the symptoms" approach and its stubborn will to employ the GLM is effectively the best direction to go, as opposed to alternative routes such as exploratory analysis (e.g. ICA, PCA, ISC), non-parametric methods or bayesian approaches. Not to say these should be exclusive, but a greater awareness of model availability could impose a beneficial competitive dimension across analysis strategies. Second, there seems to be somewhat a dyscrasy between the sophisticated research relating to statistical analysis and the actual use by non-technical groups of such tools. At least two reasons may account for this. On one side the literature is hard to evaluate. Most articles propose a new analysis strategy, either in response to a pitfall of the GLM model or to some other strategy, and typically support their proposition by "comparative validation" of the procedure on either synthetic or human data. Such validations are not uncontroversial. (For example the validation of the "pre-whitening" technique in Bullmore *et al.* [19]•• has often been questioned as applying only to the specific data and conditions used in the very validation, e.g. see [17]•• p. 197.) Comparative studies by "third parties" are completely lacking. On the other hand, of all these proposed strategies only a limited number finds implementation in the standard fMRI analysis packages (e.g. AFNI, SMP, FSL). Furthermore each group just tends to implement its own procedures. (One of the few remarkable exceptions to this is SMP2 which abandoned the Friston and Worsley alike "pre-coloring" technique used up to SPM99, see [15]••, [16]••, [17]•• in favor of a "pre-whitening" approach similar to that proposed in Bullmore *et al.* [19]••.) As a consequence a lot of the literature ends up being a "theoretical

discussion" dissociated from actual implementation and data-analysis amelioration.

References

[1] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. **Funtional brain mapping by blood oxygenation level dependent contrast magnetic resonance imaging: A comparison of signal characters with a biophisical model**. *Proc. Natl. Acad. Sci.*, 1990, **87**:9868 – 9872.

[2] K. K. Kwong, J. W. Beliveau, D. A. Chesler, I. E. Goldberg, R. M. Weisskoff, B. P. Poncelet, D. N. Kennedy, B. E. Hoppel, M. S. Cohen, R. Turner, and et al. **Dynamic Magnetic Resonance Imaging of Human Brain Activity During Primary Sensory Stimulation**. *Proc. Natl. Acad. Sci.*, 1992, **89**:5675 – 5679.

[3] S. Ogawa, D. W. Tank, R. S. Menon, J. M. Ellerman, S. G. Kim, H. Merkle, and K. Ugurbil. **Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic-resonance-imaging**. *Proc. Natl. Acad. Sci.*, 1992, **89**:5951 – 5955.

[4] C. S. Roy and C. S. Sherington. **The regulation of the blood supply of the brain**. *J. Physiol.*, 1890, **11**:85 – 108.

[5] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann. **Neurophysiological investigation of the basis of the fMRI signal**. *Nature*, 2001, **412**:150 – 157.

[6] N. K. Logothetis. **The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal**. *Philos. Trans. R. Soc. London*, 2002, **357**:1003 – 1037.

[7] K. Thomsen, N. Offenhauser, and M. Lauritzen. **Principal neuron spiking: neither necessary nor sufficient for cerebral blood flow in rat cerebellum**. *J. Physiol.*, 2004, **560**:181 – 189.

[8] D. J. Heeger, A. C. Huk, W. S. Geisler, and D. G. Albrecht. **Spikes versus BOLD: what does neuroimaging tell us about neuronal activity?** *Nat. Neurosci.*, 2000, **3**:631 – 633.

[9] D. J. Heeger and D. Ross. **What does fMRI tell us about neuronal activity?** *Nat. Rev. Neurosci.*, 2002, **3**:142 – 151.

[10] N. K. Logothetis. **The Underpinnings of the BOLD Functional Magnetic Resonance Imaging Signal**. *J. Neurosci.*, 2003, **23**:3963 – 3971.

[11] P. Jezzard, P. M. Matthews, and S. M. Smith, editors. **Functional MRI: An introduction to methods.** Oxford University Press, Oxford, UK, 2002.

[12] R. M Weisskoff, J. Baker, J. Belliveau, T. L. Davis, K. K. Kwong, M. S. Cohen, and B. R. Rosen. **Power spectrum analysis of functionally-weighted MR data: Whats in the noise?** *Proc. Soc. Magn. Reson. Med.*, 1993, **1**:[Abstract].

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

[14] ••. K. J. Friston, P. Jezzard, and R. Turner. **Analysis of functional MRI time-series**. *Hum. Brain Mapp.*, 1994, **1**:153 – 171.
This is a seminal article for use of thes GLM approach for fMRI data; the authors here propose the use of a GLM approach to fMRI data analysis.

[15] ••. K. J. Friston, A. Holmes, J.-B. Poline, P. Grasby, S. Williams, R. Frackowiak, and R. Turner. **Analysis of time-series revisited**. *NeuroImage*, 1995, **2**:45 – 53.
Together with Worsley and Friston (1995) this article establishes the use of GLM for

fMRI data analysis and first loosk into the autoregression problem of fMRI time-series.

[16] ••. K. J. Worsley and K. J. Friston. **Analysis of time-series revisited – Again**. *NeuroImage*, 1995, **2**:173 – 181.

Together with Friston *et al.* (1995) this article establishes the use of GLM for fMRI data analysis and offers a clear specification of the "pre-coloring" technique to reduce the bias in estimators' computation due to autocorrelation in the time-series.

[17] ••. K. J. Friston, O. Josephs, E. Zarahn, A. P. Holmes, S. Roquette, and J. B. Poline. **To Smooth or Not to Smooth? – Bias and Efficiency in fMRI Time-Series Analysis**. *NeuroImage*, 2000, **12**:196 – 203.

In this paper Friston and colleagues present their "pre-coloring" technique for autocorrelation modeling in response to Bullmore's (1996) "pre-whitening" tecnique. They also provide data regarding the potential problems of the Bullmore et al (1996) approach.

[18] J. D. Carew, G. Wahba, X. Xie, E. V. Nordheim, and M. E. Meyerand. **Optimal Spline Smoothing of fMRI Time Series by Generalized Cross-Validation**. *NeuroImage*, 2003, **18**:950 – 961.

[19] ••. E. Bullmore, M. Brammer, S. C. R. Williams, S. Rabe-Hesketh, N. Janot, A. David, J. Mellers, R. Howard, and P. Sham. **Statistical methods of estimation and inference for functional MR image analysis**. *MRM*, 1996, **35**:261 – 277.

The authors present their AR(1) "pre-whitening" approach to serial correlation.

[20] P. Purdon and R. Weisskoff. **Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-based false-positive rates**. *Hum. Brain Mapp.*, 1998, **6**:239 – 249.

[21] P. Purdon, V. Solo, R. M. Weisskoff, and E. N. Brown. **Locally regularized spatiotemporal modeling and model comparison for functional MRI**. *NeuroImage*, 2001, **14**:912 – 923.

[22] J. J. Locascio, J. Peggy, and S. Corkin. **A method of adjusting for temporal and spatial correlations in analysis of mean fMRI signal intensity changes**. *NeuroImage*, 1997, **3**:S76.

[23] M. W. Woolrich, B. D. Ripley, J. M. Brady, and S. M. Smith. **Temporal autocorrelation in univariate linear modelling of FMRI data**. *NeuroImage*, 2001, **14**:1370 – 1386.

[24] J. L. Marchini and B. D. Ripley. **A New Statistical Approach to Detecting Significant Activation in Functional MRI**. *NeuroImage*, 2000, **12**:366 – 380.

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

[26] E. Bullmore, C. Long, J. Suckling, J. Fadili, G. Calvert, F. Zelaya, T. A. Carpenter, and M. Brammer. **Colored noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains**. *Hum. Brain Mapp.*, 2001, **12**:61 – 78.

[27] ••. T. E. Lund, K. H. Madsen, K. Sidaros, W. L. Luo, and T. E. Nichols. **Non-White Noise in fMRI: Does modelling have an impact?** *Nat. Neurosci.*, *in press*.
In a change of perspective (with respect to "pre-whitening" *a la* Bullmore et al (1996) and "pre-coloring" *a la* Friston et al. (2000)), the authors develop a new method (Nuisance Variable Regression) that treats autocorrelation as a source of information rather than as a simple statistical nuisance. Their model explicitly models for known

sources of noise such as low-frequency hardware-related drift, residual movement effects and physiological noise (i.e. cardiac pulsation and respiration).

[28] •. W-L. Luo and T. E. Nichols. **Diagnosis & Exploration of Massively Univariate fMRI Models**. *NeuroImage*, 2002, **19**:4207 – 4221.

The authors develop a tool to run diagnostic analysis on fMRI time-series to explicitly assess whether the fundamental assumptions of the GLM model are upheld.

[29] K. M. Petersson, T. E. Nichols, J. B. Poline, and A. P. Holmes. **Statistical limitations in functional neuroimaging. *I*. Non inferential methods and statistical models**. *Philos. Trans. R. Soc. London*, 1999, **354**:1239 – 1260.

[30] M. S. Cohen. **Parametric Analysis of fMRI Data Using Linear Systems Methods**. *NeuroImage*, 1997, **6**:93 – 103.

[31] R. B. Buxton, E. C. Wong, and L. R. Frank. **Dynamics of Blood Flow and Oxygenation Changes During Brain Activation: The Balloon Model**. *MRM*, 1998, **39**:855 864.

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

[33] ••. M. D. Robson, J. L. Dorosz, and J. C. Gore. **Measurements of the Temporal fMRI Response of the Human Auditory Cortex to Trains of Tones**. *NeuroImage*, 1998, **7**:185 – 198.

In a very elegant and clear experiment the authors show that the assumption of linearity for trains of stimuli presented at a faster rate than 6 seconds leads to massive overestimation of the actual response amplitude for longer trains of stimuli.

[34] ••. K. J. Friston, O. Josephs, G. Rees, and R. Turner. **Nonlinear event-related responses in fMRI**. *MRM*, 1998, **39**:41 – 52.

Friston and colleagues here show how the use of Volterra smoothing kernels allows for empirical characterization of nonlinearities and analysis of the nonlinear system in the GLM framework.

[35] M. L. Calvisi, A. J. Szeri, D. T. J. Liley, and T. C. Ferree. **Theoretical study of BOLD response to sinusoidal input**. *IEEE*, 2004, pages 659 – 662.

[36] K. J. Friston, A. Mechelli, R. Turner, and C. J. Price. **Nonlinear Responses in fMRI: The Balloon Model, Volterra Kernels, and Other Hemodynamics**. 2000, **14**:466 477.

[37] T. D. Wager, A. Vazquez, L. Hernandez, and D. C. Noll. **Nonlinear Responses in fMRI: The Balloon Model, Volterra Kernels, and Other Hemodynamics**. *NeuroImage*, 2005, **25**:206 – 218.

[38] ••. A. P. Holmes and K. J. Friston. **Generalisability, Random Effects & Population Inference**. *NeuroImage*, 1998, **7**:S754.

The very first attempt at combining fixed and random effects approaches in a "summary statistic" framework.

[39] K. J. Worsley, C. H. Liao, J. Aston, V. Petre, G. H. Duncan, F. Morales, and A. C. Evans. **A general statistical analysis for fMRI data**. *NeuroImage*, 2002, **15**:1 – 15.

[40] J. Talairach and P. Tournoux, editors. **Co-plannar stereotaxis atlas of the human brain.** Thieme, New York, NY, 1988.

[41] A. Evans, L. Collins, C. Holmesand T. Paus, D. MacDonald, A. Zijdenbos, A. Toga, P. Fox, J. Lancaster, and J. Mazziota. **A 3D probabilistic atlas of normal human neuroanatomy.** In *Third Int. Conf. on Functional Mapping of the Human Brain*. 1997.

[42] M. Brett, I. S. Johnsrude, and A. M. Owen. **The problem of functional localization in the human brain**. *Nat. Rev. Neurosci.*, 2002, **3**:243 – 249.

[43] R. S. J. Frackowiak, K. J. Friston, C. Frith, R. Dolan, C. J. Price, S. Zeki, J. Ashburner, and W. D. Penny, editors. **Human Brain Funcion**. Academic Press, 1997.

[44] ••. K. J. Friston, A. P. Holmes, C. J. Price, C. Bruchel, and K. J. Worsley. **Multisubject fMRI studies and conjunction analyses**. *NeuroImage*, 1999a, **10**:385 – 396.
This is the seminal paper regarding conjunction analysis. All the following work proceeds from this paper.

[45] ••. W. D. Penny and A. P. Holmes. **Random-Effects analysis.** In *Human Brain Function*, pages 843 – 850. Elsevier, San Diego., 2004.
In this paper the authors present an as clear as rigorous a mathematic characterization of the Fixed Effects and Random Effects approaches and propose a synthesis in a Hierarchical model that became the basis for much of the following work.

[46] C. F. Beckmann, M. Jenkinson, and S. M. Smith. **General multilevel linear modeling for group analysis in FMRI**. *NeuroImage*, 2003, **20**:1052 – 1063.

[47] ••. C. J. Price and K. J. Friston. **Cognitive conjunction: a new approach to brain activation experiments**. *NeuroImage*, 1997, **5**:261 – 270.
Price and Friston here propose a cognitive conjunction approach to data analysis that may make use of the increased sensitivity of fixed effects analysis while retaining the generalizability of random effects.

[48] T. E. Nichols, M. Brett, J. Andersson, T. Wager, and J. B. Poline. **Valid conjunction inferences with the minimum statistic**. *NeuroImage*, 2005, **25**:653 – 660.

[49] K. J. Friston, A. P. Holmes, and K. J. Worsley. **How many subjects constitue a study?** *NeuroImage*, 1999b, **10**:1 – 5.

[50] K. J. Worsley and K. J. Friston. **A test for conjunction**. *Statistics & Probability Letters*, 2000, **47**:135 – 140.

[51] D. Caplan and L. Moo. **Cognitive conjunction and cognitive functions**. *NeuroImage*, 2004, **21**:751 – 756.

[52] V. Bosch. **Statistical analysis of multi-subject fMRI data: Assessment of focal activations**. *J Magn. Reson. Imaging*, 2000, **11**:61 – 64.

[53] D. G. Leibovici and S. Smith. **Comparing groups of subjects in fMRI studies: a review of the GLM approach.** FMRIB Technical Report, Oxford center for Functional Magnetic Resonance Imaging of the Brain (FMRIB), 2000.