Meaningful Design and Contrast Estimability in FMRI

FMRIB technical Report TR06SS1

(A related paper has been accepted for publication in NeuroImage)

Stephen Smith, Mark Jenkinson, Christian Beckmann, Karla Miller and Mark Woolrich

Oxford University Centre for Functional MRI of the Brain (FMRIB)

Abstract

Optimising the efficiency of an experimental design is known to be of great importance. However, existing methods for calculating design rank deficiency and contrast estimability (an important aspect of experimental design) relate to computational precision rather than image noise and are therefore not very meaningful. For example, a contrast between two experimental conditions may be mathematically "estimable" while requiring a huge differential BOLD response for statistical signifiance to be reached. In this paper we formulate standard efficiency equations in terms of required BOLD effect, and use this to generate measures of rank/estimability which are meaningful. This takes into account the strength and smoothness of the timeseries noise and is applicable to complex contrasts; we show how to re-express several regressors and an associated contrast vector as a single equivalent regressor, so that we can calculate the contrast's effective peak-peak height unambiguously. We also present some example results on typical designs, and characterise noise results from a range of typical FMRI acquisitions, in order to allow experimenters to apply efficiency estimation in advance of acquiring data.

Keywords

FMRI, experimental design, design efficiency, rank deficiency, estimability.

1 Introduction

FMRI analysis tools that are based on the general linear model (for example, SPM and FEAT in FSL) generally test for experimental design "rank deficiency" and "contrast estimability". For example, two different experimental conditions need to be applied with different timings if the experimenter is going to be able to distinguish between the responses that the brain has to the different conditions. Such estimability calculations should be carried out in advance of data acquisition, in order to check that the regressors (the predicted responses to the different experimental conditions) are sufficiently independent of each other, and that interesting contrasts of the conditions are mathematically estimable. However, standard calculations of estimability are purely related to computational numerical precision, and do not take into account the effect of noise in the data or reflect the efficiency of the paradigm in a meaningful way.

Some approaches require the design matrix to be invertible, which means for example that the ratio of the maximum to minimum singular values is smaller than some very large number, e.g., 10^{16}. However, strictly, all that should be necessary is that the contrasts of parameter estimates are estimable; this is a slightly looser constraint, namely that the pseudo-inverse of the design matrix is calculable (it does not have to be of full rank) and that each contrast has a calculable variance. However, such calculations do not take into account noise in the data, the presence of which can make estimation less well-conditioned than it might appear.

For example, consider a poorly designed experiment, where two different conditions are highly correlated (quite similar temporally) but still sufficiently different that the design matrix is mathematically of full rank, and a contrast between conditions theoretically estimable. However, the difference between the two regressors may be small enough that it is the noise in the data, not any underlying real signal, that mostly drives the model fitting, and the contrast will not be estimable in practice. The extent to which this is the case depends on the contrast, the regressors (including the amount of correlation between them), and the size and structure of the FMRI noise. Only if all of these things are taken into account can contrast estimability be meaningfully calculated, allowing more intelligent paradigm/contrast design.

In this paper we present the simple steps necessary to make a meaningful estimability calculation. Without such measures, an experimenter might find and report no significant effect, without realising that this is primarily because of poor experimental design. For a given experimental design, contrast vector, data noise level (and temporal smoothness) and significance threshold, we can calculate the BOLD effect size (or difference in BOLD effect in the case of differential contrasts) that is necessary in order for the activation to be robustly detected. We first present this in the simple case of a single modelled experimental effect, under the assumption of white (non-smooth) noise. We next derive the correction needed when the noise is no longer assumed to be white, and then generalise this to any number of model regressors and any contrast of parameter estimates. This includes showing how to re-express several regressors and an associated contrast vector as a single equivalent regressor, so that we can unambiguously calculate the contrast's effective peak-peak height. We discuss the interaction of these estimability calculations with the choice of highpass temporal filtering carried out. We also present example data showing typical noise characteristics, and present example estimability results on some typical designs.

2 Contrast estimability

2.1 Design Efficiency

One of the early papers to discuss the relative efficiency of designs is [5], where efficiency is quantified as the inverse of the variance of the estimated effect. An example result shown is that randomised event-related designs can be much more efficient than fixed inter-stimulus-interval designs. Such considerations were developed in considerable depth in [11, 10, 9], in particular the tradeoff between sensitivity to activation, ability to estimate shape of the haemodynamic response function (HRF), and randomisation of presentation order of multiple stimulation types. In [10, 9] there is also some discussion of the impact of temporal autocorrelation (smoothness) and of using basis functions for the HRF modelling. In [4] there is a detailed investigation of the effect of different approaches to modelling autocorrelation on estimation efficiency and false positive rate.

Related issues are discussed further, particularly in the context of power analysis, in [22]. It is shown how to estimate and predict both false positive and false negative error rates through the use of a "reference effect", for example, characterising

some aspects of the signal and noise from pilot data and then allowing the experimenter to predict the effect of varying the "neural activity" or experimental paradigm. The importance of considering false negatives, and not just false positives, is stressed, and investigated in detail.

Two software tools have been created to allow the practical optimisation of design efficiency, namely Optseq surfer.nmr.mgh.harvard.edu/optseq [5] and an approach using genetic algorithms www.columbia.edu/cu/psychology/tor [17].

Quantifying design efficiency as the inverse of the variance is closely related to quantifying it through the statistical significance of the estimated effect - for example, the t-statistic. In the next section we combine the equations for t-statistics and BOLD effect size, and derive a single simple equation for the effect required to both reach statistical significance (controlling false positive rate) and also having a specified likelihood of detectability (i.e., power - controlling false negative rate). We use this as a practical measure of both design efficiency and contrast estimability. While several of the above references have presented similar mathematical components (of efficiency calculation) as we have used, we have attempted to provide an explicit yet simple single formulation of the efficiency calculation in terms of different factors such as noise strength, noise smoothness (and its interaction with filtering) and highpass filtering, and we relate efficiency to design rank deficiency and estimability.

2.2 Simple model

The standard general linear model (GLM), in the case of a single regressor, is:

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

where Y is the (timeseries) data vector, X is the design matrix (a single regressor vector in this case), β is the regression coefficient (a single scalar, often also referred to as "effect size" or "parameter estimate") and e is the residuals (error vector, assumed to be Gaussian distributed with standard deviation σ).

The % BOLD effect, B, is given by:

$$B = \frac{100 * \text{BOLD effect}}{I_0} = \frac{100\beta h}{I_0}, \tag{2}$$

where I_0 is the baseline data intensity and h is the peak-peak height of the regressor X. We want to define practical estimability in terms of determining how large B needs to be in order to find a statistically significant effect.

Statistical significance of the BOLD effect is defined through the use of the t-statistic, which divides the effect size by its standard deviation[1]:

$$t = \frac{\widehat{\beta}}{\sqrt{\widehat{var}(\beta)}} = \frac{\widehat{\beta}}{\widehat{\sigma}/\sqrt{X'X}}. \tag{3}$$

Through any given inference method (e.g., Gaussian random field theory), one can convert a desired null-hypothesis test $p < \alpha$ into $t > t_\alpha$, where α is the critical p value (e.g., 0.05) and t_α will in general also depend on factors such as the number of timepoints in the data, the temporal smoothness of the data, and, in order to correct for multiple comparisons across voxels, the number of voxels and the spatial smoothness of the data. This controls the false positive rate ("type I error") at the required level. If one uses t_α in the efficiency calculations then *on average* statistical significance will be reached - but in practice, because of a spread of effect sizes about the mean, this will result in 50% of the measurements falling below detectability. In practice, one wants to raise the true positive rate (detectability) to say 80%, and in our calculations this can be enforced by using a critical t value that is larger than t_α (this is controlling the "type II error"). In Appendix A we discuss how to ensure that an experiment has a controlled true positive rate; for example, for typical FMRI experiments (i.e., having more than 100 timepoints, setting $\alpha = 0.05$ and correcting for multiple comparisons), one simply needs to set the critical t value t_c to $t_\alpha + 0.9$ to achieve an experimental power (true positive rate) of 80%.

We now define $D = h/\sqrt{X'X}$, a function only of the design matrix. D is proportional to the ratio of the peak-peak height of the regressor to its (temporal) standard deviation. The latter is the term of interest in design efficiency estimation discussed

[1]The $\widehat{\ }$ symbol differentiates estimated quantities from their true underlying values, though in general we do not need to make this distinction in this paper.

in [5]; all other things being equal, the higher the variance of the regressor, the more efficient the design. Thus D reflects (the inverse of) the "intrinsic" design efficiency - "intrinsic" because this measure (in contradistinction to B below) does not take into account the level of noise present in the data or the statistical significance required to find an effect.

We also define $N = 100\sigma/I_0$, i.e., the noise standard deviation expressed as a percentage fraction of the baseline intensity. N can either be quickly estimated from data, or even roughly estimated as a fixed value for a given field strength (e.g., 0.7% at 3T).

Combining all of the above, we therefore have

$$t = \frac{BI_0\sqrt{X'X}}{100\sigma h} = \frac{B}{DN} \tag{4}$$

i.e., in order to find activation with statistical significance ($t > t_c$), we require that

$$B > t_c DN. \tag{5}$$

Therefore the required BOLD effect is the product of the statistical significance required, the design efficiency and the strength of the noise in the data.

2.3 Extension for temporal smoothness & prewhitening

In FMRI statistical analysis there are problems with accounting for temporal autocorrelation (the intrinsic smoothness in each voxel's timeseries). Unless this is correctly accounted for, the timeseries analysis is at best inefficient (in terms of sensitivity to true activation) and at worst statistically invalid. Commonly, techniques have utilised temporal filtering strategies to either shape these autocorrelations, or remove them. Shaping, or "colouring", attempts to negate the effects of not accurately knowing the intrinsic autocorrelations by imposing known autocorrelation via further smoothing. Removing the autocorrelation, or "prewhitening", gives the best linear unbiased estimator, if the autocorrelation can be accurately estimated. In [20] we presented an approach for accurately and robustly estimating voxelwise autocorrelation using spectral and nonlinear spatial regularisation, and then removing the autocorrelation via a whitening step within the GLM fitting.

It is straightforward to correct the above equations to account for both temporal smoothness in the data and the associated correction for this smoothness in the model fitting ("prewhitening"). The relevant amendment to the variance of the parameter estimate is:

$$\widehat{var}(\beta) = \tilde{X}^{+} A V A' \tilde{X}^{+'} \widehat{\sigma}^2, \tag{6}$$

where V is the timeseries covariance (i.e., it specifies the autocorrelation), A is the prewhitening matrix (derived from the estimate of V), and $\tilde{X} = AX$ is the whitened design (we are still considering a single regressor design at this point). The $^+$ operator is the Moore-Penrose pseudoinverse. If we assume that the prewhitening is effective, then by definition $AVA' = I$ and we simply have to replace the above definition of D with $h/\sqrt{\tilde{X}'\tilde{X}}$.

Therefore, in order to be able to estimate B in advance of data acquisition, we ideally need to not only make a rough estimation of the size of the noise (N), but also the form of the autocorrelation present (which is estimated in order to generate the prewhitening matrix A, and applied to the design matrix to create $\tilde{X}$). One can approximately characterise this in advance of new data acquisition, on the basis of null datasets. In general N and A will depend primarily on TR (temporal sampling rate) and field strength, as well as other imaging parameters that affect data SNR.

Note that the above equations will suggest that there is an increase in estimability/efficiency when the amount of temporal autocorrelation is increased, which seems counter-intuitive. However, as pointed out in [10], this assumes that variance is being held constant, and hence is misleading when taken out of context - in fact, if one takes white data and applies temporal smoothing (thus increasing autocorrelation), variance is decreased, and estimation efficiency may well in fact decrease.

2.4 Multiple regressors and general contrasts

We now generalise the model X to being a matrix of n regressors and β to being a vector of n fitted parameter estimates. A question is asked of the experiment through the use of a contrast (vector c) of elements of β, i.e., $c'\beta$. For example, if there

are two experimental conditions and therefore two regressors in the design matrix, one would ask the question "where is the response to condition 1 greater than the response to condition 2?" via contrast c=[1 -1]′.

We also need to generalise the definition of B. In the case of simple contrasts containing just a single 1 and hence acting on a single regressor, B is the % BOLD effect associated with that regressor. In the case of a differential (or more complex) contrast, B is the differential BOLD effect, i.e., the difference in BOLD response between different experimental conditions.

It is not always obvious what "regressor height" h means in the case of differential (or more complex) contrasts. However, if we reformulate the design matrix and contrast such that the contrast being asked becomes [1 0...]′ (i.e., it is now achieved simply through the first regressor of the new design matrix), then the definition of h becomes clear. In Appendix B it is shown how to achieve this - i.e., how to reformulate a design matrix and contrast into a new design matrix whose first regressor X_{eff} has an associated parameter estimate which is equivalent to the original $c'\beta$. It is also shown how to ensure that the remainder of the new design matrix is orthogonal to X_{eff}, which means that we can for our purposes forget the rest of the design matrix. The equation derived for X_{eff} is:

$$X_{\mathit{eff}} = XQc(c'Qc)^{-1}, \text{ where } Q = (X'V^{-1}X)^{-1}. \tag{7}$$

We can therefore replace X in the definition of D with X_{eff}, and can use the peak-peak height of X_{eff} to calculate h with no ambiguity.

2.5 Interaction of efficiency with highpass temporal filtering

There is a potentially strong interaction between highpass temporal filtering (whether applied in data preprocessing or via basis functions in the design matrix) and model fitting, including prewhitening if carried out. We now discuss how this interaction relates to the issues of interest in this paper.

Highpass filtering is generally perceived as being used in order to remove noise of lower frequency than the signal of interest, in order to improve estimation efficiency. Such a view might lead one to suspect that setting highpass filtering to be as aggressive as possible, without reducing the signal of interest significantly, would result in optimal activation estimation. In fact, when seen from the viewpoint of residual whitening (i.e., when taking the Gauss-Markov approach that the best unbiased linear estimation of activation is to whiten the noise perfectly) this is not really a useful way to view highpass filtering. If one is going to whiten the data before final model fitting, the point of highpass temporal filtering should not be viewed as being to remove as much low frequency noise as possible, but should be simply removing those components of the noise which will not be well-modelled by the autocorrelation modelling to be used. For example, the very slowest trends will not be well-estimated by AR(N) modelling due to various practical factors such as limited effective sampling given the number of timepoints available. If one removes more low-frequency noise than is necessary to allow accurate autocorrelation modelling, the modelling itself will likely suffer, in the sense that the timeseries whitening will become less accurate, and optimal estimation efficiency will not be obtained.

In [20] the regularisation of autocorrelation parameters is achieved through a Tukey taper which downweights the longest lag estimation, the idea being that the slowest trends, corresponding to the longest lags, will not be well-estimated by the autocorrelation model, and should be removed in preprocessing by an appropriately set highpass filter. Unfortunately the interaction of the filtering with the whitening and subsequent GLM model-fitting is complex, and it is not straightforward to predict in advance what an optimal filter would be. As more research into temporal noise modelling takes place, it will be important to investigate what highpass filtering is appropriate for any given autocorrelation method.

In conclusion, it is important not to oversimplify the role of highpass temporal filtering, and one should keep in mind that while the equations presented in this paper may appear to show estimation efficiency changes as the temporal filtering cutoff is reduced, if the cutoff is reduced too far, timeseries whitening accuracy may suffer, and the equations such as $AVA' = I$ may well no longer hold, invalidating the efficiency calculation.

3 Examples

In this section we first present some example results characterising FMRI noise when using a few typical sets of acquisition/preprocessing parameters. This is primarily serving as an example of how one can characterise noise in advance of a large study, in order to allow one to apply the estimability calculations shown above. We then present some examples of contrast estimability calculations for a few different designs, including results showing the effect of ignoring temporal smoothness.

3.1 Noise variance and smoothness in real data as a function of field strength, TR and spatial smoothing

In order to investigate variations in noise variance and temporal smoothness we acquired 6 resting FMRI datasets, using two field strengths (1.5 and 3T Siemens scanners) and 3 different TRs (1, 3 and 5s). The flip angle was 90°, the TE was 50ms at 1.5T and 30ms at 3T, the voxel size was 3x3x3mm. The same subject (healthy female, 31 years) was used for all 6 datasets. As TR increased the number of slices obtained also increased; only the part of the field-of-view that was common across all TRs was used for quantitative analysis.

We motion-corrected each FMRI dataset using MCFLIRT [7], then smoothed with a range of spatial extents (0, 5 and 10mm FWHM), before removing very low-frequency temporal drift with a Gaussian-weighted (FWHM=100s) highpass filter. As well as investigating the noise characteristics following such standard preprocessing, we also tested the effect of removing the 10 strongest structured noise components in each dataset (using MELODIC independent component analysis [3]). This ICA-based cleanup removes spatiotemporally structured noise such as resting-state networks [1] and structured scanner artefacts, which are known to be poorly modelled by short-range/univariate filters and models such as the temporal drift removal and timeseries autocorrelation modelling. If one believes that approaches such as ICA can be effective at removing strong structured noise, the summary statistics (of noise strength and temporal smoothness) estimated in this way are arguably more informative than estimation which does not remove structured noise first.

We also acquired a 1x1x1mm T1-weighted structural image, in order to be able to create a grey-matter mask, so that we could concentrate on quantitating noise characteristics in grey matter FMRI voxels. We brain-extracted the structural image using BET [14] and then segmented the resulting brain image into different tissue types using FAST [23]. We registered the 6 FMRI datasets to the brain-extracted structural using FLIRT [8] and applied the structural-derived grey-matter mask to the FMRI data to remove non-grey-matter voxels.

Finally, we estimated noise characteristics voxelwise in the resulting datasets. We calculated variance and fit an AR(1) model to estimate temporal smoothness. (We also fit an AR(20) model, in order to confirm that the general results were not affected by the use of the simple AR(1) model, and indeed the overall pattern of results was unchanged, so we have not presented these results here.) The results for the 6 datasets are shown in figure 1. Each boxplot is over the set of grey-matter voxels. For each dataset there are 6 boxplots shown; for each of the 3 spatial smoothing extents we estimated the noise with and without ICA-based structured noise cleanup.

In order to test the dependence on the 4 factors varied (field strength, TR, spatial smoothing and structured noise removal), we ran a 4-factor ANOVA on both the noise level and the smoothness data. In all tests there was a significant factor effect ($P<0.001$).

There are several interesting aspects to these results. Noise is lower at higher-field, as expected, and the data is smoother, because the ratio of (relatively smooth) physiological noise to (non-smooth) thermal noise is greater. Noise is reduced as TR increases from 1 to 3s (because the signal level itself increases as the TR starts to exceed the T1 time) but does not reduce much from 3 to 5s. The data is less temporally smooth as TR is increased (as there are longer gaps between successive samples, hence any temporal correlation intrinsic to the signal will be less apparent in the sampled data). Spatial smoothing reduces noise and increases temporal smoothness (presumably because it reduces the thermal noise more than the physiological noise, as the former is spatially less correlated than the latter). Structured noise removal reduces apparent noise and temporal smoothness significantly.

The interquartile ranges (the spread of values across voxels) of noise level and smoothness are relatively wide compared

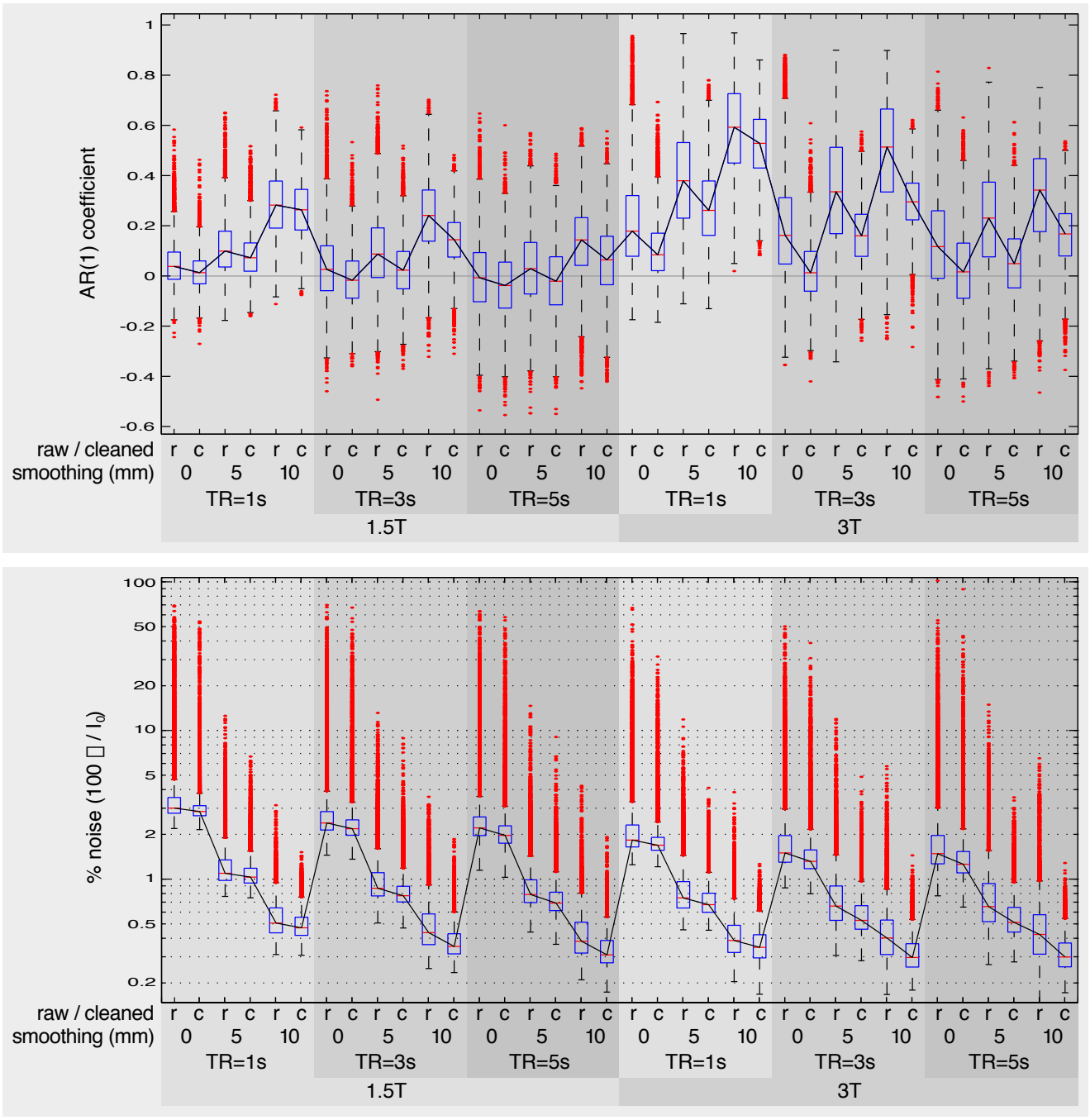

Figure 1: Noise characterisation at two MRI field strengths (1.5 and 3T), three different TRs (1, 3 and 5s), three different spatial smoothing extents (0, 5 and 10mm FWHM) and without and with ICA-based structured noise removal. For each set of acquisition+preprocessing parameters, a boxplot shows noise strength (below) and temporal smoothness (above), using FMRI voxels selected via a structural-derived grey-matter mask. Noise is plotted as fractional percentage of baseline signal (note the log scale; the median results are more variable across different acquisition+preprocessing parameters than is initially apparent).

with the median values. For example, consider one of the most "typical" sets of acquisition+preprocessing parameters: 3T, TR=3s, 5mm smoothing and no cleanup; in this case the noise level is median 0.66% (interquartile range 0.53-0.90%) and the AR(1) coefficient is median 0.34 (IQR 0.17-0.51).[2] Voxels with negative estimated autocorrelation were not obviously spatially clustered. Note that we have restricted our analysis primarily to grey matter voxels, and yet there is still considerable variability of noise level and temporal autocorrelation across the brain; this emphasises the need for autocorrelation modelling to be computed locally [20] and not summarrized via a single global estimate.

For a detailed investigation of dependence of FMRI noise as a function of field strength, voxel volume and several other factors, as well as the varying balance between thermal and physiological noise, see [16]. Note that our raw, unsmoothed estimates of noise strength at 1.5 and 3T with TR=5s are in exact agreement with the relevant measurements shown in figure 2 of [16].

3.2 Example experimental designs

3.2.1 Three multi-regressor examples

Figure 2 shows three example experimental designs. We set number of timepoints to 200, TR to 3s, the highpass filter cutoff to 100s and the critical t threshold t_c to 5.5 (corresponding, for example, to a false positive rate of $p<0.05$ corrected via GRF for 500 resels, with a detectability power of 80%). We set N=0.66% and the AR(1) coefficient to 0.34. For each design, the figure shows the original design matrix and contrasts, as well as the whitened, effective EV (explanatory variable, or regressor) relating to each contrast (i.e., $\tilde{X_{eff}}$, where X_{eff} is defined in equation 7).

With design 1, the main condition of interest has a timecourse that is clearly different than the other regressors, and the contrast is asking about this condition only. For this contrast B needs to be greater than 0.85% to reach statistical significance. If we ignored temporal smoothness in the efficiency estimation, we would get $B>0.63$% - a significantly different answer.

With design 2, the main regressor only contains 3 events; it is inefficient due to having low variance. Here $B>2.0$%, i.e., requires more than twice the signal than design 1.

Design 3 has two very similar regressors, i.e., is nearly rank deficient. The first contrast is the mean of the two regressors, so has very similar power to the first experiment: $B>0.84$%. However, the differential contrast is inefficient due to the extreme similarity of the regressors, and $B>3.36$%, i.e., requiring a very large differential response to the two conditions.

3.2.2 Single boxcar regressor - dependence on boxcar period

Figures 3 and 4 show two further examples; here efficiency for contrasts in 2 simple designs is plotted as a function of factors such as paradigm block length. These figures illustrate how non-obvious design issues can be investigated using the efficiency calculations. We set the same TR, t threshold and noise characteristics as above.

Figure 3 shows the results when the period of a simple boxcar design is varied from 4 to 200 seconds, at a range of experimental lengths (from 50 to 400 timepoints). The highpass cutoff was set at 50s. Unsurprisingly, the more timepoints the greater the efficiency. At the shortest and longest boxcar periods, the efficiency is worse than with intermediate values. For very short periods, the haemodynamic response is too slow to fully track the rapid changes in experimental condition. At the longest periods, either signal variance is lost because of highpass temporal filtering applied (that is aggressive with respect to the period), or if less aggressive filtering is used, residual long-range noise will not be modelled well by temporal whitening (see Section 2.5).

[2] Autocorrelation coefficient estimation is notoriously high-error [21]. In order to test whether the spread in noise level and smoothness seen in our data reflected a true underlying spread or simply expected estimation error, we took the median noise level and smoothness values from the 3T, TR=3s, 0mm smoothing, non-cleaned dataset and simulated the same number of timeseries with simulated AR(1) data having those median values set as input parameters. We then estimated the noise level and smoothness parameters from the simulated data, resulting in an interquartile range that was smaller than that seen in the real data by a factor of 8 for the noise level and 2 for the smoothness parameter. We can conclude that the spread of values shown in figure 1 does indeed largely reflect the true underlying reality.

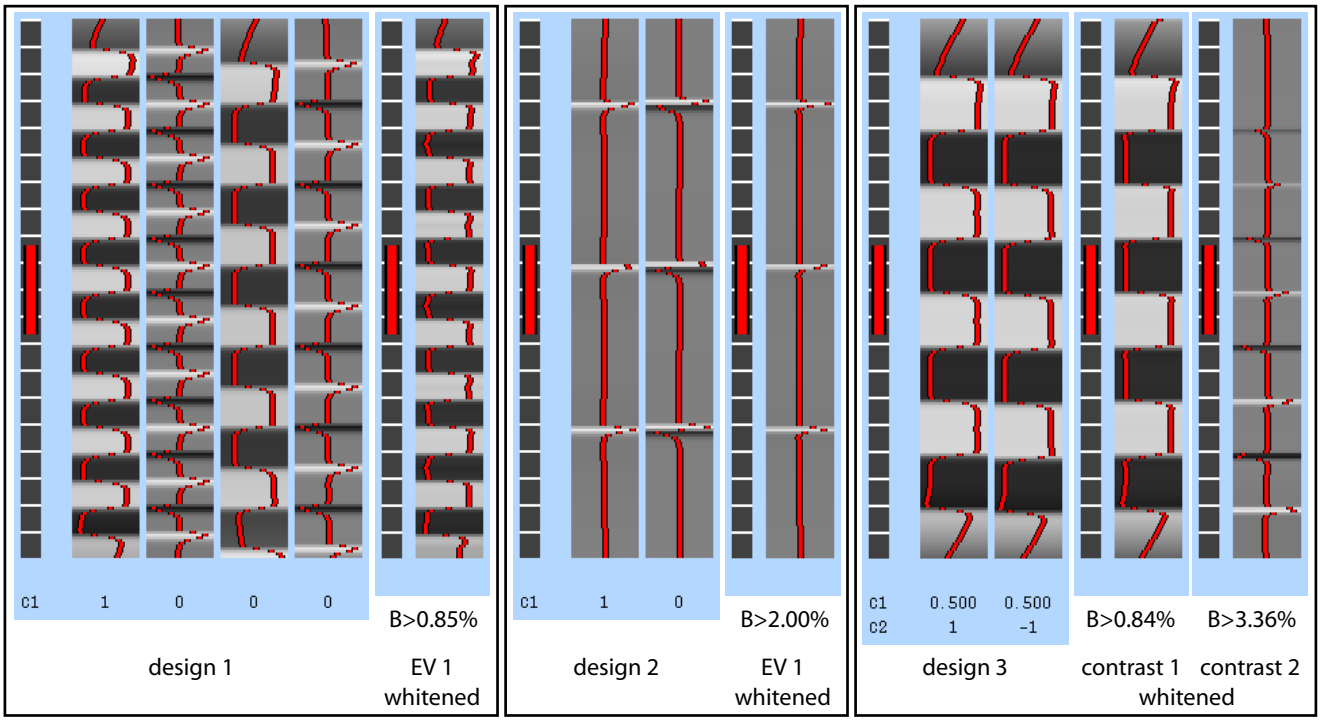

Figure 2: Three example experimental designs. The bar on the left is a representation of time, which starts at the top and points downwards; the white marks show the position of every 10th volume in time. The red bar shows the period of the longest temporal cycle which was passed by the highpass filtering. The main columns show the different regressors in the design matrix; again, time is represented vertically. The regressor's timecourse is both encoded by the lateral position of the red plots and by the underlying greyscale intensity. In designs 1 and 2 the main regressors are followed by their temporal derivatives, which are used to account for mis-specification of the haemodynamic lag. Below the regressors are shown the requested contrasts (used for testing primary or differential BOLD changes); each row is a different contrast vector and each column refers to the weighting of the relevant regressor in that contrast. The "whitened" plots show the whitened effective EV ($\tilde{X_{eff}}$) derived from each contrast.

3.2.3 Two alterating boxcar regressors - means and differential contrast

Figure 4 shows a design alternating between two conditions, with the period of rest between each activation condition varying from 0 to 60 seconds. The ON periods are fixed at 30 seconds and the highpass cutoff at 100s. Contrast 1 asks where activation condition A is greater than baseline, contrast 2 asks where A is greater than condition B, and contrast 3 asks where the mean of the two conditions is greater than baseline. Contrasts 1 and 3 are similar except at the longer rest periods, when the mean effect becomes more efficiently estimated. At the shortest rest periods, neither can be efficiently estimated, as the rest periods are too short to allow significant variation in the signal between rest and either activation condition. Contrast 2 does not depend on the rest condition, and so is most efficiently estimated with short rest, i.e., with no time wasted in the unnecessary rest condition.

Of the various effective regressor (X_{eff}) plots, shown for the two extremes of the rest duration, the one particularly interesting effective regressor is for contrast 1 at rest=0s. Here the effective regressor contains almost no variance and hence has poor estimability. Surely a [1 0] contrast would be expected to leave the first regressor as it is, when generating the effective regressor? The explanation is that when there are other regressors (or combinations of regressors) which are very similar to the one in question, the parameter estimates are driven greatly by the noise (a small change in the noise shifts the model fit around greatly between the two parameter estimates). In this case, the fitting of either of the two regressors is controlled completely by the component of that regressor that is orthogonal to the other, and because of the high collinearity, the orthogonal component has relatively low variance. Hence the error in estimating the contrast is huge - the contrast cannot be

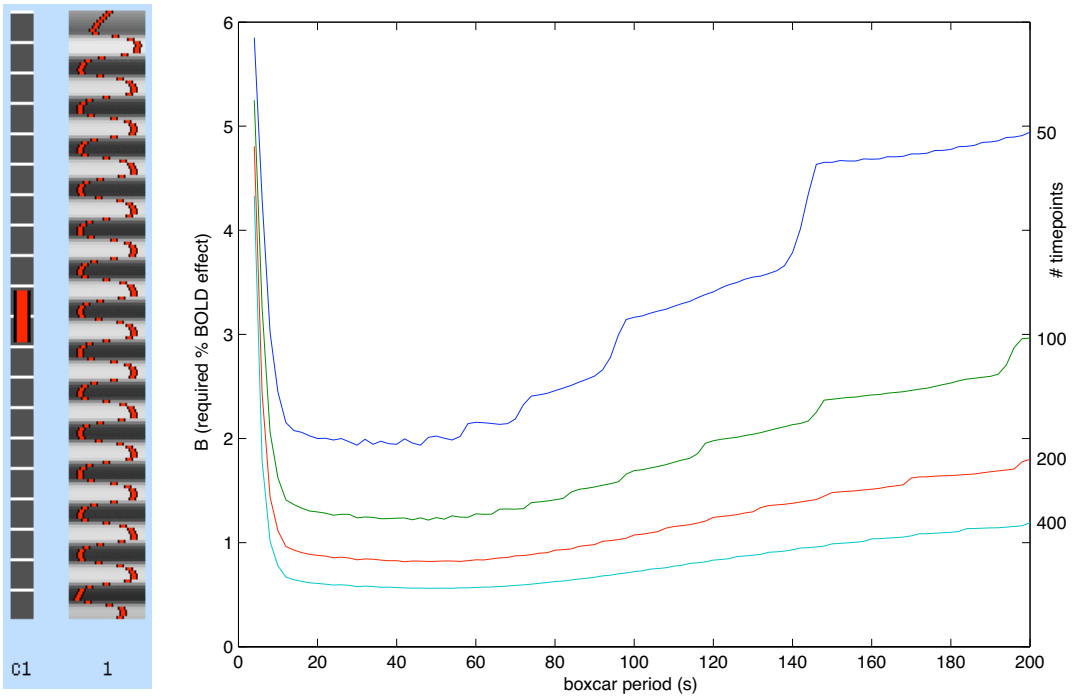

Figure 3: Required % BOLD effect for a single boxcar regressor, for a range of boxcar periods and experiment durations. The example design on the left has a boxcar period of 40s.

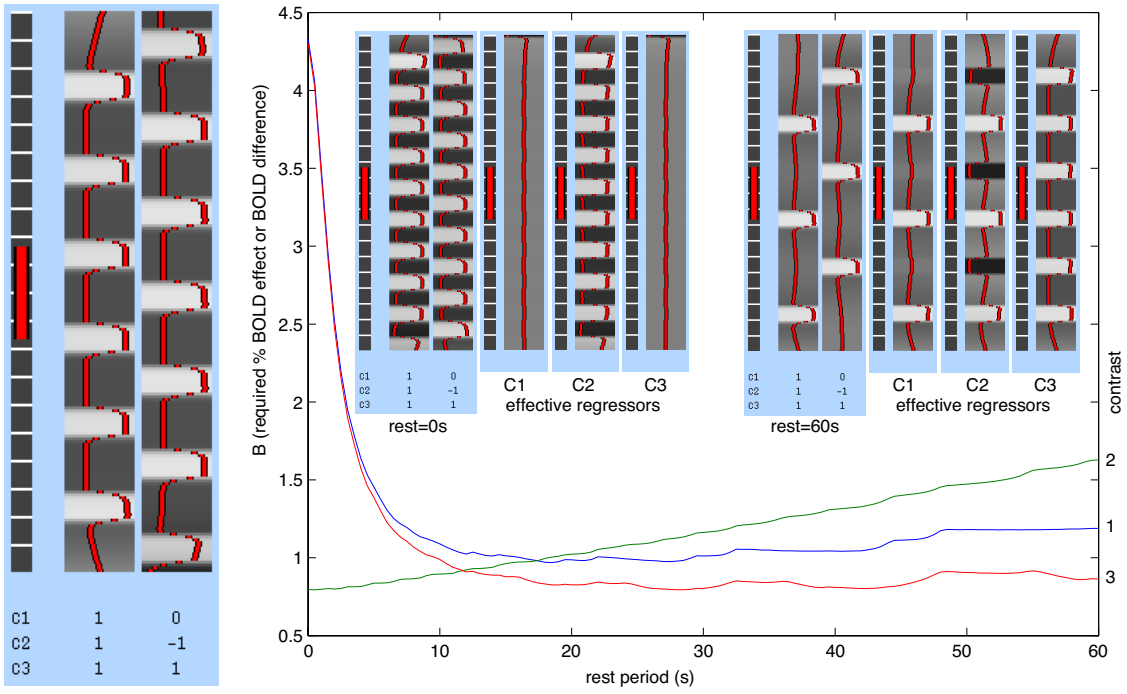

Figure 4: Required % BOLD effect for two alternating boxcar conditions, for three contrasts (condition 1, differential contrast and mean of both conditions), for a range of baseline ("rest") condition lengths, from 0 to 60s. The ON periods are fixed at 30s for both conditions. The example design on the left shows a rest duration of 15s. The insert design matrices show the full design at the two rest duration extremes (0 and 60s), as well as the effective regressor for each of the 3 contrasts.

efficiently estimated.

3.2.4 Effect of varying TR (and hence noise level and smoothness)

Figure 5 shows the effect on estimability of varying the TR. Changing TR also in general changes the noise level and smoothness, and for fixed experiment duration the number of timepoints is also changed (we do not show the effect of varying just one of these quantities in isolation, as in practice that would not be expected to occur). In order to therefore take account of all of these quantities we interpolated the data shown in Figure 1 (fitting a quadratic interpolant to the median values for 3T, 5mm smoothing and no structured noise removal) to give noise level and smoothness parameters as a function of TR, to use in the calculation of efficiency. Note that functional contrast is not expected to vary much in the TR range 1:5s. The length of the experiment was set to 550s and the highpass cutoff to 100s. Two paradigm designs were tested - a boxcar of 60s period and a dense randomised event-related design with random inter-stimulus intervals in the range 4-8s.

As expected, the boxcar design is more efficient than the event-related. In both cases the efficiency improves as TR decreases and smoothness and noise level increase, but not by huge amounts (for example, as TR is reduced from 3 to 1s, the efficiency increases by 15-30%). The improvement is mostly due to the increase in the number of timepoints as TR is reduced.

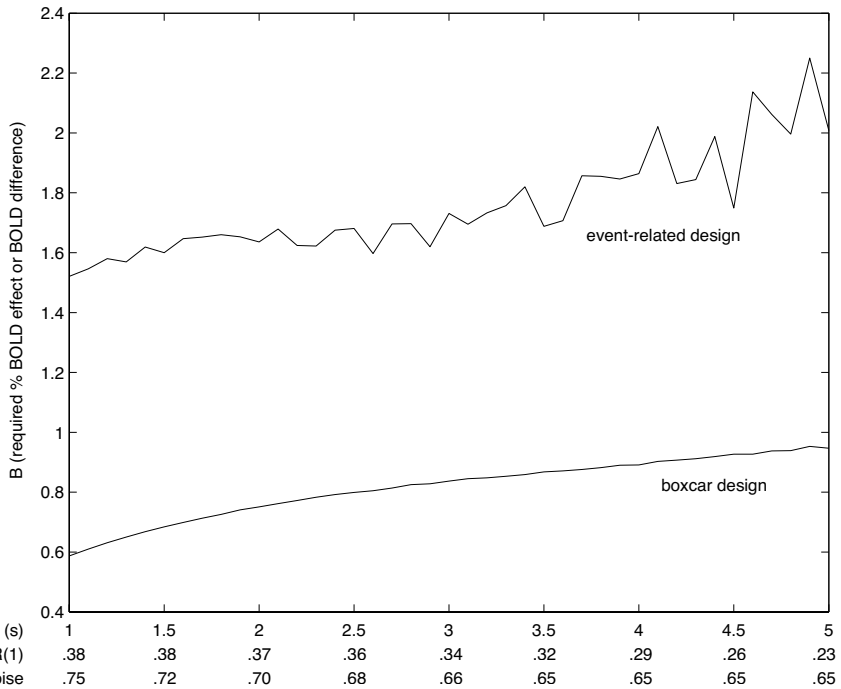

Figure 5: Required % BOLD effect as a function of TR, and therefore also as a function of typical changes induced in the AR(1) parameter and noise strength as a result of changing TR.

4 Discussion

We can easily estimate, in advance of data acquisition, the required BOLD effect size (or difference in BOLD response in different conditions), in a way that is sensitive in a meaningful way to the noise level and, crucially, the efficiency of the experimental design and contrasts. Although this final equation appears obvious if one forgets D (i.e., simply forms a t statistic as the ratio of signal over noise), D is the crucial factor, as it quantitatively informs the experimenter how design efficiency issues will affect the experimental sensitivity to activation.

When comparing different designs using the calculations described in this paper, one needs to keep in mind the fact that event-related experiments typically achieve less signal change than block designs, as the haemodynamic response to a short stimulation does not have long enough to rise to the full height reached under sufficiently long-duration activation. Therefore it is likely, all things being equal, that a given required B will be more easily attained in the block-design case. This difference is ameliorated to some extent by the nonlinearity of the BOLD response; for short stimulation/activity durations (less than approximately 1 second), the response is a fairly linear function of the duration, but as the duration increases above several seconds, the response starts to saturate. This nonlinearity of response is investigated and discussed further in [13, 17, 18], where it is shown that the response to a block design is approximately a factor of 1.5-2 less than that predicted by a linear model, extrapolating from short event-related experiments.

Furthermore, it should be remembered that the damaging effect (in the model-fitting) of misspecifying the haemodynamic response function is generally greater for event-related designs than for block - another factor meaning that a given B may be easier to attain with a block-design experiment.

Finally, note that this paper is only concerned with *within-session* analysis; cross-session and cross-subject variability/power are quite separate issues - for example, see [12, 15] for investigations of session variability *in practice*, [2, 19] on issues of inter-session/subject variance *modelling and estimation*, and [6] on issues of multi-subject power calculations.

The calculations described in this paper have been implemented in a recent release of the FEAT analysis tool (part of FSL), in the place of the less useful rank deficiency / estimability calculations that were used previously. We would hope that this will prove useful to experimenters as new experimental designs are being planned.

Appendix A - Using Efficiency Estimation in Power Calculations

If t_c is just set to t_α in the efficiency calculation summarised in equation 5 then one is defining the *exact* BOLD effect B required for statistical significance to be achieved. Of course, there is some uncertainty in estimation of B, meaning that estimated values will have a spread about the mean. For example, if the actual mean effect is B, then 50% of the estimates made will fall below the required level - we have 50% power (true positive rate). If the critical B is increased then this true positive rate (the chance of finding an effect when one is truly present) can be increased. We can easily adapt the efficiency calculation in the light of a power calculation, in order to give (for example) the standard required power of 80%.

According to standard power calculations, once one knows the distribution for the alternative hypothesis, one can estimate the true positive rate when thresholding at a given t_c level. In our case the relevant alternative ("activation") distribution for t is a non-central t distribution, though for the number of timepoints in FMRI data, this is identical to a shifted central t distribution. If one places an alternative distribution such that 20% of its lefthand tail is to the left of t_α, one has then defined the new (increased) critical t_c value which achieves the desired experimental power; see figure 6.

The exact dependence of the correction on t_α and the degrees of freedom is weak, and only causes a small fractional change in required BOLD effect; for all realistic scenarios it lies in the range 0.8-1 for a power of 80%. For example, for a false positive rate of $p < 0.05$ (uncorrected for multiple comparisons), degrees of freedom between 25 and infinity, achieving 80% power simply corresponds to setting t_c to $t_\alpha + 0.85$. For a typical multiple comparison correction to the false positive rate, the correction rises just a tiny amount, to 0.9. If required, the exact value can be easily computed, for example in MATLAB:

```
dof=200; power=80;
alpha=0.05; t_alpha=tinv(1-alpha,dof);      % uses inverse t-distribution CDF
                                            % or replace t_alpha with output from GRF, etc.
for t_c=0:0.01:100
  if nctcdf(t_alpha,dof,t_c) < 1-power/100  % noncentral t CDF
    break
  end
end
sprintf('t_alpha=%.2f t_c=%.2f',t_alpha,t_c)
```

Note that this correction assumes that one cares about controlling detection power at a particular location; if one is concerned

with global detectability rather than localisation, see [22] for mathematical details and a clear discussion of the relevant issues.

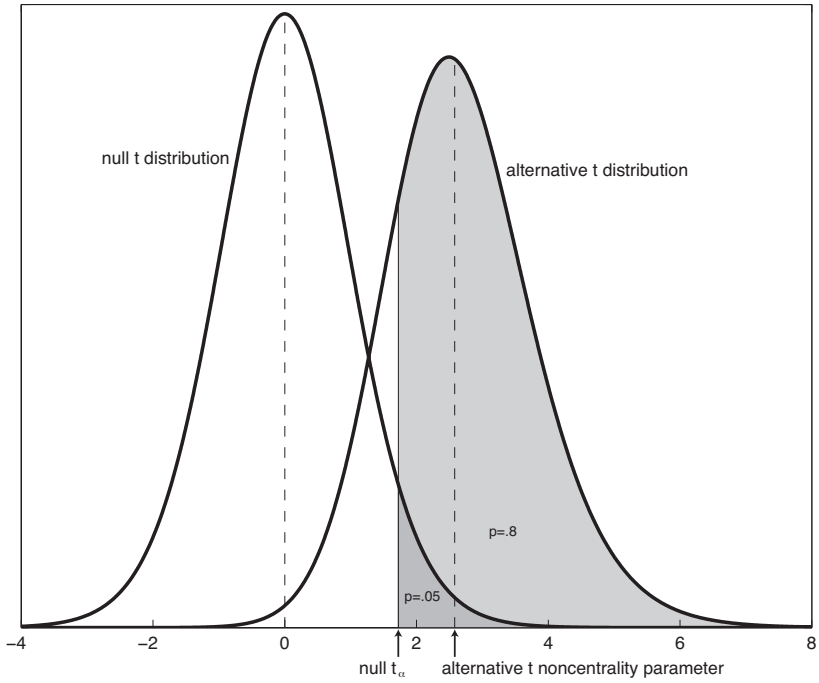

Figure 6: Illustration of controlling power (true positive rate) as well as the false positive rate. First, the null t distribution is used to set the false positive rate, for example by setting p$<$0.05 (if no corrections for multiple comparisons are being made) and finding the equivalent critical t value t_α. The alternative t distribution is then adjusted (by adjusting the noncentrality parameter, which is effectively the "centre" of the alternative distribution) until 80% of its area falls to the right of t_α. This means that if the BOLD effect B is set using the alternative t noncentrality parameter in equation 5 (instead of t_α), the experiment will have 80% power, i.e., an 80% chance of reporting the effect as statistically significant. Note that at typical degrees of freedom of FMRI timeseries, the alternative t distribution will be identical to a shifted version of the null; for this example, the degrees of freedom was set to 20 to illustrate the slight difference in shape.

Appendix B - Effective Regressor for a Complex Contrast

We show here how it is possible to replace a general design matrix and contrast with a new design matrix, for which the parameter estimate (and its variance) associated with the first regressor in the new design matrix is equal to the original contrast of parameter estimates. The purpose, for this paper, of doing this, is to make it clear how to estimate the peak-peak height and standard deviation of the new regressor, something which is not obvious in the original context of a contrast between different regressors' parameter estimates.

Given the standard GLM, $Y = X\beta + e$, and general contrast c, an equivalent model without contrasts (but with confounds) exists in the form

$$Y = \begin{bmatrix} X_{\mathit{eff}} & X_\perp \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} + e.$$

That is, $\widehat{b} = c'\widehat{\beta}$, $\mathrm{Cov}(\widehat{b}) = c'\mathrm{Cov}(\widehat{\beta})c$ and the modelled signal space is $\mathrm{Span}(X_{\mathit{eff}}) \cup \mathrm{Span}(X_\perp) = \mathrm{Span}(X)$ in the pre-whitened space.

The proof is by construction. Firstly, let c_2 be a set of contrasts that when combined with c form a complete linearly

independent set of contrasts. That is, the matrix $[c\ \ c_2]$ will be full rank (and hence invertible). Then let

$$X_{\mathit{eff}} = XQcF_1 \quad \text{and} \quad X_\perp = XQc_3F_3$$

where

$$Q = (X'X)^{-1}, \quad F_1 = (c'Qc)^{-1}, \quad c_3 = c_2 - P_c c_2, \quad P_c = c(c'Qc)^{-1}c'Q, \quad \text{and} \quad F_3 = (c_3'Qc_3)^{-1}.$$

From these definitions it is easy to see that $c'Qc_3 = 0$, which represents an orthogonality condition. It is straightforward to verify that the combined span of X_{eff} and $X_\perp$ is equal to the span of X. Consequently,

$$X_\perp' X_{\mathit{eff}} = F_3 c_3' Q X' X Q c F_1 = F_3 c_3' Q c F_1 = 0.$$

Therefore, $X_\perp$ and X_{eff} are orthogonal as well.

The estimation equations for the model become

$$\begin{aligned} \text{Cov}\left(\begin{bmatrix} \widehat{b} \\ \widehat{a} \end{bmatrix}\right) &= \begin{bmatrix} (X_{\mathit{eff}}'X_{\mathit{eff}})^{-1} & 0 \\ 0 & (X_\perp'X_\perp)^{-1} \end{bmatrix}, \\ \begin{bmatrix} \widehat{b} \\ \widehat{a} \end{bmatrix} &= \begin{bmatrix} (X_{\mathit{eff}}'X_{\mathit{eff}})^{-1}X_{\mathit{eff}}' \\ (X_\perp'X_\perp)^{-1}X_\perp' \end{bmatrix} Y. \end{aligned}$$

Thus

$$\begin{aligned} \text{Cov}(\widehat{b}) &= (F_1 c'QX'XQcF_1)^{-1} \\ &= (F_1 c'QcF_1)^{-1} \\ &= F_1^{-1} = c'(X'X)^{-1}c \\ &= c'\text{Cov}(\widehat{\beta})c \end{aligned}$$

and

$$\begin{aligned} \widehat{b} &= \text{Cov}(\widehat{b})(F_1 c'QX')Y \\ &= c'(QX')Y \\ &= c'\widehat{\beta}. \end{aligned}$$

References

[1] C.F. Beckmann, M. De Luca, J.T. Devlin, and S.M. Smith. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society*, 360(1457):1001–1013, 2005.

[2] C.F. Beckmann, M. Jenkinson, and S.M. Smith. General multi-level linear modelling for group analysis in FMRI. *NeuroImage*, 20:1052–1063, 2003.

[3] C.F. Beckmann and S.M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. on Medical Imaging*, 23(2):137–152, 2004.

[4] M. A. Burock and A. M. Dale. Estimation and detection of event-related fMRI signals with temporally correlated noise: A statistically efficient and unbiased approach. *Human Brain Mapping*, 11:249–260, 2000.

[5] A.M. Dale. Optimal experimental design for event-related fMRI. *Human Brain Mapping*, 8(2-3):109–114, 1999.

[6] J.E. Desmond and G.H. Glover. Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods*, 118:115–128, 2002.

[7] M. Jenkinson, P.R. Bannister, J.M. Brady, and S.M. Smith. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002.

[8] M. Jenkinson and S.M. Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, June 2001.

[9] T.T. Liu. Efficiency, power, and entropy in event-related fMRI with multiple trial types. Part II: design of experiments. *NeuroImage*, 21:400–413, 2004.

[10] T.T. Liu and L.R. Frank. Efficiency, power, and entropy in event-related fMRI with multiple trial types. Part I: theory. *NeuroImage*, 21:387–400, 2004.

[11] T.T. Liu, L.R. Frank, E.C. Wong, and R.B. Buxton. Detection power, estimation efficiency, and predictability in event-related fMRI. *NeuroImage*, 13:759–773, 2001.

[12] D.J. McGonigle, A.M. Howseman, B.S. Athwal, K.J. Friston, R.S.J. Frackowiak, and A.P. Holmes. Variability in fMRI: An examination of intersession differences. *NeuroImage*, 11:708–734, 2000.

[13] K.L. Miller, W.-M. Luh, T.T. Liu, A. Martinez, T. Obata, E.C. Wong, L.R. Frank, and R.B. Buxton. Nonlinear temporal dynamics of the cerebral blood flow response. *Human Brain Mapping*, 13:1–12, 2001.

[14] S.M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, November 2002.

[15] S.M. Smith, C.F. Beckmann, N. Ramnani, M.W. Woolrich, P.R. Bannister, M. Jenkinson, P.M. Matthews, and D.J. McGonigle. Variability in FMRI: A re-examination of intersession differences. *Human Brain Mapping*, 24:248–257, 2005.

[16] C. Triantafyllou, R.D. Hoge, G. Krueger, C.J. Wiggins, A. Potthast, G.C. Wiggins, and L.L. Wald. Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. *NeuroImage*, 26:243–250, 2005.

[17] T.D. Wager and T.E. Nichols. Optimization of experimental design in fMRI: A general framework using a genetic algorithm. *NeuroImage*, 18:293–309, 2003.

[18] T.D. Wager, A. Vazquez, L. Hernandez, and D.C. Noll. Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. *NeuroImage*, 25:206–218, 2005.

[19] M.W. Woolrich, T.E.J. Behrens, C.F. Beckmann, M. Jenkinson, and S.M. Smith. Multi-level linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, 21(4):1732–1747, 2004.

[20] M.W. Woolrich, B.D. Ripley, J.M. Brady, and S.M. Smith. Temporal autocorrelation in univariate linear modelling of FMRI data. *NeuroImage*, 14(6):1370–1386, 2001.

[21] 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*, 15(1):1–15, 2002.

[22] E. Zarahn and M. Slifstein. A reference effect approach for power analysis in fMRI. *NeuroImage*, 14:768–779, 2001.

[23] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. *IEEE Trans. on Medical Imaging*, 20(1):45–57, 2001.