Chapter Two

Statistic Images

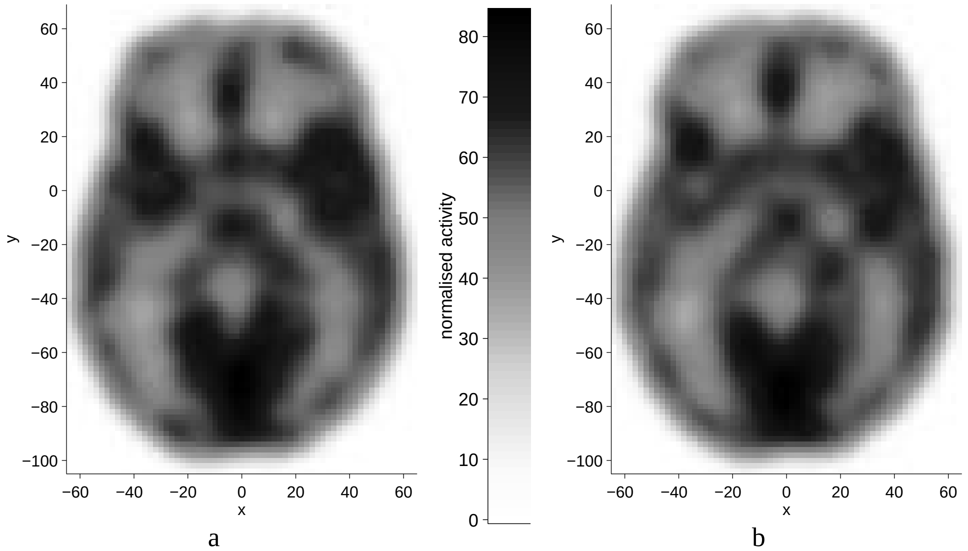
After study design, scanning, reconstruction, alignment, and possibly anatomical normalisation and primary smoothing, the adjusted images are ready for analysis. Voxel-by-voxel approaches proceed by analysing the data at each voxel, across the data, using univariate techniques. This results in the computation of a statistic for each voxel, giving an image of statistics, termed a *statistic image*.

There are various models which can be used when forming statistic images for simple activation studies. In this chapter the problems of changing global CBF are discussed, various models for single and multiple subject activation experiments introduced, and their relative merits and shortcomings considered. Particular attention is given to the models commonly used in practice. There is much disagreement in the functional neuroimaging world as to the "right" model and statistic to use, so a chapter discussing the issues is timely.

The case of the simple activation experiment with two conditions, "rest" and "activation", shall be used throughout, and the "V5" study data used as an example.

“Raw” Data

The “raw” data we shall consider for statistical analysis are the scan images after the pre-processing described in chapter 1. The images in fig.22 are of the first two scans of the second subject in the “V5” study. Although acquired under each of the conditions, there is only a small discernible difference between these two images, in the visual cortex at the posterior of the brain.



a b

Figure 22

Counts (rA) images (after pre-processing as previously described) from the first two scans of the first subject in the “V5” study, taken under: (a) condition B (“rest”), and (b) condition A (“activation”). The images have been normalised for global changes by proportional scaling to a gA of 50ml/min/dl. The colour scale is graduated in units of normalised activity. The AC-PC plane is shown, in standard Talairach co-ordinates.

Activation Study Design

Recall that in these simple activation studies each subject is scanned repeatedly under “rest” (or “baseline”) (B) and “active” (A) conditions during the course of a single scanning session of 2*M* scans. The order of allocation of conditions to the 2*M* scans for each subject is usually either alternating (BABA… or ABAB… as in the “V5” study), or alternating in pairs, beginning with a single scan under one condition (BAABBAAB… or ABBAABBA…). In such multi-subject designs subjects are randomised to a presentation order (B first or A first) in a balanced fashion.

Notation

We shall only be considering voxels of the image space that correspond to brain tissue in all the scans under consideration. These voxels, the *intracerebral voxels*, can be identified for an individual by reference to a co-registered MRI scan. In the absence of such a scan, the intracerebral volume can be fairly reliably identified directly from PET images by thresholding them at a third of their maximum value. The volume so identified usually includes the ventricles as well as the brain.

Let Y_{ijqk} denote the rCBF (or rA) measurement at voxel k (k =1,…,K); of scan j (j =1,…,M); under condition q (q = 0,1, 0 = "rest"); on subject i (i =1,…N). For brevity, we shall allow ourselves to refer to the value associated with a particular volume element simply as the value of the voxel.

Take W = {1,…,k,…,K} to be the set of indices of the intracerebral voxels. In addition to referring to voxels by their index, we shall refer to sets of voxels by the set of their indices. Thus, for U⊆W, we shall take "the voxels in U" to mean the set of voxels with indices in U, and similarly take "the region U" to mean the region of the image space that is the union of the voxels (volume elements) with indices in U.

2.1. Global Changes

Global cerebral blood flow (gCBF) varies (fig.23). Different subjects have different gCBF, and within a subject gCBF varies over the course of a scanning session, tending to decrease as the subject becomes relaxed. If "counts" images of rA are being considered as indicators of rCBF, then changes in administered dose and *head fraction*,[20] as well as changes in gCBF cause changes in gA. The latter two effects are confounded: If the head fraction remains constant then changes in gCBF cause no change in gA for the same administered dose. Clearly for reasons of specificity and sensitivity, differences in gCBF (gA) must be accounted for when examining sets of rCBF (rA) images for changes.

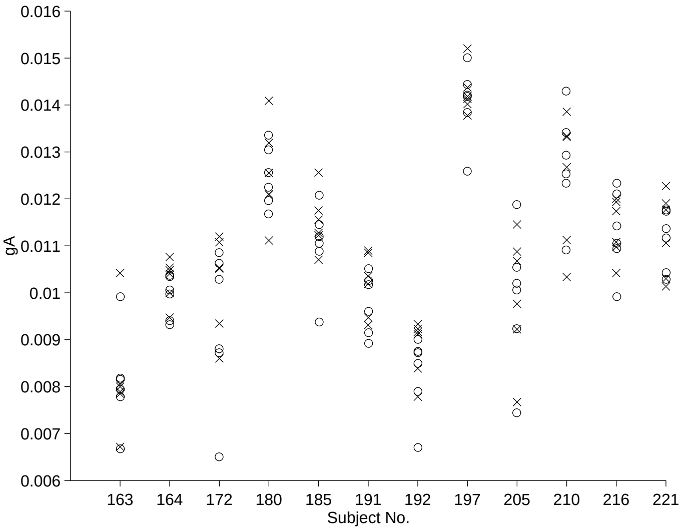


Figure 23

Global activity by subject for the "V5" subjects. Symbols indicate the condition, "o" for condition B, "×" for condition A. Global activity was measured as mean value of the intracerebral voxels as identified by 1/3rd max. thresholding (see text).

"Dose ranging"

Although most centres use very accurate techniques for tracer administration, the administered dose is sometimes deliberately adjusted during a scanning session. This is to optimise the performance of the scanner, and to ensure that the total dose over the scanning session is close to the maximum allowable. Such *dose ranging* was used in the "V5" study.

[20] The *head fraction* is the fraction of the whole body blood flow occurring in the head. This is fairly constant for an individual under normal conditions, but varies between subjects.

2.1.1. Computation of gCBF

Global (and regional) cerebral blood flow is measured in units of millilitres per minute per decilitre of brain tissue (ml/min/dl). Let x_{ijq} denote the gCBF (gA) for scan j under condition q on subject i. For rCBF (rA) images the gCBF (gA) is usually given as a mean per voxel, computed by taking the average value of voxels corresponding to brain tissue. The gCBF (gA) for scan j on subject i is:

$$x_{ijq} = \frac{1}{K}\sum_{k=1}^{K} Y_{ijqk} = \overline{Y}_{ijq\bullet} \qquad (11)$$

2.1.2. Correcting for changes in gCBF: Normalised images

As changes in rCBF were sought, the pioneers of functional neuroimaging with PET examined *subtraction images*, formed by subtracting a “rest” scan from an “active” condition scan. This required that changes in gCBF be accounted for before the subtraction.

In practice the voxel values for rCBF (rA) images are normalised by dividing by the global flow (activity) and then multiplying by the normal gCBF of 50ml/min/dl, to give the images a physiologically relevant scale (eqn.12). This *proportional scaling* method has the apparent advantage of simplicity; the normalised images are easy to compute, and can be interpreted as regional activity relative to the global level.

$$Y'_{ijqk} = Y_{ijqk} \,/\, (x_{ijq} \,/\, 50) \qquad (12)$$

The assumption is that for a stable brain state, the rCBF depends proportionally on the gCBF. Alternatively, the assumption can be discarded on the understanding that relative values are to be analysed as indicators of neuronal activity.

2.2. Single Subject Activation Experiments

We begin our discussion of the formation of statistic images with the single study activation experiment. Dropping the subject index (i) from our notation, let Y_{jqk} denote the rCBF (rA) measurement at voxel k of scan j under condition q. Let x_{jq} denote the corresponding gCBF (gA).

2.2.1. Two sample *t*-statistic

The normalised images Y'_{jqk} constitute M observations under each of the two conditions for each voxel. If the stimulus of the "active" ($q = 1$) condition causes an increase in neuronal activity at a particular voxel location, then the rCBF values at that voxel should be increased in the "active" scans. The conditions are presented to the subject in pairs (sometimes with randomly assigned order within pairs) to allow for time effects, so an appropriate test would reflect this blocked design. However, in practice the blocking is ignored, and a two sample *t*-test used rather than a paired test. We shall return to this point later.

Computing the value T_k of the two sample *t*-statistic for each voxel gives an image of statistics $\boldsymbol{T} = (T_1,\ldots,T_K)$:

$$T_k = \frac{\overline{Y'_{\bullet 1k}} - \overline{Y'_{\bullet 0k}}}{\sqrt{S_k^2\left(\frac{1}{M} + \frac{1}{M}\right)}} \qquad (13)$$

for S_k^2 the pooled variance estimate: $S_k^2 = \dfrac{1}{2M-2}\displaystyle\sum_{q=0}^{1}\sum_{j=1}^{M}\left(Y'_{jqk} - \overline{Y'_{\bullet qk}}\right)^2$

Assuming $Y'_{jqk} \sim N(\mu_q,\sigma_k^2)$, then $T_k \sim t_{2M\text{-}2}$ under $\mathrm{H}_k{:}\mu_0 = \mu_1$. $\boldsymbol{T}$ is then referred to as a *t*-statistic image with $2M$-2 degrees of freedom, since each voxel k has associated value distributed as a Student's *t*-distribution with $2M$-2 degrees of freedom under H_k. Since it is activation we are interested in, it is usual to test against the one-sided alternative hypothesis $\overline{\mathrm{H}}_k{:}\mu_k > 0$. A *p*-value for each voxel can then be computed, and arranged into an (unadjusted) *p*-value image, indicating evidence of activation.

Two sample model: Proportional regression

The model implicit in the use of normalised images is one of proportionality. The two sample *t*-test on normalised images above is equivalent to a one way ANOVA for two treatments at each voxel, with the model for the data at voxel k given by:

$$Y'_{jqk} = \alpha_{qk} + \varepsilon'_{jqk} \qquad \text{where } \varepsilon'_{jqk} \overset{\text{iid}}{\approx} N(0, \sigma_k^2) \qquad (14)$$

For the purposes of this discussion, we shall take the normalised images Y'_{jqk} to be Y_{jqk}/x_{jq} , rather than $Y_{jqk}/(x_{jq}\,/50)$ (eqn.12). Assuming that the global flow (activity) x_{jq} is measured with negligible error, multiplying the above model by x_{jq} gives eqn.15, a weighted proportional regression model, with condition dependent regression coefficient. The proposed *t*-test is a test of equal slope in this model.

$$Y_{jqk} = \alpha_{qk}\, x_{jq} + \varepsilon_{jqk} \qquad \text{where } \varepsilon_{jqk} \overset{\text{iid}}{\approx} N(0, x_{jq}^2\sigma_k^2) \qquad (15)$$

It is apparent from the PET literature that many researchers overlook the weighting of the error variance, and consider the two sample *t*-test model as equivalent to a proportional regression model with homogeneous variance:

$$Y_{jqk} = \alpha_{qk}\, x_{jq} + \varepsilon_{jqk} \qquad \text{where } \varepsilon_{jqk} \overset{\text{iid}}{\sim} N(0, \sigma'^{2}_{k}) \qquad (16)$$

In this form the model can be easily compared with other linear models. Scheffé (1959, §1.5) notes that the method of least squares with inappropriate weights still leads to unbiased estimates of the parameters, though not the same estimates. Clearly estimates of the variance will be biased. Although an assumption of constant variance at each voxel in models 15 and 16 cannot be true simultaneously, it is perhaps equally feasible in either one, particularly for subjects whose gCBF (gA) remains roughly constant, where there is little difference between the models.[21] Thus, this oversight may not be too serious.

[21] For example, consider the second subject in the "V5" study: The maximum absolute difference between the two sample *t*-statistic and a *t*-statistic for equality of slope in eqn.16 (over all the voxels in the AC-PC plane) is 0.300 (3dp).

2.2.2. Friston’s model: AnCova

Friston *et al.* (1990) proposed a more general model for the relationship between regional and global cerebral blood flow.

A standard linear regression model for cross sectional data

The actual relationship between regional and global CBF is unlikely to be linear, but is possibly well approximated by a straight line for the normal range of global values (see fig.24). Since the normal range of global values is far from zero, the relationship is likely to be better approximated by an arbitrary line rather than one constrained to the origin. For scans acquired under the rest condition ($q = 0$), Friston proposed a standard simple regression model at each voxel:

$$Y_{j0k} = \alpha_k + \beta_k\,(x_{j0} - \overline{x}_{\bullet\bullet}) + \varepsilon_{j0k} \qquad \text{where } \varepsilon_{j0k} \overset{\text{iid}}{\sim} N(0,\sigma_k^2) \tag{17}$$

The parameters in model 17 (intercept, regression coefficient and variance) are distinct for each voxel, since the global changes in CBF are likely to affect distinct regions differently[22]. Thus we have K simultaneous regressions, one for each voxel.

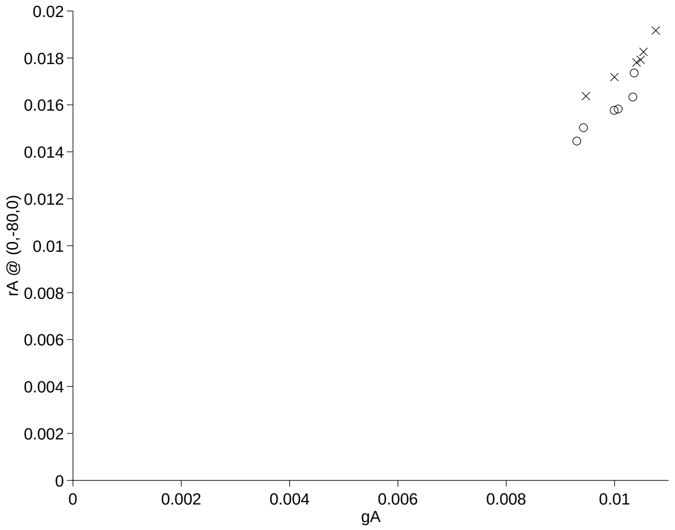


Figure 24
Regional activity for voxel at Talairach co-ordinates (0,-80,0) plotted against global activity, for the twelve scans of the second subject in the “V5” study. The conditions are indicated by the symbols, “o” for condition A (the rest condition) and “×” for condition B (the active condition). This voxel was chosen because it is in the middle of the visual cortex, which is expected to be stimulated by the activation condition.

Covariate is mean of response variables

By way of an aside, note that if all the voxels used to compute the global values are included in the ANCOVA analysis, then this imposes a loose constraint on the parameters. Writing model 17 non-centred, and dropping the $q = 0$ subscripts (so that Y_{jk} is the rCBF (rA) value at voxel k of scan j obtained under the rest condition, with x_j the corresponding gCBF (gA)) we obtain the simultaneous regressions of eqn.17a:

$$Y_{jk} = \alpha_k + \beta_k\, x_j + \varepsilon_{jk} \qquad \text{where } \varepsilon_{jk} \overset{\text{iid}}{\sim} N(0,\sigma_k^2) \tag{17a}$$

[22] Specifically, “it is physiologically likely that sensitivity of rCBF to gCBF in any one area depends on the neuronal projections to that area”. (From Friston *et. al.*, 1990.)

Suppose that all K voxels are used to compute the global values, then, summing eqn.17a over $k=1,\ldots,K$ we obtain:

$$K\overline{Y}_{j\bullet} = K x_j = \sum_{k=1}^{K}\alpha_k + x_j\sum_{k=1}^{K}\beta_k + \sum_{k=1}^{K}\varepsilon_{jk}$$

$$\Leftrightarrow x_j = \overline{\alpha}_{\bullet} + x_j\,\overline{\beta}_{\bullet} + \overline{\varepsilon}_{j\bullet}$$

Interpreting this as a regression of the x_j's on themselves, rather than an equation derived from separate regressions, Clark & Carson (1993) asserted that $\overline{\alpha}_{\bullet}$ must be zero, and $\overline{\beta}_{\bullet}$ must be one. In fact this is true for the fitted values, as pointed out by Friston (1994). Consider the fitted values for the simple regression of eqn.17a:

$$\sum_{k=1}^{K}\hat{\beta}_k = \sum_{k=1}^{K}\frac{\sum_i (x_i - \overline{x}_{\bullet})(Y_{ik} - \overline{Y}_{\bullet k})}{\sum_i (x_i - \overline{x}_{\bullet})^2}$$

$$= \frac{\sum_i (x_i - \overline{x}_{\bullet})(K\overline{Y}_{i\bullet} - K\overline{Y}_{\bullet\bullet})}{\sum_i (x_i - \overline{x}_{\bullet})^2} = K$$

$$\sum_{k=1}^{K}\hat{\alpha}_k = \sum_{k=1}^{K}\left(\overline{Y}_{\bullet k} - \hat{\beta}_k\,\overline{x}_{\bullet}\right)$$

$$= K\overline{Y}_{\bullet\bullet} - K\,\overline{x}_{\bullet} = 0$$

So, $\overline{\hat{\alpha}}_{\bullet}$ must be zero, and $\overline{\hat{\beta}}_{\bullet}$ must be one. Given the large number of voxels under consideration, this restriction is hardly of any consequence to the regression at an individual voxel. Since the model is a means to inference rather than an end in itself, most ignore this subtlety. If global values are computed over more voxels than are considered for the regression models, the problem disappears.

Considering the regression at a single voxel, one might be concerned that the covariate is measured with error. This concern is ill founded. Firstly, regarding the measured gCBF (gA) as a random variable with mean the true global value; its variance is negligible compared to that of the estimated rCBF (rA) at a single voxel, since the global value is the mean of the voxel values. Secondly, for fixed effects models it doesn't matter anyway: Scheffé (1959, p195) states (for fixed effects models), that if the distribution of the response variables conditional on the explanatory variables is "assumed to hold for all possible values of the observations on the [explanatory] variables, then, regardless of the joint distribution of observations on the [explanatory] variables, the conditional significance levels and conditional confidence coefficients are constant, and hence the same unconditionally".

One way AnCova model

To include activation scans in the model, Friston *et al.* (1990) proposed a (balanced) one way analysis of covariance (AnCova) model. The model for voxel k is then:

$$Y_{jqk} = \alpha_{qk} + \beta_k\,(x_{jq} - \overline{x}_{\bullet\bullet}) + \varepsilon_{jqk} \qquad \text{where } \varepsilon_{jqk} \overset{\text{iid}}{\sim} N(0,\sigma_k^2) \tag{18}$$

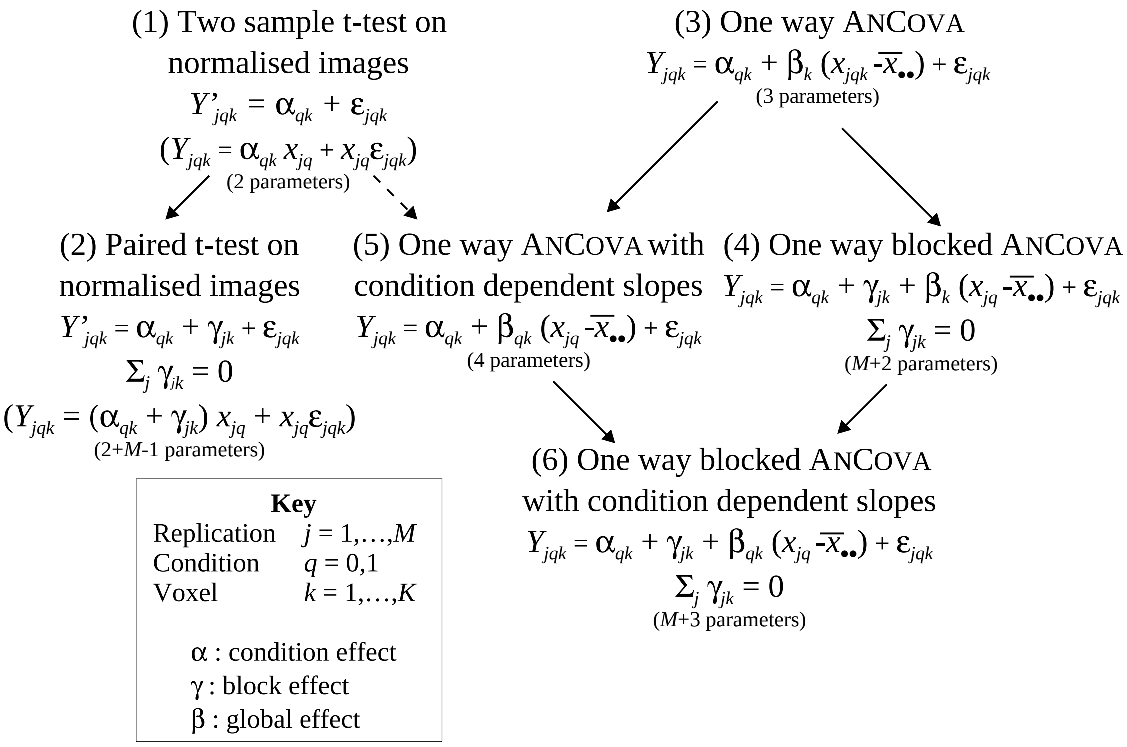
The effect of an activation is assumed to be additive. rCBF (rA) is increased by local neuronal activity by an amount independent of the underlying gCBF (gA). In this centred model the images of condition effects $\alpha_q = (\alpha_{q1},\ldots,\alpha_{qK})$ can be viewed as mean images for the condition, adjusted to a gCBF (gA) of $\bar{x}_{\bullet\bullet}$.

The null hypothesis of no activation effect at voxel k: H_k: $\alpha_{0k} = \alpha_{1k}$, can then be assessed. Since we are interested in activation, it is usual to test against the alternative hypothesis $\bar{H}_k$:$\alpha_{1k} > \alpha_{0k}$ via a t-test of H_k: α_{1k}-α_{0k} = 0 against $\bar{H}_k$:α_{1k} -α_{0k} > 0.

2.2.3. Model selection for single subject

The two test statistics proposed for a single subject activation experiment are the two sample t-test on normalised data (eqn.13, p54), and the simple one way ANCOVA (eqn.18, p57). The relationship between these two approaches can be seen in the following taxonomy of possible models for a balanced single subject activation experiment (eqns.19).

Taxonomy of single subject models



(1) Two sample t-test on normalised images

$Y'_{jqk} = \alpha_{qk} + \varepsilon_{jqk}$

$(Y_{jqk} = \alpha_{qk}\, x_{jq} + x_{jq}\varepsilon_{jqk})$

(2 parameters)

(3) One way ANCOVA

$Y_{jqk} = \alpha_{qk} + \beta_k\,(x_{jqk} - \bar{x}_{\bullet\bullet}) + \varepsilon_{jqk}$

(3 parameters)

(2) Paired t-test on normalised images

$Y'_{jqk} = \alpha_{qk} + \gamma_{jk} + \varepsilon_{jqk}$

$\Sigma_j\, \gamma_{jk} = 0$

$(Y_{jqk} = (\alpha_{qk} + \gamma_{jk})\, x_{jq} + x_{jq}\varepsilon_{jqk})$

(2+M-1 parameters)

(5) One way ANCOVA with condition dependent slopes

$Y_{jqk} = \alpha_{qk} + \beta_{qk}\,(x_{jq} - \bar{x}_{\bullet\bullet}) + \varepsilon_{jqk}$

(4 parameters)

(4) One way blocked ANCOVA

$Y_{jqk} = \alpha_{qk} + \gamma_{jk} + \beta_k\,(x_{jq} - \bar{x}_{\bullet\bullet}) + \varepsilon_{jqk}$

$\Sigma_j\, \gamma_{jk} = 0$

(M+2 parameters)

(6) One way blocked ANCOVA with condition dependent slopes

$Y_{jqk} = \alpha_{qk} + \gamma_{jk} + \beta_{qk}\,(x_{jq} - \bar{x}_{\bullet\bullet}) + \varepsilon_{jqk}$

$\Sigma_j\, \gamma_{jk} = 0$

(M+3 parameters)

Key

Replication	$j = 1,\ldots,M$
Condition	$q = 0,1$
Voxel	$k = 1,\ldots,K$

α : condition effect
γ : block effect
β : global effect

Equations (19)

Recall that $j = 1,\ldots,M$ indexes the replication for scans acquired under condition $q = 0,1$, and $k = 1,\ldots,K$ indexes the voxels. Proportional scaling models have been written without the scaling to gCBF of 50ml/min/dl. Arrows indicate logical extensions of models. The dashed arrow indicates that model 19.5 is an extension of model 19.1 if the variance weights are ignored. Clearly the parameters may be different in each model, though the same symbols have been used. In all the models it is assumed that $\varepsilon_{jqk} \overset{\text{iid}}{\sim} N(0,\sigma_k^2)$, where σ_k is unique for each model.

The block effect γ_{jk} (eqns.19.2, 19.4 & 19.6) would model regional changes in neuronal activity (under both conditions) between successive pairs of scans that is not accounted for by global changes, such as localised decreases in rCBF in areas associated with anxiety.

A condition by replication interaction (α_{qjk}) would be justifiable on physiological terms as a *habituation* effect, modelling the reduced increase in brain activity as the subject becomes accustomed to the task set in the "active" condition. However, as there are no replications within treatment-block pairs, this cannot be considered, unless a simple parameterised trend is assumed.

2.2.3.1. Model selection for images

Model selection in this context is rather problematical.

Simultaneous model for all voxel

We are seeking a model that will simultaneously describe the relationship between rCBF and gCBF and design factors at all voxels. Separate regions may require separate models. Fitting a richer model than is appropriate, with redundant terms, reduces the degrees of freedom available for variance estimation, and hence reduces the power of ensuing tests at that voxel. Fitting a smaller model than is appropriate leaves the omitted effects in the residuals, introducing structure into the residuals in defiance of their assumed independence, and leads to increased variance estimates with more degrees of freedom than under the appropriate model. The validity of ensuing tests cannot be guaranteed: If the appropriate model has few degrees of freedom available for variance estimation then the spurious "extra" degrees of freedom from a smaller model may result in more lenient tests despite the increase in variance.

The cautious approach is to choose, as the model for all the voxels, the richest of the models appropriate for the voxels individually. Thus, consider a backward elimination procedure (Draper and Smith, 1981, §6.3). Starting with a saturated model, discard terms if there is insufficient evidence against the (omnibus) hypothesis that the terms in question are all zero for all voxels at significance level α.

Multiple comparisons

There remains the large multiple comparisons problem in using univariate statistics to compare models for each voxel. Multiple comparisons in this image setting is a major topic of this thesis, and is explored in detail in later chapters.

Since we are interested in finding a model that fits for all the voxels, in comparing two (nested) models the relevant question is: "Is there any evidence, at all, against the omnibus hypothesis that the extra parameters in the richer model are zero for all voxels." If there is evidence, then the richer model should be chosen for all the voxels. Thus, an omnibus multiple comparisons procedure with weak control over familywise error is required. The exceedence proportion test of Worsley (§3.4.2.) will be used, with a (probability) threshold of $\eta = 0.01$.

To facilitate the discussion, it is convenient to adopt some notation not fully utilised until we formally discuss multiple comparisons in chapter 3 (§3.1.1.). Recall that $\mathrm{W}=\{1,\ldots,K\}$ is the set of (indices of) voxels that correspond to the brain region of interest, assumed to be the whole intracerebral volume. For computational reasons, we shall examine the AC-PC plane only. Let $\mathrm{W}_{\mathrm{P8}} \subset \mathrm{W}$ denote the (indices) of the intracerebral voxels in this plane. If H_k, $k \in \mathrm{U}$ are a set of voxel hypotheses, then the intersection of these hypotheses is the omnibus hypothesis over voxels U, which we shall denote H_{U}.

H_W is then the overall omnibus hypothesis, and $H_{W_{P8}}$ the omnibus hypothesis for the AC-PC plane.

2.2.3.2. Model selection: Single subject activation studies

Model selection for "V5" subjects

For each of the "V5" subjects, table 25 gives *p*-values for the omnibus hypotheses that all the additional parameters in the richer model are redundant, for comparisons of all models in the taxonomy (eqns.19). For instance the column labelled "19.5 vs. 19.3" contains omnibus *p*-values for the hypotheses H_k:$\beta_{0k} = \beta_{1k}$, $k \in W_{P8}$, in model 19.5.

F-Statistic images for the hypotheses were computed voxel-by-voxel (for the AC-PC plane only) by comparing the residual sums of squares under the two models, the so called "extra sum of squares" likelihood ratio test (Draper and Smith, 1981, §2.7). (Practicalities of computation are discussed in §2.5.) This *F*-statistic image was then "transformed" to a standard Gaussian statistic image, replacing each voxel statistic with a standard Gaussian ordinate with identical probability of being exceeded (see §3.3.3.). The resulting Gaussian statistic image was then assumed to be a strictly stationary discrete Gaussian random field, and the variance-covariance matrix of partial derivatives estimated directly from the positive part of the image (see §3.3.5., and footnote[23]). The omnibus *p*-value was then computed using Worsley's exceedence proportion test (§3.4.2.) with a threshold of $-\Phi^{-1}(0.01)$, at significance level $\alpha = 0.05$.

The last comparison, labelled "19.5 vs. 19.1", was effected by fitting the two-sample *t*-test model for normalised images (eqn.19.1) via the proportional regression model (in parentheses in the taxonomy), ignoring the weighting of the error terms. Thus the validity of these comparisons is in doubt. (Recall discussion of the proportional regression view of §2.2.1., p.54.)

Subject	19.2 vs.	19.5 vs.	19.4 vs.	19.6 vs.	19.6 vs.	19.5 vs. 19.1
	19.1	19.3	19.3	19.5	19.4	
n163	**0.0051**	0.7501	0.1797	0.1834	0.1328	0.6750
n164	0.1192	0.6541	0.1030	0.4404	0.6587	0.7668
n172	0.1701	0.8757	0.2190	0.1875	0.8555	0.2916
n180	**0.0000**	0.5243	0.1280	0.2215	0.7144	**0.0000**
n185	0.5705	0.7921	0.1173	0.2210	0.8616	0.7049
n191	0.2451	0.9352	**0.0116**	0.0521	0.8583	0.6350
n192	0.0821	0.3869	0.7890	0.8745	0.7358	**0.0002**
n197	0.6245	0.5737	0.6066	0.7731	0.6828	0.7373
n205	0.0625	0.2243	0.1878	0.0608	0.2792	0.1076
n210	**0.0000**	0.7742	**0.0000**	**0.0004**	0.5646	0.7960
n216	0.3817	0.2407	0.1683	0.6032	0.6446	0.2729
n221	**0.0000**	0.7214	**0.0072**	**0.0216**	0.3107	0.6834

Table (25)
p-values (to 4dp) for the omnibus hypotheses comparing the possible models for individual subjects from the "V5" study. *p*-values less than $\alpha = 0.05$ have been emphasised with bold type.

In the taxonomy (eqns.19), there are two "saturated" models to start from, the paired *t*-test model and the two way blocked ANCOVA with condition dependent slopes.

[23] These "Gaussianised" *F*-statistic images have a fairly smooth surface above the X-Y plane, but a distinctly rougher one below. Since we are interested in high values of the Gaussian statistic image, the smoothness was estimated using only voxels with associated positive values.

Starting with the paired *t*-test model (eqn.19.2), we see that the two sample *t*-test model is adequate for eight of the twelve subjects; there is evidence of a block effect in the remaining five subjects. Considering the ANCOVA models, we see that there is insufficient evidence of condition dependent slopes (19.5 vs.19.3 & 19.6 vs.19.4). This leaves us with the one way blocked ANCOVA (eqn.19.4), within which there is significant evidence of a block effect in three subjects.

Block effects?

The two models in routine use for the analysis of single subject activation data (namely the two sample *t*-test model (§2.2.1.) and the one way ANCOVA (§2.2.2.)), do not consider block effects. The paired *t*-test and the blocked ANCOVA leave very few degrees of freedom for variance estimation (*M*-1 and *M*-2 respectively), and can therefore be highly conservative.

In most cases the block effects are likely to be slight relative to the activation effect, and can safely be ignored. Large block effects mask the condition effects, and add to the variance (albeit in a structured manner), making the two sample *t*-test and the one way ANCOVA conservative. For very large block effects the paired *t*-test and the blocked ANCOVA may be more powerful than their one way counterparts, despite the low degrees of freedom.

There remains the choice between proportional scaling and *t*-tests and an ANCOVA approach.

Proportional scaling vs. linear modelling for rest scans

It is difficult to compare the proportional scaling models and the AnCova models, since they are not directly related. Friston *et al*. (1990) justified the AnCova approach over proportional scaling by fitting a simple regression model (eqn.17) to the rest scans and assessing H_k:α_{0k}= 0 against $\overline{H}_k$:$\alpha_{0k} \neq 0$ with the usual *t*-statistic. (This disregards the weighting of the variance terms implicit in the proportional scaling approach.) They found significant evidence against the omnibus hypothesis H_W at the 5% level using their exceedence proportion test (§3.4.1.).[24] This is perhaps to be expected, given the distance of the data from the origin and the observation that the relationship between rCBF and gCBF is unlikely to be linear.

However, if uncalibrated “counts” images of rA are used as indicators of rCBF, then variations in the administered dose[25] may cause variations in gA that swamp those caused by changing gCBF, resulting in a relationship between rA and gA that *is* proportional. This appears to be the case for the majority of the “V5” subjects, as illustrated in the following table (table 26) of omnibus *p*-values for $H_{W_{P8}}$, where H_k:α_{0k}= 0 in eqn.17.

Subject	*p*-value
n163	0.3560
n164	0.7534
n172	0.6295
n180	**0.0000**
n185	0.8810
n191	0.7327
n192	0.4809
n197	0.7051
n205	**0.0009**
n210	0.7171
n216	0.6045
n221	0.7086

Table (26)

p-Values (to 4dp) for the omnibus hypothesis $H_{W_{P8}}$ of voxel hypothesis H_k:α_{0k} = 0 for all intracerebral voxels in the AC-PC plane, for the rest scans from each subject from the “V5” study. The model is a simple regression (eqn.17). These *p*-values were computed in a similar fashion to those in table 25 above. *t*-statistic images for the hypotheses were computed voxel-by-voxel for the AC-PC plane of each subject. Since we are interested in a two sided alternative hypothesis, the *t*-statistic image was “transformed” to a standard Gaussian statistic such that the standard Gaussian ordinate associated with each voxel is as likely to be exceeded (by a standard Gaussian random variable) as the *t*-statistic is to be exceeded *in absolute value* by an appropriately distributed *t* random variable. (See §3.3.3.). These Gaussian statistic images were then assessed by Worsley’s exceedence proportion test as before, at threshold $-\Phi^{-1}(0.01)$. *p*-values less than α = 0.05 have been emphasised with bold type.

[24] Friston *et al.* (1990), due to having a limited amount of data, assessed the proportionality hypothesis on only eight rCBF scans, these being the two “rest” scans on each subject from an experiment on four subjects. Subject effects were not considered, the resulting correlations in the errors and probable underestimation of the variance were ignored. (See Miller, 1986, §5.5.) However, Friston *et al.* report that a test for proportionality test had been a routine part of their analysis method for a period of time, and that H_W was consistently rejected.

[25] For example, due to “dose ranging”, which was used in the “V5” study.

2.2.4. Conclusions: Single subject statistic images

AnCova for calibrated rCBF data

For the analysis of single subject activation experiments with true (calibrated) rCBF data, this author's recommendation is the one way AnCova approach (as proposed by Friston *et al.*, 1990). If the AnCova model fits well, with regression coefficient far from unity, then the *t*-test approach on normalised images may be insensitive to activation. If the proportional regression model implicit in the two sample *t*-test on normalised data (ignoring the weighting of the error terms) fits well, then the AnCova may be expected to give a slightly conservative test, since it has an extra parameter to fit. Thus an AnCova approach would appear to possess the best all round properties. The use of the one way AnCova may result in a loss of power in the presence of large block effects, in which case the one way blocked AnCova should be used.

The choice of a sensitive but robust method is particularly pertinent since any one PET centre will analyse studies with a set method, rather than selecting a model for the data at hand. The apparent complexity (to the statistically naive) of AnCova, and the increased difficulty of computation of statistic images (as compared to a *t*-test), are the main barriers to its routine use.

Proportional scaling for large variations in introduced dose

If no arterial sampling is undertaken, and relative activity is being examined as an indicator of CBF, then a proportional model may well be appropriate. In this situation the *t*-test on normalised data is likely to be more powerful then the one way AnCova, since the *t*-test has an additional degree of freedom (of some consequence when the degrees of freedom are so low), and since the assumption of constant regression coefficient across conditions in the AnCova model is likely to be false.

Since the majority of functional mapping experiments use "counts" images of relative activity, this conjecture deserves further examination on real data sets.

2.3. Multiple Subject Activation Experiments

Consider now a multiple subject activation experiment. Recall our notation: Y_{ijqk} denotes the rCBF (rA) measurement at voxel $k=1,\dots,K$, of scan $j=1,\dots,M$, under condition $q=0,1$ (0 = "rest", 1 = "active"), on subject $i=1,\dots N$; and x_{ijq} is the corresponding gCBF (gA).

2.3.1. Proportional scaling approach

Proponents of the proportional scaling approach for the normalisation of rCBF (rA) images for global changes, analyse multiple subject activation studies using a paired *t*-statistic at each voxel, pairing the mean of the (normalised) rest scans with the mean of the (normalised) active scans for each subject.

2.3.1.1. t-statistic on subject difference images

Specifically, the data for each subject is collapsed into a *subject difference image* $\Delta_i = (\Delta_{i1},\dots,\Delta_{iK})$ by subtracting (for each voxel) the mean of the "rest" scans from the mean of the "active" scans, after normalisation for global changes (eqn.20). This constitutes the pairing, and *t*-statistic $\boldsymbol{T} = (T_1,\dots,T_K)$ is computed in the usual way (eqn.21).

$$\Delta_{ik} = \overline{Y'}_{i\bullet 1k} - \overline{Y'}_{i\bullet 0k} \tag{20}$$

$$T_k = \frac{\overline{\Delta}_{\bullet k}}{\sqrt{S_k^2/N}} \tag{21}$$

where $\overline{\Delta}_{\bullet k} = \frac{1}{N}\sum_{i=1}^{N}\Delta_{ik}$ is the *study mean difference* at voxel *k*,(22)

and $S_k^2 = \frac{1}{N\text{-}1}\sum_{i=1}^{N}\left(\Delta_{ik} - \overline{\Delta}_{\bullet k}\right)^2$ is the variance estimate (23)

Distributional results

Assuming $\Delta_{ik} \sim N(\mu_k,\sigma_k^2)$, then $\overline{\Delta}_{\bullet k} \sim N(\mu_k, \sigma_k^2 / N)$, and $S_k^2 \sim \frac{\sigma^2\chi^2_{N-1}}{N-1}$. Under $\mathrm{H}_k{:}\mu_k = 0$, $T_k \sim t_{N-1}$, a Student's *t*-distribution with *N*-1 degrees of freedom. For the one-sided alternative hypotheses $\overline{\mathrm{H}}_k{:}\mu_k > 0$, a *p*-value for each voxel can be computed, giving an (unadjusted) *p*-value image indicating evidence of activation.

2.3.1.2. Discussion of t-test on subject difference images

Robustness: Assumptions

The collapsing of the M scans for each condition to mean rest and activation images $\overline{Y'}_{i\bullet 0k}$ and $\overline{Y'}_{i\bullet 1k}$ for each subject doubtless gives a robust test. The only assumptions are $\Delta_{ik} \sim N(\mu_k,\sigma_k^2)$, which appear to be reasonable for the "V5" study (see §2.6.1.). Averaging the data lends increased credence to assumptions of normality, by appeal to the Central Limit Theorem. No assumption needs to be made about the intra-subject variation of rCBF values.

Robustness: Subject, block, habituation and linear trend effects absorbed

Since the test is paired, subject effects are also absorbed. The test assesses the mean activation (after normalisation) over a scanning session, and thus block effects cancel out, and habituation causes no problem. (Recall that each consecutive pair of scan slots forms a block, and that habituation is a block by condition interaction.) Localised linear time effects also cancel out, provided they are constant across subjects and that subjects have been allocated to condition presentation order (ABAB… or BABA…) in a balanced fashion. Global changes in images are removed by the proportional scaling.

Robustness: Random subject effects incorporated

If different subjects respond to the stimulus differently (a condition by subject interaction), then an appropriate model is $\Delta_{ik} \sim N(\mu_{ik},\sigma_k^2)$, that is, a different mean activation for each subject. If the subjects are randomly sampled from a target population, then μ_{ik} may be treated as a simple random effect, $\mu_{ik} \sim N(\mu_k, \tau_k^2)$, where μ_k is the population mean activation effect. Then $\Delta_{ik} \sim N(\mu_k, \sigma_k^2 + \tau_k^2)$, and the t-test on subject difference images provides a valid test for a hypothesised zero mean activation effect (H_k:μ_k=0) for the population.

Model for t-test on subject difference images

Since a paired t-test is equivalent to a two way blocked ANOVA with two treatments, the t-statistic proposed may be viewed as the test statistic for H_k:α_{1k}-α_{0k} = 0 within the model:

$$\overline{Y'}_{i\bullet qk} = \alpha_{qk} + \gamma_{ik} + \varepsilon_{iqk} \qquad \text{where } \varepsilon_{iqk} \overset{\text{iid}}{\sim} N(0,\sigma_k^2) \qquad (24)$$

Further, it can be shown that the (two-sided) t-test on subject difference images is equivalent to the F-test for no main effect in a two-way mixed effects ANOVA model for the normalised images. In this model (eqn.25) the main effects (the condition effects α) are assumed fixed, and the block effects (the subject effects γ) are assumed random. The subject by condition interaction effects $\alpha\gamma$ are also considered random.

$$Y'_{ijqk} = \alpha_{qk} + \gamma_{ik} + \alpha\gamma_{qik} + \varepsilon_{ijqk} \qquad \text{where } \varepsilon_{ijqk} \overset{\text{iid}}{\sim} N(0,\sigma_k^2) \qquad (25)$$

The F-statistic for main effect is the ratio of the (mean) sums of squares for the main effect and for the interaction (Scheffé, 1959; Miller, 1986). This F-statistic is the square of the t-statistic of eqn.21.

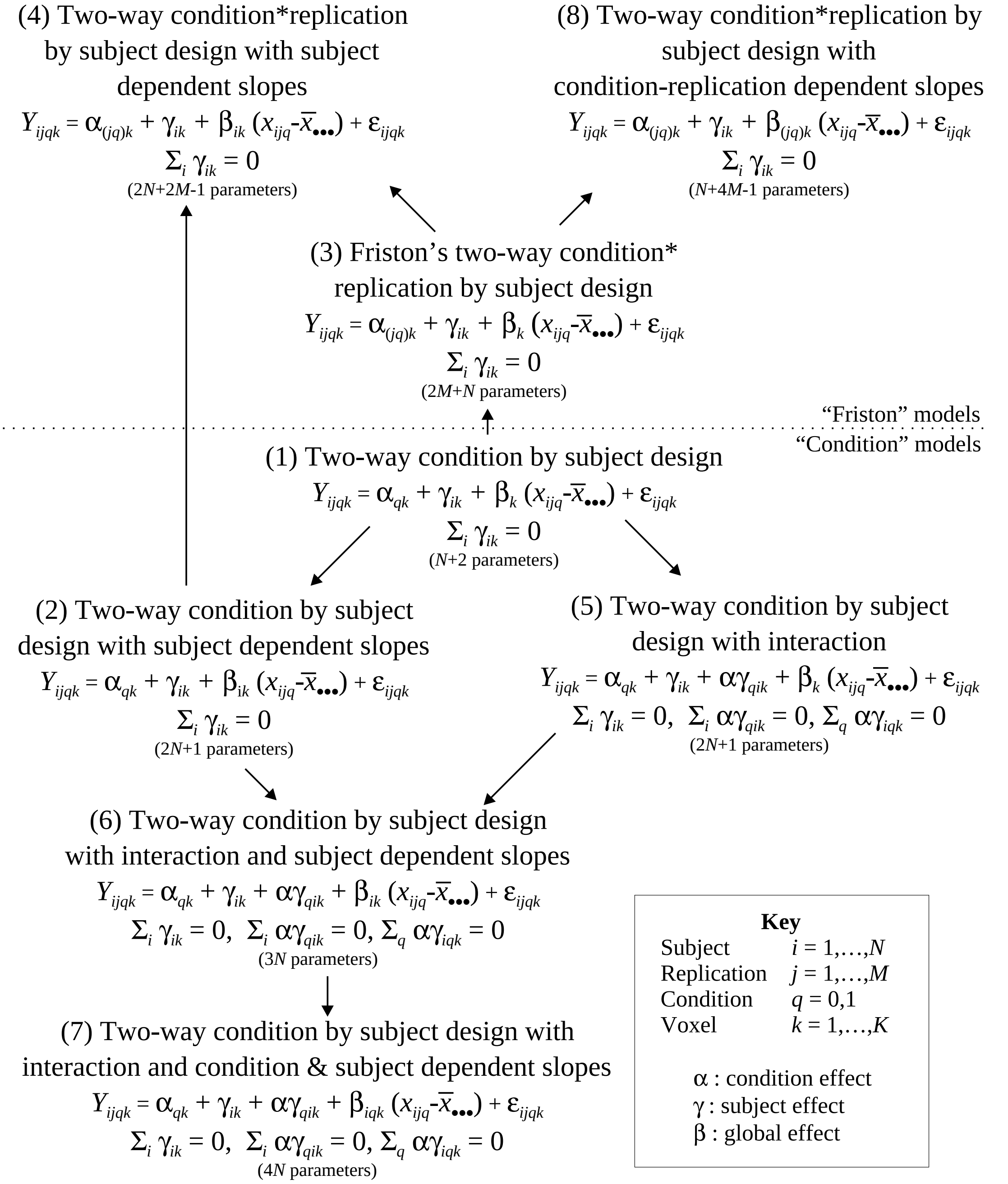
Drawbacks: Low degrees of freedom, assumption of proportionality

Most activation studies only have six to twelve subjects, leaving the proposed *t*-statistic with very few degrees of freedom. As we shall see (§3.3.6.5.), the random field methods for assessing the significance of statistic images don't work well for *t*-statistic images with low degrees of freedom. Considering the equivalent *F*-statistic for the mixed-effects ANOVA model (eqn.25), greater degrees of freedom may be acquired by dropping effects for which there is little evidence.

For the single subject experiment (§2.2.3.2., p60), we saw that the assumption of proportionality inherent in the use of normalised images was most likely false for true rCBF data, and that the use of normalised images and two sample *t*-statistics to assess activation may not be as sensitive as an ANCOVA approach. This would suggest that (at least for true rCBF data) that an ANCOVA approach could be more powerful than an ANOVA on normalised images. We now turn our attention to ANCOVA models for multiple subject activation experiments.

2.3.2. ANCOVA models

2.3.2.1. Models



(4) Two-way condition*replication by subject design with subject dependent slopes

$$Y_{ijqk} = \alpha_{(jq)k} + \gamma_{ik} + \beta_{ik}\,(x_{ijq} - \bar{x}_{\bullet\bullet\bullet}) + \varepsilon_{ijqk}$$

$$\Sigma_i\, \gamma_{ik} = 0$$

(2*N*+2*M*-1 parameters)

(8) Two-way condition*replication by subject design with condition-replication dependent slopes

$$Y_{ijqk} = \alpha_{(jq)k} + \gamma_{ik} + \beta_{(jq)k}\,(x_{ijq} - \bar{x}_{\bullet\bullet\bullet}) + \varepsilon_{ijqk}$$

$$\Sigma_i\, \gamma_{ik} = 0$$

(*N*+4*M*-1 parameters)

(3) Friston's two-way condition* replication by subject design

$$Y_{ijqk} = \alpha_{(jq)k} + \gamma_{ik} + \beta_{k}\,(x_{ijq} - \bar{x}_{\bullet\bullet\bullet}) + \varepsilon_{ijqk}$$

$$\Sigma_i\, \gamma_{ik} = 0$$

(2*M*+*N* parameters)

"Friston" models

"Condition" models

(1) Two-way condition by subject design

$$Y_{ijqk} = \alpha_{qk} + \gamma_{ik} + \beta_{k}\,(x_{ijq} - \bar{x}_{\bullet\bullet\bullet}) + \varepsilon_{ijqk}$$

$$\Sigma_i\, \gamma_{ik} = 0$$

(*N*+2 parameters)

(2) Two-way condition by subject design with subject dependent slopes

$$Y_{ijqk} = \alpha_{qk} + \gamma_{ik} + \beta_{ik}\,(x_{ijq} - \bar{x}_{\bullet\bullet\bullet}) + \varepsilon_{ijqk}$$

$$\Sigma_i\, \gamma_{ik} = 0$$

(2*N*+1 parameters)

(5) Two-way condition by subject design with interaction

$$Y_{ijqk} = \alpha_{qk} + \gamma_{ik} + \alpha\gamma_{qik} + \beta_{k}\,(x_{ijq} - \bar{x}_{\bullet\bullet\bullet}) + \varepsilon_{ijqk}$$

$$\Sigma_i\, \gamma_{ik} = 0,\ \ \Sigma_i\, \alpha\gamma_{qik} = 0,\ \Sigma_q\, \alpha\gamma_{iqk} = 0$$

(2*N*+1 parameters)

(6) Two-way condition by subject design with interaction and subject dependent slopes

$$Y_{ijqk} = \alpha_{qk} + \gamma_{ik} + \alpha\gamma_{qik} + \beta_{ik}\,(x_{ijq} - \bar{x}_{\bullet\bullet\bullet}) + \varepsilon_{ijqk}$$

$$\Sigma_i\, \gamma_{ik} = 0,\ \ \Sigma_i\, \alpha\gamma_{qik} = 0,\ \Sigma_q\, \alpha\gamma_{iqk} = 0$$

(3*N* parameters)

(7) Two-way condition by subject design with interaction and condition & subject dependent slopes

$$Y_{ijqk} = \alpha_{qk} + \gamma_{ik} + \alpha\gamma_{qik} + \beta_{iqk}\,(x_{ijq} - \bar{x}_{\bullet\bullet\bullet}) + \varepsilon_{ijqk}$$

$$\Sigma_i\, \gamma_{ik} = 0,\ \ \Sigma_i\, \alpha\gamma_{qik} = 0,\ \Sigma_q\, \alpha\gamma_{iqk} = 0$$

(4*N* parameters)

Key

Subject	$i = 1,\ldots,N$
Replication	$j = 1,\ldots,M$
Condition	$q = 0,1$
Voxel	$k = 1,\ldots,K$

α : condition effect
γ : subject effect
β : global effect

Equations (26)

Recall that *j* (= 1,…,*M*) indexes the replication for scans acquired under condition *q* (= 0,1), on subject *i* (= 1,…,*N*), and that *k* (= 1,…,*K*) indexes the voxels.

In the taxonomy the parenthesised subscripts $\alpha_{(jq)k}$ (in models 26.3, 26.4 & 26.8) indicate that replication (*j*) and condition (*q*) are to be considered in combination as a single factor. For each voxel *k*, there is a separate main effect, α, for each combination of condition (q) and replication (j). This arrangement was proposed by Friston *et al.* (1990, 1994b), and we shall refer to models 26.3, 26.4 & 26.8 as “Friston” models.

Arrows indicate logical extensions of models. Clearly the parameters may be different in each model, though the same symbols have been used. In all the models it is assumed that $\varepsilon_{jqk} \overset{\text{iid}}{\approx} N(0,\sigma_k^2)$, where σ_k is unique for each model. Appropriate sum-to-zero constraints have been suggested where necessary, to give unique parameter estimates for fixed effects models. For consideration as mixed effects models it is usual to either omit constraints altogether, or to only constrain random effects to sum to zero over the levels of the fixed effects.

For models with condition dependent slope (26.7 & 26.8) the effect of activation depends on the value of the covariate gCBF (gA), but is tested at $\bar{x}_{\bullet\bullet\bullet}$. The exact form of this dependency must be examined to ascertain whether an effect is meaningful. (In its simplest guise this is the “non-parallel lines” ANCOVA problem.)

Random effects for population inference

If these models are treated as fixed effects models, then we can only assume that the scans are drawn from the “population” of all (hypothetical) realisations of scans of these subjects, under identical conditions, and inference extends only to the current study group, under identical conditions. This inference is of limited value in models with subject dependent slopes or subject by condition interaction: For example, in models 26.5 & 26.6, a significant positive contrast of the condition effects (significant evidence against H_k:α_{1k} -α_{0k}>0 for some voxel(s) k), indicates only that there is an evidence of an average activation over all the subjects.

If the subjects in a study are randomly drawn from some population about which it is desired to infer, then subject effects (and hence any subject by condition interaction effects) should be considered as random, giving mixed effects models. Models 26.6 & 26.7 have subject dependent regression coefficients, which then should also be considered random.

2.3.2.2. Model selection

Fixed effects for model selection

For the purpose of model selection, we propose that all effects be considered as fixed. In model selection we are seeking a model for the subjects at hand. If there is evidence of an effect *for these subjects*, that is, as a fixed effect, then such an effect should be considered. When making inference about the population from which our subjects were drawn, *then* the appropriate effects should be considered as random, in the model previously chosen.

This fixed effects approach for model selection gives greater degrees of freedom for testing the presence of certain effects that would be available were they considered random from the outset, giving more powerful tests which are more straightforward to perform. This is in line with the proposed model selection policy of considering the richest model necessary, including effects if there is any evidence for them at all, anywhere.[26]

[26] Recall §2.2.3.1. “Model selection for images”, p.59.

2.3.2.3. Model selection for "V5" study

Omnibus p-values for comparisons of models

Omnibus p-values for pairwise comparisons of the models (eqns.26) for the "V5" study data are given in table 27. As in the single subject case, these are omnibus p-values for the hypotheses H_k that all additional parameters in the richer model are redundant, for all (intracerebral) voxels in the AC-PC plane, $k \in W_{P8}$. For instance, the row labelled "26.3 vs. 26.1" contains the p-value for the omnibus hypothesis $H_{W_{P8}}$, where $H_k{:}\alpha_{(1q)k}=\ldots=\alpha_{(Mq)k}$, q=0,1. F-Statistic images for the hypotheses were computed voxel-by-voxel, "transformed" to a standard Gaussian statistic image, and p-values obtained by Worsley's exceedence proportion test with a threshold of $-\Phi^{-1}(0.01)$. This is the same procedure as used for the single subject model selection, where details were given (text preceding table 25, §2.2.3.2., p.60).

Comparison	p-value
26.2 vs. 26.1	**0.0000**
26.7 vs. 26.6	0.4404
26.6 vs. 26.5	**0.0000**
26.6 vs. 26.2	**0.0000**
26.4 vs. 26.3	**0.0000**
26.4 vs. 26.2	0.2988
26.3 vs. 26.1	**0.0000**
26.5 vs. 26.1	**0.0000**
26.8 vs. 26.3	0.0916

Table (27)
p-values (to 4dp) for the omnibus hypotheses (over the intracerebral voxels in the AC-PC plane) comparing possible models for multiple subject simple activation experiments (eqns.26) on the "V5" study data. p-values less than $\alpha = 0.05$ have been emphasised in bold face.

Leaving the "Friston" models aside for the moment, the proposed backwards selection method (§2.2.3.1., p.59) starts with the richest model in the proposed taxonomy, the two-way condition by subject design with interaction and subject & condition dependent slopes, model 26.7, which we consider as a fixed effects model. Since there is insufficient evidence against an omnibus hypothesis of constant slope across conditions ("26.7 vs. 26.6" in table 27, H_k: $\beta_{i0k} = \beta_{i1k}\ \forall\ k \in W_{P8}$), we accept this hypothesis (for the subjects at hand), and consider model 26.6.

Considering this two-way condition by subject design with interaction and subject dependent slopes as a fixed effects model, we have significant evidence against the omnibus hypotheses of no interaction ("26.6 vs. 26.2" in table 27, H_k: $\alpha\gamma_{qik} = 0\ \forall\ k \in W_{P8}$), and also have significant evidence against the omnibus hypothesis of constant slope ($H_k{:}\beta_{ik}=\beta_k\ \forall\ k \in W_{P8}$).

Our selection procedure stops here, at model 26.6. It is comforting to note that comparisons "26.2 vs. 26.1" and "26.5 vs. 26.1", with voxel hypotheses $H_k{:}\beta_{ik} = \beta_k\ \forall i$ and $H_k{:}\alpha\gamma_{qik} = 0\ \forall q,i$ respectively, also give significant evidence against the respective omnibus hypotheses. This again indicates the presence of a subject dependent regression parameter and a subject by condition interaction, respectively.

Selected model for "V5" study

Although models have been compared only over the AC-PC plane (for computational reasons), there is little to suggest that comparisons over the whole intracerebral volume would give substantially different results. Thus, for these data, with subjects considered as sampled from a population, the appropriate model would be model 26.6, with the subject effect (and hence the subject by condition interaction and the slope parameters) considered as random.

This model is perhaps to be expected. It seems unlikely that different subjects, under the same conditions, will exhibit the same relationship between regional and global values. Hence we have a subject effect and subject dependent regression parameter for the global flow (activity). Similarly, it seems unlikely that different subjects respond to a stimulus with the same increase in rCBF (rA) (after global changes have been accounted for), hence a subject by condition interaction. Model 26.6 would therefore appear to be the minimal justifiable model.

It is interesting to note that model 26.6 reduces to a two-way mixed effects model on omission of the regression terms (β_{ik}). This is precisely the model of the simple *t*-statistic on subject difference images (eqn.25), which is applied to globally normalised rCBF (rA) data.

In conclusion, model 26.6 is probably the most suitable of the proposed models for most multiple subject simple activation data sets. Unfortunately, as a mixed effects model it is rather complicated, and the presence of a main effect is difficult to test. The model appears to fit into the Multilevel Modelling framework of Goldstein (1986). This observation appears to create as many problems as it solves, and we shall not adopt this line of investigation.

The ANCOVA model in widespread use is that proposed by Friston *et al*. (1990), which we now consider.

2.3.2.4. SPM and Friston's ANCOVA

Friston's ANCOVA

Friston *et al*. (1990) proposed modelling activation studies with model 26.3, a two-way ANCOVA with subject as blocking factor, and a combination of the condition and replication as the main treatment factor. There is a separate main effect α for each combination of condition (q) and replication (j) at each voxel (k). If all subjects are presented with conditions in the same sequence, then the condition & replication factor is equivalent to one indicating the sequential number of the scan within the session. Alternatively, scans can be re-ordered to a common presentation order. We shall refer to this condition & replication factor as the condition*replication factor.

With this model the hypothesis of no activation at voxel k is then expressed as a contrast of the condition*replication effects: H_k: $\overline{\alpha}_{(\bullet 1)k} - \overline{\alpha}_{(\bullet 0)k} = 0$, where $\overline{\alpha}_{(\bullet q)k}$ is the mean of the condition*replication effects for scans acquired under condition q.

SPM

This model deserves special attention because of its widespread use. Friston *et al*. (1991b) developed a software package for the analysis of functional PET data. This "*Statistical Parametric Mapping*" (SPM) package was (and still is) the only complete package for this type of work, and a large number of sites acquired the package for routine use. The method implemented for creating statistic images for multiple subject activation studies was the *t*-statistic for the above contrast of condition*replication effects in model 26.3. To most PET practitioners, this is *the* ANCOVA.[27]

History

Originally, not appreciating the regression approach (using matrix methods), Friston *et al.* considered only models for which computations via sums of squares were readily available, and in this respect appear to have been limited to the models covered in their primary reference for ANCOVA, that of Wildt & Ahtola (1978). This elementary text covers inference for one-way (completely randomised) designs, two-way (completely randomised block) designs without interaction for data without replication within each cell (including a test of homogeneity of the regression parameter), and, briefly, a two-way (factorial) design with interaction and constant regression parameter.

The omission, of Wildt & Ahtola, to consider replications within each treatment/block combination perhaps explains why Friston *et al.* arranged their model accordingly, considering condition and replication jointly as the treatment factor. This arrangement has its advantages, as pointed out by Friston (1994b). In particular, it provides a general model that allows analysis of various experiments with more than two conditions, permits post-hoc (one-sided) testing of interactions between conditions, or of time period/activation interactions via appropriate contrasts of the condition*replication effects.

Discussion

Friston's proposed model (26.3) is the largest model in the taxonomy for which there is no difference in the analyses under fixed and random assumptions on the subject effects. But is the model big enough?

Firstly, model (26.3) assumes a constant regression parameter across conditions and subjects. As we have seen, homogeneity of regression parameter across conditions within subjects appears reasonable, but not across subjects. Consider subject dependent slopes for model 26.3, giving model 26.4. For voxel hypotheses H_k: $\beta_{1k}=\ldots=\beta_{Nk}$ the *p*-value for the omnibus hypothesis (over all the intracerebral voxels in the AC-PC plane), assuming fixed effects, is 0.0000 (to 4dp), significant evidence against homogeneity of regression (table 27, comparison "26.4 vs. 26.3").[28] The assumption of constant condition*replication effect across subjects would also seem questionable, for the same reasons that constant condition effect was questionable for the "condition" models, a hypothesis we rejected for the "V5" study.

Thus, from a modelling point of view, Friston's model (26.3) is inadequate for the "V5" study, and perhaps for simple activation studies in general. Assuming constant regression parameter and condition*replication effect substantially increases the degrees of freedom available to test the significance of a condition effect, over what would be available under a more appropriate mixed effects model (26.6), and quite possibly leads to over-sensitive tests, a point noted by Ford (1994).

In the models considered when selecting an ANCOVA for group data the condition effects were assumed to be constant across replications. In Friston's model they are not.

[27] The SPM package has recently been re-written, released in November 1994 as SPM94. In this version any fixed effects model can be analysed. However, model 26.3 is still recommended (Friston *et al.*, 1994b).

[28] Friston *et al.* did not report such a test. Rather, they tested the homogeneity of regression across the condition*replication combination factor. Using data from four subjects, each of whom were scanned twice under each of three conditions, they found little evidence against the homogeneity hypothesis (using their omnibus test). This is perhaps not surprising considering the size of the data set and that between subject variation dominates any within subject variation of regression parameter. Similar results are obtained for the "V5" study (table 27, comparison "26.8 vs. 26.3", omnibus *p*=0.0916).

To assess the importance of this, consider the voxel hypotheses $H_k{:}\alpha_{(1q)k}=\ldots=\alpha_{(Mq)k}$, $q = 0,1$. For the "V5" study we find that there is evidence against this hypothesis in Friston's model (eqn.26.3, omnibus $p = 0.0000$ (4dp), comparison "26.3 vs. 26.1" in table 27), but not when subject dependent slopes are considered (omnibus $p = 0.2988$ to 4dp in table 27, comparison "26.4 vs. 26.2"). We conclude that, for these data, there is insufficient evidence against homogeneity of condition effect across replications when an appropriate model is used.

2.3.3. Conclusions

An approach to global normalisation has to be chosen, and a model selected for inference. In practice this has resulted in two methods being adopted almost exclusively, namely the *t*-statistic on subject difference images (after normalisation for global changes by proportional scaling), and Friston's AnCova.

2.3.3.1. t-statistic on subject difference images

Advantages

The *t*-statistic is attractive for routine use because of its robustness and simplicity. The assumptions required for its use are minimal and easy to verify. The statistic is easy to compute. The formulation of the *t*-statistic in terms of subject difference images makes the statistic accessible, and easy to visualise.

This simplicity hides a rather complicated model, a two-way mixed effects Anova (eqn.25). In the "simple" *t*-statistic, subject, block and habituation effects cancel out, as do (local) linear trend effects. (The latter provided they are constant across subjects, and that subjects have been allocated to condition presentation order in a balanced fashion.) Further, random subject by condition interaction is incorporated.

Disadvantages

The criticisms of the *t*-statistic are the insensitivity of the ensuing tests (usually in comparison to Friston's AnCova model, eqn.26.6), and the assumption of proportionality implicit in global scaling for the normalisation of global effects.

Considering the ensuing tests, these are insensitive because of the low degrees of freedom available. As we shall see in chapter 3 (§3.3.6.5.), methods for testing *t*-statistic images using results for continuous random fields are conservative for *t*-statistic images with low degrees of freedom. In itself, this increases confidence in any significant results. However, there are other techniques available for testing statistic images with low degrees of freedom that are sensitive, namely variance smoothing and the non-parametric approach of chapter 6, or the use of "secondary smoothing" which we shall return to in chapter 3 (§3.3.6.6.).

Turning to the assumption of proportionality. For true rCBF data the evidence suggests that the relationship between regional and global flow over the normal range of gCBF (across scans on the same individual under the same conditions) is not proportional. For uncalibrated "counts" data the case is less clear cut, and depends on the variability of the introduced activity. A proportional model appears to be acceptable for the majority of the subjects in the "V5" experiment analysed here. Whether this is true for other "counts" data sets remains to be seen. Nonetheless, the real question is whether departures from proportionality compromise the validity of tests based on the paired *t*-statistic. The opinion of this author is that the opposite is true, namely that departures from proportionality merely add to the error variance (in a random manner), and thereby decrease the power of the approach.

In summary, the *t*-statistic, for subject difference images, leads to a robust test, albeit a relatively insensitive one. Assurance can be placed in the results of a such an analysis, even if the initial assumptions are not checked.

2.3.3.2. Friston's AnCova

More accurate modelling of the relationship between regional and global values is the only motivation for considering an AnCova approach. However, for the "V5" study, model selection for the AnCova models leads to a model (26.6) that is complicated and difficult to apply to data. The AnCova model in widespread use is that proposed by Friston *et al*. (1990), a two-way condition*replication by subject design, with constant regression slope.

Validity?

It has been shown that, from a modelling point of view, this model is inappropriate for the "V5" study, and its inappropriateness for other similar studies conjectured at. However, the key issue is how serious the routine use of this model is in terms of false positives. The degrees of freedom available for testing in Friston's AnCova are substantially greater than those available within a more appropriate mixed effects model, such as 26.6, but the effects omitted in Friston's model inflate the variance term. The validity of tests for activation based on this model depends on the actual magnitude and structure of the omitted effects. In the presence of such effects, all that can be said is that the assumptions of the model are not true, and therefore that the validity of ensuing tests cannot be guaranteed.

Empirical validation

It would be interesting to compare the results of analyses using Friston's model, and models including subject effects, subject by condition interactions and subject dependent slopes. This would give some insight into the importance of such terms, and the consequences of their omission.

Ideally, one would like to apply the various models to a number of null data sets, where the "rest" and "active" experimental conditions are identical.

This has been done with Friston's model in a few cases, together with Friston's "Bonferroni" method for testing statistic images (Friston *et al.*, 1991d)[29]. Only a few false positives have been found, indicating that the method leads to valid tests. Also, for "gold standard" activation experiments where the actual activation site is known from previous work, the activated areas are correctly identified, with only a few isolated false positives. For many this is sufficient empirical validation to justify the approach. However, methods of empirical validation such as these only address the issue of validity for the experimental paradigm at hand. The subjects must be assumed to be representative of the population about which it is desired to infer.

It would be interesting to continue these investigations further.

[29] Friston's "Bonferroni" method for testing statistic images is discussed in §3.3.2.

2.4. Additional Comments

Inter-group comparisons

The single subject models (eqns.19, p58) are also used to assess changes in rCBF patterns between groups, where each subject in each group is scanned once. Here there is no notion of a block (a time pairing) and attention is restricted to models 19.1, 19.3 & 19.5; where *q* indexes the group, and *j* the subject within each group.

In this situation the inter-subject variation in gCBF is likely to be great. For "counts" images of rA, variations in gA due to differences in gCBF may swamp those due to changes in the administered dose.

In addition, there may be many physiological or neurological reasons to suggest that the sensitivity of regional to global flow differ between the two groups for some regions of the brain, particularly if the groups are distinguished by some physiological or mental trait, as is usually the case. That is, the assumption of constant regression coefficient in the one way ANCOVA model (eqn.19.3) may be inappropriate, in which case 19.5 should be considered, and the presence of a condition effect interpreted with caution. Also of concern is the possibility that mean gCBF differs between the groups.

Condition dependent gCBF

In allowing for changes in gCBF (gA) (either by proportional scaling or ANCOVA) when assessing condition specific changes in rCBF (rA), it should be borne in mind that changes in gCBF (gA) across conditions can adversely affect the analysis.

Since gCBF is calculated as the average rCBF across the intracerebral voxels, an increase in rCBF in a particular brain region must cause an increase in gCBF, unless there is a corresponding decrease in rCBF elsewhere in the brain. There are physiological and neurological theories for such a corresponding decrease, explaining how increased blood flow and/or neuronal activity in one region of the brain can inhibit flow and/or activity in another. De-activation observed in the presence of activation is taken by some to represent some form of functional connectivity, the increased neuronal activity in one area inhibiting activity in the other. This view is not universally held.

If gCBF varies with the condition then care must be taken. If gCBF is increased by a large activation that doesn't have a corresponding de-activation, comparison at a common gCBF (gA) will make non-activated regions of the brain (whose rCBF (rA) remained constant) falsely appear as de-activated, and the magnitude of the activation will similarly be reduced. In these circumstances a better measure of background change should be sought. Such an estimate can be obtained by examining the flow (activity) in brain regions known to be unaffected by the stimulus. If such unaffected regions cannot be specified, then another possibility would be to fit a background global value to each rCBF (rA) scan. This could be achieved using the stochastic sign change criteria, the background gCBF (gA) estimated as the threshold level which is crossed most frequently by the rCBF (rA) image.

If gCBF varies considerably between conditions, as in pharmacological activation experiments, then testing for a condition effect after allowing for global changes compares the two brain states at a gCBF (gA) unattainable in at least one of the brain states. This involves extrapolating the relationship between regional and global values outside the range of the observed data, an extrapolation which might not be valid.

As a precaution, it is usual to test for changes in gCBF across conditions. In the "V5" study, there is insufficient evidence of a change in gA across conditions.[30] John Watson, the primary researcher in the "V5" study, considers the depressed background of the *t*-statistic image for the "V5" study (See §2.6.1.) to indicate a true decrease of rCBF induced by the large increases in the visual cortex (personal communication).

Outliers due to "dose ranging" in non-calibrated studies

In many studies arterial sampling is not carried out, and "counts" images of relative activity rA are obtained as indicating rCBF. Here, *dose ranging*[31] can lead to anomalous gA for some scans, outliers adversely affecting the AnCova through large leverage. To avoid this, images are sometimes scaled before AnCova according to the amount of tracer introduced in a scan. If z_{ijq} is the measured amount of activity introduced into subject i during scan j under condition q, then the adjusted images and global activities are given by

$$Y'_{ijqk} = \frac{Y_{ijqk}}{\left(z_{ijq}/\bar{z}_{i\bullet\bullet}\right)}, \qquad x'_{ijq} = \frac{x_{ijq}}{\left(z_{ijq}/\bar{z}_{i\bullet\bullet}\right)}$$

This clearly alters the variance assumptions. An alternative approach would be to weight observations according to the administered dose. Outlying gA have no effect on the proportional scaling approach, since the first step is to divide by the measured global activity.

[30] Consider initially the model $x_{ijq} = \alpha_q + \gamma_i + \alpha\gamma_{iq} + \varepsilon_{ijq}$ where $\varepsilon_{ijq} \overset{\text{iid}}{\sim} N(o,\sigma^2)$. This model is chosen since it is analogous to the proposed AnCova design. Since the p-value for non-zero interaction terms is 0.977 (this test is valid whether subject effects are considered random or not), we assume no interaction and consider the model $x_{ijq} = \alpha_q + \gamma_i + \varepsilon_{ijq}$. The p-value for $H\alpha_0 = \alpha_1$ is 0.106 (again, this is so whether or not subject effects are considered random, see Miller, 1980, §4.5). Thus there is insufficient evidence against H.

[31] Recall §2.1. (p52) for details.

2.5. A Multivariate Perspective

Abstract

In our discussion of linear models relating rCBF to design factors and gCBF, we have not considered the computational problems of simultaneously fitting the model for thousands of voxels.[32] Viewing the problem from a multivariate perspective provides efficient computation, and offers some insight into the problem. In this section we demonstrate the relationship between the simultaneous general linear models for PET data (which we term *image regression*), and multivariate regression. Some basic multivariate regression theory is reviewed. Under an assumption of multivariate normality for the rCBF images the PET scenario *is* a multivariate regression, but the high dimensionality of the data precludes any multivariate analysis. It is for this reason that we concentrate on simultaneous univariate tests.

Whilst there is nothing new in this section for the statistical reader, the advantages of the multivariate perspective are only just dawning on the PET community, and this section is included for the benefit of the reader in the latter category.

2.5.1. Sums of squares approaches

We have a single model to be fitted at each voxel. In a typical data set there are 77000 voxels, making individual fitting on a voxel-by-voxel basis using statistical packages prohibitive. Working with the images as row vectors, and using matrix manipulation routines, the appropriate sums of squares can be computed for all voxels simultaneously. This is the approach taken in older versions of the SPM software (Friston *et al.*, 1991a), using the exposition of AnCova provided by Wildt & Atholla (1978).

This approach is rather inelegant, being rather slow and requiring a purpose written program for each possible design. Clearly direct fitting of general linear models for images is possible by viewing the problem from a multivariate perspective and using matrix methods. This point was noted by the author, passed on to Friston *et al.* (1994b), and is implemented in SPM94.

[32] A substantial (and unseen) part of the work undertaken during this Ph.D. has been the development of software for interactive analysis of PET image data. A number of the author's routines are part of the SPM software.

2.5.2. Multivariate regression formulation

Consider the general linear model for the data at voxel k:

$$Y_{jk} = x_{j1}\,\beta_{1k} + \ldots + x_{jQ}\,\beta_{Qk} + \varepsilon_{jk} \quad \text{where } \varepsilon_{jk} \overset{\text{iid}}{\sim} N(0,\sigma_k^2) \tag{27}$$

Where Y_{jk} denotes the rCBF (rA) measurement at voxel k =1,…,K; of scan j =1,…,N; and let x_{jq} q = 1,…,Q be a set of Q explanatory variables for scan j, either covariates (such as gCBF), dummy variables indicating levels of a factor, or a combination (for interactions or for covariates with effect dependent on the level of a factor).

In matrix form the model is:

$$\begin{array}{ccccccc}
\begin{pmatrix} Y_{1k} \\ \vdots \\ Y_{Nk} \end{pmatrix} & = & \begin{pmatrix} x_{11} & \cdots & x_{1Q} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NQ} \end{pmatrix} & \begin{pmatrix} \beta_{1k} \\ \vdots \\ \beta_{Qk} \end{pmatrix} & + & \begin{pmatrix} \varepsilon_{1k} \\ \vdots \\ \varepsilon_{Nk} \end{pmatrix} \\
\boldsymbol{Y}^k & = & \boldsymbol{X} & \boldsymbol{\beta}^k & + & \boldsymbol{\varepsilon}^k
\end{array}$$

Since the design matrix, $\boldsymbol{X}$, is the same for every voxel, we can write all the voxel models simultaneously in a multivariate linear model:

$$\left(\boldsymbol{Y}^1 \mid \cdots \mid \boldsymbol{Y}^K\right) = \begin{pmatrix} x_{11} & \cdots & x_{1Q} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NQ} \end{pmatrix} \left(\boldsymbol{\beta}^1 \mid \cdots \mid \boldsymbol{\beta}^K\right) + \left(\boldsymbol{\varepsilon}^1 \mid \cdots \mid \boldsymbol{\varepsilon}^K\right)$$

$$\begin{array}{ccccc}
\begin{pmatrix} Y_{11} & \cdots & Y_{1K} \\ \vdots & \ddots & \vdots \\ Y_{N1} & \cdots & Y_{NK} \end{pmatrix} & = \begin{pmatrix} x_{11} & \cdots & x_{1Q} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NQ} \end{pmatrix} & \begin{pmatrix} \beta_{11} & \cdots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{Q1} & \cdots & \beta_{QK} \end{pmatrix} & + & \begin{pmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1K} \\ \vdots & \ddots & \vdots \\ \varepsilon_{N1} & \cdots & \varepsilon_{NK} \end{pmatrix} \\
\boldsymbol{Y} & = \quad \boldsymbol{X} & \boldsymbol{\beta} & + & \boldsymbol{\varepsilon}
\end{array}$$

Let $\boldsymbol{Y}_j = (Y_{j1},\ldots,Y_{jK})$; $\boldsymbol{\beta}_q = (\beta_{q1},\ldots,\beta_{qK})$ & $\boldsymbol{\varepsilon}_j = (\varepsilon_{j1},\ldots,\varepsilon_{jK})$ be the rCBF images, the images of the coefficients, and images of the errors respectively, all as row vectors. Then the matrices $\boldsymbol{Y}$ and $\boldsymbol{\beta}$ are stacks of the images, and stacks of the coefficient images respectively:

$$\underbrace{\boldsymbol{Y}}_{N\times K} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \hline \vdots \\ \hline \boldsymbol{Y}_N \end{pmatrix} \qquad \underbrace{\boldsymbol{\beta}}_{Q\times K} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \hline \vdots \\ \hline \boldsymbol{\beta}_Q \end{pmatrix} \qquad \underbrace{\boldsymbol{\varepsilon}}_{N\times K} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \hline \vdots \\ \hline \boldsymbol{\varepsilon}_N \end{pmatrix}$$

Image regression

In multivariate regression it is usually assumed that the error vectors $\boldsymbol{\varepsilon}_j$ are drawn independently from a multivariate normal distribution with zero mean and variance-covariance matrix Σ. In the present context of simultaneous regressions (eqn.27) we are only assuming that the marginal distributions are normal, $\varepsilon_{jk} \sim N(0,\sigma_k^2)$. It is for this reason that we differentiate *image regression* from standard multivariate regression.

2.5.3. Multivariate regression

Since the difference between image regression and standard multivariate regression lies only in the distributional assumptions, computation of least squares estimates in the two situations is identical.

Least squares estimates

The usual matrix results for univariate regression continue to hold for multivariate regression. (See Krzanowski, 1988, ch.15.) The least squares principle gives normal equations:

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})\,\hat{\boldsymbol{\beta}}$$

If $\boldsymbol{X}$ is of full rank then $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is invertible and the $p \times K$ matrix of parameter estimates $\hat{\boldsymbol{\beta}}$ (each row is the image of a fitted parameter) is given by:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\,\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{28}$$

Since $\boldsymbol{X}$ is only of dimension $N \times Q$ computation of $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ is not prohibitive. For non-unique designs, constraints on the parameters can be imposed to give a design matrix of full rank, leading to a unique least squares estimate, or an algebraic inverse can be used to obtain least squares estimates. See Scheffé (1959). For ease of computation, we shall take the former course of action, and henceforth assume that$\boldsymbol{X}$ has rank Q.

Fitted values, residuals

The $N \times K$ matrix of fitted values $\hat{\boldsymbol{Y}}$ (rows are fitted images) can be obtained as:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\,\hat{\boldsymbol{\beta}}$$

and $N \times K$ matrix of residuals $\hat{\boldsymbol{\varepsilon}}$, estimates of the error matrix $\boldsymbol{\varepsilon}$, as:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$$

2.5.4. Image regression

That the multivariate formulation addresses the simultaneous regressions of image regression can now be readily seen: Partitioning $\hat{\beta}$ and $\boldsymbol{Y}$ in eqn.28 into $\left(\hat{\beta}^1 \mid \cdots \mid \hat{\beta}^K\right)$ and $\left(\boldsymbol{Y}^1 \mid \cdots \mid \boldsymbol{Y}^K\right)$ it is clear that the multivariate approach is simultaneously fitting the K general linear models for each voxel. In particular, $\hat{\beta}^k$ is the least squares estimate of β^k for the regression at voxel k. Since we have assumed (univariate) normality, the univariate theory then gives us that $\hat{\beta}^k$ is also the maximum likelihood estimate of β^k (Scheffé, 1959), with Q-variate normal distribution:

$$\hat{\beta}^k \sim N_Q(\beta, \sigma_k^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})$$

Note that the fitted parameters, considered together over all voxels, do not necessarily have a multivariate normal distribution, since no multivariate assumption is made about the data in the image regression setting.

Distributional results

The image regression formulation allows easy computation of univariate results at all the voxels simultaneously.

The residual sums of squares for the voxels, arranged as an image ($1\times K$ row vector), are given by:

$$\boldsymbol{R} = \mathrm{diag}(\hat{\varepsilon}^{\mathrm{T}}\,\hat{\varepsilon}) \qquad \text{(the diagonal elements of } \hat{\varepsilon}^{\mathrm{T}}\,\hat{\varepsilon}\text{, arranged as a row vector)}$$

(Since the residual matrix $\hat{\varepsilon}$ is of dimension $N\times K$, the residual sums of squares is more efficiently calculated directly, by summing the squares of the elements for each voxel, i.e. summing within columns, the squares of the elements of $\hat{\varepsilon}$).

Similarly, the image of variances, $\boldsymbol{v} = (\sigma_1^2,\ldots,\sigma_K^2)$, is estimated by $\hat{\boldsymbol{v}}$ as:

$$\hat{\boldsymbol{v}} = \frac{\boldsymbol{R}}{N - \mathrm{rank}(\boldsymbol{X})}$$

$$\text{with } \frac{\hat{\boldsymbol{v}}\,(N - \mathrm{rank}(\boldsymbol{X}))}{\boldsymbol{v}} \overset{elementwise}{\sim} \chi^2_{N - \mathrm{rank}(\boldsymbol{X})}$$

Here, the division operator is understood to act element by element on matrices and vectors. The “elementwise” qualification on the tilde indicates that the distributional result is to be interpreted as applying to the individual elements of the vector on the left. So, element k of $\hat{\boldsymbol{v}}$, $\hat{\sigma}_k^2$, is distributed as $\sigma_k^2/(N\text{-rank}(\boldsymbol{X}))$ times a chi-squared variate on N-rank($\boldsymbol{X}$) degrees of freedom.

Consider a fixed effects model, and a uniquely specified design (perhaps by imposed constraints) with Q (= rank($\boldsymbol{X}$)) parameters. Then, for $\boldsymbol{x}$ a (Q×1) vector defining an estimable function $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$:

$$\frac{\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{\hat{\boldsymbol{v}}\, \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}}} \overset{elementwise}{\sim} t_{N-Q}$$

Here the square root operator is understood to act element by element on a matrices. For unique designs, with rank($\boldsymbol{X}$) = Q, all parametric functions are estimable (see Scheffé, 1959, Th.4).

For fixed effects models the familiar "extra sum of squares" likelihood ratio test (Draper and Smith, 1981, §2.7) for general linear hypotheses constraining the parameters β_{qk} , is also easily computed, provided the hypotheses H_k constrain the parameters identically for each voxel k. In this case the model under the null hypotheses is the same for all voxels, and the F-statistics for each voxel are given simultaneously in an F-statistic image $\boldsymbol{F}$ by:

$$\boldsymbol{F} = \frac{(\boldsymbol{R} - \boldsymbol{R}_{\mathrm{H}})\,/\,df}{\hat{\boldsymbol{v}}} \overset{elementwise}{\sim} F_{df,\ \mathrm{rank}(\boldsymbol{X})}$$

where $\boldsymbol{R}_{\mathrm{H}}$ is the image of residual sums of squares under all H_k, and df is the reduction in degrees of freedom imposed by the constraints.

Thus, considering the regression for all voxels together as a multivariate regression enables simultaneous fitting of the models for each voxel, and allows easy computation using a matrix manipulation package on a large computer.[33]

2.5.5. Multivariate regression revisited

We have concentrated on simultaneous univariate methods, and used the multivariate perspective of image regression for computational efficiency. As we shall now demonstrate, a full multivariate analysis is precluded.

Multivariate hypothesis testing precluded

Consider the problem as a multivariate regression. Assume that the error vectors $\boldsymbol{\varepsilon}_j$ are drawn from a K-variate normal distribution with variance-covariance matrix $\boldsymbol{\Sigma}$. Under this assumption, the joint distribution of the fitted parameters $\hat{\beta}_{qk}$ *is* multivariate normal; the expected value of $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{\beta}$, and the covariance between $\hat{\beta}_{qk}$ and $\hat{\beta}_{q'k'}$ is $(\Sigma)_{k\,k'} \times ((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})_{q\,q'}$. The full sum of squares $\boldsymbol{S} = \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\,\hat{\boldsymbol{\varepsilon}}$ [34] leads to maximum likelihood estimate of the variance-covariance matrix $\boldsymbol{\Sigma}$ as $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}/N$, and unbiased estimate $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}/(N\,$-rank$(\mathbf{X}))$.

Since the dimensionality of the data, K, far exceeds the number of replications, N, the estimated variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$ has linearly dependent rows/columns, and is

[33] The current work was undertaken using MatLab (The MathWorks Inc., Natick), a matrix manipulation package with extensive programming and visualisation features. The platform used was a Sun Sparc2, with 48Mb of ram and 160Mb of virtual memory.

[34] $\boldsymbol{S}$ has a Wishart distribution with N -rank($\boldsymbol{X}$) degrees of freedom and parameter Σ, and is independent of $\hat{\boldsymbol{\beta}}$.

therefore singular (see Healy, 1986). This precludes any of the standard multivariate analyses, whose statistics are functions of the eigenvalues of $\hat{\Sigma}$.

For instance, for fixed effects models, the likelihood ratio test for a general linear hypothesis lead to Wilks' Lambda as a test statistic (see Krzanowski, 1988):

$$\Lambda = \left|\hat{\Sigma}\right| \Big/ \left|\hat{\Sigma}_0\right|$$

where $\hat{\Sigma}_0$ is the maximum likelihood estimate of $\mathbf{\Sigma}$ under the null hypothesis. (This is the multivariate analogue of the "extra sum of squares principle") If $\hat{\Sigma}_0$ is singular, then so is $\hat{\Sigma}$, and the statistic is not defined.

2.6. Example–“V5” Study

Presentation of statistic images

Throughout this thesis, statistic images shall be depicted by mesh plots of a single transverse plane, usually the AC-PC plane. This form is chosen in preference to grey-scale images because details in the images are shown more clearly. In particular, rugosities in an image are readily discernible.

In these plots the X-Y plane corresponds to the relevant transverse slice, with scales graduated in millimetres according to the standard Talairach co-ordinate system (to which the images have been aligned). Thus the bottom left of the mesh corresponds to the posterior of the brain, and the top left to the left of the brain. The vertices of the mesh are located above the voxel centres, with heights indicating the value of the statistic at that voxel.

2.6.1. Proportional scaling approach

A proportional scaling approach will be illustrated, using a *t*-statistic formed from subject difference images, as described in §2.3.1.

2.6.1.1. Statistic images

Below are mesh plots of the AC-PC planes of various statistic images for the “V5” data.

Subject difference image

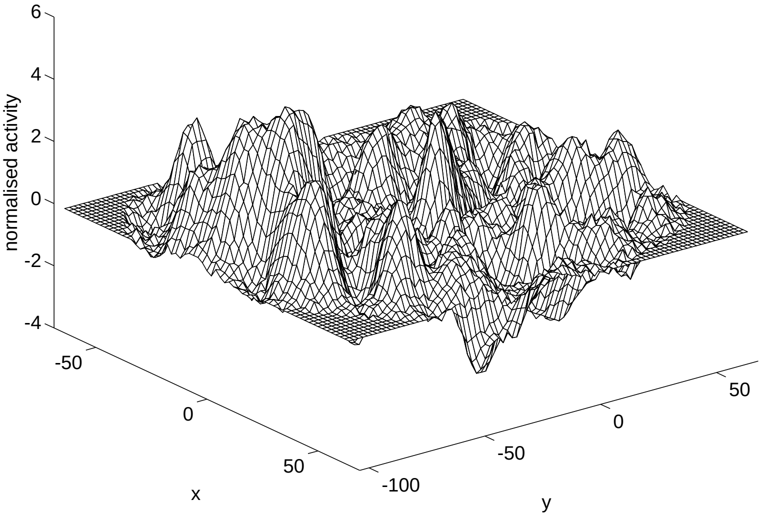


Figure 28
Mesh plot of subject difference image $\mathbf{\Delta}_1$ for first subject in the “V5” study. (Eqn.20) The X-Y plane is the AC-PC plane. The heights of the vertices indicate the value of Δ_{1k}. for the appropriate voxels. The Z axis is graduated normalised counts. The “flat” border at the edges corresponds to voxels outside the intracerebral volume, whose values have been set to zero.

Study mean difference image

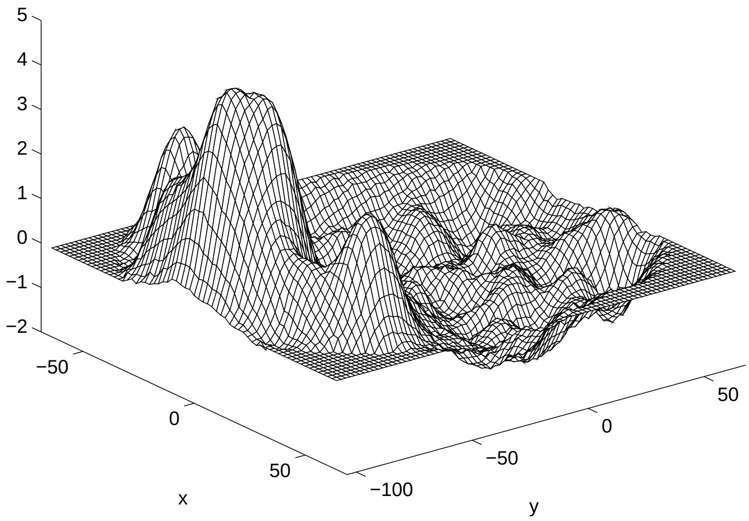


Figure 29
Mesh plot of study mean difference image for the "V5" study. (Eqn.22) The Z axis is graduated normalised counts. The AC-PC plane is shown.

Sample variance image

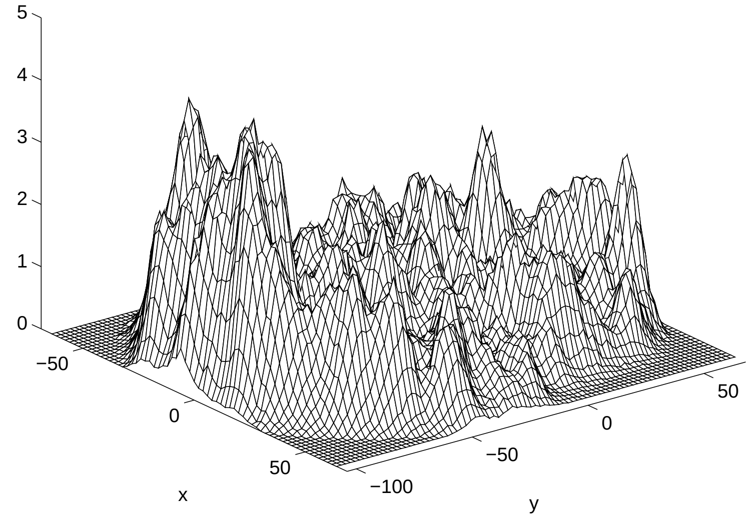


Figure 30
Mesh plot of sample variance at voxel *k*, S_k^2, of the subject mean differences for the "V5" study (eqn.23). The Z axis is graduated normalised counts (squared). The AC-PC plane is shown. Note how the sample variance image is quite noisy, whereas the mean difference image of fig.29 is smooth.

t-statistic image

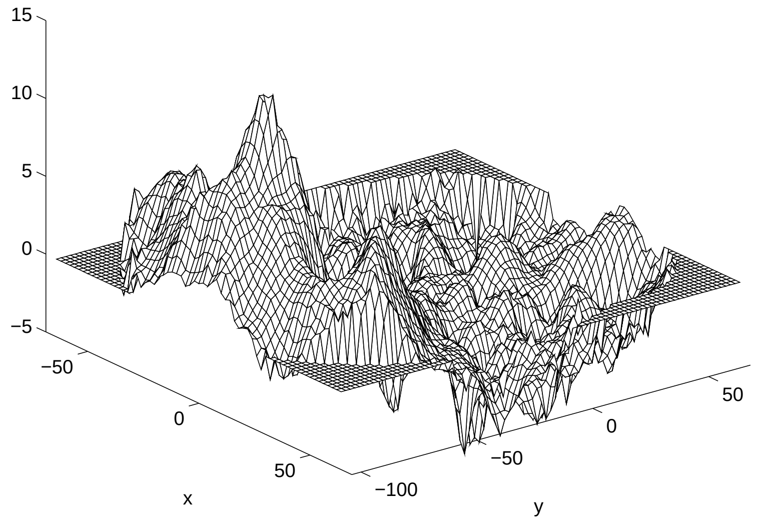


Figure 31
Mesh plot of *t*-statistic image ***T*** for the "V5" study (eqn.21). Each voxel statistic is distributed as a Student's *t* variate with 11 degrees of freedom, under the hypothesis of no activation at that voxel, H_k:μ_k= 0. $\Delta_{ik} \sim N(\mu_k, \sigma_k^2)$ is assumed. The AC-PC plane is shown. The roughness of the statistic image is due to noise in the sample variance image (fig.30).

Unadjusted p-value image

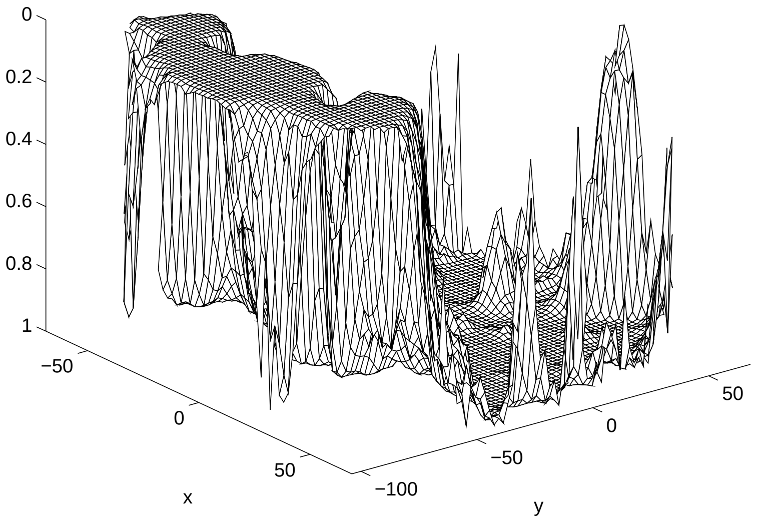


Figure 32
Mesh plot of (unadjusted) one-sided *p*-values for the voxel hypotheses H_k:μ_k= 0 of no activation at voxel *k*, computed from the *t*-statistic image (with 11 degrees of freedom) of fig.31. The *p*-value axis is graduated in reverse to depict activated voxels as high. Voxels outside the intracerebral volume have been removed. The AC-PC plane is shown.

2.6.1.2. A crude Bonferroni analysis

From the *t*-statistic image and the corresponding *p*-value image (figs.31 & 32 respectively), there would appear to be evidence against the hypotheses of no activation for voxels at the rear of the brain. (Voxel hypotheses $H_k{:}\mu_k = 0$, assuming $\Delta_{ik} \sim N(\mu_k,\sigma_k^2)$). The maximum *t*-statistic in the AC-PC plane is 13.692 (to 3dp) at Talairach location (2,-74,0), and the maximum *t*-statistic in the whole brain volume is 20.147 (to 3dp) at (-20,-80,12), with *p*-values of 1.48×10^{-8} and 2.47×10^{-10} (to 3sf) respectively.

A crude Bonferroni assessment

At overall significance level α for all the K voxels, a (highly conservative) Bonferroni[35] correction for the K simultaneous tests would reject H_k if the *p*-value at voxel k, P_k, was less than α/K. This leads to Bonferroni single step adjusted *p*-values of $\tilde{P}_k = \min\{K_W P_k, 1\}$. The Bonferroni approach at level α rejects the null hypotheses for voxels with adjusted *p*-values less than α. Here, the intracerebral volume consists of $K = 77189$ voxels, leading to the adjusted *p*-value image of figure 33a. Thresholding the *p*-value image at $\alpha = 0.05$ at this value reveals that there is evidence of activation at the posterior of the brain (fig.33b).

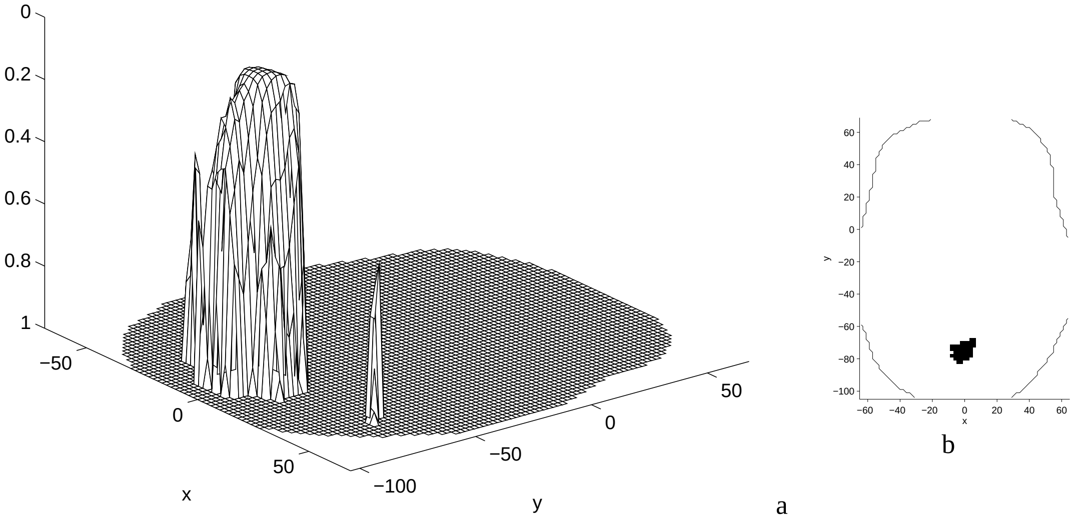


Figure 33
(a) Mesh plot of Bonferroni single step adjusted one-sided *p*-values, computed from the *p*-values of fig.32. Voxels outside the intracerebral volume have been removed. (b) Voxels with adjusted *p*-value below level 0.05. The outline of the intracerebral area is superimposed. The AC-PC plane is shown.

[35] "The" Bonferroni correction for multiple comparisons problems is discussed in §3.2.1.

2.6.1.3. Empirical examination of assumptions

Q-Q plot for single voxel

The analysis of the "V5" study just seen relies on the assumption that the subject difference images have normally distributed voxel values: $\Delta_{ik} \sim N(\mu_k,\sigma_k^2)$. Considering a single voxel, this assumption can be examined by plotting the observed values against expected order statistics (normal scores), in a Q-Q probability plot (fig.34). The approximate linearity of this plot suggests that the assumption is reasonable. Summarising the linearity via the correlation coefficient leads to a simple Shapiro-Wilk type test for normality (Filliben, 1975). The null hypothesis is H_k: $\overline{\Delta}_{ik} \sim N(\mu_k,\sigma_k^2)$. The correlation here is 0.972 (to 3dp), above the critical threshold,[36] giving insufficient evidence against the hypothesis H_k at the 5% level.

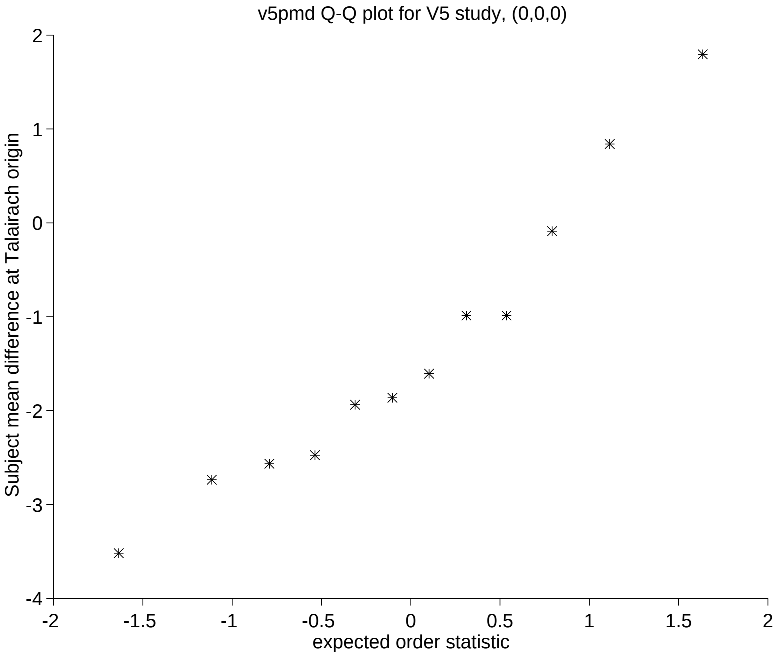


Figure 34

Q-Q plot for the data at the Talairach origin. The values of the subject difference images Δ_{ik} for voxel k at (0,0,0) are plotted against the corresponding expected order statistics (normal scores) from a standard normal distribution.

[36] A high correlation is consistent with normality. The MINITAB reference manual ("arithmetic" section, NSCORES command) gives critical values of 0.9180 and 0.9383 below which the correlation coefficient must fall to suggest evidence against the null hypothesis of normality at the 5% level, for samples of sizes 10 and 15 respectively.

Correlation coefficient of Q-Q plots for AC-PC plane

Summarising the linearity of Q-Q plots using the correlation coefficient allows us to examine the assumption over all other voxels simultaneously. Fig.35 shows the correlation coefficients so computed for the AC-PC plane. Less than 5% of the voxels have correlation coefficient less than the critical value for a sample of size 12 at the 5% level, an exceedence proportion indicating insufficient evidence against the omnibus hypothesis of normality ($\Delta_{ik} \sim N(\mu_k,\sigma_k^2)$) at all voxels. However, it should be borne in mind that these Shapiro-Wilk type tests have extremely low power.

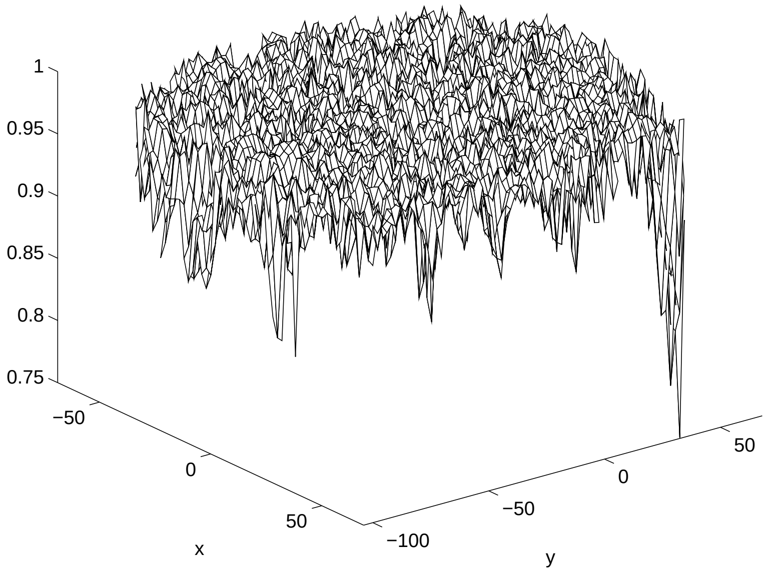


Figure 35

Mesh plot showing the correlation between the data Δ_{ik} (subject difference images) at each voxel k and the expected order statistics from a standard Normal distribution. The AC-PC plane is shown. (Voxels outside the intracerebral volume have been removed.) Note that the Z-axis is truncated at 0.75.

2.6.2. Friston’s ANCOVA

In §2.3.2.4., the two way ANCOVA design (model 26.3) proposed by Friston *et al.*, (1990) was discussed. To illustrate this discussion, consider applying the method to the “V5” study data.

Recall that the hypothesis of no activation at voxel k is then expressed as an appropriate contrast of the condition*replication effects: H_k: $\overline{\alpha}_{(\bullet 1)k}$ - $\overline{\alpha}_{(\bullet 0)k}$ = 0, where $\overline{\alpha}_{(\bullet q)k}$ is the mean of the condition*replication effects for scans acquired under condition q. Fitting the model to the “V5” data and evaluating the contrast for the estimated effects leads a *t*-statistic image with 120 degrees of freedom, the AC-PC plane of which is depicted below (fig.36).

t-Statistic image

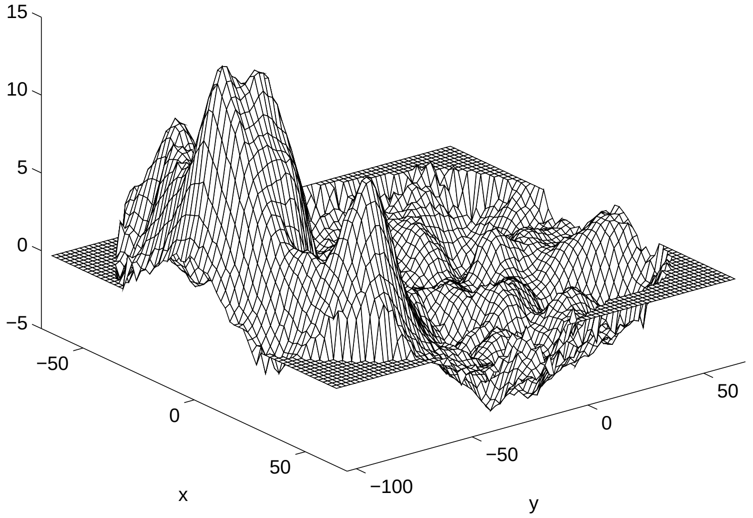


Figure 36

Mesh plot of *t*-statistic for the contrast of condition*replication effects in Friston’s ANCOVA (model 26.3), computed for the “V5” study data. Each voxel statistic is distributed as a Student’s *t* with 120 degrees of freedom, under the hypothesis of no activation at that voxel: H_k: $\overline{\alpha}_{(\bullet 1)k}$ - $\overline{\alpha}_{(\bullet 0)k}$ = 0 The model is assumed to fit. The AC-PC plane is shown.

Unadjusted p-value image

Referring these t-statistics to a Student’s t-distribution with 120 degrees of freedom gives the unadjusted p-value image below (fig.37).

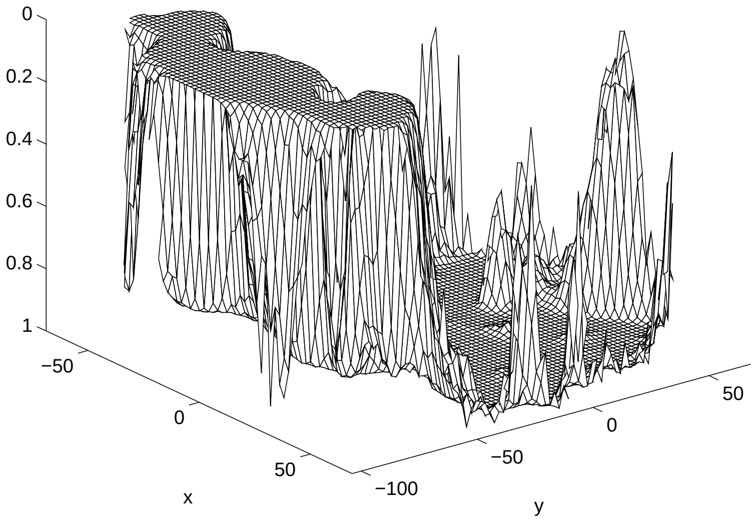


Figure 37
Mesh plot of (unadjusted) one-sided p-values computed from the t-statistic image of fig.36. Voxels outside the intracerebral volume have been removed. The AC-PC plane is shown.

Adjusted p-value image

Adjusting the one-sided p-values for the $K = 77189$ intracerebral voxels leads to Bonferroni single step one-sided adjusted p-values (fig.38a). As can be seen, the significant region for a level $\alpha = 0.05$ test is much larger than for a proportional scaling approach with t-statistic computed from subject difference images (fig.38b, compare with fig.33b, p86)

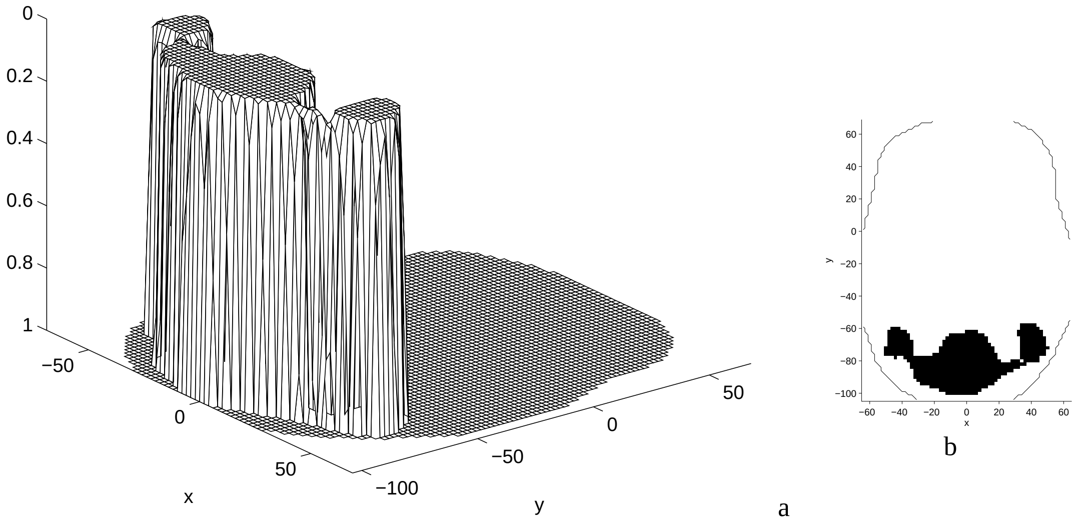


Figure 38
(a) Mesh plot of Bonferroni single step adjusted one-sided p-values, computed from the p-values of fig.37. Voxels outside the intracerebral volume have been removed. (b) Voxels with adjusted p-value below level 0.05. The outline of the intracerebral area is superimposed. The AC-PC plane is shown.

2.7. Pooled Variance

The low numbers of subjects and replications per subject in PET experiments frequently leaves very few degrees of freedom for the estimation of variance. This is particularly a problem in single subject analyses, and in the *t*-statistic approach, where the data for each subject is collapsed into a difference image. These gives tests at the voxel level with very low degrees of freedom, and hence low power. In addition, as we shall see in chapter 3, the multiple comparisons procedures based on random fields are conservative when the voxel statistics have low degrees of freedom.

Homogeneous variance

If it may be assumed that the variance is the same at all the voxels under consideration (*homoscedasticity*), then Worsley *et al*. (1992) advocate pooling the variance estimates across all the voxels. Consider the example of a multiple subject simple activation study, to be assessed using proportional scaling and a *t*-statistic on subject difference images (§2.3.1.). Assume $\Delta_{ik} \sim N(\mu_k,\sigma^2)$, That is, that the variance of the subject difference images across subjects is constant across the voxels. Since S_k^2 (eqn.23) is computed for each voxel using the same number of observations (the number of subjects, *N*) the degrees of freedom is the same for each voxel. The pooled sample variance is simply the mean over all the voxels, $\overline{S}_\bullet^2$. The *t*-statistic image is then formed using $\overline{S}_\bullet^2$ in place of S_k^2 in eqn.21, and is essentially just a normalised mean difference image:

$$T_k = \frac{\overline{\Delta}_k}{\sqrt{\overline{S}_\bullet^2 / n}} \tag{29}$$

Pooled variance regarded as known

Sample variance estimates at individual voxels are not independent, since the subject difference images Δ_i are smooth. Thus, the distribution of the pooled sample variance $\overline{S}_\bullet^2$ cannot be determined. However, if the estimate is formed by pooling over a very large number of voxels, and the smoothness of the subject difference images is much less in extent than the dimensions of the volume covered by these voxels, then the estimate effectively has large enough degrees of freedom that it can be regarded as known. Worsley *et al*. (1992) argue as follows (p901, c2): "…If we can find *R* voxels sufficiently separated so that they are independent, then the effective degrees of freedom is at least $(N\text{-}1)R$. Typically $R \approx 300$ and $N \approx 10$, so the effective degrees of freedom is large enough…" The value of *R* used is the number of *resolution elements*, a concept we shall return to when reviewing Worsley's method for assessing the significance of statistic images (§3.3.1.).

Distributional results: Gaussian statistic images

Regarding the variance as estimated almost exactly, $T_k \dot{\sim} N(\mu_k,1)$, giving a *Gaussian statistic image*. The hypothesis of no activation is then $\mathrm{H}_k{:}\mu_k = 0$, to be tested against the one-sided alternative $\overline{\mathrm{H}}_k{:}\mu_k > 0$. A one-sided p-value is then $1\text{-}\Phi(T_k)$.

2.7.1. Example–“V5” Study

Assuming homogeneity of variance, i.e. that $\sigma_k^2 = \sigma^2$ for all voxels $k = 1,\ldots,K$ in the “V5” study (K=77189), the pooled estimate of the common variance is $\overline{S}_\bullet^2 = 0.812$ (to 3dp). From the (voxel) sample variance image (fig.30, p84) it appears that the sample variance is not constant.

Chi-squared statistic image for homogeneity of variance

To assess the assumption of homogeneity of variance, consider the usual Chi-squared statistic for testing $\mathrm{H}_k{:}\sigma_k^2 = \sigma^2$ where σ^2 is known:

$$C_k = \frac{(n\text{-}1)\; S_k^2}{\sigma^2}$$

Taking $\overline{S}_\bullet^2$ as σ^2, under H_k $C_k \dot{\sim} \chi^2_{n\text{-}1}$. Since it is underestimation of variance that is of consequence, consider the one sided alternative hypothesis $\overline{\mathrm{H}}_k{:}\sigma_k^2 > \sigma^2$ (but note that with $\overline{S}_\bullet^2$ as σ^2, $\sigma_k^2 > \sigma^2$ for some voxels implies $\sigma_k^2 < \sigma^2$ for others). For the “V5” data this gives the statistic image ***C***, the AC-PC plane of which is depicted in figure 39. Over the whole intracerebral volume, 6.63% (to 2dp) of the voxels have associated statistics significant at the 1% level (unadjusted for multiple comparisons). This exceedence proportion gives a *p*-value of 0.0000 (to 4dp) for the omnibus hypothesis $\mathrm{H_W}$, computed by Worsley’s exceedence proportion test.[37] Thus we may conclude that the variance is not homogeneous for the “V5” data.

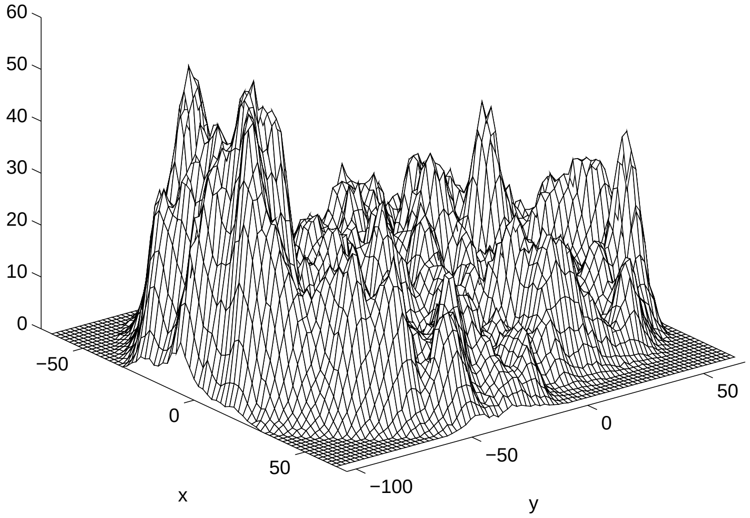


Figure 39
Mesh plot of Chi-squared statistic ***C***, indicating evidence against homogeneity of variance. The AC-PC plane is shown.

[37] The Chi-squared statistic image was “transformed” to a standard Gaussian statistic image. (By replacing each voxel statistic with the standard Gaussian ordinate, with identical probability of being exceeded. See §3.3.3.). The resulting Gaussian statistic image was then assumed to be a strictly stationary discrete Gaussian random field, and the variance-covariance matrix of partial derivatives estimated directly from the positive part of the image (see §3.3.5.). The omnibus *p*-value was then computed using Worsley’s exceedence proportion test (§3.4.2.), with a threshold of $-\Phi^{-1}(0.01)$.

2.7.2. Inappropriate use of pooled variance

The assumption of constant variance has often been made when analysing activation studies. One reason for this is that the resulting Gaussian statistic image is more amenable to analysis, as we shall see in chapter 3. The validity of this assumption is seldom checked,[38] and many experiments are analysed falsely assuming homogeneity of variance. One of the pitfalls is that variance may be underestimated, leading to falsely inflated *t*-statistics, and possibly false positives. To illustrate, consider the use of a pooled variance estimate for the "V5" data.

Underestimation of variance at site of activation

From the sample variance image (figure 30, p84) (or equivalently from the Chi-squared statistic image for homogeneous variance, figure 39), it appears that the sample variance is increased at the posterior of the brain. When considering the straightforward *t*-statistic image, computed with voxel level sample variance, the posterior of the brain was identified as activated (figure 32 and following text, p85). Thus the variance is increased at the site of the activation[39]. The use of a pooled variance *t*-statistic in this case falsely inflates the statistic at the site of the activation, as can be seen in figure 40 (compare with figure 31, p85). In addition, since the variance is now assumed to be known, the voxel *p*-values are greatly increased in significance (figure 41, compare with figure 33, p86). The inappropriate use of a pooled variance estimate can easily lead to false positives.

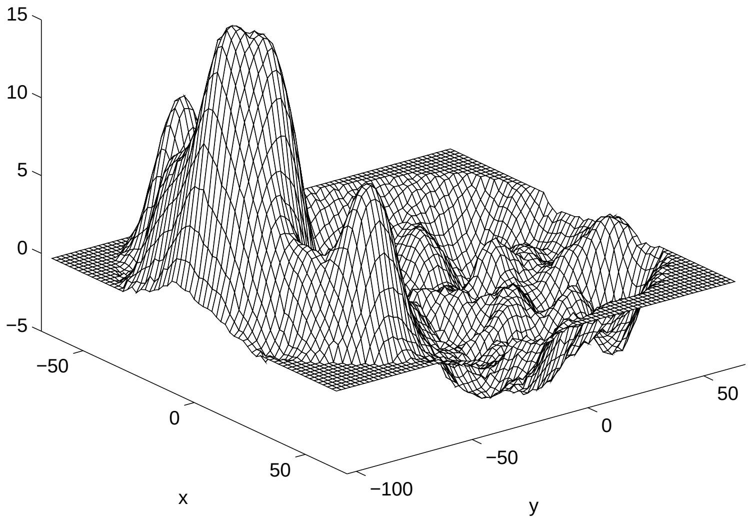


Figure 40
Mesh plot of *t*-statistic image ***T*** computed with pooled variance estimate for "V5" study. (Eqn.29) Each voxel statistic is distributed as standard Gaussian variate under the hypothesis of no activation at that voxel, H_k:μ_k= 0, where $\Delta_{ik} \sim N(\mu_k,\sigma^2)$ is assumed. The AC-PC plane is shown.

[38] Researchers usually use pre-written analysis software that do not include such checks. Furthermore, many of the tests themselves have low power, and can only detect gross deviations from the assumptions.

[39] This is perhaps to be expected. The "rest" and "active" conditions in the "V5" study both involve visual stimuli, so the non-visual areas of the brain should remain fairly stable. In the visual areas of the brain, different subjects can be expected to exhibit different increases in blood flow between the two conditions, so mean difference images can be expected to vary across subjects more in the challenged area than in those unaffected by the conditions.

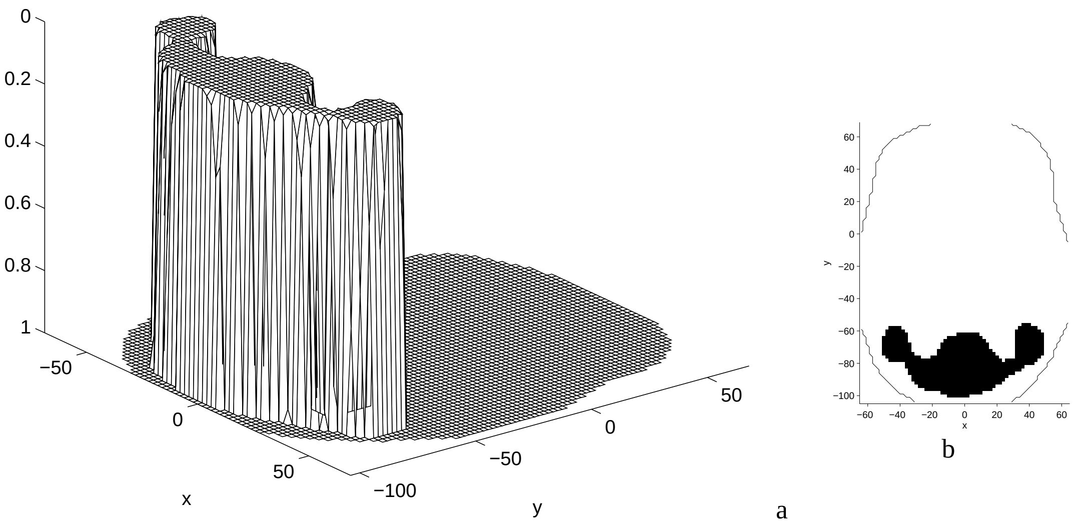


Figure 41
(a) Mesh plot of Bonferroni single step adjusted one-sided *p*-values, for the pooled variance *t*-statistic image of figure 40. Voxels outside the intracerebral volume have been removed. (b) Voxels with adjusted *p*-value below level 0.05. The outline of the intracerebral area is superimposed. The AC-PC plane is shown.

Underestimation of variance for grey matter due to white matter and ventricles

A further cause for underestimation of variance is due to the non-homogeneity of the brain itself. The brain consists of grey matter and white matter, and *ventricles* that are filled with spinal fluid. High level processing takes place in the grey matter on the surface of the brain, so it is only the grey matter that is interesting. Freely diffusible blood flow tracers get into the spinal fluid, but only in small quantities, so the ventricular regions of the brain appear to have a constant but low rCBF (rA) when measured with PET, and hence exhibit low variation between subject difference images. The ventricles are of sufficient size that a variance estimate pooled over the whole intracerebral volume underestimates the variance in grey matter regions.