14 | Statistical analysis of activation images

K.J. Worsley

14.2 Introduction—Why do we need statistics?

Statistical analysis is concerned with making inference about underlying patterns in data that often contain a large amount of random error. This is certainly the case with fMRI data, where the effect of a stimulus may be as little as 1 per cent of the BOLD signal. However, by careful averaging of the data over time, such as averaging the BOLD response at the times when the stimulus is ON, and subtracting the average when the stimulus is OFF, we are often able to detect such a small signal in the presence of considerable background noise.

More complex experimental designs require more complex analysis. The type of analysis can be guided by constructing a ***model*** for the way in which the BOLD response depends on the stimulus. Such a model must include a component of random error which explains how the observations vary even if the experiment is repeated on the same subject under exactly the same conditions. Statistics can be used to best estimate the parameters in the model, including the variability of the errors. It is this random variability of the errors that can then be used to assess the random variability, or standard error, of the estimated parameters themselves. This key quantity allows the experimenter not only to say that the BOLD response is increased by say 20 units due to a particular stimulus, but also how accurate that estimate is, say ± 8 units.

Finally, comparing the size of the increase to its standard error allows the experimenter to decide if any increase has really taken place; after all, the 20 unit increase could have occurred by chance alone when in fact the true increase (after a very large number of repetitions of the experiment) was very close to zero. However, the fact that the increase is $20/8 = 2.5$ times larger than its standard error makes this extremely unlikely; in fact this would happen with a probability of less than 1 per cent (under certain reasonable assumptions) if in fact there really was no true increase. We usually report this as 'z = 2–5, $P < 0.01$' or we sometimes give the exact P-value, or probability of a more extreme value of z than that observed, which is 0.0062.

Motivated by the above discussion, our first step in this chapter is build up a model of the fMRI data, beginning with the haemodynamic response to the stimulus (Section 14.2), then the random error (Section 14.3). The remainder of the chapter then deals with estimating the parameters of these models, assessing their variability, and making decisions about whether the fMRI data shows any evidence of a BOLD response to the stimulus.

Theoretical statistics material in this chapter, that can be skipped by the non-technical reader, is marked *.

14.2 Modelling the response to the stimulus

In this section we model the way in which the BOLD response depends on an external stimulus. Time will be denoted by t and the external stimulus will be denoted by $s(t)$. For example, $s(t)$ could take the value 1 when the stimulus is ON and 0 when the stimulus is OFF (see Figure 14.1(a)). The BOLD response at a particular voxel, denoted by $x(t)$, usually occurs between 3 and 10 s after the stimulus, peaking at about 6 s. This delay and blurring is modelled by a haemodynamic response function (HRF) $h(t)$ (Figure 14.1(b)) which weights past stimulus values by a *convolution* as follows:

$$x(t) = \int_0^\infty h(u)\, s(t-u)\, \mathrm{d}u. \qquad (14.1)$$

The HRF can be modelled as a simple gamma function

(Lange and Zeger, 1997). Friston (1998a) proposes a difference of two gamma functions that captures the fact that there is a small dip after the HRF has returned to zero:

$$h(t) = \left(\frac{t}{d_1}\right)^{a_1} \exp\left(\frac{-(t-d_1)}{b_1}\right) - c\left(\frac{t}{d_2}\right)^{a_2} \exp\left(\frac{-(t-d_2)}{b_2}\right) \quad (14.2)$$

where $d_j = a_j b_j$ is the time to the peak, and $a_1 = 6$, $a_2 = 12$, $b_1 = b_2 = 0.9$ s, and $c = 0.35$ (Glover, 1999). This particular choice of HRF is shown in Figure 14.1(b). The resulting convolution of $h(t)$ with $s(t)$ is shown in Figure 14.1(c). This is then sampled at the n FMRI volume acquisition times $t_1, \ldots t_n$ to give the response $x_i = x(t_i)$ at volume i.

In many experiments several different stimuli are presented. In the experiment used to illustrate the methods in this chapter, a subject was given a painful heat stimulus (49°C) to the left forearm for 9 s, followed by a neutral stimulus for 9 s, interspersed with 9 s when no stimulus was presented. These two stimuli $s_1(t)$ and $s_2(t)$ are shown in Figure 14.1(a). We can denote their corresponding responses by $x_1(t)$ and $x_2(t)$ by convolution with $h(t)$ as in (14.1). We usually assume that the responses have different magnitudes, denoted by β_1 and β_2 and that they add together to

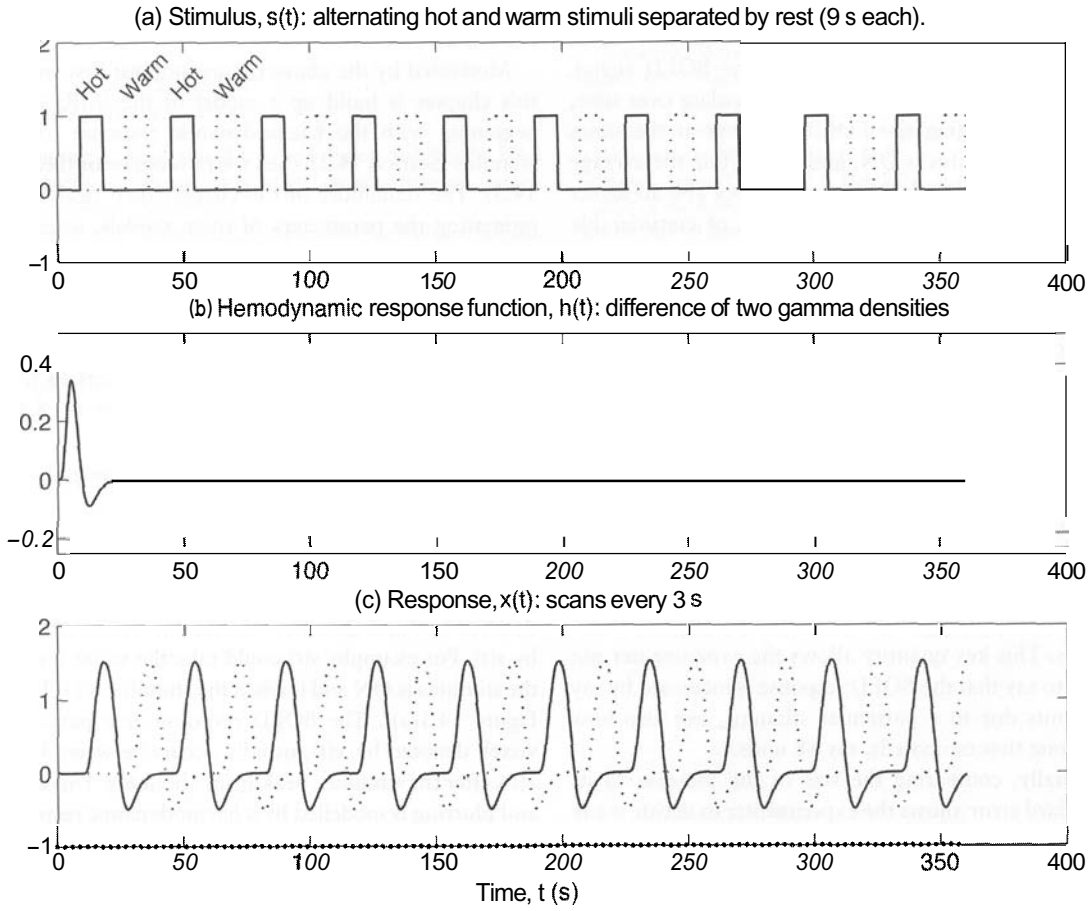


Fig. 14.1 (a) The hot and neutral stimuli $s(t)$, (b) the hemodynamic response function $h(t)$ and (c) its convolution with $s(t)$ to give the response $x(t)$. The time between volumes is $\Delta t = 3$s, so $x(t)$ is then subsampled at the $n = 118$ volume acquisition times $t_i = 3i$ to give the response $x_i = x(t_i)$ at time index $i = 1, \ldots n$.

produce the final BOLD response. In general, the effect of m different responses in volume i, denoted by $x_{i1}, \ldots x_{im}$ is modelled as the linear model (Friston *et al.*, **1995**)

$$x_{i1}\beta_1 + \ldots + x_{im}\beta_m. \tag{14.3}$$

Many voxels in fMRI data also show a slow variation over time, known as drift. The removal of drift, or low frequency noise, was described in Chapter **12.** Recall that drift can be removed either by high-pass filtering or by introducing low frequency drift terms, such as cosines, polynomials, or splines, into the linear model. However, drift also appears in this chapter, as the problem of drift interacts with model design; the presence of drift limits the type of stimulus design that can be used with fMRI experiments. Any stimulus that behaves like drift, such as a steadily increasing stimulus intensity, cannot be easily distinguished from drift and is either impossible or very difficult to estimate (i.e. estimable with very high error). This includes block designs with very long blocks, such as presenting a stimulus continually during the second half of an experiment. This type of design should be avoided. The best designs should try to present the stimulus fairly rapidly so that its effect can be assessed over a short period where the drift has little effect (see Section **14.8** for further discussion of optimal design).

14.3 Modelling the random error

Statistical models usually contain two parts: the fixed effects and the random error. The fixed effects are the parts of the model that do not vary if the experiment is repeated; they capture the underlying scientific 'truth' that we hope to discover. The random error is the part left over that varies every time new data is obtained. Random error is very important for two reasons: first, it tells us how to best estimate the effect of the stimulus, and second, and more importantly, it gives us a way of assessing the error in the effect. This then allows us to compare the effect with its random error, and select those voxels where the effect is much larger than its random error, that is, voxels with high signal-to-noise ratio.

The way to do this is to first combine the response and the drift terms, if high-pass temporal filtering has not been applied, into a single linear model for the fixed effects as in (14.3). Then a random error ε_i is added to obtain the observed fMRI data, Y_i, at time index i:

$$Y_i = x_{i1}\beta_1 + \ldots + x_{im}\beta_m + \varepsilon_i. \tag{14.4}$$

The observations tend to be correlated in time, particularly in cortical regions, with correlations up to **0.4** between time points 3 s apart (Fig. 14.2(a)). This effect is known as temporal autocorrelation (correlation of the errors separated by a fixed time lag) or smoothness; it can be caused, for example, by simply blurring the data in time, but is most likely due to some influence of the random error of the preceding time points on that of the current time point.

14.3.1 *Modelling the temporal correlation

Why is the temporal correlation structure important? The reason has to do not so much with how to estimate the signal strengths β_i, but with how to assess the standard errors of these estimates, and hence how to detect the presence of the signal. For this reason, we must take some care to model the correlation structure.

The simplest is the first order autoregressive model. This is generated by combining the error from the previous time point with a new error term to produce the error for the current time point:

$$\varepsilon_i = \rho\varepsilon_{i-1} + \chi_{i1},$$

where $|\rho| < 1$ and χ_{i1} is a 'white noise' sequence of independent and identically distributed normal random variables with mean 0 and standard deviation σ_1, written as $\chi_{i1} \sim N(0, \sigma_1^2)$. With such a model, the temporal correlation decays exponentially as the lag l increases:

$$\mathrm{Cor}(\varepsilon_i, \varepsilon_{i-l}) = \rho^{|l|}.$$

More complex oscillatory behaviour as well as exponential decay can be obtained by adding more terms to give autoregressive models of order p, known as AR(p) models:

$$\varepsilon_i = \alpha_1\varepsilon_{i-1} + \ldots + \alpha_p\varepsilon_{i-p} + \chi_{i1},$$

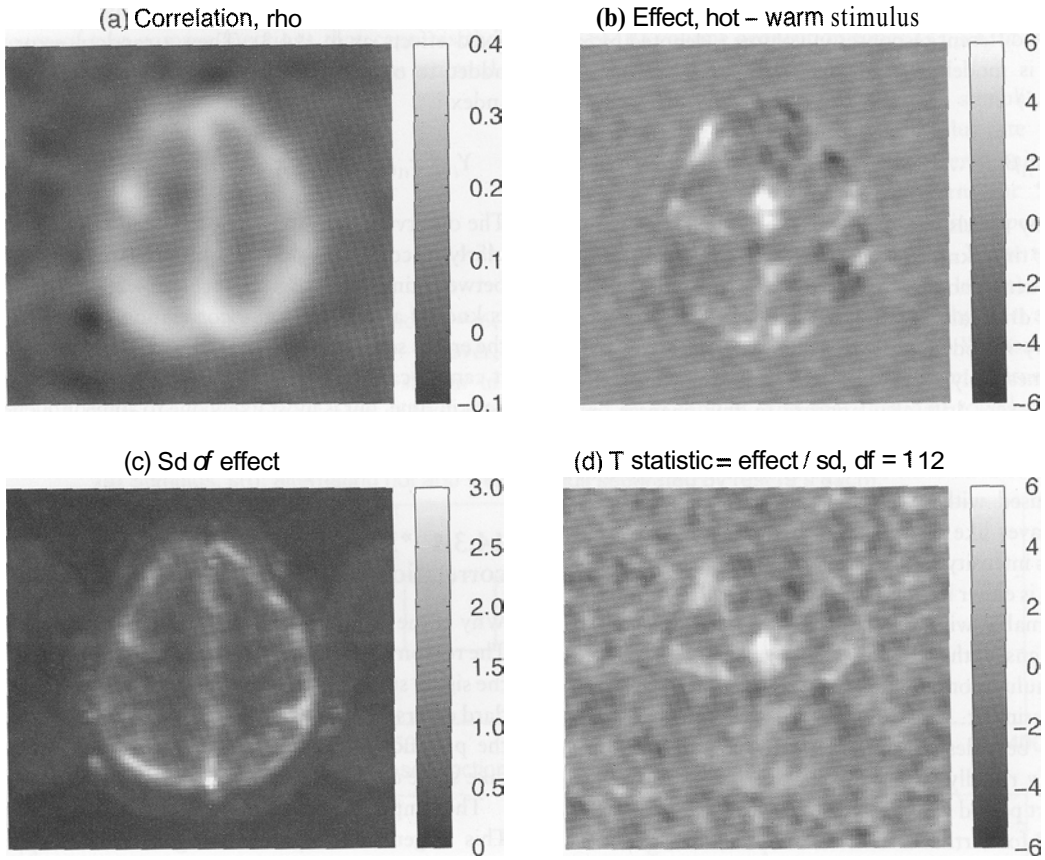


Fig. 14.2 Statistical analysis of the fMRI data. (a) The estimated AR(1) autocorrelation parameter ρ after bias correction and spatial smoothing with a 15mm FWHM Gaussian filter. Note that the correlation is high in cortical regions. (b) The effect of the hot stimulus minus the neutral stimulus, $c'\hat{\beta}$. (c) The estimated standard deviation of the effect $(\hat{\text{Var}}(c'\hat{\beta}))^{1/2}$. Note that it is much higher in cortical regions than elsewhere in the brain. (d) The T statistic T equal to (b) divided by (c), with $\theta = 112$ degrees of freedom.

To take into account white noise from the scanner, Purdon *et al.* (1998) has extended the AR(1) model as follows:

$$\delta_i = \rho\eta_{i-1} + \chi_{i1}, \tag{14.5}$$

$$\varepsilon_i = \eta_i + \chi_{i2}, \tag{14.6}$$

in which a second independent white noise term $\chi_{i2} \sim N(0,\sigma_2^2)$ is added to an AR(1) component. This extra component χ_{i2} accounts for the scanner white noise which is added to physiological 'coloured' (temporally correlated) noise η_i from the brain itself. The correlation then becomes

$$\text{Cor}(\varepsilon_i, \varepsilon_{i-l}) = \frac{\rho^{|l|}}{1 + (1-\rho^2)\sigma_2^2/\sigma_1^2},$$

if $l \neq 0$ and 1 if $l = 0$. In other words, there is a jump at zero lag, known in the geostatistics literature as a 'nugget effect'. Further autoregressive terms can be added. This is a special type of state space model (Caines 1988) in which (14.5) is the state equation, and (14.6) is the observation equation. State space models are extremely powerful at capturing complex dynamic relationships, including drift.

14.4 Estimating the response magnitudes

So far we have built a simple model for the BOLD response (Section 14.2) and the random error that is added to that response (Section 14.3). The magnitudes β_j of the responses are still unknown, and the purpose of this section is to find good estimates of them (Sections 14.5 and 14.9 look at how we estimate the noise). We present three methods: the 'best possible' (fully efficient, that is, most accurate) method, the potentially more robust SPM'99 method, and the Fourier space method. Finally we compare the three methods.

14.4.1 "Notation

To do the theoretical work in this section we shall need matrix notation:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \text{h} \\ Y_n \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \text{h} & & \text{h} \\ x_{n1} & \cdots & x_{nm} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_1 \\ \text{h} \\ \beta_m \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \text{h} \\ \varepsilon_n \end{pmatrix}$$

As before, n is the number of time points (volumes) and m is the number of different stimuli or drift terms (explanatory variables) being modelled.

We will denote the variance of the vector of errors ε by the matrix $\mathbf{V}\sigma^2$, where σ^2 is an unknown scalar, and the element of V in row i and column j, multiplied by σ^2, is the covariance between ε_i and ε_j. For the AR(1) model, for example,

$$\mathbf{V} = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \text{h} & \text{h} & \text{h} & & \text{h} \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}, \quad \sigma^2 = \frac{\sigma_1^2}{1-\rho^2}$$

We further assume that the distribution of ε is multivariate normal, so that we can write the entire linear model as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim \mathrm{N}_n(0, \mathbf{V}\sigma^2). \tag{14.7}$$

A general unbiased estimator of β can be found by first multiplying (14.7) through by an $n \times n$ matrix A (various possible choices of A will be discussed below) to give:

$$\tilde{\mathbf{Y}} = \mathbf{AY}, \quad \tilde{\mathbf{X}} = \mathbf{AX},$$

$$\tilde{\mathbf{Y}} = \mathbf{X}\beta + \tilde{\varepsilon} \quad \tilde{\varepsilon} = \mathbf{A}\varepsilon \sim \mathrm{N}_n(0, \mathbf{AVA}'\sigma^2). \tag{14.8}$$

The least squares estimator of β is the value of β that minimizes the sum of squared errors in (14.8), that is,

$$\min \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta\|^2.$$

We shall adopt conventional statistical notation and denote estimators by ^ throughout. It can be shown that

$$\hat{\beta} = \tilde{\mathbf{X}}^+ \tilde{\mathbf{Y}},$$

where + denotes the Moore–Penrose pseudoinverse, the 'best possible' inverse that minimizes the error sum-of-squares. For any choice of A, $\hat{\beta}$ is unbiased and its variance matrix is given by:

$$\mathrm{E}(\beta) = \hat{\beta}, \quad \mathrm{Var}(\hat{\beta}) = \tilde{\mathbf{X}}^+ \mathbf{AVA}' \tilde{\mathbf{X}}^{+\prime} \sigma^2.$$

Note that the estimation of β (by calculating the pseudoinverse of the design matrix and then using $\hat{\beta} = \tilde{\mathbf{X}}^+ \tilde{\mathbf{Y}}$) is what is referred to generally as fitting the model to the data.

14.4.2 *The fully efficient estimator

The fully efficient (most accurate, i.e. minimum variance) estimator of β is obtained by choosing A so that the variance of the errors is proportional to the identity matrix, equivalent to 'whitening' the errors, by the Gauss–Markov Theorem. This process removes any temporal smoothness in the data, whether caused by the intrinsic smoothness of the random errors, or by pre-filtering. This is accomplished by factoring V, for example by a Cholesky factorization, then inverting the transpose of the factor:

$$\mathbf{V} = \mathbf{H'H}, \quad \mathbf{A} = (\mathbf{H'})^{-1}, \quad \mathbf{AVA'} = \mathbf{I},$$

where **I** is the $n \times n$ identity matrix. Doing this in practice can be very time consuming if it is repeated at every voxel. Fortunately thcrc are computationally efficient ways of finding **A** if the errors are generated by an AR(p) process or a state space model (using the Kalman filter). For the AR(1) model, for example,

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -\rho R & R & 0 & \dots & 0 \\ 0 & -\rho R & R & \text{j} & \text{h} \\ \text{h} & \text{j} & & \text{j} & 0 \\ 0 & \dots & 0 & -\rho R & R \end{pmatrix}$$

where $R = 1/\sqrt{(1-\rho^2)}$, so that $\tilde{Y}_1 = Y_1$ and $\tilde{Y}_i = (Y_i - \rho Y_{i-1})/\sqrt{(1-\rho^2)}$, $i = 2, \dots n$.

14.4.3 "The more robust estimator of SPM *'99*

An alternative, adopted by SPM '99, is to 'precolour', or smooth, the data (and the model). This yields an unbiased estimator of β, but with slightly increased variance. This small loss of efficiency is offset by a more robust estimator of the variance, that is, an estimator of β whose estimated variance is less sensitive to departures from the assumed form of **V**. Rather than modelling the correlation structure of the original observations, SPM'99 adopts an AR(1) model for the smoothed data.

14.4.4 "Estimation in Fourier space

Still another possibility is to choose **A** so that the transformed observations ? are independent though not necessarily equally variable (as for the fully efficient estimator). It is a remarkable fact that if the error process is stationary (the same correlation structure at any time) then this can be achieved by choosing the rows of **A** to be the Fourier transform sine and cosine basis functions (the reason is that these basis functions are almost eigenvectors of **V**). This would be exactly so if the correlation structure were periodic; non-periodicity is less important if the sequence is long. Multiplying by **A** is then equivalent to taking the Fourier transform, a very rapid operation.

Thc advantage is that the resulting errors $\tilde{\varepsilon}$ become almost independent, but with variances equal to the spectrum of the process (in engineering terms, the expected power of the process at each frequency). This simplifies the analysis; fitting the model (14.8) is then equivalent to weighted least squares, with weights inversely proportional to the spectrum. From this point of view, the SPM '99 method can be seen as weighted least squares with weights proportional to the spectrum of the haemodynamic response function, which gives more weight to the frequencies that are passed by the haemodynamic response, and less weight to those that are damped by the haemodynamic response.

An added advantage of working in Fourier space is that convolution of the stimulus with the haemodynamic response function (14.1) becomes simple multiplication of their Fourier transforms. We make use of this to estimate the haemodynamic response itself in Section 14.11.

14.4.5 Comparison of the methods

The fully efficient method, based on pre-whitening thc data, produces the best estimators if the correlation structure is correctly modelled. The SPM method is more robust to biases in modeling and estimating the correlation structure, at the expensive of losing a little accuracy (again if the correlation structure is correctly modelled). It is not easy to choose between them, but fortunately both methods give very similar answers in most situations. The Fourier space method is simply a convenient way of implementing either the fully efficient or the SPM methods, particularly when the design is periodic.

It should be noted that parameter estimation for some types of experimental design is unaffected by the choice of **A.** It can be shown that if the columns of **X** are linear combinations of p eigenvectors of **A'A,** then the same estimator can be obtained by using least squares in model (14.4), i.e. ignoring multiplication by **A** altogether. For fully efficient estimation, $\mathbf{A'A} = \mathbf{V}^{-1}$, which has the same eigenvectors as **V.** Now as remarked above, the eigenvectors of **V** are the Fourier sine and cosine functions, provided the error process is stationary.

This implies that stimuli whose intensity varies as a sine or cosine function can be estimated with full efficiency by ignoring **A.** Furthermore, for the SPM '99

method, A'A is a Toeplitz matrix whose eigenvectors are almost the Fourier sines and cosines, so here again a design of this sort is estimated with full efficiency by the SPM '99 method. The reason should now be clear: for this design, the regressors are just 1 at a subset of p of the frequencies, and zero elsewhere. Data at only these m frequencies are used to estimate m parameters, so any weighting scheme yields the same parameter estimates.

Block designs are almost sine functions, so these are estimated with almost full efficiency by the SPM '99 method. Random event-related designs have a more complex spectrum so these are most affected by the choice of method; precolouring can result in a fairly inefficient analysis in the case of dense event-related designs.

14.5 *Estimating the variance

In this section we look at how to estimate the error variance σ^2, assuming for the moment that the error correlation structure V is known (we shall look at estimating V a little later in Section 14.9). The estimator is based on the residuals, defined as the difference between the data and the estimated fixed effects

$$\mathbf{r} = \tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\beta} = \mathbf{R}\tilde{\mathbf{Y}},$$

where $\mathbf{R} = \mathbf{I} - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{+}$. The estimator of σ^2 is the sum of squares of the residuals divided by a constant chosen so that $\hat{\sigma}^2$ is unbiased:

$$\hat{\sigma}^2 = \mathbf{r}'\mathbf{r}/\mathrm{tr}(\mathbf{RAVA}').$$

Its effective degrees of freedom, based on matching second moments, is

$$\nu = \mathrm{tr}(\mathbf{RAVA}')^2/\mathrm{tr}((\mathbf{RAVA}')^2),$$

so that the distribution of $\nu\hat{\sigma}^2/\sigma^2$ is well approximated by a χ^2 distribution with ν degrees of freedom, known as the Satterthwaite approximation. If the estimation is fully efficient, so that AVA' = **I**, then the degrees of freedom becomes the usual $\nu = n - \tilde{m}$, where $\tilde{m}$ is the rank of $\tilde{\mathbf{X}}$, and the Satterthwaite approximation is exact.

For example, say we have $n = 118$ volumes, two parameters for the fMRI response, and four parameters for a polynomial drift of degree 3, giving $m = 6$ total parameters. If the fully efficient estimator is used (i.e. 'prewhitening'), the degrees of freedom is $\nu = 118 - 6 = 112$.

14.6 Detecting an effect

In the previous two sections we have given methods for estimating both the signal and noise parameters. Usually we are more interested in comparing signal parameters, such as whether the hot stimulus gives a bigger response than the neutral stimulus. In other words, we are interested in the difference between the hot stimulus and the neutral stimulus, $\beta_1 - \beta_2$. This is known as an effect (an effect could of course be just a single magnitude, say β_1). In this section we give estimates for an effect and its variance. We then turn to the crucial question of detecting an effect, that is, whether or not there is any evidence for an effect. In this way we can detect those voxels where there is evidence that the pain stimulus produces a BOLD response over and above that produced by the neutral stimulus.

14.6.1 T-tests

The particular differences of the parameters β that make up the effect are specified by a *contrast* vector **c**, a vector of the same length as β, which specifies a linear combination of the parameters $\mathbf{c}'\beta$. First of all, this is estimated by the same linear combination of the estimated parameters $\mathbf{c}'\hat{\beta}$ (from now on we shall use the term effect to refer to the estimator as well as to the combination of parameters to be estimated).

The simplest contrast involves only one explanatory variable. For example, to test activation in the hot condition versus rest, the contrast vector is $\mathbf{c}' = (1\,0\,0\,0\,0\,0)$ (the first zero excludes the neutral condition from the contrast and the other four exclude the cubic drift). This means that the estimate of the effect is simply $\mathbf{c}'\hat{\beta} = \hat{\beta}_1$. If, however, we are interested in the difference between the hot stimulus (β_1) and the neutral stimulus (β_2), then the contrast vector is $\mathbf{c}' = (1\,-1\,0\,0\,0\,0)$. This means that the estimate of the effect is $\mathbf{c}'\hat{\beta} = \hat{\beta}_1 - \hat{\beta}_2$ (Figure 14.2(b)).

We have now defined the effect via the contrast vector, and given a natural estimator of it. As usual, we

would also like to estimate the variance of the effect, which is given by

$$\hat{\mathrm{Var}}(\mathbf{c}'\hat{\beta}) = \mathbf{c}'\tilde{\mathbf{X}}^{+}\mathbf{AVA}'\tilde{\mathbf{X}}^{+\prime}\,\mathbf{c}\hat{\sigma}^2. \qquad (14.9)$$

An example of the estimated standard deviation (the square root of the estimated variance (14.9)) is shown in Figure 14.2(c). Note that it is much higher in cortical regions than elsewhere in the brain.

Finally we test whether or not there is any evidence for the effect, that is whether the effect differs from zero, using the ratio of the effect to its standard error, called the T statistic:

$$T = \frac{\mathbf{c}'\hat{\beta}}{\sqrt{\hat{\mathrm{Var}}(\mathbf{c}'\hat{\beta})}}$$

This has an approximate t distribution with v degrees of freedom (exact if $\mathbf{AVA}' = \mathbf{I}$) when there is no effect, that is, $\mathbf{c}'\beta = 0$. An example is shown in Figure 14.2(d), where large values of T indicate evidence for an effect.

Note that T (or F) statistic images are often converted into Z statistic images, where the theoretical T (or F) null distribution (given the relevant degrees-of-freedom) is converted into a unit-variance Gaussian distribution (which does not depend on degrees-of-freedom). Thus Z statistic images are often referred to as 'Gaussianised T-statistics' etc. (Note however that the random field theory in Sections 14.12.1 and 14.12.2 does not apply to an image of Gaussianised statistics, and must be applied to the original non-Gaussianised images of T or F statistics.)

14.6.2 F-tests for several contrasts

Sometimes we may wish to make a simultaneous test of several contrasts at once. For example, we may wish to detect *any* difference between the hot and neutral stimuli and rest. This can be done by using a contrast matrix

$$\mathbf{c}' = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The first row compares hot to rest, the second compares neutral to rest. To test $K > 1$ contrasts at the same time, that is, if c is a $K \times m$ matrix, the T statistic is replaced by an F statistic defined by

$$F = \frac{\hat{\beta}'\mathbf{c}(\hat{\mathrm{Var}}(\mathbf{c}'\hat{\beta}))^{-1}\mathbf{c}'\hat{\beta}}{K}$$

which has an approximate F distribution with K and v degrees of freedom (exact if $\mathbf{AVA}' = \mathbf{I}$) when there is no effect, $\mathbf{c}'\beta = 0$. The effects are then detected simultaneously by large values of F. If $K = 1$, then $F = T^2$, so the F-test is equivalent to the T-test.

An alternative is to use the increase in error sum of squares when the model is restricted so that $\mathbf{c}'\beta = 0$. One way of doing this is to replace $\mathbf{X}$ with $\mathbf{X}(\mathbf{I} - \mathbf{c}\mathbf{c}^{+})$. If $\mathbf{R}_0$ is the equivalent of R under the restricted model, so that the restricted residuals are $\mathbf{r}_0 = \mathbf{R}_0\mathbf{Y}$, then the resulting F statistic is

$$\bar{F} = (\mathbf{r}_0'\mathbf{r}_0 - \mathbf{r}'\mathbf{r})/(\mathrm{tr}\{(\mathbf{R} - \mathbf{R}_0)\,\mathbf{AVA}'\}\hat{\sigma}^2)$$

which has an approximate F distribution with $\bar{K}$ and v degrees of freedom, where

$$\bar{K} = \mathrm{tr}[(\mathbf{R} - \mathbf{R}_0)\mathbf{AVA}']^2/\mathrm{tr}\{[(\mathbf{R} - \mathbf{R}_0)\mathbf{AVA}']^2\}.$$

If the estimation is fully efficient ($\mathbf{AVA}' = \mathbf{I}$) then the two F-tests are identical ($\bar{F} = \mathrm{F}$).

14.6.3 When to use F-tests

F-tests should only be used when we are interested in any linear combination of the contrasts. For example, an F-test would be appropriate for detecting regions with high polynomial drift, since we would be interested in either a linear, quadratic or cubic trend, or any linear combination of these. In this case we could use the contrast matrix

$$\mathbf{c}' = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Another good use of the F-test is for detecting effects when the haemodynamic response is modelled by a set of basis functions (see Section 14.11).

The F-test could also be used for detecting differences between a set of stimuli as in (14.10), but a significant result would simply say that there were *some* differences between the stimuli, without saying which ones were different. Researchers would probably be more interested in comparing each stimulus

would also like to estimate the variance of the effect, which is given by

$$\hat{\mathrm{Var}}(\mathbf{c}'\hat{\beta}) = \mathbf{c}'\tilde{\mathbf{X}}^{+}\mathbf{A}\mathbf{V}\mathbf{A}'\tilde{\mathbf{X}}^{+\prime}\mathbf{c}\hat{\sigma}^2. \qquad (14.9)$$

An example of the estimated standard deviation (the square root of the estimated variance (14.9))is shown in Figure 14.2(c). Note that it is much higher in cortical regions than elsewhere in the brain.

Finally we test whether or not there is any evidence for the effect, that is whether the effect differs from zero, using the ratio of the effect to its standard error, called the T statistic:

$$T = \frac{\mathbf{c}'\hat{\beta}}{\sqrt{\hat{\mathrm{Var}}(\mathbf{c}'\hat{\beta})}}$$

This has an approximate t distribution with ν degrees of freedom (exact if $\mathbf{AVA}' = \mathbf{I}$) when there is no effect, that is, $\mathbf{c}'\beta = 0$. An example is shown in Figure 14.2(d), where large values of T indicate evidence for an effect.

Note that T (or F) statistic images are often converted into Z statistic images, where the theoretical T (or F) null distribution (given the relevant degrees-of-freedom) is converted into a unit-variance Gaussian distribution (which does not depend on degrees-of-freedom). Thus Z statistic images are often referred to as 'Gaussianised T-statistics' etc. (Note however that the random field theory in Sections 14.12.1 and 14.12.2 does not apply to an image of Gaussianised statistics, and must be applied to the original non-Gaussianised images of T or F statistics.)

14.6.2 F-tests for several contrasts

Sometimes we may wish to make a simultaneous test of several contrasts at once. For example, we may wish to detect *any* difference between the hot and neutral stimuli and rest. This can be done by using a contrast matrix

$$\mathbf{c}' = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The first row compares hot to rest, the second compares neutral to rest. To test $K \geq 1$ contrasts at the same time, that is, if c is a K x m matrix, the T statistic is replaced by an F statistic defined by

$$F = \frac{\hat{\beta}'\mathbf{c}(\hat{\mathrm{Var}}(\mathbf{c}'\hat{\beta}))^{-1}\mathbf{c}'\hat{\beta}}{K}$$

which has an approximate F distribution with K and ν degrees of freedom (exact if $\mathbf{AVA}' = \mathbf{I}$) when there is no effect, $\mathbf{c}'\beta = 0$. The effects are then detected simultaneously by large values of F. If $K = 1$, then $F = T^2$, so the F-test is equivalent to the T-test.

An alternative is to use the increase in error sum of squares when the model is restricted so that $\mathbf{c}'\beta = 0$. One way of doing this is to replace $\mathbf{X}$ with $\mathbf{X}(\mathbf{I} - \mathbf{c}\mathbf{c}^{+})$. If $\mathbf{R}_0$ is the equivalent of R under the restricted model, so that the restricted residuals are $\mathbf{r}_0 = \mathbf{R}_0\mathbf{Y}$, then the resulting F statistic is

$$\bar{F} = (\mathbf{r}_0'\mathbf{r}_0 - \mathbf{r}'\mathbf{r})/(\mathrm{tr}((\mathbf{R} - \mathrm{R},)\ \mathbf{AVA}')\hat{\sigma}^2)$$

which has an approximate F distribution with $\bar{K}$ and ν degrees of freedom, where

$$\bar{K} = \mathrm{tr}[(\mathbf{R} - \mathbf{R}_0)\mathbf{AVA}']^2/\mathrm{tr}\{[(\mathbf{R} - \mathbf{R}_0)\mathbf{AVA}']^2\}.$$

If the estimation is fully efficient ($\mathbf{AVA}' = \mathbf{I}$) then the two F-tests are identical ($\bar{F} = F$).

14.6.3 When to use F-tests

F-tests should only be used when we are interested in any linear combination of the contrasts. For example, an F-test would be appropriate for detecting regions with high polynomial drift, since we would be interested in either a linear, quadratic or cubic trend, or any linear combination of these. In this case we could use the contrast matrix

$$\mathbf{c}' = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Another good use of the F-test is for detecting effects when the haemodynamic response is modelled by a set of basis functions (see Section 14.11).

The F-test could also be used for detecting differences between a set of stimuli as in (14.10), but a significant result would simply say that there were *some* differences between the stimuli, without saying which ones were different. Researchers would probably be more interested in comparing each stimulus

with a baseline, or paired comparisons between all pairs of stimuli, using a simple *T*-test. If desired, a Bonferroni correction could be used to correct for multiple T-tests, in which the *P* values are multiplied by the number of tests made. In other words, most scientific questions can be handled by a few well-chosen *T*-tests, rather than an F-test.

14.7 Setting up the model—an example

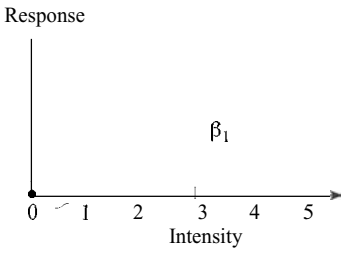
For the experimenter, specifying the the stimuli $s(t)$ and the contrasts **c** are the most difficult steps in the analysis, because these relate the design of the experiment to the scientific questions. A few examples should help clarify some important issues.

14.7.1 A linear intensity effect

Suppose a single stimulus is compared to a baseline, but the intensity of the stimulus varies. We are interested first in whether the stimulus was detected, and then whether the effect increases with stimulus intensity. Suppose the stimulus is presented sequentially in the first 10 blocks, with intensity values 0 (no stimulus), 1, 2, 3, 4, *5,* and two stimulus functions (explanatory variables) $s_1(t)$ and $s_2(t)$ are set up to capture a linear intensity effect, as follows (recall that $x(t)$ is $s(t)$ convolved with $h(t)$):

Linear model. orthogonalized

Block	1	2	3	4	5	6	7	8	9	10
Intensity	0	1	0	2	0	3	0	4	0	5
$s_1(t)$	0	1	0	1	0	1	0	1	0	1
$s_2(t)$	0	−2	0	−1	0	0	0	1	0	2

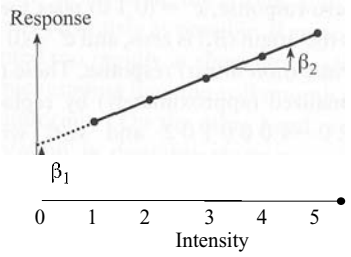


Each block might comprise say 3 volumes of 3 s each, and these 10 blocks might be repeated say 4 times in one run to give 120 volumes in all. Recall that the constant term, which models the baseline level (horizontal axis), would normally be included in the drift terms.

Note that the second stimulus model $s_2(t)$ has been centered by subtracting its mean. This allows us to look at two (nearly) orthogonal contrasts, that is, contrasts whose estimators are statistically independent: c' = (1 0) which tests for an overall effect of the stimulus compared to baseline (β_1), and **c'** = (0 1) which tests for a linear effect of stimulus intensity (β_2). These contrasts are not quite orthogonal because of the temporal correlation. Note that if $s_2(t)$ had been replaced by 0 1 0 2 0 3 0 4 0 5 as follows:

Linear model, not orthogonalized

Block	1	2	3	4	5	6	7	8	9	10
Intensity	0	1	0	2	0	3	0	4	0	5
$s_1(t)$	0	1	0	1	0	1	0	1	0	1
$s_2(t)$	0	1	0	2	0	3	0	4	0	5



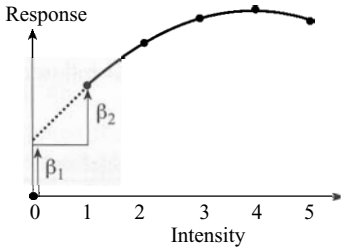
then the fit of the model would be identical, but the interpretation of the tests would be different. The second contrast **c'** = (0 1) would still test for a linear effect of stimulus intensity (β_2), but the first contrast c' = (1 0) would now test for the intercept of the intensity response (β_1), that is, whether zero intensity gives zero response. Replacing it with **c'** = (1 3), where 3 is the average of the non-zero intensities, would once again test for an overall effect of the stimulus as above.

14.7.2 A quadratic intensity effect

Sometimes the relationship of intensity to response may be non-linear. This can be captured by adding higher order terms, such as the quadratic term $s_3(t) = s_2(t)^2$:

Ouadratic model

Block	1	2	3	4	5	6	7	8	9	10
Intensity	0	1	0	2	0	3	0	4	0	5
$s_1(t)$	0	1	0	1	0	1	0	1	0	1
$s_2(t)$	0	1	0	2	0	3	0	4	0	5
$s_3(t)$	0	1	0	4	0	9	0	1 6	0	25



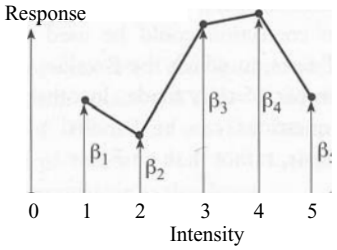
The contrast $c' = (1\ 0\ 0)$ tests for whether zero intensity gives zero response, $c' = (0\ 1\ 0)$ tests for whether the slope at the origin (β_2) is zero, and $c' = (0\ 0\ 1)$ tests for a quadratic (non-linear) response. These terms can be orthogonalized (approximately) by replacing $s_2(t)$ with $0\ -2\ 0\ -1\ 0\ 0\ 0\ 1\ 0\ 2$ and $s_3(t)$ with $0\ 2\ 0\ -1\ 0\ -2\ 0\ -1\ 0\ 2$, in which case $c' = (1\ 0\ 0)$ tests for an overall effect, $c' = (0\ 1\ 0)$ tests for a linear effect, and $c' = (0\ 0\ 1)$ tests for the same quadratic effect as before.

14.7.3 Intensity as a factor

The factor model assigns a separate parameter to each level of the intensity, allowing for an arbitrary relationship between stimulus intensity and response:

Factor model

Block	1	2	3	4	5	6	7	8	9	1	0
Intensity	0	1	0	2	0	3	0	4	0	5	
$s_1(t)$	0	1	0	0	0	0	0	0	0	0	
$s_2(t)$	0	0	0	1	0	0	0	0	0	0	
$s_3(t)$	0	0	0	0	0	1	0	0	0	0	
$s_4(t)$	0	0	0	0	0	0	0	1	0	0	
$s_5(t)$	0	0	0	0	0	0	0	0	0	1	



The coefficients $\beta_1,\beta_2,\beta_3,\beta_4,\beta_5$ are now the effects of each stimulus level relative to the baseline (0 stimulus). To test for an arbitrary unspecified non-linear effect, use an F-test with contrast matrix

$$c' = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The factor model is identical to a fourth degree polynomial because a fourth degree polynomial can be fitted exactly through any five points. We can still test for polynomial effects using the factor model: to test for an overall effect, use $c' = (1\ 1\ 1\ 1\ 1)$; to test for a linear effect, use $c' = (-2\ -1\ 0\ 1\ 2)$; to test for a quadratic effect, use $c' = (2\ -1\ -2\ -1\ 2)$.

What is the difference between testing for a linear effect using the contrast $c' = (-2\ -1\ 0\ 1\ 2)$ in the factor model and the contrast $c' = (0\ 1)$ in the linear model? The estimated effect is identical in both cases, but their standard deviations might be different. The reason is that the linear model only allows only for a linear effect, whereas the factor model allows for more polynomial effects.

If the effect is predominantly linear, then the F-test may fail to detect it, in other words, it has less sensitivity than a T-test with the contrast $c' = (-2\ -1\ 0\ 1\ 2)$. This is the price paid for not knowing where to look; the F-test looks for all possible effects: overall, linear, quadratic, cubic and quartic, so naturally it has to sacrifice some sensitivity against a directed search for any one of them. The usual advice applies here: first look in the direction where you expect to see something (T-test), then look in all possible directions for the unanticipated (*F*-test).

14.7.4 The design

Two comments on the design of this experiment. First, the design could be improved by rearranging the temporal order of the intensity levels, because a linear intensity effect could be confounded with drift. The reason is that if an effect of interest looks like drift, then it will be partially removed by the drift terms in the linear model. A good choice might be to present the stimulus intensities in the order 4 1 3 5 2 which is orthogonal to a linear drift. Second, since the primary interest is probably detection of the stimulus, it has been alternated with the baseline, rather than say putting all the baselines together at one end. The next section, 14.8 on optimal design, gives a justification for this.

14.7.5 The baseline or rest condition

In the factor model there is no explicit indicator variable, such as 1 0 1 0 1 0 1 0 1 0 for the baseline or rest condition of no stimulus; the same is true for the hot/neutral example. What is special about this condition? What would happen if we simply had the five intensity levels, or the hot and neutral stimulus, with no rest or baseline in between? Which one would we treat as the baseline? The answer is that the baseline is the condition of the subject *before* volume acquisition commenced. In the varying intensity experiment, we are assuming that the condition of 0 stimulus persisted before scanning commenced, so this is the baseline and it is not modelled as a separate condition. In the hot/neutral experiment, the subject was at rest with no heat stimulus before scanning commenced, so this is the baseline.

If the hot/neutral experiment consisted of alternating hot and neutral stimuli applied at the start of the first volume, with no rest in between, then the hot and neutral conditions should still be modelled with a separate stimulus response, as in Figure 14.1 but without the gaps. The baseline or rest condition would appear in the first few volumes, carried over by the haemodynamic response function. The first coefficient β_1 would then measure the difference between the hot stimulus and this small amount of pre-scanning baseline; the same would be true for the neutral stimulus coefficient β_2. Obviously by themselves these coefficients would not be very informative (due to high standard deviation; one explanatory variable is very nearly the inverse of the other, leading to poorly conditioned estimation of the individual parameters). However, the main interest is the difference between the hot and the neutral, and this would still be well estimated (low standard deviation) by the difference of the coefficients $\beta_1 - \beta_2$. If on the other hand the neutral stimulus was continuously applied preceding the first volume, then it would become the baseline and only the hot stimulus would be used; the coefficient β_1 of the hot stimulus would then measure the difference hot–neutral.

14.8 Optimal experimental design

The question arises of how to optimally design the experiment in order for the data to contain the maximum possible amount of extractable information. In other words, how should we choose the frequency and duration of the stimuli in order to have the greatest sensitivity in detecting the effects, and to estimate the effects as accurately as possible. If an on-off stimulus is presented too rapidly in short blocks then the haemodynamic response function will smooth the response to near-uniformity. On the other hand, a short stimulus presentation is desirable since it capitalises on the temporal correlation, which reduces the variance of the on minus the off volumes. Optimal designs have been investigated by Friston *et al.* (19996).

The problem comes down to finding the stimulus that minimizes $\mathrm{Var}(\mathbf{c}'\hat{\beta})$ from (14.9). To simplify the discussion, assume that there is just one parameter, $c = 1$, no drift, $\sigma = 1$, and we use the fully efficient estimator so that AVA' = I. Then

$$\mathrm{Var}\,(\hat{\beta}) = \frac{1}{\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}} = \frac{1}{\sum_{f=0}^{n-1} |\tilde{s}_f|^2 |\tilde{h}_f|^2/\nu_f},$$

where $\tilde{h}_f$ and $\tilde{s}_f$ are the Fourier transforms of the haemodynamic response function and the stimulus at frequency $2\pi f/n$, and ν_f is the variance of the Fourier transform of the errors (spectrum) at that frequency, e.g. $\nu_f = 1/(1 - 2a_1 \cos(2\pi f/n) + a_1^2)$ for a (periodic) AR(1) process. For fixed total stimulus sum of squares

$$\sum_{i=1}^{n} s(t_i)^2 = \sum_{j=0}^{n-1} |\tilde{s}_j|^2,$$

(14.11) is minimized by placing all of the weight of $|\tilde{s}_j|$ at the value of j that maximizes $|\tilde{h}_j|^2/v_j$, and zero elsewhere. In other words, the optimal design should be a sine wave with frequency that maximizes the spectrum of the haemodynamic response function divided by the spectrum of the noise. Interestingly enough, this is precisely the stimulus whose estimation is unaffected by the choice of **A.**

The block design with equal on and off periods should be close to optimal since it closely matches a sine wave. For the haemodynamic response function (14.2) and an AR(1) process with $0 < p < 0.5$ at 3 s volume intervals, the optimal period of the block design is 21 to 16 s, or about 4 to 3 times the delay of the haemodynamic response. This optimal period is not changed greatly by drift removal, which mainly affects low frequency stimuli. For comparing the response between two stimuli the same result applies: the two stimuli should be presented alternatively in equal blocks with a period of 21 to 16 **s.**

For event-related designs, in which the stimulus duration is one volume or less, optimal design depends on the volume interval. For the example analysed here, the optimal design is one event every 5 to 4 volumes, or 15 to 12 s, as ρ varies between 0 and 0.5.

14.9 "Estimating the correlation structure

Getting the correct correlation structure, specified by V, is very important for three reasons: first, it guides us to the best estimator (see Section 14.4), second, it tells us how to best design the experiment, (see Section 14.8), but third, and most importantly, it leads to the correct estimator of the variance of the estimator, vital for getting the correct T or *F* statistic. In this section we look at how to estimate V, which we have so far assumed is known.

Estimating V, or the parameters such as p that make **up V,** is not as straightforward as estimating β and σ^2. There are no simple methods that give best unbiased answers; the better methods all involve costly iterative calculations that are expensive to compute. One of the simplest methods is the Cochrane–Orcutt method that first estimates β by least squares for the original unsmoothed data (14.7), that is with A = I. This estimator is always unbiased, though perhaps not the most accurate, but at least it ensures that the residuals **r** contain only error and no signal, due to the fact that the expectation of the residuals is zero, that is, averaged over all random instances of the errors. Moreover the correlation structure of the residuals is closely approximated by the matrix V that we wish to estimate.

The parameters of an AR(p) model are easily estimated from the autocorrelations of the residuals via the Yule–Walker equations, but $\mathbf{A} = \mathbf{V}^{-1/2}$ can be estimated directly from the autocorrelations (see Worsley *et al.* 2000). This is based only on an estimate of the first $(p + 1) \times (p + 1)$ elements of V, given by the sample autocorrelation out to lag p:

$$\hat{\mathbf{V}} = \mathrm{Cor}\,(\varepsilon_i, \varepsilon_j) = \frac{\sum_{l=|i-j|+1}^{n} r_l r_{l-|i-j|}}{\sum_{l=1}^{n} r_l^2}$$

$$1 \le i, j \le p+1$$

A slight bias creeps into these estimators due to the correlation of the residuals induced by removing an estimated linear effect from the observations. Typically, they are about 0.05 lower than expected. Worsley *et al.* (2000) gives a simple method of correcting this. Using the estimated A, the parameters β can be re-estimated from (14.8) and the above procedure can be iterated to convergence, but in practice just one iteration seems to be enough.

The parameters of a state space model can be estimated by using a Kalman predictor to obtain the likelihood of the parameters. This must then be maximized by iterative methods to find maximum likelihood estimators. Purdon *et al.* (1998) avoided this by estimating the white noise variance from outside the brain, where the AR(1) contribution is assumed to be zero, then estimating the AR(1) component from voxels inside the brain, assuming the white noise variance is the same as outside.

Lange and Zeger (1997) took a non-parametric approach. Noting that the Fourier transform of a stationary error sequence diagonalizes V (that is, makes the variance matrix into a diagonal matrix), they assumed that the diagonal components (the spec-

trum) is a smooth function of the frequency. Instead of fitting a model to the spectrum, they simply smoothed it avoiding the frequencies which contained the signal. This works well for a periodic stimulus, since the signal is then confined to the frequency of the signal and its higher harmonics. Taking this further, other authors have proposed simply averaging the spectrum either side of the main harmonic, in effect using linear interpolation as a form of smoother (Marchini and Ripley 2000). These approaches are much more complicated to implement in the case of non-periodic stimuli.

14.10 Spatial smoothing

so far no information has been used from neighbouring voxels, and all our models have been fitted independently at each voxel. If it is felt that the signal extends over a certain predetermined region (region of interest, ROI) in space then it can be shown that signal detection is optimal if the data is simply averaged over all voxels in that region (the ROI approach).

Since we do not usually know the location of the ROI, a reasonable compromise is to smooth the data with a kernel whose shape matches the assumed spatial activation pattern. The most common choice is a Gaussian shaped kernel. For example, if it is felt that the signal covers a 10 mm region, then the data should be smoothed with a 10 mm wide kernel (see also Chapter 12). The ROI approach can be seen as a special case in which we smooth the data with a 'box' kerriel whose shape matches the ROI.

14.10.1 Scale space

Smoothing the data has been criticised because it sacrifices resolvability for detectability. Moreover, we need to know in advance the width of the signal to be detected; smoothing with a 10 mm kernel will be optimal for 10 mm signals but less optimal for 5 or 20 mm signals. A way out of this was first proposed by Poline and Mazoyer (1994) in which a range of filter widths is used to create an extra scale dimension to the data, known as 'scale space'. To maintain constant resolution in scale space, filter widths should be chosen to be equally spaced on a log scale, e.g. 8, 12, 18, 27 mm. The data is now searched in location as well as scale, though there is a small price to pay in terms of an increase in the critical threshold of the resulting T statistics (see the end of Section 14.12.1).

14.10.2 Spatial information

Solo *et al.* (2000) have proposed a novel approach to overcoming the problem of incorporating spatial information without smoothing the data. The idea is to estimate the signal parameters without smoothing, but to estimate the noise parameters by smoothing the likelihood, not the data. An information criterion is used to set the extent of the smoothing, producing an adaptive smoother. The result is that local information is used to estimate the variability of the signal, but not the signal itself.

Worsley *et al.* (2000) took a similar approach by advocating smoothing the parameters $\alpha_1, \ldots \alpha_p$ of an AR(p) model for the noise, without smoothing the estimators of the signal β or the variance σ^2 (Figure 14.2).

SPM '99 takes this idea to the limit by averaging the AR parameters over all voxels, to produce a global estimate common to the whole brain. The robustness conferred by high frequency filtering offsets the bias in this estimator.

14.11 Estimating the haemodynamic response function

So far we have assumed a fixed parametric form for the haemodynamic response function. Although the parameters are usually reasonably well known, it is still worth estimating these parameters. For some types of experiment, the parameters themselves, such as the delay, are of intrinsic interest.

First we shall present some methods for estimating the haemodynamic response function parameters, then in Sections 14.11.1 and 14.11.2 we shall look at the cost of over and under estimating these parameters. Finally in Section 14.11.3 we shall look at non-linear alternatives to the basic convolution model (14.1).

The problem with estimating parameters of the haemodynamic response is that its parameters enter the model in a non-linear fashion, requiring time-consuming iterative estimation methods. Lange and Zeger (1997) take this route, estimating the para-

meters of a gamma model by non-linear regression techniques.

Rajapakse *et al.* (1999) instead modified the form of the haemodynamic response to make it easier to estimate. They chose a Gaussian function for the haemodynamic response, whose mean and variance parameters can then be estimated by least squares in the frequency domain. Even though the Gaussian is not a realistic model for the haemodynamic response, since its support includes negative lags, this method appears to give reasonable results.

Liao *et al.* (2001), inspired by SPM'99, have proposed linearizing the scale of the haemodynamic response by expanding $h(t)$ as a Taylor series in an unknown scale change δ:

$$e^{-\delta}h(te^{-\delta}) \approx h(t) + (-h - t\dot{h}(t))\delta,$$

where $\dot{h}(t) = \partial h(t)/\partial t$. We can then convolve the stimuli with $-h - t\dot{h}(t)$ and add these to the model, which allows for different scales for different types of stimuli. It is then possible to estimate δ from the ratio of the two coefficients *to* produce a 3D image of the delay of the haemodynamic response, and a 3D image of its standard error.

To give yet greater flexibility, SPM'99 proposes modeling the haemodynaniic response function as **a** linear combination of a set of J basis functions $b_j(t)$, j = 1, ... J that capture possible differential delays and dispersions:

$$h(t) = \sum_{j=1}^{J} \gamma_j b_j(t), \tag{14.12}$$

where γ_j, $j = 1, \ldots J$ are unknown parameters to be estimated. One such set of basis functions is a set of gamma density functions with different delays and dispersions, or a single gamma density modulated by one period of a sine wave with different frequencies.

The advantage of this approach is that the response (14.4) is still linear in the unknown parameters:

$$x(t) = \int_0^\infty h(u)\, s(t-u)\, du$$

$$= \sum_{j=1}^{J} \gamma_j \left(\int_0^\infty b_j(u)\, s(t-u)\, du \right).$$

If we allow different parameters γ_j for different stimuli, then the resulting model (14.4) is still a linear model, now in Jm instead of m parameters. This means that all the above methods can be used to rapidly estimate the parameters and test for activation.

Burock and Dale (2000) have taken this further by replacing the integral in (14.1) by a sum over the first few lags, then simply modeling the haemodynamic response by arbitrary coefficients. In effect, they propose modeling the haemodynamic response function by a linear combination of basis functions as above, with one basis function for each lag taking the value 1 at that lag and 0 elsewhere. This highly parameterized linear model is easy to estimate, but there is an attendant loss of sensitivity at detecting activation, relative to knowing the haemodynamic response exactly, which we shall discuss in the next section 14.11.1.

Finally, Genovese *et al.* (2000) have taken the most sophisticated approach. Each part of the haemodynamic response is modelled separately: the time to onset, the rate of increase, the duration of the response, the rate of decline, the undershoot, and the recovery. Priors are constructed for each of these parameters, and all the other signal and noise parameters, and the entire model is estimated by Bayesian methods using the Gibbs sampler. This makes it possible to generate the posterior distribution of any combination of parameters, though the time required for such an analysis is forbidding.

14.11.1 Over-specifying the haemodynamic response function

The reason for the loss of sensitivity when using a large number J of basis functions for the haemodynamic response is quite simple. The null model with no activation due to one stimulus has no effect of that stimulus and hence no convolution with a haemodynamic response function. It therefore has J less parameters than the model with activation, so we must use the F statistic with J and $\nu - J$ degrees of freedom to detect activation. The sensitivity of the F test decreases as J increases. It can be shown that this translates into having only about $n/J^{0.4}$ observations instead of n observations, for large n. In other words, the extra parameters dilute the effect of the activation, making it harder *to* detect.

However, it must be remembered that a haemodynamic response function with too few parameters may be biased, resulting also in a loss of sensitivity because

it will fail to capture all of the response. Obviously one should try to strike a balance between too few parameters to adequately capture the response, and too many parameters that overfits the response.

The lesson is that the more flexibility allowed for the response, the more difficult it is to detect it. The best strategy is to try to model the haemodynamic response with a small number of well chosen basis functions, or preferably just one basis function. These comments apply equally well to the non-linear models in Section 14.11.3.

14.11.2 Misspecifying the haemodynamic response function

What is the cost of mis-specifying the haemodynamic response function? First, there is no effect at all on the validity of the analysis. P-values for detecting pure activation are still correct even if the haemodynamic response is wrong because they are based on the null model in which there is no activation and hence no haemodynamic response. However, it is still important to get the haemodynamic response correct when comparing activations, because now the null model does contain a haemodynamic response, equal for the conditions to be compared.

The main cost of misspecifying the haemodynamic response is a loss of sensitivity. This is more pronounced for event-related designs than for block designs, because the block stimulus with long blocks is less affected by convolution with the haemodynamic response function. In fact some stimuli are completely unaffected by convolution with the haemodynamic response function. One such is the sine wave stimulus with arbitrary amplitude and phase, that is, a linear combination of sine and cosine with the same known frequency ω:

$$s(t) = \beta_1 \sin(\omega t) + \beta_2 \cos(\omega t).$$

Convolution of $s(t)$ with any haemodynamic response function changes β_1 and β_2 but leaves the form of the model unchanged. This means that for this design, there is no cost to misspecifying the haemodynamic response — in fact it can be ignored altogether.

14.11.3 Non-linear haemodynamic response and stimulus non-additivity

The linearity of the haemodynamic response, and hence the additivity of signals closely separated in time, has been questioned by several authors. Is the response always a simple convolution of stimuli with a haemodynamic response function? Friston et al. (1998b) have addressed this by expanding the haemodynamic convolution itself as a set of Volterra kernels. The second-order model is:

$$x(t) = \int_0^\infty h_1(u_1)s(t-u_1)\,du_1 + \int_0^\infty\int_0^\infty h_2(u_1,u_2)s(t-u_1)\,s(t-u_2)\,du_1\,du_2. \quad (14.13)$$

The first term is the simple convolution model (14.1) in which past stimuli have a linear effect on the current response. The second term is the second-order Volterra kernel in which past stimuli have a quadratic (including interactions) effect on the current response. (Note that without loss of generality h_2 is symmetric: $h_2(u_1, u_2) = h_2(u_2, u_1)$.) In other words, this model allows for the possibility that the effect of stimuli may not be purely additive; the response to two stimuli in close succession may be different from the sum of the separate responses if the two stimuli are far apart in time.

It might be possible to estimate the second-order kernel by extending the method of Burock and Dale (2000) (Section 14.11). The integrals in (14.13) could be replaced by summations over the first few lags, and the discrete kernels become arbitrary unknown parameters. The result is once again a large linear model including linear and quadratic terms in the first few lags of the stimulus. However, the large number of parameters to be estimated might make this method prohibitive.

A more practical suggestion, due to Friston *et* al. (1998b), is to model the first and second order kernels by a linear combination of a small number of basis functions $b_j(t), j = 1, \ldots J$, extending (14.12):

$$h_1(u_1) = \sum \gamma_j b_j(u_1),$$

$$h_2(u_1, u_2) = \sum_{j=1}^{J} \sum_{k=1}^{J} \gamma_{jk} b_j(u_1) b_k(u_2)$$

where γ_j, γ_{jk}, $1 \leqslant j \leqslant k \leqslant J$ are unknown parameters to be estimated. The convolution of each basis function with the stimulus is $z_j(t) = \int b_j(u) s(t-u)$, so that the response becomes:

$$x(t) = \sum_{j=1}^{J} \gamma_j z_j(t) + \sum_{j=1}^{J} \sum_{k=j}^{J} \gamma_{jk} z_j(t) z_k(t),$$

which is once again linear in the unknown parameters, so it can be fitted by the linear models methods above. Linearity of the haemodynamic response and stimulus additivity can now be tested by an F statistic for the bivariate terms γ_{jk}, $1 \leqslant j \leqslant k \leqslant J$ as in Section 14.6.

14.12 Detecting an effect at an unknown location

In this section we shall look at the question of detecting an effect $c'\beta$ or activation ($c'\beta > 0$) at an unknown spatial location, rather than at a known location as in Section 14.6. Very often we do not know in advance where to look for an effect, and we are interested in searching the whole brain, or part of it. This presents special statistical problems related to the problem of multiple comparisons, or multiple tests. Two methods have been proposed, the first based on the maximum of the T or F statistic, the second based on the spatial extent of the region where these statistics exceed some threshold value. Both involve results from random field theory (Adler 1981).

14.12.1 The maximum test statistic

An obvious method is to select those locations where a test statistic Z (which could alternatively be the T statistic or F statistic of Section 14.6) is large, that is, to threshold the image of Z at a height z. The problem is then to choose the threshold z to exclude false positives with a high probability, say 0.95. Setting z to the usual (uncorrected) $P = 0.05$ critical value of Z (1.64 in the Gaussian case) means that 5 per cent of the unactivated parts of the brain will show false positives. We need to raise z so that the probability of finding any activation in the non-activated regions is 0.05. This is a type of multiple comparison problem, since we are testing the hypothesis of no activation at a very large number of voxels.

A simple solution is to apply a Bonferroni correction. The probability of detecting any activation in the unactivated locations is bounded by assuming that the unactivated locations cover the entire search region. By the Bonferroni inequality, the probability of detecting any activation is further bounded by

$$\mathrm{P}(\max Z > z) \leqslant N\,\mathrm{P}(Z > z), \tag{14.14}$$

where the maximum is taken over all N voxels in the search region. For a $P = 0.05$ test of Gaussian statistics, critical thresholds of 4–5 are common. This procedure is conservative if the image is smooth, although for fMRI data it often gives very accurate thresholds.

Random field theory gives a less conservative (lower) P-value if the image is smooth. As with time series analysis, if the statistic image is smooth, then there are less truly independent voxels than the original voxel count. Thus the N used above should be reduced to the correct number of independent voxels, giving less conservative thresholding. The smoothness of the statistic image is estimated and a 'resel' size is derived, where a resel is larger than a voxel and represents the size of 'independent voxels'. The resulting thresholding is thus:

$$\mathrm{P}(\max Z > z) \approx \sum_{d=0}^{D} \mathrm{Resels}_d \mathrm{EC}_d(z) \tag{14.15}$$

where D is the number of dimensions of the search region, Resels_d is the number of d-dimensional resels (resolution elements) in the search region, and $\mathrm{EC}_d(z)$ is the d-dimensional Euler characteristic density. The approximation (14.15)is based on the fact that the left hand side is the exact expectation of the Euler characteristic of the region above the threshold z. The Euler characteristic counts the number of clusters if the region has no holes, which is likely to be the case if z is large. Details can be found in Worsley *et* al. (1996*a*).

The approximation (14.15) is accurate for search regions of any size or shape, even a single point, but it is best for search regions that are not too concave. Sometimes it is better to surround a highly convoluted search region, such as the cortical surface, by a convex hull with slightly higher volume but less surface area, to get a lower and more accurate P-value.

For large search regions, the last term $(d = 3)$ is the most important. The number of resels is

$$\text{Resels}_3 = V/\text{FWHM}^3,$$

where V is the volume of the search region and FWHM is the effective full width at half maximum of a Gaussian kernel used to smooth the data. The corresponding EC density for a T statistics image with ν degrees of freedom is

$$\text{EC}_3(z) = \frac{(4 \log_e 2)^{3/2}}{(2\pi)^2} \left(\frac{\nu - 1}{\nu} z^2 - 1\right) \left(1 + \frac{z^2}{\nu}\right)^{-(1/2)(\nu-1)}$$

For small search regions, the lower dimensional terms $d < 3$ become important. However, the P-value (14.15) is not very sensitive to the shape of the search region, so that assuming a spherical search region gives a very good approximation. In practice, it is better to take the minimum of the the two P-values (14.14) and (14.15). Figure 14.3 shows the T statistic thresholded at the $P = 0.05$ value of $z = 4.86$, found by equating (14.15) to 0.05 and solving for z.

Extensions of the result (14.15) to scale space random fields are given in Worsley *et al.* (1996*b*). Here the search is over all spatial filter widths as well over location, so that the width of the signal is estimated as well as its location. The price to pay is an increase in critical threshold of about 0.5.

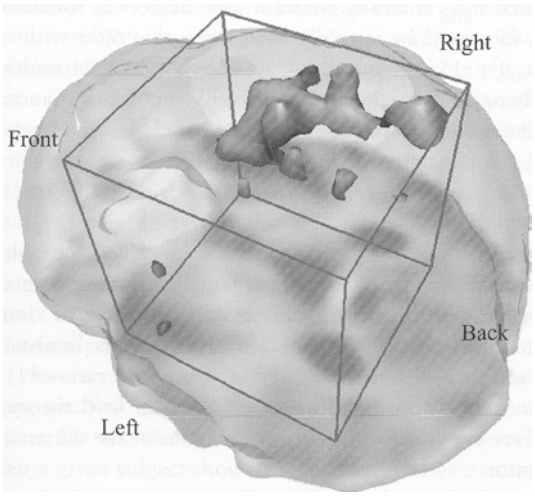


Fig. 14.3 The T statistic thresholded at the $P = 0.05$ value of 4.86.

14.12.2 The maximum spatial extent of the test statistic

An alternative test can be based on the spatial extent of clusters of connected components of supra threshold voxels where $Z > z$ (Friston *et al.*, 1994). Typically z is chosen to be about 3 for a Gaussian random field. Once again the image must be a smooth stationary random field. The idea is to approximate the shape of the image by a quadratic with a peak at the local maximum. For a Gaussian random field, it can be shown that the second spatial derivative of this quadratic is well approximated by $Z = -z\mathbf{\Lambda}$, where $\mathbf{\Lambda} = \text{Var}(\dot{Z})$, for large z. The spatial extent S is then approximated by the volume of the quadratic of height H above z:

$$S \approx cH^{D/2},$$

where

$$c = \frac{(2\pi/z)^{D/2}}{\det(\mathbf{\Lambda})^{1/2}\Gamma(D/2 + 1)} \tag{14.16}$$

For large z, the upper tail probability of H is well approximated by

$$P(H > h) = \frac{\text{P}(\max Z > z + h)}{\text{P}(\max Z > z)} \approx \exp(-zh), \tag{14.17}$$

from which we conclude that H has an approximate exponential distribution with mean $1/z$. From this we can find the approximate P-value of the spatial extent S of a single cluster:

$$\text{P}(S > s) \approx \exp(-z(s/c)^{2/D}). \tag{14.18}$$

The P-value for the largest spatial extent is obtained by a simple Bonferroni correction for the expected number of clusters N:

$$\text{P}(\max S > s) \approx \text{E}(N)\ \text{P}(S > s),$$

$$\text{where } \text{E}(N) \approx \text{P}(\max Z > z) \tag{14.19}$$

from (14.15).

We can substantially improve the value of the constant c by equating the expected total spatial extent, given by $V\,\mathrm{P}(Z > z)$, to that obtained by summing up the spatial extents of all the clusters $S_1, \ldots S_N$:

$$V\,\mathrm{P}(Z > z) = \mathrm{E}(S_1 + \ldots + S_N) = \mathrm{E}(N)\,\mathrm{E}(S).$$

Using the fact that

$$\mathrm{E}(S) \approx c\Gamma(D/2 + 1)/z^{D/2}$$

from (14.16), it follows that

$$c \approx \frac{\mathrm{FWHM}^D z^{D/2} \mathrm{P}(Z > z)}{\mathrm{EC}_D(z)\,\Gamma(D/2 + 1)}$$

Cao (1999) has extended these results to T and F fields, but unfortunately there are no theoretical results for non-smooth fields such as raw fMRI data.

14.12.3 Searching in small regions

For small pre-specified search regions such as the cingulate, the P-values for the maximum test statistic are very well estimated by (14.15), but the results in section 14.12.2 only apply to large search regions. Friston (1997) has proposed a fascinating method that avoids the awkward problem of pre-specifying a small search region altogether. We threshold the image of test statistics at z, then simply pick the nearest peak to a point or region of interest. The clever part is this. Since we have identified this peak based only on its spatial location and not based on its height or extent, there is now no need to correct for searching over all peaks. Hence, the P-value for its spatial extent S is simply $\mathrm{P}(S > s)$ from (14.18), and the P-value for its peak height H above z is simply $\mathrm{P}(H > h)$ from (14.17).

14.13 Multiple runs, sessions, and subjects

fMRI experiments are often repeated for several runs in the same session, several sessions on the same subject, and for several subjects drawn from a population. We shall assume that all the images have been aligned to a common stereotactic space (see Chapter 15), so that anatomical variability is not a problem. Nevertheless, there remains a very different sort of statistical problem.

It has long been recognized that a simple fixed effects analysis, in which we assume that the signal strength β is identical in all runs, sessions and subjects, is incorrect (Holmes and Friston, 1998). A random effects analysis seems the most appropriate, in which the error of the effect is calculated from independent repetitions, not from the noise error σ. Unfortunately this leads to an awkward practical problem: usually the number of repetitions (runs, sessions, subjects) is small, so the available degrees of freedom is small. For most purposes this would not be too serious, but in brain mapping we are often looking in the extreme tails of the distribution, where low degrees of freedom give very large critical thresholds for maximum test statistics, which substantially reduces the sensitivity of detecting any activation. Added to this is the problem of the Gaussian assumption for the errors; although the Central Limit Theorem assures good normality for test statistics, it is not clear that normality is maintained far into the tails of the distribution.

In PET data, degrees of freedom can be increased by spatially smoothing the random effects variance to produce a global estimate for the entire brain. Unfortunately this cannot be done for fMRI data because the variance is much too spatially structured. Instead, Worsley *et al.* (2000) assume that the ratio of random effects variance $\sigma^2_{\mathrm{random}}$ to fixed effects variance $\sigma^2_{\mathrm{fixed}}$ is locally constant. The degrees of freedom is increased by spatially smoothing this ratio with a $\omega_{\mathrm{ratio}} = 15$ mm FWHM Gaussian kernel, then multiplying back by the unsmoothed fixed effects variance. The residual variance is then estimated by

$$\sigma^2_{\mathrm{residual}} = \sigma^2_{\mathrm{fixed}} \mathrm{smooth}(\sigma^2_{\mathrm{random}}/\sigma^2_{\mathrm{fixed}}).$$

The result is a slightly biased but much less variable estimate of the variance of an effect, that comes midway between a random effects analysis (no smoothing, $\omega_{\mathrm{ratio}} = 0$) and a fixed effects analysis (complete smoothing, $\omega_{\mathrm{ratio}} = \infty$, to a global ratio of 1).

A simple formula, based on random field theory, gives the effective degrees of freedom of the variance ratio:

$$\nu_{\mathrm{ratio}} = \nu_{\mathrm{random}}(2(\omega_{\mathrm{ratio}}/\omega_{\mathrm{data}})^2 + 1)^{3/2},$$

where ν_{random} is the random effects degrees of freedom and ω_{data} is the FWHM of the fMRI signal, usually taken to be that of the raw data (typically 6 mm). The final effective degrees of freedom of the residuals, $\nu_{residual}$, is estimated by

$$1/\nu_{residual} = 1/\nu_{ratio} + 1/\nu_{fixed},$$

where ν_{fixed} is the fixed effects degrees of freedom. In practice we choose the amount of smoothing ω_{ratio} so that the final degrees of freedom $\nu_{residual}$ is at least 100, ensuring that errors in its estimation do not greatly affect the distribution of test statistics.

14.13.1 Conjunctions

An alternative method of dealing with multiple subjects is through conjunctions. A conjunction is simply the locations where all the subjects' test statistics exceed a fixed threshold (Friston *et al.*, 1999*a*). We are interested in the P-value of this event if in fact there is no activation for any of the subjects, which is equivalent to the P-value of the maximum (over location) of the minimum (over subjects) of the test statistic images. There is a neat formula for this based on random field theory (Worsley and Friston 2000).

It is useful to compare this with the above regularized random effects analysis. As it stands, conjunction analysis is still a fixed effects analysis, since the distribution of the test statistic is based on errors estimated within subjects, rather than between subjects. The random effects analysis assumes that there is an effect for each subject that is zero when averaged over all subjects. In other words, the random effects analysis is using a much weaker null hypothesis than the fixed effects analysis; the random effects analysis assumes that there is an effect, but this effect is randomly distributed about zero; the fixed effects analysis demands in addition that the variability of this random effect is zero, forcing the effect on each subject to be identically zero.

However, Friston *et al.* (1999*a*) turns the conjunction analysis into a neat test for a type of random effect. He asks the following question: suppose we say that a given subject shows an effect if it passes a usual $P = 0.05$ test based on a fixed effect; what is the probability that all subjects will show this type of effect in some small region (i.e. a conjunction), if in fact a proportion γ do, and the rest do not? The paper then gives a lower bound for γ, based on the data, such that the true γ is larger than the lower bound with a probability of at least *0.95*. In other words, we obtain a type of (conservative) confidence interval for the proportion of subjects that show a fixed effect.

14.14 Conclusion

This chapter has presented a review of methods for setting up a model for fMRI data, described how to estimate the parameters of this model, and how to assess the errors in these estimates. It obviously presupposes that the experimenter knows quite a lot about how and when the stimulus affects the BOLD response, but it does not suppose that we know which regions of the brain are affected. Thresholding and looking at cluster size of activated regions, will detect those regions that are affected above background noise.

There are other approaches that make far fewer assumptions about the time course of the expected BOLD response. Most of these are based on some sort of decomposition of the data into time courses and spatial patterns that are uncorrelated (singular value decomposition (SVD), principal components analysis (PCA)), or independent (independent components analysis (ICA)). These methods can be extremely useful at suggesting or generating hypotheses that can be captured and confirmed by the models presented in this chapter.

The idea of using a hypothesis test to detect activated regions does contain a fundamental flaw that all experimenters should be aware of. Think of it this way: if we had enough data, T statistics would increase (as the square root of the number of time points or subjects) until *all* voxels were 'activated'! In reality, *every* voxel must be affected by the stimulus, perhaps by a very tiny amount; it is impossible to believe that $\beta = 0.000000000$ exactly. So thresholding simply excludes those voxels where we do not yet have enough evidence to distinguish their effects from zero. If we had more evidence, perhaps with better scanners, or simply more time points, we would surely be able to do so. But then we would probably not want to detect activated regions. As for satellite images, the job for statisticians would then be signal *enhancement* rather

than signal detection. The distinguishing feature of our fMRI data is that there is so little signal to enhance. Even with the advent of better scanners this is still likely to be the case, because neuroscientists will surely devise yet more subtle experiments that will push the signal to the limits of detectability.

References

Adler, R.J. *(1981).The geometry of random fields.* Wiley, New York.

Caines, P.E. *(1988). Linear stochastic systems.* Wiley, New York.

Cao, J. *(1999).*The size of the connected components of excursion sets of χ^2, t and F fields. *Advances in Applied Probability,* **31**, 577–93.

Burock, M.A., and Dale, A.M. *(2000).*Estimation and detection of event-related fMRI signals with temporally correlated noise: a statistically efficient unbiased approach. *Human Brain Mapping,* **11**(4), 249–60.

Friston, K.J. *(1997).*Testing for anatomically specified regional effects. *Human Brain Mapping, 5,* 133–6.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., and Evans, A.C. *(1994).*Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping,* **1**, 214–20.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J-B., Frith, C.D., and Frackowiak, R.S.J. *(1995).*Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping, 2,* 189–210.

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A.P., Rugg, M.D., and Turner, R. (1998*a*). Event-related fMRI: Characterising differential responses. *Neurolmage,* **7**, 30–40.

Friston, K.J., Josephs, O., Rees, G., and Turner, R. (1998*b*). Non-linear event-related responses in fMRI. *Magnetic Resonance in Medicine, 39, 41–52.*

Friston, K.J., Holmes, A.P., Price, C.J., Buchel, C., and Worsley, K.J. (1999*a*). Multi-subject fMRI studies and conjunction analyses. *NeuroImage, 10,* 385–96.

Friston, K.J., Zarahn, E., Josephs, O., Henson, R.N., and Dale, A.M. *(19996).* Stochastic designs in event-related fMRI. *Neurolmage,* **10**, 607–19.

Friston, K.J., Josephs, O., Zarahn, E., Holmes, A.P., Rouquette, S. and Poline, J.-B. *(2000).*To smooth or not to smooth: Bias and efficiency in fMRI time series analysis. *Neurolmage, 12, 196–208.*

Genovese, C.R. (2000).A Bayesian time-course model for functional magnetic resonance imaging data (with discussion). *Journal of the American Statistical Association,* **95**, 691–719.

Glover, G.H. *(1999).* Deconvolution of impulse response in event-related BOLD fMRI. *Neurolmage,* **9**, 416–29.

Holmes, A.P., and Friston, K.J. *(1998).*Generalizability, random effects, and population inference. *Neurolmage, 7,* S754.

Lange, N. and Zeger, S.L. *(1997).* Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging (with discussion). *Applied Statistics, 46, 1–29.*

Liao, C., Worsley, K.J., Poline, J-B., Duncan, G.H., and Evans, A.C. *(2001).* Estimating the delay of the haemodynamic response of fMRI data. *Neurolmage,* **13**, S185.

Marchini, J.L. and Ripley, B.D. *(2000).* A new statistical approach to detecting significant activation in functional MRI. *Neurolmage,* **12**, 366–80.

Poline, J-B., and Mazoyer, B.M. *(1994).*Enhanced detection in activation maps using a multifiltering approach. *Journal of Cerebral Blood* Flow *and Metabolism 14,* 690–9.

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

Rajapakse, J.C., Kruggel, F., Maisog, J.M., and von Cramon, D.Y. *(1998).*Modeling haemodynamic response for analysis of functional MRI time-series. *Human Brain Mapping, 6,* 283–300.

Solo, V., Purdon, P., Brown, E., and Weisskoff, R. *(2001).*A signal estimation approach to functional MRI. *IEEE Transactions on Medical Imaging,* **20**(1), *26–35.*

Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., and Evans, A.C. (1996*a*). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping, 4, 58–73.*

Worsley, K.J., Marrett, S., Neelin, P., and Evans, A.C. (1996*b*). Searching scale space for activation in PET images. *Human Brain Mapping, 4, 74–90.*

Worsley, K.J., Liao, *C.,* Grabove, M., Petre, V., Ha, B., and Evans, A.C. *(2000).*A general statistical analysis for fMRI data. *Neurolmage, 11,* S648.