1 Generalised Inverse

For GLM

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

where X is a $N \times k$ design matrix and $p(e) = \mathsf{N}(0, \sigma^2 I_N)$, we can estimate the coefficients from the normal equations

$$(X^T X)\beta = X^T y \tag{2}$$

If rank of X, denoted $r(X)$, is k (ie. full rank) then $X^T X$ has an inverse (it is 'nonsingular') and

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

But if $r(x) < k$ we can have $X\beta_1 = X\beta_2$ (ie. same predictions) with $\beta_1 \neq \beta_2$ (different parameters). The parameters are then not therefore 'unique', 'identifiable' or 'estimable'.

For example, a design matrix sometimes used in the

Analysis of Variance (ANOVA)

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \tag{4}$$

has $k = 3$ columns but rank $r(X) = 2$ ie. only two linearly independent columns (any column can be expressed as a linear combination of the other two).

For models such as these $X^T X$ is not invertible, so we must resort to the *generalised inverse*, X^-. This is defined as any matrix X^- such that $XX^-X = X$. It can be shown that in the general case

$$\begin{aligned} \hat{\beta} &= (X^T X)^- X^T y \\ &= X^- y \end{aligned} \tag{5}$$

If X is full-rank, $X^T X$ is invertible and $X^- = (X^T X)^{-1} X^T$.

There are many generalise inverses. We would often choose the pseudo-inverse (pinv in MATLAB)

$$\hat{\beta} = X^{+}y \tag{6}$$

Take home message: avoid rank-deficient designs. If X is full rank, then $X^{+} = X^{-} = (X^{T}X)^{-1}X^{T}$.

2 Estimating error variance

An *unbiased* estimate for the error variance σ^2 can be derived as follows. Let

$$X\hat{\beta} = Py \tag{7}$$

where P is the *projection matrix*

$$\begin{aligned} P &= X(X^{T}X)^{-}X^{T} \\ &= XX^{-} \end{aligned} \tag{8}$$

Py projects the data y into the space of X. P has two important properties (i) it is symmetric $P^{T} = P$, (ii)

PP=P. This second property follows from it being a projection. If what is being projected is already in X space (ie. Py) then looking for that component of it that is in X space will give the same thing ie. $PPy = Py$.

Then residuals are

$$\begin{aligned} \hat{e} &= y - X\hat{\beta} \\ &= (I - P)y \\ &= Ry \end{aligned} \tag{9}$$

where $R = I_N - XX^-$ is the *residual-forming* matrix. Remember, $\hat{e}$ is that component of the data, orthogonal to the 'space' X. Ry is another projection matrix, but one that projects the data y into the orthogonal complement of X. Similarly, R has the two properties (i) $R^T = R$ and (ii) $RR = R$.

We now look seek an unbiased estimator of the variance by first looking at the expected sum of squares

$$\begin{aligned} E[\hat{e}^T\hat{e}] &= E[y^T R^T Ry] \\ &= E[y^T Ry] \end{aligned} \tag{10}$$

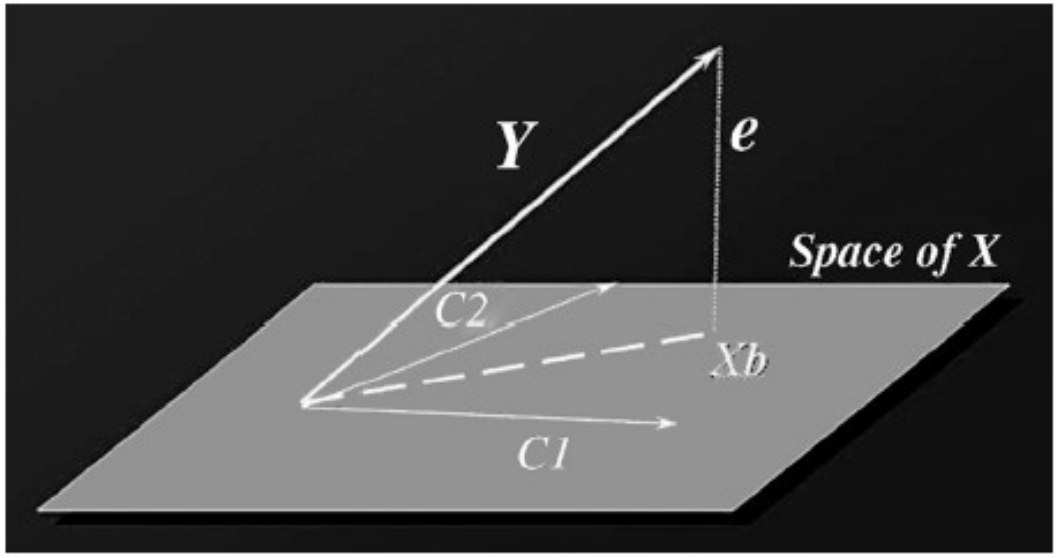


FIGURE 9.15 Geometrical perspective: estimation. The data Y are projected orthogonally onto the space of the design matrix (X) defined by two regressors $C1$ and $C2$. The error e is the distance between the data and the smallest possible within the model space.

We now use the standard result: If $p(a) = N(\mu, V)$ then

$$E[a^T B a] = \mu^T B \mu + Tr(BV)$$

So, if $p(y) = N(X\hat{\beta}, \sigma^2 I_N)$ then

$$\begin{aligned} E[y^T R y] &= \hat{\beta}^T X^T R X \hat{\beta} + Tr(\sigma^2 R) \\ &= \hat{\beta}^T (X^T X - X^T X X^- X)\hat{\beta} + Tr(\sigma^2 R) \\ &= Tr(\sigma^2 (I - P)) \\ &= \sigma^2 (N - r(P)) \\ &= \sigma^2 (N - k) \end{aligned} \tag{11}$$

So, an *unbiased* estimate of the variance is

$$\begin{aligned} \hat{\sigma^2} &= (y^T R y)/(N - k) \\ &= RSS/(N - k) \end{aligned} \tag{12}$$

where the RSS is 'Residual Sum of Squares'. Remember,

the ML variance estimate is

$$\hat{\sigma^2}_{ML} = (y^T R y)/N \tag{13}$$

3 Comparing nested GLMs

Full model:

$$y = X_0\beta_0 + X_1\beta_1 + e \tag{14}$$

Reduced model:

$$y = X_0\beta_0 + e_0 \tag{15}$$

Consider the test-statistic

$$f = \frac{(RSS_{red} - RSS_{full})/(k-p)}{RSS_{full}/(N-k)} \tag{16}$$

where 'Residual Sum of Squares (RSS)' are

$$RSS_{full} = \hat{e}^T\hat{e} \tag{17}$$

$$RSS_{red} = \hat{e}_0^T\hat{e}_0 \tag{18}$$

We can re-write in terms of 'Extra Sum of Squares'

$$f = \frac{ESS/(k-p)}{RSS_{full}/(N-k)} \tag{19}$$

where

$$ESS = RSS_{red} - RSS_{full} \tag{20}$$

We can compute these quantities using

$$\begin{aligned} RSS_{full} &= y^T R y \\ RSS_{red} &= y^T R_0 y \end{aligned} \tag{21}$$

We expect the denominator to be

$$E[RSS_{full}/(N-k)] = \sigma^2 \tag{22}$$

and, under the null ($\beta_1 = 0$), we have $\sigma_0^2 = \sigma^2$ and therefore expect the numerator to be

$$E[(RSS_{red} - RSS_{full})/(k-p)] = \sigma^2 \tag{23}$$

where $r(R_0 - R) = k - p$ (mirroring the earlier expectation calculation). Under the null, we therefore expect a

test statistic of unity

$$< f >= \frac{\sigma^2}{\sigma^2} \tag{24}$$

as both numerator and denominator are unbiased estimates of error variance. We might naively expect to get a numerator of zero, under the null. But this is not the case because, in any finite sample, ESS will be non zero. When we then divide by $(k-p)$ we get $E[ESS/(k-p)] = \sigma^2$.

When the full model is better we get a larger f value.

4 Partial correlation and R^2

The square of the partial correlaton coefficient

$$R^2_{y,X_1|X_0} = \frac{RSS_{red} - RSS_{full}}{RSS_{red}} \tag{25}$$

is the (square) of the correlation between y and $X_1\beta_1$ after controlling for the effect of $X_0\beta_0$. Abbreviating the

above to R^2, the F-statistic can be re-written as

$$f = \frac{R^2/(k-p)}{(1-R^2)/(N-k)} \tag{26}$$

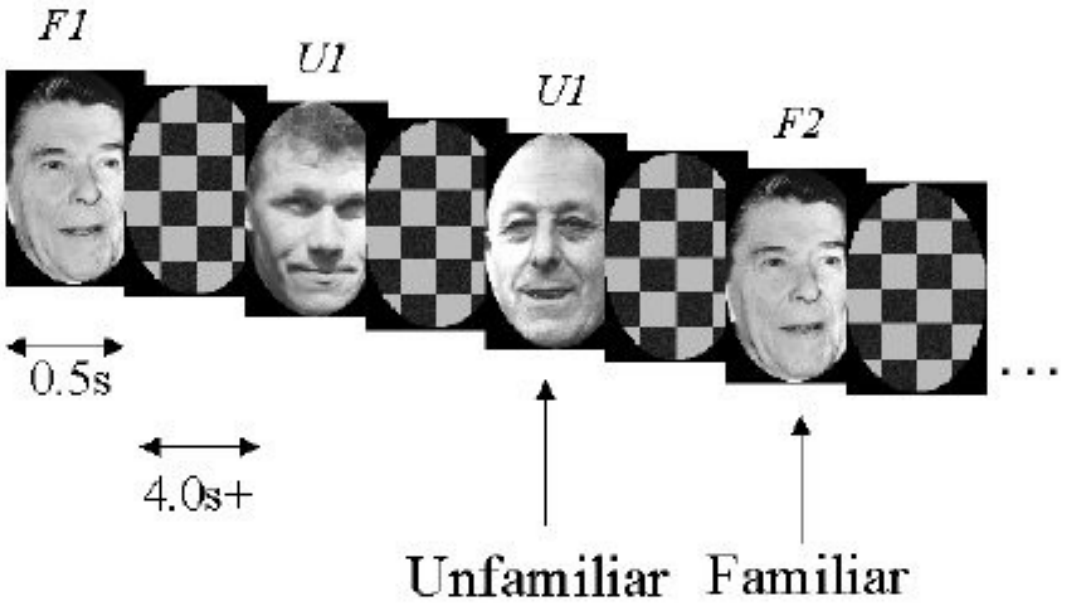
Model comparison tests are identical to tests of partial correlation.

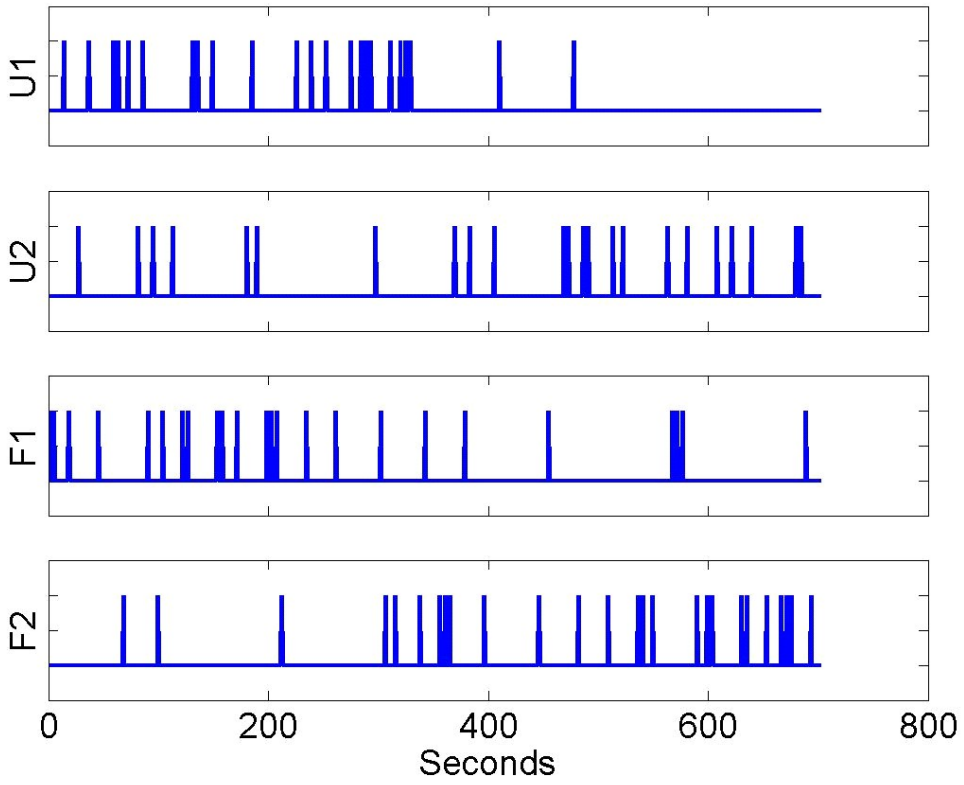
In X_0 explains no variance eg. it is a constant or empty matrix then

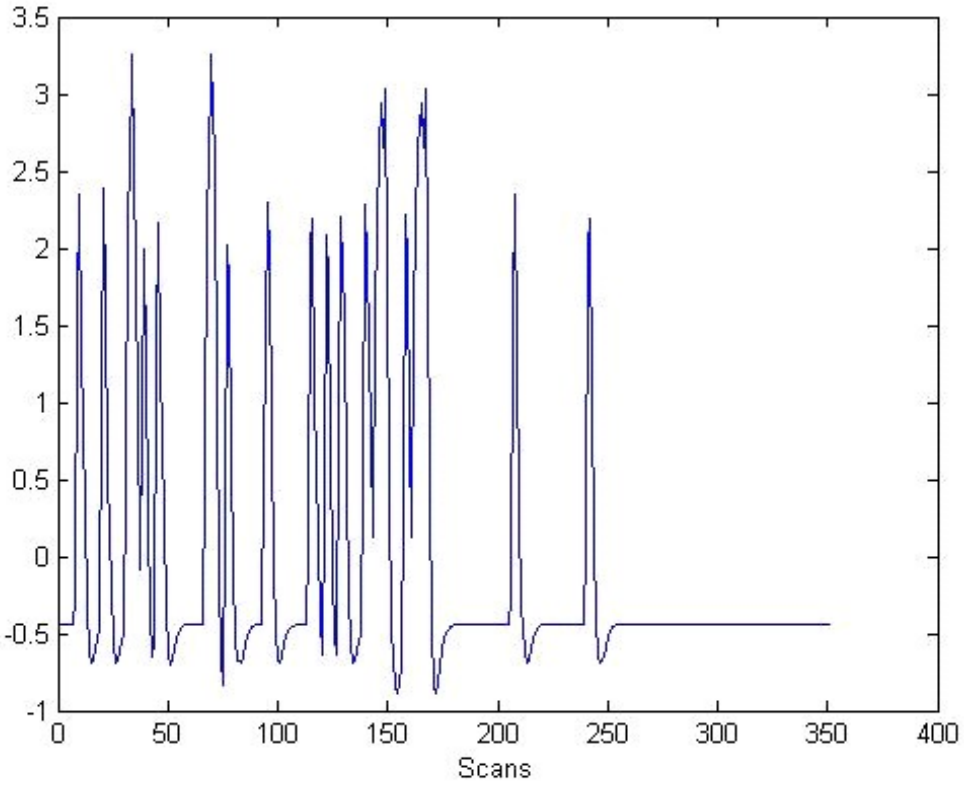
$$R^2 = \frac{Y^T Y - Y^T R Y}{Y^T Y} \tag{27}$$

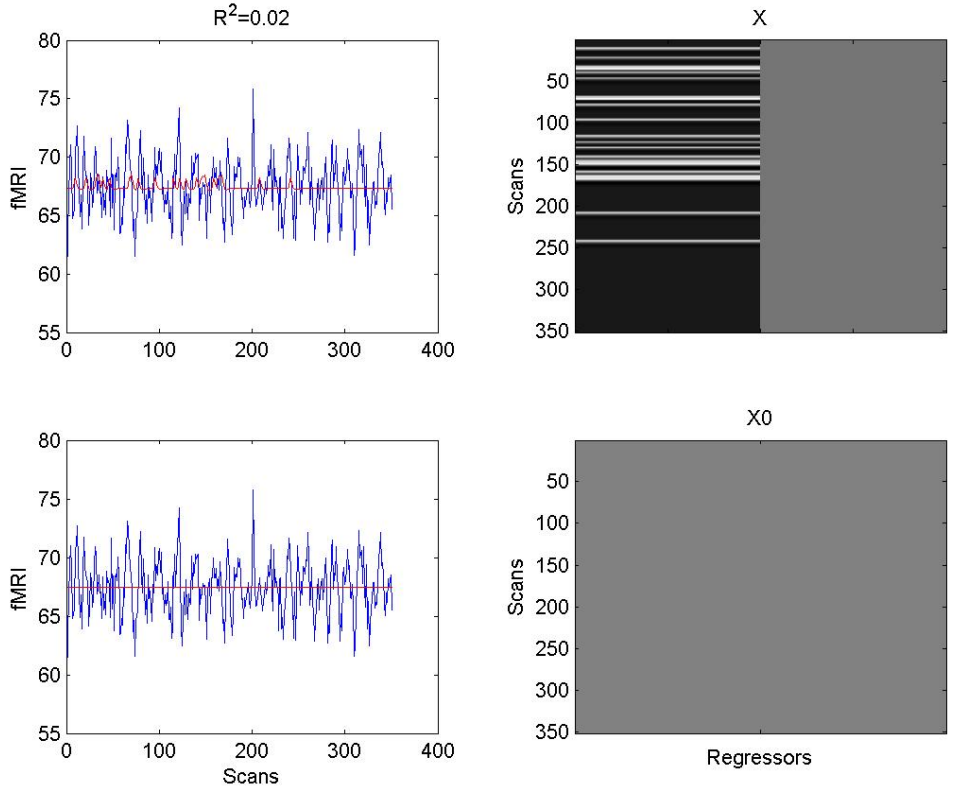
which is the proportion of variance explained by the model with design matrix X. More generally, if X_0 is not the empty matrix then R^2 is that proportion of the variability unexplained by the reduced model X_0 that is explained by the full model X.

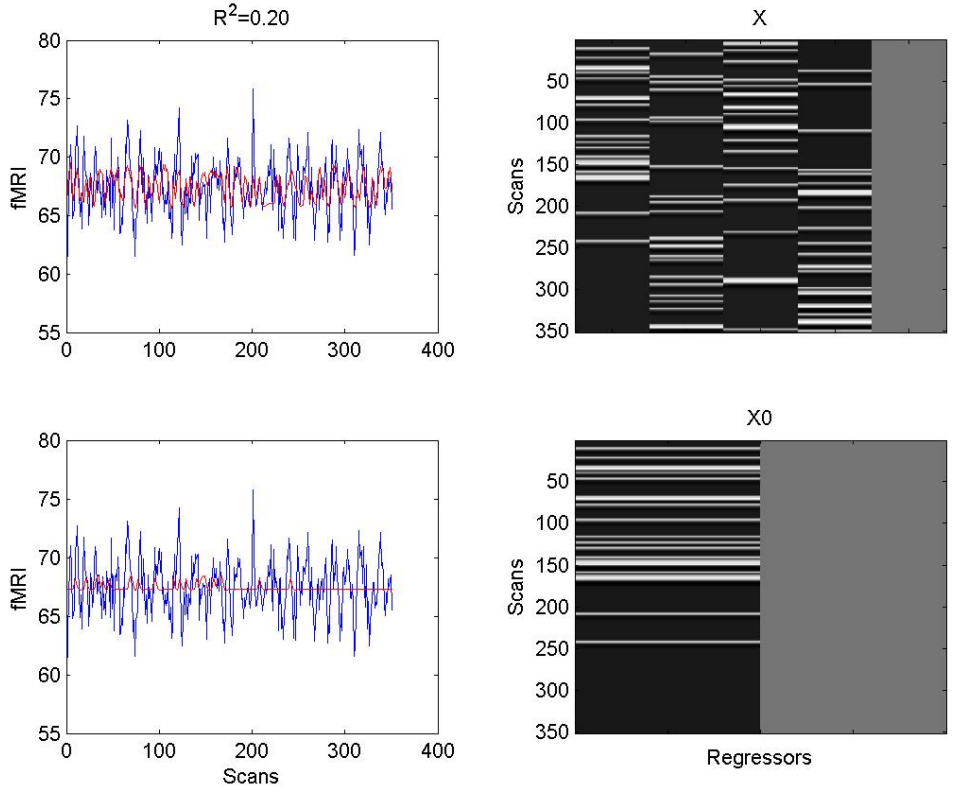
5 Examples

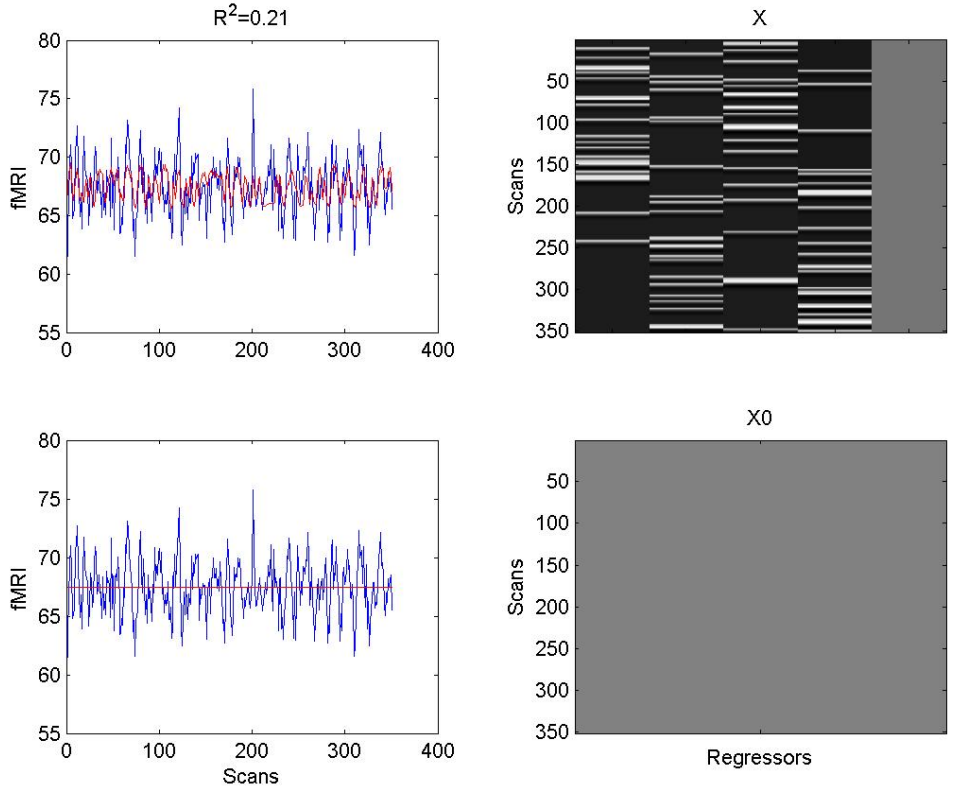

F1
U1
U1
F2
0.5s
4.0s+
Unfamiliar
Familiar


U1
U2
F1
F2
0
200
400
600
800
Seconds


3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
0
50
100
150
200
250
300
350
400
Scans


R2=0.02
fMRI
Scans
X
X0
Regressors


R2=0.20
fMRI
Scans
X
X0
Regressors


R2=0.21
fMRI
80
75
70
65
60
55
0
100
200
300
400
X
Scans
50
100
150
200
250
300
350
fMRI
Scans
X0
Scans
Regressors

6 How large must f be for a 'significant' improvement ?

Under the null ($\beta_1 = 0$), f follows an F-distribution with $k - p$ numerator degrees of freedom (DF) and $N - k$ denominator DF.

Info on PDFs and transforming them.

7 Contrasts

We can also compare nested models using *contrasts*. This is more efficient, as we only need to estimate parameters of the *full* model.

For a contrast matrix C we wish to test the hypothesis $C^T\beta = 0$. This can correspond to a model comparison, as before, if C is chosen appropriately. But it is also more general, as we can test any effect which can be expressed as

$$C^T\beta = H^T X\beta \tag{28}$$

for some H. This defines a space of estimable contrasts.

The contrast C defines a subspace $X_c = XC$. As before, we can think of the hypothesis $C^T\beta = 0$ as comparing a full model, X, versus a reduced model which is now given by $X_0 = XC_0$ where C_0 is a contrast orthogonal to C ie.

$$C_0 = I_k - CC^- \tag{29}$$

A test statistic can then be generated as before where

$R_0 = I_N - X_0 X_0^-$, $M = R_0 - R$ and

$$f = \frac{y^T M y / r(M)}{y^T R y / r(R)} \tag{30}$$

In fMRI, the use of contrasts allows us to test for (i) main effects and interactions in factorial designs, (ii) choice of hemodynamic basis sets. Importantly, we do not need to refit models.

The numerator can be calculated efficiently as

$$y^T M y = \hat{c}^T \left[C^T (X^T X)^- C \right]^- \hat{c} \tag{31}$$

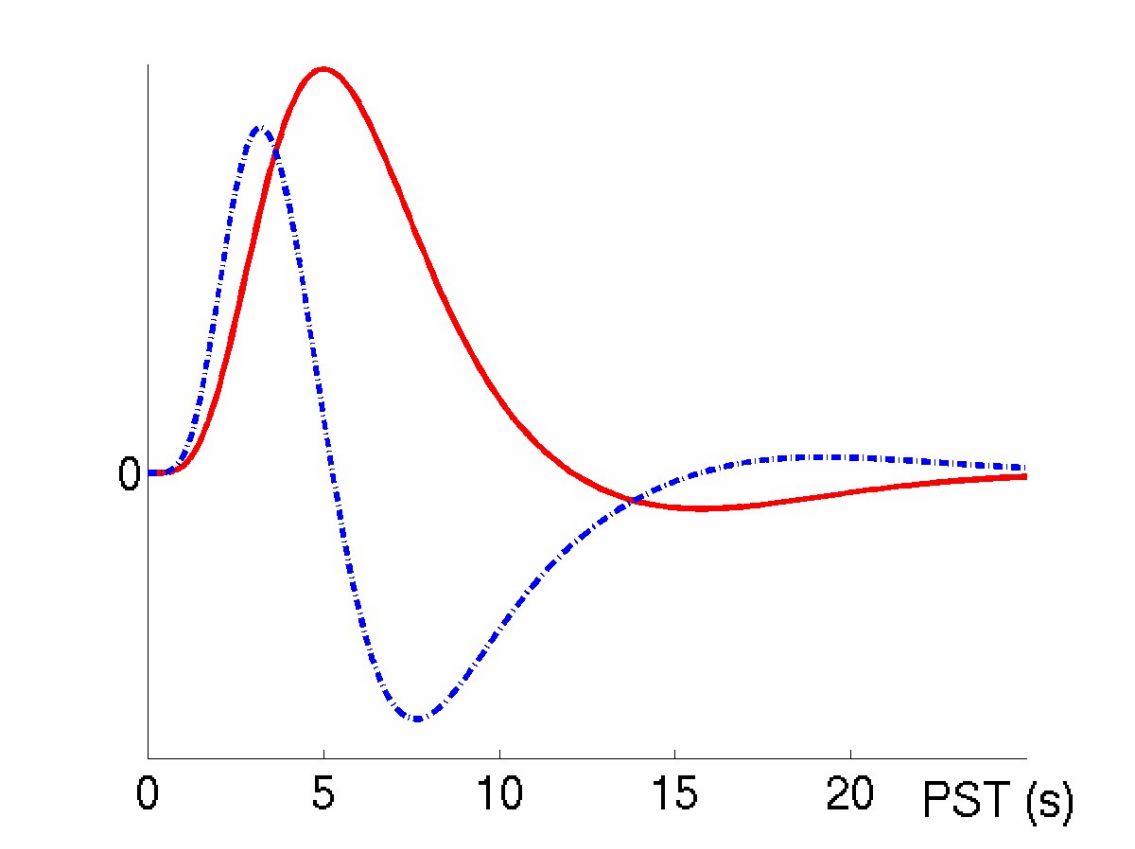
where $\hat{c} = C^T \hat{\beta}$ is the estimated effect size. See Christensen [1] for details.

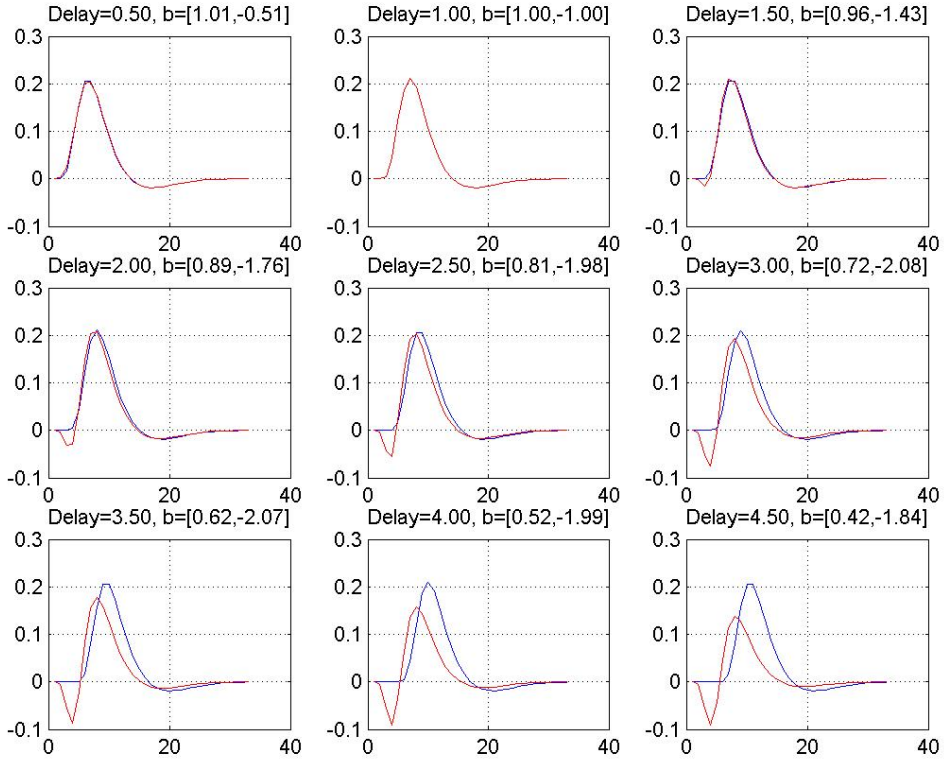
8 Hemodynamic basis functions

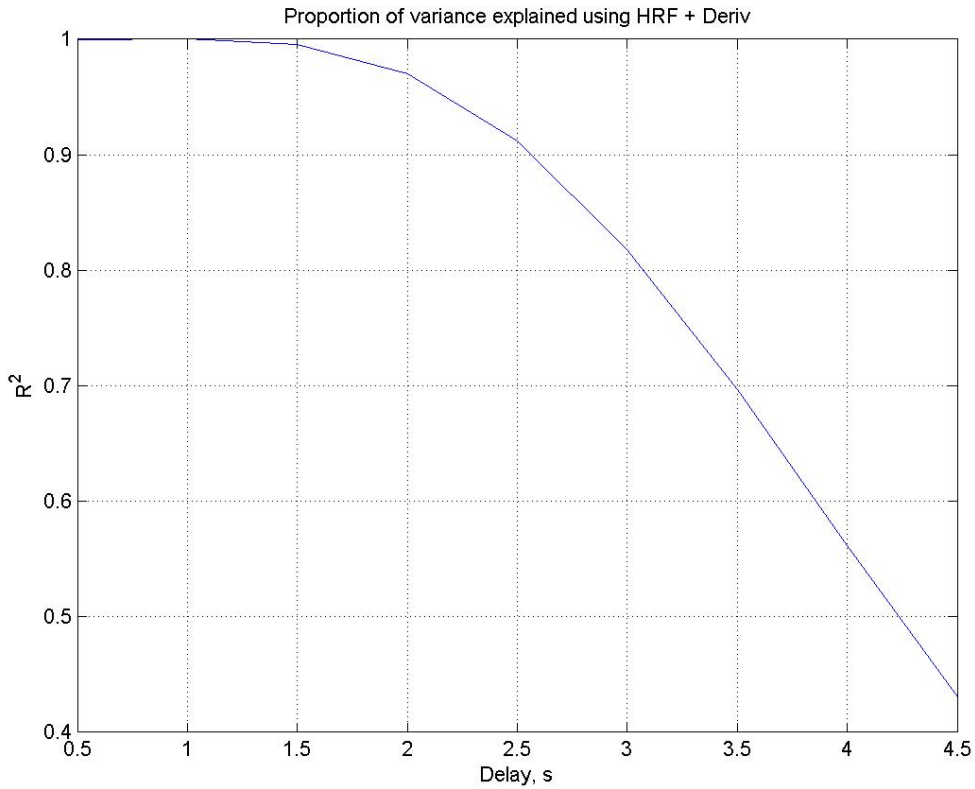
If $C(t,u)$ is the 'Canonical' basis function for event offset u then, using a first-order Taylor series approximation

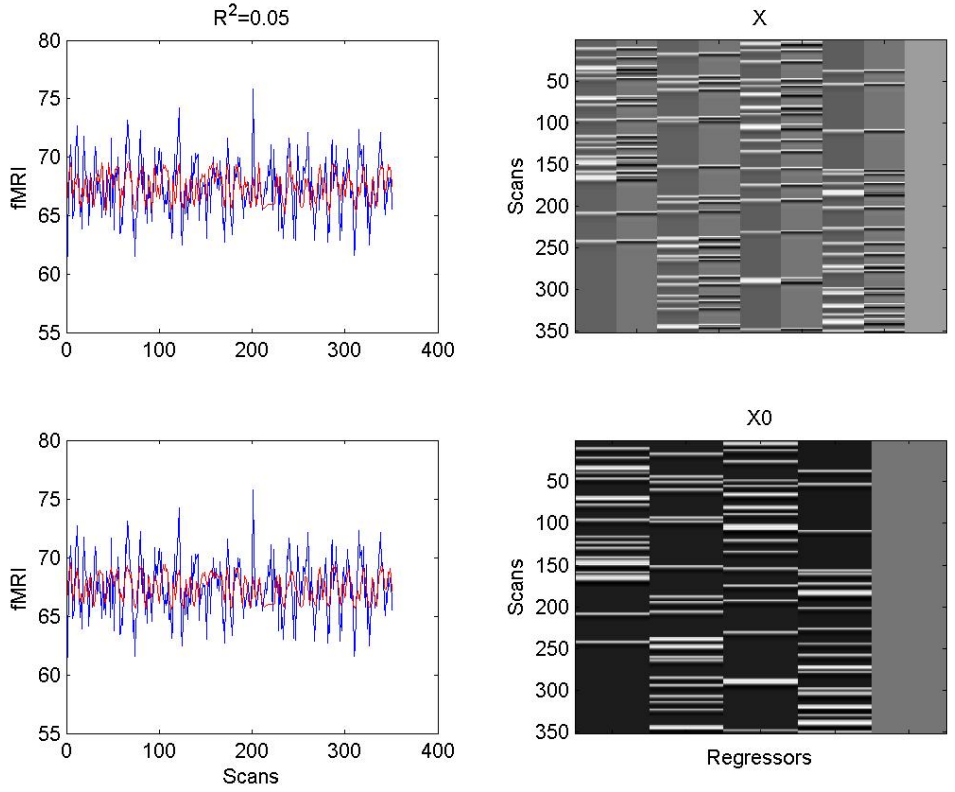
$$\begin{aligned} C(t,u_0+h) &\approx C(t,u_0) + h\frac{dC(t,u)}{du} \\ &\approx C(t,u_0) + hD(t,u_0) \end{aligned} \tag{32}$$

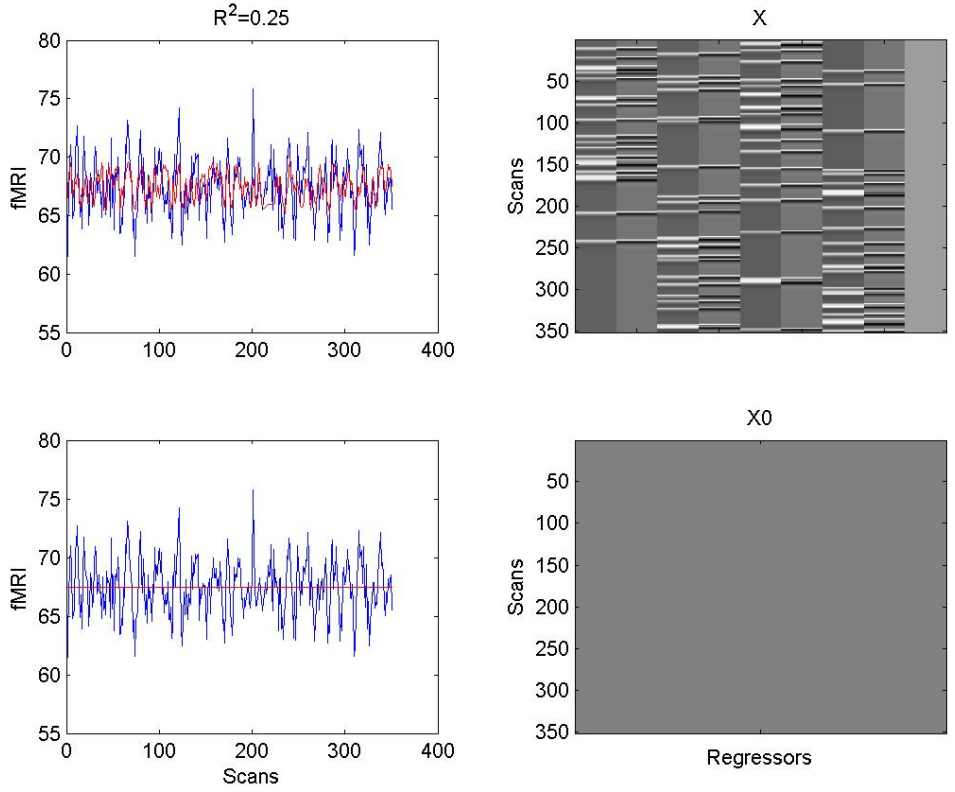
where the derivative is evaluated at $u = u_0$. This will allow us to accomodate small errors in event timings, or earlier/later rises in the hemodynamic response.


0
0
5
10
15
20
PST (s)


Delay=0.50, b=[1.01,-0.51]
Delay=1.00, b=[1.00,-1.00]
Delay=1.50, b=[0.96,-1.43]
Delay=2.00, b=[0.89,-1.76]
Delay=2.50, b=[0.81,-1.98]
Delay=3.00, b=[0.72,-2.08]
Delay=3.50, b=[0.62,-2.07]
Delay=4.00, b=[0.52,-1.99]
Delay=4.50, b=[0.42,-1.84]
0.3
0.2
0.1
0
-0.1
0
20
40


Proportion of variance explained using HRF + Deriv
1
0.9
0.8
0.7
0.6
0.5
0.4
R2
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Delay, s


R2=0.05
X
X0
fMRI
Scans
Regressors
80
75
70
65
60
55
0
100
200
300
400
50
150
250
350


R2=0.25
X
X0
fMRI
Scans
Regressors
80
75
70
65
60
55
0
100
200
300
400
50
150
250
350

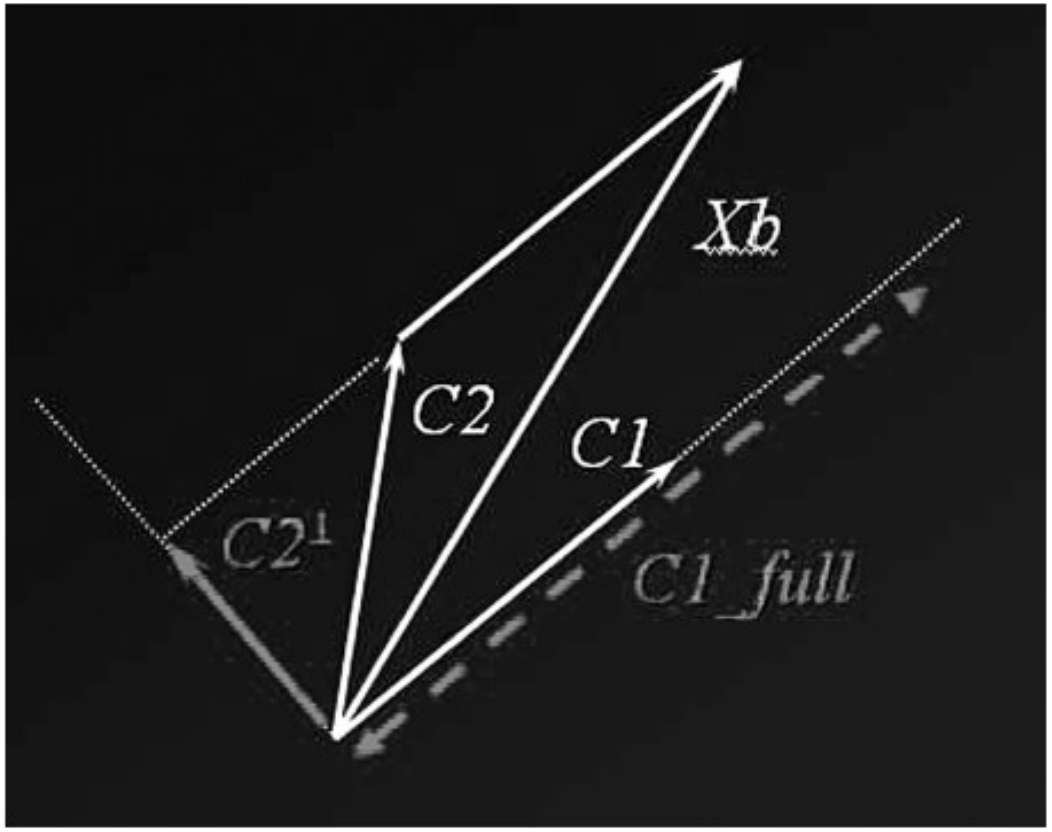


FIGURE 9.16 Hypothesis testing: the geometrical perspective. With a model defined by the two regressors $C1$ and $C2$, testing for $C2$ in effect measures its part orthogonal to C1. If the model is explicitly orthogonalized, (i.e. $C2$ is replaced by $C2^{orth}$), the test of $C2$ is unchanged, but the test of $C1$ is, and will capture more variablity, as indicated by $C1_{full}$.

9 References

[1] R. Christensen. Plane Answers to Complex Questions: The Theory of Linear Models. Springer, New York, 2002.