There are a variety of goals in the statistical analysis of fMRI data.

These include localizing brain areas activated by a task, determining networks corresponding to brain function and making predictions about psychological or disease states.

Often these objectives are related to understanding how the application of certain stimuli leads to changes in neuronal activity.

The most common use of fMRI to date has been in localizing areas of the brain that activate in response to a certain task.

Brain mapping studies are necessary for the development of biological markers and increasing our understanding of brain function.

These types of studies tend to be exploratory in nature.

Typically data analysis is performed by constructing a separate model at each voxel:

- The 'massive univariate approach'.
- Assumes an improbable independence between voxel pairs.

Typically dependencies between voxels are dealt with at a later stage using random field theory, which makes assumptions about the spatial dependencies between voxels.
General Linear Model

The general linear model (GLM) approach treats the data as a linear combination of model functions (predictors) plus noise (error).

- The model functions are assumed to have known shapes, but their amplitudes are unknown and need to be estimated.
- The GLM framework encompasses many of the commonly used techniques in fMRI data analysis (and data analysis more generally).

Illustration

- Consider an experiment of alternating blocks of finger-tapping and rest.
- Construct a model to study data from a single voxel for a single subject.

GLM

A standard GLM can be written:

\[ Y = X\beta + \varepsilon \quad \varepsilon \sim N(0, V) \]

where

\[ Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \]

\( V \) is the covariance matrix whose format depends on the noise model.

Estimation

- If \( \varepsilon \) is i.i.d., then Ordinary Least Square (OLS) estimate is optimal

\[ \hat{\beta} = (X'X)^{-1}X'Y \]

- If \( \text{Var}(\varepsilon) = \sigma^2 I \), then Generalized Least Squares (GLS) estimate is optimal

\[ \hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y \]

Model Refinement

- This model has a number of shortcomings.

  - We want to use our understanding of the signal and noise properties of BOLD fMRI to aid us in constructing appropriate models.
  - This includes deciding on an appropriate design matrix, as well as an appropriate noise model.
1. BOLD responses have a delayed and dispersed form.

2. The fMRI signal includes substantial amounts of low-frequency noise.

3. The data are serially correlated which needs to be considered in the model.

**Issues**

**BOLD Response**

- Predict the shape of the BOLD response to a given stimulus pattern. Assume the shape is known and the amplitude is unknown.

- The relationship between stimuli and the BOLD response is typically modeled using a linear time invariant (LTI) system.

- In an LTI system the impulse (the neuronal activity) is convolved with the impulse response function (the HRF).

**Convolution Examples**

**Multiple Conditions**

**Assumptions:**

- Assume neural activity function is correct
- Assume HRF is correct
- Assume LTI system

**HRF Models**

- Often a fixed canonical HRF is used to model the response to neuronal activity
  - Linear combination of 2 gamma functions.
  - Optimal if correct.
  - If wrong, leads to bias and power loss.
    - Unlikely that the same HRF is valid for all voxels
    - True response may be faster/slower
    - True response may have smaller/bigger undershoot

**Problems**

The HRF shape depends both on the vasculature and the time course of neural activity.

Assuming a fixed HRF is usually not appropriate.
Temporal Basis Functions

• To allow for different types of HRFs in different brain regions, typically it is better to use temporal basis functions.

• A linear combination of functions can be used to account for delays and dispersions in the HRF.
  – The stimulus function is convolved with each of the basis functions to give a set of regressors.
  – The parameter estimates give the coefficients that determine the combination of basis functions that best models the HRF for the trial type and voxel in question.

Temporal Basis Functions

• In an LTI system the BOLD response is modeled
  \[ x(t) = (s * h)(t) \]
  where \( s(t) \) is a stimulus function.

• Let \( f_i(t) \) be a set of temporal basis functions such that
  \[ h(t) = \sum \beta_i f_i(t) \]

Temporal Basis Functions

• The BOLD response can be rewritten:
  \[ x(t) = \sum \beta_i (s * f_i)(t) \]

• In the GLM framework the convolution of the stimulus function with each basis function makes up a separate column of the design matrix.

• Each corresponding \( \beta_i \) describes the weight of that component.

Typically-used models vary in the degree they make a priori assumptions about the shape of the response.

• In the most extreme case, the shape of the HRF is fixed and only the amplitude is allowed to vary.

• By contrast, a finite impulse response (FIR) basis set, contains one free parameter for every time-point following stimulation in every cognitive event type modeled.

Canonical HRF + Derivatives

Including the derivatives allows for a shift in delay and dispersion.

Motivation

• Assume that the actual response, \( y(t) \), is a scaled version of the canonical hrf, \( h(t) \), delayed by a small amount \( dt \), i.e.
  \[ y(t) = cbh(t + dt) \]

• A first order Taylor expansion:
  \[ y(t) = cbh(t) + cbh'(t)dt \]
  \[ = \beta_1 h(t) + \beta_2 h'(t) \]
The model estimates an HRF of arbitrary shape for each event type in each voxel of the brain.

**Finite Impulse Response**

**Basis sets**

**Example**

Suppose we have the following time course (blue) corresponding to a set of stimuli (red).

Estimate the HRF using the FIR basis set.

**Smooth FIR**

- The standard FIR method gives rise to noisy estimates of the HRF.

- One can impose smoothness constraints on $h$ by specifying a Gaussian prior $p(h)$ on the filter parameters.

- The maximum a posteriori estimate of $h$ gives a smoothed version of the HRF.

**Example**

Red – FIR

Blue – Smooth FIR
Simulation Study
• We performed simulations to compare various models ability to handle shifts in onset and duration with respect to bias and power-loss.
• The models we studied were:
  – The canonical HRF
  – The canonical HRF + temporal derivative
  – The canonical HRF + temporal & dispersion derivative
  – The FIR model
  – The Smooth FIR model
  – Inverse Logit model

Inverse Logit Model
• Superposition of three inverse logit (sigmoid) functions.

Simulation
• TR=1, ISI = 30, 10 epochs, 15 “subjects”, Cohen’s d = 0.5
• Estimates of amplitude were obtained and averaged across the 15 subjects.

Results

Parametric Modulation
• Often a stimulus can be parametrically varied across repetitions, and it is thought that this may be reflected in the strength of the neuronal response.
• In these types of situations the parametric modulation can be modeled by including an additional regressor in the design matrix that accounts for the possible variation in neural response.

Parametric Modulators

Parametric modulators are often used to model trial-to-trial variation in a psychological process.
Other Components

• Often model factors associated with known sources of variability, but that are not related to the experimental hypothesis, are also included in the GLM.

• Examples of possible ‘nuisance’ regressors:
  – Physiological (e.g., respiration) artifacts
  – Head motion, e.g. six regressors comprising of three translations and three rotations.
    • Sometimes transformations of the six regressors also included.

Illustration

fMRI Temporal Autocorrelation

Characteristics:
– “1/f” in frequency domain
– Nearby time-points exhibit positive correlation

Implications
1. There is a slow drift in the signal that will add noise to your estimates unless removed.
2. The independence assumption is violated in fMRI data. Single subject statistics are not valid without an accurate model of the noise!
3. We need to include drift and autocorrelated noise in our model.

Modeling Drift

• Drift is a nuisance source of variability.

• We can model it using any slowly varying curve
  – Splines
  – Linear, quadratic functions
  – Discrete Cosine Transform

Model with Drift

\[ Y = X \times \beta + \epsilon \]
**High Pass Filtering**

**fMRI Noise**

- The noise in fMRI is typically modeled using either an AR(p) or an ARMA(1,1) process.

- The autocorrelation is considered to be due to physiological noise and low frequency drift, that has not been appropriately modeled.

- After removal of these terms there is evidence that the resulting error term is white noise (Lund et al. (2006)).

**AR(1) model**

- Often serial correlation is modeled using a first-order autoregressive model, i.e.

  \[ \epsilon_t = \phi \epsilon_{t-1} + u_t \quad u_t \sim WN(0, \sigma) \]

- The error term \( \epsilon_t \) depends on the previous error term \( \epsilon_{t-1} \) and a new disturbance term \( u_t \).

**Properties**

- The autocorrelation function for an AR(1) process is given by

  \[ \rho(h) = \begin{cases} 1 & \text{if } h = 0 \\ \phi^h & \text{if } h \neq 0 \end{cases} \]

**Error Term**

- The format of \( V \) will depend on what noise model is used.

  **IID Case**
  
  \[
  V = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\
  0 & 1 & 0 & \ldots & 0 \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & 0 & \ldots & 1 \end{bmatrix}
  \]

  **AR(1) Case**
  
  \[
  V = \begin{bmatrix} 1 & \phi & \phi^2 & \ldots & \phi^{t-1} \\
  \phi & 1 & \phi & \ldots & \phi^{t-2} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  \phi^{t-1} & \phi^{t-2} & \phi^{t-3} & \ldots & 1 \end{bmatrix}
  \]

**Estimation**

- If \( \epsilon \) is i.i.d., then Ordinary Least Square (OLS) is optimal

  \[
  Y = X\beta + \epsilon \quad \hat{\beta} = (X'X)^{-1}X'Y
  \]

- If \( \text{Var}(\epsilon) = \sigma^2 I \), then Generalized Least Squares (GLS) is optimal

  \[
  Y = X\beta + \epsilon \quad \hat{\beta} = (XV^{-1}X')^{-1}XV^{-1}Y
  \]
Estimating \( V \)

- In general the form of the covariance matrix is unknown, which means it has to be estimated.
- Estimating \( V \) depends on \( \beta \)'s, and estimating \( \beta \)'s depends on \( V \). Need iterative procedure.
- Methods for estimating variance components:
  - Method of moments
  - Maximum likelihood
  - Restricted maximum likelihood

Iterative Procedure

1. Assume that \( V = I \) and calculate the OLS solution.
2. Estimate the parameters of \( V \) using the residuals.
3. Re-estimate the \( \beta \) values using the estimated covariance matrix \( \hat{V} \) from step 2.
4. Iterate until convergence.

Yule-Walker Estimates

Assume \( \{ \varepsilon \}_t \) is an AR(1) process.

\[ \varepsilon_t = \phi \varepsilon_{t-1} + u_t \quad t = 0, 1, \ldots \]

where \( \{ u \}_t \sim WN(0, \sigma^2) \)

The Yule-Walker estimates are:

\[ \hat{\phi} = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} \]

Auto Covariance Function

\[ \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\phi}^2(1) \]

ReML

- Restricted maximum likelihood (ReML) requires maximizing the restricted log-likelihood given by

\[ l'(\lambda) = -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} \log |\mathbf{X}' \mathbf{V} \mathbf{X}| - \frac{1}{2} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \hat{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \hat{\beta}) \]

Extra ReML variance term

where \( \lambda \) are parameters associated with \( V \).

ML vs ReML

- Maximum Likelihood
  - Maximize likelihood of data \( y \)
  - Used to estimate "mean" parameters \( \beta \)
  - But can produce biased estimates of variance

\[ \hat{\sigma}^2_{ML} = \frac{1}{n} \sum (y_i - \hat{\gamma})^2 \]

- Restricted Maximum Likelihood
  - Maximize likelihood of residuals \( e = y - X \hat{\beta} \)
  - Used to estimate variance parameters
  - Provides unbiased estimates

\[ \hat{\sigma}^2_{ReML} = \frac{1}{n-1} \sum (e_i - \hat{\gamma})^2 \]

SPM Approach

- In SPM the correlation matrix is expressed as a linear combination of known matrices,

\[ \mathbf{V} = \sum \lambda \mathbf{Q} \]

where \( \mathbf{Q} \) represents some basis set for the covariance matrices.

- In single-subject fMRI studies \( V \) is typically expressed as the linear combination of two matrices.
Approximate AR(1):
Taylor expansion: \( \text{Cor}(\varepsilon_i, \varepsilon_j) \approx \rho_0 |i-j| + |i-j| \rho_0 |i-j|^{-1} (\rho_0 - \rho) \)
- Assume fixed scalar \( \rho_0 \)
- \( \rho_0 \) is scaled linearly by \( \lambda \)
- This approximation is now linear in \( \rho \)

\[ V = \sum \lambda_i Q_i \]

Details
- Only most significant voxels used to estimate \( V \) (F-stat \( p < 0.001 \)).
- Once \( V \) estimated, same estimate used over whole brain.
  - Assumes same autocorrelation over brain.
- Requires two passes
  - 1st pass: OLS, est. \( V \) via REML at selected voxels
  - 2nd pass: GLS using \( V \).

FSL Approach
- In contrast, FSL estimates the autocorrelation non-parametrically at each lag.
- Since estimates at high lags are based on relatively few points, and are extremely variable, they are down-weighted using a Tukey taper.
- The autocorrelation is estimated locally but smoothed across voxels.

Procedure
1. For each voxel estimate raw autocorrelations
   \[ \hat{\rho}(\tau) = \frac{1}{\sigma^2} \sum_{t=1}^{N} \varepsilon(t) \varepsilon(t-\tau)/(N-\tau) \]
2. Within each voxel, smooth the correlation estimates using a Tukey taper.
3. Spatially smooth the estimates across voxels.

Example

Inference
- Once we fit our model we use the estimated parameters to determine whether there is significant activation present in the voxel.
- Inference is based on the fact that:
  \[ \hat{\beta} \sim N(\beta, (X'V^{-1}X)^{-1}) \]
- Use t and F procedures to perform tests on effects of interest.
Contrasts

• It is often of interest to see whether a linear combination of the parameters are significant.

• The term $c\beta$ specifies a linear combination of the estimated parameters, i.e.

$$c^T\beta = c_1\beta_1 + c_2\beta_2 + \ldots + c_n\beta_n$$

• In fMRI terminology, $c$ is called a contrast vector.

Example

Event-related experiment with two types of stimuli.

$$H_0 : c^T\beta = 0$$

$$H_a : c^T\beta \neq 0$$

T-test

• To test

$$H_0 : c^T\beta = 0 \quad H_a : c^T\beta \neq 0$$

use the $t$-statistic:

$$T = \frac{c^T\hat{\beta}}{\sqrt{\text{Var}(c^T\hat{\beta})}}$$

• Under $H_0$, $T$ is approximately $t(\nu)$ with

$$\nu = \frac{\text{var}(c^T\epsilon)}{\text{var}(c^T\hat{\beta})}$$

Multiple Contrasts

• We are often interested in making a simultaneous test of several contrasts at once.

• $c$ is now a contrast matrix.

• Suppose

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

then

$$c^T\hat{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Example

Recall the model with box-car shaped activation and drift modeled using the discrete cosine basis.

$$Y = X\beta + \epsilon$$

Example

Do the drift components add anything to the model?

Test: $H_a : c^T\hat{\beta} = 0$

where

$$c = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$
This is equivalent to testing:
\[ H_0 : (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^T = 0 \]
To understand what this implies, we split the design matrix into two parts:
\[
\begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1n} \\
  x_{21} & x_{22} & \cdots & x_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{i1} & x_{i2} & \cdots & x_{in}
\end{bmatrix}
\]

F-test

- Test the hypothesis using the F-statistic:
  \[
  F = \frac{\left( e^T R e \right)}{\sigma^2 (e^T (R - R_0) V)^T}
  \]
- Assuming the errors are normally distributed, F has an approximate F-distribution with \((\nu_0, \nu)\) degrees of freedom, where
  \[
  \nu_0 = \frac{\nu (R - R_0)^T V}{\nu (R V)^2}
  \]
  and \(\nu = \frac{\nu (R V)^2}{\nu (R^2)^2}\)

Statistical Images

How do we determine which voxels are actually active?

Problems:
- The statistics are obtained by performing a large number of hypothesis tests.
- Many of the test statistics will be artificially inflated due to the noise.
- This leads to many false positives.

Statistical Images

Multiple Comparisons

- Which of 100,000 voxels are significant?
  \[ - \alpha=0.05 \Rightarrow 5,000 \text{ false positive voxels} \]
- Choosing a threshold is a balance between sensitivity (true positive rate) and specificity (true negative rate).
  \[
  x_{i1} x_{i2} \cdots x_{in}
  \]

Example

- Do the drift components add anything to the model?
- The \(X_1\) matrix explains the drift. Does it contribute in a significant way to the model?
- Compare the results using the full model, with design matrix \(X\), with those obtained using a reduced model, with design matrix \(X_0\).

Example

- For each voxel a hypothesis test is performed. The statistic corresponding to that test is used to create a statistical image over all voxels.