The General Linear Model

Guillaume Flandin
Wellcome Trust Centre for Neuroimaging
University College London
Image time-series

Realignement → Smoothing → Design matrix → General Linear Model → Parameter estimates → Statistical Inference → Statistical Parametric Map

Spatial filter

Normalisation

Anatomical reference

Realignment

Smoothing

Design matrix

General Linear Model

Parameter estimates

Statistical Inference

Statistical Parametric Map

RFT

p < 0.05
A very simple fMRI experiment

One session

Passive word listening versus rest

7 cycles of rest and listening

Blocks of 6 scans with 7 sec TR

Question: Is there a change in the BOLD response between listening and rest?
Modelling the measured data

Why?

Make inferences about effects of interest

1. Decompose data into effects and error

How?

2. Form statistic using estimates of effects and error
Voxel-wise time series analysis

Model specification
Parameter estimation
Hypothesis
Statistic

BOLD signal

single voxel time series
Single voxel regression model

\[ y = x_1 \beta_1 + x_2 \beta_2 + e \]
Mass-univariate analysis: voxel-wise GLM

\[ y = X\beta + e \]

\[ e \sim N(0, \sigma^2 I) \]

Model is specified by:
1. Design matrix \( X \)
2. Assumptions about \( e \)

\( N: \) number of scans  
\( p: \) number of regressors

The design matrix embodies all available knowledge about experimentally controlled factors and potential confounds.
GLM: mass-univariate parametric analysis

- one sample \( t \)-test
- two sample \( t \)-test
- paired \( t \)-test
- Analysis of Variance (ANOVA)
- Factorial designs
- correlation
- linear regression
- multiple regression
- \( F \)-tests
- fMRI time series models
- Etc..
Parameter estimation

\[ y = X\beta + e \]

Objective: estimate parameters to minimize

\[ \sum_{t=1}^{N} e_t^2 \]

Ordinary least squares estimation (OLS) (assuming i.i.d. error):

\[ \hat{\beta} = (X^T X)^{-1} X^T y \]
A geometric perspective on the GLM

Design space defined by $X$

Smallest errors (shortest error vector) when $e$ is orthogonal to $X$

\[
X^T e = 0
\]

\[
X^T (y - X\hat{\beta}) = 0
\]

\[
X^T y = X^T X\hat{\beta}
\]

\[
\hat{\beta} = (X^T X)^{-1} X^T y
\]

Ordinary Least Squares (OLS)
Example: Two sample T-test

Compare the average difference between two groups
Example: Two sample T-test

\[
\begin{pmatrix}
\hat{\beta}_1 \\
\hat{\beta}_2 \\
\end{pmatrix} = (X^T X)^{-1} X^T Y = 
\begin{pmatrix}
\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\end{pmatrix}
\begin{pmatrix}
A_1 \\
A_2 \\
A_3 \\
A_4 \\
B_1 \\
B_2 \\
B_3 \\
B_4 \\
\end{pmatrix}
\]

Compare the average difference between two groups
What are the problems of this model?

1. BOLD responses have a delayed and dispersed form.

2. The BOLD signal includes substantial amounts of low-frequency noise (eg due to scanner drift).

3. Due to breathing, heartbeat & unmodeled neuronal activity, the errors are serially correlated. This violates the assumptions of the noise model in the GLM.
Problem 1: Shape of BOLD response
Solution: Convolution model

$$f \otimes g(t) = \int_0^t f(\tau)g(t-\tau)d\tau$$

- expected BOLD response
- = input function \( \otimes \) impulse response function (HRF)
Convolution model of the BOLD response

Convolve stimulus function with a canonical hemodynamic response function (HRF):

$$ f \otimes g(t) = \int_{0}^{t} f(\tau) g(t - \tau) d\tau $$

![Graph showing the convolution process and intensity over scans]
Problem 2: Low-frequency noise
Solution: High pass filtering

blue = data
black = mean + low-frequency drift
green = predicted response, taking into account low-frequency drift
red = predicted response, NOT taking account low-frequency drift

discrete cosine transform (DCT) set
High pass filtering

Frequency domain
128 second High-pass filter

relative spectral density

0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45
0 0.02 0.04 0.06

Frequency (Hz)

discrete cosine transform (DCT) set
Problem 3: Serial correlations

\[ e_t = ae_{t-1} + \varepsilon_t \text{ with } \varepsilon_t \sim N(0, \sigma^2) \]

1st order autoregressive process: AR(1)
Multiple covariance components

\[ e_i \sim N(0, C_i) \]

enhanced noise model at voxel i

\[ C_i = \sigma_i^2 V \]

\[ V = \sum \lambda_j Q_j \]

error covariance components \( Q \) and hyperparameters \( \lambda \)

Estimation of hyperparameters \( \lambda \) with ReML (Restricted Maximum Likelihood).
Summary

- Mass univariate approach.

- Fit GLMs with design matrix, $X$, to data at different points in space to estimate local effect sizes, $\beta$

- GLM is a very general approach

- Hemodynamic Response Function

- High pass filtering

- Temporal autocorrelation
Statistical Inference

Guillaume Flandin
Wellcome Trust Centre for Neuroimaging
University College London
Image time-series → Spatial filter → Design matrix → Statistical Parametric Map

Realignment → Smoothing → General Linear Model → Statistical Inference

Normalisation → Anatomical reference

Parameter estimates → p < 0.05

RFT
A mass-univariate approach
Estimation of the parameters

i.i.d. assumptions: $\varepsilon \sim N(0, \sigma^2 I)$

OLS estimates: $\hat{\beta} = (X^T X)^{-1} X^T y$

$\hat{\beta}_1 = 3.9831$

$\hat{\beta}_2 = 0.6871, 1.9598, 1.3902, 166.1007, 76.4770, -64.8189$

$\hat{\beta}_8 = 131.0040$

$\hat{\varepsilon} = 
\hat{\varepsilon} \sim N(\beta, \sigma^2 (X^T X)^{-1})$

$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{N-p}$
Contrasts

- A contrast selects a specific effect of interest.
  - A contrast \( c \) is a vector of length \( p \).
  - \( c^T \beta \) is a linear combination of regression coefficients \( \beta \).

\[
c = [1 \ 0 \ 0 \ 0 \ ...]^T
\]

\[
c^T \beta = 1 \times \beta_1 + 0 \times \beta_2 + 0 \times \beta_3 + 0 \times \beta_4 + \cdots
= \beta_1
\]

\[
c = [0 \ 1 \ -1 \ 0 \ ...]^T
\]

\[
c^T \beta = 0 \times \beta_1 + 1 \times \beta_2 + -1 \times \beta_3 + 0 \times \beta_4 + \cdots
= \beta_2 - \beta_3
\]

\[
c^T \hat{\beta} \sim N(c^T \beta, \sigma^2 c^T (X^T X)^{-1} c)
\]
Hypothesis Testing

To test an hypothesis, we construct “test statistics”.

- **Null Hypothesis** $H_0$
  Typically what we want to disprove (no effect).
  ⇒ The Alternative Hypothesis $H_A$ expresses outcome of interest.

- **Test Statistic** $T$
  The test statistic summarises evidence about $H_0$.
  Typically, test statistic is small in magnitude when the hypothesis $H_0$ is true and large when false.
  ⇒ We need to know the distribution of $T$ under the null hypothesis.
Hypothesis Testing

- **Significance level $\alpha$:**
  Acceptable *false positive rate* $\alpha$.
  $\Rightarrow$ threshold $u_\alpha$

  Threshold $u_\alpha$ controls the false positive rate
  $$\alpha = p(T > u_\alpha \mid H_0)$$

- **Conclusion about the hypothesis:**
  We reject the null hypothesis in favour of the alternative hypothesis if $t > u_\alpha$

- **$p$-value:**
  A *$p$-value* summarises evidence against $H_0$.
  This is the chance of observing value more extreme than $t$ under the null hypothesis.

$$p(T > t \mid H_0)$$
**T-test - one dimensional contrasts – SPM\{t\}**

$ c^T = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 $  

Question: box-car amplitude > 0 ?  

$ \beta_1 = c^T \beta > 0 $  

Null hypothesis:  

$ H_0: c^T \beta = 0 $  

Test statistic:  

$ T = \frac{c^T \hat{\beta}}{\sqrt{\text{var}(c^T \hat{\beta})}} = \frac{c^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}} \sim t_{N-p} $
**T-contrast in SPM**

- For a given contrast \( c \):

\[
\hat{\beta} = (X^T X)^{-1} X^T y
\]

\( \hat{\beta} \) images

\[
\hat{\sigma}^2 = \frac{\hat{\mathcal{E}}^T \hat{\mathcal{E}}}{N - p}
\]

ResMS image

\( c^T \hat{\beta} \)

con_????? image

\( \text{spmT_????? image} \)

SPM\{t\}
**T-test:** a simple example

- Passive word listening versus rest

**c^T = [1 0 0 0 0 0 0 0]**

**Q:** activation during listening?

**Null hypothesis:** $\beta_1 = 0$

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

**SPM results:**
Height threshold $T = 3.2057 \ \{p<0.001\}$

<table>
<thead>
<tr>
<th>$T$</th>
<th>$Z$</th>
<th>$p$</th>
<th>mm</th>
<th>mm</th>
<th>mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.94</td>
<td>Inf</td>
<td>0.000</td>
<td>-63</td>
<td>-27</td>
<td>15</td>
</tr>
<tr>
<td>12.04</td>
<td>Inf</td>
<td>0.000</td>
<td>-48</td>
<td>-33</td>
<td>12</td>
</tr>
<tr>
<td>11.82</td>
<td>Inf</td>
<td>0.000</td>
<td>-66</td>
<td>-21</td>
<td>6</td>
</tr>
<tr>
<td>13.72</td>
<td>Inf</td>
<td>0.000</td>
<td>57</td>
<td>-21</td>
<td>12</td>
</tr>
<tr>
<td>12.29</td>
<td>Inf</td>
<td>0.000</td>
<td>63</td>
<td>-12</td>
<td>-3</td>
</tr>
<tr>
<td>9.89</td>
<td>7.83</td>
<td>0.000</td>
<td>57</td>
<td>-39</td>
<td>6</td>
</tr>
<tr>
<td>7.39</td>
<td>6.36</td>
<td>0.000</td>
<td>36</td>
<td>-30</td>
<td>-15</td>
</tr>
<tr>
<td>6.84</td>
<td>5.99</td>
<td>0.000</td>
<td>51</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td>6.36</td>
<td>5.65</td>
<td>0.000</td>
<td>-63</td>
<td>-54</td>
<td>-3</td>
</tr>
<tr>
<td>6.19</td>
<td>5.53</td>
<td>0.000</td>
<td>-30</td>
<td>-33</td>
<td>-18</td>
</tr>
<tr>
<td>5.96</td>
<td>5.36</td>
<td>0.000</td>
<td>36</td>
<td>-27</td>
<td>9</td>
</tr>
<tr>
<td>5.84</td>
<td>5.27</td>
<td>0.000</td>
<td>-45</td>
<td>42</td>
<td>9</td>
</tr>
<tr>
<td>5.44</td>
<td>4.97</td>
<td>0.000</td>
<td>48</td>
<td>27</td>
<td>24</td>
</tr>
<tr>
<td>5.32</td>
<td>4.87</td>
<td>0.000</td>
<td>36</td>
<td>-27</td>
<td>42</td>
</tr>
</tbody>
</table>
**T-test**: summary

- **T-test** is a *signal-to-noise* measure (ratio of estimate to standard deviation of estimate).

- Alternative hypothesis:

  \[
  H_0: \quad c^T \beta = 0 \quad \text{vs} \quad H_A: \quad c^T \beta > 0
  \]

- **T-contrasts** are simple combinations of the betas; the T-statistic does not depend on the scaling of the regressors or the scaling of the contrast.
**F-test** - the extra-sum-of-squares principle

- **Model comparison:**

  \[
  \text{Null Hypothesis } H_0: \text{ True model is } X_0 \text{ (reduced model)}
  \]

  **Test statistic:** ratio of explained variability and unexplained variability (error)

  \[
  F \propto \frac{RSS_0 - RSS}{RSS} \quad \text{or} \quad F \propto \frac{ESS}{RSS} \sim F_{v_1,v_2}
  \]

  \[
  v_1 = \text{rank}(X) - \text{rank}(X_0) \quad v_2 = N - \text{rank}(X)
  \]
**F-test** - multidimensional contrasts – SPM\{F\}

- Tests multiple linear hypotheses:

H₀: True model is \( X_0 \)

\( H_0: \beta_4 = \beta_5 = \ldots = \beta_9 = 0 \)

**test \( H_0: \ c^T \beta = 0 \)?**

\[
\begin{align*}
X_0 & \quad \ | \quad \ X_1 (\beta_{4-9}) & \quad \ | \quad \ X_0 \\
\begin{array}{c}
\begin{array}{ccccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
\end{array}
\end{array}
\end{align*}
\]

Full model? Reduced model?
F-contrast in SPM

\[ \hat{\beta} = (X^T X)^{-1} X^T y \]

ResMS image
\[ \hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{N - p} \]

ess_???? images
\( (RSS_0 - RSS) \)

spmF_???? images

SPM{F}
**F-test** example: movement related effects

- Contrast(s)

- Design matrix

[Images of brain scans and statistical matrices]

- Contrast(s)

- Design matrix
**F-test: summary**

- F-tests can be viewed as testing for the additional variance explained by a larger model wrt a simpler (nested) model \( \Rightarrow \) **model comparison**.

- F tests a weighted **sum of squares** of one or several combinations of the regression coefficients \( \beta \).

- In practice, we don’t have to explicitly separate \( X \) into \([X_1X_2]\) thanks to **multidimensional contrasts**.

- Hypotheses:

  \[
  \begin{bmatrix}
  1 & 0 & 0 & 0 \\
  0 & 1 & 0 & 0 \\
  0 & 0 & 1 & 0 \\
  0 & 0 & 0 & 0 \\
  \end{bmatrix}
  \]

  Null Hypothesis \( H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \)

  Alternative Hypothesis \( H_A : \) at least one \( \beta_k \neq 0 \)

- In testing uni-dimensional contrast with an \( F \)-test, for example \( \beta_1 - \beta_2 \), the result will be the same as testing \( \beta_2 - \beta_1 \). It will be exactly the square of the \( t \)-test, testing for both positive and negative effects.
Orthogonal regressors

Variability described by $X_1$

Testing for $X_1$

Variability described by $X_2$

Testing for $X_2$

Variability in $Y$
Correlated regressors

Variability described by $X_1$

Shared variance

Variability described by $X_2$

Variability in $Y$
Correlated regressors

Variability described by $X_1$

Testing for $X_1$

Variability described by $X_2$

Variability in $Y$
Correlated regressors

Testing for $X_2$

Variability described by $X_1$

Variability described by $X_2$

Variability in $Y$
Correlated regressors

Variability described by $X_1$

Variability described by $X_2$

Variability in $Y$
Correlated regressors

Testing for $X_1$

Variability described by $X_1$

Variability described by $X_2$

Variability in $Y$
Correlated regressors

Variability described by $X_1$

Variability described by $X_2$

Testing for $X_2$

Variability in $Y$
Correlated regressors

Testing for $X_1$ and/or $X_2$

Variability described by $X_1$

Variability described by $X_2$

Variability in $Y$
For each pair of columns of the design matrix, the orthogonality matrix depicts the magnitude of the cosine of the angle between them, with the range 0 to 1 mapped from white to black.

If both vectors have zero mean then the cosine of the angle between the vectors is the same as the correlation between the two variates.

**Measure**: abs. value of cosine of angle between columns of design matrix

**Scale**
- black - colinear (cos=+1/-1)
- white - orthogonal (cos=0)
- gray - not orthogonal or colinear
Topological Inference

Guillaume Flandin
Wellcome Trust Centre for Neuroimaging
University College London
Pre-processings

General Linear Model

Statistical Inference

\[ y = \beta + \varepsilon \]

\[ \varepsilon \sim N(0, \sigma^2 I) \]

\[ \hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\text{rank}(X)} \]

Random Field Theory
Inference at a single voxel

Null distribution of test statistic $T$

Decision rule (threshold) $u$: determines false positive rate $\alpha$

$\Rightarrow$ Choose $u$ to give acceptable $\alpha$ under $H_0$

$\hat{\beta}_1 = 3.9831$

Null Hypothesis $H_0$: zero activation
Multiple tests

If we have 100,000 voxels, \(\alpha=0.05\ \Rightarrow 5,000\) false positive voxels.

This is clearly undesirable; to correct for this we can define a null hypothesis for a collection of tests.
Multiple tests

If we have 100,000 voxels, $\alpha=0.05 \Rightarrow 5,000$ false positive voxels.

This is clearly undesirable; to correct for this we can define a null hypothesis for a collection of tests.

Use of ‘uncorrected’ $p$-value, $\alpha =0.1$

Percentage of Null Pixels that are False Positives
Family-Wise Null Hypothesis

*Family-Wise Null Hypothesis:* 
*Activation is zero everywhere*

If we reject a voxel null hypothesis at *any* voxel, we reject the family-wise Null hypothesis.

A FP *anywhere* in the image gives a *Family Wise Error* (FWE).

Family-Wise Error rate (FWER) = ‘*corrected*’ $p$-value

Use of ‘uncorrected’ $p$-value, $\alpha = 0.1$

Use of ‘corrected’ $p$-value, $\alpha = 0.1$
Bonferroni correction

The Family-Wise Error rate (FWER), $\alpha_{FWE}$, for a family of $N$ tests follows the inequality:

$$\alpha_{FWE} \leq N\alpha$$

where $\alpha$ is the test-wise error rate.

Therefore, to ensure a particular FWER choose:

$$\alpha = \frac{\alpha_{FWE}}{N}$$

This correction does not require the tests to be independent but becomes very stringent if dependence.
Spatial correlations

100 x 100 independent tests

Spatially correlated tests (FWHM=10)

Discrete data

Spatially extended data

Bonferroni is too conservative for spatial correlated data.

10,000 voxels $\Rightarrow \alpha_{BONF} = \frac{0.05}{10,000} \Rightarrow u_c = 4.42$ (uncorrected $u = 1.64$)
Topological inference

Topological feature: Peak height

- Significant local maxima
- Non-significant local maxima

$u_\alpha$
Topological inference

Topological feature: Cluster extent

$u_{\text{clus}}$: cluster-forming threshold (arbitrary)

$k_\alpha$: $\alpha$-level extent threshold

significant cluster

non significant clusters
Here, $c=1$, only one cluster larger than $k$. 

Topological inference
RFT and Euler Characteristic

Variability described by $X_1$

- Search volume
- Roughness (1/smoothness)
- Threshold
Expected Euler Characteristic

Testing for $X_1$

100 x 100 Gaussian Random Field with FWHM=10 smoothing

$\alpha_{FWE} = 0.05 \Rightarrow u_{RFT} = 3.8$

($u_{BONF} = 4.42, u_{uncorr} = 1.64$)
Random Field Theory

- The statistic image is assumed to be a good lattice representation of an underlying continuous stationary random field. Typically, FWHM > 3 voxels (combination of intrinsic and extrinsic smoothing).

- Smoothness of the data is unknown and estimated: very precise estimate by pooling over voxels \( \Rightarrow \) stationarity assumptions (esp. relevant for cluster size results).

- RFT conservative for low degrees of freedom (always compare with Bonferroni correction). Afford littles power for group studies with small sample size.

- *A priori* hypothesis about where an activation should be, reduce search volume \( \Rightarrow \) Small Volume Correction:
  - mask defined by (probabilistic) anatomical atlases
  - mask defined by separate "functional localisers"
  - mask defined by orthogonal contrasts
  - (spherical) search volume around previously reported coordinates
Conclusion

- There is a **multiple testing problem** and **corrections** have to be applied on $p$-values (for the volume of interest only (see SVC)).

- Inference is made about **topological features** (peak height, spatial extent, number of clusters). Use results from the **Random Field Theory**.

- **Control of FWER** (probability of a false positive anywhere in the image): very specific, not so sensitive.

- **Control of FDR** (expected proportion of false positives amongst those features declared positive (the discoveries)): less specific, more sensitive.
References


http://www.fil.ion.ucl.ac.uk/spm/doc/biblio/Keyword/RFT.html